Compressing Positive Semidefinite Operators with Sparse/Localized Bases

Thesis by
Pengchuan Zhang

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
Defended May 17, 2017

Pengchuan Zhang
ORCID: 0000-0003-1155-9507

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my adviser, Prof. Thomas Y. Hou, for his continuous support, guidance, and encouragement. Prof. Hou guided me to the interdisciplinary research between PDE analysis and data science, and I have truly enjoyed my research in this area. Prof. Hou gave me the right amount of freedom to explore, which really helped me find my interests and trained my ability to do independent research. He also provided insightful advice and enormous help when I got stuck or lost. Moreover, Prof. Hou's great personality has deeply influenced me and continuously drives me to be a better person, both in research and in life.

I want to thank Prof. Houman Owhadi. I took several of his courses, worked with him as TA, got inspired by his research, and got his instructions on research. I would like to thank the rest of thesis committee – Profs. James L. Beck and Andrew Stuart – for taking time to review my thesis. I also want to thank Prof. Qin Li. She helped me quickly get into research mode in my second year at Caltech. She is a great mentor, collaborator, and friend.

I want to thank everyone in our CMS department. I had lots of stimulating discussions with many great mathematicians here, especially Prof. Joel Tropp, Prof. Venkat Chandrasekaran, Prof. Zhiwen Zhang from HKU, Prof. Lei Zhang from SJTU, Dr. Pengfei Liu, Mr. De Huang, Dr. Ka Chun Lam, and more. Their knowledge truly inspired me! I would also like to thank the staff of CMS department – Carmen Nemer-Sirois, Sydney Gastang, and Maria Lopez – for their help over the past years.

I would like to thank Prof. Dr. XiaohuiWu and Dr. Mulin Cheng for their supports when I was at ExxonMobil. I would like to thank Dr. Alex Deng and Dr. Jiannan Lu for their supports when I was at Microsoft A&E team. In addition, I want to thank Prof. Li Deng and Prof. Xiaodong He for their supports when I was working on this thesis in Deep Learning Technology Center at Microsoft Research.

I thank my friends at Caltech – Peng He, Pengfei Sui, Yichen Huang, Yong Sheng Soh, Yanan Sui, Xiaoqi Ren, Armeen Taeb, Lingwen Gan, Niangjun Chen, Yorie Nakahira, and more– for the help, the fun, and the good time we had together.

Most importantly, I would like to thank my beloved parents, Mr. Jiahua Zhang and Mrs. Suzhen Zou, for their endless love and support. No words can express how grateful I am for you and how much I love you.

Last but not the least, my thanks to my lovely fiancée, Suiqing Bao. You have supported me all along the way and created lots of happiness for me. Without you, I could not have survived this strenuous Ph.D. journey. This thesis is dedicated to you and my parents with all my heart.

ABSTRACT

Given a positive semidefinite (PSD) operator, such as a PSD matrix, an elliptic operator with rough coefficients, a covariance operator of a random field, or the Hamiltonian of a quantum system, we would like to find its best finite rank approximation with a given rank. One way to achieve this objective is to project the operator to its eigenspace that corresponds to the smallest or largest eigenvalues, depending on the setting. The eigenfunctions are typically global, i.e. nonzero almost everywhere, but our interest is to find the sparsest or most localized bases for these subspaces. The sparse/localized basis functions lead to better physical interpretation and preserve some sparsity structure in the original operator. Moreover, sparse/localized basis functions also enable us to develop more efficient numerical algorithms to solve these problems.

In this thesis, we present two methods for this purpose, namely the sparse operator compression (Sparse OC) and the intrinsic sparse mode decomposition (ISMD). The Sparse OC is a general strategy to construct finite rank approximations to PSD operators, for which the range space of the finite rank approximation is spanned by a set of sparse/localized basis functions. The basis functions are energy minimizing functions on local patches. When applied to approximate the solution operator of elliptic operators with rough coefficients and various homogeneous boundary conditions, the Sparse OC achieves the optimal convergence rate with nearly optimally localized basis functions. Our localized basis functions can be used as multiscale basis functions to solve elliptic equations with multiscale coefficients and provide the optimal convergence rate $O(h^k)$ for $2k$'th order elliptic problems in the energy norm. From the perspective of operator compression, these localized basis functions provide an efficient and optimal way to approximate the principal eigen-space of the elliptic operators. From the perspective of the Sparse PCA, we can approximate a large set of covariance functions by a rank-n operator with a localized basis and with the optimal accuracy. While the Sparse OC works well on the solution operator of elliptic operators, we also propose the ISMD that works well on low rank or nearly low rank PSD operators. Given a rank-n PSD operator, say a N-by-N PSD matrix A ($n \leq N$), the ISMD *decomposes* it into n rank-one matrices $\sum_{i=1}^{n} g_i g_i^T$, where the modes $\{g_i\}_{i=1}^{n}$ are required

to be as sparse as possible. Under the regular-sparse assumption (see Definition 1.3.2), we have proved that the ISMD gives the optimal patchwise sparse decomposition, and is stable to small perturbations in the matrix to be decomposed. We provide several applications in both the physical and data sciences to demonstrate the effectiveness of the proposed strategies.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] T. Y. Hou, Q. Li, and P. Zhang. "A Sparse Decomposition of Low Rank Symmetric Positive Semidefinite Matrices". In: *Multiscale Modeling & Simulation* 15.1 (2017), pp. 410–444. DOI: 10.1137/16M107760X. eprint: http://dx.doi.org/10.1137/16M107760X. URL: http://dx.doi.org/10.1137/16M107760X.

P.Z. proposed this project, proposed the patchwise sparseness minimization problem, designed algorithms to solve the optimization problem, conducted the numerical experiments, and proved the main theoretical results. P.Z. also participated in the writing of the manuscript.

[2] T. Y. Hou, Q. Li, and P. Zhang. "Exploring the Locally Low Dimensional Structure in Solving Random Elliptic PDEs". In: *Multiscale Modeling & Simulation* 15.2 (2017), pp. 661–695. DOI: 10.1137/16M1077611. eprint: http://dx.doi.org/10.1137/16M1077611. URL: http://dx.doi.org/10.1137/16M1077611.

P.Z. participated in proposing this project, designing algorithms, conducting the numerical experiments, proving the main theoretical results, and writing the manuscript.

[3] T. Y. Hou and P. Zhang. "Sparse operator compression of higher order elliptic operators with rough coefficients". In: *Research in the Mathematical Sciences* (2017). in press.

P.Z. proposed this project, proposed the concept of sparse operator compression, proved the main theoretical results, and conducted the numerical experiments. P.Z. also participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

In the last decade, significant progress has been made in a variety of fields of data sciences using ideas centered around sparsity. Examples include least absolute shrinkage and selection operator (lasso) (see e.g. [125, 126]), compressed sensing (see e.g. [39, 19]), sparse principal component analysis (see e.g. [137, 131]), matrix completion (see e.g. [20, 112]), robust principal component analysis (see e.g. [22, 25]), and phase retrieval (see e.g. [21]), etc. A key step in these examples is use of an optimization formulation with a constraint or penalty term that uses the l_1 or related norms. Sparse structures also prevail in physical sciences and partial differential equations (PDEs), mostly in two forms: localized and low rank structures. They have been explored in several methods, such as localized Wannier functions in quantum physics (see e.g. [90, 42, 91, 42]), localized multiscale finite element bases for the multiscale finite element method (MsFEM) and numerical homogenization for PDEs with multiscale coefficients (see e.g. [66, 70, 104, 43, 99]), fast multiple methods and hierarchical matrices (see e.g. [35, 57, 59, 11]), sparse grid stochastic finite element methods for stochastic partial differential equations (SPDEs) (see e.g. [96, 97, 33, 34]), etc. Several attempts to extend the optimization and l_1 techniques to physical sciences and PDEs have also appeared recently, including the l_1-optimization in solving SPDEs [41, 134], the l_1-optimization to explore sparse dynamics in PDEs [116], the compressed modes for variational problems [107, 108, 78], etc.

In this thesis, we will explore both sparse structures (localization and low rank) of positive semidefinite (PSD) operators or matrices that arise in the physical and data sciences. More specifically, given a PSD operator $\mathcal{A}$ or a large PSD matrix A, such as the Hamiltonian of a many-body system, the covariance operator of a random field, a differential operator, or a large sparse matrix that comes from the discretization of a differential operator, we would like to find its finite rank approximation with the smallest possible rank. Given a rank n, it is well-known that the best rank-n approximation is the projection of the operator to the eigen-subspaces corresponding to the n largest or smallest eigenvalues, depending on the problem. The representation of such subspaces,

however, is not unique. How do we represent these subspaces most efficiently, i.e., how do we find basis functions/vectors of these eigen-subspaces that require the fewest degrees of freedom to represent? Moreover, if a small sacrifice of the approximation accuracy is allowed, is it possible to get approximate subspaces with much more efficient representations, i.e., subspaces with much more localized basis functions/vectors? If so, what is the optimal trade-off between approximation accuracy and basis localization? Questions of this nature arise in many different contexts from both the physical and data sciences.

In this chapter, we first use the second order elliptic operator with rough coefficients as a model problem to motivate our research on operator compression with sparse/localized basis functions. Then we summarize our first method, i.e., the sparse operator compression (Sparse OC), and present its results when compressing different kinds of elliptic operators. After that, we summarize our second method, i.e., the intrinsic sparse mode decomposition (ISMD), which decomposes low rank operators into optimally sparse rank-one operators. Finally, we summarize different applications of our operator compression methods, including solving elliptic equations with multiscale coefficients, conducting the sparse principal component analysis, constructing localized Wannier functions for Hamiltonians and solving elliptic equations with high dimensional random coefficients.

1.1 A Model Problem of Operator Compression Solving Elliptic Equations With Rough Coefficients

Consider the elliptic equation with multiscale coefficients:

$$\begin{aligned} (\mathcal{L}u)(x) := -\nabla \cdot (a(x)\nabla u(x))u = f(x), \qquad & x \in D \subset \mathbb{R}^d, \\ u(x) = 0, \qquad & x \in \partial D, \end{aligned} \tag{1.1}$$

where D is a bounded Lipschitz domain in $\mathbb{R}^d$ and $f \in L^2(D)$. Here, we only assume that a is symmetric and uniformly elliptic on D and with entries in $L^\infty(D)$, i.e., there exist $0 < a_{min} \leq a_{max}$ such that

$$a_{min} I_d \preceq a(x) \preceq a_{max} I_d, \qquad x \in D. \tag{1.2}$$

The Lax–Milgram Lemma (see e.g. [31]) implies that Eqn. (1.1) has a unique weak solution in $H_0^1(D)$, denoted as $\mathcal{L}^{-1}f$. It is important to point out that we do not impose any assumption on the structure of a, such as periodicity, scale separation, or ergodicity at fine scales, as in the classical homogenization

literature; see [12]. We only make the minimal assumption that ensures the existence of a unique weak solution.

Under the framework of the Galerkin finite element method, given n basis functions $\Psi = [\psi_1, \ldots, \psi_n]$ in $H_0^1(D)$, we can solve the elliptic equation (1.1) approximately by projecting it onto the subspace spanned by Ψ, i.e.,

$$u_\Psi = \Psi L_n^{-1} \Psi^T f.$$

Here, $L_n \in \mathbb{R}^{n\times n}$ and $\Psi^T f \in \mathbb{R}^n$ are the stiffness matrix and the loading vector respectively, which are formally defined as follows:

$$L_n(i,j) := \int_D \psi_i \mathcal{L} \psi_j, \quad (\Psi^T f)(i) := \int_D \psi_i f, \quad \forall 1 \le i,j \le n. \tag{1.3}$$

We would like to choose the basis Ψ so that it minimizes the solution error measured by the following norm:

$$\|\mathcal{L}^{-1} - \Psi L_n^{-1} \Psi^T\|_2 := \sup_{\|f\|_{L^2(D)}\le 1} \|\mathcal{L}^{-1} f - \Psi L_n^{-1} \Psi^T f\|_{L^2(D)},$$

where $\|A\|_2$ is the largest eigenvalue for a PSD operator A.

Therefore, for a self-adjoint, positive definite elliptic operator $\mathcal{L}$, we define the operator compression error of the basis Ψ as

$$E_{oc}(\Psi; \mathcal{L}^{-1}) := \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{L}^{-1} - \Psi K_n \Psi^T\|_2, \tag{1.4}$$

which is the optimal approximation error of $\mathcal{L}^{-1}$ among all positive semidefinite operators with range space spanned by Ψ. It is easy to verify that the operator compression error only depends on the subspace spanned by Ψ, and is independent of the functions that are used to represent this subspace.

A Brief Review of Existing Results

The compression error $E_{oc}(\Psi; \mathcal{L}^{-1})$ achieves its minimum, which is $\mathcal{O}(n^{-2/d})$, when Ψ are taken as the eigenfunctions of $\mathcal{L}$ corresponding to its smallest n eigenvalues; see e.g. [93, 36]. However, the eigenfunctions are expensive to compute, and there are no localized basis functions that can span the eigensubspace; see e.g. [137, 107, 64]. In practice, localized/sparse basis functions are preferred to preserve the sparsity in the (discretized) elliptic operator, and thus to achieve better numerical complexity. For example, when the coefficient a is constant in Eqn. (1.1), piecewise linear elements on a uniform mesh

achieve the optimal compression error up to a constant multiplier (see e.g. [31]). Meanwhile, these elements are optimally localized, the corresponding stiffness matrix (1.3) is sparse, and thus efficient numerical methods such as conjugate gradient method (SC) or multigrid method can be applied; see e.g. [2]. However, the compression error given by the piecewise linear elements can be arbitrarily large if the coefficient is rough; see [5]. This motivates the generalized finite element method [6, 123], multiscale finite element method [66, 70, 67, 44, 43, 89, 29] and numerical homogenization [102, 103, 105, 99]. One of the major objectives is to construct localized basis functions without loss of the approximation accuracy (up to a constant multiplier) for Eqn. (1.1).

In [56], on a regular finite element mesh with mesh size h, the authors construct a localized finite element basis (AL basis) with the same support of the piecewise linear finite elements. They have proved that with $\mathcal{O}\left((\frac{1}{h})^d(\log\frac{1}{h})^{d+1}\right)$ basis functions, the basis achieves an $\mathcal{O}(h^2)$ compression error. Let n be the number of elements in the whole domain with mesh size h. We have $n=\mathcal{O}((1/h)^d)$. Their results imply that an $\mathcal{O}(n^{-2/d})$ compression error can be achieved with $\mathcal{O}\left(n(\log n)^{d+1}\right)$ localized basis functions. Although the construction provided in [56] involves solving global problems, its theoretical results support the possibility of constructing localized basis functions while achieving the optimal compression error up to a logarithmic multiplier.

In the localizable orthogonal decomposition (LOD) [89], the authors introduce a modified Clément interpolation $\mathcal{I}_h$ on a uniform mesh with mesh size h. They define V^f as the kernel of $\mathcal{I}_h$ (i.e., the set of functions u such that $\mathcal{I}_h u=0$), and identify the finite element space Ψ as the orthogonal complement of V^f with respect to the inner product defined by $a(u,v)=\int_D u\mathcal{L}v$ for $u,v\in H_0^1(D)$. The finite element basis ψ_i is identified by $\varphi_i-\mathcal{P}^a_{V^f}\varphi_i$, where φ_i is the nodal piecewise linear element and $\mathcal{P}^a_{V^f}\varphi_i$ is its projection onto the space V^f with respect to (w.r.t.) the inner product $a(u,v)$. The work [89] shows that this finite element basis Ψ achieves the optimal compression rate. Moreover, they show that the finite element basis function ψ_i decays exponentially fast away from its associated node, and thus can be localized to local patches of size $\mathcal{O}(h\log(1/h))$ without loss of accuracy.

In [98, 99], the basis functions (also called gamblets) are derived from a Bayesian perspective, by conditioning certain Gaussian random fields with some measurements of the solution (e.g., the average on a local patch). More

specifically, for $f \in L^2(D)$, one can think that the solution $\boldsymbol{\xi}$ follows a distribution of a Gaussian measure with mean 0 and covariance operator $\mathcal{L}^{-1}$ *a priori*. With n measurements of the solution, say $\{\int_D \varphi_i \boldsymbol{\xi}\}_{i=1}^n$, where $\{\varphi_i\}_{i=1}^n \subset L^2(D)$, the best guess of the solution (posterior mean) is

$$\mathbb{E}\left[\boldsymbol{\xi} \mid \int_D \varphi_i \boldsymbol{\xi}, 1 \le i \le n\right] = \sum_{i=1}^n \left(\int_D \varphi_i \boldsymbol{\xi}\right) \psi_i,$$

where $\{\psi_i\}_{i=1}^n$ are the basis functions to be constructed. Intuitively, φ_i is the best guess of the solution for the measurements $\int_D \varphi_i \boldsymbol{\xi} = 1$ and $\int_D \varphi_j \boldsymbol{\xi} = 0$ for $j \neq i$. In [98, 99], it is proved that the basis function ψ_i can be obtained from the following quadratic optimization problem:

$$\begin{aligned} \min_{\psi \in H_0^1(D)} \quad & \int_D \nabla\psi \cdot a \nabla\psi \\ \text{subject to (s.t.)} \quad & \int_D \psi\varphi_j = \delta_{i,j}, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{1.5}$$

In [99], by partitioning the physical domain D with a regular mesh with mesh size h and taking $\{\varphi_i\}_{i=1}^n$ as the indicator function of local cells (therefore, n is the number of cells in the partition), the author proved that $\{\psi_i\}_{i=1}^n$ achieves the optimal compression rate, and that ψ_i decays exponentially fast away from its associated patch. Therefore, ψ_i can be approximated accurately by solving Eqn. (1.5) on a local patch with patch size $\mathcal{O}(h \log(1/h))$, and without loss of the optimal compression rate. It is worth mentioning that this construction can be implemented hierarchically on a multi-level grid and leads to a multigrid algorithm to solve Eqn. (1.1) with complexity $N \log^{3d} N$ as a direct solver, and with complexity $N \log^{d+1} N$ for subsequent solves with a different right hand side. This work has recently been extended to solve elliptic equations with nonzero potentials, and hyperbolic and parabolic ODEs/PDEs with rough coefficients in [101].

The existing results imply that for second order elliptic operators with rough coefficients, although the eigen-subspaces do not have localized basis functions, we can still localize the basis functions if an optimal compression accuracy up to a constant multiplier is acceptable. Moreover, we can obtain the optimal compression accuracy (up to a constant multiplier) with nearly optimally localized basis functions (up to a logarithmic multiplier) in the trade-off between the approximation accuracy and basis localization. While we have these exciting advances for second order elliptic operators, there is little literature on

operator compression for higher order elliptic operators with rough coefficients. Moreover, the current methods for second order elliptic operators scale poorly with the contrast of the coefficients (defined as a_{max}/a_{min} in Eqn. (1.2)). More precisely, the constants in both the compression error and the localization depend polynomially on the contrast in [56, 89, 98, 99]), which makes the existing methods inefficient for coefficients with high contrast. These defects motivate us to propose a general strategy to construct localized basis functions and to perform the error analysis for a large class of elliptic operators.

Compressing elliptic operators with localized basis functions (also known as multiscale finite element method and numerical homogenization) has been an active research area and new results are still appearing. In particular, there are new results in [110, 60] in which the LOD method has been extended to tackle the high contrast coefficient problem. In [100], the work in [99] has been generalized to compress a large class of bounded linear operators, including the solution operator of higher order elliptic operators. We review these methods in detail when we present our methods and then we compare these different results.

1.2 The Sparse Operator Compression (Sparse OC)

Problem Setting

In this thesis, we develop a general method called sparse operator compression to compress a bounded self-adjoint positive semidefinite (PSD) operator $\mathcal{K} : X \to X$ with sparse/localized basis functions, where X can be any separable Hilbert space with inner product $(\cdot, \cdot)$. More precisely, given the operator $\mathcal{K} : X \to X$ and a positive integer n, the Sparse OC constructs n sparse/localized basis functions $\Psi^{loc} := [\psi_1^{loc}, \ldots, \psi_n^{loc}] \subset X$ that achieves a small operator compression error measured in the following norm:

$$E_{oc}(\Psi^{loc}; \mathcal{K}) := \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{K} - \Psi^{loc} K_n \Psi^{loc,T}\|_{X,X}. \tag{1.6}$$

Here, $\Psi^{loc,T}$ is an operator that maps $f \in X$ to $[(\psi_1^{loc}, f), \ldots, (\psi_n^{loc}, f)]^T \in \mathbb{R}^n$, and $\|\mathcal{K}\|_{X,X} := \sup_{\|f\|_X \leq 1} \frac{\|\mathcal{K}f\|_X}{\|f\|_X}$ is the induced operator norm for the bounded operator $\mathcal{K} : X \to X$.

The Cameron-Martin space H, defined as the completion of $\mathcal{K}(X)$ with the inner product

$$(\mathcal{K}\varphi_1, \mathcal{K}\varphi_2)_H = (\mathcal{K}\varphi_1, \varphi_2) \qquad \forall \varphi_1, \varphi_2 \in X, \tag{1.7}$$

plays an important role in both the construction and the analysis. In Section 2.2, we will prove that H is a Hilbert space that can be continuously embedded into X. In the model problem of compressing a second order elliptic operator $\mathcal{L}$ in the last section, $X = L^2(D)$, $\mathcal{K}$ plays the role of the solution operator $\mathcal{L}^{-1}$, and H is the solution space $H_0^1(D)$ equipped with the inner product $(\psi_1, \psi_2)_H = \int_D \nabla\psi_1 \cdot a\nabla\psi_2$ for any $\psi_1, \psi_2 \in H_0^1(D)$. We will also apply the Sparse OC to other bounded self-adjoint PSD operators. For example, when compressing the solution operator of a second order elliptic operator with high contrast coefficients $a \in L^\infty(D)$, we choose $X = L_a^2(D)$ ($L^2(D)$ with the a-weighed inner product), and H is a subspace of the Sobolev space $H^1(D)$ equipped with the associated energy norm; and when compressing the solution operator of a $2k$'th ($k \geq 1$) order elliptic operator, we choose $X = L^2(D)$, and H is a subspace of the Sobolev space $H^k(D)$ equipped with the associated energy norm; when compressing the covariance operator of a Gaussian measure over $L^2(D)$, we choose $X = L^2(D)$, and H is exactly the Cameron-Martin space associated with this Gaussian measure (see [15]). This is the reason why we call H the Cameron-Martin space.

Constructing Basis Functions

The Sparse OC follows three steps to construct sparse/localized basis functions $\Psi^{loc} := [\psi_1^{loc}, \ldots, \psi_n^{loc}]$ that achieve the optimal operator compression rate. In the first step, we partition the physical domain D with a regular finite element mesh with mesh size $h > 0$, and denote all the elements (local patches) as $\{\tau_i\}_{i=1}^m$. On each local patch τ_i, we choose Q_i ($Q_i \in \mathbb{N}$) measurement functions that are only supported on τ_i, denoted as $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$. In the second step, we construct non-local basis functions $\{\psi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}$ by the following minimizing problem:

$$\begin{aligned} \psi_{i,q} = \underset{\psi \in H}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_{j,q'}) = \delta_{iq,jq'}, \quad \forall 1 \leq j \leq m,\, 1 \leq q' \leq Q_j, \end{aligned} \tag{1.8}$$

where $\delta_{iq,jq'}$ is the Kronecker delta that is 1 when $i = j$ and $q = q'$ and 0 in all other cases. Collecting all the nonlocal basis functions $\psi_{i,q}$ for $1 \leq i \leq m$ and $1 \leq q \leq Q_i$, we get a nonlocal basis Ψ. Although $\psi_{i,q}$ is not localized, we will show that it decays exponentially fast away from its associated patch. Therefore, in the final step, we restrict the global construction onto a neighborhood of the patch τ_i with diameter r, denoted as $S_r(\tau_i)$, and obtain a localized basis

function:

$$\begin{aligned} \psi_{i,q}^{loc} = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_{j,q'}) = \delta_{iq,jq'}, \quad \forall 1 \le j \le m,\, 1 \le q' \le Q_j, \\ & \psi(x) \equiv 0, \quad x \in D \backslash S_r. \end{aligned} \tag{1.9}$$

Collecting all the $\psi_{i,q}^{loc}$ for $1 \le i \le m$ and $1 \le q \le Q_i$ together, we get our local basis Ψ^{loc}. In most of our applications, we take $r = O(h \log(1/h))$. We will discuss the choice of r in details when we analyze the compression error of Ψ^{loc}.

A General Analysis Framework

We provide a general framework to analyze the compression error $E_{oc}(\Psi^{loc}; \mathcal{K})$ and the localization of the basis Ψ^{loc}. The big picture is (1) to analyze the compression error and the decay property of the nonlocal basis Ψ, and (2) to choose sufficiently large $S_r(\tau_i)$ such that the *compression rate* remains the same for the localized basis Ψ^{loc} (although the actual compression error may be amplified by a constant).

First of all, we show that a local *projection-type* approximation property suffices to guarantee an estimate of the compression error $E_{oc}(\Psi; \mathcal{K})$. More precisely, we have the error estimate for the nonlocal basis

$$E_{oc}(\Psi; \mathcal{K}) \le \epsilon^2 , \tag{1.10}$$

if the following local *projection-type* approximation property

$$\inf_{\varphi \in \Phi_i} \|u - \varphi\|_{X(\tau_i)} \le \epsilon \|u\|_{H(\tau_i)} \tag{1.11}$$

holds true for every patch τ_i and every $u \in \mathcal{K}(X) \subset H$. Here, Φ_i is the space spanned by the local measurements $\{\varphi_{i,q} : 1 \le q \le Q_i\}$; $\|u\|_{X(\tau_i)}$ and $\|u\|_{H(\tau_i)}$ are the X-norm and H-norm of u restricted to τ_i, respectively. Moreover, if Ψ is used as the finite element basis to solve the corresponding linear system $\mathcal{L}u = f$ (whose solution is $u = \mathcal{K}f$), we have the following error estimate in the energy norm:

$$\|u - u_\Psi\|_H \le \epsilon \|f\|_X \quad \forall f \in X,$$

where u_Ψ is the corresponding Galerkin finite element solution. When $X = L^2(D)$, $H \subset H^1(D)$ and $\{\varphi_{i,q}\}_{q=1}^{Q_i}$ only contains the constant function, the

local *projection-type* approximation property (1.11) can be obtained from the Poincare inequality. These local constant measurement functions are used in [99, 101], and thus our Sparse OC can be viewed as a generalization of their gamblet method for second order elliptic operators. We emphasize that the local *projection-type* approximation property (1.11) serves as the criterion to pick the measurement functions when compressing a general PSD operator, which is one of the key ideas to generalize the gamblet method [99, 101]. For example, when compressing elliptic operators of order $2k$ ($k \geq 1$), the solution space H is a subset of $H^k(D)$. Since the polynomial space has a good *projection-type* approximation property in $H^k(D)$ (see Theorem 3.2.1), it is natural to pick Φ_i as the local polynomial space. When compressing elliptic operators with high contrast coefficients, we take $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$ as the first Q_i eigenfunctions of the elliptic operator with the homogeneous Neumann boundary condition on $\partial\tau_i$. The number of local measurements Q_i is roughly the number of disconnected high permeability regions (where the coefficient a is large) contained in the local patch τ_i. With this construction, we are able to achieve a compression error independent of the contrast; see Section 4.2.

Secondly, we show that a local *inverse energy estimate* guarantees the exponential decay of all the nonlocal basis functions $\{\psi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}$. More precisely, suppose we have the following estimate on every local patch τ_i:

$$\|\mathcal{L}\psi\|_{H(\tau_i)} \leq \frac{C_{inv}}{\epsilon}\|\psi\|_{H(\tau_i)} \quad \forall \psi \in \Psi, \tag{1.12}$$

where $\mathcal{L}$, i.e. the inverse of $\mathcal{K}$, is an elliptic operator in this thesis, $\epsilon > 0$ is the same as that in the local approximation property (1.11), and $C_{inv} > 0$ is some constant that may depend on the contrast of the coefficients in $\mathcal{L}$. Then we can prove that every nonlocal basis function $\psi_{i,q}$ decays exponentially fast away from its associated patch τ_i, with a decay rate at the order of C_{inv}. Roughly speaking, we can prove that there exists $x_i \in \tau_i$ such that $|\psi_{i,q}(x)| \overset{\leq}{\approx} |\psi_{i,q}(x_i)| \exp(-\frac{|x-x_i|}{C_{inv}h})$ holds true for any $x \in D$. Notice that the "load" $\mathcal{L}\psi$ is bounded by the energy of the "solution" ψ in Eqn. (1.12), which is in an inverse direction of the standard energy estimate for elliptic equations. Therefore, we call it an *inverse energy estimate*. It is definitely not true for all functions in the solution space H, but it can be proved for all functions in the finite dimensional space Ψ. This exponential decay is proved for second order uniformly elliptic operators with the homogeneous Dirichlet boundary condi-

tion in [99]. We prove that this exponential decay is true for second order uniformly elliptic operators with various homogeneous boundary conditions and with nonzero potentials. Other boundary conditions, like periodic and Neumann boundary conditions, and nonzero potentials are of interest when we apply the Sparse OC to compress the Hamiltonian in quantum chemistry; see Section 2.5. Furthermore, we have proved this local inverse energy estimate for higher order elliptic operator and for elliptic operators with high contrast coefficients, by using $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$ that we choose to satisfy the local approximation property (1.11). Therefore, the nonlocal basis Ψ satisfies the error estimate (1.10) and decays exponentially fast at the same time! Finally, for elliptic operators with high contrast coefficients, under some geometric assumptions of the coefficients (see Section 4.5), we have shown that the exponential decay rate (which is of order C_{inv}) is independent of the contrast under our construction. Therefore, the basis functions can be localized such that its support depends on the contrast of the coefficients only logarithmically. This partially resolves the issue that current methods, e.g. the LOD [89] and the gamblet [99, 101], scale poorly with the contrast. Recently, improved LOD methods have appeared to tackle the high contrast coefficients (see [110, 60]). We review these LOD-based methods and compare them with our method in Section 4.1.

Finally, the exponential decay justifies the localized construction of Ψ^{loc}. On a regular finite element mesh with mesh size h, constructing $\psi_{i,q}^{loc}$ on a local domain with diameter $\mathcal{O}(h\log(1/h))$ is sufficient to preserve the compression rate of the nonlocal basis Ψ. When compressing a second order elliptic operator with high contrast coefficients, although the compression error is more sensitive to the truncation, by a small modification of our method we can prove that a local domain with diameter $\mathcal{O}\left(h\left(\log(1/h)+\log\left(\frac{a_{max}}{a_{min}}\right)\right)\right)$ is sufficient to preserve the compression error given by Ψ, which is independent of the contrast by construction. We summarize the choice of local measurements Φ_i, the compression rate and the support size of the localized basis function $\psi_{i,q}$ for different kinds of elliptic operators in Table 1.1. Notice that the number of elements in a regular mesh with mesh size h is $\mathcal{O}(h^{-d})$. Since the nth largest eigenvalue of $\mathcal{L}^{-1}$ is $\mathcal{O}(n^{-2d/k})$, i.e. $\lambda_n(\mathcal{L}^{-1}) = \mathcal{O}(n^{-2d/k})$, one can easily check that the localized basis Ψ^{loc} indeed achieves the optimal compression rate and is nearly optimally localized (up to a logarithmic multiplier). Therefore, for elliptic operators with low contrast coefficients, one can achieve the optimal

compression error (up to a constant multiplier) with nearly optimally localized basis functions (up to a logarithmic multiplier) in the trade-off between the approximation accuracy and basis localization.

Order of $\mathcal{L}$	2	$2k(k \geq 2)$	2
Contrast	low	low	high
X	$L^2(D)$	$L^2(D)$	$L_a^2(D)$
Φ_i or $\{\varphi_{i,q}\}_{q=1}^{Q_i}$	$\mathbf{P}_0$	$\mathbf{P}_{k-1}$	$\mathcal{L}\varphi_{i,q} = \lambda_{i,q} a \varphi_{i,q}$
Support size of $\psi_{i,q}^{loc}$	$\mathcal{O}(h \log(1/h))$	$\mathcal{O}(h \log(1/h))$	$\mathcal{O}\left(h\left(\log \frac{1}{h} + \log \frac{a_{max}}{a_{min}}\right)\right)$
GFEM error (in energy norm)	$\mathcal{O}(h)$	$\mathcal{O}(h^k)$	$\mathcal{O}(h)$
Compression error	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{2k})$	$\mathcal{O}(h^2)$
Contrast dependence	Yes	Yes	No

Table 1.1: Sparse OC applied to different kinds of elliptic operators: the space X where is the space for the right hand side f, the local measurement function space Φ_i ($\mathbf{P}_k$ denotes the polynomial space with degree no more than k), the support size of $\psi_{i,q}^{loc}$, the convergence rate of the Galerkin finite element solution in energy norm (i.e. H-norm), the operate compression rate $E_{oc}(\Psi^{loc}; \mathcal{K})$, and the last row indicates whether the constants in the estimates depend on the contrast or not.

Our Contributions

The Sparse Operator Compression is directly inspired by the Bayesian homogenization [98] and the gamblet method [99], including the ideas of constructing the space Ψ from $\mathcal{K}\Phi$ and of constructing the basis functions from energy minimizing problems. The recursive argument to prove the exponential decay and basis localization has been used in [89] and [99]. Based on these existing works, we have made the following main contributions in our Sparse OC.

First of all, our Sparse OC, which is purely based on functional analysis, generalizes the probabilistic framework for the Bayesian numerical homogenization [98] and the gamblet method [99]. We have identified the local *projection-type* approximation property and the local *inverse energy estimate* as the two key components in proving the error estimate and the exponential decay. These two inequalities serve as the criteria to choose the local measurement functions in the Sparse OC.

Secondly, for *strongly elliptic* operators of $2k$'th order, we have constructed nearly optimally localized basis functions with support size $\mathcal{O}(h\log(1/h))$ and with optimal operator compression rate $\mathcal{O}(h^{2k})$. The strong ellipticity is equivalent to the standard uniform ellipticity in the cases: (1) $k = 1$ (2)$d = 1$ or 2 and (3) $(d, k) = (3, 2)$, and is only slightly stronger than the uniform ellipticity in other cases. We have also conducted numerical experiments to demonstrate that the fractional Laplacian operators cannot be localized using the same approach. More precisely, the global basis functions constructed in Eqn. (1.8) do not have exponential decay for most fractional Laplacian operators.

Thirdly, for second order elliptic operators with high contrast coefficients, we have constructed localized basis functions with support size of order

$$h\left(\log(1/h) + \log\left(\frac{a_{max}}{a_{min}}\right)\right).$$

For the two-phase coefficients with smooth inclusions/channels, we have shown that the decay rate of the basis functions is independent of the contrast by an asymptotic analysis. Moreover, the error in energy norm of the corresponding finite element solution is of order h, and is independent of the contrast. Compared with recent results on the high contrast problems [110, 60], our result requires weaker assumptions on the coefficients. For example, our method allows multiple high-conductivity inclusions in a local patch but neither of the methods in [110, 60] works in this case.

Finally, we have applied the Sparse OC to the problem of sparse principal component analysis in statistics and constructing localized Wannier functions in quantum chemistry; see more details of these applications in Section 1.4. For the problem of sparse principal component analysis, the Sparse OC is especially suitable for random fields with the Matérn class covariance operators [92], which can be seen as the solution operator of certain higher order elliptic operators. We have compared the Sparse OC with the l_1 regularization approach [107] on simple model problems (sparse principal component analysis of a Matérn class covariance operator and constructing localized Wannier functions for the 1D free-electron model), and our results have demonstrated the effectiveness and efficiency of the Sparse OC in these applications.

Recently, the authors of [100] introduce the Gaussian cylinder measure and successfully generalize the work in [98, 99] to a much broader class of operators. With a slightly different construction from our construction (1.9), they are able

to obtain the same order of compression accuracy and basis localization as we have achieved (the second column in Table 1.1), but without requiring the strong ellipticity. The Sparse OC and the work [100] generalize the the work in [98, 99] from different perspectives, resulting in different conditions for the measurement functions and slightly different constructions of localized basis functions. These new results in numerical methods for PDEs are likely to find more applications in both data and physical sciences.

1.3 The Intrinsic Sparse Mode Decomposition (ISMD)

Problem Setting

In the Sparse OC above, we look for rank-n approximations in the form of

$$\mathcal{K} \approx \Psi K_n \Psi^T, \tag{1.13}$$

where $\Psi = [\psi_1, \ldots, \psi_n]$ is a basis of the range space of the rank-n approximation and K_n is any PSD n-by-n matrix that minimizes the approximation error. The corresponding error, i.e., $E_{oc}(\Psi; \mathcal{K}) = \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{K} - \Psi K_n \Psi^T\|_{X,X}$, is invariant with respect to any non-degenerate linear transformation of Ψ. In some applications, one is interested in approximating $\mathcal{K}$ with n rank-one operators, i.e.,

$$\mathcal{K} \approx \sum_{i=1}^{n} \psi_i \psi_i^T. \tag{1.14}$$

The corresponding error, $\|\mathcal{K} - \Psi\Psi^T\|_{X,X}$ is only invariant to unitary transformations of Ψ.[1] As in the Sparse OC, we require that the decomposed modes $\{\psi_i\}_{i=1}^n$ be as sparse/localized as possible. Compared with the eigendecomposition, the decomposed modes in Eqn. (1.14) are required to be sparse/localized instead of orthogonal.

In this part of our study, we only consider the case when the operator $\mathcal{K}$ is rank-n, i.e., $\lambda_n(\mathcal{K}) > \lambda_{n+1}(\mathcal{K}) = 0$, or is nearly rank-$n$, i.e., $\lambda_n(\mathcal{K}) \gg \lambda_{n+1}(\mathcal{K}) \approx 0$. In this case, there is no need to consider the trade-off between the approximation accuracy and basis sparsity: on the approximation accuracy side, we fix the number of modes in Ψ to n and impose the hard constraint $\mathcal{K} = \Psi\Psi^T$ when $\mathcal{K}$ is rank-n; on the sparsity/localization side, we would like to obtain the optimally sparse/localized basis functions Ψ. More precisely, we

[1] Mathematically, we get the kind of approximation in Eqn. (1.14) if we limit K_n to be diagonal in the Sparse OC approximation (1.13).

want to solve the following optimization problem:

$$\min_{\psi_1,\ldots,\psi_n} \quad \sum_{i=1}^{n} |\mathrm{supp}(\psi_i)| \quad \text{s.t.} \quad \mathcal{K} = \sum_{i=1}^{n} \psi_i \psi_i^T , \tag{1.15}$$

where $|\mathrm{supp}(\psi_i)|$ is a measure of the volume of ψ_i's support, such as the number of nonzero entries (also known as the l^0 norm) for a vector in the discrete case or the volume of a function support in the continuous case. Compared with the problem solved by the Sparse OC, although problem (1.15) does not consider the trade-off between accuracy and sparsity, it is much more difficult in the sense that we want to find the optimally localized basis functions instead of finding nearly optimally localized basis functions up to a logarithmic multiplier as we have done in Sparse OC.

Patchwise Sparseness and the Surrogate Problem

In most cases, minimizing support (l^0 norm in the discrete setting) results in a combinatorial problem and is computationally intractable. Therefore, we introduce the following patchwise sparseness as a surrogate of $|\mathrm{supp}\psi_i|$ and make the problem computationally tractable.

Definition 1.3.1 (Patchwise sparseness)**.** *Suppose that $\mathcal{P} = \{\tau_i\}_{i=1}^m$ is a partition of the physical domain D or the N nodes, i.e., $\overline{D} = \cup_{i=1}^m \overline{\tau_i}$ in the continuous setting or $\{1, 2, 3, \ldots, N\} = \cup_{i=1}^m \tau_i$ in the discrete setting. The patchwise sparseness of ψ with respect to the partition $\mathcal{P}$, denoted by $s(\psi; \mathcal{P})$, is defined as*

$$s(\psi; \mathcal{P}) = \#\{\tau \in \mathcal{P} : \psi|_\tau \neq \mathbf{0}\}.$$

Here, $\psi|_\tau \neq \mathbf{0}$ means that ψ does not completely vanish on the patch τ_i. Once the partition $\mathcal{P}$ is fixed, smaller $s(\psi; \mathcal{P})$ means that ψ is nonzero on fewer patches, which implies a sparser or more localized function. With the patchwise sparseness as a surrogate of $|\mathrm{supp}\psi_i|$, the sparse decomposition problem (1.15) is relaxed to

$$\min_{\psi_1,\ldots,\psi_n} \quad \sum_{i=1}^{n} s(\psi_i; \mathcal{P}) \quad \text{s.t.} \quad \mathcal{K} = \sum_{i=1}^{n} \psi_i \psi_i^T . \tag{1.16}$$

If $\{g_i\}_{i=1}^n$ is an optimizer for problem (1.16), we call them a set of *intrinsic sparse modes* for $\mathcal{K}$ under partition $\mathcal{P}$. Since the objective function of problem (1.16) only takes nonnegative integer values, we know that for a symmetric PSD operator $\mathcal{K}$ with rank n, there exists at least one set of intrinsic sparse modes.

Theoretical Results of the ISMD

It is obvious that the intrinsic sparse modes depend on the domain partition $\mathcal{P}$. Two extreme cases would be $m \to \infty$ and $m = 1$. For $m \to \infty$, $s(\psi; \mathcal{P})$ recovers $|\text{supp}\psi|$ and the patchwise sparseness minimization problem (1.16) recovers the original support minimization problem (1.15). Unfortunately, it is computationally intractable. For $M = 1$, every non-zero vector has sparseness one, and the support size makes no difference. However, in this case problem (1.16) is computationally tractable. For instance, a set of (unnormalized) eigenfunctions is one of the optimizers. We are interested in the sparseness defined in between, namely, a partition with a meso-scale patch size. Compared to $|\text{supp}\psi|$, the meso-scale partition sacrifices some resolution when measuring the support, but makes the optimization (1.16) efficiently solvable. Specifically, problem (1.16) with the following *regular-sparse* partitions enjoys many good properties. These properties enable us to design a very efficient algorithm to solve problem (1.16).

Definition 1.3.2 (Regular-sparse partition)**.** *The partition $\mathcal{P}$ is regular-sparse w.r.t. $\mathcal{K}$ if there exists a decomposition $\mathcal{K} = \sum_{i=1}^{n} g_i g_i^T$ such that all nonzero modes on each patch τ_i $(1 \le i \le m)$ are linearly independent.*

If two intrinsic sparse modes are non-zero on exactly the same set of patches, which are called unidentifiable modes in Definition 5.3.4, it is easy to see that any rotation of these unidentifiable modes forms another set of intrinsic sparse modes. From a theoretical point of view, if a partition is regular-sparse w.r.t. $\mathcal{K}$, the intrinsic sparse modes are unique up to rotations of unidentifiable modes; see Theorem 5.3.5. Moreover, as the partition gets refined, the original identifiable intrinsic sparse modes remain unchanged, while the original unidentifiable modes become identifiable and become sparser; see Theorem 5.3.6. In this sense, the intrinsic sparse modes are essentially independent of the partition that we use.

From a computational point of view, a regular-sparse partition ensures that the restrictions of the intrinsic sparse modes on each patch τ_i can be constructed from rotations of local eigenvectors. Following this idea, we propose the intrinsic sparse mode decomposition (ISMD); see Algorithm 2. The ISMD follows the "local-modes-construction + patching-up" procedure. The key step is to construct local pieces of the intrinsic sparse modes by a joint diagonalization problem. Thereafter, a pivoted Cholesky decomposition is utilized to glue

these local pieces together. In Theorem 5.3.5, we prove that the ISMD solves problem (1.16) exactly on regular-sparse partitions. We point out that, even when the partition is not regular-sparse, numerical experiments show that the ISMD still generates a sparse decomposition of $\mathcal{K}$.

Computational Complexity of the ISMD

The ISMD has very low computational complexity. There are two reasons for its efficiency. First of all, instead of computing the expensive global eigendecomposition, we compute only the local eigendecompositions on every patch. Secondly, there is an efficient algorithm to solve the joint diagonalization problems. For partitions with a large range of patch sizes, the computational cost of the ISMD is comparable to that of the partial eigendecomposition [117, 82]. For certain partitions, the ISMD could be ten times faster than the partial eigendecomposition. We have also compared the ISMD with the convex relaxation of the Sparse PCA [78, 128]. Our numerical results indicate that the convex relaxation of Sparse PCA fails to capture the long range correlation. Moreover, it needs to perform (partial) eigendecomposition on matrices repeatedly many times. As a result, the convex relaxation of Sparse PCA is thus much slower than the ISMD. Finally, because both performing the local eigendecompositions and solving the joint diagonalization problems can be done independently on each patch, the ISMD is embarrassingly parallelizable.

Our Contributions

Our ISMD is a novel approach to decompose a low rank operator into several sparse/localized rank-one operators, which is important in many applications, such as uncertainty quantification (see e.g. [7, 4, 69, 106]) and latent factor models (see e.g. [48, 129, 1, 25]). First of all, we use a domain partition with meso-scales to define the patchwise sparseness, similar to the group sparsity [135, 71] or structured sparsity [72]. We then propose the ISMD algorithm, which consists of local eigendecompositions and joint diagonalization across patches, to solve the patchwise sparseness minimization problem (1.16). We have conducted numerical experiments to show the efficiency and robustness of the ISMD. We have also compared the ISMD with other existing methods, e.g., eigendecomposition, pivoted Cholesky decomposition, and convex relaxation of the Sparse PCA [78, 128].

Secondly, we prove that the ISMD solves problem (1.16) exactly under the

regular-sparse assumption. Moreover, we show that the results of the ISMD are consistent when the partition is refined, which means that the original identifiable intrinsic sparse modes remain unchanged and the original unidentifiable modes become identifiable and become sparser as the partition is refined. Finally, we prove the stability of ISMD under small perturbation of the input low rank operators. All our theoretical results are validated by numerical experiments using covariance matrices from porous media problems.

Last but not least, we have used the ISMD to obtain locally low dimensional parametrization of a random field. Based on this locally low dimensional parametrization, we propose a stochastic multiscale finite element method (StoMsFEM) to solve the second order elliptic equations with high dimensional random coefficients. The proposed method shows significant computational saving compared to the traditional MC methods or the gPC based methods; see Section 1.4 for a brief overview and our paper [65] for details.

1.4 Applications of The Operator Compression Methods

In addition to solving elliptic equations with rough coefficients, we have explored three more applications of the proposed operator compression methods. In the first application, we apply both the Sparse OC and the ISMD to solve the problem of sparse principal component analysis in statistics. These two methods look for different forms of sparse principal component analysis (the ISMD imposes principal factors to be uncorrelated while the Sparse OC does not), and are suitable for different kinds of covariance operators. In the second application, we apply the Sparse OC to construct localized Wannier functions in quantum chemistry. In the third application, we apply the ISMD to explore the locally low dimensional structure of the solutions of second order elliptic equations with high dimensional random coefficients, resulting in a stochastic multiscale finite element method (StoMsFEM).

Principal Component Analysis With Sparse/Localized Loadings

Given a random field $\kappa : D \times \Omega \to \mathbb{R}$, where $D \subset \mathbb{R}^d$ is the physical domain and $(\Omega, \mathcal{F}, P)$ is a probability space, its mean field $\bar{\kappa} : D \to \mathbb{R}$ and covariance function $K : D \times D \to \mathbb{R}$ are defined by

$$\bar{\kappa}(x) := \mathbb{E}[\kappa(x, \omega)], \quad K(x, y) := \mathbb{E}[(\kappa(x, \omega) - \bar{\kappa}(x))(\kappa(y, \omega) - \bar{\kappa}(y))].$$

Its covariance operator, denoted as $\mathcal{K} : L^2(D) \to L^2(D)$, is the Hilbert-Schmidt operator with kernel $K(x, y)$. Many basic operations on a random field (e.g., sampling, factorization, marginalization, and conditioning) require a finite rank approximation of its covariance operator, which can be obtained in different ways.

Eigendecomposition

The eigendecomposition is the standard method to obtain the best low rank approximation of a PSD operator. One such example is the truncated Karhenen–Loève (KL) expansion [75, 85], which is the most widely used method to factorize a random field. The truncated KL expansion first computes the best rank-n approximation of the covariance operator by the truncated eigendecomposition, i.e.,

$$\mathcal{K} \approx \sum_{i=1}^{n} \lambda_i e_i e_i^T,$$

where $\{\lambda_i\}_{i=1}^n$ are the largest n eigenvalues and $\{e_i\}_{i=1}^n$ are the corresponding eigenfunctions. Without loss of generality, we assume that the random field is centered, i.e., $\bar{\kappa}(x) \equiv 0$, in this subsection. Then the random field is approximately factorized by its truncated KL expansion

$$\kappa(x, \omega) \approx \sum_{i=1}^{n} \sqrt{\lambda_i} e_i(x) \boldsymbol{\xi}_i, \tag{1.17}$$

where $\boldsymbol{\xi}_i := \int_D k(x, \omega) e_i(x) \mathrm{d}x$ is the principal factor corresponding to the eigenfunction e_i. The truncated KL expansion enjoys the bi-orthogonality property, i.e.,

$$(e_i, e_j)_{L^2(D)} = \delta_{i,j}, \quad \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_j] = \delta_{i,j}. \tag{1.18}$$

The discrete version of the truncated KL expansion is the famous principal component analysis in statistics. However, the eigenfunctions are typically global, i.e., nonzero almost everywhere, and are sometimes difficult to interpret. The global eigenfunctions imply that even on a local patch, all the factors $\{\boldsymbol{\xi}_i\}_{i=1}^n$ have influence there. This seems to be counter intuitive. It seems to make more sense that a random variable would impact only a small number of patches. In this case, we can apply our ISMD or Sparse OC to obtain the low rank approximation with sparse/localized modes.

The ISMD

When the covariance operator $\mathcal{K}$ is (nearly) rank-n, we can use the ISMD to decompose it into n sparse rank-one components

$$\mathcal{K} \approx \sum_{i=1}^{n} \psi_i \psi_i^T,$$

where $\{\psi_i\}_{i=1}^n$ are the intrinsic sparse modes. After projecting the random field to the space spanned by $\{\psi_i\}_{i=1}^n$, we obtain a sparse-orthogonal factorization

$$\kappa(x, \omega) \approx \sum_{i=1}^{n} \psi_i(x) \boldsymbol{\eta}_i, \tag{1.19}$$

where $\boldsymbol{\eta}_i$ is the random factor associated with the intrinsic sparse mode ψ_i. One can easily check that the random factors $\{\boldsymbol{\eta}_i\}_{i=1}^n$ are uncorrelated with each other, i.e.,

$$\mathbb{E}[\boldsymbol{\eta}_i \boldsymbol{\eta}_j] = \delta_{i,j}. \tag{1.20}$$

Therefore, compared with the bi-orthogonality (1.18) in the truncated KL expansion, the factorization (1.19) obtained from the ISMD requires orthogonality only in the stochastic space but sparsity in the physical space.

The sparse-orthogonal factorization (1.19) is closely related to the latent factor model with sparse loadings, which has found many applications ranging from DNA microarray analysis [48], facial and object recognition [129], and web search models [1], etc. In some scenarios, the uncorrelated latent factors make lots of sense, but is not guaranteed by many existing factorization methods, e.g., the non-negative matrix factorization (NMF) [80], the sparse principal component analysis (SPCA) [73, 137, 37], the structured SPCA [72]. We recommend the ISMD for sparse factorization problems where the covariance operator is (nearly) low rank and the uncorrelated constraint on the factors is imposed.

In Section 5.5, we provide such an application using covariance matrices from porous media problems [50, 46]. We compare the ISMD with other existing methods, e.g., eigendecomposition, pivoted Cholesky decomposition and convex relaxation of the SPCA (see [78, 128]). Our results demonstrate the superiority of the ISMD when decomposing (nearly) low-rank PSD matrices with sparse/localized modes.

The Sparse OC

We can also apply our Sparse OC to obtain a rank-n approximation for the covariance operator, i.e.,

$$\mathcal{K} \approx \Psi^{loc} K_n \Psi^{loc,T} = \sum_{i,j=1}^{n} K_n(i,j) \psi_i^{loc} \psi_j^{loc,T},$$

which achieves both high approximation accuracy and basis localization at the same time. After projecting the random field to the space spanned by Ψ^{loc}, we obtain a sparse factorization

$$\kappa(x,\omega) \approx \sum_{i=1}^{n} \psi_i^{loc}(x) \boldsymbol{\eta}_i, \tag{1.21}$$

where $\boldsymbol{\eta}_i$ is the random factor associated with the sparse mode ψ_i^{loc}. In this case, the random factors $\{\boldsymbol{\eta}_i\}_{i=1}^n$ are correlated with each other, with correlation

$$\mathbb{E}[\boldsymbol{\eta}_i \boldsymbol{\eta}_j] = K_n(i,j). \tag{1.22}$$

The Sparse OC is especially suitable for the Matérn class covariance [92]. In spatial statistics, geostatistics, machine learning and image analysis, the Matérn class covariance is used to model random fields with smooth samples; see e.g. [121, 58, 53]. The Matérn class covariance between two points $x, y \in D \subset \mathbb{R}^d$ is given by

$$K_\nu(x,y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right), \tag{1.23}$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and ρ and ν are non-negative parameters of the covariance. Two special cases are the exponential kernel when $\nu = 1/2$, i.e., $K_{1/2}(x,y) = \sigma^2 \exp(-|x-y|/\rho)$, and the Gaussian kernel when $\nu \to \infty$, i.e., $\lim_{\nu\to\infty} K_\nu(x,y) = \sigma^2 \exp(-\frac{|x-y|^2}{2\rho^2})$. The Fourier transform of Matérn class covariance is given by

$$\widehat{k}(\omega) = c_{\nu,\lambda} \sigma^2 \left(\frac{2\nu}{\lambda^2} + |\omega|^2 \right)^{-(\nu+d/2)}, \quad c_{\nu,\lambda} := \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)(2\nu)^\nu}{\Gamma(\nu)\lambda^{2\nu}}, \tag{1.24}$$

where we use the convention $\widehat{f}(\omega) = \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\omega} \mathrm{d}x$ for the Fourier transform. Recent studies [83, 16] show that the Matérn covariance and the elliptic operators are closely connected. With proper homogeneous boundary conditions, the Matérn covariance operator with $\nu + d/2$ as an integer is the solution

operator of an elliptic operator of order $2\nu + d$. For example, the exponential kernel (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 $(8\pi\rho^3\sigma^2)^{-1}(1 - 2\rho^2\Delta + \rho^4\Delta^2)$ when $d = 3$. The Matérn covariance operator with $\nu = 1$ is the solution operator of the fourth order elliptic operator $(4\pi\rho^2\sigma^2)^{-1}(1 - 2\rho^2\Delta + \rho^4\Delta^2)$ when $d = 2$. Note that the elliptic operator that is associated with the Matérn covariance contains lower order terms. Thus, it is essential that our Sparse OC can accommodate lower order terms and various boundary conditions.

In Section 2.4, we have applied the Sparse OC to compress the 1D Matérn kernel with $\nu = 1/2$. We have demonstrated that the Sparse OC achieves the optimal approximation accuracy and the nearly optimal localization. In this application, we have also shown that the logarithmic factor in the localization (i.e., localized basis functions with support diameter $\mathcal{O}(h\log(1/h))$) is necessary to obtain the optimal approximate accuracy under the framework of Sparse OC.

Constructing Localized Wannier Functions

Motivated by the localized Wannier functions developed in solid state physics and quantum chemistry, our Sparse OC can serve as a computationally efficient method to construct the localized Wannier functions. We begin by reviewing the basic ideas for obtaining spatially localized basis functions of the independent-particle Schrödinger's equation. For simplicity, we consider a finite system with n electrons and neglect the electron spin. The ground state energy is given by $E_n = \sum_{i=1}^{n} \lambda_i$, where λ_i are the eigenvalues of Hamiltonian, $\mathcal{H} = -\frac{1}{2}\Delta + V(x)$, arranged in increasing order and satisfying $\mathcal{H}e_i = \lambda_i e_i$, with e_i being the corresponding eigenfunctions. A basic task in computational chemistry is to compute the ground state energy E_n, which in turn requires the eigendecomposition of the Hamiltonian $\mathcal{H}$. Notice that the Hamiltonian is nothing but a second order elliptic operator with a nontrivial potential $V(x)$.

In most cases, the eigenfunctions e_i are nonzero almost everywhere, i.e., they are "dense". This presents challenges for computational efficiency since the eigendecomposition requires $\mathcal{O}(n^3)$ operations, dominating the computational effort for $n \approx 10^3$ and above. Moreover, the screened correlations in con-

densed matter are usually short-ranged (see [111]). It is well known that an appropriate linear transformation of the eigenfunctions could lead to a set of more spatially localized eigenfunctions that span the same eigenspace of $\mathcal{H}$. Methods for obtaining such functions have been developed in solid state physics and quantum chemistry, where they are known as Wannier functions (see [130, 76]). However, simply applying a linear transformation on the eigenfunctions has two disadvantages. First of all, in most cases, one can obtain at most polynomially decaying functions by using linear transformations on eigenfunctions. For example, the eigenfunctions of the 1D free electron are harmonic waves, and no linear combinations of the first n harmonic waves can result in compactly supported or exponentially decaying functions! Secondly, it still requires obtaining the eigenfunctions at the first place, which does not help the computational efficiency.

To achieve truly localized Wannier functions and computational efficiency, one needs to look for an approximate n-dimensional subspace that is spanned by n localized basis functions and gives an accurate approximation of the eigenspace at the same time. Our Sparse OC is a natural choice to achieve this goal. We point out that there is no consensus on the norm to measure the distance between the constructed n-dimensional subspace Ψ and the eigenspace $V_n := \text{span}\{e_1, \ldots, e_n\}$ in the existing literature. Our Sparse OC measures the distance by the operator compression error $E_{oc}(\Psi; \mathcal{H}^{-1})$ (1.6) (we can always add a constant to $\mathcal{H}$ such that it is invertible, and this does not change the eigenspace of $\mathcal{H}$). Another natural choice to define the distance between the approximate space Ψ and the eigenspace V_n is

$$\widetilde{E}_{oc}(\Psi) = \|\mathcal{P}_{V_n} - \mathcal{P}_{\Psi}\|_2, \tag{1.25}$$

where $\mathcal{P}_V$ is the orthogonal projection from $L^2(D)$ to the subspace V. To compare these two criteria, we rewrite them in a form that makes the comparison easier:

$$E_{oc}(\Psi; \mathcal{H}^{-1}) = \min_{K_n \succeq 0} \left\| \sum_{i=1}^{\infty} \frac{1}{\lambda_i} e_i e_i^T - \Psi \mathcal{K}_n \Psi^T \right\|_2, \quad \widetilde{E}_{oc}(\Psi) = \left\| \sum_{i=1}^{n} e_i e_i^T - \mathcal{P}_{\Psi} \right\|_2.$$

We believe that $E_{oc}(\Psi; \mathcal{H}^{-1})$ is a better criterion for operator compression because it takes into consideration the relative importance $(1/\lambda_i)$ of different eigenfunctions e_i. In contrast, the quantity $\|\mathcal{P}_{V_n} - \mathcal{P}_{\Psi}\|_2$ gives equal weight to all the first n eigenfunctions and does not count the rest of the eigenfunctions at all.

One of the most commonly used methods to construct localized Wannier functions is the l_1 approach, which is inspired by the compressed sensing and is briefly reviewed in Section 2.5. Our Sparse OC is inspired by the recent advances in numerical homogenization [98, 101], and has a very different philosophy from the l_1 approach. We compare our Sparse OC with the l_1 approach on the free-electron model in Section 2.5. We summarize the main conclusions here.

1. With roughly the same support size of the localized basis functions, the results given by the l_1 approach and the Sparse OC are very similar, in terms of the shape of the function and the approximate eigenvalues.

2. The total computation cost of the Sparse OC is comparable to the cost of *each iteration* in the l_1 approach. The l_1 approach needs hundreds (sometimes even thousands) of iterations to solve the nonconvex problem, depending on the choice of the algorithm parameters.

Application in Stochastic Multiscale Model Reduction

Finally, we have applied the ISMD and other sparse factorization methods to solve the following random elliptic equation

$$\begin{cases} -\nabla_x \cdot (\kappa(x,\omega)\nabla_x u(x,\omega)) = f(x), \quad x \in D, \omega \in \Omega, \\ u(x,\omega) = 0, \quad x \in \partial D \end{cases} \qquad P\text{-almost surely}, \tag{1.26}$$

where the random coefficients $\kappa(x,\omega)$ have a large stochastic dimension and have multiscale features in the physical space. Here, the "stochastic dimension" means the number of the factors/parameters in the factorization of the random coefficients. As usual, we assume that $\kappa(x,\omega)$ satisfies $\kappa(x,\omega) \geq \alpha > 0$, for every $x \in D$ and $\omega \in \Omega$. The random elliptic equation (1.26) is the fundamental model to simulate flows in heterogeneous porous media, whose permeability is often modeled as a multiscale random field. The parametrization of a multiscale random medium requires a large number of random variables, leading to a random elliptic equation with a high stochastic dimension, which is challenging to solve numerically.

In [65], we have proposed a stochastic multiscale finite element method (StoMsFEM) that combines a localized factorization method and a deterministic model reduction method. The StoMsFEM can significantly speed up the exist-

ing *non-intrusive* stochastic methods. By "non-intrusive stochastic methods", we mean those methods that can call a deterministic PDE solver as a blackbox, e.g., Monte Carlo, multilevel Monte Carlo [51, 9, 32], (sparse grid) stochastic collocation [4, 96, 97, 133], least-squares methods [40, 120] and compressed sensing methods [41, 134]. The StoMsFEm is based on the following observation: most deterministic model reduction methods only require solving local problems (see e.g. [66, 67, 105, 89, 99]), and the local problems often have much lower stochastic dimensions. More precisely, the random coefficients restricted to a local subdomain can be factorized by a much smaller number of parameters, which depends only on the ratio between the subdomain size and the correlation length of the random coefficients. Therefore, the deterministic model reduction methods result in solving local subproblems with low stochastic dimensions, whose solutions can be efficiently precomputed by the gPC based methods (see e.g., [49, 132, 7, 4]) in the offline stage.

Based on this observation, the proposed StoMsFEM solves the random PDEs in three steps. The first two steps are in the offline stage and the third step is in the online stage. In the first step, we obtain a locally low dimensional parametrization of the random coefficients $\kappa(x, \omega)$. This can be achieved by our ISMD or Sparse OC, as well as other localized factorization methods like the local truncated KL expansion [27] and the SPCA [137, 37, 128]. In the second step, we apply a deterministic local upscaling method to obtain a *parametric upscaled system*. We provide two methods to do the parametric upscaling: random interpolation method and reduced basis method. The random interpolation method takes advantage of the fact that the local upscaled coefficients are analytic functions of the local random parameters, and we introduce an interpolation scheme for each upscaled coefficient at the coarse-grid level. The random interpolation method can be viewed as a local reduced-order method in the stochastic space. The reduced basis method makes use of the low rank property of the solutions for the local upscaling problems, and prepares a small set of spatial basis functions for each local upscaling problem. The reduced basis method can be viewed as a local reduced-order method in the physical space. In the online stage (i.e., the third step), for each sample of the random parameters, we either interpolate the upscaled coefficients in the random interpolation setting, or solve the small reduced-order systems to obtain the upscaled coefficients. A numerical coarse-grid solution for this sample can be obtained by solving the upscaled system.

In Section 5.5, we utilize our ISMD to obtain a locally low dimensional parametrization of the random coefficients $\kappa(x, \omega)$, which is the first step of the StoMsFEM. We refer to our paper [65] for detailed information about the analysis and implementation of the StoMsFEM.

1.5 Roadmap of the Thesis

The thesis is organized as follows.

- In Chapter 2, we present the general framework of the sparse operator compression (Sparse OC) and its three applications. In Section 2.3, we provide the application of the Sparse OC to second order elliptic operators. In Section 2.4, we apply the Sparse OC to compress the 1D exponential kernel. In Section 2.5, we apply the Sparse OC to construct localized Wannier functions for the 1D free-election model.

- In Chapter 3, we apply the Sparse OC to compress the higher order strongly elliptic operators. We first prove the local projection-type approximation in the Sobolev spaces H^k and the corresponding local inverse energy estimate. Then we introduce the concept of strong ellipticity, and show its relation with the uniform ellipticity. After that, we prove the exponential decay of the global energy minimizing basis functions $\psi_{i,q}$ and then localize it to obtain the localized basis function $\psi_{i,q}^{loc}$. Finally, both 1D and 2D examples are provided to validate our theoretical results.

- In Chapter 4, we apply the Sparse OC to the second order elliptic operators with high contrast coefficients. We first obtain a local projection-type approximation property from a local generalized eigenvalue problem, and prove the corresponding local inverse energy estimate for the two-phase coefficients. Then we prove that the global energy minimizing basis functions $\psi_{i,q}$ decays exponentially fast away from its associated patch, and the decay rate is independent of the contrast. Finally, a 2D example with high permeability channels is provided to demonstrate the contrast-independent decay rate.

- In Chapter 5, we present the intrinsic sparse mode decomposition (ISMD) for low rank PSD operators. We first present our ISMD algorithm for low rank matrices and analyze its computational complexity. Then our main theoretical results are presented. After that, we discuss the stability of

the ISMD by performing perturbation analysis. Finally, we present a few numerical examples to demonstrate the efficiency of the ISMD and compare its performance with other existing methods.

- We make some concluding remarks in Chapter 6 and outline several future directions.

Chapter 2

SPARSE OPERATOR COMPRESSION AND ITS APPLICATIONS

The main purpose of this chapter is to develop a general framework, i.e., the sparse operator compression (Sparse OC), to compress positive semidefinite (PSD) operators with localized basis functions. We will also talk about three applications of the Sparse OC: to solve elliptic equations with rough coefficients, to solve the the problem of sparse principal component analysis (SPCA), and to construct localized Wannier functions.

2.1 Problem Setting

Suppose $\mathcal{K} : X \to X$ is a bounded self-adjoint PSD operator, where X can be any separable Hilbert space with inner product $(\cdot, \cdot)$. We ask the question: given an integer n, what is the best rank-n approximation of the operator $\mathcal{K}$ with localized basis functions? This question arises in many different contexts.

Consider the elliptic equation with the homogeneous Dirichlet boundary conditions

$$\mathcal{L}u := -\nabla \cdot (a\nabla u) = f, \qquad u \in H_0^1(D), \tag{2.1}$$

where the load $f \in L^2(D)$. If a is symmetric and uniformly elliptic on D, i.e., there exist $0 < a_{min} \leq a_{max}$ such that

$$a_{min} I_d \preceq a(x) \preceq a_{max} I_d, \qquad x \in D.$$

Then the Lax–Milgram Lemma (see e.g. [31]) implies that Eqn. (1.1) has a unique weak solution in $H_0^1(D)$, denoted as $\mathcal{L}^{-1} f$. This defines the solution operator $\mathcal{L}^{-1} : L^2(D) \to L^2(D)$. Under the framework of the Galerkin finite element method, given n basis functions $\Psi = [\psi_1, \ldots, \psi_n]$ in $H_0^1(D)$, we can solve the elliptic equation (1.1) approximately by projecting it onto the subspace spanned by Ψ, i.e.,

$$u_\Psi = \Psi L_n^{-1} \Psi^T f,$$

$L_n \in \mathbb{R}^{n \times n}$ and $\Psi^T f \in \mathbb{R}^n$ are the stiffness matrix and the loading vector, respectively. As a candidate for the finite element basis, Ψ should minimize

the solution error in the worst scenario:

$$\|\mathcal{L}^{-1} - \Psi L_n^{-1} \Psi^T\|_2 := \sup_{\|f\|_{L^2(D)} \leq 1} \|\mathcal{L}^{-1} f - \Psi L_n^{-1} \Psi^T f\|_{L^2(D)},$$

where $\|A\|_2$ is the largest eigenvalue for a PSD operator A.

Therefore, for a self-adjoint, positive definite operator $\mathcal{K} : X \to X$ (the solution operator $\mathcal{L}^{-1} : L^2(D) \to L^2(D)$ when solving elliptic equations), we define the operator compression error of the basis Ψ as

$$E_{oc}(\Psi; \mathcal{K}) := \min_{K_n \in \mathbb{R}^{n \times n},\ K_n \succeq 0} \|\mathcal{K} - \Psi K_n \Psi^T\|_{X,X}, \tag{2.2}$$

where $\|\mathcal{K}\|_{X,X} := \sup_{\|f\|_X \leq 1} \frac{\|\mathcal{K} f\|_X}{\|f\|_X}$ is the induced operator norm for the bounded operator $\mathcal{K} : X \to X$. The operator compression error is the optimal approximation error of $\mathcal{K}$ among all positive semidefinite operators with range space spanned by Ψ. It is easy to verify that the operator compression error only depends on the subspace spanned by Ψ, still denoted as Ψ, and is independent of the choice of its basis function Ψ. When solving the the problem of sparse principal component analysis (SPCA), $\mathcal{K}$ will be the covariance operator of a random field/vector; see Section 2.4. When constructing localized Wannier functions in quantum chemistry, $\mathcal{K}$ will be the inverse of the Hamiltonian; see Section 2.5. In all the current applications, X is a Hilbert space over some bounded Lipschitz domain $D \subset \mathbb{R}^d$, such as $L^2(D)$ and $L^2_a(D)$ ($L^2(D)$ with a-weighted inner product).

Without imposing the sparsity constraints on the basis Ψ, the compression error $E_{oc}(\Psi; \mathcal{K})$ achieves its minimum $\lambda_{n+1}(\mathcal{K})$ if we use the first n eigenfunctions (corresponding to the largest n eigenvalue) of $\mathcal{K}$ to form Ψ. However, the eigenfunctions are expensive to compute, and do not have localized support, [137, 107, 64]. In many cases, localized/sparse basis functions are preferred. For example, in the multiscale finite element method [45], localized basis functions lead to sparse linear systems, and thus result in more efficient algorithms; see e.g. [6, 66, 123, 67, 3, 44, 43, 89, 105, 99, 29]. In quantum chemistry, localized basis functions like the Wannier functions have better interpretability of the local interactions between particles (see e.g. [90, 42, 91, 107, 78]), and also lead to more efficient algorithms [54]. In statistics, the sparse principal component analysis (SPCA) looks for sparse vectors to span the eigenspace of the covariance matrix, which in many cases leads to better interpretability

compared with the dense principal components from the PCA; see e.g. [73, 137, 37, 131, 128].

Discussions On The Definition Of Operator Compression Error

Suppose the eigen-pairs of $\mathcal{K}$ are $\{(\lambda_i, e_i)\}_{i=1}^{\infty}$, where $\{\lambda_i\}_{i=1}^{\infty}$ are in a descending order. Another natural choice to define the compression error is the distance between the constructed space Ψ and the first n-dimensional eigenspace $V_n = \text{span}\{e_1, \ldots, e_n\}$:

$$\widetilde{E}_{oc}(\Psi) = \|\mathcal{P}_{V_n} - \mathcal{P}_{\Psi}\|_2, \tag{2.3}$$

where $\mathcal{P}_V$ is the orthogonal projection from $L^2(D)$ to its subspace V. To compare criterion (2.2) with criterion (2.3), we rewrite them in a form that makes the comparison easier:

$$E_{oc}(\Psi; \mathcal{K}) = \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \left\| \sum_{i=1}^{\infty} \lambda_i e_i e_i^T - \Psi \mathcal{K}_n \Psi^T \right\|_2,$$
$$\widetilde{E}_{oc}(\Psi) = \left\| \sum_{i=1}^{n} e_i e_i^T - \mathcal{P}_{\Psi} \right\|_2.$$

We believe that $E_{oc}(\Psi; \mathcal{K})$ is a better criterion for operator compression because it takes into consideration the decay of the eigenvalues of the solution operator $\mathcal{K}$. In contrast, the quantity $\|\mathcal{P}_{V_n} - \mathcal{P}_{\Psi}\|_2$ gives equal weight to all eigenfunctions and does not take into account the relative importance of different eigenfunctions.

Due to the unboundedness of the elliptic operators $\mathcal{L}$, we use $\mathcal{L}^{-1}$ to define its operator compression error. The compression error $E_{oc}(\Psi; \mathcal{L}^{-1})$ can be extended to any uniform elliptic operator. By the Garding's inequality [113], there exists $\lambda_G > 0$ such that $\mathcal{L} + \lambda_G$ satisfies the assumptions of the Lax-Milgram lemma, and thus its inverse operator $(\mathcal{L} + \lambda_G)^{-1}$ exists. Using $E_{oc}(\Psi; (\mathcal{L} + \lambda_G)^{-1})$ to quantify the compression error is useful for operators that are not invertible, such as $-\Delta$ with periodic boundary conditions.

Constructing Basis Functions

The Sparse OC follows three steps to construct sparse/localized basis functions $\Psi^{loc} := [\psi_1^{loc}, \ldots, \psi_n^{loc}]$. In the first step, we partition the physical domain D with a regular finite element mesh with mesh size $h > 0$, and denote all the elements (local patches) as $\{\tau_i\}_{i=1}^{m}$. On each local patch τ_i, we choose Q_i

($Q_i \in \mathbb{N}$) measurement functions that are only supported on τ_i, denoted as $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$.

In the second step, we construct non-local basis functions $\{\psi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}$ by the following minimizing problem:

$$\begin{aligned} \psi_{i,q} = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_{j,q'}) = \delta_{iq,jq'}, \quad \forall 1 \leq j \leq m,\, 1 \leq q' \leq Q_j, \end{aligned} \tag{2.4}$$

where $\delta_{iq,jq'}$ is the Kronecker delta that is 1 when $i = j$ and $q = q'$ and 0 in all other cases. Here, H is a Hilbert space that is called the Cameron-Martin space and is formally defined in Section 2.2. When $\mathcal{K} = \mathcal{L}^{-1}$, the Cameron-Martin 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 be applied to any self-adjoint PSD operator $\mathcal{K}$. Collecting all the nonlocal basis functions $\psi_{i,q}$ for $1 \leq i \leq m$ and $1 \leq q \leq Q_i$, we get a nonlocal basis Ψ. Although $\psi_{i,q}$ is not localized, we will see that it decays exponentially fast away from its associated patch.

In the final step, we restrict the global construction onto a neighborhood of the patch τ_i with radius r, denoted as $S_r(\tau_i)$ (see Figure 2.1), and obtain a localized basis function:

$$\begin{aligned} \psi_{i,q}^{loc} = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_{j,q'}) = \delta_{iq,jq'}, \quad \forall 1 \leq j \leq m,\, 1 \leq q' \leq Q_j, \\ & \psi(x) \equiv 0, \quad x \in D \backslash S_r. \end{aligned} \tag{2.5}$$

Collecting all the $\psi_{i,q}^{loc}$ for $1 \leq i \leq m$ and $1 \leq q \leq Q_i$ together, we get our local basis Ψ^{loc}. In all our applications, we take $r = Ch(\log(1/h) + \log(\text{Contrast}))$, in which C is a constant independent of the coefficients and Contrast is the contrast of the coefficients. We will discuss the choice of r in detail when we analyze the compression error of Ψ^{loc}.

We point out that when implementing the Sparse OC, there is no need to compute the global basis Ψ (2.4). Conversion of the global basis Ψ to the local basis Ψ^{loc} is useful when we do theoretical analysis of the Sparse OC.

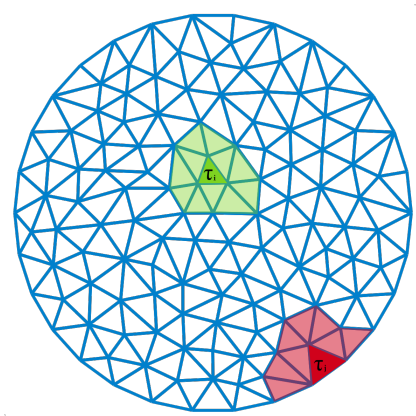


Figure 2.1: A regular partition, local patch τ_i and its associated S_r.

Comparison With Other Existing Methods

Our approach for operator compression originates in 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 [6, 66, 123, 45, 3, 43, 89, 105, 98, 99, 29]. Specifically, our work is inspired by the work presented in [89, 99], 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 [99] 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 [99] as the direct template for our method, to the best of our knowledge, the Local Orthogonal Decomposition (LOD) [89], in the context of multi-dimensional numerical homogenization, contains the first rigorous proof of optimal exponential decay rates with a priori estimates (leading to localization to sub-domains 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 [56], although it did not provide a constructive local basis. In addition to establishing the exponential decay of the basis (for general non-conforming 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 [99] 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 3.3.1 is an essential step to obtaining the inverse energy estimate for higher order PDEs that is not found in [89] nor [99]. We remark that Lemma 3.12 in [99] 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 [99] 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 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 [100] introduce the Gaussian cylinder measure and successfully generalize the probabilistic framework in [98, 99] to a much broader class of operators, including higher order elliptic operators without requiring strong ellipticity.

As in [89, 99], the error bound in our convergence analysis blows up for fixed oversampling ratio r/h. To achieve the desired $O(h)$ accuracy, 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. [61, 109]. In particular, the authors of [61, 109] 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 [61, 109] proved that the 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. [3, 63]. However, they need to use $O(\log(1/h))$ basis functions associated with each coarse finite element to

recover the $O(h)$ accuracy. It is worth mentioning that the authors of [63] use a local oversampling operator to construct the optimal local boundary conditions for the nodal multi-scale basis and enrich the nodal multis-scale basis with optimal edge multi-scale basis. Moreover, the method in [63] allows an explicit control of the approximation accuracy in the offline stage by truncating the SVD of the oversampling operator. In [63], 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 framework (GMsFEM) [43, 29] 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:

$$\begin{aligned} \min_{\Psi} \quad & \sum_{i=1}^{n} \|\psi_i\|_H^2 + \lambda \sum_{i=1}^{n} \|\psi_i\|_1, \\ \text{s.t.} \quad & (\psi_i, \psi_j) = \delta_{i,j} \quad \forall 1 \leq i, j \leq n, \end{aligned} \tag{2.6}$$

where $\|\psi_i\|_H$ is the energy norm induced by the operator $\mathcal{L}$. In problem (2.6), small $\|\psi_i\|_H$ leads to a small compression error, small $\|\psi_i\|_1$ enforces a sparse basis function, and $\lambda > 0$ is a parameter to control the trade-off between the accuracy and sparsity.

The sparse principal component analysis (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 Eqn. (2.6):

$$\begin{aligned} \min_{\Psi} \quad & -\sum_{i=1}^{n} (\psi_i, \mathcal{K}\psi_i) + \lambda \sum_{i=1}^{n} \|\psi_i\|_1, \\ \text{s.t.} \quad & (\psi_i, \psi_j) = \delta_{i,j} \quad \forall 1 \leq i, j \leq n, \end{aligned} \tag{2.7}$$

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 (2.7), 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 (2.6) nor the SPCA (2.7) is convex, and one needs to use some sophisticated techniques to solve the non-convex optimization problem or its convex relaxation; see e.g. [137, 37, 107, 128, 78].

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 Eqn. (2.5). Moroever, our construction (2.5) is a quadratic optimization with linear constraints, which can be solved as efficiently as solving an elliptic problem on the local domain S_τ. The computational complexity to obtain all n localized basis functions $\{\psi_i^{loc}\}_{i=1}^n$ 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 [99]. In contrast, the orthogonality constraint in Eqn. (2.6) is not convex, which introduces additional difficulties in solving the problem. Finally, our construction of $\{\psi_i^{loc}\}_{i=1}^n$ is completely decoupled, while all the basis functions in Eqn. (2.6) are coupled together. This decoupling leads to a simple parallel execution, and thus makes the computation of $\{\psi_i^{loc}\}_{i=1}^n$ even more efficient.

Outline Of This Chapter

In Section 2.2, we provide an abstract and general framework to construct the finite element space Ψ, to analyze the approximation accuracy, and to construct a localized basis of the space Ψ. In Section 2.3, we provide the application of the Sparse OC to second order elliptic operators, and reproduce the result that has been obtained in [99]. In Section 2.4, we apply the Sparse OC to compress the 1D exponential kernel. In Section 2.5, we apply the Sparse OC to construct localized Wannier functions for the 1D free-election model.

2.2 An Abstract Framework of Sparse 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 operation 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. A probabilistic framework for Bayesian numerical homogenization has been proposed in [98], but it requires the existence of a Gaussian measure with certain given covariance operator. Our following framework is purely based on functional analysis, which applies to any bounded self-adjoint positive semidefinite operators.

First of all, we introduce the Cameron–Martin space, which plays the role of

the range space of $\mathcal{K}$. Secondly, we provide our main theorem to estimate the compression error. Thirdly, we give the abstract form to construct basis functions by minimizing their energies. Finally, we give the abstract form to construct localized basis functions. We use this abstract framework to compress elliptic operators in the rest of the thesis.

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 $\{e_n)\}_{n=1}^{\infty}$ can be chosen to be an orthonormal basis of X.

Lemma 2.2.1. *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*

$$(\mathcal{K}\varphi_1, \mathcal{K}\varphi_2)_H = (\mathcal{K}\varphi_1, \varphi_2) \qquad \forall \varphi_1, \varphi_2 \in X. \tag{2.8}$$

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.* $\text{null}(\mathcal{K}) = \{\mathbf{0}\}$.

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 non-negativity 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 $\varphi \in \text{null}(\mathcal{K})$, i.e. $\mathcal{K}\varphi = 0$.

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

$$\sqrt{(\mathcal{K}\varphi, \mathcal{K}\varphi)} \leq \sqrt{\lambda_1 (\mathcal{K}\varphi, \varphi)} = \sqrt{\lambda_1} \sqrt{(\mathcal{K}\varphi, \mathcal{K}\varphi)_H}, \tag{2.9}$$

   where we have used the definition of $(\cdot, \cdot)_H$ in Eqn. (2.8) in the last step.

3. If $\text{null}(\mathcal{K}) = \{\mathbf{0}\}$, we have $\text{span}\{e_n, n \geq 1\} \subset \mathcal{K}(X)$. 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 2.2.2. *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

$$(f, \mathcal{K}\psi)_H = (f, \psi). \tag{2.10}$$

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 Eqn. (2.9) and the the continuity of norms, we have $(\psi, \psi) \leq \lambda_1(\psi, \psi)_H$ for any $\psi \in H$.

2. It is obvious from item 3 in Lemma 2.2.1.

3. If $f \in \mathcal{K}(X)$, Eqn. (2.10) is exactly the definition of $(\cdot,\cdot)_H$ in Eqn. (2.8). By the continuity of inner product, Eqn. (2.10) is true for any $f \in H$.

□

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 thesis, $\mathcal{P}_\Phi^{(H)}$ denotes the orthogonal projection from a Hilbert space H to its subspace Φ. With this notation, we present our theorem for error estimate as follows.

Theorem 2.2.1. *Suppose there is a n-dimensional subspace $\Phi \subset X$ with basis $\{\varphi_i\}_{i=1}^n$ such that*

$$\|u - \mathcal{P}_\Phi^{(X)} u\|_X \leq k_n \|u\|_H \qquad \forall u \in \mathcal{K}(X) \subset H. \tag{2.11}$$

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

$$\|u - \mathcal{P}_\Psi^{(H)} u\|_H \leq k_n \|f\|_X \,. \tag{2.12}$$

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

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_X \leq k_n^2 \|f\|_X \,. \tag{2.13}$$

3. *We have*

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \leq k_n^2 \,, \tag{2.14}$$

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 2.2.1, based on a projection type approximation property of Φ in H, i.e. Eqn. (2.11), we obtain the error estimates of the GFEM with finite element basis $\{\mathcal{K}\varphi_i\}_{i=1}^n$ in the energy norm, i.e. Eqn. (2.12). 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 is pioneered by the GFEM [6, 123] and the multiscale finite element method (MsFEM) [66, 70], and is pervasive in recent developments in finite element methods.

Our error analysis is very different from the traditional error analysis for the GFEM from 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:

$$|u - \mathcal{I}_h u|_{1,2,D} \leq Ch|u|_{2,2,D} \quad \forall u \in H_0^2(D), \tag{2.15}$$

where $\mathcal{I}_h u$ is the piecewise linear interpolation of the solution u. In Eqn. (2.15), one assumes $u \in H^2(D)$, but it is not the case for elliptic operators with rough coefficients. Secondly, in our projection type approximation property (2.11) 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 [99].

Proof. **[Proof of Theorem 2.2.1]**

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 2.2.2, we have

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

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

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

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,

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

On the other hand,

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

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 let's 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

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

We used the definition of H-norm again in the last step.

□

Energy Minimizing Basis Functions

By Theorem 2.2.1, the space Ψ spanned by $\{\mathcal{K}\varphi_i\}_{i=1}^n$ can compress the operator $\mathcal{K}$ well. In this subsection, we will construct another set of basis functions $\{\psi_i\}_{i=1}^n$ for Ψ via a variational approach. Although these two sets of basis functions span the same space Ψ, their decaying property is very different as we will see later.

For any given $i \in \{1, 2, \ldots, n\}$, consider the following quadratic optimization problem

$$\begin{aligned} \psi_i = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_j) = \delta_{i,j}, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{2.16}$$

Due to the strong convexity of $\|\cdot\|_H$, the minimizer of Eqn (2.16) is unique if there exists one. For the existence, it is important to notice that the constraints in (2.16) are equivalent to $(\psi, \mathcal{K}\varphi_j)_H = \delta_{i,j}$ for all $j = 1, 2, \ldots, n$. Then it is easy to verify that if and only if $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent, i.e. $\text{null}(\mathcal{K}) \cap \Phi = \{\mathbf{0}\}$, the constraints in (2.16) are consistent for all $i = 1, 2, \ldots, n$. Therefore, when $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent, the unique minimizer of Eqn. (2.16), denoted as ψ_i, is well-defined. For completeness, we provide the proof of existence here.

Proposition 2.2.2. *For any $1 \le i \le n$, there exists ψ_i such that $(\psi_i, \mathcal{K}\varphi_j)_H = \delta_{i,j}$ for all $1 \le j \le n$ if and only if $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent, i.e.* $\text{null}(\mathcal{K}) \cap \Phi = \{\mathbf{0}\}$.

Proof. If $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent, it is obvious that for any $i = 1, 2, \ldots, n$ there exists ψ_i such that $(\psi_i, \mathcal{K}\varphi_j)_H = \delta_{i,j}$ for all $j = 1, 2, \ldots, n$. In the other direction, we assume that for any $i = 1, 2, \ldots, n$ there exists ψ_i such that $(\psi_i, \mathcal{K}\varphi_j)_H = \delta_{i,j}$ for all $j = 1, 2, \ldots, n$. Suppose we have $\sum_{j=1}^n \alpha_j \mathcal{K}\varphi_j = \mathbf{0}$. Combined with the constraints, we have for any $i = 1, 2, \ldots, n$, $0 = \sum_{j=1}^n \alpha_j (\psi_i, \mathcal{K}\varphi_j)_H = \sum_{j=1}^n \alpha_j \delta_{i,j} = \alpha_i$. Therefore, we have $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent. □

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

$$\Theta_{i,j} := (\mathcal{K}\varphi_i, \varphi_j). \tag{2.17}$$

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 $\Theta_{i,j}^{-1}$ the (i,j)-th entry of

Θ^{-1}. Finally, we have the following explicit formula to characterize ψ_i, which is defined as the unique minimizer of Eqn (2.16).

Theorem 2.2.3. *If* $\text{null}(\mathcal{K}) \cap \Phi = \{\mathbf{0}\}$ *holds true, then we have*

1. *The optimization problem* (2.16) *admits a unique minimizer* ψ_i*, which can be written as*
$$\psi_i = \sum_{j=1}^{n} \Theta^{-1}_{i,j} \mathcal{K}\varphi_j. \tag{2.18}$$

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*
$$\|\psi\|_H^2 = \|\sum_{i=1}^{n} w_i\psi_i\|_H^2 + \|\psi - \sum_{i=1}^{n} w_i\psi_i\|_H^2. \tag{2.19}$$

3. $(\psi_i, \psi_j)_H = \Theta^{-1}_{i,j}$.

The intuition of Eqn. (2.18) is that we linearly transform the basis $\{\mathcal{K}\psi_i\}_{i=1}^n$ to form another basis for Ψ, and make sure that the new basis functions satisfy the constraints in (2.16). Noting that the constraints of ψ_i are equivalent to $(\psi_i, \mathcal{K}\varphi_j)_H = \delta_{i,j}$ for all $j = 1, 2, \ldots, n$, then it is obvious that ψ_i given in (2.18) solves the energy minimization problem (2.16).

Proof. Thanks to $\text{null}(\mathcal{K}) \cap \Phi = \{\mathbf{0}\}$, the problem (2.16) is feasible for all $i = 1, 2, \ldots, n$. Let Ψ be the n-dimensional subspace of H and $\Psi^\perp$ be the space of vectors orthogonal to it in H.

1. Write $\psi_i = \sum_{k=1}^n \alpha_{i,k}\mathcal{K}\varphi_k + \psi_i^\perp$ where $\psi_i^\perp \in \Psi^\perp$. Then the constraints of ψ_i become $\sum_k \Theta_{j,k}\alpha_{i,k} = \delta_{i,j}$ for all $j = 1, 2, \ldots, n$. Then we obtained $\alpha_{i,k} = \Theta^{-1}_{i,k}$. Note that $\|\psi_i\|_H^2 = \|\sum_{k=1}^n \alpha_{i,k}\mathcal{K}\varphi_k\|_H^2 + \|\psi_i^\perp\|_H^2$. Therefore, the minimizer of Eqn. (2.16), still denoted as ψ_i, is unique and can be explicitly written as (2.18).

2. One can verify that $\psi_w := \sum_{i=1}^n w_i\psi_i$ is the only function in Ψ satisfying the constraints $(\varphi_j, \psi) = w_j$ for $j = 1, 2, \ldots, n$. Therefore, for any ψ which satisfies $(\varphi_j, \psi) = w_j$ for $j = 1, 2, \ldots, n$, we have $(\varphi_j, \psi - \psi_w) = 0$, i.e. $(\mathcal{K}\varphi_j, \psi - \psi_w)_H = 0$, for $j = 1, 2, \ldots, n$. Therefore, $\psi - \psi_w$ is

orthogonal to the n-dimensional subspace Ψ, which contains ψ_w, in H, and thus Eqn. (2.19) holds true. Then the optimality of ψ_w naturally follows.

3. Combined (2.18) and the H-norm definition, we have

$$(\psi_i, \psi_j)_H = (\sum_k \Theta^{-1}_{i,k} \mathcal{K}\varphi_k, \psi_j).$$

By the constraints of ψ_j, i.e. $(\mathcal{K}\varphi_k, \psi_j)_H = \delta_{k,j}$ for all $k = 1, 2, \ldots, n$, we have $(\psi_i, \psi_j)_H = \sum_k \Theta^{-1}_{i,k}\delta_{k,j} = \Theta^{-1}_{i,j}$.

□

Remark 2.2.1. *If $\mathcal{K}$ is nuclear, i.e. $\sum_{i=1}^n \lambda_i < \infty$, the basis functions $\{\psi_i\}_{i=1}^n$ can be interpreted by conditioning a Gaussian measure on Hilbert space X; see [99]. In this case, there exists a Gaussian measure γ on X with mean $\mathbf{0}$ and covariance operator $\mathcal{K}$. Suppose $\boldsymbol{\xi}$ is a random vector in X and is distributed as γ. It can be proved that for any $\boldsymbol{w} \in \mathbb{R}^n$*

$$\mathbb{E}[\boldsymbol{\xi} \mid (\varphi_i, \boldsymbol{\xi}) = w_i \,,\, j = 1, 2, \ldots, n] = \sum_{i=1}^n w_i \psi_i,$$

where $\psi_i = \mathbb{E}[\boldsymbol{\xi} \mid (\varphi_j, \boldsymbol{\xi}) = \delta_{i,j} \,,\, j = 1, 2, \ldots, n]$ is exactly the minimizer of Eqn. (2.16). Therefore, ψ_i can be viewed as the optimal guess of $\boldsymbol{\xi}$ (in the least square sense) conditioning on the measurements $(\varphi_j, \boldsymbol{\xi}) = \delta_{i,j} \,,\, j = 1, 2, \ldots, n$. The very recent work [100] extends this probabilistic approach beyond the nuclear class by introducing the Gaussian cylinder measure.

Remark 2.2.2. *Using the abstract framework above, we are able to rewrite the l_1 operator compression (2.6) and the SPCA (2.7) as follows to make the comparison easier. We assume $\mathcal{L}$ is the positive definite elliptic operator, and one wants to construct basis functions that approximates the range space of the solution operator $\mathcal{K} = \mathcal{L}^{-1}$ well. Thanks to Lemma 2.2.1 and 2.2.2, we have the inclusion $\mathcal{K}(X) \subset H \subset X$. The l_1 operator compression (2.6) looks for a set of sparse basis functions $\{\psi_i\}_{i=1}^n$ in the smaller space H:*

$$\begin{aligned} \min_{\{\psi_i\}_{i=1}^n \subset H} \quad & \sum_{i=1}^n (\psi_i, \mathcal{L}\psi_i) + \lambda \sum_{i=1}^n \|\psi_i\|_1, \\ s.t. \quad & (\psi_i, \psi_j) = \delta_{i,j} \quad \forall 1 \le i, j \le n. \end{aligned} \tag{2.20}$$

On the contrary, the SPCA (2.7) *looks for sparse basis functions* $\{\varphi_i\}_{i=1}^n$ *in the bigger space* X*:*

$$\begin{aligned} \min_{\{\varphi_i\}_{i=1}^n \subset X} \quad & -\sum_{i=1}^n (\varphi_i, \mathcal{K}\varphi_i) + \lambda \sum_{i=1}^n \|\varphi_i\|_1, \\ s.t. \quad & (\varphi_i, \varphi_j) = \delta_{i,j} \quad \forall 1 \le i, j \le n. \end{aligned} \tag{2.21}$$

On one hand, thanks to the variational terms, i.e. $(\psi_i, \mathcal{L}\psi_i)$ *or* $-(\varphi_i, \mathcal{K}\varphi_i)$*, both Problem* (2.20) *and Problem* (2.21) *enforce their solutions, i.e.* $\{\psi_i\}_{i=1}^n$ *or* $\{\varphi_i\}_{i=1}^n$*, to align with the eigenspace corresponding to the small eigenvalues of* $\mathcal{L}$*. On the other hand, Problem* (2.20) *is more efficient because it searches in the smaller space* H*. Suppose* $\{(\lambda_i, e_i)\}_{i=1}^\infty$ *are the eigen-pairs of the elliptic operator* $\mathcal{L}$ *with* λ_i *in an ascending order. Due to the fact that* $\lim_{i\to\infty} \lambda_i = \infty$*, the variational term* $(\psi_i, \mathcal{L}\psi_i)$ *will be very sensitive when* ψ_i *does not align with the eigenvectors with small eigenvalues. On the contrary, the eigenvalues of* $-\mathcal{K}$*, i.e.* $-\frac{1}{\lambda_i}$*, cluster around 0 when* i *is large, and thus the variational term* $(\varphi_i, \mathcal{K}\varphi_i)$ *sees no difference between* e_n *and* e_m *for large* m *and* n*. In the case when* $X = L^2(D)$ *and* $H \subset H^k(D)$*, it means that* $(\psi_i, \mathcal{L}\psi_i)$ *penalizes more on rough* ψ_i*, but* $(\varphi_i, \mathcal{K}\varphi_i)$ *is not very sensitive to the roughness of* φ_i*. Recall that the eigenspace* $V_n := span\{e_1, \ldots, e_n\}$ *contains "smooth" functions in* $H^k(D)$*, and thus Problem* (2.20) *has a better accuracy to locate this subspace, especially when* n *is large.*

Localized Basis Functions

In the case to compress an elliptic operator $\mathcal{L}$, the positive semi-definite operator $\mathcal{K}$ plays the role of the solution operator $\mathcal{L}^{-1}$ and $X = L^2(D)$. The Cameron–Martin space H plays the role of the solution space of $\mathcal{L}$, which is a subset of $H^k(D)$, equipped with the energy norm $\|\cdot\|_H$. By an appropriate choice of the basis $\Phi \equiv [\varphi_1, \ldots, \varphi_n]$ of Φ, the energy minimizing basis functions in Eqn. (2.16) enjoy good localization properties.

Let $\{\tau_i\}_{1\le i\le m}$ be a partition of D such that each τ_i is Lipschitz convex and of diameter at most h. We also assume that the partition is regular [31]. It means that if h_i denotes the diameter of τ_i, there exists $\delta \in (0, 1)$ such that τ_i contains a ball centered at x_i with diameter

$$\rho_i \ge \delta h_i \qquad \forall i = 1, 2, \ldots, m. \tag{2.22}$$

For second order elliptic operators, Φ can be chosen as the space of piecewise constant functions, with basis φ_i to be the indicator function of the patch

τ_i. In Theorem 2.3.1, we will show that the basis function ψ_i, defined in Eqn. (2.16), decays exponentially fast away from its associated patch τ_i. For elliptic operators of order $2k$ ($k \geq 1$), Φ can be chosen as the space of (discontinuous) piecewise polynomials, with degree no more than $k-1$. We choose $\{\varphi_{i,q}\}_{i=1,q=1}^{m,Q}$ to be the basis of Φ, 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. In Theorem 4.3 and Theorem 4.4 in Part II, we will show that the basis function $\psi_{i,q}$, defined in Eqn. (2.16), decays exponentially fast away from its associated patch τ_i for every $1 \leq i \leq Q$.

The exponentially decaying property justifies the following local construction of the basis functions:

$$\begin{aligned} \psi_i^{loc} = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_j) = \delta_{i,j}, \quad j = 1, 2, \ldots, n, \\ & \psi(x) \equiv 0, \quad x \in D \backslash S_i, \end{aligned} \tag{2.23}$$

where $S_i \subset D$ is a neighborhood of the patch that ψ_i is associated with. Compared with Eqn. (2.16), the localized basis ψ_i^{loc} is obtained by solving exactly the same quadratic problem but on a localized domain S_i. Because the basis function ψ_i decays exponentially fast away from its associated patch, the localized basis function ψ_i^{loc} approximates ψ_i accurately, and the compression rate $E(\Psi^{loc}; \mathcal{K})$ is at the same order of $E(\Psi; \mathcal{K})$. Please refer to Theorem 3.6.1, Theorem 3.6.2 and Corollary 3.6.3 for details.

2.3 Sparse Operator Compression of Second Order Elliptic Equations

An important class of the operator $\mathcal{K}$ is the solution operator of elliptic operators, denoted as $\mathcal{L}$. In this section, we consider the following second order elliptic equation:

$$\begin{aligned} & \mathcal{L}u := -\nabla \cdot (a(x)\nabla u(x)) + c(x)u(x) = f(x) \qquad x \in D, \\ & u \in H_0^1(D), \end{aligned} \tag{2.24}$$

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

$$a_{min} I_d \preceq a(x) \preceq a_{max} I_d, \quad 0 \leq c(x) \leq c_{max} \qquad x \in D. \tag{2.25}$$

By the Lax-Milgram lemma, Eqn. (2.24) has a unique weak solution $u \in H_0^1(D)$, denoted as $\mathcal{L}^{-1} f$ or $\mathcal{K} f$. It is well-known that the operator $\mathcal{K} : L^2(D) \to L^2(D)$ is symmetric, positive definite and compact, and its eigenvalues decay like $\lambda_m(\mathcal{K}) \sim m^{-2/d}$; see e.g. [93]. For any $m \in \mathbb{N}$, we want to construct a m-rank operator, denoted as $\mathcal{K}_m$, such that (1) $\|\mathcal{K} - \mathcal{K}_m\| \sim \lambda_m(\mathcal{K})$, (2) there exists m basis functions that span the range space of $\mathcal{K}_m$ and have exponentially decaying tails.

In this case, $X = L^2(D)$. Following the definition of the Cameron–Martin space, we have $H = H_0^1(D)$ with inner product $(u, v)_H = \int_D \nabla u \cdot a \nabla v + cuv$.

Construction Of Basis Functions And The Approximation Rate

Let $\{\tau_i\}_{1 \leq i \leq m}$ be a regular partition of D such that each τ_i is Lipschitz convex and of diameter at most h. Following the strategy in Section 2.2, we take

$$\Phi = \text{span}\{\varphi_i, 1 \leq i \leq m\}, \qquad \Psi = \mathcal{K}(\Phi), \tag{2.26}$$

where φ_i is the characteristic function of the patch τ_i, i.e., φ_i is equal to one on τ_i and zero elsewhere. From the Poincare inequality, we can easily get that:

$$\|u - \mathcal{P}_\Phi^{(X)} u\|_{L^2(D)} \leq \frac{h}{\pi \sqrt{a_{min}}} \|u\|_H \qquad \forall u \in H^1(D). \tag{2.27}$$

Proof. By the construction of Φ, $u - \mathcal{P}_\Phi^{(X)} u = u - \int_{\tau_i} u/|\tau_i|$ on patch τ_i. By the Poincare inequality,

$$\|u - \int_{\tau_i} u/|\tau_i|\|_{L^2(\tau_i)} \leq h/\pi \|\nabla u\|_{L^2(\tau_i)}, \tag{2.28}$$

where the Poincare constant is taken to be $1/\pi$ since τ_i is Lipschitz convex. Therefore, we have

$$\|u - \mathcal{P}_\Phi^{(X)} u\|_{L^2(D)} \leq h/\pi \|\nabla u\|_{L^2(D)} \leq \frac{h}{\pi \sqrt{a_{min}}} \left(\int_D \nabla u \cdot a \nabla u \right)^{1/2} \leq \frac{h}{\pi \sqrt{a_{min}}} \|u\|_H.$$

We used $a \succeq a_{min} I_d$ in the last second inequality. □

According to Theorem 2.2.1, we have

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

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_H \leq \frac{h}{\pi\sqrt{a_{min}}} \|f\|_{L^2(D)} \,. \tag{2.29}$$

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

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_{L^2(D)} \leq \frac{h^2}{\pi^2 a_{min}} \|f\|_{L^2(D)} \,. \tag{2.30}$$

3. We have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \leq \frac{h^2}{\pi^2 a_{min}} \,. \tag{2.31}$$

Since $m \approx 1/h^d$ where m is the number of local patches τ_i, we have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \underset{\sim}{<} \lambda_m(\mathcal{K}) \,. \tag{2.32}$$

Therefore, the m-dimensional subspace Ψ compresses the solution operator $\mathcal{K}$ at the optimal rate.

Exponential Decay Of Basis Functions

In this subsection, 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 [99]. When $c \neq 0$, it has been recently studied in [101] independently of our work. The results presented in this second order case are not new. 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.

To simply our notations, for any $\psi \in H$ and any subdomain $S \subset D$, $\|\psi\|_{H(S)}$ denotes $\left(\int_S \nabla\psi \cdot a\nabla\psi + c\psi^2\right)^{1/2}$. In this chapter (second order elliptic operator with low contrast coefficients), the projection-type approximation property is simply the Poincare inequality, as we have used in the last subsection to obtain the error estimate. The following lemma provides us the inverse energy estimate, which is a special case of Lemma 3.5.1.

Lemma 2.3.1. *For any domain partition with $h \leq h_0 \equiv \pi\sqrt{\frac{a_{max}}{2c_{max}}}$, we have*

$$\|\mathcal{L}v\|_{L^2(\tau_j)} \leq \sqrt{a_{max}} C(d,\delta) h^{-1} \|v\|_{H(\tau_j)}, \quad \forall v \in \Psi \,, \forall j = 1, 2, \ldots, m, \tag{2.33}$$

where $C(d,\delta) = \sqrt{8d(d+2)}\delta^{-1-d/2}$. *If* $c_{max} = 0$, *i.e.* $c(x) \equiv 0$, *Eqn.* (2.33) *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 2.3.1. *For* $h \le h_0 \equiv \pi\sqrt{\frac{a_{max}}{2c_{max}}}$, *it holds true that*

$$\|\psi_i\|^2_{H(D\cap(B(x_i,r))^c)} \le \exp(1 - \frac{r}{lh})\|\psi_i\|^2_{H(D)} \tag{2.34}$$

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.* $c(x) \equiv 0$, *Eqn.* (2.34) *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 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 Figure 2.2.

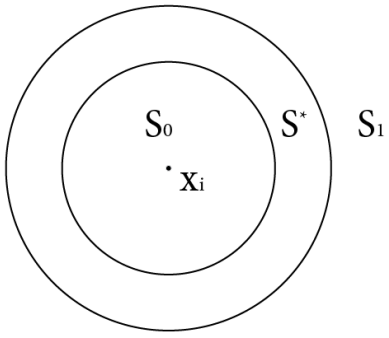


Figure 2.2: Illustration of S_0, S_1, and S^*.

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 $k \ge 1$, there exists constant C such that $b_{k+1} \le C(b_k - b_{k+1})$. Then we have $b_{k+1} \le \frac{C}{C+1}b_k$ for any $k \ge 1$ and thus we get the exponential decay $b_k \le (\frac{C}{C+1})^{k-1}b_1 \le (\frac{C}{C+1})^{k-1}b_0$. We will choose l such that $C \le \frac{1}{\mathrm{e}-1}$ and thus get $b_k \le e^{1-k}b_0$, which gives the result (2.34). 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, let's 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 $\eta(x) = \text{dist}(x, S_0)/\left(\text{dist}(x, S_0) + \text{dist}(x, S_1)\right)$. 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

$$\int_D \eta\nabla\psi_i \cdot a\nabla\psi_i + \int_D \eta c|\psi_i|^2 = \underbrace{\int_D \eta\psi_i(-\nabla\cdot(a\nabla\psi_i) + c\psi_i)}_{I_2} \underbrace{- \int_D \psi_i\nabla\eta \cdot a\nabla\psi_i}_{I_1}. \tag{2.35}$$

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-Schwartz inequality, we obtain

$$\begin{aligned} I_1 &\leq \|\nabla\eta\|_{L^\infty(D)}\|\psi_i\|_{L^2(S^*)}\left(\int_{S^*}\nabla\psi_i \cdot a\nabla\psi_i\right)^{1/2}\sqrt{a_{max}} \\ &\leq \frac{1}{lh}\|\psi_i\|_{L^2(S^*)}\|\psi_i\|_{H(S^*)}\sqrt{a_{max}}. \end{aligned} \tag{2.36}$$

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 (2.16), we have $\int_D \psi_i\varphi_j = 0$ for $i \neq j$. Thanks to (2.18), 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

$$\begin{aligned} I_2 &= -\int_{S^*}\eta\psi_i(-\nabla\cdot(a\nabla\psi_i) + c\psi_i) = -\sum_{\tau_j\in S^*}\int_{\tau_j}(\eta - \eta_j)\psi_i(-\nabla\cdot(a\nabla\psi_i) + c\psi_i) \\ &\leq \frac{1}{l}\sum_{\tau_j\in S^*}\|\psi_i\|_{L^2(\tau_j)}\|\mathcal{L}\psi_i\|_{L^2(\tau_j)}. \end{aligned} \tag{2.37}$$

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 piece-wise constant function space. Specifically, Lemma 2.3.1 says $\|\mathcal{L}\psi_i\|_{L^2(\tau_j)} \leq \sqrt{a_{max}}C(d,\delta)h^{-1}\|\psi_i\|_{H(\tau_j)}$ when $h \leq h_0 \equiv \pi\sqrt{\frac{a_{max}}{2c_{max}}}$. Then, we obtain

$$I_2 \leq \frac{\sqrt{a_{max}}C(d,\delta)}{lh}\|\psi_i\|_{L^2(S^*)}\|\psi_i\|_{H(S^*)} \quad \forall h \leq h_0. \tag{2.38}$$

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

By the construction of ψ_i (2.16), we have $\int_{\tau_j} \psi_i = 0$ for all $\tau_j \in S^*$. By the Poincare inequality, we have $\|\psi_i\|_{L^2(\tau_j)} \leq \|\nabla\psi_i\|_{L^2(\tau_j)} h/\pi$, and then we obtain

$$\|\psi_i\|^2_{H(S_1)} \leq I_1 + I_2 \leq \frac{1 + C(d,\delta)}{\pi l}\sqrt{\frac{a_{max}}{a_{min}}}\|\psi_i\|^2_{H(S^*)}. \tag{2.39}$$

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

Remark 2.3.1. *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; see our application in Section 2.4. The periodic boundary condition is useful in compressing a Hamiltonian with periodic boundary condition in quantum physics; see our numerical example in Section 2.5.*

The above proof can be applied to the operator $\mathcal{L}$ in (2.24) *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 conditions, the Cameron–Martin space is not $H^1_0(D)$ but the solution space associated with the corresponding boundary condition. The proof of Theorem 2.3.1 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* (2.35)*. For the homogeneous Neumann boundary condition and the periodic boundary condition, the proof is exactly the same because the integration by parts* (2.35) *can be carried out in exactly the same way. For the problems with the Robin boundary condition, i.e.*

$$\begin{aligned} \mathcal{L}u := -\nabla\cdot(a(x)\nabla u(x)) + c(x)u(x) &= f(x) \qquad x \in D, \\ \frac{\partial u}{\partial n} + \alpha(x)u(x) &= 0 \qquad x \in \partial D, \end{aligned} \tag{2.40}$$

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

$$\|u\|^2_H = \int_D \nabla u \cdot a\nabla u + \int_D cu^2 + \int_{\partial D} \alpha u^2. \tag{2.41}$$

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

$$\|u\|^2_{H(S)} = \int_S \nabla u \cdot a \nabla u + \int_S cu^2 + \int_{\partial D \cap \partial S} \alpha u^2. \tag{2.42}$$

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

Localization Of The Basis Functions

Theorem 2.3.1 allows us to localize the construction of basis functions ψ_i as follows. For $r > 0$, let S_r be the union of the subdomains τ_j intersecting $B(x_i, r)$ (recall that $B(x_i, \delta h_i/2) \subset \tau_i$) and let ψ_i^{loc} be the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_i^{loc} = \underset{\psi \in H_0^1(S_r)}{\arg\min} \quad & \|\psi\|^2_H \\ \text{s.t.} \quad & \int \varphi_j \psi = \delta_{i,j}, \quad \forall 1 \le j \le m. \end{aligned} \tag{2.43}$$

We will naturally identify ψ_i^{loc} with its extension to $H_0^1(D)$ by setting $\psi_i^{loc} = 0$ outside of S_r.

If the elliptic operator $\mathcal{L}$ is given with some other homogeneous boundary condition, the localized problem (2.43) should be modified slightly as follows such that the basis function ψ_i honors the given boundary condition on ∂D:

$$\begin{aligned} \psi_i^{loc} = \underset{\psi \in H}{\arg\min} \quad & \|\psi\|^2_H \\ \text{s.t.} \quad & \int \varphi_j \psi = \delta_{i,j}, \quad \forall 1 \le j \le m, \\ & \psi(x) \equiv 0 \quad x \in D \backslash S_r. \end{aligned} \tag{2.44}$$

When $\partial S_r \cap \partial D = \emptyset$, Eqn. (2.44) is equivalent to Eqn. (2.43). However, when $\partial S_r \cap \partial D \neq \emptyset$, Eqn. (2.44) only enforces the zero Dirichlet boundary condition on $\partial S_r \backslash \partial D$, but honors the original boundary condition on ∂D.

Thanks to the exponential decay of the energy minimizing basis functions $\{\psi_i\}_{i=1}^m$, S_r with radius $r = \mathcal{O}(h \log(1/h))$ is sufficient to guarantee that the localized basis functions $\{\psi_i^{loc}\}_{i=1}^m$ have the same compression accuracy as the exponentially decaying basis functions. The following three theorems demonstrate such properties of the localized basis functions $\{\psi_i^{loc}\}_{i=1}^m$. These theorems are the special case ($k = 1$) of Theorem 3.6.1, Theorem 3.6.2, and

Corollary 3.6.3 in Chapter 3. We refer to Section 3.6 in Chapter 3 for their proofs.

Theorem 2.3.2. *Under the same assumptions as those in Theorem 2.3.1, for any* $1 \leq i \leq m$ *and* $h \leq h_0 \equiv \pi\sqrt{\frac{a_{max}}{2c_{max}}}$, *it holds true that*

$$\|\psi_i - \psi_i^{loc}\|_{H(D)} \leq C_3 h^{-d/2-1} \exp(-\frac{r-2h}{2lh}), \tag{2.45}$$

where

$$C_3 = C(d,\delta)\left(\frac{e2^{d+1}a_{max}}{V_d\delta^d}\right)^{1/2}\left(\left(\frac{2}{\pi}\sqrt{\frac{a_{max}}{a_{min}}}+1\right)^2 + \frac{2}{\pi}\sqrt{\frac{a_{max}}{a_{min}}}C(d,\delta)\right)^{1/2}.$$

Here, the constants $C(d,\delta)$ *and* l *are from Theorem 2.3.1, and* V_d *is the volume of the unit* d*-dimensional ball.*

When $c(x) \equiv 0$, *i.e.* $\mathcal{L}u = -\nabla\cdot(a(x)\nabla u)$, *Eqn.* (2.45) *holds true for all* $h > 0$. *In this case, the constant* C_3 *can be taken as*

$$C_3 = C(d,\delta)\left(\frac{e2^{d}a_{max}}{V_d\delta^d}\right)^{1/2}\left(\left(\frac{1}{\pi}\sqrt{\frac{a_{max}}{a_{min}}}+1\right)^2 + \frac{1}{\pi}\sqrt{\frac{a_{max}}{a_{min}}}C(d,\delta)\right)^{1/2}.$$

Theorem 2.3.3. *Let* $u \in H_0^1(D)$ *be the weak solution of* $\mathcal{L}u = f$ *and* ψ_i^{loc} *be the localized basis functions defined in Eqn.* (2.43). *Then for* $r \geq (d+4)lh\log(1/h) + 2(1+l\log C_4)h$, *we have*

$$\inf_{v\in\Psi^{loc}} \|u-v\|_{H(D)} \leq \frac{2h}{\pi\sqrt{a_{min}}}\|f\|_{L^2(D)}. \tag{2.46}$$

The constants $C_4 = \pi a_{min}^{1/2} C_3 C_e$. *Here,* C_3 *is defined in Theorem 2.3.2, and* C_e *is the constant such that* $\|u\|_{L^2(D)} \leq C_e\|f\|_{L^2(D)}$ *holds true.*

Theorem 2.3.3 shows that we can obtain a linear convergence rate in the energy norm when our localized basis functions $\{\psi_i^{loc}\}_{i=1}^m$ are used as basis functions in the multiscale finite element method. By applying the Aubin-Nistche duality argument, we can get the following corollary.

Corollary 2.3.4. *Let* $\psi_{i,q}^{loc}$ *be the localized basis functions defined in Eqn.* (2.43). *Then for* $r \geq (d+4)lh\log(1/h) + 2(1+l\log C_4)h$, *we have*

$$\|\mathcal{K} - \mathcal{P}_{\Psi^{loc}}^{(H)}\mathcal{K}\| \leq \frac{4h^2}{\pi^2 a_{min}}, \tag{2.47}$$

where all the constants are the same as those defined in Theorem 2.3.3.

Corollary 2.3.4 shows that we can compress the symmetric positive semidefinite operator $\mathcal{K}$ with the optimal rate h^2 and with the nearly optimal localized basis (with support size of order $h\log(1/h)$).

Connections To The LOD Method

In the localizable orthogonal decomposition (LOD) [89], the authors introduce a modified Clément interpolation $\mathcal{I}_h$ on a uniform mesh with mesh size h, write V^f the kernel of $\mathcal{I}_h$ (i.e., the set of functions u such that $\mathcal{I}_h u = 0$), and identify the finite element space Ψ as the orthogonal complement of V^f with respect to the inner product defined by $a(u,v) = \int_D u\mathcal{L}v$ for $u, v \in H_0^1(D)$. The finite element basis ψ_i is identified by $\lambda_i - \mathcal{P}^a_{V^f}\lambda_i$, where λ_i is the nodal piecewise linear element and $\mathcal{P}^a_{V^f}\lambda_i$ is its projection onto the space V^f with respect to (w.r.t.) the inner product $a(u,v)$. The work [89] shows that this finite element basis Ψ achieves the optimal compression rate. Moreover, they showed that the finite element basis function ψ_i decays exponentially fast away from its associated node, and thus can be localized to local patches of size $\mathcal{O}(h\log(1/h))$ without loss of accuracy. The authors of [89] have also considered other types of Clément-type quasi-interpolation in the LOD method, and have used them to solve different kinds of second-order elliptic equations; see e.g. [109, 110, 60].

The general LOD method can be interpreted in the framework of the Sparse OC. Let $\mathcal{T}_h$ denote a regular triangulation of D into closed simplices, $\mathcal{N}_h = \{z_i\}_{i=1}^m$ denote the set of all interior mesh nodes in $\mathcal{T}_h$ and $V_h \subset H_0^1(D)$ the corresponding piecewise linear finite element space. Given $\{\varphi_i\}_{i=1}^m \subset L^2(D)$, a Clément-type quasi-interpolation operator $\mathcal{I}_h : H_0^1(D) \to V_h$ is defined by

$$\mathcal{I}_h v := \sum_{z_i \in \mathcal{N}_h} \Big(\int_D \varphi_i v\Big)\lambda_i, \tag{2.48}$$

where $\lambda_i \in L^2(D)$ is the piecewise linear element centered at z_i. In [89], φ_i is taken as the normalized nodal piecewise linear element λ_i, i.e., $\varphi_i = \frac{\lambda_i}{\int_D \lambda_i}$. With the Clément-type quasi-interpolation operator given in Eqn. (2.48), the global exponentially decaying basis functions $\{\psi_i\}_{i=1}^m$ are the unique solution of the following energy-minimizing problem:

$$\begin{aligned} \psi_i = \underset{\psi \in H_0^1(D)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int \varphi_j \psi = \int \varphi_j \lambda_i, \quad \forall 1 \le j \le m. \end{aligned} \tag{2.49}$$

Like what we have done in the Sparse OC, the construction of ψ_i can be localized onto a neighborhood of z_i, denoted by S_r. The localized basis function ψ_i^{loc} in the LOD method is the unique solution of the following local energy-minimizing problem:

$$\begin{aligned} \psi_i^{loc} = \underset{\psi \in H_0^1(S_r)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int \varphi_j \psi = \int \varphi_j \lambda_i, \quad \forall 1 \leq j \leq m. \end{aligned} \tag{2.50}$$

Comparing Eqn. (2.43) and (2.50), we can clearly see the similarity between the LOD method and the Sparse OC. Moreover, our error analysis (Theorem 2.2.1) can be directly used to prove the optimal linear convergence rate of the global basis $\{\psi_i\}_{i=1}^m$ given by Eqn. (2.49). Although our proof (following the proof in [99]) is different from the proof in [89], we share some essential elements. First of all, we both use a recursive argument to prove the exponential decay of basis functions. As far as we know, this kind of recursive argument first appeared in [89]. Secondly, the Poincare inequality plays an essential role in both proofs, i.e., the local projection-type approximation property in our proof and the assumption (2.5.a) in [89]. Thirdly, the stability condition of the Clément-type quasi-interpolation operator (see assumption (2.5.b) in [89]) plays a similar role as our inverse energy estimate.

Finally, although [89] contains the first rigorous proof of exponential decay for such energy-minimizing basis functions, the idea of its proof uses in a crucial way the projection properties of the Clement interpolation operator (associated with the underlying implicit measurement functions used in [89]), which hinders its generalization. The proof of exponential decay provided in [99] uses a combination of energy and inverse energy estimates instead and enables the generalization beyond measurements derived from the Clement interpolation operator as acknowledged in [109] (Page 8).

> "In a setting with a modified trial space, further generalisations are possible. Since V_H does not appear any more in the method, its conformity can be relaxed as it was recently proposed in [99] in the context of a multilevel solver for Poisson-type problems with L^∞ coefficients. This approach enables one to compute very general quantities of the solution such as piecewise mean values."

Therefore, we choose to follow [99] to generalize this idea to high-order elliptic operators and to second-order elliptic operator with high-contrast coefficients.

2.4 Application in Sparse Principal Component Analysis

In spatial statistics, geostatistics, machine learning and image analysis, the Matérn covariance [92] is used to model random fields with smooth samples; see e.g. [121, 58, 53]. The Matérn covariance between two points $x, y \in D \subset \mathbb{R}^d$ is given by

$$K_\nu(x,y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x-y|}{\rho} \right), \tag{2.51}$$

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

$$\widehat{k}(\omega) = c_{\nu,\lambda}\sigma^2 \left(\frac{2\nu}{\lambda^2} + |\omega|^2 \right)^{-(\nu+d/2)}, \quad c_{\nu,\lambda} := \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)(2\nu)^\nu}{\Gamma(\nu)\lambda^{2\nu}}, \tag{2.52}$$

where we use the convention $\widehat{f}(\omega) = \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\omega} \mathrm{d}x$ for the Fourier transform. For both sampling from the random fields and performing basic computations like marginalization and conditioning, we need to compress the Matérn covariance operator $\mathcal{K} : L^2(D) \to L^2(D)$, the Hilbert-Schmidt operator with kernel $K_\nu(x,y)$, with rank-n covariance operators:

$$E_{oc}(\Psi; \mathcal{K}) := \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{K} - \Psi K_n \Psi^T\|_2, \tag{2.53}$$

where $\Psi = [\psi_1, \ldots, \psi_n]$ span the range space of the approximate operator $\Psi K_n \Psi^T$. Recent study [83, 16] shows that the Matérn covariance and the elliptic operators are closely connected. With proper homogeneous boundary conditions, the Matérn covariance operator with $\nu + d/2$ as an integer is the solution operator of an elliptic operator of order $2\nu + d$. For example, 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 $(8\pi\rho^3\sigma^2)^{-1}(1 - 2\rho^2\Delta + \rho^4\Delta^2)$ when $d = 3$. The Matérn covariance operator with $\nu = 1$ is the solution operator of the fourth order elliptic operator $(4\pi\rho^2\sigma^2)^{-1}(1 - 2\rho^2\Delta + \rho^4\Delta^2)$ when $d = 2$. Note that the elliptic operator that is associated with the Matérn covariance contains lower order terms. Thus, it

is essential that our analysis can accommodate lower order terms and various boundary conditions.

Based on Eqn. (2.17) and (2.18), 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

$$K(x,y) = \exp(-|x-y|) \qquad x,y \in [0,1], \tag{2.54}$$

which is exactly the Matérn covariance (2.51) with $\nu = 1/2$, $\sigma = 1$ and $\rho = 1$. This problem has been studied by different groups; see e.g. [52, 38, 64, 8]. 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 Eqn. (2.17) and (2.18). In Figure 2.3, 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 Figure 2.4. We can see that the exponentially decaying basis functions Ψ has nearly the same compression rate with the eigendecomposition.

One can easily verify that the exponential kernel (2.54) is the Green's function of the following second order elliptic equation

$$\begin{aligned} -\frac{1}{2}u''(x) + \frac{1}{2}u &= f(x), \qquad 0 < x < 1, \\ u(0) - u'(0) = 0, &\quad u(1) + u'(1) = 0, \end{aligned} \tag{2.55}$$

whose associated energy norm is

$$\|u\|^2_{H(D)} = \frac{1}{2}\left(u(0)^2 + u(1)^2 + \int_0^1 (u')^2 + \int_0^1 u^2\right). \tag{2.56}$$

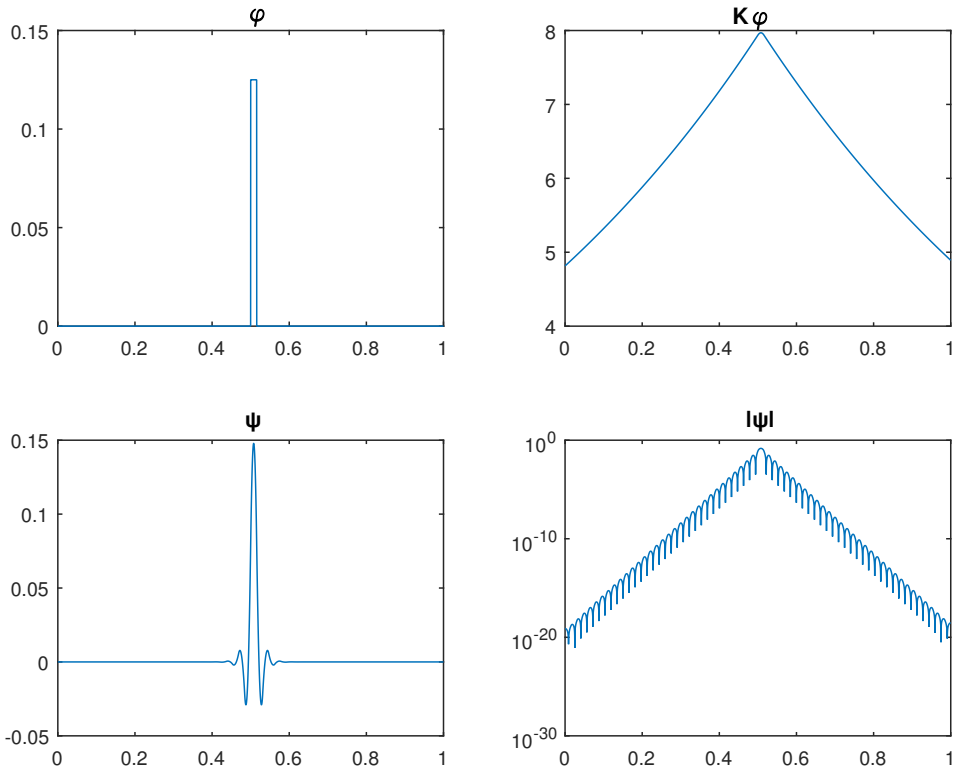


Figure 2.3: The basis function associated with patch $[1/2 - h, 1/2]$.

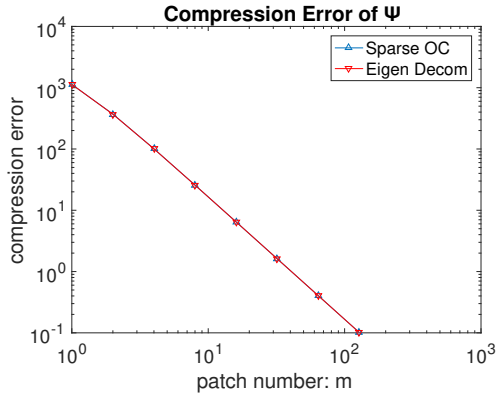


Figure 2.4: The operator compression error $E(\Psi; \mathcal{K})$ (2.53) for the exponential kernel (2.54) with exponentially decaying basis functions Ψ. They have nearly the same compression error as the eigenfunctions of $\mathcal{K}$.

Solving the localized variational problem (2.44), we can get localized basis functions Ψ^{loc}. With different sizes of the support S_r, we compute the compression error $E(\Psi^{loc}; \mathcal{K})$ for $m = 2^i$ $(0 \le i \le 7)$. The results are summarized in Figure 2.5. In the left subfigure of Figure 2.5, we take the support with size Ch, for $C = 3, 5, 7, 9$, and 11. In the right subfigure of Figure 2.5, 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 following $C\log_2(1/h)$. We can see that the constant oversampling strategy does not work well, while the $h\log_2(1/h)$ oversampling strategy has the optimal second order convergence rate as our Corollary 2.3.4 predicted. 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. We show several basis functions ψ_i^{loc} in Figure 2.6. One can see that the basis functions on the boundary honor the Robin boundary conditions.

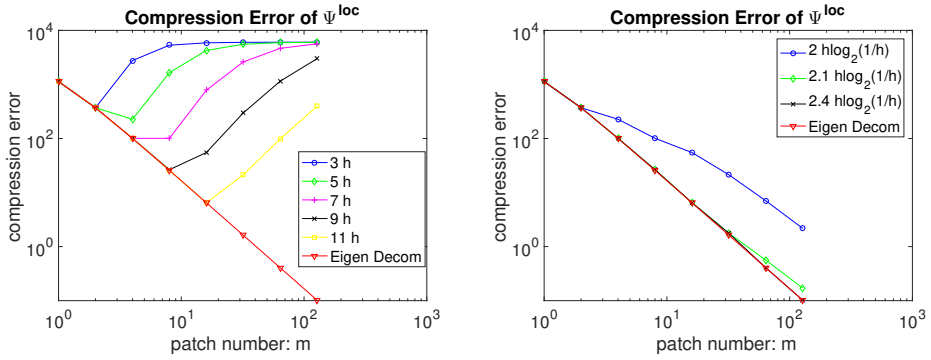


Figure 2.5: The operator compression error $E(\Psi^{loc}; \mathcal{K})$ (2.53) with localized basis functions Ψ^{loc}. The constant oversampling strategy (left) does not work well, while the $h\log_2(1/h)$ oversampling strategy (right) has the optimal second order convergence rate.

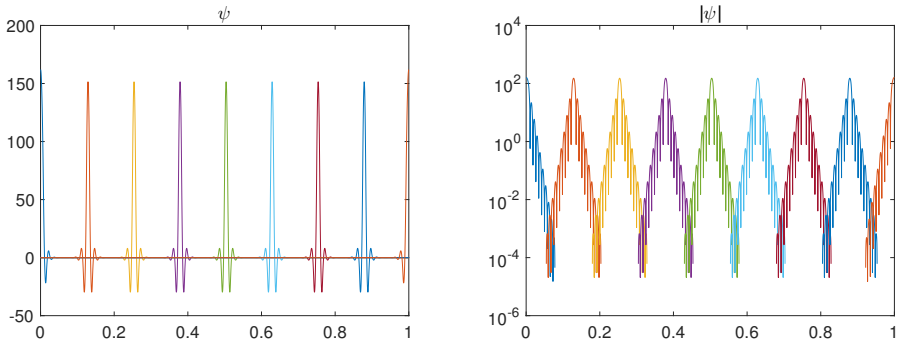


Figure 2.6: A few basis functions for the case $m = 2^7$ and $r = 2.4h\log_2(1/h)$.

2.5 Application in Constructing Localized Wannier Functions

In this section, we consider the Hamiltonian of a free particle in a bounded domain $D = [-\pi, \pi]$ with periodic boundary condition

$$\mathcal{H} = -\Delta. \tag{2.57}$$

Suppose its eigen decomposition is given by $\mathcal{H}e_i = \lambda_i e_i$, where λ_i are the eigenvalues in increasing order and e_i are the corresponding eigenfunctions. We denote $V_n = \text{span}\{v_i : 1 \le i \le n\}$, the n-dimensional low-lying eigenspace. We want to construct n localized basis functions $\{\psi_i\}_{i=1}^n$ that can span V_n accurately. This problem is studied extensively before; see e.g. [90, 42, 91, 107, 78]. In this section, we propose to use the operator compression error $E(\Psi; \mathcal{H}^{-1})$ (2.2) to quantify the compression error. Compared with other existing methods, our variational construction is guaranteed to obtain the optimal compression error with nearly optimally localized basis functions, and is much more efficient due to its convexity and decoupling in computing different basis functions. We will briefly review the compressed modes [107] from the l_1 approach and compare it with our Sparse OC.

Construction via The l_1 Approach

In [107], the authors proposed a novel method to create a set of localized functions $\{\psi_i\}_{i=1}^n$, which are compressed modes, such that $\sum_{i=1}^n \psi_i^T \mathcal{H} \psi_i$ approximates $E_n = \sum_{i=1}^n \lambda_i$. The locality is accomplished by introducing an l_1 regularization of the basis functions into the variational formulation of eigendecomposition:

$$\begin{aligned} E = \min_{\Psi_n} \quad & \sum_{i=1}^n \left(\frac{1}{\mu} \|\psi_i\|_1 + \langle \psi_i, \mathcal{H}\psi_i \rangle \right) \\ \text{s.t.} \quad & \langle \psi_i, \psi_j \rangle = \delta_{ij} \quad \forall 1 \le i, j \le n, \end{aligned} \tag{2.58}$$

where $\Psi_n = \{\psi_i\}_{i=1}^n$ and the l_1 norm is defined as $\|\psi_i\|_1 = \int_D |\psi_i(x)| \mathrm{d}x$. The parameter μ controls the trade-off between sparsity and accuracy: larger values of μ gives solutions that better minimize the total energy at the expense of more extended basis functions, while a smaller μ will give highly localized wave functions at the expense of larger errors in the calculated ground state energy E_n.

The authors of [107] solve the non-convex problem (2.58) using the algorithm of splitting orthogonality constraint (SOC) proposed in [79]. By discretizing

$\mathcal{H}$ into an N-by-N Hermitian matrix, still denoted as $\mathcal{H}$, and introducing auxiliary variables $Q = \Psi$ and $P = \Psi$, Eqn. (2.58) is equivalent to the following constrained problem:

$$\begin{aligned} \min_{\Psi,P,Q\in\mathbb{R}^{N\times n}} \quad & \frac{1}{\mu}\|Q\|_1 + \mathrm{Tr}(\Psi^T\mathcal{H}\Psi) \\ \text{s.t.} \quad & Q = \Psi, P = \Psi, P^TP = I_n, \end{aligned} \tag{2.59}$$

where $\|Q\|_1$ is the entry-wise l_1 norm of the matrix Q. Problem (2.59) can be solved by the following SOC algorithm based on split Bregman iteration, see Algorithm 1. The above subminimization problems can be easily solved as

Algorithm 1 The algorithm of splitting orthogonality constraint (SOC)

1: Initialize $\Psi_n^0 = P^0 = Q^0$, $b^0 = B^0 = 0$.
2: **while** "not converged" **do**
3: $\Psi_n^k = \arg\min_\Psi \mathrm{Tr}(\Psi^T\mathcal{H}\Psi) + \frac{\lambda}{2}\|\Psi - Q^{k-1} + b^{k-1}\|_F^2 + \frac{r}{2}\|\Psi - P^{k-1} + B^{k-1}\|_F^2$.
4: $Q^k = \arg\min_Q \frac{1}{\mu}\|Q\|_1 + \frac{\lambda}{2}\|\Psi_n^k - Q + b^{k-1}\|_F^2$.
5: $P^k = \arg\min_P \frac{r}{2}\|\Psi_n^k - P + B^{k-1}\|_F^2 \quad \text{s.t.} \quad P^TP = I_n$.
6: $b^k = b^{k-1} + \Psi_n^k - Q$.
7: $B^k = B^{k-1} + \Psi_n^k - P$.
8: **end while**

followings

$$(2\mathcal{H} + \lambda + r)\Psi_n^k = r(P^{k-1} - B^{k-1}) + \lambda(Q^{k-1} - b^{k-1}), \tag{2.60}$$

$$Q^k = \mathrm{Shrink}(\Psi_n^k + b^{k-1}, 1/(\lambda\mu)), \tag{2.61}$$

$$P^k = (\Psi_n^k + B^{k-1})U\Lambda^{-1/2}S^T, \tag{2.62}$$

where $U\Lambda S^T = \mathrm{svd}\left((\Psi_n^k + B^{k-1})^T(\Psi_n^k + B^{k-1})\right)$ and the "Shrink" operator is defined as $\mathrm{Shrink}(u,\delta) = \mathrm{sgn}(u)\max(0, |u| - \delta)$.

To resolve the small scales in the basis functions, we typically discretize $\mathcal{H}$ such that $N = Cn$, say $N = 16n$ or $N = 32n$, where C is number of nodes to resolve small scales in each localized basis function. In each iteration, the most time consuming part is Eqn. (2.62), which involves an SVD factorization and can be straightforwardly solved with an $\mathcal{O}(n^3)$ algorithm. If we are allowed to use the support as prior knowledge, the orthogonality constraint $P^TP = I$ can be replaced by a system of banded orthogonality constraints:

$$\int \psi_j\psi_k = \delta_{jk}, \quad , j = 1, \ldots, n,\ k = j, j \pm 1, \ldots, j \pm p,$$

where p is the band width. Taking advantage of this banded structure, the complexity in the SVD step can be reduced to $\mathcal{O}(8p^3n)$. See [107] for more details. Typically, Algorithm 1 takes hundreds of iterations to converge, depending on the choice of the parameters λ and r and the convergence criterion.

Construction via the Sparse OC

Applying the Sparse OC to $\mathcal{H} = -\frac{1}{2}\Delta + V(x)$, the localized basis functions $\{\psi_i\}_{i=1}^n$ are constructed as follows. First of all, we partition the physical domain D using a regular mesh $\{\tau_i\}_{i=1}^n$. We denote the mesh size h. Second, we choose $r > 0$, say $r = 2h\log(1/h)$. For each patch τ_i, let S_r be the union of the subdomains $\tau_{i'}$ intersecting $B(x_i, r)$ (for some $x_i \in \tau_i$), see Figure 2.1. Finally, let φ_i be the indicator function of the patch τ_i. The basis function ψ_i is obtained from the following convex optimization problem, which has a quadratic objective and several linear constraints:

$$\begin{aligned} \psi_i = \underset{\psi \in H^1_{\mathcal{B}}(D)}{\arg\min} \quad & \langle \psi, \mathcal{H}\psi \rangle \\ \text{s.t.} \quad & \int_{S_r} \psi\varphi_j = \delta_{i,j}, \quad \forall 1 \le j \le n, \\ & \psi(x) \equiv 0, \quad x \in D \backslash S_r. \end{aligned} \tag{2.63}$$

In Eqn. (2.63), $H^1_{\mathcal{B}}(D)$ is a subspace of $H^1(D)$ that contains the functions satisfying the prescribed boundary condition $\mathcal{B}$, such as the periodic boundary condition. The parameter r directly controls the size of the support of ψ_i. For $V(x) \ge 0$, we have proved that by choosing $r = Ch\log(h)$, we can achieve the optimal operator compression error and nearly optimally localize the basis functions simultaneously.

Eqn. (2.63) is a convex optimization problem with a quadratic objective and several linear constraints, and it can be solved very efficiently as follows. First of all, we discretize $\mathcal{H}$ and ψ on a fine mesh (a refined mesh over the partition $\{\tau_i\}_{i=1}^n$) with linear nodal basis functions, and we get the discretized $\mathcal{H} \in \mathbb{R}^{N\times N}$ and $\psi \in \mathbb{R}^N$. Suppose $\mathcal{I}_i$ is the set of fine mesh nodes that lies in S_r, and I_i^c is the set of the other nodes. Therefore, the constraint $\psi(x) \equiv 0$ for $x \in D\backslash S_r$ is $\psi|_{\mathcal{I}_i^c} \equiv 0$. The linear constraints $\int_{S_r} \psi\varphi_j = \delta_{i,j}$ for all $1 \le i \le n$ are written as $\Phi^T\psi = \boldsymbol{e}_i$, where $\boldsymbol{e}_i \in \mathbb{R}^n$ is the i-th column of the identity

matrix I_n. Therefore, we can obtain ψ_i by solving the following problem:

$$\begin{aligned} \psi_i = \underset{\psi \in \mathbb{R}^N}{\arg\min} \quad & \psi^T \mathcal{H} \psi \\ \text{s.t.} \quad & \Phi^T \psi = \boldsymbol{e}_i, \quad \psi|_{\mathcal{I}_i^c} \equiv 0. \end{aligned} \tag{2.64}$$

By the method of Lagrange multipliers, the nonzero part of ψ_i, i.e., $\psi_i|_{\mathcal{I}_i}$, can be efficiently solved by solving

$$\begin{bmatrix} \mathcal{H}_i & \Phi_i \\ \Phi_i^T & 0 \end{bmatrix} \begin{bmatrix} \psi \\ l \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{e}_i \end{bmatrix}, \tag{2.65}$$

where $\mathcal{H}_i = \mathcal{H}(\mathcal{I}_i, \mathcal{I}_i)$, $\Phi_i = \Phi(\mathcal{I}_i, :)$, and $l \in \mathbb{R}^n$ is the Lagrange multiplier. Since we take $r = Ch\log(1/h)$, say $r = 2h\log(1/h)$, the number of fine grid nodes in S_r, i.e. $|\mathcal{I}_i|$, is $\mathcal{O}(\frac{N}{n}\log n)$. The problem (2.64) or Eqn. (2.65) can be solved efficiently by the multigrid method, with complexity $\mathcal{O}\left(\frac{N}{n}\log n(\log\frac{N}{n} + \log\log n)^c\right)$ for some constant c. Since every ψ_i is solved independently, the complexity to obtain all the localized basis functions $\{\psi_i\}_{i=1}^n$ is simply $\mathcal{O}\left(N\log n(\log\frac{N}{n} + \log\log n)^c\right)$. Compared with the SOC algorithm to solve the l_1 penalized problem (2.58), the complexity of the Sparse OC is comparable to that of a single iteration in the SOC algorithm.

Numerical Results

In this section, we compare the l_1 approach with the Sparse OC by the free-electron model in [107]. The Hamiltonian of the free-electron model is $\mathcal{H} = -\frac{1}{2}\Delta$. We consider $D = [0, 50]$ as the physical space, as in [107]. We discretize D into with a fine mesh $h_f = 1/1024$, and the resulting discretized Hamiltonian $\mathcal{H} \in \mathbb{R}^{N\times N}$ where $N = 1024$. We are interested in approximating the first $n = 128$ low-lying eigenspace. In our comparison, with the same support size for ψ_i, we compare their performance in approximating the first n eigenvalues of $\mathcal{H}$, i.e. $\{\lambda_i\}_{i=1}^n$, and the density $\rho(x) = \sum_{i=1}^n e_i^2(x)$.

We would like to thank Professor Rongjie Lai for providing his code to solve the l_1 penalized problem (2.58). All the computations are performed in Matlab 2016a on a Macbook Pro 10.1 with 2.3 GHz Intel Core i7 processor. Since every ψ_i in the Sparse OC is solved independently, it is embarrassingly easy to implement it in a parallel fashion. However, in order to compare the total computational complexity of two approaches, we do not use parallel computing in the following comparison.

The l_1 approach

In this subsection, we recreate the result in [107] with $\mu = 0.84$. We will see that this μ gives roughly the same support size as that given by the Sparse OC. We pick

$$\lambda = r = 1/h_f^2 = 419.43$$

in Algorithm 1.

After 390 iterations, the l_1 approach achieves 1e-7 relative energy decrease, and the iteration is stopped. The total time is 4.426 secs. Every iteration takes 0.013 sec. Eight of the compressed modes are shown in Figure 2.7. It seems that the compressed modes also have exponential decay near its peak.

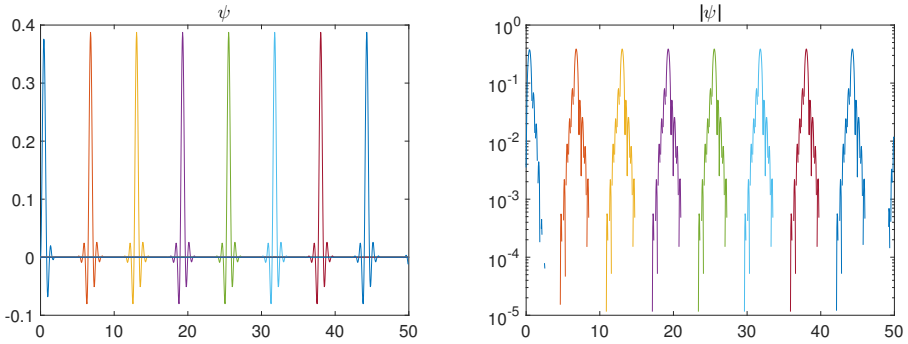


Figure 2.7: A few compressed modes from the l_1 approach, $m = 2^7, \mu = 0.84$

We can approximate eigenvalues of $\mathcal{H}$ by the eigenvalues of $\Psi^T \mathcal{H} \Psi$. The approximate eigenvalues are plotted in Figure 2.8.

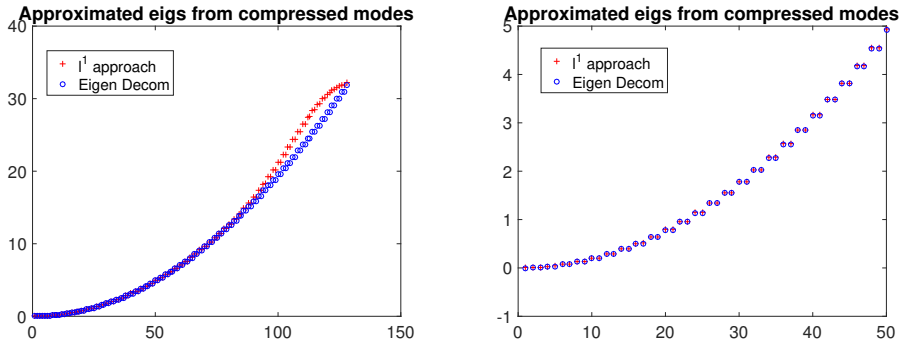


Figure 2.8: The eigenvalues of $\Psi^T H \Psi$, $m = 2^7, \mu = 0.84$

We also plot the approximate density $\rho(x) = \sum_{i=1}^{n} \psi_i^2(x)$ in Figure 2.9. We can see that the density approximation is also not very accurate.

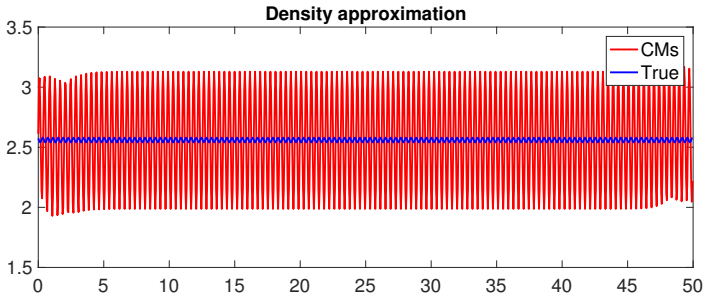


Figure 2.9: Density approximation by the l_1 approach.

The sparse operator compression

In this subsection, we show the results given by the Sparse OC. We take the support size of the localized basis as $r = h \log_2(h)$. It takes 0.035 sec to obtain all the 128 localized basis functions, without parallel computing. We can see that the total cost of Sparse OC is smaller than the cost per iteration in the l_1 approach. Eight of the localized modes are shown in Figure 2.10. We can see that the localized modes look very similar to the compressed modes in Figure 2.7.

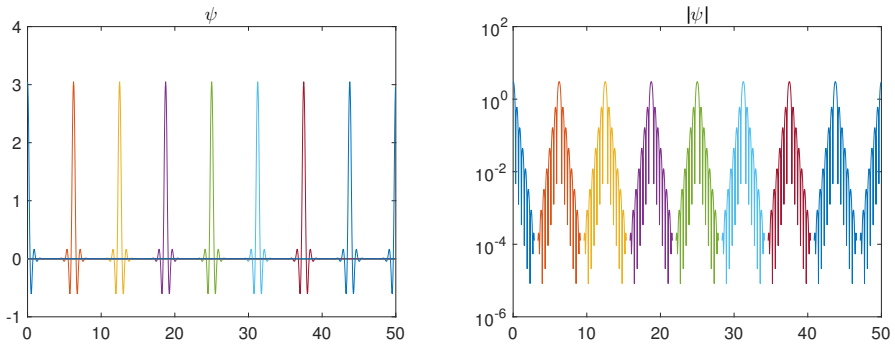


Figure 2.10: A few basis functions for the case $m = 2^7$ and $r = h \log_2(1/h)$.

We can approximate eigenvalues of $\mathcal{H}$ by the eigenvalues of $Q^T \mathcal{H} Q$, where Q is an orthonormal basis spanning Ψ. The approximate eigenvalues are plotted in Figure 2.11. The approximate eigenvalues are very similar to those in Figure 2.8.

We also plot the approximate density $\rho(x) = \sum_{i=1}^{n} \psi_i^2(x)$ in Figure 2.12. We can see that the density approximation is inaccurate, as for the l_1 approach in Figure 2.9.

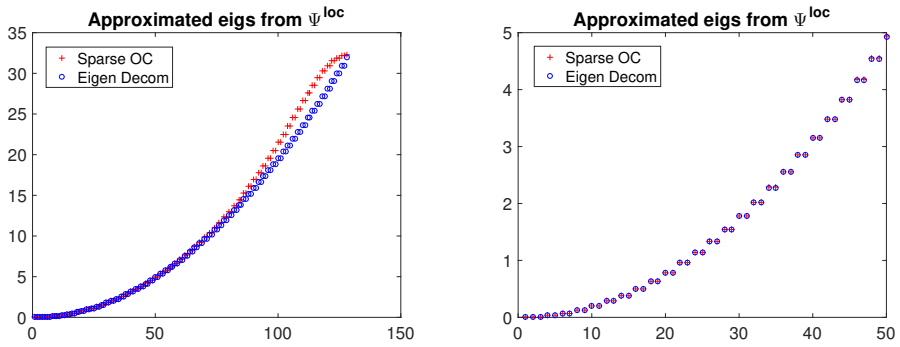


Figure 2.11: The eigenvalues of $Q^T H Q$ and H; Q is an orthonormal basis of Ψ.

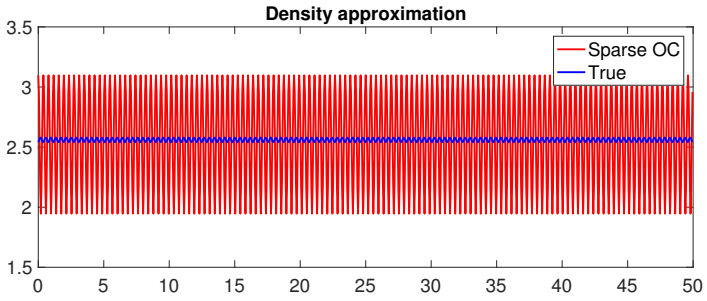


Figure 2.12: Density approximation by the Sparse OC.

The operator compression error of the Sparse OC is plotted in Figure 2.13, with different support sizes. As Corollary 2.3.4 predicts, the operator compression error is nearly optimal, i.e., decays like h^2, when the support size is taken as $r = Ch\log(1/h)$ for some constant $C > 0$.

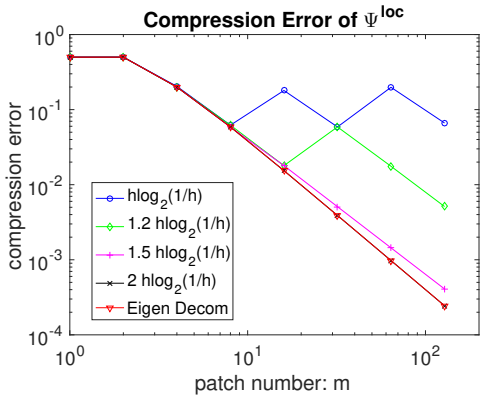


Figure 2.13: The operator compression error $E(\Psi; (\mathcal{L}+1)^{-1})$ for the Hamiltonian with localized basis functions Ψ^{loc}.

Conclusions of The Comparison

We summarize our comparison with the following four points.

1. With roughly the same width of the localized modes, the results given by the l_1 approach and the Sparse OC are very similar, in terms of the shape of the localized modes, the approximate eigenvalues, and the approximate density function.

2. The computation cost of each iteration for the l_1 approach is comparable to the Sparse OC. The l_1 approach needs many iterations to solve the nonconvex problem, although the number of iterations are only in the hundreds for a good choice of parameters λ and r.

3. Further speed up of the SOC algorithm is possible if one is allowed to use the support as prior knowledge. Parallel computing is also possible in solving both Ψ_n^k and P^k. In the Sparse OC, we can directly localize the support because we have proved that the localization will not affect the operator compression error. The Sparse OC can be easily executed in parallel, due to the decoupling of construction of every basis. Therefore, the Sparse OC is expected to be even faster if executed in parallel.

4. The SOC algorithm will converge for any choice of μ although we might need to be more careful about parameters. As the proposed l_1 regularized problem is nonconvex, the ADMM type methods only converge when the parameter r is greater than a certain number (according to Prof. Rongjie Lai).

Chapter 3

SPARSE OPERATOR COMPRESSION OF HIGHER ORDER ELLIPTIC OPERATORS

In Chapter 2, we introduced the sparse operator compression to compress a self-adjoint positive semi-definite operator $\mathcal{K} : L^2(D) \to L^2(D)$ by localized basis functions, where D is a bounded domain in $\mathbb{R}^d$. We applied our method to second order elliptic operators with rough and multiscale coefficients and various boundary conditions. We showed that on a regular mesh with mesh size h, our localized basis functions have supports of diameter $h\log(1/h)$, and give optimal compression rate of the solution operator. The main purpose of this chapter is to apply our sparse operator compression to higher order elliptic operators, and to show that our localized basis functions are able to give the optimal approximation property of the solution operator.

3.1 Problem Setting

Let $\mathcal{L}$ be a self-adjoint elliptic operator

$$\mathcal{L}u = \sum_{0\leq|\sigma|,|\gamma|\leq k} (-1)^{|\sigma|} D^{\sigma}(a_{\sigma\gamma}(x)D^{\gamma}u), \tag{3.1}$$

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 multi-index. Consider the elliptic equation with the homogeneous Dirichlet boundary conditions

$$\mathcal{L}u = f, \qquad u \in H_0^k(D), \tag{3.2}$$

where the load $f \in L^2(D)$. Here, we only consider the case when $\mathcal{L}$ (thus $\mathcal{K}$) is self-adjoint, i.e.

$$\int_D (\mathcal{L}u)v = \int_D u(\mathcal{L}v) \qquad \forall u, v \in H_0^k(D). \tag{3.3}$$

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

$$B(u, v) = \sum_{0\leq|\sigma|,|\gamma|\leq k} \int_D a_{\sigma\gamma}(x)D^{\sigma}uD^{\gamma}v. \tag{3.4}$$

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

$$a_{min}|u|_{k,2,D}^2 \leq B(u,u) \leq a_{max}|u|_{k,2,D}^2 \qquad \forall u \in H_0^k(D). \tag{3.5}$$

Thanks to the Riesz representation lemma, Eqn. (3.1) has a unique weak solution in $H_0^k(D)$ for $f \in L^2(D)$.

Construction Of The Basis Functions

Under the framework of Sparse OC introduced in Section 2.2, 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 Eqn. (3.1).

First of all, we divide D 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 [31]. 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

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

Then We choose the local measurement functions $\{\varphi_{i,q}\}_{q=1}^{Q}$ to 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 $k-1$. Thereafter, we have

$$\Phi = \text{span}\{\varphi_{i,q} : 1 \leq q \leq Q, 1 \leq i \leq m\}, \quad \Psi = \mathcal{K}\Phi. \tag{3.6}$$

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

$$\int_{\tau_i} \varphi_{i,q}\varphi_{i,q'} = |\tau_i|\delta_{q,q'}. \tag{3.7}$$

After that, a set of global energy-minimizing basis functions of Ψ is defined by Eqn. (2.16) accordingly, i.e.,

$$\begin{aligned} \psi_{i,q} = &\underset{\psi\in H_0^k(D)}{\arg\min} \quad \|\psi\|_H^2 \\ &\text{s.t.} \quad \int_D \psi_{i,q}\varphi_{j,q'} = \delta_{iq,jq'}, \forall 1 \leq q' \leq Q, 1 \leq j \leq m. \end{aligned} \tag{3.8}$$

We will prove that $\psi_{i,q}$ decays exponentially fast away from τ_i and thus we can localize its construction 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 $\psi_{i,q}^{loc}$ be the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_{i,q}^{loc} = \arg\min_{\psi \in H_0^k(S_r)} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int \varphi_{j,q'} \psi = \delta_{iq,jq'}, \quad \forall 1 \le j \le m,\, 1 \le q' \le Q. \end{aligned} \tag{3.9}$$

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

If the elliptic operator $\mathcal{L}$ is given with some other homogeneous boundary condition, the localized problem (3.9) should be slightly modified as follows such that the basis function $\psi_{i,q}$ honors the given boundary condition on ∂D:

$$\begin{aligned} \psi_{i,q}^{loc} = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int \varphi_{j,q'} \psi = \delta_{iq,jq'}, \quad \forall 1 \le j \le m,\, 1 \le q' \le Q, \\ & \psi(x) \equiv 0 \quad x \in D \backslash S_r. \end{aligned} \tag{3.10}$$

When $\partial S_r \cap \partial D = \emptyset$, Eqn. (3.10) is equivalent to Eqn. (3.9). However, when $\partial S_r \cap \partial D \neq \emptyset$, Eqn. (3.10) only enforces the zero Dirichlet boundary condition on $\partial S_r \backslash \partial D$, but honors the original boundary condition on ∂D.

Collecting all the $\psi_{i,q}^{loc}$ for $1 \le i \le m$ and $1 \le q \le Q$ together, we get our localized basis Ψ^{loc}.

Summary Of Our Main Results

To simplify the expression of constants, we will assume without loss of generality that the domain is rescaled so that $\mathrm{diam}(D) \le 1$. In this chapter, we prove that for $r = \mathcal{O}(h \log(1/h))$:

1. Ψ^{loc} achieve the optimal convergence rate to solve the elliptic equation, i.e.,

$$\|\mathcal{L}^{-1} f - \Psi^{loc} L_n^{-1} (\Psi^{loc})^T f\|_H \le C_e h^k \|f\|_2 \quad \forall f \in L^2(D), \tag{3.11}$$

   where the constant C_e is independent of n.

2. Ψ^{loc} achieve the optimal approximation error to approximate the elliptic operator, i.e.,
$$E_{oc}(\Psi^{loc}; \mathcal{L}^{-1}) \leq C_e^2 h^{2k}. \tag{3.12}$$

For $n = mQ$, we can show that the nth largest eigenvalue of $\mathcal{L}^{-1}$ is of the order h^{2k}, i.e., $\lambda_n(\mathcal{L}^{-1}) = \mathcal{O}(h^{2k})$; see [93, 36]. Therefore, the optimality above implies that the constructed localized basis Ψ^{loc} achieves nearly optimal performance on both ends in the accuracy-sparsity trade-off (2.6).

1. They are optimally localized up to a logarithmic factor, i.e.,
$$\left|\text{supp}(\psi_i^{loc})\right| \leq \frac{C_l \log(n)}{n}, \quad \forall 1 \leq i \leq n. \tag{3.13}$$
Here, $|\text{supp}(\psi_i^{loc})|$ 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 the Galerkin finite element method to solve the elliptic equations, we achieve the optimal convergence rate in the energy norm, i.e.,
$$\|\mathcal{L}^{-1} f - \Psi^{loc} L_n^{-1} (\Psi^{loc})^T f\|_H \leq C_e \sqrt{\lambda_n(\mathcal{L}^{-1})} \|f\|_2 \quad \forall f \in L^2(D), \tag{3.14}$$
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.,
$$E_{oc}(\Psi^{loc}; \mathcal{L}^{-1}) \leq C_e^2 \lambda_n(\mathcal{L}^{-1}), \tag{3.15}$$
where $E_{oc}(\Psi^{loc}; \mathcal{L}^{-1})$ is the operator compression error defined in Eqn. (2.2).

Outline Of This Chapter

The outline of this chapter is as follows. In Section 3.2, we prove the local projection-type approximation in the Sobolev spaces $H^k(\tau)$ (where $k \geq 1$ and τ is a subdomain of D). Based on this projection-type approximation, we provide the operator compression error estimate for the global energy minimizing basis Ψ. In Section 3.3, we prove the local inverse energy estimate. Then in Section 3.4, we introduce the concept of strong ellipticity, and show its relation with the uniform ellipticity. In Section 3.5, we prove that the global energy minimizing basis functions $\psi_{i,q}$ decays exponentially fast away from

its associated patch. In Section 3.6, we prove that localized basis functions $\psi_{i,q}^{loc}$ approximate $\psi_{i,q}$ accurately and preserve the optimal $\mathcal{O}(h^{2k})$ operator compression error. In Section 3.7, both 1D and 2D examples are provided to validate our theoretical results.

3.2 The Projection-type Polynomial Approximation Property and Error Estimates

The Projection-type Polynomial Approximation Property

When we compress the second order elliptic operator in Section 2.3, the Poincare inequality plays an essential role in both obtaining the optimal approximation error and proving the exponential decay of the energy minimizing basis functions. To prove the same kind of results for high-order elliptic operators, we first introduce the following projection-type polynomial approximation property in the Sobolev space $H^k(D)$, which can be viewed as a generalized Poincare inequality.

Theorem 3.2.1. *Suppose $\Omega \subset \mathbb{R}^d$ is affine equivalent to $\widehat{\Omega}$, i.e., there exists an invertible affine mapping*

$$F : \widehat{x} \in \widehat{\Omega} \to F(\widehat{x}) = B\widehat{x} + b \in \Omega \tag{3.16}$$

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 $0 \leq p \leq k+1$

$$|u - \Pi u|_{p,2,\Omega} \leq C(k, \widehat{\Omega})\delta^{-p} h^{k-p+1} |u|_{k+1,2,\Omega}. \tag{3.17}$$

To prove Theorem 3.2.1, 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 [31]), to prove Theorem 3.2.1.

Theorem 3.2.2. *For some integers $k \geq 0$ and $m \geq 0$, 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 Eqn.* (3.16)*), 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*,*

$$|v - \Pi_\Omega v|_{m,2,\Omega} \le C(\widehat{\Pi}, \widehat{\Omega})\delta^{-m} h^{k-m+1} |v|_{k+1,2,\Omega}, \qquad \forall v \in H^{k+1}(\Omega), \tag{3.18}$$

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 3.2.1.

Proof of Theorem 3.2.1. 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, 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 $\Pi_\Omega v \in \mathcal{P}_k(\Omega)$. Second, 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 $v \in \mathcal{P}_k(\Omega)$. Third, 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 Π_Ω is the orthogonal projection from $H^{k+1}(\Omega)$ to $\mathcal{P}_k(\Omega)$ in $L^2(\Omega)$.

Finally, applying Theorem 3.2.2 with $\widehat{\Pi}$ and Π_Ω above, we prove Theorem 3.2.1 with the constant $C(k,\widehat{\Omega}) := C(\widehat{\Pi},\widehat{\Omega})$ in Eqn. (3.18). □

We also give the following theorem, which is a direct result of the Friedrichs' inequality; see e.g., [95].

Theorem 3.2.3. *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*

$$|u|_{p,2,\Omega_h} \leq C_f h^{k-p} |u|_{k,2,\Omega_h} \quad \forall u \in H_0^k(\Omega_h). \tag{3.19}$$

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

The Error Estimate Of The Global Basis Ψ

Applying Theorem 3.2.1 to $\Omega = \tau_j$, for any $u \in H^k(D)$ and any $0 \leq p \leq k$, we have

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

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 $\{\tau_i\}_{1\leq i\leq m}$, because all elements in $\{\tau_i\}_{1\leq i\leq m}$ 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 $1 \leq i \leq m$ and any $0 \leq p \leq k$, we have

$$|u - \Pi_i u|_{p,2,\tau_i} \leq C_p h^{k-p} |u|_{k,2,\tau_i}. \tag{3.20}$$

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

$$\|u - \widetilde{u}\|_{L^2(D)} \leq C_p h^k |u|_{k,2,D}. \tag{3.21}$$

Combining Eqn. (3.5) and (3.21), we have

$$\|u - \mathcal{P}_\Phi^{(X)} u\|_{L^2(D)} \leq \frac{C_p h^k}{\sqrt{a_{min}}} \|u\|_H, \quad \forall u \in H. \tag{3.22}$$

Applying Theorem 2.2.1 with X and H defined above, we have:

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

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_H \leq \frac{C_p h^k}{\sqrt{a_{min}}} \|f\|_{L^2(D)} \,. \tag{3.23}$$

   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

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_{L^2(D)} \leq \frac{C_p^2 h^{2k}}{a_{min}} \|f\|_{L^2(D)} \,. \tag{3.24}$$

3. We have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \leq \frac{C_p^2 h^{2k}}{a_{min}} \,. \tag{3.25}$$

Notice that the eigenvalues of the operator $\mathcal{L}$ (with the homogeneous Dirichlet boundary conditions) in (3.1) grow like $\lambda_n(\mathcal{L}) \sim n^{2k/d}$ (see [93, 36]), 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 Eqn. (3.25), we have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \leq \frac{C_p^2 Q^{2k/d}}{a_{min}} n^{-2k/d} \underset{\sim}{<} \lambda_n(\mathcal{K}) \,. \tag{3.26}$$

Therefore, our construction of the m-dimensional subspace Ψ approximates $\mathcal{K}$ at the optimal rate.

3.3 The Inverse Energy Estimate

In the sparse operator compression, we will show that the global energy minimizing basis Ψ have exponentially decaying tails, which makes localization of these basis functions possible.

The Main Results

The following lemma plays a key role in proving such exponential decay property.

Lemma 3.3.1. *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.*

$$\begin{aligned} (-1)^k \sum_{|\sigma|=k} D^{2\sigma} u_h(x) = f(x) \qquad & x \in \Omega_h, \\ u_h \in H_0^k(\Omega_h). & \end{aligned} \tag{3.27}$$

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

$$\|\mathcal{L}u_h\|_{L^2(\Omega_h)} \leq C(k,s,d,\delta)h^{-k}|u_h|_{k,2,\Omega_h}, \quad \forall u_h \in \mathcal{L}^{-1}\mathcal{P}_{s-1}. \tag{3.28}$$

Proof. Let G_h be the Green's function of Eqn. (3.27). After multiplying u_h on both sides of Eqn. (3.27) 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 Eqn. (3.28) is equivalent to

$$\int_{\Omega_h} p^2(x)\mathrm{d}x \leq \left(C(k,s,d,\delta)\right)^2 h^{-2k} \int_{\Omega_h}\int_{\Omega_h} G_h(x,y)p(x)p(y)\mathrm{d}x\mathrm{d}y, \quad \forall p \in \mathcal{P}_{s-1}. \tag{3.29}$$

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 Eqn. (3.27) with right hand side p_i, and $S_h, M_h \in \mathbb{R}^{Q\times Q}$ be defined as follows:

$$S_h(i,j) = \int_{\Omega_h}\int_{\Omega_h} G_h p_i p_j = \int_{\Omega_h} u_{h,i}p_j, \qquad M_h(i,j) = \int_{\Omega_h} p_i p_j. \tag{3.30}$$

Then, Eqn. (3.29) is equivalent to

$$M_h \preceq \left(C(k,s,d,\delta)\right)^2 h^{-2k} S_h, \tag{3.31}$$

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 $\Omega_1 \equiv \{x/h \ : \ x \in \Omega_h\}$:

$$\begin{aligned} (-1)^k \sum_{|\sigma|=k} D^{2\sigma}u_{1,i}(x) = p_i(x) \qquad & x \in \Omega_1, \\ u_{1,i} \in H_0^k(\Omega_1), & \end{aligned} \tag{3.32}$$

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

$$S_h(i,j) = h^{2k+o_i+o_j+d}S_1(i,j), \qquad M_h(i,j) = h^{o_i+o_j+d}M_1(i,j), \tag{3.33}$$

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

$$C(k,s,d,\Omega_1) = \sqrt{\lambda_{max}(M_1,S_1)}, \tag{3.34}$$

we have

$$M_1 \preceq \left(C(k,s,d,\Omega_1)\right)^2 S_1. \tag{3.35}$$

Combining (3.33) and (3.35), Eqn. (3.31) naturally follows. In Proposition (3.3.1) in next subsection, 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 (3.3.1). 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 [99] 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 3.3.1 is much tighter. Moreover, Lemma 3.3.1 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.

More On Lemma 3.3.1

In this subsection, we prove that $C(k,s,d,\Omega_1)$ can be bounded by $C(k,s,d,\delta)$,and we 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 3.3.2. *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 $f \in L^2(\Omega)$,*

$$\int_S \int_S G_S(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y \leq \int_\Omega \int_\Omega G_\Omega(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y. \tag{3.36}$$

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

$$\begin{aligned} \psi_\Omega = \operatorname*{arg\,min}_{u\in H_0^k(\Omega)} I_\Omega(u,f), \qquad \psi_S = \operatorname*{arg\,min}_{u\in H_0^k(S;\Omega)} I_\Omega(u,f) \\ H_0^k(S;\Omega) := \{u \in H_0^k(\Omega) : u \equiv 0 \text{ on } \Omega\backslash S\}. \end{aligned} \tag{3.37}$$

Moreover, we have

$$\begin{aligned} I_\Omega(\psi_\Omega, f) = -\frac{1}{2}\int_\Omega \psi_\Omega f = -\frac{1}{2}\int_\Omega\int_\Omega G_\Omega(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y, \\ I_\Omega(\psi_S, f) = -\frac{1}{2}\int_S \psi_S f = -\frac{1}{2}\int_S\int_S G_S(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{3.38}$$

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

$$I_\Omega(\psi_\Omega, f) \le I_\Omega(\psi_S, f), \tag{3.39}$$

which proves the lemma. □

Notice that Lemma 3.3.2 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 3.3.2, 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 3.3.1. *$C(k,s,d,\Omega)$ (defined in Eqn. (3.34)) can be bounded by $C(k,s,d,\delta)$ which only depends on k,s,d and δ. Moreover, we can set*

$$C(1,1,d,\delta) = 2\sqrt{d(d+2)}\delta^{-1-d/2}. \tag{3.40}$$

Proof. From the definition (3.34), we have

$$\left(C(k,s,d,\Omega)\right)^2 = \lambda_{max}(M,S) = \max_{p\in\mathcal{P}_{s-1}} \frac{\int_\Omega p^2(x)\mathrm{d}x}{\int_\Omega\int_\Omega G(x,y)p(x)p(y)\mathrm{d}x\mathrm{d}y}, \tag{3.41}$$

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 3.3.2, 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

$$\widehat{S}(i,j) = \int_{B(0,\delta/2)}\int_{B(0,\delta/2)} G_{\delta/2}p_ip_j = \int_{B(0,\delta/2)} u_{_{\delta/2,i}}p_j, \qquad \widehat{M}(i,j) = \int_{B(0,1)} p_ip_j. \tag{3.42}$$

Here, $\{p_1,p_2,\ldots,p_Q\}$ are all the monomials defined in Lemma 3.3.1 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. Therefore, we can choose

$$C(k,s,d,\delta) = \sqrt{\lambda_{max}(\widehat{M},\widehat{S})}. \tag{3.43}$$

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 $\int_\Omega p^2(x)\mathrm{d}x \le \int_{B(x_0,1/2)} p^2(x)\mathrm{d}x$. Therefore, 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 this $s=1$ case, Eqn. (3.40) naturally follows. □

3.4 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 3.4.1. *An operator in 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* $\theta_{min} > 0$ *such that*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge \theta_{min}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2, \quad \forall x\in D, \boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d-1}{k}}, \tag{3.44}$$

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 k-th 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 divergence form $\widetilde{\mathcal{L}}$ such that $\mathcal{L}u = \widetilde{\mathcal{L}}u$ for all $u \in C^{2k}(D)$.

Remark 3.4.1. *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* (3.4.1) *while the other does. For example, the biharmonic operator $\mathcal{L} = \Delta^2$ in 2d physical domain have the following two different divergence forms:*

$$\mathcal{L}u = \sum_{|\sigma|=|\gamma|=2} D^\sigma(a_{\sigma\gamma} D^\gamma u) = \sum_{|\sigma|=|\gamma|=2} D^\sigma(\widetilde{a}_{\sigma\gamma}(x) D^\gamma u), \tag{3.45}$$

where

$$(a_{\sigma\gamma}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (\widetilde{a}_{\sigma\gamma}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \tag{3.46}$$

when $\{D^\sigma u : |\sigma| = 2\}$ is ordered as $(\partial^2_{x_1}, \partial^2_{x_2}, \partial_{x_1}\partial_{x_2})$. Obviously, the first one does not satisfy the strong ellipticity condition (3.4.1) *while the second one does. These two divergence forms correspond to two bilinear forms on $H^2_0(D)$:*

$$B(u,v) = \int_D \Delta u \Delta v, \quad \widetilde{B}(u,v) = \int_D D^2 u : D^2 v, \tag{3.47}$$

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

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

Lemma 3.4.1. *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.*

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge 0, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}, \tag{3.48}$$

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

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \leq \theta_{k,max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 + \theta_{0,max}\sum_{|\sigma|<k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall\boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d}{k}}, \tag{3.49}$$

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

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \geq \theta_{min}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall\boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d-1}{k}}. \tag{3.50}$$

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

$$\|\psi\|^2_{H(S)} = \sum_{0\leq|\sigma|,|\gamma|\leq k}\int_S a_{\sigma\gamma}(x)D^\sigma\psi D^\gamma\psi. \tag{3.51}$$

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*

$$|\psi|_{k,2,S} \leq \theta_{min}^{-1/2}\|\psi\|_{H(S)}, \qquad \forall\psi\in H^k(D). \tag{3.52}$$

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

$$|\psi_{i,q}|_{k,2,S} \leq (2/\theta_{min})^{1/2}\,\|\psi_{i,q}\|_{H(S)}, \qquad \forall i\notin\mathcal{S}, 1\leq q\leq Q. \tag{3.53}$$

  Here, Λ is any subset of $\{1,2,\ldots,m\}$, and $\psi_{i,q}$ is defined by Eqn. (3.8).

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

$$\begin{aligned}\|\psi\|^2_{H(S)} = &\underbrace{\sum_{|\sigma|=|\gamma|=k}\int_S a_{\sigma\gamma}D^\sigma\psi D^\gamma\psi}_{J_1} + \underbrace{\sum_{|\sigma|,|\gamma|<k}\int_S a_{\sigma\gamma}D^\sigma\psi D^\gamma\psi}_{J_2}\\ &+\underbrace{\sum_{|\sigma|=k,|\gamma|<k}\int_S(a_{\sigma\gamma}+a_{\gamma\sigma})D^\sigma\psi D^\gamma\psi}_{J_3}.\end{aligned} \tag{3.54}$$

From the strong ellipticity (3.50), we have

$$J_1 \geq \theta_{min}|\psi|^2_{k,2,S}. \tag{3.55}$$

From the nonnegativity (3.48), we have

$$J_2 \geq 0. \tag{3.56}$$

Combining the nonnegativity (3.48) and the boundedness (3.49), we can prove that

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

Therefore, using the Cauchy-Schwartz inequality, we obtain

$$|J_3| \leq 2\theta^{1/2}_{0,max}\theta^{1/2}_{k,max}|\psi|_{k,2,S}\|\psi\|_{k-1,2,S}. \tag{3.57}$$

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

$$\|\psi_{i,q}\|^2_{k-1,2,S} \leq C_p^2 \frac{h^2(1-h^{2k})}{1-h^2} |\psi_{i,q}|^2_{k,2,S}. \tag{3.58}$$

Combining Eqn. (3.57) and (3.58), for $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta^2_{min}}{16\theta_{0,max}\theta_{k,max}C_p^2}$, we have

$$|J_3| \leq \frac{\theta_{min}}{2}|\psi|^2_{k,2,S}. \tag{3.59}$$

Combining Eqn. (3.54), (3.55), (3.56), and (3.59), we prove the second point. □

Remark 3.4.2. *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 Eqn.* (3.52) *for all $h > 0$.*

The strong ellipticity condition above is different from the standard uniformly elliptic condition (see Definition 9.2 in [113]), i.e., a linear partial differential operator $\mathcal{L}u = (-1)^k \sum_{|\alpha|\leq 2k} a_\alpha D^\alpha u$ is uniformly elliptic if there exists a constant $\theta_{min} > 0$ such that

$$\sum_{|\alpha|=2k} a_\alpha(x)\boldsymbol{\xi}^\alpha \geq \theta_{min}|\boldsymbol{\xi}|^{2k}, \quad \forall x \in D, \boldsymbol{\xi} \in \mathbb{R}^d. \tag{3.60}$$

On 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 Eqn. (3.44). 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 (3.44) is equivalent to that there exists $\theta_{min} > 0$ such that

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

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 Appendix A.1.

Theorem 3.4.2. *Let $a_\alpha \in C^{|\alpha|-k}(\overline{D})$ for $k < |\alpha| \le 2k$, $a_\alpha \in C(\overline{D})$ 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.

3.5 Exponential Decay of The Basis Functions

Exponential decay of basis functions I

In this subsection, we prove the exponential decay of basis functions constructed in Eqn. (3.8) 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 3.5.1. *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*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \le \theta_{k,max} \sum_{|\sigma|=k} \boldsymbol{\zeta}_\sigma^2, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}, \tag{3.61}$$

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

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge \theta_{k,min} \sum_{|\sigma|=k} \boldsymbol{\zeta}_\sigma^2, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}. \tag{3.62}$$

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

$$\|\psi_{i,q}\|^2_{H(D\cap(B(x_i,r))^c)} \le \exp(1 - \frac{r}{lh})\|\psi_{i,q}\|^2_{H(D)} \tag{3.63}$$

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 Eqn.* (3.20) *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 2.3.1 and [99] (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 2.3.1 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 $D^\sigma(\eta\psi_{i,q}) = \eta D^\sigma\varphi_{i,q} + \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}$, we obtain

$$\begin{aligned}\sum_{|\sigma|=|\gamma|=k} \int_D \eta a_{\sigma\gamma}(x) D^\sigma(\psi_{i,q}) D^\gamma \psi_{i,q} &= \underbrace{\int_D \eta\psi_{i,q}\mathcal{L}\psi_{i,q}}_{I_2} \\ \underbrace{- \sum_{|\sigma|=|\gamma|=k} \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} \int_D a_{\sigma\gamma}(x) D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} D^\gamma\psi_{i,q}\,.}_{I_1} &\end{aligned} \tag{3.64}$$

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

$$I_1 = - \sum_{|\sigma|=|\gamma|=k} \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} \int_{S^*} a_{\sigma\gamma}(x) D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} D^\gamma\psi_{i,q} \tag{3.65}$$

$$\leq \left(\sum_{|\sigma|=k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \sqrt{\theta_{k,max}} \tag{3.66}$$

$$\leq C_1 C_\eta \left(\sum_{s'=1}^{k} (lh)^{-2s'} |\psi_{i,q}|^2_{k-s',2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \sqrt{\theta_{k,max}}\,. \tag{3.67}$$

Here, C_1 is a constant only dependent on k and d. We have used the Cauchy-Schwarz inequality and the bound (3.61) in Eqn. (3.66). We will defer the proof of the last step in Eqn. (3.67) to the Appendix. Since $\psi_{i,q} \perp \mathcal{P}_{k-1}$ locally in L^2, we obtain from Theorem 3.2.1 that

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

Therefore, we get

$$I_1 \leq C_1 C_\eta \sqrt{\theta_{k,max}} C_p \left(\sum_{s'=1}^{k} l^{-2s'} |\psi_{i,q}|^2_{k,2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \tag{3.68}$$

$$\leq \frac{C_1 C_\eta \sqrt{\theta_{k,max}} C_p}{\sqrt{l^2-1}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{3.69}$$

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

By the construction of $\psi_{i,q}$ given in (3.8), we have $\int_D \psi_{i,q}\varphi_{j,q'} = 0$ for $i \neq j$. Thanks to (2.18), 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

$$I_2 = \int_{S^*} \eta\psi_{i,q}\mathcal{L}\psi_{i,q} = \sum_{\tau_j \in S^*} \int_{\tau_j} (\eta - \eta_j)\psi_{i,q}\mathcal{L}\psi_{i,q} \leq \frac{C_\eta}{l} \sum_{\tau_j \in S^*} \|\psi_{i,q}\|_{L^2(\tau_j)} \|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)}. \tag{3.70}$$

By using Lemma 3.5.1, which is stated in the beginning of Section 3.5, we have $\|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)} \leq \sqrt{\theta_{k,max}} C(k,d,\delta) h^{-k} \|\psi_{i,q}\|_{H(\tau_j)}$ for any $h > 0$ because $\mathcal{L}$ contains only the highest order derivatives. Then we obtain

$$\begin{aligned} I_2 &\leq \frac{\sqrt{\theta_{k,max}} C_\eta C(k,d,\delta)}{l h^k} \|\psi_{i,q}\|_{L^2(S^*)} \|\psi_{i,q}\|_{H(S^*)} \\ &\leq \frac{\sqrt{\theta_{k,max}} C_\eta C(k,d,\delta) C_p}{l} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}, \end{aligned} \tag{3.71}$$

where we have used Eqn. (3.20) in the last step.

Combining Eqn. (3.69) and (3.71), we obtain

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

By the strong ellipticity (3.62) and Eqn. (3.52), we have $|\psi_{i,q}|_{k,2,S^*} \leq \theta_{k,min}^{-1/2} \|\psi_{i,q}\|_{H(S^*)}$. Therefore, we have

$$\|\psi_{i,q}\|^2_{H(S^1)} \leq \sqrt{\frac{\theta_{k,max}}{(l^2-1)\theta_{k,min}}} C_\eta C_p (C_1 + C(k,d,\delta)) \|\psi_{i,q}\|^2_{H(S^*)}. \tag{3.72}$$

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. □

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 3.5.2, we need the polynomial approximation property (3.20) and the Friedrichs' inequality (3.19) to bound the lower order terms, and we get an extra factor of 2 in our error bound.

Theorem 3.5.2. *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.*

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge 0, \qquad \forall x\in D, \quad \forall \boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d}{k}}, \tag{3.73}$$

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

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \le \theta_{k,max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2+\theta_{0,max}\sum_{|\sigma|<k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall x\in D, \quad \forall \boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d}{k}}, \tag{3.74}$$

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

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge \theta_{k,min}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall \boldsymbol{\zeta}\in\mathbb{R}^{\binom{k+d-1}{k}}. \tag{3.75}$$

Then 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*

$$\|\psi_{i,q}\|^2_{H(D\cap(B(x_i,r))^c)} \le \exp(1-\frac{r}{lh})\|\psi_{i,q}\|^2_{H(D)} \tag{3.76}$$

with $\sqrt{l^2-1}\ge 2(e-1)C_\eta C_p(C_1+C(k,d,\delta))\sqrt{\frac{\theta_{k,max}}{\theta_{k,min}}}$. *Here,* C_1 *and* C_η *depend on* k *and* d *only,* C_p *is the constant given in Eqn.* (3.20), $C(k,d,\delta):=C(k,k,d,\delta)$ *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*

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

where C_f *is the constant in the Friedrichs' inequality* (3.19).

Proof. The proof follows the same structure as the proof of Theorem 3.5.1. All we need to do is to use the polynomial approximation property (3.20) and the Friedrichs' inequality (3.19) to bound the lower order terms when they appear. First, the I_1 in Eqn. (3.64) contains all the lower order terms and its estimation should be modified as follows:

$$\begin{aligned} I_1 &= -\sum_{0\le|\sigma|,|\gamma|\le k}\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*}a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma\psi_{i,q} && (3.77)\\ &\le \left(\sum_{|\sigma|\le k}\int_{S^*}\left|\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}\right|^2\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,max}} && (3.78)\\ &\le C_1C_\eta\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}(lh)^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,max}}. && (3.79)\end{aligned}$$

Here, $\theta_{k,max} := \max(\theta_{0,max}, \theta_{k,max})$. We have used the Cauchy-Schwarz inequality and the bound (3.74) in Eqn. (3.78). We will defer the proof of the last step in Eqn. (3.79) to the Appendix. Since $\psi_{i,q} \perp \mathcal{P}_{k-1}$ locally in L^2, we obtain from Theorem 3.2.1 that

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

Therefore, we have

$$\begin{aligned} I_1 &\leq C_1 C_\eta \sqrt{\theta_{k,max}} C_p \left(\sum_{s=1}^{k} \sum_{s'=1}^{s} l^{-2s'} |\psi_{i,q}|^2_{s,2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} && (3.80) \\ &\leq \frac{C_1 C_\eta \sqrt{\theta_{k,max}} C_p}{\sqrt{l^2-1}} \left(\sum_{s=1}^{k} |\psi_{i,q}|^2_{s,2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} && (3.81) \\ &\leq \frac{C_1 C_\eta \sqrt{2\theta_{k,max}} C_p}{\sqrt{l^2-1}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. && (3.82) \end{aligned}$$

If we compare the above estimate with Eqn. (3.69), we conclude that Eqn. (3.81) contains all the lower order terms. We will use the polynomial approximation property (3.20) and take $\frac{h^2-h^{2k}}{1-h^2} \leq 1/C_p^2$ to guarantee that Eqn. (3.82) is valid. When $\mathcal{L}$ contains lower order terms, by Lemma 3.5.1, we have $\|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)} \leq \sqrt{2\theta_{k,max}} C(k,d,\delta) h^{-k} \|\psi_{i,q}\|_{H(\tau_j)}$ for any $h > 0$ satisfying $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,max}}{2\theta_{0,max} C_f^2}$. Therefore, using Eqn. (3.71) we get

$$I_2 \leq \frac{\sqrt{2\theta_{k,max}} C_\eta C(k,d,\delta) C_p}{l} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}, \tag{3.83}$$

when h satisfies $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,max}}{2\theta_{0,max} C_f^2}$. Finally, we need to use Eqn. (3.53) instead of Eqn. (3.52) to bound $|\psi_{i,q}|_{k,2,S^*}$. We get

$$\|\psi_{i,q}\|^2_{H(S^1)} \leq 2 \sqrt{\frac{\theta_{k,max}}{(l^2-1)\theta_{k,min}}} C_\eta C_p (C_1 + C(k,d,\delta)) \|\psi_{i,q}\|^2_{H(S^*)}, \tag{3.84}$$

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

Remark 3.5.1. *As we have pointed out in Remark 3.4.2, when $\mathcal{L}$ contains low order terms but there is no crossing term between $D^\sigma u$ ($|\sigma| = k$) and $D^\sigma u$ ($|\sigma| < k$), Eqn.* (3.52) *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

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

Lemmas

In this subsection, we will prove the following lemma, which is used in the proof of Theorem 3.5.1 and Theorem 3.5.2.

Lemma 3.5.1. *$\mathcal{L}$ is defined in Eqn.* (3.1) *and the space Ψ is defined as above. Assume that for any $x \in D$*

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \leq \theta_{k,max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 + \theta_{0,max}\sum_{|\sigma|<k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}. \quad (3.85)$$

Let C_f be the constant in the Friedrichs' inequality (3.19)*. 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*

$$\|\mathcal{L}v\|_{L^2(\tau_j)} \leq \sqrt{2\theta_{k,max}}C(k,d,\delta)h^{-k}\|v\|_{H(\tau_j)}, \quad \forall v \in \Psi,\ \forall j = 1,2,\ldots,m, \tag{3.86}$$

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 3.5.2), and to use the polynomial approximation property to deal with the low order terms; see Lemma 3.5.3. For this purpose, we first introduce the following two lemmas.

Lemma 3.5.2. *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

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \le \sum_{0\le|\sigma|,|\gamma|\le k} b_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}. \tag{3.87}$$

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

$$\int_\Omega\int_\Omega G_{\mathcal{M}}(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y \le \int_\Omega\int_\Omega G_{\mathcal{L}}(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y. \tag{3.88}$$

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

$$\begin{aligned} I_{\mathcal{L}}(u,f) &= \frac{1}{2}\sum_{0\le|\sigma|,|\gamma|\le k}\int_D a_{\sigma\gamma}(x)D^\sigma u D^\gamma u - \int_\Omega uf \quad u \in H_0^k(\Omega), \\ I_{\mathcal{M}}(u,f) &= \frac{1}{2}\sum_{0\le|\sigma|,|\gamma|\le k}\int_D b_{\sigma\gamma}(x)D^\sigma u D^\gamma u - \int_\Omega uf \quad u \in H_0^k(\Omega). \end{aligned} \tag{3.89}$$

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

$$\begin{aligned} I_{\mathcal{L}}(\psi_{\mathcal{L}},f) &= -\frac{1}{2}\int_\Omega \psi_{\mathcal{L}} f = -\frac{1}{2}\sum_{0\le|\sigma|,|\gamma|\le k}\int_D a_{\sigma\gamma}(x)D^\sigma\psi_{\mathcal{L}}D^\gamma\psi_{\mathcal{L}}, \\ I_{\mathcal{M}}(\psi_{\mathcal{M}},f) &= -\frac{1}{2}\int_\Omega \psi_{\mathcal{M}} f = -\frac{1}{2}\sum_{0\le|\sigma|,|\gamma|\le k}\int_D a_{\sigma\gamma}(x)D^\sigma\psi_{\mathcal{M}}D^\gamma\psi_{\mathcal{M}}. \end{aligned} \tag{3.90}$$

Observe that

$$I_{\mathcal{L}}(\psi_{\mathcal{L}},f) \le I_{\mathcal{L}}(\psi_{\mathcal{M}},f) \le I_{\mathcal{M}}(\psi_{\mathcal{M}},f), \tag{3.91}$$

where the first inequality is true because $\psi_{\mathcal{L}}$ is the minimizer of $I_{\mathcal{L}}$, and the second inequality is true because $I_{\mathcal{L}}(u,f) \le I_{\mathcal{M}}(u,f)$ for any $u \in H_0^k(\Omega)$. Combining Eqn. (3.90) and (3.91), we obtain $\int_\Omega \psi_{\mathcal{M}} f \le \int_\Omega \psi_{\mathcal{L}} f$. This proves the lemma. □

Lemma 3.5.3. *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 $\mathcal{L}_0 u = (-1)^k\sum_{|\sigma|=k} D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\Omega_h$. Here, $c > 0$ is a positive constant. Then for any $f \in L^2(\Omega_h)$*

$$\lim_{h\to 0}\frac{\int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y} = 1. \tag{3.92}$$

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

$$\begin{aligned} I_{\mathcal{L}}(u,f) &= \frac{1}{2}|u|^2_{k,2,\Omega_h} + \frac{c}{2}\|u\|^2_{k-1,2,\Omega_h} - \int_{\Omega_h} uf, \\ I_{\mathcal{L}_0}(u,f) &= \frac{1}{2}|u|^2_{k,2,\Omega_h} - \int_{\Omega_h} uf. \end{aligned} \tag{3.93}$$

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

$$\begin{aligned} I_{\mathcal{L}}(\psi_h,f) &= -\frac{1}{2}\int_{\Omega_h}\psi_h f = -\frac{1}{2}\left(|\psi_h|^2_{k,2,\Omega_h} + c\|\psi_h\|^2_{k-1,2,\Omega_h}\right), \\ I_{\mathcal{L}_0}(\psi_{\mathcal{L}_0},f) &= -\frac{1}{2}\int_{\Omega_h}\psi_{h,0} f = -\frac{1}{2}|\psi_{h,0}|^2_{k,2,\Omega_h}. \end{aligned} \tag{3.94}$$

Note that Eqn. (3.94) implies that $I_{\mathcal{L}_0}(\psi_{h,0},f) < 0$. By the definition of Green's function, we further have

$$\begin{aligned} \int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y &= \int_{\Omega_h}\psi_h f = -2I_{\mathcal{L}}(\psi_h,f) = |\psi_h|^2_{k,2,\Omega_h} + c\|\psi_h\|^2_{k-1,2,\Omega_h}, \\ \int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y &= \int_{\Omega_h}\psi_{h,0} f = -2I_{\mathcal{L}_0}(\psi_{h,0},f) = |\psi_{h,0}|^2_{k,2,\Omega_h}. \end{aligned} \tag{3.95}$$

Since $I_{\mathcal{L}_0}(u,f) \leq I_{\mathcal{L}}(u,f)$ for any $u \in H^k_0(\Omega)$, we have $\frac{\int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\mathrm{d}y} \leq 1$ for any $h > 0$. Applying the Friedrich's inequality (3.19) to $\|\psi_{h,0}\|^2_{k-1,2,\Omega_h}$, we get

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

Here, we have used Eqn. (3.95) in the last equality. Therefore, we have

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

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

Now we are ready to prove Lemma 3.5.1.

Proof of Lemma 3.5.1. Let $v = \sum_{i=1}^{m} \sum_{q=1}^{Q} c_{i,q}\psi_{i,q}$. Thanks to Eqn. (2.18), 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

$$\|\mathcal{L}v\|^2_{L^2(\tau_j)} = \|g_j\|^2_{L^2(\tau_j)}. \tag{3.96}$$

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 $v_1 \in H^k_0(\tau_j)$, and v_2 solves $\mathcal{L}v_2 = 0$ with $v_2 - v \in H^k_0(\tau_j)$. 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 3.5.2, we have

$$\|v_1\|^2_{H(\tau_j)} \geq \frac{1}{\theta_{k,max}} \int_{\tau_j}\int_{\tau_j} G^*_j(x,y)g_j(x)g_j(y)\mathrm{d}x\mathrm{d}y, \tag{3.97}$$

where G^*_j is the Green's function of the operator $(-1)^k \sum\limits_{|\sigma|=k} D^{2\sigma}u + \frac{\theta_{k,max}}{\theta_{0,max}} \sum\limits_{|\sigma|<k} (-1)^\sigma D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\tau_j$. Thanks to Lemma 3.5.3, 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

$$\int_{\tau_j}\int_{\tau_j} G^*_j(x,y)g_j(x)g_j(y)\mathrm{d}x\mathrm{d}y \geq \frac{1}{2}\int_{\tau_j}\int_{\tau_j} G^*_{j,0}(x,y)g_j(x)g_j(y)\mathrm{d}x\mathrm{d}y, \tag{3.98}$$

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

$$\|g_j\|^2_{L^2(\tau_j)} \leq \left(C(k,k,d,\delta)\right)^2 h^{-2} \int_{\tau_j}\int_{\tau_j} G^*_{j,0}(x,y)g_j(x)g_j(y)\mathrm{d}x\mathrm{d}y. \tag{3.99}$$

Combining Eqn. (3.97), (3.98), and (3.99), we have

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

Therefore, we have proved Lemma 3.5.1. 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 do not need to pay a factor of 2 in Eqn. (3.98), 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:

$$\mathcal{L}u = f(x) \qquad x \in \tau_i, \quad u \in H_0^k(\tau_i). \tag{3.100}$$

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

$$M_0(q,q') = \int_{\tau_i} \varphi_{i,q}\varphi_{i,q'}, \qquad A_0(q,q') = \int_{\tau_i} \varphi_{i,q} \mathcal{L}_0^{-1}(a\varphi_{i,q'}). \tag{3.101}$$

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

$$\lambda_{max}(M_0, A_0) = \sup_{v\in\mathbb{R}^Q} \frac{v^T M_0 v}{v^T A_0 v} = \sup_{\varphi\in\mathcal{P}_k(\tau_i)} \frac{\|\varphi\|^2_{L^2(\tau_i)}}{\|\mathcal{L}_0^{-1}\varphi\|^2_{H(\tau_i)}}. \tag{3.102}$$

The proof of Lemma 3.5.1 also implies that

$$\sqrt{\lambda_{max}(M_0, A_0)} \leq \sqrt{2\theta_{k,max}} C(k,d,\delta) h^{-k}. \tag{3.103}$$

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

$$\sqrt{\lambda_{max}(M_0, A_0)} \leq \sqrt{\theta_{k,max}} C(k,d,\delta) h^{-k}. \tag{3.104}$$

3.6 Localization of The Basis Functions

Lemma 3.6.1. *For any domain partition with* $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,max}}{2\theta_{0,max} C_f^2}$, *it holds true that*

$$\|\psi^{loc}_{i,q}\|_H \leq C(k,d,\delta) \left(\frac{2^{d+1}\theta_{k,max}}{V_d \delta^d}\right)^{1/2} h^{-d/2-k}. \tag{3.105}$$

If the operator $\mathcal{L}$ *contains only the highest order terms, it holds true that* $\|\psi^{loc}_{i,q}\|_H \leq C(k,d,\delta) \left(\frac{2^d\theta_{k,max}}{V_d\delta^d}\right)^{1/2} h^{-d/2-k}$ *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 Eqn. (3.103)) and $\mathcal{L}_0^{-1}\varphi_{i,q'}$ is the weak solution of the local problem (3.100) 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 $\psi_{i,q}^{loc}$ (see Eqn. (3.9)), and thus we get

$$\|\psi_{i,q}^{loc}\|_H \leq \|\zeta_{i,q}\|_H. \tag{3.106}$$

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

$$\|\zeta_{i,q}\|_H^2 = A_0^{-1}(q,q) \leq \lambda_{max}(A_0^{-1}) = \frac{\lambda_{max}(M_0, A_0)}{|\tau_i|}, \tag{3.107}$$

where we have used $M_0(q,q') = |\tau_i|\delta_{i,j}$ (due to the normalization (3.7)) in the last identity. Combining Eqn. (3.104) (or (3.103)), (3.106), and (3.107) and $|\tau_i| \geq V_d(\delta h/2)^d$, we complete the proof of Eqn. (3.105). □

Theorem 3.6.1. *Under the same assumptions as those in Theorem 3.5.2, 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*

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)} \leq C_3 h^{-d/2-k}\exp(-\frac{r-2h}{2lh}), \tag{3.108}$$

where

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

Here, all the parameters are the same as those in Theorem 3.5.2.

When the operator $\mathcal{L}$ contains only the highest order terms, i.e. $\mathcal{L}u = (-1)^k\sum_{|\sigma|=|\gamma|=k} D^\sigma(a_{\sigma\gamma}D^\gamma u)$, Eqn. (3.108) holds true for all $h > 0$. In this case, the constant C_3 can be taken as

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

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 $r \geq 2h$ 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

Figure 3.1. 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 Figure 2.2.

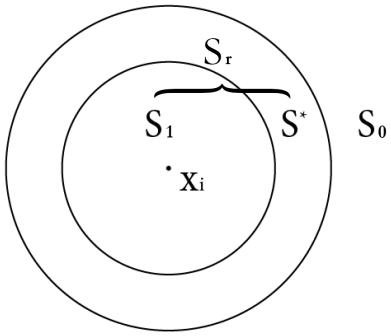


Figure 3.1: Illustration of S_r, S_0, S_1 and S^*.

Let η be a smooth cut-off function such that $0 \leq \eta \leq 1$, $\eta|_{S_1} \equiv 1$, $\eta|_{S_0} \equiv 0$ and $\|D^\sigma \eta\|_{L^\infty(D)} \leq \frac{C_\eta}{h^{|\sigma|}}$ for all σ. Since $\psi_{i,q}^{loc}$ satisfies the same constraints as those in the definition of $\psi_{i,q}$, thanks to Eqn. (2.19) we have

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)}^2 = \|\psi_{i,q}^{loc}\|_{H(D)}^2 - \|\psi_{i,q}\|_{H(D)}^2. \tag{3.109}$$

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

$$\begin{aligned} \psi_{j,q}^{i,r} := \underset{\psi \in H_0^k(S_r)}{\arg\min} \quad & \|\psi\|_{H(S_r)}^2 \\ \text{s.t.} \quad & \int_{S_r} \psi \varphi_{j',q'} = \delta_{jq,j'q'}, \quad \forall 1 \leq j' \leq m,\, 1 \leq q' \leq Q. \end{aligned} \tag{3.110}$$

Note that $\psi_{i,q}^{loc} = \psi_{i,q}^{i,r}$. Let $w_{jq'} = \int_D \eta \psi_{i,q} \varphi_{j,q'}$ and $\psi_w^{iq,r} = \sum_{j=1}^m \sum_{q'=1}^Q w_{jq'} \psi_{j,q'}^{i,r}$. Thanks to the orthogonality between $\psi_{i,q}$ and $\varphi_{j,q'}$, i.e. the constraints in Eqn. (3.8), we have

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

Using (3) of Theorem 2.2.3, we have $(\psi_{i,q}^{loc}, \psi_{j,q'}^{i,r})_H = \Theta_{iq,jq'}^{i,-1}$, where Θ^i is defined by Eqn. (2.17) 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

$$\|\psi_w^{iq,r}\|_H^2 = \|\psi_{i,q}^{loc}\|_H^2 + \|\sum_{\tau_j \subset S^*} \sum_{q'=1}^Q w_{jq'} \psi_{j,q'}^{i,r}\|_H^2 + 2 \sum_{\tau_j \subset S^*} \sum_{q'=1}^Q w_{jq'} \Theta_{iq,jq'}^{i,-1}. \tag{3.111}$$

By (2) of Theorem 2.2.3, we know that $\psi_w^{iq,r}$ is the minimizer of the following quadratic problem:

$$\begin{aligned}\psi_w^{iq,r} = \underset{\psi\in H_0^k(S_r)}{\arg\min} \quad & \|\psi\|^2_{H(S_r)} \\ \text{s.t.} \quad & \int_{S_r}\psi\varphi_{j,q'} = \int_D \eta\psi_{i,q}\varphi_{j,q'}, \quad \forall 1\le j\le m,\, 1\le q'\le Q.\end{aligned} \tag{3.112}$$

Noting that $\eta\psi_{i,q}$ satisfies the same constraint, we have $\|\psi_w^{iq,r}\|_H^2 \le \|\eta\psi_{i,q}\|_H^2$. By using this estimate with (3.109) and (3.111), we obtain

$$\|\psi_{i,q}-\psi_{i,q}^{loc}\|^2_{H(D)} \le \underbrace{\|\eta\psi_{i,q}\|_H^2 - \|\psi_{i,q}\|_H^2}_{I_1} + \underbrace{2\Big|\sum_{\tau_j\subset S^*}\sum_{q'=1}^{Q} w_{jq'}\Theta^{i,-1}_{iq,jq'}\Big|}_{I_2}. \tag{3.113}$$

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 3.5.2 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 A.2, we give a bound for $\|\eta\psi_{i,q}\|_{H(S^*)}$ using a similar technique that we used to obtain Eqn. (3.82) from Eqn. (3.77) in the proof of Theorem 3.5.2. With this bound, we obtain

$$I_1 \le \left(\frac{C_3}{2}|\psi_{i,q}|_{k,2,S^*} + \sqrt{\frac{C_3^2}{4}|\psi_{i,q}|^2_{k,2,S^*} + C_3|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}\|_{H(S^*)} + \|\psi_{i,q}\|^2_{H(S^*)}}\right)^2, \tag{3.114}$$

where $C_3 = C_1C_\eta C_p\sqrt{2k\theta_{k,max}}$. With the strong ellipticity (3.75) and the bound (3.53), we conclude

$$I_1 \le \left(2C_1C_\eta C_p\sqrt{\frac{k\theta_{k,max}}{\theta_{k,min}}} + 1\right)^2 \|\psi_{i,q}\|^2_{H(S^*)}. \tag{3.115}$$

Applying the exponential decay of Theorem 3.5.2 to $\|\psi_{i,q}\|_{H(S^*)}$, we get

$$I_1 \le \left(2C_1C_\eta C_p\sqrt{\frac{k\theta_{k,max}}{\theta_{k,min}}} + 1\right)^2 e^{1-\frac{r-2h}{lh}} \|\psi_{i,q}\|^2_{H(D)}. \tag{3.116}$$

We now estimate I_2. Combining (3) of Theorem 2.2.3 with the definition of H-norm (2.8), we have

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

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

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

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

$$\|\mathcal{L}\psi_{i,q}^{loc}\|_{L^2(\tau_j)} = |\tau_j|^{1/2}\left(\sum_{q'=1}^{Q}(\Theta_{iq,jq'}^{i,-1})^2\right)^{1/2}. \tag{3.117}$$

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

$$|\tau_j|^{-1/2}\left(\sum_{q'=1}^{Q}|w_{jq'}|^2\right)^{1/2} \leq \|\eta\psi_{i,q}\|_{L^2(\tau_j)} \leq \|\psi_{i,q}\|_{L^2(\tau_j)}, \tag{3.118}$$

where we have made use of $0 \leq \eta \leq 1$ in the last step. Combining (3.117) and (3.118), we get

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

Now, we arrive at exactly the same situation as I_2 (see (3.70)) in the proof of Theorem 3.5.1. With the same derivation from Eqn. (3.70) to Eqn. (3.71), i.e. applying Lemma 3.5.1 to $\|\mathcal{L}\psi_{i,q}^{loc}\|_{L^2(\tau_j)}$ and Theorem 3.2.1 to $\|\psi_{i,q}\|_{L^2(\tau_j)}$, we obtain

$$\begin{aligned} I_2 &\leq 2\sqrt{2\theta_{k,max}}C(k,d,\delta)C_p|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}^{loc}\|_{H(S^*)} \\ &\leq 4\sqrt{\frac{\theta_{k,max}}{\theta_{k,min}}}C(k,d,\delta)C_p\|\psi_{i,q}\|_{H(S^*)}\|\psi_{i,q}^{loc}\|_{H(S^*)}, \end{aligned} \tag{3.119}$$

where we have used $\theta_{k,max} := \max(\theta_{0,max}, \theta_{k,max})$, the strong ellipticity (3.75) and the bound (3.53) in the last step. Applying the exponential decay of Theorem 3.5.2 to both $\|\psi_{i,q}\|_{H(S^*)}$ and $\|\psi_{i,q}^{loc}\|_{H(S^*)}$, we obtain

$$I_2 \leq 2\sqrt{\frac{\theta_{k,max}}{\theta_{k,min}}}C(k,d,\delta)C_p e^{1-\frac{r-2h}{lh}}\|\psi_{i,q}\|_{H(D)}\|\psi_{i,q}^{loc}\|_{H(D)}. \tag{3.120}$$

Combining Eqn. (3.113), (3.116) and (3.120), and using Eqn. (3.105) to bound $\|\psi_{i,q}^{loc}\|_{H(D)}$ and $\|\psi_{i,q}\|_{H(D)}$ (recall $\|\psi_{i,q}\|_{H(D)} \leq \|\psi_{i,q}^{loc}\|_{H(D)}$), we complete the proof of Eqn. (3.108).

When the operator $\mathcal{L}$ contains only the highest order terms, i.e. $\mathcal{L}u = (-1)^k \sum_{|\sigma|=|\gamma|=k} D^\sigma(a_{\sigma\gamma} D^\gamma u)$, Eqn. (3.116) and (3.120) hold true for all $h > 0$. In this case, we can get rid of the factor "2" in both Eqn. (3.116) and (3.120). Therefore, we obtain the estimate on C_3 stated in the theorem. □

Theorem 3.6.2. *Let $u \in H_0^k(D)$ be the weak solution of $\mathcal{L}u = f$ and $\psi_{i,q}^{loc}$ be the localized basis functions defined in Eqn.* (3.9)*. Then for $r \geq (d + 4k)lh\log(1/h) + 2(1 + l\log C_4)h$, we have*

$$\inf_{v\in\Psi^{loc}} \|u - v\|_{H(D)} \leq \frac{2C_p}{\sqrt{a_{min}}} h^k \|f\|_{L^2(D)}, \tag{3.121}$$

where $C_4 = \frac{C_3 C_e}{C_p}(Qa_{min})^{1/2}$, and C_3 is defined in Theorem 3.6.1, a_{min} comes from the norm-equivalence (3.5)*, and C_e is the constant such that $\|u\|_{L^2(D)} \leq C_e\|f\|_{L^2(D)}$ holds true.*

Proof. Let $v_1 := \sum_{i=1}^m \sum_{q=1}^Q c_{iq}\psi_{i,q}$ and $v_2 := \sum_{i=1}^m \sum_{q=1}^Q c_{iq}\psi_{i,q}^{loc}$ with $c_{iq} = \int_D u\varphi_{i,q}$. Estimation (3.23) gives that

$$\|u - v_1\|_H \leq \frac{C_p h^k}{\sqrt{a_{min}}}\|f\|_{L^2(D)}. \tag{3.122}$$

Using the Cauchy inequality, we have

$$\|v_1 - v_2\|_H \leq \max_{i,q} \|\psi_{i,q} - \psi_{i,q}^{loc}\|_H \sum_{i=1}^m \sum_{q=1}^Q |c_{iq}| \leq \max_{i,q} \|\psi_{i,q} - \psi_{i,q}^{loc}\|_H \sum_{i=1}^m Q^{1/2} (\sum_{q=1}^Q |c_{iq}|^2)^{1/2}.$$

Thanks to the orthogonality of $\{\varphi_{i,q}\}_{q=1}^Q$ (3.7), we have $|\tau_i|^{-1/2}(\sum_{q=1}^Q |c_{iq}|^2)^{1/2} \leq \|u\|_{L^2(\tau_i)}$. Then we obtain

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

Using the energy estimation $\|u\|_{L^2(D)} \leq C_e\|f\|_{L^2(D)}$ and Theorem 3.6.1, we obtain

$$\|v_1 - v_2\|_H \leq C_3 C_e Q^{1/2} h^{-\frac{d}{2}-k} \exp(-\frac{r - 2h}{2lh})\|f\|_{L^2(D)}. \tag{3.123}$$

Combining Eqn. (3.122) and (3.123) together, we conclude the proof. □

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

Corollary 3.6.3. *Let $\psi_{i,q}^{loc}$ be the localized basis functions defined in Eqn.* (3.9)*. Then for $r \geq (d+4k)lh\log(1/h) + 2(1+l\log C_4)h$, we have*

$$\|\mathcal{K} - \mathcal{P}_{\Psi^{loc}}^{(H)}\mathcal{K}\| \leq \frac{4C_p^2}{a_{min}}h^{2k}, \tag{3.124}$$

where all the constants are the same as those defined in Theorem 3.6.2.

Corollary 3.6.3 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 3.6.1. *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 3.6.1. 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 3.6.1 still works for other homogeneous boundary conditions.*

3.7 Numerical Examples

In this section, we present two numerical results to support the theoretical findings, and to show how the sparse operator compression is utilized in higher order elliptic operators. In Section 3.7, 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 Section 3.7, 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.

The 1D Fourth Order Elliptic Operator

Consider the solution operator of the Euler-Bernoulli equation

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(a(x)\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\right) &= f(x), \qquad 0<x<1, \\ u(0) = u'(0) = 0, &\quad u(1) = u'(1) = 0, \end{aligned} \tag{3.125}$$

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

$$a(x) := 1 + \frac{1}{2}\sin\left(\sum_{k=1}^{K} k^{-\alpha}(\zeta_{1k}\sin(kx) + \zeta_{2k}\cos(kx))\right), \tag{3.126}$$

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 [66, 94, 105], and has no scale separation. We choose $\alpha = 0$ and $K = 40$ in the numerical experiment. A sample coefficient is shown in Figure 3.2.

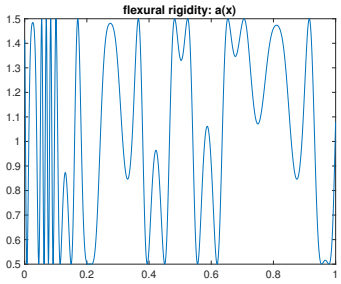
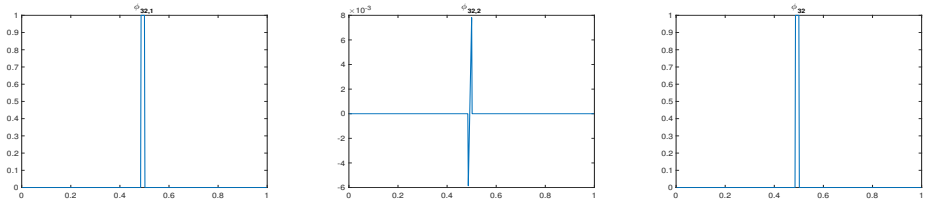


Figure 3.2: Highly oscillatory flexural rigidity without scale separation.

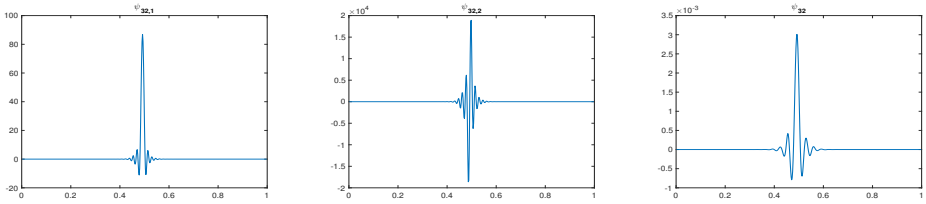
We partition the physical space $[0,1]$ uniformly into $m = 2^6$ patches, where the ith patch $I_i = [(i-1)h, ih]$ with $h = 1/m$. In this fourth-order case, our theory requires the piecewise polynomial space Φ be the space of (discontinuous) piecewise linear functions, which has dimension $n = 2m$. 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 (3.8), 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 Figure 3.3 A. In Figure 3.3(B-C), we plot the basis functions $\psi_{32,1}$ and $\psi_{32,2}$, which clearly show exponential decay.

To demonstrate the necessity for Ψ to contain all piecewise linear functions, in the third column of Figure 3.3, 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 Figure 3.3(A) and (B), we plot φ_{32} and ψ_{32}. Solving the quadratic optimization problem (3.8), we obtain only one basis function ψ, denoted as

[$\varphi_{32,1}, \varphi_{32,2}$ for piecewise linear Φ and φ_{32} for piecewise constant Φ]



[$\psi_{32,1}, \psi_{32,2}$ for piecewise linear Φ and ψ_{32} for piecewise constant Φ]



[In log-scale: $\psi_{32,1}, \psi_{32,2}$ for piecewise linear Φ and ψ_{32} for piecewise constant Φ]

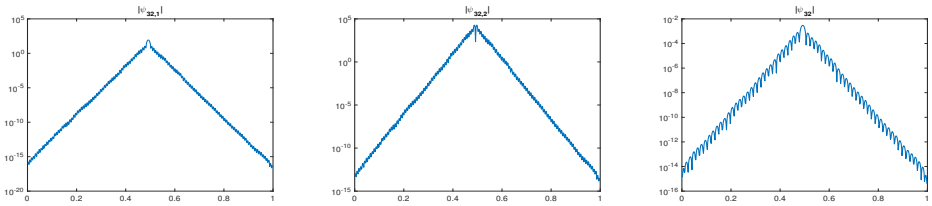


Figure 3.3: One dimensional fourth order elliptic operator (3.125).

ψ_i, associated with the patch I_i. In Figure 3.3(C), 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 (3.126) 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 (3.125), 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 Figure 3.4. 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 2.2.1 and Eqn. (3.23).

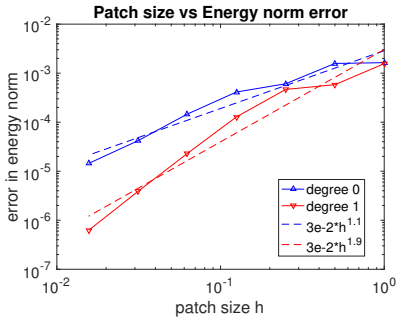


Figure 3.4: Error of the finite element solutions: $\|u_{h,0} - u\|_H$ and $\|u_{h,1} - u\|_H$.

The 2D Fourth Order Elliptic Operator

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

$$\partial_x^2(a_{20}(x,y)\partial_x^2 u(x,y)) + \partial_y^2(a_{02}(x,y)\partial_y^2 u(x,y)) + 2\partial_{xy}(a_{11}(x,y)\partial_{xy}u(x,y)) = f(x,y),$$
$$u \in H_0^2(D), \tag{3.127}$$

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

$$\begin{aligned} a_{20}(x,y) = a_{02}(x,y) =& \frac{1}{6}\Big(\frac{1.1+\sin(2\pi x/\epsilon_1)}{1.1+\sin(2\pi y/\epsilon_1)} + \frac{1.1+\sin(2\pi y/\epsilon_2)}{1.1+\cos(2\pi x/\epsilon_2)} + \\ & \frac{1.1+\cos(2\pi x/\epsilon_3)}{1.1+\sin(2\pi y/\epsilon_3)} + \frac{1.1+\sin(2\pi y/\epsilon_4)}{1.1+\cos(2\pi x/\epsilon_4)} + \sin(4x^2y^2) + 1\Big), \\ a_{11}(x,y) =& 1 + \frac{1}{2}\sin\left(\sum_{k=1}^{K} k^{-\alpha}(\zeta_{1k}\sin(kx) + \zeta_{2k}\cos(ky))\right), \end{aligned} \tag{3.128}$$

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 (3.8) with the weighted extended B-splines (Web-splines [62]) 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 Figure 3.5. We also show them in the log-scale in Figure 3.6. We can clearly see that the basis functions decay exponentially fast away from its associated patch, which validates our Theorem 3.5.1.

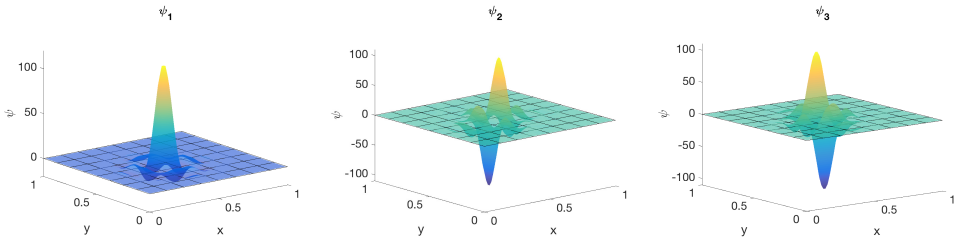


Figure 3.5: The three basis functions associated with patch $[1/2 - h_x, 1/2] \times [1/2 - h_y, 1/2]$.

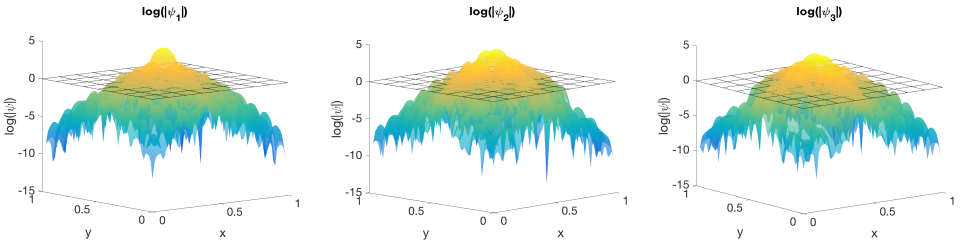


Figure 3.6: The three basis functions associated with patch $[1/2 - h_x, 1/2] \times [1/2 - h_y, 1/2]$ in log-scale.

We point out that the stiffness matrix for the fourth order elliptic operator (3.127) gets ill-conditioned very quickly when we refine the grid size. Specifically, for fine grid size $h_{x,f} = h_{y,f} = \frac{1}{64}$, its condition number is at the order of 10^9, and the exponential decay is heavily polluted by the numerical error. High-precision computing is required here to further refine the domain partition and to validate the optimal convergence rate. We will leave this as our future work.

Chapter 4

SPARSE OPERATOR COMPRESSION OF ELLIPTIC OPERATORS WITH HIGH CONTRAST COEFFICIENTS

In Chapter 2, we introduced the sparse operator compression to compress a self-adjoint positive semi-definite operator $\mathcal{K} : L^2(D) \to L^2(D)$ by localized basis functions, where D is a bounded domain in $\mathbb{R}^d$. In this chapter, we apply the Sparse OC framework to construct localized multiscale finite basis functions for the elliptic equations with high contrast coefficients.

4.1 Problem Setting

Consider the second-order elliptic equation with the homogeneous Dirichlet boundary condition

$$\mathcal{L}u := -\nabla \cdot (a\nabla u) = f, \qquad u \in H_0^1(D), \tag{4.1}$$

where the load $f \in L^2(D)$. We consider the scalar rough coefficient $a \in L^\infty(D)$ that is positive and uniformly bounded below and above, i.e.,

$$a_{min} \leq a(x) \leq a_{max} \qquad \forall x \in D, \tag{4.2}$$

but the contrast $\frac{a_{max}}{a_{min}}$ can be arbitrarily large!

Key Ideas

To design a multiscale finite element method (MsFEM) whose convergence rate is independent of the rough coefficient and its contrast is important for many practical applications. For example, in porous media applications, the permeability of subsurface regions often has multiscale features and contrast. Existing methods for second order elliptic operators scale badly with the contrast of the coefficients. More precisely, in the existing methods [56, 89, 98, 99] and in the application of Sparse OC to higher order elliptic operators in Chapter 3, the coefficient contrast $\frac{a_{max}}{a_{min}}$ enters the proof of the exponential decay via the following norm equivalence and Cauchy-Schwartz inequality

$$\begin{aligned} a_{min}\int_D u^2 \leq \int_D au^2 \leq a_{max}\int_D u^2 \quad \forall u \in L^2(D), \\ \int_D auv \leq a_{max}\|u\|_{L^2(D)}\|v\|_{L^2(D)} \quad \forall u, v \in L^2(D). \end{aligned} \tag{4.3}$$

These inequalities lead to polynomial dependence between the decay rate and the contrast.

In this chapter, we avoid using these contrast-dependent inequalities by introducing the a-weighted L^2 and H^1 spaces

$$L^2_a(D) = \left\{ f : \int_D af^2 < \infty \right\}, \quad H^1_{0,a} = \left\{ u \in H^1_0(D) : \int_D a|\nabla u|^2 < \infty \right\}, \tag{4.4}$$

and their associated norms/inner products

$$\|f\|_{L^2_a(D)} := \|f\|_{0,2,D,a} = \left(\int_D af^2 \right)^{1/2}, \quad H^1_{0,a} := \|u\|_{1,2,D,a} = \left(\int_D a|\nabla u|^2 \right)^{1/2}. \tag{4.5}$$

Then we can avoid the contrast by using the Cauchy-Schwartz inequality in these a-weighted spaces, e.g.,

$$\int_D auv \leq \|u\|_{L^2_a(D)} \|v\|_{L^2_a(D)} \quad \forall u, v \in L^2_a(D).$$

By selecting the local measurement functions from the local eigenvalue problem (see Eqn. (4.6)), we naturally obtain a local projection-type approximation property. By picking possibly more than one measurement function per patch, the local projection-type approximation property has an approximation rate independent of the contrast. To prove that the decay rate of the basis functions is independent of the contrast, we still need an inverse energy estimate independent of the contrast. We have proved this contrast-independent inverse energy estimate for the two-phase coefficient model (first proposed in [46]) using asymptotic analysis. Moreover, we provide an efficient way to obtain the decay rate of the constructed basis functions, and we can localize the basis functions based on this known decay rate.

Our Construction

To construct such localized basis functions $\Psi^{loc} = [\psi_1^{loc}, \ldots, \psi_n^{loc}]$, we first partition the physical domain D using a regular partition $\{\tau_i\}_{i=1}^m$ with mesh size h. On every patch τ_i, we solve the following eigenvalue problem:

$$\begin{aligned} -\nabla \cdot (a \nabla \varphi_{i,q}) &= \lambda_{i,q} a(x) \varphi_{i,q} \\ \boldsymbol{n} \cdot \nabla \varphi_{i,q} &= 0 \quad \text{on } \partial\tau_i, \end{aligned} \tag{4.6}$$

where $\boldsymbol{n}$ is the normal vector on the boundary of τ_i. Assume that the eigenvalues are ordered as

$$0 = \lambda_{i,1} \leq \lambda_{i,2} \leq \cdots \leq \lambda_{i,q} \leq \ldots. \tag{4.7}$$

Let Q_i be the smallest number such that $1/\lambda_{i,Q_i+1} \leq C_P^2 h^2$, i.e.,

$$Q_i = \min\{q : 1/\lambda_{i,q+1} \leq C_P^2 h^2\}, \tag{4.8}$$

where C_P is a user-specific constant. For example, one can take $C_P = 1/\pi$, which is the Poincare constant for bounded convex domains.

Let $\Phi_i \equiv \text{span}\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$, and we choose $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$ is an orthonormal basis of Φ_i in $L_a^2(\tau_i)$. Then we apply the framework of the Sparse OC with $X = L_a^2(D)$, $H = H_{0,a}^1(D)$ and local measurement functions $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$. Specifically, for $r > 0$, let S_r be the union of the subdomains τ_j intersecting $B(x_i, r)$ (for some $x_i \in \tau_i$) and let $\psi_{i,q}^{loc}$ be the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_{i,q}^{loc} = \underset{\psi \in H_0^1(D)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int_{S_r} a\psi\varphi_{j,q'} = \delta_{iq,jq'}, \quad \forall 1 \leq j \leq m,\, 1 \leq q' \leq Q_j, \\ & \psi(x) \equiv 0, \quad x \in D \backslash S_r. \end{aligned} \tag{4.9}$$

Collecting all the $\psi_{i,q}^{loc}$ for $1 \leq i \leq m$ and $1 \leq q \leq Q_i$ together, we get our basis Ψ^{loc}.

In this chapter, we prove that for $r = (C_{r,1} d \log(1/h) + C_{r,2})h$, the localized basis functions Ψ^{loc} achieve the linear convergence rate to solve the elliptic equation, i.e.,

$$\left\| u - \Psi^{loc} L_n^{-1} (\Psi^{loc})^T f \right\|_H \leq 2C_P h \left(\int_D a^{-1} f^2 \right)^{1/2} \leq \frac{2C_P h}{\sqrt{a_{min}}} \|f\|_{L^2(D)} \quad \forall f \in L^2(D), \tag{4.10}$$

where u is the solution of the elliptic equation (4.1), L_n is the stiffness matrix associated with Ψ^{loc}, the constant $C_{r,1}$ is independent of n, and the constant $C_{r,2}$ depends on the contrast at most logarithmically, i.e., $C_{r,2} = \mathcal{O}(\log(a_{max}/a_{min}))$.

For the two-phase coefficient model where its inclusions have smooth boundaries, our asymptotic analysis shows that (1) Q_i can be taken as the number of high-conductivity inclusions in the local patch τ_i, and that (2) the constant $C_{r,1}$ is independent of the contrast of the coefficient. In this case, the radius $r = (C_{r,1} d \log(1/h) + C_{r,2})h$ only depends on the contrast logarithmically, which significantly improves the polynomial dependence in existing methods [56, 89,

98, 99]. Although we have not proved the independence between the constant $C_{r,1}$ and the coefficient contrast for any coefficients with high contrast and multiscale features, we provide an efficient way to compute the constant $C_{r,1}$. Therefore, one can still localize the basis functions on a local patch with radius $r = (C_{r,1}d\log(1/h) + C_{r,2})h$.

Comparison With Other Existing Methods

Our method is inspired by the work in [47, 46], where the authors proposed a domain decomposition preconditioner and the resulting preconditioned conjugate gradient method converges independent of the contrast. Our work differs from their work in two aspects. First of all, in [47, 46], the local eigenfunctions $\varphi_{i,q}$ are directly used as multiscale basis functions after multiplied by a set of partition of unity functions, while in our method $\varphi_{i,q}$ are local measurement functions in our method and the energy-minimizing basis functions $\psi_{i,q}$ are our multiscale basis functions. Secondly, in [47, 46], the multiscale basis functions are used to construct a domain decomposition preconditioner, and they show that the preconditioned system has a condition number independent of the contrast. In our method, the multiscale basis functions are directly used in the Galerkin projection framework to solve the linear system, and we prove that the solution error (in the energy norm) is independent of the contrast.

Recently, improved numerical methods based on the local orthogonal decomposition (LOD [89], see section 2.3 for a brief review) have appeared to tackle the high contrast coefficients, such as [110, 60]. Both methods use the LOD framework and propose new Clément-type quasi-interpolation operators to tackle the high contrast problem. Let $\mathcal{T}_h$ denote a regular triangulation of D into closed simplices, $\mathcal{N}_h = \{z_i\}_{i=1}^m$ denote the set of all interior mesh nodes in $\mathcal{T}_h$ and $V_h \subset H_0^1(D)$ the corresponding piecewise linear finite element space. The method in [110] also makes use of the a-weighted $L^2(D)$ space, and their localized basis function ψ_i^{loc} is the unique solution of the following local energy-minimizing problem:

$$\begin{aligned} \psi_i^{loc} = \underset{\psi \in H_0^1(D)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int a\lambda_j\psi = \int a\lambda_j\lambda_i, \quad \forall 1 \le j \le m, \\ & \psi(x) \equiv 0, \quad x \in D\backslash S_r, \end{aligned} \tag{4.11}$$

where S_r is a neighborhood of the mesh node z_i and λ_i is the nodal piecewise

linear element centered at node z_j. In [110], the authors have proved that ψ_i in Eqn. (4.11) has exponential decay rate independent of the contrast for locally quasi-monotone coefficients (which essentially requires at most one connected high-conductivity inclusion in a local patch). Despite its similarity with our construction (4.9), especially the a-weighted inner product, we emphasize that they only have one local measurement function (i.e. λ_i) per node, while we may have multiple local measure functions $\{\varphi_{i,q} : 1 \leq q \leq Q_i\}$. This choice of local measure functions allows us to work for local patches with multiple high-conductivity inclusions while the modified LOD method in [110] cannot. We point out although locally quasi-monotone coefficients cannot contain two high-conductivity inclusions, they can take multiple values. From this aspect, the proof in [110] applies to some cases that are not covered by our results.

In [60], the authors introduce a Clément-type quasi-interpolation operator (based on Scott-Zhang node variables) for two-phase coefficients, and their method does not need the a-weighted $L^2(D)$ space. The proposed quasi-interpolation operator forces the basis function decay within channels, and they prove that the localization error for this operator is independent of the contrast. The basic idea is to select the local measurement function for each node in such a way that the operator admits a contrast independent Poincare-type inequality in every local patch. In practice, this requires that each connected high-conductivity channel and inclusion include a mesh node in $\mathcal{N}_h = \{z_i\}_{i=1}^m$, which is not true when the coefficient has many fine scale high-conductivity inclusions but the mesh $\mathcal{T}_h$ is not fine enough to resolve them. This limitation of the method in [60] originates from its construction where there is only one measurement function per node. It is interesting to see whether they can deal with the general case as we do here by allowing more local measurement functions. Another limitation is that the construction of their quasi-interpolation operator makes use of the two-phase coefficient model in a crucial way. It is not clear whether their construction can be extended to work for general coefficients.

The flux norm approach [13] is another way to avoid the norm equivalence and Cauchy-Schwartz inequality (4.3), which leads to a contrast-independent error estimate. The flux norm is also used in [10] to achieve a contrast-robust $\mathcal{H}$-matrix approximation of the solution operator of second-order elliptic operators.

Outline Of This Chapter

The rest of the chapter is organized as follows. In Section 4.2, we obtain a local projection-type approximation property from the local generalized eigenvalue problem. We also provide a local inverse energy estimate that depends on a computable constant. In Section 4.3, we prove that the global energy minimizing basis functions $\psi_{i,q}$ decay exponentially fast away from its associated patch, and the decay rate only depends on the constant in the inverse energy estimate. In Section 4.4, we prove that localized basis functions $\psi_{i,q}^{loc}$ approximate $\psi_{i,q}$ accurately and preserve the $\mathcal{O}(h)$ convergence rate in the energy norm. In Section 4.5, for the two-phase coefficients with smooth inclusions/channels, we show the constant in the local inverse energy estimate is independent of the contrast using an asymptotic expansion. In Section 4.6, a 2D example with high permeability channels is provided to demonstrate the contrast-independent decay rate.

4.2 The Projection-type Approximation Property And Inverse Energy Estimate

We consider the following equation:

$$\mathcal{L}u := -\nabla \cdot (a\nabla u) = af, \qquad u \in H^1_{0,a}(D), \tag{4.12}$$

where the load $f \in L^2_a(D)$. The model equation (4.1) is just the above equation with a weighted load $a^{-1}f$. $u \in H^1_{0,a}(D)$ is a weak solution of Eqn. (4.12) if and only if

$$\int_D a\nabla u \cdot \nabla v = \int_D afv \qquad \forall v \in H^1_{0,a}(D). \tag{4.13}$$

We define

$$C_E^2 := \sup_{\substack{u \in H^1_{0,a}(D) \\ u \neq 0}} \frac{\int_D a|u|^2}{\int_D a|\nabla u|^2}, \tag{4.14}$$

which is similar to the Friedrich's constant in the Friedrich's inequality $\int_D |u|^2 \leq C_F^2 \int_D |\nabla u|^2$ for any $u \in H^1_0(D)$. A simple bound for C_E is

$$C_E \leq C_F \left(\frac{a_{max}}{a_{min}}\right)^{1/2}, \tag{4.15}$$

which depends on the contrast a_{max}/a_{min}. We note that Eqn. (4.15) is very crude in practice. Thanks to the Riesz lemma, we can prove that Eqn. (4.12) has a unique solution u, and u satisfies the following energy estimate:

$$\left(\int_D a|\nabla u|^2\right)^{1/2} \leq C_E \left(\int_D af^2\right)^{1/2}. \tag{4.16}$$

We define $\mathcal{K} : L^2_a(D) \to L^2_a(D)$ as

$$\mathcal{K}f = u, \tag{4.17}$$

which is exactly the solution operator for Eqn. (4.12). Intuitively, we can think that $\mathcal{K}f = \mathcal{L}^{-1}(af)$. It is easy to check that $\mathcal{K}$ is self-adjoint and positive definite, and thus we can apply the general framework introduced in Chapter 2 with $X = L^2_a(D)$ and $H = H^1_{0,a}(D)$.

A Projection-type Approximation Property

Recall that $\{(\lambda_{i,q}, \varphi_{i,q})\}_{q=1}^{\infty}$ are the eigen-pairs of the eigenvalue problem $\mathcal{L}u = a\lambda u$ with the homogeneous Neumann boundary conditions. From the completeness of the eigenfunctions, we know that $\{\varphi_{i,q}\}_{q=1}^{\infty}$ is a complete orthonormal basis of $L^2_a(\tau_i)$, and they are orthogonal in $H^1_{0,a}(\tau_i)$. Therefore, we have the following lemma.

Lemma 4.2.1. *Let $\Phi_i = span\{\varphi_{i,q} : 1 \le q \le Q_i\}$. Then for any $Q_i \ge 1$, we have*

$$\|u - \mathcal{P}^{(L^2_a)}_{\Phi_i} u\|_{L^2_a} \le \frac{1}{\sqrt{\lambda_{i,Q_i+1}}} \left(\int_{\tau_i} a|\nabla u|^2 \right)^{1/2}, \quad \forall u \in H^1_a(\tau_i). \tag{4.18}$$

Proof. For any $u \in H^1_a(\tau_i)$, we decompose it as $u = u_\varphi + u_{\varphi^\perp}$, where $u_\varphi \in \text{span}\{\varphi_{i,q} : q \ge 1\}$ and $u_{\varphi^\perp} \perp \text{span}\{\varphi_{i,q} : q \ge 1\}$ in $H^1_a(\tau_i)$, which has an inner product as $\langle u, v\rangle = \int_D a(uv + \nabla u \cdot \nabla v)$. Thanks to the zero Neumann BC of $\varphi_{i,q}$, we have

$$\int_D a\nabla u_{\varphi^\perp} \cdot \nabla \varphi_{i,q} = \int_D u_{\varphi^\perp} \mathcal{L}\varphi_{i,q} = \lambda_{i,q} \int_D a u_{\varphi^\perp} \varphi_{i,q}, \quad \forall q \ge 1.$$

Then, we have

$$0 = \langle u_{\varphi^\perp}, \varphi_{i,q}\rangle = (1 + \lambda_{i,q}) \int_D a u_{\varphi^\perp} \varphi_{i,q}, \quad \forall q \ge 1.$$

Since $1 + \lambda_{i,q} > 0$ for any q and $\{\varphi_{i,q} : q \ge 1\}$ is an orthogonal basis in $L^2_a(\tau_i)$, we conclude $u_{\varphi^\perp} = 0$. Therefore, we write

$$u = u_\varphi = \sum_{q=1}^{\infty} c_q \varphi_{i,q},$$

and then we have

$$\|u - \mathcal{P}^{(L^2_a)}_{\Phi_i} u\|^2_{L^2_a} = \sum_{q=Q_i+1}^{\infty} c_q^2, \qquad \int_{\tau_i} a|\nabla u|^2 = \sum_{q=2}^{\infty} c_q^2 \lambda_{i,q}.$$

Therefore, for $Q_i \ge 1$, Eqn. (4.18) naturally follows. □

In the following, we will pick Q_i as in Eqn. (4.8), i.e., $Q_i = \min\{q : 1/\lambda_{i,q+1} \leq C_P^2 h^2\}$. We extend $\varphi_{i,q}$ to the whole physical domain D by setting $\varphi_{i,q} = 0$ outside τ_i, and define Φ as the space spanned by all these functions, i.e.,

$$\Phi := \text{span}\{\varphi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}. \tag{4.19}$$

Then we have the following projection type approximation property, which is useful in the error estimate, see Eqn. (2.11) in Theorem 2.2.1.

Lemma 4.2.2. *Suppose that* $Q_i = \min\{q : 1/\lambda_{i,q+1} \leq C_P^2 h^2\}$ *and that* $\Phi := \text{span}\{\varphi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}$. *We have*

$$\|u - \mathcal{P}_\Phi^{(L_a^2)} u\|_{L_a^2} \leq C_P h \|u\|_{H_{0,a}^1}, \quad \forall u \in H_{0,a}^1(D). \tag{4.20}$$

Proof. By the construction of Φ, $u - \mathcal{P}_\Phi^{(L_a^2)} u = u - \mathcal{P}_{\Phi_i}^{(L_a^2)} u$ on patch τ_i. Combining Lemma 4.2.1 with the choice of Q_i, we have

$$\|u - \mathcal{P}_\Phi^{(L_a^2)} u\|_{L_a^2(\tau_i)} \leq C_P h \left(\int_{\tau_i} a |\nabla u|^2 \right)^{1/2}. \tag{4.21}$$

Therefore, we have

$$\begin{aligned} \|u - \mathcal{P}_\Phi^{(L_a^2)} u\|_{L_a^2(D)} &= \left(\sum_{i=1}^m \|u - \mathcal{P}_\Phi^{(L_a^2)} u\|_{L_a^2(\tau_i)}^2 \right)^{1/2} \\ &\leq \left(\sum_{i=1}^m C_P^2 h^2 \int_{\tau_i} a |\nabla u|^2 \right)^{1/2} = C_P h \|u\|_{H_{0,a}^1(D)}. \end{aligned}$$

□

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

$$\Psi = \{\mathcal{K}\varphi : \varphi \in \Phi\} = \text{span}\{\mathcal{K}\varphi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}. \tag{4.22}$$

Thanks to Theorem 2.2.1, we have

$$\|u - \mathcal{P}_\Psi^{(H_{0,a}^1)} u\|_{H_{0,a}^1} \leq C_P h \|f\|_{L_a^2}, \tag{4.23}$$

where u is the unique weak solution of Eqn. (4.12). Eqn. (4.23) says that MsFEM with basis $\{\mathcal{K}\varphi_{i,q} : 1 \leq i \leq m, 1 \leq q \leq Q_i\}$ can achieve the linear convergence rate when solving the elliptic equation (4.12). However, $\mathcal{K}\varphi_{i,q}$ are

typically global, and have the same computational cost as solving the original equation. Moreover, the resulting stiffness is dense, which further increases the complexity to solve the reduced linear system.

Following the our framework, we can define the energy minimizing basis

$$\begin{aligned} \psi_{i,q} = \underset{\psi \in H^1_{0,a}(D)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int_D a\psi_{i,q}\varphi_{j,q'} = \delta_{iq,jq'}, \quad \forall 1 \le q' \le Q_j, 1 \le j \le m. \end{aligned} \tag{4.24}$$

Thanks to Theorem (2.2.3) and the positive definiteness of $\mathcal{K}$, we know that $\Psi = \{\psi_{i,q} : 1 \le i \le m, 1 \le q \le Q_i\}$ is another basis of Ψ. Using Eqn. (2.18), we conclude that

$$\mathcal{L}\psi \mid_{\tau_j} \in a\Phi_j := \{a\varphi : \varphi \in \Phi_j\}, \qquad \forall \psi \in \Psi, 1 \le j \le m. \tag{4.25}$$

Combining Eqn. (4.25) and the orthogonality constraints in Eqn. (4.24), we obtain that

$$\int_{\tau_j} \psi_{i,q}\mathcal{L}\psi = 0, \qquad \forall j \ne i, \psi \in \Psi. \tag{4.26}$$

The above orthogonality will be very useful in the rest of the paper.

An Inverse Energy Estimate In Ψ

In this section, we propose an inverse energy bound for functions in Ψ, see Lemma 4.2.3, which is crucial in proving the exponential decay and the localization of $\psi_{i,q}$. Lemma 3.12 in [99] provides such an estimate for second order uniformly elliptic operators, by utilizing a relation between the Laplacian operator Δ and the d-dimensional Brownian motion. In [68], we proved an inverse energy estimate that is valid for any $2k$'th order elliptic operators, but the estimation depends on the contrast of the coefficients. In this section, we postulate the inverse energy estimate, and use numerical computations to verify the bound.

Let $\mathcal{L}_0^{-1}(af) \in H^1_{0,a}(\tau_i)$ be the unique weak solution of the following elliptic equation with the homogeneous Dirichlet boundary condition

$$\begin{aligned} \mathcal{L}u = -\nabla \cdot (a\nabla u) = af(x) \qquad & x \in \tau_i, \\ u = 0 \qquad & x \in \partial\tau_i. \end{aligned} \tag{4.27}$$

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

$$M_0(q,q') = \int_{\tau_i} a\varphi_{i,q}\varphi_{i,q'}, \qquad A_0(q,q') = \int_{\tau_i} a\varphi_{i,q}\mathcal{L}_0^{-1}(a\varphi_{i,q'}), \tag{4.28}$$

and denote $\lambda_{max}(M_0, A_0)$ be the maximal generalized eigenvalue of the eigenvalue problem $M_0\alpha = \lambda A_0\alpha$. We define

$$C_i := \sup_{\varphi\in\Phi_i\setminus\{0\}} \frac{\|\varphi\|_{L^2_a(\tau_i)}}{\sqrt{\lambda_{i,Q_i+1}}\|\mathcal{L}_0^{-1}(a\varphi)\|_{H^1_{0,a}(\tau_i)}} = \sqrt{\frac{\lambda_{max}(M_0, A_0)}{\lambda_{i,Q_i+1}}}, \qquad C_{inv} := \max_{1\le i\le m} C_i. \tag{4.29}$$

Lemma 4.2.3. *For every $\psi \in \Psi$, we have*

$$\|\mathcal{L}\psi\|_{L^2_{1/a}(\tau_i)} \le \sqrt{\lambda_{max}(M_0, A_0)}\|\psi\|_{H^1_{0,a}(\tau_i)} = C_i\sqrt{\lambda_{i,Q_i+1}}\|\psi\|_{H^1_{0,a}(\tau_i)}. \tag{4.30}$$

Proof. For every $\psi \in \Psi$, we have $\mathcal{L}\psi = a\varphi$ for some $\varphi \in \Phi_i$ when restricted to patch τ_i. Consider the decomposition $\psi = \psi_1 + \psi_2$ where $\mathcal{L}\psi_1 = a\varphi$, $\psi_1 = 0$ on $\partial\tau_i$, and $\mathcal{L}\psi_2 = 0$, $\psi_2 = \psi$ on $\partial\tau_i$. It is easy to check that $\int_{\tau_i} a\nabla\psi_1\cdot\nabla\psi_2 = 0$, and thus $\|\psi\|^2_{H^1_{0,a}(\tau_i)} = \|\psi_1\|^2_{H^1_{0,a}(\tau_i)} + \|\psi_2\|^2_{H^1_{0,a}(\tau_i)}$, which implies that $\|\psi_1\|_{H^1_{0,a}(\tau_i)} \le \|\psi\|_{H^1_{0,a}(\tau_i)}$.

Notice that $\psi_1 = \mathcal{L}_0^{-1}(a\varphi)$. From the definition of M_0 and A_0, we have $\|\varphi\|^2_{L^2_a(\tau_i)} \le \lambda_{max}(M_0, A_0)\|\psi_1\|^2_{H^1_{0,a}(\tau_i)}$. Therefore, we have $\|\mathcal{L}\psi\|_{L^2_{1/a}(\tau_i)} = \|\varphi\|_{L^2_a(\tau_i)} \le \sqrt{\lambda_{max}(M_0, A_0)}\|\psi_1\|_{H^1_{0,a}(\tau_i)}$. We conclude the proof by using that $\|\psi_1\|_{H^1_{0,a}(\tau_i)} \le \|\psi\|_{H^1_{0,a}(\tau_i)}$, and thus we conclude the proof. □

In Theorem 4.3.1, we prove that the exponential decay rate of $\psi_{i,q}$ is bounded by $(\mathrm{e}-1)(C_P + C_{inv})$. Therefore, if C_{inv} (i.e., every C_i for $1 \le i \le m$) can be bounded independent of the contrast of the coefficients, we obtain a decay rate independent of the contrast. In section 4.5 we show that C_i can be bounded by a constant independent of the contrast for the two-phase coefficient model, and in section 4.6 our numerical examples confirm this result. For the general L^∞ coefficients with multiscale and high-contrast features, one only needs to solve Q_i local linear systems, i.e. Eqn. (4.27) with $\{\varphi_{i,q}\}_{i=1}^{Q_i}$, to obtain the constant C_i. Compared with the computational cost to solve the local eigenvalue problem (4.6), this extra cost is affordable.

4.3 Exponential Decay of The Basis Functions

In this section, we will prove that the basis function $\psi_{i,q}$ decays exponentially fast away from its associated patch τ_i. The proof follows the same structure as that of Theorem 4.3.1 and [99] (Thm. 3.9). One important difference is that we make use of the Cauchy-Schwatz inequality in $L^2_a(D)$, i.e.

$\int_D auv \leq \|u\|_{L^2_a(D)}\|v\|_{L^2_a(D)}$ instead of $\int_D auv \leq a_{max}\|u\|_{L^2(D)}\|v\|_{L^2(D)}$, to obtain a contrast-independent decay rate.

To simplify our notations, for any $\psi \in H$ and any subdomain $S \subset D$, we define $\|\psi\|_{H(S)} := \left(\int_S a|\nabla\psi|^2\right)^{1/2}$.

Theorem 4.3.1. *For any $h > 0$, $1 \leq i \leq m$ and $1 \leq q \leq Q_i$, it holds true that*

$$\|\psi_{i,q}\|^2_{H(D\cap(B(x_i,r))^c)} \leq \exp(1 - \frac{r}{lh})\|\psi_{i,q}\|^2_{H(D)} \tag{4.31}$$

with $l = (\mathrm{e} - 1)(C_P + C_{inv})$.

Proof. Let $k \in \mathbb{N}$, $l > 0$ and $i \in \{1, 2, \ldots, m\}$. Define S_0 as the union of all the domains τ_j that are contained in the closure of $B(x_i, klh) \cap D$, S_1 as 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 $S^* = S_0^c \cap S_1^c \cap D$ (be the union of all the remaining elements τ_j not contained in S_0 or S_1).

Let $b_k := \|\psi_{i,q}\|^2_{H(S_0^c)}$, and by definition we have $b_0 = \|\psi_{i,q}\|^2_{H(D)}$, $b_{k+1} = \|\psi_{i,q}\|^2_{H(S_1)}$ and $b_k - b_{k+1} = \|\psi_{i,q}\|^2_{H(S^*)}$. 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 (4.31). 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., $\|\psi_{i,q}\|^2_{H(S_1)} \leq C\|\psi_{i,q}\|^2_{H(S^*)}$. Let η be the function on D defined by $\eta(x) = \mathrm{dist}(x, S_0)/\left(\mathrm{dist}(x, S_0) + \mathrm{dist}(x, S_1)\right)$. 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

$$\int_D \eta a|\nabla\psi_{i,q}|^2 = \underbrace{\int_D \eta\psi_{i,q}(-\nabla\cdot(a\nabla\psi_{i,q}))}_{I_2} \underbrace{- \int_D a\psi_{i,q}\nabla\eta\cdot\nabla\psi_{i,q}}_{I_1}. \tag{4.32}$$

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

$$I_1 \leq \|\nabla\eta\|_{L^\infty(D)}\|\psi_{i,q}\|_{L^2_a(S^*)}\|\psi_{i,q}\|_{H(S^*)} \leq \frac{1}{lh}\|\psi_{i,q}\|_{L^2_a(S^*)}\|\psi_{i,q}\|_{H(S^*)}. \tag{4.33}$$

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

Thanks to Eqn. (4.26), we have $\int_{S_1} \eta\psi_{i,q}(-\nabla\cdot(a\nabla\psi_{i,q})) = 0$. Denoting η_j as the volume average of η over τ_j, we have

$$\begin{aligned} I_2 &= -\int_{S^*} \eta\psi_{i,q}(-\nabla\cdot(a\nabla\psi_{i,q})) = -\sum_{\tau_j\in S^*}\int_{\tau_j}(\eta-\eta_j)\psi_{i,q}(-\nabla\cdot(a\nabla\psi_{i,q})) \\ &\leq \frac{1}{l}\sum_{\tau_j\in S^*}\|\psi_{i,q}\|_{L^2_a(\tau_j)}\|\mathcal{L}\psi_{i,q}\|_{L^2_{1/a}(\tau_j)}. \end{aligned} \tag{4.34}$$

Up to now, I_1 and I_2 are some quantities that depend $\psi_{i,q}$ only on S^*, and we only need to prove that both of them can be bounded by $\|\psi_{i,q}\|^2_{H(S^*)}$ (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 $\|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)}$ which in general may not be bounded by the lower order term $\|\psi_{i,q}\|_{H(S^*)}$. Fortunately, Lemma 4.2.3 shows that $\|\mathcal{L}\psi_{i,q}\|_{L^2_{1/a}(\tau_j)} \leq C_j\lambda^{1/2}_{j,Q_j+1}\|\psi_{i,q}\|_{H(\tau_j)}$. Then, we obtain

$$I_2 \leq \frac{1}{l}\sum_{\tau_j\in S^*} C_j\lambda^{1/2}_{j,Q_j+1}\|\psi_{i,q}\|_{L^2_a(\tau_j)}\|\psi_{i,q}\|_{H(\tau_j)}. \tag{4.35}$$

By the construction of $\psi_{i,q}$ (4.24), we have $\int_{\tau_j} a\psi_{i,q}\varphi_{j,q'} = 0$ for all $\tau_j \in S^*$ and $1 \leq q' \leq Q_j$. By the approximation property (4.18), we have $\|\psi_{i,q}\|_{L^2_a(\tau_j)} \leq \lambda^{-1/2}_{j,Q_j+1}\|\psi_{i,q}\|_{H(\tau_j)}$. Combined with the choice of Q_i (4.8), we obtain

$$\begin{aligned} I_1 &\leq \frac{C_P}{l}\|\psi_{i,q}\|^2_{H(S^*)}, \\ I_2 &\leq \frac{1}{l}\sum_{\tau_j\in S^*} C_j\|\psi_{i,q}\|^2_{H(\tau_j)} \leq \frac{C_{inv}}{l}\|\psi_{i,q}\|^2_{H(S^*)}. \end{aligned}$$

Finally, we have

$$\|\psi_{i,q}\|^2_{H(S_1)} \leq I_1 + I_2 \leq \frac{C_P + C_{inv}}{l}\|\psi_{i,q}\|^2_{H(S^*)}. \tag{4.36}$$

By taking $l \geq (\mathrm{e}-1)(C_P+C_{inv})$, we conclude that the constant $\frac{C_P+C_{inv}}{l} \leq \frac{1}{\mathrm{e}-1}$. Using the same iterative given above, we prove that the basis functions have exponential decay away from its associated patch, τ_i. □

4.4 Localization of The Basis Functions

Theorem 4.3.1 allows us to localize the construction of basis functions ψ_i as follows. For $r > 0$, let S_r be the union of the subdomains τ_j intersecting

$B(x_i, r)$ (recall that $B(x_i, \delta h_i/2) \subset \tau_i$) and let ψ_i^{loc} be the minimizer of the following quadratic optimization problem:

$$\begin{aligned} \psi_{i,q}^{loc} = \underset{\psi \in H^1_{0,a}(S_r)}{\arg\min} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int_D a\varphi_{j,q'}\psi = \delta_{ij,qq'}, \quad \forall 1 \le q' \le Q_j, 1 \le j \le m. \end{aligned} \tag{4.37}$$

We will naturally identify $\psi_{i,q}^{loc}$ with its extension to $H^1_{0,a}(D)$ by setting $\psi_i^{loc} = 0$ outside of S_r.

Thanks to the exponential decay of the energy minimizing basis functions $\psi_{i,q}$, choosing S_r with radius $r = \mathcal{O}(h \log(1/h))$ is sufficient to guarantee that the localized basis functions $\psi_{i,q}^{loc}$ have the same compression accuracy as the exponentially decaying basis functions. The following three theorems demonstrate such properties of the localized basis functions $\{\psi_{i,q}^{loc}\}_{i=1,q=1}^{m,Q_i}$.

We recall that $\{\varphi_{i,q}\}_{q=1}^{Q}$ are orthogonal in $L_a^2(\tau_i)$. Without loss of generality, we normalize them such that

$$\int_{\tau_i} a\varphi_{i,q}\varphi_{i,q'} = |\tau_i|\delta_{q,q'}. \tag{4.38}$$

Therefore, for M_0 and A_0 defined in Eqn. (4.28), we have

$$M_0 = |\tau_i| I_Q, \qquad \lambda_{max}(M_0, A_0) = |\tau_i| \lambda_{max}(A_0^{-1}). \tag{4.39}$$

Lemma 4.4.1. *It holds true that*

$$\|\psi_{i,q}^{loc}\|_H \le \frac{C_i}{C_P} \delta^{-d/2} V_d^{-1/2} h^{-d/2-1}. \tag{4.40}$$

Proof. Define

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

where $\mathcal{L}_0$ is the elliptic operator $-\nabla \cdot (a\nabla u)$ with the homogeneous Dirichlet boundary condition on $\partial\tau_i$. From the definition of A_0, we know that $\int_{\tau_i} a\varphi_{i,q}\zeta_{i,q'} = \delta_{q,q'}$. Notice that $\zeta_{i,q} \in H^1_{0,a} \subset H^1_{0,a}(S_r)$ is only supported on τ_i, and thus $\zeta_{i,q}$ satisfies all constraints of $\psi_{i,q}^{loc}$, see Eqn. (4.37). Therefore, we have

$$\|\psi_{i,q}^{loc}\|_H \le \|\zeta_{i,q}\|_H. \tag{4.41}$$

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

$$\|\zeta_{i,q}\|_H^2 = A_0^{-1}(q, q) \le \lambda_{max}(A_0^{-1}). \tag{4.42}$$

We have used $\lambda_{max}(A_0^{-1}) = \sup_{\|v\|_2=1} v^T A_0^{-1} v$ in the last inequality. Combining Eqn. (4.41) and (4.42), we have

$$\|\psi_{i,q}^{loc}\|_H \leq \lambda_{max}^{1/2}(A_0^{-1}) = \frac{C_i \lambda_{i,Q_i+1}^{1/2}}{|\tau_i|^{1/2}},$$

where A_0 and C_i are defined in Eqn. (4.28) and Eqn. (4.29), respectively. With the choice of Q_i, we have $\lambda_{i,Q_i+1} \approx C_P^{-2} h^{-2}$. Because $|\tau_i| \geq V_d(\delta h)^d$ where V_d is volume of the d-dimensional unit ball, we conclude the proof of Eqn. (4.40). □

Theorem 4.4.1. *It holds true that*

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)} \leq C_3 h^{-d/2-1} \exp(-\frac{r-2h}{2lh}), \tag{4.43}$$

where

$$l = (\mathrm{e}-1)(C_P + C_{inv}), \qquad C_3 = \left(\frac{e(1+C_P)^2 + 2eC_{inv}}{\delta^d V_d}\right)^{1/2} \frac{C_i}{C_P}.$$

Proof. Let S_0 be the union of the subdomains τ_j 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 prove is trivial. We can choose $r \geq 2h$ 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.

Let η be a smooth cut-off function such that $0 \leq \eta \leq 1$, $\eta|_{S_1} \equiv 1$, $\eta|_{S_0} \equiv 0$ and $\|\nabla\eta\|_{L^\infty(D)} \leq 1/h$.

Since $\psi_{i,q}^{loc}$ satisfies the same constraints as $\psi_{i,q}$, thanks to Eqn. (2.19) we have

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)}^2 = \|\psi_{i,q}^{loc}\|_{H(D)}^2 - \|\psi_{i,q}\|_{H(D)}^2. \tag{4.44}$$

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

$$\begin{aligned} \psi_{j,q}^{i,r} := \underset{\psi \in H_{0,a}^1(S_r)}{\arg\min} \quad & \|\psi\|_{H(S_r)}^2 \\ \text{s.t.} \quad & \int_{S_r} a\psi\varphi_{j',q'} = \delta_{jq,j'q'}, \quad \forall 1 \leq j' \leq m,\, 1 \leq q' \leq Q_{j'}. \end{aligned} \tag{4.45}$$

Note that $\psi_{i,q}^{loc} = \psi_{i,q}^{i,r}$. Let $w_{jq'} = \int_D \eta a \psi_{i,q} \varphi_{j,q'}$ and $\psi_w^{iq,r} = \sum_{j=1}^m \sum_{q'=1}^{Q_j} w_{jq'} \psi_{j,q'}^{i,r}$. Thanks to the orthogonality between $\psi_{i,q}$ and $\varphi_{j,q'}$, see the constraints in Eqn. (4.24), we have

$$\psi_w^{iq,r} = \psi_{i,q}^{loc} + \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q_j} w_{jq'} \psi_{j,q'}^{i,r}.$$

Using (3) of Theorem 2.2.3, we have $(\psi_{i,q}^{loc}, \psi_{j,q'}^{i,r})_H = \Theta_{iq,jq'}^{i,-1}$, where Θ^i is defined by Eqn. (2.17) with $\mathcal{K} : L_a^2(S_r) \to L_a^2(S_r)$ being the solution operator of $\mathcal{L}u = af$ with the homogeneous Dirichlet boundary condition on ∂S_r. Therefore, we have

$$\|\psi_w^{iq,r}\|_H^2 = \|\psi_{i,q}^{loc}\|_H^2 + \|\sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q_j} w_{jq'} \psi_{j,q'}^{i,r}\|_H^2 + 2 \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q_j} w_{jq'} \Theta_{iq,jq'}^{i,-1}. \quad (4.46)$$

By (2) of Theorem 2.2.3, we know that $\psi_w^{iq,r}$ is the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_w^{iq,r} = \underset{\psi \in H_{0,a}^1(S_r)}{\arg\min} \quad & \|\psi\|_{H(S_r)}^2 \\ \text{s.t.} \quad & \int_{S_r} a\psi\varphi_{j,q'} = \int_{S_r} \eta a \psi_{i,q} \varphi_{j,q'}, \quad \forall 1 \le j \le m,\ 1 \le q' \le Q_j. \end{aligned} \quad (4.47)$$

Noting that $\eta\psi_{i,q}$ satisfies the same constraints, we have $\|\psi_w^{iq,r}\|_H^2 \le \|\eta\psi_{i,q}\|_H^2$. Combined this estimate with (4.44) and (4.46), we obtain

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)}^2 \le \underbrace{\|\eta\psi_{i,q}\|_H^2 - \|\psi_{i,q}\|_H^2}_{I_1} + \underbrace{2|\sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q_j} w_{jq'} \Theta_{iq,jq'}^{i,-1}|}_{I_2}. \quad (4.48)$$

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 4.3.1 and can be estimated in a similar way. We will estimate these two terms as follows.

Let's first estimate I_1. Since $\eta|_{S_1} \equiv 1$ and $\eta|_{S_0} \equiv 0$, we have $I_1 = \|\eta\psi_{i,q}\|_{H(S^*)}^2 - \|\psi_{i,q}\|_{H(S^* \cup S_0)}^2 \le \|\eta\psi_{i,q}\|_{H(S^*)}^2$. Using the same trick to estimate I_1 in the proof of Theorem 4.3.1, we have

$$\begin{aligned} \|\eta\psi_{i,q}\|_{H(S^*)}^2 &= \int_{S^*} a\eta\nabla\psi_{i,q} \cdot \nabla(\eta\psi_{i,q}) + \int_{S^*} a\psi_{i,q}\nabla\eta \cdot \nabla(\eta\psi_{i,q}) \\ &\le \|\eta\|_{L^\infty(S^*)} \|\psi_{i,q}\|_{H(S^*)} \|\eta\psi_{i,q}\|_{H(S^*)} + \|\nabla\eta\|_{L^\infty(S^*)} \|\psi_{i,q}\|_{L_a^2(S^*)} \|\eta\psi_{i,q}\|_{H(S^*)} \\ &\le (1 + C_P) \|\psi_{i,q}\|_{H(S^*)} \|\eta\psi_{i,q}\|_{H(S^*)}. \end{aligned} \quad (4.49)$$

In the last step, we have used $0 \le \eta \le 1$, $|\nabla\eta| \le 1/h$ and $\|\psi_{i,q}\|_{L_a^2(S^*)} \le C_P h$ due to the choice of Q_i, see Eqn. (4.20). Therefore, we have

$$I_1 \le \|\eta\psi_{i,q}\|_{H(S^*)}^2 \le (1 + C_P)^2 \|\psi_{i,q}\|_{H(S^*)}^2. \quad (4.50)$$

Let's now estimate I_2. Combining (3) of Theorem 2.2.3 with the definition of H-norm (2.8), we have

$$\Theta^{i,-1}_{iq,jq'} = (\psi^{loc}_{i,q}, \psi^{i,r}_{j,q'})_{H(S_r)} = \int_{S_r} \psi^{i,r}_{j,q'} \mathcal{L}\psi^{loc}_{i,q}.$$

Thanks to $\mathcal{L}\psi^{loc}_{i,q} \mid_{\tau_j} \in \text{span}\{a\varphi_{j,q'} : 1 \leq q' \leq Q_j\}$ and the orthogonality between Φ_j and $\psi^{i,r}_{j,q'}$, see the constraints in Eqn. (4.45), we have

$$\mathcal{L}\psi^{loc}_{i,q} \mid_{\tau_j} = a \sum_{q'=1}^{Q_j} \Theta^{i,-1}_{iq,jq'} \varphi_{j,q'}.$$

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

$$\|\mathcal{L}\psi^{loc}_{i,q}\|_{L^2_{1/a}(\tau_j)} = |\tau_i|^{1/2} \left(\sum_{q'=1}^{Q} (\Theta^{i,-1}_{iq,jq'})^2 \right)^{1/2}. \tag{4.51}$$

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

$$|\tau_i|^{-1/2} \left(\sum_{q'=1}^{Q_j} |w_{jq'}|^2 \right)^{1/2} \leq \|\eta\psi_{i,q}\|_{L^2_a(\tau_j)} \leq \|\psi_{i,q}\|_{L^2_a(\tau_j)}. \tag{4.52}$$

Here, we have made use of $0 \leq \eta \leq 1$ in the last step. Combining (4.51) and (4.52), we get

$$\begin{aligned} I_2 &= 2| \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q_j} w_{jq'} \Theta^{i,-1}_{iq,jq'} | \leq 2 \sum_{\tau_j \subset S^*} \left(\sum_{q'=1}^{Q_j} (\Theta^{i,-1}_{iq,jq'})^2 \right)^{1/2} \left(\sum_{q'=1}^{Q_j} |w_{jq'}|^2 \right)^{1/2} \\ &\leq 2 \sum_{\tau_j \subset S^*} \|\mathcal{L}\psi^{loc}_{i,q}\|_{L^2_{1/a}(\tau_j)} \|\psi_{i,q}\|_{L^2_a(\tau_j)}. \end{aligned}$$

Now, we arrive at exactly the same situation as I_2 (see (4.34)) in the proof of Theorem 4.3.1. Using the same argument from Eqn. (4.35) to Eqn. (4.36), we obtain

$$I_2 \leq 2C_{inv} \|\psi^{loc}_{i,q}\|_{H(S^*)} \|\psi_{i,q}\|_{H(S^*)}. \tag{4.53}$$

Applying the exponential decay of Theorem 4.3.1 to both $\|\psi_{i,q}\|_{H(S^*)}$ and $\|\psi^{loc}_{i,q}\|_{H(S^*)}$, we obtain

$$\begin{aligned} I_1 + I_2 \leq & (1 + C_P)^2 \|\psi_{i,q}\|^2_{H(D)} \exp(1 - \frac{r - 2h}{lh}) \\ & + 2C_{inv} \|\psi_{i,q}\|_{H(D)} \|\psi^{loc}_{i,q}\|_{H(D)} \exp(1 - \frac{r - 2h}{lh}). \end{aligned} \tag{4.54}$$

Combining $\|\psi_{i,q}\|_{H(D)} \leq \|\psi_{i,q}^{loc}\|_{H(D)}$, Eqn. (4.40), and Eqn. (4.48), we obtain that

$$\|\psi_{i,q} - \psi_{i,q}^{loc}\|_{H(D)} \leq \left(\frac{e(1+C_P)^2 + 2eC_{inv}}{\delta^d V_d}\right)^{1/2} \frac{C_i}{C_P} h^{-d/2-1} \exp(-\frac{r-2h}{2lh}). \tag{4.55}$$

This completes the proof of Eqn. (4.43). □

Theorem 4.4.2. *Let $u \in H^1_{0,a}(D)$ be the weak solution of $-\nabla \cdot (a\nabla u) = af$ with the homogeneous Dirichlet boundary condition and $\psi_{i,q}^{loc}$ be the localized basis functions defined in Eqn.* (4.37)*. Then for*

$$r \geq (d+4)lh\log(1/h) + 2(1 + l\log C_4 + 2l\log C_E + \frac{l}{2}\log\overline{Q})h, \tag{4.56}$$

we have

$$\inf_{v\in\Psi^{loc}} \|u - v\|_{H(D)} \leq 2C_P h\|f\|_{L^2_a(D)}. \tag{4.57}$$

In Eqn. (4.56)*, we have constants*

$$l = (\mathrm{e}-1)(C_P + C_{inv}), \qquad C_4 = \left(\frac{e\,Vol(D)\,((1+C_P)^2 + 2C_{inv})}{\delta^d V_d}\right)^{1/2} \frac{C_{inv}}{C_P^2}, \tag{4.58}$$

which are independent of the contrast a_{max}/a_{min}, and contrast-depending constants

$$C_E = \sup_{u\in H^1_{0,a}(D), u\neq 0} \frac{\|u\|_{L^2_a(D)}}{\|u\|_{H^1_{0,a}(D)}}, \qquad \overline{Q} = \frac{\sum_{i=1}^m Q_i|\tau_i|}{Vol(D)}. \tag{4.59}$$

The quantity $\overline{Q}$ is the average number of basis functions on each patch, and is typically of order 1. Therefore, the term $\log\overline{Q}$ is nearly independent of the contrast a_{max}/a_{min} and can be ignored in practice. The contrast enters only through $\log(C_E)$. Even with the crude bound $C_E \leq C_F(a_{max}/a_{min})^{1/2}$, see Eqn. (4.15), we have $\log C_E \leq \log C_F + \frac{1}{2}\log\left(\frac{a_{max}}{a_{min}}\right)$, which grows logarithmically with the contrast. This term is typically dominated by the $(d+4)lh\log(1/h)$ term as the partition is refined. Therefore, the support size of the localized multiscale basis functions is nearly independent of the contrast, or at most depends on the contrast logarithmically, and we achieve the optimal linear convergence rate with the constant C_P essentially independent of the contrast.

Proof. Let $v_1 := \sum_{i=1}^m \sum_{q=1}^{Q_i} c_{iq}\psi_{i,q}$ and $v_2 := \sum_{i=1}^m \sum_{q=1}^{Q_i} c_{iq}\psi_{i,q}^{loc}$ with $c_{iq} = \int_D au\varphi_{i,q}$. It is easy to verify that $v_1 = \mathcal{P}_\Psi^{(H)} u$ and Eqn. (4.23) gives

$$\|u - v_1\|_H \leq C_P h\|f\|_{L^2_a(D)}. \tag{4.60}$$

Using the Cauchy inequality, we have

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

Since $\{\varphi_{j,q}\}_{q=1}^{Q_i}$ is orthogonal and normalized such that $\int_{\tau_i} a\varphi_{j,q}\varphi_{j,q'} = |\tau_i|\delta_{q,q'}$, we have $(\sum_{q=1}^{Q_i} |c_{iq}|^2)^{1/2} \leq |\tau_i|^{1/2}\|u\|_{L_a^2(\tau_i)}$. Then we obtain

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

Using the energy estimate $\|u\|_{L_a^2(D)} \leq C_E\|\nabla u\|_{L_a^2(D)} \leq C_E^2\|f\|_{L_a^2(D)}$ and Theorem 4.4.1, we obtain

$$\|v_1 - v_2\|_H \leq C_3 C_E^2 (\sum_{i=1}^{m} Q_i|\tau_i|)^{1/2} h^{-\frac{d}{2}-1} \exp(-\frac{r-2h}{2lh}) \|f\|_{L_a^2(D)}. \tag{4.61}$$

Combining Eqn. (4.60) and (4.61) together, we conclude the proof. □

4.5 Asymptotic Analysis of The Two-phase Coefficient Model

In this section, under some geometric assumptions, we use asymptotic analysis to show that C_i (defined in Eqn. (4.29)) can be bounded by a constant independent of the contrast for the two-phase coefficient model.

First of all, it is easy to check that C_i is invariant under the isotropic rescaling and translation of the coordinates x. Specifically, we can rescale the local domain τ_i to $\widehat{\tau}_i = \{(x - x_i)/h : x \in \tau_i\}$ by the dilation $\widehat{x} = (x - x_i)/h$, and the rescaled domain $\widehat{\tau}_i$ has diameter one and $B(0, \delta/2) \subset \widehat{\tau}_i$. After the rescaling, we still have

$$C_i = \sup_{\widehat{\varphi} \in \widehat{\Phi}_i \setminus \{0\}} \frac{\|\widehat{\varphi}\|_{L_{\widehat{a}}^2(\widehat{\tau}_i)}}{\sqrt{\widehat{\lambda}_{i,Q_i+1}} \|\widehat{\mathcal{L}}_0^{-1}(\widehat{a}\widehat{\varphi})\|_{H_{0,\widehat{a}}^1(\widehat{\tau}_i)}},$$

where $\widehat{a}(\widehat{x}) := \widehat{a}((x - x_i)/h) = a(x)$, $\widehat{\Phi}_i$ is the first Q_i-dimensional eigenspace of the rescaled local eigenvalue problem

$$\begin{aligned}-\nabla \cdot (\widehat{a}\nabla\widehat{\varphi}_q) &= \widehat{\lambda}_{i,q}\widehat{a}\widehat{\varphi}_q \\ \boldsymbol{n} \cdot \nabla\widehat{\varphi}_q &= 0 \quad \text{on } \partial\widehat{\tau}_i,\end{aligned}$$

$\widehat{\lambda}_{i,Q_i+1}$ is the corresponding (Q_i+1)'th smallest eigenvalue, and $\widehat{\mathcal{L}}_0^{-1}(\widehat{a}\widehat{\varphi})$ is the solution of the following rescaled local problem with the Dirichlet Boundary condition

$$\begin{aligned} -\nabla\cdot(\widehat{a}\nabla u) &= \widehat{a}\widehat{\varphi}(x) \qquad x\in\widehat{\tau}_i,\\ u &= 0 \qquad x\in\partial\widehat{\tau}_i. \end{aligned}$$

Therefore, when analyzing C_i, we can simply assume that τ_i has diameter one and $B(0,\delta/2)\subset\tau_i$. Our analysis in this section can be applied to every local patch, and thus we drop the subscript i to simplify our notation.

Therefore, we consider the (rescaled) local eigenvalue problem

$$\begin{aligned} -\nabla\cdot(a\nabla\varphi_q) &= \lambda_q a\varphi_q\\ \boldsymbol{n}\cdot\nabla\varphi_q &= 0 \quad \text{on } \partial\tau, \end{aligned} \tag{4.62}$$

where the (rescaled) local domain τ has diameter 1 and $B(0,\delta/2)\subset\tau$. For a given $Q\in\mathbb{N}$, define $\Phi=\text{span}\{\varphi_q : 1\le q\le Q\}$. We want to show that the following quantity

$$C_i^2 = \sup_{\varphi\in\Phi\setminus\{0\}} \frac{\int_\tau a\varphi^2}{\lambda_{Q+1}\int_\tau a|\nabla\psi|^2}, \tag{4.63}$$

can be bounded by a constant independent of the contrast of the coefficient. In Eqn. (4.63), ψ is the solution of the following (rescaled) local problem with the Dirichlet Boundary condition

$$\begin{aligned} -\nabla\cdot(a\nabla\psi) &= a\varphi(x) \qquad x\in\tau,\\ u &= 0 \qquad x\in\partial\tau. \end{aligned} \tag{4.64}$$

We will only show this result for the two-phase coefficient model, in which we assume that the coefficient a only takes two values, i.e., 1 and $\eta>0$, on the (rescaled) local domain τ. We further assume that τ is the disjoint union of a background domain and inclusions, i.e., $\tau = D_0\cup(\cup_{q=1}^Q D_q)$ and

$$a(x) = \begin{cases} \eta & \text{if } x\in\cup_{q=1}^Q D_q\\ 1 & \text{if } x\in D_0=\tau\setminus\overline{\cup_{q=1}^Q D_q} \end{cases}. \tag{4.65}$$

We assume that $D_0, D_1,\ldots,D_Q$, are polygonal domains (or domains with smooth boundaries). We also assume that each D_q is a connected domain, $q=1,\ldots,Q$. Let D_0 represent the background domain and the subdomains $\{D_q\}_{q=1}^Q$ represent the inclusions.

We first analyze the eigenvalue problem (4.6) in the first subsection. Then we analyze the case of high-conductivity inclusions ($\eta > 1$) in the second subsection and the case of low-conductivity inclusions ($\eta < 1$) in the third subsection. Following the method in [18], we use asymptotic expansion with respect to the contrast η to analyze the magnitude order of different objects. In this thesis, we compute only the first (dominant) term in the expansion to get the order of the quantity of interest C_i^2. We do not provide the proof of the convergence of the asymptotic expansion in $H^1(\tau)$. However, as in [18], one can continue to compute higher order terms in the expansion, and can also prove the convergence of the asymptotic expansion in $H^1(\tau)$. Finally, We point out that the assumption that the background domain D_0 is connected can be relaxed, but we will not elaborate on this issue in this thesis.

Local Eigenvalue Problems

In this subsection, we will show that for the local eigenvalue problem (4.62) the number of small eigenvalues is the number of disconnected high-conductivity inclusions or channels. We will also give the principal component of the eigenfunctions. In this subsection, we assume that we have Q disconnected high-conductivity inclusions, i.e., $\eta > 1$ in Eqn. (4.65), but we do *not* assume that the background domain D_0 is connected. We assume that there exists $\chi_q^{(n)} \in H^1(\tau)$ such that

$$\chi_q^{(n)} \equiv \delta_{ql} \quad \text{on } D_l \text{ for } l = 1, 2, \ldots, Q, \tag{4.66}$$

and $\chi_q^{(n)}$ is defined as the harmonic extension of its boundary data in D_0, i.e.,

$$\begin{aligned} \int_{D_0} \nabla \chi_q^{(n)} \cdot \nabla z = 0, \quad & \text{for all } z \in H_0^1(D_0), \\ \chi_q^{(n)} = \delta_{ql} \quad & \text{on } \partial D_l \text{ for } l = 1, 2, \ldots, Q, \\ \boldsymbol{n} \cdot \nabla \chi_q^{(n)} = 0 \quad & \text{on } \partial D_0 \cap \partial \tau. \end{aligned} \tag{4.67}$$

We call $\chi_q^{(n)}$ the Neumann harmonic characteristic function of D_q. We define

$$V_\chi^{(n)} = \text{span}\{\chi_q^{(n)} : 1 \le q \le Q\} \tag{4.68}$$

as the space spanned by the Neumann harmonic characteristic functions.

Similarly, *when D_q is an interior inclusion*, we define its Dirichlet harmonic characteristic function $\chi_q^{(d)} \in H_0^1(\tau)$ that satisfies

$$\chi_q^{(d)} \equiv \delta_{ql} \quad \text{on } D_l \text{ for } l = 1, 2, \ldots, Q, \tag{4.69}$$

and, in D_0, $\chi_q^{(d)}$ is defined as the harmonic extension of its boundary data in D_0, i.e.,

$$\begin{aligned} \int_{D_0} \nabla\chi_q^{(d)} \cdot \nabla z = 0, \quad \text{for all } z \in H_0^1(D_0), \\ \chi_q^{(d)} = \delta_{ql} \quad \text{on } \partial D_l \text{ for } l = 1, 2, \ldots, Q, \\ \chi_q^{(d)} = 0 \quad \text{on } \partial\tau. \end{aligned} \tag{4.70}$$

We also define

$$V_\chi^{(d)} = \text{span}\{\chi_q^{(d)} : 1 \le q \le Q,\ D_q \text{ is an interior inclusion in } \tau.\} \tag{4.71}$$

as the space spanned by the Neumann harmonic characteristic functions.

We point out that the boundaries of τ and D_q $(0 \le q \le Q)$ should be smooth enough so that both the Neumann and Dirichlet harmonic characteristic functions exist.

On the number of small contrast-dependent eigenvalues

We prove that $\lambda_q = \mathcal{O}(1/\eta)$ for $q = 1, \ldots, Q$ and $\lambda_{Q+1} = \mathcal{O}(1)$ (i.e., it is bounded below independent of η). The previous statement implies that if we take the number of local measurement functions to be equal to the number of high-conductivity inclusions and channels in τ_i, we obtain an error estimate independent of the contrast. We point out that this proof follows the idea presented in Appendix A in [46], with some minor modification.

First, we prove that there are at least Q small eigenvalues. Since we have

$$\lambda_Q = \min_{\dim(V)=Q} \max_{v \in V\setminus\{0\}} R(v), \quad \text{where} \quad R(v) = \frac{\int_\tau a|\nabla v|^2}{\int_\tau a|v|^2},$$

we need to find a Q-dimensional subspace $V \subset H^1(\tau)$ where the quotient $R(\cdot)$ is of order $1/\eta$. Let $v \in V_\chi^{(n)}$ and assume that $v = \sum_q v_q \chi_q^{(n)}$. Then we have

$$R(v) = \frac{\int_\tau a|\nabla v|^2}{\int_\tau a|v|^2} = \frac{\int_{D_0} |\nabla v|^2}{\int_{D_0} v^2 + \eta \sum_{q=1}^Q \int_{D_q} v^2} \le \frac{\int_{D_0} |\nabla v|^2}{\eta \sum_{q=1}^Q v_q^2 |D_q|}.$$

Notice that $C = \max_{\boldsymbol{v}\in\mathbb{R}^Q} \frac{\int_{D_0} |\nabla v|^2}{\sum_{q=1}^Q v_q^2 |D_q|}$ is a constant that depends on the geometries of $D_0, D_1, \cdots, D_Q$, but is independent on η. Therefore, we have $\lambda_Q = \mathcal{O}(1/\eta)$.

On the other hand, we have that

$$\lambda_{Q+1} = \min_{\dim(V)=Q+1} \max_{v \in V\setminus\{0\}} R(v).$$

Then we have to show that for every $(Q+1)$-dimensional subspace $V \subset H^1(\tau)$, there exists a function $v \in V$ such that $R(v) = \mathcal{O}(1)$. Define the space

$$V_{poin} = \{v \in H^1(\tau) : \int_{D_q} v = 0, \quad q = 1, 2, \ldots, Q\}.$$

The subspace V_{poin} is of codimension Q. Note that for every $v \in V_{poin}$, we can apply the Poincare inequality in every D_q. Then we can write

$$\sum_{q=1}^{Q} \int_{D_q} |v|^2 \leq C_1 \sum_{q=1}^{Q} \int_{D_q} |\nabla v|^2, \quad \text{for all } v \in V_{poin},$$

where C_1 is independent of η, or equivalently

$$\sum_{q=1}^{Q} (\eta - 1) \int_{D_q} |v|^2 \leq C_1 \sum_{q=1}^{Q} (\eta - 1) \int_{D_q} |\nabla v|^2, \quad \text{for all } v \in V_{poin}, \tag{4.72}$$

We can apply the standard Poincare inequality to functions in V_{poin}:

$$\int_{\tau} |v|^2 \leq C_2 \int_{\tau} |\nabla v|^2, \quad \text{for all } v \in V_{poin}, \tag{4.73}$$

where C_2 is independent of η. Adding (4.72) and (4.73), we obtain that

$$\int_{\tau} a|v|^2 \leq C \int_{\tau} a|\nabla v|^2, \quad \text{for all } v \in V_{poin},$$

where the constant C is independent of η, but depends on the geometries of D_q $(1 \leq q \leq Q)$. Let $V \subset H^1(\tau)$ be a subspace of dimension $M + 1$. We have that the intersection between V and V_{poin} is a subspace of dimension at least one. Then we can select $v \in V \cap V_{poin}$ with $v \neq 0$, and for this vector we have

$$R(v) = \frac{\int_{\tau} a|\nabla v|^2}{\int_{\tau} a|v|^2} \geq \frac{1}{C} = \mathcal{O}(1).$$

This completes the proof.

Expansions for eigenvalues and eigenvectors

For $1 \leq q \leq Q$, now we use the asymptotic analysis to obtain the asymptotic expansion of the eigenvalue λ_q and the eigenfunction φ_q. Due to the above analysis, we expand λ_q as

$$\lambda_q = \frac{C_1}{\eta} + \frac{C_2}{\eta^2} + \ldots. \tag{4.74}$$

Here, C_1 can be 0, since we only know that $\lambda_q = \mathcal{O}(1/\eta)$. We also expand the corresponding eigenfunction φ_q as

$$\varphi_q = \varphi_{q,0} + \frac{\varphi_{q,1}}{\eta} + \frac{\varphi_{q,2}}{\eta^2} + \dots. \tag{4.75}$$

Plugging Eqn. (4.74) and (4.75) into the following variational form of the local eigenvalue problem, i.e.,

$$\int_{D_0} \nabla\varphi_q \cdot \nabla v + \eta \int_{\cup_{q=1}^{Q} D_q} \nabla\varphi_q \cdot \nabla v = \lambda_q \int_{D_0} \varphi_q v + \eta\lambda_q \int_{\cup_{q=1}^{Q} D_q} \varphi_q v \qquad \forall v \in H^1(\tau),$$

the terms corresponding to η^1 and η^0 give the following two equations:

$$\int_{\cup_{q=1}^{Q} D_q} \nabla\varphi_{q,0} \cdot \nabla v = 0, \quad \forall v \in H^1(\tau), \tag{4.76}$$

and

$$\int_{D_0} \nabla\varphi_{q,0} \cdot \nabla v + \int_{\cup_{q=1}^{Q} D_q} \nabla\varphi_{q,1} \cdot \nabla v = C_1 \int_{\cup_{q=1}^{Q} D_q} \varphi_{q,0} v, \quad \forall v \in H^1(\tau). \tag{4.77}$$

From Eqn. (4.76), we know that $\varphi_{q,0} \equiv c_q$ for some $c_q \in \mathbb{R}$ and for any $1 \le q \le Q$. By taking $v \in H_0^1(D_0)$ in Eqn. (4.77), we know that

$$\int_{D_0} \nabla\varphi_{q,0} \cdot \nabla v = 0, \quad \forall v \in H_0^1(D_0),$$

and thus $\varphi_{q,0}$ is harmonic in D_0. Therefore, we conclude that

$$\varphi_{q,0} \in V_\chi^{(n)} \quad \forall 1 \le q \le Q. \tag{4.78}$$

Expansions For High-Conductivity Inclusions

In this subsection, we consider the case of high-conductivity inclusions, i.e. $\eta \gg 1$. From the analysis in the last subsection, we can take the number of local measurement functions to be the number of high-conductivity inclusions (i.e. Q), and achieve an error estimate independent of the contrast. Because we have proved that λ_{Q+1} is $\mathcal{O}(1)$ independent of the contrast η, we only need to prove that $\frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2} = \mathcal{O}(1)$ for all $\varphi \in \Phi\backslash\{0\}$ to prove that $C_i^2 = \frac{\int_\tau a\varphi^2}{\lambda_{Q+1}\int_\tau a|\nabla\psi|^2}$ is $\mathcal{O}(1)$. In fact, we will show

$$\sup_{\varphi\in\Phi\backslash\{0\}} \frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2} = \mathcal{O}(1/\eta)$$

when we have (at least) one interior inclusion, and

$$\sup_{\varphi\in\Phi\backslash\{0\}} \frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2} = \mathcal{O}(1)$$

when all the inclusions intersect with the domain boundary, i.e., $\partial D_q \cap \partial\tau \neq \emptyset$ for all $1 \leq q \leq Q$. Therefore, the constant C_i is bounded by an $\mathcal{O}(1)$ constant independent of the contrast η. Thanks to Eqn. (4.75) and (4.78), for any $\varphi \in \Phi\backslash\{0\}$, we write

$$\varphi_q = \varphi_0 + \frac{\varphi_1}{\eta} + \frac{\varphi_2}{\eta^2} + \ldots, \text{ where } \varphi_0 = \sum_{q=1}^{Q} c_q \chi_q^{(n)} \not\equiv 0. \tag{4.79}$$

We will plug φ_q into the variational form of Eqn. (4.64), i.e.,

$$\int_{D_0} \nabla\psi \cdot \nabla v + \eta \int_{\cup_{q=1}^{Q} D_q} \nabla\psi \cdot \nabla v = \int_{D_0} \varphi v + \eta \int_{\cup_{q=1}^{Q} D_q} \varphi v \qquad \forall v \in H_0^1(\tau) \tag{4.80}$$

to obtain the asymptotic expansion of ψ.

The case when all inclusions are interior inclusions

When all inclusions are interior inclusions, we seek to determine $\{\psi_i : -1 \leq i \leq +\infty\} \subset H_0^1(\tau)$ such that

$$\psi = \eta\psi_{-1} + \psi_0 + \frac{\psi_1}{\eta} + \frac{\psi_2}{\eta^2} + \ldots. \tag{4.81}$$

Plugging Eqn. (4.79) and (4.81) into Eqn. (4.80), the terms corresponding to η^2 and η^1 give the following two equations:

$$\int_{\cup_{q=1}^{Q} D_q} \nabla\psi_{-1} \cdot \nabla v = 0, \quad \forall v \in H_0^1(\tau), \tag{4.82}$$

and

$$\int_{D_0} \nabla\psi_{-1} \cdot \nabla v + \int_{\cup_{q=1}^{Q} D_q} \nabla\psi_0 \cdot \nabla v = \int_{\cup_{q=1}^{Q} D_q} \varphi_0 v, \quad \forall v \in H_0^1(\tau). \tag{4.83}$$

From Eqn. (4.82), we know that $\psi_{-1} \equiv z_q$ for some $z_q \in \mathbb{R}$ and for any $1 \leq q \leq Q$. By taking $v \in H_0^1(D_0)$ in Eqn. (4.83), we know that

$$\int_{D_0} \nabla\psi_{-1} \cdot \nabla v = 0, \quad \forall v \in H_0^1(D_0),$$

and thus ψ_{-1} is harmonic in D_0. Therefore, we conclude that

$$\psi_{-1} = \sum_{q=1}^{Q} z_q \chi_q^{(d)} \in V_\chi^{(d)}. \tag{4.84}$$

Since $\varphi_0 = \sum_{q=1}^{Q} c_q \chi_q^{(n)} \not\equiv 0$, we know that $\psi_{-1} = \sum_{q=1}^{Q} z_q \chi_q^{(d)} \not\equiv 0$. Therefore, we have

$$\frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2} = \frac{\eta \sum_{q=1}^{Q} \int_{D_q} \varphi_0^2 + \mathcal{O}(1)}{\eta^2 \int_{D_0} |\nabla\psi_{-1}|^2 + \mathcal{O}(\eta)} = \mathcal{O}(\frac{1}{\eta}), \quad \forall \varphi \in \Phi\backslash\{0\}.$$

The case when all inclusions intersect with the boundary

When all inclusions intersect with the boundary, we seek to determine $\{\psi_i : 0 \leq i \leq +\infty\} \subset H_0^1(\tau)$ such that

$$\psi = \psi_0 + \frac{\psi_1}{\eta} + \frac{\psi_2}{\eta^2} + \dots. \tag{4.85}$$

Plugging Eqn. (4.79) and (4.85) into Eqn. (4.80), the terms corresponding to η^1 and η^0 give the following two equations:

$$\int_{\cup_{q=1}^{Q} D_q} \nabla\psi_0 \cdot \nabla v = \int_{\cup_{q=1}^{Q} D_q} \varphi_0 v, \quad \forall v \in H_0^1(\tau), \tag{4.86}$$

and

$$\int_{D_0} \nabla\psi_0 \cdot \nabla v + \int_{\cup_{q=1}^{Q} D_q} \nabla\psi_1 \cdot \nabla v = \int_{D_0} \varphi_0 v + \int_{\cup_{q=1}^{Q} D_q} \varphi_1 v, \quad \forall v \in H_0^1(\tau). \tag{4.87}$$

Since D_q is connected and intersects with the boundary and $\{D_q\}_{q=1}^{Q}$ are not connected to each other, Eqn. (4.86) uniquely determines ψ_0 restricted on D_q:

$$\int_{D_q} \nabla\psi_0 \cdot \nabla v = \int_{D_q} \varphi_0 v, \quad \forall v \in H_0^1(D_q),$$

$$\psi_0 \mid_{\partial\tau \cap \partial D_q} = 0, \quad \boldsymbol{n} \cdot \nabla\psi_0 \mid_{\partial\tau \setminus \partial D_q} = 0.$$

Since $\varphi_0 = \sum_{q=1}^{Q} c_q \chi_q^{(n)} \not\equiv 0$, there exists at least one q such that $\psi_0 \mid_{D_q} \not\equiv 0$. By taking $v \in H_0^1(D_0)$ in Eqn. (4.87), we know that

$$\int_{D_0} \nabla\psi_0 \cdot \nabla v = \int_{D_0} \varphi_0 v, \quad \forall v \in H_0^1(D_0).$$

Together with the Dirichlet boundary condition given on $\partial\tau$ and ∂D_q ($1 \leq q \leq Q$), $\psi_0 \mid_{D_0}$ can be uniquely determined. Therefore, we have

$$\frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2} = \frac{\eta \sum_{q=1}^{Q} \int_{D_q} \varphi_0^2 + \mathcal{O}(1)}{\eta \sum_{q=1}^{Q} \int_{D_q} |\nabla\psi_0|^2 + \mathcal{O}(1)} = \mathcal{O}(1), \quad \forall \varphi \in \Phi \setminus \{0\}.$$

The case when there is at least one interior inclusion

For the case when there is at least one interior inclusion, the asymptotic expansion of ψ and the order of $\frac{\int_\tau a\varphi^2}{\int_\tau a|\nabla\psi|^2}$ are similar to the case when all inclusions are interior inclusions. We will not elaborate its derivations here. To summarize, we conclude that for the case of high-conductivity inclusions, we have

$$C_i^2 = \sup_{\varphi \in \Phi \setminus \{0\}} \frac{\int_\tau a\varphi^2}{\lambda_{Q+1} \int_\tau a|\nabla\psi|^2} = \mathcal{O}(1),$$

which implies that the constant C_i can be bounded by an $\mathcal{O}(1)$ constant independent of the contrast η.

Expansions For Low-Conductivity Inclusions

In this subsection, we consider the case of low-conductivity inclusions, i.e. $\eta \ll 1$. Because we are considering the weighted operator $-\frac{1}{a}\nabla \cdot (a\nabla)$, the two-phase coefficient with low-conductivity (η) inclusions and "1" background is equivalent to the two-phase coefficient with high-conductivity ($1/\eta$) background and "1" inclusions. Thanks to the assumption that the background domain D_0 is connected and the argument in section 4.5, there is only one small eigenvalue, which is 0, and the second eigenvalue $\lambda_2 = \mathcal{O}(1)$. Therefore, $Q = 1$ is sufficient to achieve an error estimate independent of the contrast. In this case, the one-dimensional eigenspace Φ is the space consisting of constant functions. In this subsection, using the asymptotic analysis, we show that

$$C_i^2 = \frac{\int_\tau a}{\lambda_2 \int_\tau a|\nabla\psi|^2} = \mathcal{O}(1),$$

where ψ is the solution of the following (rescaled) local problem with the Dirichlet Boundary condition

$$\begin{aligned} -\nabla \cdot (a\nabla\psi) = a \qquad & x \in \tau, \\ u = 0 \qquad & x \in \partial\tau. \end{aligned}$$

Since we already have $\lambda_2 = \mathcal{O}(1)$, we only need to show $\frac{\int_\tau a}{\int_\tau a|\nabla\psi|^2} = \mathcal{O}(1)$. Before, we go to the details, we point out that the following analysis can be generalized to the case when the background domain D_0 is not connected and contains several connected subdomains. In this case, the number of local measurement functions should be the number of the connected subdomains in D_0.

Now we show that $\frac{\int_\tau a}{\int_\tau a|\nabla\psi|^2} = \mathcal{O}(1)$. We seek to determine $\{\psi_i : 0 \le i \le +\infty\} \subset H_0^1(\tau)$ such that

$$\psi = \psi_0 + \eta\psi_1 + \eta^2\psi_2 + \dots. \tag{4.88}$$

Plugging (4.88) into Eqn. (4.80), the terms corresponding to η^0 and η^1 give the following two equations:

$$\int_{D_0} \nabla\psi_0 \cdot \nabla v = \int_{D_0} v, \quad \forall v \in H_0^1(\tau), \tag{4.89}$$

and

$$\int_{D_0} \nabla\psi_1 \cdot \nabla v + \int_{\cup_{q=1}^{Q} D_q} \nabla\psi_0 \cdot \nabla v = \int_{\cup_{q=1}^{Q} D_q} v, \quad \forall v \in H_0^1(\tau). \tag{4.90}$$

Since D_0 is connected and intersects with the boundary, Eqn. (4.89) uniquely determines ψ_0 restricted on D_0:

$$\int_{D_0} \nabla\psi_0 \cdot \nabla v = \int_{D_0} v, \quad \forall v \in H_0^1(D_0),$$
$$\psi_0 \mid_{\partial\tau\cap\partial D_0} = 0, \quad \boldsymbol{n} \cdot \nabla\psi_0 \mid_{\partial\tau\setminus\partial D_0} = 0.$$

Since we have the constant "1" on the right hand side, we have $\psi_0 \not\equiv 0$. By taking $v \in H_0^1(D_q)$ in Eqn. (4.90), we know that

$$\int_{D_q} \nabla\psi_0 \cdot \nabla v = \int_{D_q} v, \quad \forall v \in H_0^1(D_q).$$

Together with the Dirichlet boundary condition given on $\partial\tau$ and ∂D_q ($1 \leq q \leq Q$), $\psi_0 \mid_{D_q}$ can be uniquely determined. Therefore, we have

$$\frac{\int_\tau a}{\int_\tau a|\nabla\psi|^2} = \frac{|D_0| + \mathcal{O}(\eta)}{\int_{D_0} |\nabla\psi_0|^2 + \mathcal{O}(\eta)} = \mathcal{O}(1).$$

It is obvious to see the similarity between the case when all high-conductivity inclusions intersect with the boundary and the current case when we have low-conductivity inclusions. The reason for this similarity is from the a-weighted L^2 formulation of the local eigenvalue problem (4.62) and the local elliptic problem (4.64). With this a-weighted L^2 formulation, multiplying the coefficient a by any constant (like η or $1/\eta$) does not change the problem. Therefore, the problem whose coefficients have low-conductivity inclusions is equivalent to the problem whose coefficients have high-conductivity background.

4.6 Numerical Examples

In this section, we apply our method to a 2D second-order elliptic equation with high contrast coefficients, and show that the exponential decay rate of the energy minimizing basis function remains the same as the contrast of the coefficients increases.

Consider the following 2D second-order elliptic equation with the homogeneous Dirichlet boundary condition

$$\begin{aligned} -\nabla \cdot (a(x,y)\nabla u(x,y)) &= f(x,y), \quad (x,y) \in D := (0,1)^2 \\ u(x,y) &= 0, \quad (x,y) \in \partial D. \end{aligned} \tag{4.91}$$

The coefficient

$$a(x,y) = 1 + (\eta - 1)\sum_{k=1}^{3} \chi_{D_k}(x,y) \qquad \eta \geq 1 \tag{4.92}$$

takes only two values, i.e., 1 in the background domain and η in three channels $\{D_k\}_{k=1}^3$. The contrast of the coefficient is thus η. We use a uniform mesh with mesh size $h_x = h_y = 1/13$ to partition the physical domain, resulting in $13 \times 13 = 169$ local patches. For $\eta = 10^6$, the coefficient and the partition are shown in Figure 4.1.

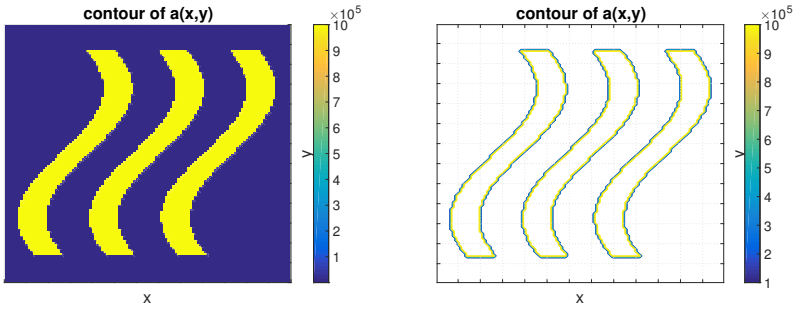


Figure 4.1: The high-contrast ($\eta = 10^6$) coefficient with three high-conductivity channels. The 13×13 partition is shown in the contour plot.

We first focus on the local patch $[h_x, 2 * h_x] \times [8 * h_x, 9 * h_x]$. In Figure 4.2, we plot (in log scale) the global energy-minimizing basis function ψ_i constructed in section 2.3, where we do not use the a-weighted L^2 norm and the local constant function is the only local measurement function. We also plot (in log scale) their energy norm distributions on all the local patches. On the left column, we have ψ_i for $\eta = 1$, where the basis function decays exponentially fast away from its patch, as we proved in Theorem 2.3.1. On the right column, we have ψ_i for $\eta = 10^6$, where the basis function shows nearly no decay along the high-conductivity channels. In the right-bottom figure, we can also see that the energy norm shows nearly no decay along with the channels.

Using the method proposed in this chapter, we take the number of local measurement functions, denoted as $Q_i = \min\{q : 1/\lambda_{i,q+1} < \frac{h_x^2}{0.99 * \pi^2}\}$, which guarantees an error estimate independent of the contrast. We have two measurement functions on the patch $[h_x, 2 * h_x] \times [8 * h_x, 9 * h_x]$, and the corresponding $\psi_{i,1}$ and $\psi_{i,2}$ (in log scale) are plotted in Figure 4.3. We can see that the exponential decay along the high-conductivity channels is as fast as the decay in the background domain. Comparing the energy norm distribution in Figure 4.3 with that in Figure 4.2, we can also see that our new construction improves the decay rate along the high-conductivity channels.

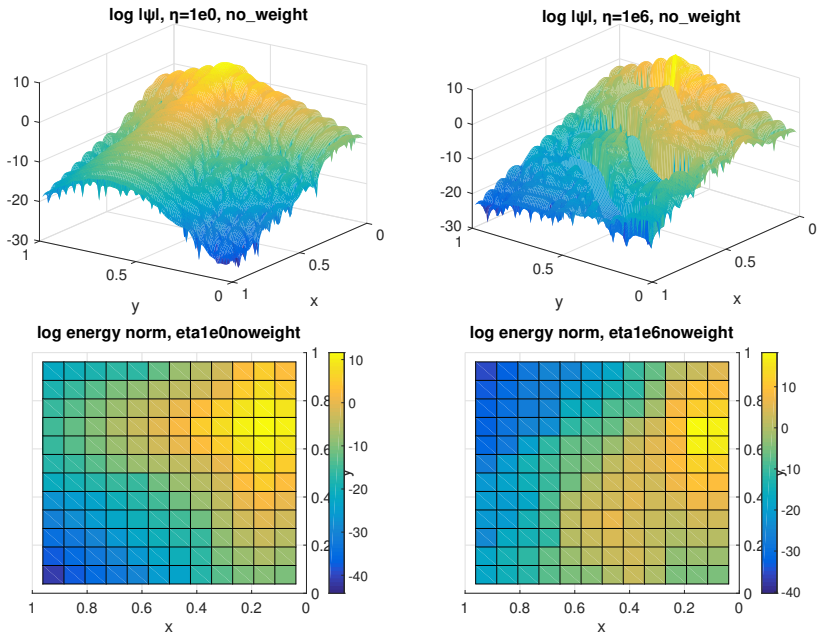


Figure 4.2: The basis function ψ_i and its energy norm (both in log scale) with local piecewise constant measurement functions (construction in section 2.3). $\eta = 1$ on the left column, and $\eta = 10^6$ on the right column.

In Figure 4.4, we show the number of basis functions (left) and the constant C_i in the inverse energy estimate (right) on all the local patches. There are at most two basis functions per patch. We point out that the number of the basis functions does not increase any more even if we further increase the contrast η. The constant C_i, which is associated with the decay rate of the basis function, is bounded by 10, although the contrast has now reached 10^6.

In Figure 4.5, we plot the $C_i(\eta)$ for three different patches. The label $(2, 9)$ means the patch located at $[h_x, 2 * h_x] \times [8 * h_y, 9 * h_y]$, which is the patch we analyzed above. The patch $(4, 5)$ is the patch with the largest C_i in Figure 4.4 (right). We also include the patch $(6, 6)$ which has an intermediate C_i. As we analyzed in section 4.5, C_i does not increase in a polynomial fashion as the contrast η increases. In fact, C_i can be bounded by a constant as the contrast goes to infinity.

Finally, we point out that for the 13×13 partition, there is at most one connected high-conductivity region per patch. If we change to a 9×9 partition, we will have patches that contain two disconnected high-conductivity regions. We have also run experiments on the 9×9 partition, and obtained similar results as above. We choose to present the result on the 13×13 partition

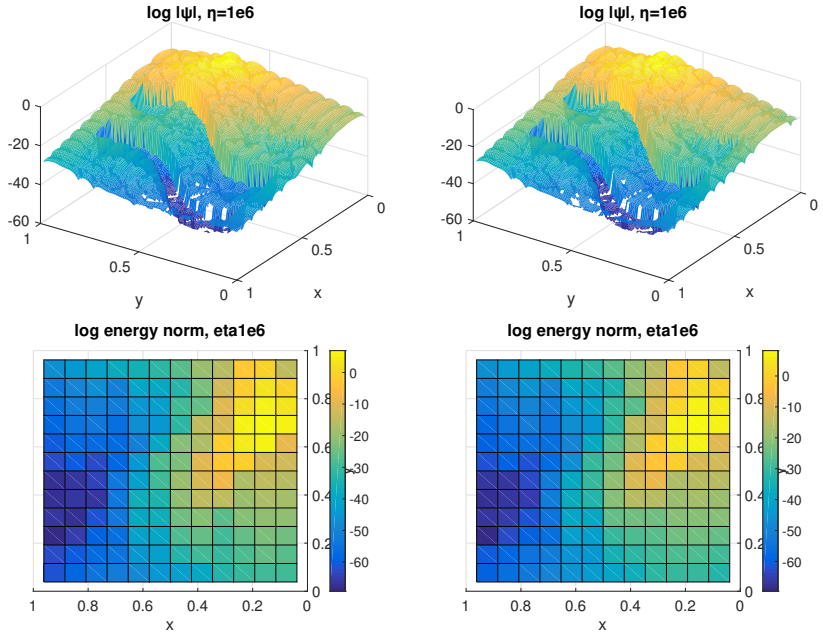


Figure 4.3: The two basis functions $\psi_{i,q}$ and their energy norm (both in log scale) constructed by Eqn. (4.24).

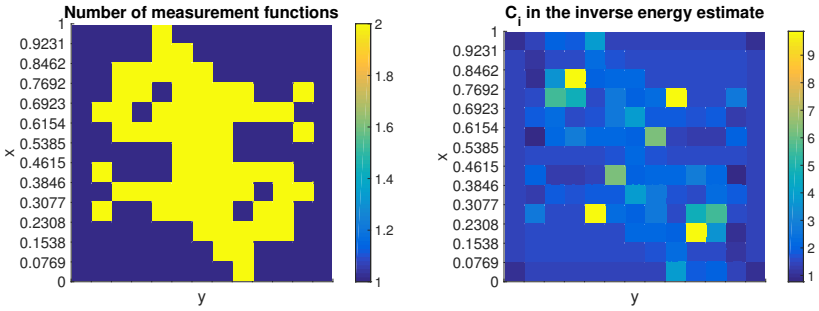


Figure 4.4: Number of basis functions (left) and C_i in the inverse energy estimate (right) per patch.

because the exponential decay is visually clearer when we have more patches per axis.

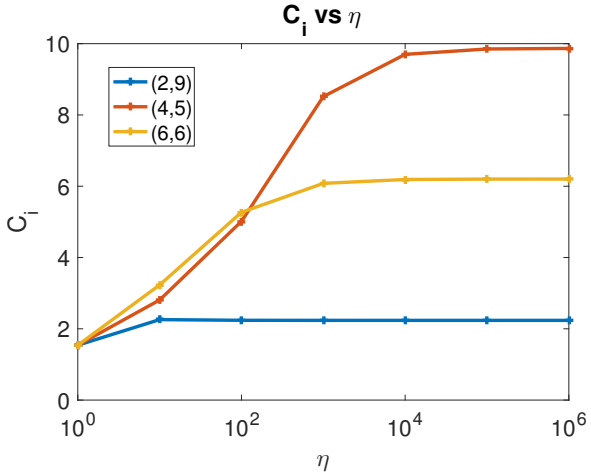


Figure 4.5: The trace of C_i as the contrast η increases. $(2, 9)$ refers to the patch located at $[h_x, 2 * h_x] \times [8 * h_y, 9 * h_y]$

Chapter 5

INTRINSIC SPARSE MODE DECOMPOSITION FOR LOW RANK PSD MATRICES

Many problems in science and engineering lead to huge *symmetric* and *positive semidefinite* (PSD) matrices. Often they arise from the discretization of self-adjoint PSD operators or their kernels, especially in the context of data science and partial differential equations.

Consider a symmetric PSD matrix of size $N \times N$, denoted as A. Since N is typically large, this causes serious obstructions when dealing numerically with such problems. Fortunately, in many applications, the matrix A is low-rank or approximately low-rank, i.e., there exists $\{\psi_1, \ldots, \psi_K\} \subset \mathbb{R}^N$ for $K \ll N$ such that

$$A = \sum_{k=1}^{K} \psi_k \psi_k^T \quad \text{or} \quad \|A - \sum_{k=1}^{K} \psi_k \psi_k^T\|_2 \leq \epsilon,$$

respectively. Here, $\epsilon > 0$ is some small number and $\|A\|_2 = \lambda_{max}(A)$ is the largest eigenvalue of A. To obtain such a low-rank decomposition/approximation of A, the most natural method is perhaps the eigendecomposition with $\{\psi_k\}_{k=1}^K$ as the eigenvectors corresponding to the largest K eigenvalues of A. An additional advantage of the eigendecomposition is the fact that eigenvectors are orthogonal to each other. However, eigenvectors are typically dense vectors, i.e., every entry is typically nonzero.

For a symmetric PSD matrix A with rank $K \ll N$, the aim of this chapter is to find an alternative decomposition

$$A = \sum_{k=1}^{K} g_k g_k^T. \tag{5.1}$$

Here the number of components is still its rank K, which is optimal, and the modes $\{g_k\}_{k=1}^K$ are required to be as sparse as possible. In this chapter, we work on symmetric PSD matrices, which are typically discretized representation of self-adjoint PSD operators or their kernels. We could have just as well worked on the self-adjoint PSD operators themselves, which would correspond to the case where $N = \infty$. Much of what will be discussed below applies equally well to this case.

5.1 Our Results

The number of nonzero entries of a vector $\psi \in \mathbb{R}^N$ is called its l_0 norm, denoted by $\|\psi\|_0$. Since the modes in (5.1) are required to be as sparse as possible, the sparse decomposition problem is naturally formulated as the following optimization problem:

$$\boxed{\min_{\psi_1,\ldots,\psi_K \in \mathbb{R}^N} \quad \sum_{k=1}^{K} \|\psi_k\|_0 \quad \text{s.t.} \quad A = \sum_{k=1}^{K} \psi_k \psi_k^T .} \tag{5.2}$$

However, this problem is rather difficult to solve because: first, minimizing the l_0 norm results in a combinatorial problem and is computationally intractable in general; second, the number of unknown variables is $K \times N$ where N is typically a huge number. Therefore, we introduce the following patchwise sparseness as a surrogate of $\|\psi_k\|_0$ and make the problem computationally tractable.

Definition 5.1.1 (Patchwise sparseness)**.** *Suppose that $\mathcal{P} = \{P_m\}_{m=1}^M$ is a disjoint partition of the N nodes, i.e., $[N] \equiv \{1, 2, 3, \ldots, N\} = \sqcup_{m=1}^M P_m$. The patchwise sparseness of $\psi \in \mathbb{R}^N$ with respect to (w.r.t.) the partition $\mathcal{P}$, denoted by $s(\psi; \mathcal{P})$, is defined as*

$$s(\psi; \mathcal{P}) = \#\{P \in \mathcal{P} : \psi|_P \neq \mathbf{0}\}.$$

Throughout this chapter, $[N]$ denotes the index set $\{1, 2, 3, \ldots, N\}$; $\mathbf{0}$ denotes the vectors with all entries equal to 0; $|P|$ denotes the cardinality of a set P; $\psi|_P \in \mathbb{R}^{|P|}$ denotes the restriction of $\psi \in \mathbb{R}^N$ on patch P. Once the partition $\mathcal{P}$ is fixed, smaller $s(\psi; \mathcal{P})$ means that ψ is nonzero on fewer patches, which implies a sparser vector. With patchwise sparseness as a surrogate of the l_0 norm, the sparse decomposition problem (5.2) is relaxed to

$$\boxed{\min_{\psi_1,\ldots,\psi_K \in \mathbb{R}^N} \quad \sum_{k=1}^{K} s(\psi_k; \mathcal{P}) \quad \text{s.t.} \quad A = \sum_{k=1}^{K} \psi_k \psi_k^T .} \tag{5.3}$$

If $\{g_k\}_{k=1}^K$ is an optimizer for (5.3), we call them a set of *intrinsic sparse modes* for A under partition $\mathcal{P}$. Since the objective function of problem (5.3) only takes nonnegative integer values, we know that for a symmetric PSD matrix A with rank K, there exists at least one set of intrinsic sparse modes.

It is obvious that the intrinsic sparse modes depend on the domain partition $\mathcal{P}$. Two extreme cases would be $M = N$ and $M = 1$. For $M = N$, $s(\psi; \mathcal{P})$ recovers $\|\psi\|_0$ and the patchwise sparseness minimization problem (5.3) recovers the original l_0 minimization problem (5.2). Unfortunately, it is computationally intractable. For $M = 1$, every non-zero vector has sparseness one, and thus the number of nonzero entries makes no difference. However, in this case problem (5.3) is computationally tractable. For instance, a set of (unnormalized) eigenvectors is one of the optimizers. We are interested in the sparseness defined in between, namely, a partition with a meso-scale patch size. Compared to $\|\psi\|_0$, the meso-scale partition sacrifices some resolution when measuring the support, but makes the optimization (5.3) efficiently solvable. Specifically, problem (5.3) with the *regular-sparse* partitions (see Definition 1.3.2) enjoys many good properties. These properties enable us to design a very efficient algorithm to solve problem (5.3).

If two intrinsic sparse modes are non-zero on exactly the same set of patches, which are called unidentifiable modes in Definition 5.3.4, it is easy to see that any rotation of these unidentifiable modes forms another set of intrinsic sparse modes. From a theoretical point of view, if a partition is regular-sparse w.r.t. A, the intrinsic sparse modes are unique up to rotations of unidentifiable modes; see Theorem 5.3.5. Moreover, as the partition gets refined, the original identifiable intrinsic sparse modes remain unchanged, while the original unidentifiable modes become identifiable and become sparser (in the sense of l_0 norm); see Theorem 5.3.6. In this sense, the intrinsic sparse modes are independent of the partition that we use. From a computational point of view, a regular-sparse partition ensures that the restrictions of the intrinsic sparse modes on each patch P_m can be constructed from rotations of local eigenvectors. Following this idea, we propose the intrinsic sparse mode decomposition (ISMD); see Algorithm 2. In Theorem 5.3.5, we have proved that the ISMD solves problem (5.3) exactly on regular-sparse partitions. We point out that, even when the partition is not regular-sparse, numerical experiments show that the ISMD still generates a sparse decomposition of A.

The ISMD consists of three steps. In the first step, we perform eigendecomposition of A restricted on local patches $\{P_m\}_{m=1}^M$, denoted as $\{A_{mm}\}_{m=1}^M$, to get $A_{mm} = H_m H_m^T$. Here, columns of H_m are the unnormalized local eigenvectors of A on patch P_m. In the second step, we recover the local pieces

of intrinsic sparse modes, denoted by G_m, by rotating the local eigenvectors $G_m = H_m D_m$. The method to find the right local rotations $\{D_m\}_{m=1}^M$ is the core of the ISMD. All the local rotations are coupled by the decomposition constraint $A = \sum_{k=1}^K g_k g_k^T$ and it seems impossible to solve $\{D_m\}_{m=1}^M$ from this big coupled system. Surprisingly, when the partition is regular-sparse, this coupled system can be decoupled and every local rotation D_m can be solved independently by a joint diagonalization problem (5.13). In the last "patch-up" step, we identify correlated local pieces across different patches by the pivoted Cholesky decomposition of a symmetric PSD matrix Ω and then glue them into a single intrinsic sparse mode. Here, Ω is the projection of A onto the subspace spanned by all the local pieces $\{G_m\}_{m=1}^M$, see Eqn. (5.15). This step is necessary to reduce the number of decomposed modes to the optimal K, i.e., the rank of A. The last step also equips the ISMD with the power to identify long range correlations and to honor the intrinsic correlation structure hidden in A. The popular l_1 approach typically does not have this property.

The ISMD has very low computational complexity. There are two reasons for its efficiency: first of all, instead of computing the expensive global eigendecomposition, we compute only the local eigendecompositions of $\{A_{mm}\}_{m=1}^M$; second, there is an efficient algorithm to solve the joint diagonalization problems for the local rotations $\{D_m\}_{m=1}^M$. Moreover, because both performing the local eigendecompositions and solving the joint diagonalization problems can be done independently on each patch, the ISMD is embarrassingly parallelizable.

The stability of the ISMD is also explored when the input data A is mixed with noise. We study the small perturbation case, i.e., $\widehat{A} = A + \epsilon \widetilde{A}$. Here, A is the noiseless rank-K symmetric PSD matrix, $\widetilde{A}$ is the symmetric additive perturbation, and $\epsilon > 0$ quantifies the noise level. A simple thresholding step is introduced in the ISMD to achieve our aim: *to clean up the noise $\epsilon\widetilde{A}$ and to recover the intrinsic sparse modes of A*. Under some assumptions, we can prove that sparse modes $\{\widehat{g}_k\}_{k=1}^K$, produced by the ISMD with thresholding, exactly capture the supports of A's intrinsic sparse modes $\{g_k\}_{k=1}^K$ and the error $\|\widehat{g}_k - g_k\|$ is small; see Section 5.4 for a precise description.

We have verified all the theoretical predictions with numerical experiments on several synthetic covariance matrices of high dimensional random vectors.

Without parallel execution, for partitions with a large range of patch sizes, the computational cost of the ISMD is comparable to that of the partial eigendecomposition [117, 82]. For certain partitions, the ISMD could be ten times faster than the partial eigendecomposition. We have also implemented the convex relaxation of SPCA [78, 128] and compared these two methods. It turns out that the convex relaxation of SPCA fails to capture the long range correlation, needs to perform (partial) eigendecomposition on matrices repeatedly for many times and is thus much slower than the ISMD. Moreover, we demonstrate the robustness of the ISMD on partitions which are not regular-sparse and on inputs which are polluted with small noise.

Applications

The ISMD leads to a sparse-orthogonal matrix factorization for any matrix. Given a matrix $X \in \mathbb{R}^{N\times M}$ of rank K and a partition $\mathcal{P}$ of the index set $[N]$, the ISMD tries to solve the following optimization problem:

$$\min_{\substack{g_1,\ldots,g_K\in\mathbb{R}^N\\ u_1,\ldots,u_K\in\mathbb{R}^M}} \sum_{k=1}^{K} s(g_k;\mathcal{P}) \quad \text{s.t.} \quad X=\sum_{k=1}^{K} g_k u_k^T, \quad u_k^T u_{k'}=\delta_{k,k'} \qquad \forall 1\le k,k'\le K, \tag{5.4}$$

where $s(g_k;\mathcal{P})$ is the patchwise sparseness defined in Definition (5.1.1). Compared to the biorthogonal property of SVD, the ISMD requires orthogonality only in one dimension and requires sparsity in the other dimension. The method to obtain the decomposition (5.4) consists of three steps: first, compute $A = XX^T$; second, apply the ISMD to A to get $\{g_k\}_{k=1}^K$; third, project X on to $\{g_k\}_{k=1}^K$ to obtain $\{u_k\}_{k=1}^K$.

The sparse-orthogonal matrix factorization (5.4) has potential applications in statistics, machine learning, and uncertainty quantification. In statistics and machine learning, latent factor models with sparse loadings have found many applications ranging from DNA microarray analysis [48], facial and object recognition [129], web search models [1], etc. Specifically, latent factor models decompose a data matrix $X \in \mathbb{R}^{N\times M}$ by product of the loading matrix $G \in \mathbb{R}^{N\times K}$ and the factor value matrix $U \in \mathbb{R}^{M\times K}$, with possibly small noise $E \in \mathbb{R}^{N\times M}$, i.e.,

$$X = GU^T + E. \tag{5.5}$$

The sparse-orthogonal matrix factorization (5.4) tries to find the optimal sparse loadings G under the condition that latent factors are *normalized* and

uncorrelated, i.e., columns in U are orthonormal. In practice, using uncorrelated latent factors makes lots of sense, but is not guaranteed by many existing matrix factorization methods, e.g., non-negative matrix factorization (NMF) [80], SPCA [73, 137, 37], structured SPCA [72].

In uncertainty quantification (UQ), we often need to parametrize a random field, denoted as $\kappa(x, \omega)$, with a finite number of random variables. Applying the ISMD to its covariance function, denoted by $\mathrm{Cov}(x, y)$, we can get a parametrization with K random variables:

$$\kappa(x, \omega) = \bar{\kappa}(x) + \sum_{k=1}^{K} g_k(x)\eta_k(\omega), \tag{5.6}$$

where $\bar{\kappa}(x)$ is the mean field, the physical modes $\{g_k\}_{k=1}^K$ are sparse/localized, and the random variables $\{\eta_k\}_{k=1}^K$ are *centered*, *uncorrelated*, and have *unit variance*. The parametrization (5.6) has a form similar to the widely used Karhenen–Loève (KL) expansion [75, 85], but in the KL expansion the physical modes $\{g_k\}_{k=1}^K$ are eigenfunctions of the covariance function and are typically nonzero everywhere. Obtaining a sparse parametrization is important to uncover the intrinsic sparse features in a random field and to achieve computational efficiency for further scientific experiments. In [65], such sparse parametrization methods are used to design efficient algorithms to solve partial differential equations with random inputs.

Connection With The Sparse Matrix Factorization Problem

Given a matrix $X \in \mathbb{R}^{N\times M}$ of M columns corresponding to M observations in $\mathbb{R}^N$, a sparse matrix factorization problem is to find a matrix $G = [g_1, \ldots, g_r] \in \mathbb{R}^{N\times r}$, called a *dictionary*, and a matrix $U = [u_1, \ldots, u_r] \in \mathbb{R}^{M\times r}$, called *decomposition coefficients*, such that GU^T approximates X well and the columns in G are sparse.

In [81, 131, 88], the authors formulated this problem as an optimization problem by penalizing the l1 norm of G, i.e. $\|G\|_1 := \sum_{k=1}^r \|g_k\|_1$, to enforce the sparsity of the dictionary. This can be written as

$$\boxed{\min_{G\in\mathbb{R}^{N\times r}, U\in\mathbb{R}^{M\times r}} \quad \|X - GU^T\|_F^2 + \lambda\|G\|_1 \quad \text{s.t.} \quad \|u_k\|_2 \le 1 \qquad \forall 1 \le k \le r,} \tag{5.7}$$

where the parameter $\lambda > 0$ controls to what extent the dictionary G is regularized. We point out that the l1 penalty can be replaced by other penalties. For

example, the structured SPCA [72] uses certain l1/l2 norms of G to enforce sparsity with specific structures, e.g. rectangular structure on a grid. Problem (5.7) is not jointly convex in (G, U). Certain specially designed algorithms have been developed to solve this optimization problem. We will discuss one of these methods in Section 5.2.

There are two major differences between the optimization problem (5.4) and the optimization problem (5.7). First, the ISMD, which is designed to solve (5.4), requires that the decomposition coefficients U be orthonormal, while many other methods, including SPCA and structured SPCA, which are designed to solve (5.7), only normalize every columns in U. One needs to decide whether the orthogonality in U is necessary in a specific application and choose the appropriate method. Second, the number of modes K in the ISMD must be the rank of the matrix, while the number of modes r in problem (5.7) is picked by users and can be any number. In other words, the ISMD is seeking an *exact matrix decomposition*, while other methods make a trade-off between the accuracy $\|X - GU^T\|_F$ and the sparsity $\|G\|_1$ by recovering the matrix approximately instead of obtaining an exact recovery. Although the ISMD can be modified to do matrix approximation (with the orthogonality constraint on U), see Algorithm 4, the optimal sparsity of the dictionary G is no longer guaranteed anymore. *Based on these two differences, we recommend the ISMD for sparse matrix factorization problems where the orthogonality in decomposition coefficients U is required and an exact (or nearly exact) decomposition is desired.*

Outline

In Section 5.2 we present our ISMD algorithm for low rank matrices, analyze its computational complexity and talk about its relation with other methods for sparse decomposition or approximation. In Section 5.3 we present our main theoretical results, i.e., Theorem 5.3.5 and Theorem 5.3.6. In Section 5.4, we discuss the stability of the ISMD by performing perturbation analysis. We also provide two modified ISMD algorithms: Algorithm 3 for low rank matrices with small noise, and Algorithm 4 for sparse matrix approximation. Finally, we present a few numerical examples in Section 5.5 to demonstrate the efficiency of the ISMD and compare its performance with other existing methods.

5.2 Intrinsic Sparse Mode Decomposition

In this section, we present the algorithm of the ISMD and analyze its computational complexity. Its relation with other matrix decomposition methods is discussed at the end of this section. In the rest of the chapter, $\mathbb{O}(n)$ denotes the set of real unitary matrices of size $n \times n$; $\mathbb{I}_n$ denotes the identity matrix with size $n \times n$.

The Algorithm Of ISMD

Suppose that we have a symmetric positive symmetric matrix, denoted as $A \in \mathbb{R}^{N\times N}$, and a partition of the index set $[N]$, denoted as $\mathcal{P} = \{P_m\}_{m=1}^M$. The partition typically originates from the physical meaning of the matrix A. For example, if A is the discretized covariance function of a random field on domain $D \subset \mathbb{R}^d$, $\mathcal{P}$ is constructed from certain domain partition of D. The submatrix of A, with row index in P_m and column index in P_n, is denoted as A_{mn}. To simplify our notation, we assume that indices in $[N]$ are rearranged such that A is written as below:

$$A = \left[\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1M} \\ \hline A_{21} & A_{22} & \cdots & A_{2M} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline A_{M1} & A_{M2} & \cdots & A_{MM} \end{array}\right]. \tag{5.8}$$

Notice that when implementing the ISMD, there is no need to rearrange the indices as above. The ISMD tries to find the optimal sparse decomposition of A w.r.t. partition $\mathcal{P}$, defined as the minimizer of problem (5.3). The ISMD consists of three steps: local decomposition, local rotation, and global patch-up.

In the first step, we perform the local eigendecomposition

$$A_{mm} = \sum_{i=1}^{K_m} \gamma_{m,i} h_{m,i} h_{m,i}^T \equiv H_m H_m^T, \tag{5.9}$$

where K_m is the rank of A_{mm} and $H_m = [\gamma_{m,1}^{1/2} h_{m,i}\,, \gamma_{m,2}^{1/2} h_{m,2}\,, \ldots\, \gamma_{m,K_m}^{1/2} h_{m,K_m}]$. If A_{mm} is ill-conditioned, we truncate the small eigenvalues and a truncated eigendecomposition is used as follows:

$$A_{mm} \approx \sum_{i=1}^{K_m} \gamma_{m,i} h_{m,i} h_{m,i}^T \equiv H_m H_m^T. \tag{5.10}$$

Let $K_{(t)} \equiv \sum_{m=1}^{M} K_m$ be the total local rank of A. We extend columns of H_m into $\mathbb{R}^N$ by adding zeros, and get the block diagonal matrix

$$H_{ext} = \text{diag}\{H_1, H_2, \cdots, H_M\}.$$

The correlation matrix with basis H_{ext}, denoted by $\Lambda \in \mathbb{R}^{K_{(t)} \times K_{(t)}}$, is the matrix such that

$$A = H_{ext} \Lambda H_{ext}^T. \tag{5.11}$$

Since columns of H_{ext} are orthogonal and span a space that contains range(A), Λ exists and can be computed blockwise as follows:

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \cdots & \Lambda_{1M} \\ \Lambda_{21} & \Lambda_{22} & \cdots & \Lambda_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{M1} & \Lambda_{M2} & \cdots & \Lambda_{MM} \end{bmatrix}, \quad \Lambda_{mn} = H_m^\dagger A_{mn} \left(H_n^\dagger\right)^T \in \mathbb{R}^{K_m \times K_n}, \tag{5.12}$$

where $H_m^\dagger \equiv (H_m^T H_m)^{-1} H_m^T$ is the (Moore-Penrose) pseudo-inverse of H_m.

In the second step, on every patch P_m, we solve the following joint diagonaliziation problem to find a local rotation D_m:

$$\boxed{D_m = \underset{V \in \mathbb{O}(K_m)}{\arg\min} \quad \sum_{n=1}^{M} \sum_{i \neq j} |(V^T \Sigma_{n;m} V)_{i,j}|^2,} \tag{5.13}$$

in which

$$\Sigma_{n;m} \equiv \Lambda_{mn} \Lambda_{mn}^T. \tag{5.14}$$

We rotate the local eigenvectors with D_m and get $G_m = H_m D_m$. Again, we extend columns of G_m into $\mathbb{R}^N$ by adding zeros, and get the block diagonal matrix

$$G_{ext} = \text{diag}\{G_1, G_2, \cdots, G_M\}.$$

The correlation matrix with basis G, denoted by $\Omega \in \mathbb{R}^{K_{(t)} \times K_{(t)}}$, is the matrix such that

$$A = G_{ext} \Omega G_{ext}^T. \tag{5.15}$$

With Λ in hand, Ω can be obtained as follows:

$$\Omega = D^T \Lambda D, \qquad D = \text{diag}\{D_1, D_2, \cdots, D_M\}. \tag{5.16}$$

Joint diagonalization has been well studied in the blind source separation (BSS) community. We present some relevant theoretical results in supplementary materials B.2. A Jacobi-like algorithm [23, 17], see Algorithm 5, is used in our chapter to solve problem (5.13). For most cases, we may want to normalize the columns of G_{ext} and put all the magnitude information in Ω, i.e.,

$$G_{ext} = \bar{G}_{ext}E, \quad \bar{\Omega} = E\Omega E^T, \tag{5.17}$$

where E is a diagonal matrix with E_{ii} being the l_2 norm of the i-th column of G_{ext}, $\bar{G}_{ext}$ and $\bar{\Omega}$ will substitute the roles of G and Ω in the rest of the algorithm.

In the third step, we use the pivoted Cholesky decomposition to patch up the local pieces G_m. Specifically, suppose the pivoted Cholesky decomposition of Ω is given as

$$\Omega = PLL^TP^T\,, \tag{5.18}$$

where $P \in \mathbb{R}^{K_{(t)}\times K_{(t)}}$ is a permutation matrix and $L \in \mathbb{R}^{K_{(t)}\times K}$ is a lower triangular matrix with positive diagonal entries. Since A has rank K, both Λ and Ω have rank K. This is why L only has K nonzero columns. However, we point out that the rank K is automatically identified in the algorithm instead of given as an input parameter. Finally, A is decomposed as

$$A = GG^T \equiv G_{ext}PL(G_{ext}PL)^T\,. \tag{5.19}$$

The columns in $G = G_{ext}PL$ are our decomposed sparse modes.

The full algorithm is summarized in Algorithm 2. We point out that there are two extreme cases for the ISMD:

- The coarsest partition $\mathcal{P} = \{[N]\}$. In this case, the ISMD is equivalent to the standard eigendecomposition.
- The finest partition $\mathcal{P} = \{\{i\} : i \in [N]\}$. In this case, the ISMD is equivalent to the pivoted Cholesky factorization on $\bar{A}$ where $\bar{A}_{ij} = \frac{A_{ij}}{\sqrt{A_{ii}A_{jj}}}$. If the normalization (5.17) is applied, the ISMD is equivalent to the pivoted Cholesky factorization of A in this case.

In these two extreme cases, there is no need to use the joint diagonalization step and it is known that, in general, neither the ISMD nor the pivoted

Cholesky decomposition generates a sparse decomposition. When $\mathcal{P}$ is neither of these two extreme cases, the joint diagonalization is applied to rotate the local eigenvectors and thereafter the generated modes are patchwise sparse. Specifically, when the partition is regular-sparse, the ISMD generates the optimal patchwise sparse decomposition as stated in Theorem 5.3.5.

Algorithm 2 Intrinsic sparse mode decomposition

Require: $A \in \mathbb{R}^{N\times N}$: symmetric and PSD; $\mathcal{P} = \{P_m\}_{m=1}^{M}$: partition of index set $[N]$

Ensure: $G = [g_1, g_2, \cdots, g_K]$: K is the rank of A, $A = GG^T$

###Local eigendecomposition
for $m = 1, 2, \cdots, M$ **do**
    Local eigendecomposition: $A_{mm} = H_m H_m^T$
end for
###Assemble correlation matrix Λ
Assemble $\Lambda = H_{ext}^{\dagger} A \left(H_{ext}^{\dagger}\right)^T$ blockwisely as in Eqn. (5.12)
###Joint Diagonalization
for $m = 1, 2, \cdots, M$ **do**
    for $n = 1, 2, \cdots, M$ **do**
        $\Sigma_{n;m} = \Lambda_{mn}\Lambda_{mn}^T$
    end for
    Solve the joint diagonalization problem (5.13) for D_m ▷ Use Algorithm 5
end for
###Assemble correlation matrix Ω and its pivoted Cholesky decomposition
$\Omega = D^T \Lambda D$
$\Omega = PLL^T P^T$
###Assemble the intrinsic sparse modes G
$G = H_{ext} DPL$

Remark 5.2.1. *One can interpret H_m as the patchwise amplitude and D_m as the patchwise phase. The patchwise amplitude is easy to obtain using a local eigendecomposition* (5.9)*, while the patchwise phase is obtained by the joint diagonalization* (5.13)*.*

In fact, the ISMD solves the following optimization problem where we jointly

diagonalize A_{mn}*:*

$$\boxed{\begin{aligned} \min_{G_m \in \mathbb{R}^{|P_m| \times K_m}} \quad & \sum_{n=1}^{M} \sum_{i \neq j} |B_{n;m}(i,j)|^2 \\ s.t. \quad & G_m G_m^T = A_{mm}\,, \\ & G_m B_{n;m} G_m^T = A_{mn} A_{nn}^{\dagger} A_{mn}^T, \end{aligned}} \tag{5.20}$$

in which $A_{nn}^{\dagger} = \sum_{i=1}^{K_n} \gamma_{n,i}^{-1} h_{n,i} h_{n,i}^T$ *is the (Moore–Penrose) pseudo-inverse of* A_{nn}*. Eqn.* (5.20) *is not a unitary joint diagonalization problem, i.e., the variable* G_m *is not unitary. The ISMD solves this non-unitary joint diagonalization problem in two steps:*

1. *Perform a local eigendecomposition* $A_{mm} = H_m H_m^T$*. Then the feasible* G_m *can be written as* $H_m D_m$ *with a unitary matrix* D_m*.*

2. *Find the rotation* D_m *that solves the unitary joint diagonalization problem* (5.13)*.*

Computational Complexity

The main computational cost of the ISMD comes from the local KL expansion, the joint diagonalization, and the pivoted Cholesky decomposition. To simplify the analysis, we assume that the partition $\mathcal{P}$ is uniform, i.e., each group has $\frac{N}{M}$ nodes. On each patch, we perform the eigendecomposition of A_{mm} of size N/M and rank K_m. Then, the cost of the local eigendecomposition step is

$$\text{Cost}_1 = \sum_{m=1}^{M} \mathcal{O}\left((N/M)^2 K_m\right) = (N/M)^2 \mathcal{O}(\sum_{m=1}^{M} K_m).$$

For the joint diagonalization, the computational cost of Algorithm 5 is

$$\sum_{m=1}^{M} N_{corr,m} K_m^3 N_{iter,m}\,.$$

Here, $N_{corr,m}$ is the number of nonzero matrices in $\{\Sigma_{n;m}\}_{n=1}^{M}$. Notice that $\Sigma_{n;m} \equiv \Lambda_{mn} \Lambda_{mn}^T = 0$ if and only if $A_{mn} = 0$. Therefore, $N_{corr,m}$ may be much smaller than M if A is sparse. Nevertheless, we take an upper bound M to estimate the cost. $N_{corr,m} K_m^3$ is the computational cost for each sweep in Algorithm 5 and $N_{iter,m}$ is the number of iterations needed for convergence. The asymptotic convergence rate is shown to be quadratic [17], and no more

than six iterations are needed in our numerical examples. Therefore, we can take $N_{iter,m} = \mathcal{O}(1)$ and in total we have

$$\text{Cost}_2 = \sum_{m=1}^{M} M\mathcal{O}(K_m^3) = M\mathcal{O}(\sum_{m=1}^{M} K_m^3).$$

Finally, the pivoted Cholesky decomposition of Ω, which is of size $\sum_{k=1}^{M} K_m$, has cost

$$\text{Cost}_3 = \mathcal{O}\left((\sum_{k=1}^{M} K_m)K^2\right) = K^2\mathcal{O}(\sum_{m=1}^{M} K_m).$$

Combining the computational costs in all three steps, we conclude that the total computational cost of the ISMD is

$$\text{Cost}_{\text{ISMD}} = \left((N/M)^2 + K^2\right)\mathcal{O}(\sum_{m=1}^{M} K_m) + M\mathcal{O}(\sum_{m=1}^{M} K_m^3)\,. \tag{5.21}$$

Making use of $K_m \leq K$, we have an upper bound for Cost_{ISMD},

$$\text{Cost}_{\text{ISMD}} \leq \mathcal{O}(N^2K/M) + \mathcal{O}(M^2K^3)\,. \tag{5.22}$$

When $M = \mathcal{O}((N/K)^{2/3})$, $\text{Cost}_{\text{ISMD}} \leq \mathcal{O}(N^{4/3}K^{5/3})$. Compared with the cost of partial eigendecomposition [117, 82], which is about $\mathcal{O}(N^2K)$ [1], the ISMD is more efficient for low-rank matrices.

For matrix A which has a sparse decomposition, the local ranks K_m are much smaller than its global rank K. An extreme case is $K_m = \mathcal{O}(1)$, which is, in fact, true for many random fields; see [28, 65]. In this case,

$$\text{Cost}_{\text{ISMD}} = \mathcal{O}(N^2/M) + \mathcal{O}(M^2) + \mathcal{O}(MK^2)\,. \tag{5.23}$$

When the partition gets finer (M increases), the computational cost first decreases due to the saving in local eigendecompositions. The computational cost achieves its minimum around $M = \mathcal{O}(N^{2/3})$ and then increases due to the increasing cost for the joint diagonalization. This trend is observed in our numerical examples; see Figure 5.4.

We point out that the M local eigendecompositions (5.9) and the joint diagonalization problems (5.13) are solved independently on different patches. Therefore, our algorithm is embarrassingly parallelizable. This will save the computational cost in the first two steps by a factor of M, which makes the ISMD even faster.

[1] The cost can be reduced to $\mathcal{O}(N^2 \log(K))$ if a randomized SVD with some specific technique is applied.

Connection With Other Matrix Decomposition Methods

Sparse decompositions of symmetric PSD matrices have been studied in different fields for a long time. There are, in general, two approaches to achieve sparsity: rotation or l_1 minimization.

The rotation approach begins with eigenvectors. Suppose that we have decided to retain and rotate K eigenvectors. Define $H = [h_1, h_2, \ldots, h_K]$ with h_k being the k'th eigenvector. We postmultiply H by a matrix $T \in \mathbb{R}^{K\times K}$ to obtain the rotated modes $G = [g_1, g_2, \ldots, g_K] = HT$. The choice of T is determined by the rotation criterion we use. In data science, for the commonly used varimax rotation criterion [77, 74], T is an orthogonal matrix chosen to maximize the variance of squared modes within each column of G. This drives entries in G towards 0 or ± 1. In quantum chemistry, every column in H and G corresponds to a function over a physical domain D and certain specialized sparse modes (localized modes) are sought after. The most widely used criterion to achieve maximally localized modes is the one proposed in [90]. This criterion requires T to be unitary, and then minimizes the second moment:

$$\sum_{k=1}^{K} \int_D (x - x_k)^2 |g_k(x)|^2 \mathrm{d}x \,, \tag{5.24}$$

where $x_k = \int_D x|g_k(x)|^2 \mathrm{d}x$. More recently, a method weighted by higher degree polynomials is discussed in [42]. While these criteria work reasonably well for simple symmetric PSD functions/operators, they all suffer from non-convex optimization, which requires a good starting point to converge to the global minimum. In addition, these methods only care about the eigenspace spanned by H instead of the specific matrix decomposition, and thus they cannot be directly applied to solve our problem (5.3).

The ISMD proposed in this chapter follows the rotation approach. The ISMD implicitly finds a unitary matrix $T \in \mathbb{R}^{K\times K}$ to construct the intrinsic sparse modes

$$[g_1, g_2, \ldots, g_K] = [\sqrt{\lambda_1}h_1, \sqrt{\lambda_2}h_2, \ldots, \sqrt{\lambda_K}h_K]\, T. \tag{5.25}$$

Notice that we rotate the unnormalized eigenvector $\sqrt{\lambda_k}h_k$ to satisfy the decomposition constraint $A = \sum_{k=1}^{K} g_k g_k^T$. The criterion of the ISMD is to minimize the total patchwise sparseness as in (5.3). The success of the ISMD lies in the fact that as long as the domain partition is regular-sparse, the optimization problem (5.3) can be exactly and efficiently solved by Algorithm 2.

Moreover, the intrinsic sparse modes produced by the ISMD are optimally localized because we are directly minimizing the total patchwise sparseness of $\{g_k\}_{k=1}^K$.

The l_1 minimization approach, pioneered by ScotLass [73], has a rich literature in solving the sparse matrix factorization problem (5.7); see [137, 37, 136, 128, 107, 78]. Problem (5.7) is highly non-convex in (G, U), and there has been a lot of effort (see e.g. [37, 128, 78]) in relaxing it to a convex optimization. First of all, since there are no essential constraints on U, one can get rid of U by considering the variational form [73, 137, 107]:

$$\boxed{\min_{G\in\mathbb{R}^{N\times K}} \quad -\mathrm{Tr}(G^TAG)+\mu\|G\|_1 \quad \text{s.t.} \quad G^TG=\mathbb{I}_K,} \tag{5.26}$$

where $A = XX^T$ is the covariance matrix as in the ISMD (5.3) and Tr is the trace operator on square matrices. Notice that the problem is still non-convex due to the orthogonality constraint $G^TG = \mathbb{I}_K$. In the second step, the authors in [128] proposed the following semidefinite programming to obtain the sparse density matrix $W \in \mathbb{R}^{n\times n}$, which plays the same role as GG^T in (5.26):

$$\boxed{\min_{W\in\mathbb{R}^{N\times N}} \quad -\mathrm{Tr}(AW)+\mu\|W\|_1 \quad \text{s.t.} \quad 0\preceq W\preceq \mathbb{I}_N,\ \mathrm{Tr}(W)=K.} \tag{5.27}$$

Here, $0 \preceq W \preceq \mathbb{I}_N$ means that both W and $\mathbb{I}_N - W$ are symmetric and positive semidefinite. Finally, the first K eigenvectors of W are used as the sparse modes G. An equivalent formulation was proposed in [78], and the authors proposed to pick K columns of W as the sparse modes G.

We will compare the advantages and disadvantages of the ISMD and the convex relaxation of SPCA in Section 5.5.

5.3 Theoretical Results With Regular-Sparse Partitions

In this section, we present the main theoretical results of the ISMD, i.e., Theorem 5.3.5, Theorem 5.3.6 and its perturbation analysis. We first introduce a domain-decomposition type presentation of any feasible decomposition $A = \sum_{k=1}^K \psi_k\psi_k^T$. Then we discuss the regular-sparse property and use it to prove our main results. When no ambiguity arises, we denote patchwise sparseness $s(g_k; \mathcal{P})$ as s_k.

A Domain-decomposition Type Representation

For an arbitrary decomposition $A = \sum_{k=1}^K \psi_k\psi_k^T$, denote $\Psi \equiv [\psi_1, \ldots, \psi_K]$ and $\Psi|_{P_m} \equiv [\psi_1|_{P_m}, \ldots, \psi_K|_{P_m}]$. For a sparse decomposition, we expect that most

columns in $\Psi|_{P_m}$ are zero, and thus we define the local dimension on patch P_m as follows.

Definition 5.3.1 (Local dimension)**.** *The local dimension of a decomposition $A = \sum_{k=1}^{K} \psi_k \psi_k^T$ on patch P_m is the number of nonzero modes when restricted to this patch, i.e.,*

$$d(P_m; \Psi) = |S_m|, \qquad S_m = \{k \ : \ \psi_k|_{P_m} \neq 0\}.$$

When no ambiguity arises, $d(P_m; \Psi)$ is written as d_m. We enumerate all the elements in S_m as $\{k_i^m\}_{i=1}^{d_m}$, and group together all the nonzero local pieces on patch P_m and obtain

$$\Psi_m \equiv [\psi_{m,1}, \ldots, \psi_{m,d_m}], \qquad \psi_{k_i^m}|_{P_m} = \psi_{m,i}. \tag{5.28}$$

Therefore, we have

$$\Psi|_{P_m} = \Psi_m L_m^{(\psi)}, \tag{5.29}$$

where $L_m^{(\psi)}$ is a matrix of size $d_m \times K$ with the k_i^m-th column being $\boldsymbol{e}_i$ for $i \in [d_m]$ and other columns being $\mathbf{0}$. Here, $\boldsymbol{e}_i$ is the i-th column of $\mathbb{I}_{d_m}$. $L_m^{(\psi)}$ is called the *local indicator matrix* of Ψ on patch P_m. Restricting the decomposition constraint $A = \Psi\Psi^T$ to patch P_m, we have $A_{mm} = \Psi|_{P_m} \left(\Psi|_{P_m}\right)^T$, where A_{mm} is the restriction of A on patch P_m, as in (5.8). Since Ψ_m is obtained from $\Psi|_{P_m}$ by deleting zero columns, we have

$$A_{mm} = \Psi_m \Psi_m^T. \tag{5.30}$$

We stack up Ψ_m and $L_m^{(\psi)}$ as follows,

$$\Psi_{ext} \equiv \mathrm{diag}\{\Psi_1, \Psi_2, \cdots, \Psi_M\}, \quad L^{(\psi)} \equiv \left[L_1^{(\psi)}; L_2^{(\psi)}; \cdots; L_M^{(\psi)}\right],$$

and then we have:

$$\Psi = [\Psi|_{P_1}; \ldots; \Psi|_{P_M}] = \Psi_{ext} L^{(\psi)}. \tag{5.31}$$

The intuition in Eqn. (5.31) is that the local pieces Ψ_m are linked together by the indicator matrix $L^{(\psi)}$ and the modes Ψ on the entire domain $[N]$ can be recovered from Ψ_{ext} and $L^{(\psi)}$. We call $L^{(\psi)}$ the *indicator matrix* of Ψ.

We use a simple example to illustrate the patchwise sparseness, the local dimension and Eqn. (5.31). In this case, $\Psi \in \mathbb{R}^{N \times K}$ $(N = 100, K = 2)$ is the

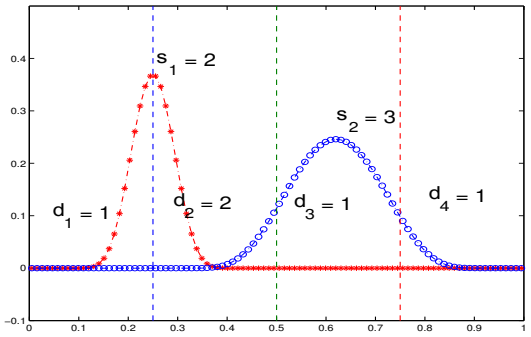


Figure 5.1: Illustration of sparseness, local dimension and $\Psi = \Psi_{ext} L^{(\psi)}$.

discretized version of two functions on $[0,1]$ and $\mathcal{P}$ partitions $[0,1]$ uniformly into four intervals as shown in Figure 5.3. ψ_1, the red starred mode, is nonzero on the left two patches and ψ_2, the blue circled mode, is nonzero on the right three patches. The sparseness of ψ_1 is 2, the sparseness of ψ_2 is 3, and the local dimensions of the four patches are 1, 2, 1, and 1 respectively, as we comment in Figure 5.3. Following the definitions above, we have $\Psi_1 = \psi_1|_{P_1}$, $L_1^{(\psi)} = [1,0]$, $\Psi_2 = [\psi_1|_{P_2}, \psi_2|_{P_2}]$, $L_2^{(\psi)} = [1,0;0,1]$, $\Psi_3 = \psi_2|_{P_3}$, $L_3^{(\psi)} = [0,1]$, $\Psi_4 = \psi_2|_{P_4}$, and $L_4^{(\psi)} = [0,1]$. Finally, we get

$$[\psi_1, \psi_2] = \Psi_{ext} L^{(\psi)} \equiv \left[\begin{array}{c|cc|c|c} \psi_{1,1} & 0 & 0 & 0 & 0 \\ \hline 0 & \psi_{1,2} & \psi_{2,2} & 0 & 0 \\ \hline 0 & 0 & 0 & \psi_{2,3} & 0 \\ \hline 0 & 0 & 0 & 0 & \psi_{2,4} \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ \hline 1 & 0 \\ 0 & 1 \\ \hline 0 & 1 \\ \hline 0 & 1 \end{array}\right].$$

With this domain-decomposition type representation of Ψ, the decomposition constraint is rewritten as:

$$A = \Psi\Psi^T = \Psi_{ext}\Omega^{(\psi)}\Psi_{ext}^T, \qquad \Omega^{(\psi)} \equiv L^{(\psi)} \left(L^{(\psi)}\right)^T. \tag{5.32}$$

Here, $\Omega^{(\psi)}$ has a role similar to that of Ω in the ISMD. It can be viewed as the correlation matrix of A under basis Ψ_{ext}, just like how Λ and Ω are defined.

Finally, we provide two useful properties of the local indicator matrices $L_m^{(\psi)}$, which are direct consequences of their definitions.

Proposition 5.3.2. *For an arbitrary decomposition $A = \Psi\Psi^T$,*

1. *The k-th column of $L^{(\psi)}$, denoted as $l_k^{(\psi)}$, satisfies $\|l_k^{(\psi)}\|_1 = s_k$ where s_k is the patchwise sparseness of ψ_k, as in Definition 5.1.1. Moreover, different columns in $L^{(\psi)}$ have disjoint supports.*

2. *Define*
$$B_{n;m}^{(\psi)} \equiv \Omega_{mn}^{(\psi)} \left(\Omega_{mn}^{(\psi)}\right)^T, \tag{5.33}$$
where $\Omega_{mn}^{(\psi)} \equiv L_m^{(\psi)}(L_n^{(\psi)})^T$ is the (m,n)-th block of $\Omega^{(\psi)}$. $B_{n;m}^{(\psi)}$ is diagonal with diagonal entries either 1 or 0. Moreover, $B_{n;m}^{(\psi)}(i,i) = 1$ if and only if there exists $k \in [K]$ such that $\psi_k|_{P_m} = \psi_{m,i}$ and $\psi_k|_{P_n} \neq \mathbf{0}$.

Proof. 1. $l_k^{(\psi)}$, divided into patches, can be written as $l_k^{(\psi)} = [l_{1,k}; l_{2,k}; \cdots; l_{M,k}]$. From the definition (5.29), we have $\|l_{m,k}\|_1 = 1$ if $\psi_k|_{P_m} \neq \mathbf{0}$ and 0 otherwise. Therefore, we obtain
$$\|l_k^{(\psi)}\|_1 = \sum_{m=1}^{M} \|l_{m,k}\|_1 = s_k(\psi_k; \mathcal{P}).$$
Moreover, on patch P_m different ψ_k's correspond to different local pieces in Ψ_m (when they are identical, we keep both when constructing Ψ_m), and thus different columns in $L_m^{(\psi)}$ have disjoint supports. Therefore, different columns in $L^{(\psi)}$ have disjoint supports.

2. From the definition (5.29), the j-th row of $L_n^{(\psi)}$ is equal to $\boldsymbol{e}_{k_j^m}^T$, where $\boldsymbol{e}_{k_j^m}$ is the k_j^m-th column of $\mathbb{I}_K$. Then we have $(L_n^{(\psi)})^T L_n^{(\psi)} = \sum_{j=1}^{d_n} \boldsymbol{e}_{k_j^n} \boldsymbol{e}_{k_j^n}^T$. Therefore, we obtain
$$B_{n;m}^{(\psi)} \equiv L_m^{(\psi)}(L_n^{(\psi)})^T L_n^{(\psi)}(L_m^{(\psi)})^T = \sum_{j=1}^{d_n} L_m^{(\psi)} \boldsymbol{e}_{k_j^n} (L_m^{(\psi)} \boldsymbol{e}_{k_j^n})^T = \sum_{j=1}^{d_n} l_{m,k_j^n} l_{m,k_j^n}^T, \tag{5.34}$$
where l_{m,k_j^n} is the k_j^n-th column of $L_m^{(\psi)}$.

From the definition (5.29), l_{m,k_i^m}, the k_i^m-th column of $L_m^{(\psi)}$, is equal to $\boldsymbol{e}_i$ for $i \in [d_m]$ and all other columns are $\mathbf{0}$. Therefore,
$$\sum_{k=1}^{K} l_{m,k} l_{m,k}^T = \sum_{i=1}^{d_m} l_{m,k_i^m} l_{m,k_i^m}^T = \sum_{i=1}^{d_m} \boldsymbol{e}_i \boldsymbol{e}_i^T = \mathbb{I}_{d_m}. \tag{5.35}$$

Eqn. (5.34) sums over $k \in \{k_j^n\}_{j=1}^{d_n} \subset [K]$ and then we conclude that $B_{n;m}^{(\psi)}$ is diagonal with diagonal entries either 1 or 0. Moreover, if $B_{n;m}^{(\psi)}(i,i) = 1$ the term $\boldsymbol{e}_i \boldsymbol{e}_i^T$ has to be included in the summation in (5.34). Among

all terms $\{l_{m,k}l_{m,k}^T\}_{k=1}^K$, only $l_{m,k_i^m}l_{m,k_i^m}^T$ is equal to $\boldsymbol{e}_i\boldsymbol{e}_i^T$ due to the definition of $L_m^{(\psi)}$. Therefore, the term $l_{m,k_i^m}l_{m,k_i^m}^T$ has to be included in the summation in (5.34). Therefore, there exists $j \in [d_n]$ such that $k_j^n = k_i^m$. In other words, there exist $k \in [K]$ and $j \in [d_n]$ such that $\psi_k|_{P_m} = \psi_{m,i}$ and $\psi_k|_{P_n} = \psi_{n,j}$.

□

Since different columns in $L^{(\psi)}$ have disjoint supports, $\Omega^{(\psi)} \equiv L^{(\psi)}\left(L^{(\psi)}\right)^T$ has a block-diagonal structure with K blocks. The k-th diagonal block is the one contributed by $l_k^{(\psi)}\left(l_k^{(\psi)}\right)^T$. Therefore, as long as we obtain $\Omega^{(\psi)}$, we can use the pivoted Cholesky decomposition to efficiently recover $L^{(\psi)}$. The ISMD follows this rationale: we first construct local pieces $\Psi_{ext} \equiv \mathrm{diag}\{\Psi_1, \Psi_2, \cdots, \Psi_M\}$ for a certain set of intrinsic sparse modes Ψ. Then from the decomposition constraint (5.32) we are able to compute $\Omega^{(\psi)}$. Finally, the pivoted Cholesky decomposition is applied to obtain $L^{(\psi)}$ and the modes are assembled by $\Psi = \Psi_{ext}L^{(\psi)}$. Obviously, the key step is to construct Ψ_{ext}, which are local pieces of a set of intrinsic sparse modes; this is exactly where the regular-sparse property and the joint diagonalization come into play.

Regular-sparse Property and Local Modes Construction

In this and the next subsections (Section 5.3 and 5.3), we assume that the submatrices A_{mm} are well conditioned and thus the exact local eigendecomposition (5.9) is used in the ISMD.

Combining the local eigendecomposition (5.9) and local decomposition constraint (5.30), there exists $D_m^{(\psi)} \in \mathbb{R}^{K_m \times d_m}$ such that

$$\Psi_m = H_m D_m^{(\psi)}. \tag{5.36}$$

Moreover, since the local eigenvectors are linearly independent, we have

$$d_m \geq K_m\,, \qquad D_m^{(\psi)}\left(D_m^{(\psi)}\right)^T = \mathbb{I}_{K_m}. \tag{5.37}$$

We see that $d_m = K_m$ if and only if columns in Ψ_m are also linearly independent. In this case, $D_m^{(\psi)}$ is unitary, i.e., $D_m^{(\psi)} \in \mathbb{O}(K_m)$. This is exactly what is required by the regular-sparse property, see Definition 1.3.2. It is easy to see that we have the following equivalent definitions of regular-sparse property.

Proposition 5.3.3. *The following assertions are equivalent.*

1. *The partition $\mathcal{P}$ is regular-sparse w.r.t. A.*
2. *There exists a decomposition $A = \sum_{k=1}^{K} \psi_k \psi_k^T$ such that on every patch P_m its local dimension d_m is equal to the local rank K_m, i.e., $d_m = K_m$.*
3. *The minimum of problem* (5.3) *is $\sum_{m=1}^{M} K_m$.*

The proof is elementary and is omitted here. By Proposition 5.3.3, for regular-sparse partitions, local pieces of a set of intrinsic sparse modes can be constructed from rotating local eigenvectors, i.e., $\Psi_m = H_m D_m^{(\psi)}$. All the local rotations $\{D_m^{(\psi)}\}_{m=1}^{M}$ are coupled by the decomposition constraint $A = \Psi\Psi^T$. At first glance, it seems impossible to find such D_m from this big coupled system. However, the following lemma gives a necessary condition that $D_m^{(\psi)}$ must satisfy so that $H_m D_m^{(\psi)}$ are local pieces of a set of intrinsic sparse modes. More importantly, this necessary condition turns out to be sufficient, and thus provides us a criterion to find the local rotations.

Lemma 5.3.1. *Suppose that $\mathcal{P}$ is regular-sparse w.r.t. A and that $\{\psi_k\}_{k=1}^{K}$ is an arbitrary set of intrinsic sparse modes. Denote the transformation from H_m to Ψ_m as $D_m^{(\psi)}$, i.e., $\Psi_m = H_m D_m^{(\psi)}$. Then $D_m^{(\psi)}$ is unitary and jointly diagonalizes $\{\Sigma_{n;m}\}_{n=1}^{M}$, which are defined in* (5.14)*. Specifically, we have*

$$B_{n;m}^{(\psi)} = \left(D_m^{(\psi)}\right)^T \Sigma_{n;m} D_m^{(\psi)}, \quad m = 1, 2, \ldots, M, \tag{5.38}$$

where $B_{n;m}^{(\psi)} \equiv \Omega_{mn}^{(\psi)} \left(\Omega_{mn}^{(\psi)}\right)^T$, defined in (5.33)*, is diagonal with diagonal entries either 0 or 1.*

Proof. From item 3 in Proposition 5.3.3, any set of intrinsic sparse modes must have local dimension $d_m = K_m$ on patch P_m. Therefore, the transformation $D_m^{(\psi)}$ from H_m to Ψ_m must be unitary. Combining $\Psi_m = H_m D_m^{(\psi)}$ with the decomposition constraint (5.32), we get

$$A = H_{ext} D^{(\psi)} \Omega^{(\psi)} \left(D^{(\psi)}\right)^T H_{ext},$$

where $D^{(\psi)} = \mathrm{diag}\{D_1^{(\psi)}, D_2^{(\psi)}, \ldots, D_M^{(\psi)}\}$. Recall that $A = H_{ext} \Lambda H_{ext}$ and that H_{ext} has linearly independent columns. We obtain

$$\Lambda = D^{(\psi)} \Omega^{(\psi)} \left(D^{(\psi)}\right)^T, \tag{5.39}$$

or blockwisely,

$$\Lambda_{mn} = D_m^{(\psi)} \Omega_{mn}^{(\psi)} \left(D_n^{(\psi)}\right)^T. \tag{5.40}$$

Since $D_n^{(\psi)}$ is unitary, Eqn. (5.38) naturally follows the definitions of $B_{n;m}^{(\psi)}$ and $\Sigma_{n;m}$. By item 2 in Proposition 5.3.2, we know that $B_{n;m}^{(\psi)}$ is diagonal with diagonal entries either 0 or 1. □

Lemma 5.3.1 guarantees that $D_m^{(\psi)}$ for an arbitrary set of intrinsic sparse modes is the minimizer of the joint diagonalization problem (5.13). In the other direction, the following lemma guarantees that any minimizer of the joint diagonalization problem (5.13), denoted as D_m, transforms local eigenvectors H_m to G_m, which are the local pieces of certain intrinsic sparse modes.

Lemma 5.3.2. *Suppose that $\mathcal{P}$ is regular-sparse w.r.t. A and that D_m is a minimizer of the joint diagonalization problem* (5.13)*. As in the ISMD, define $G_m = H_m D_m$. Then there exists a set of intrinsic sparse modes such that its local pieces on patch P_m are equal to G_m.*

Before we prove this lemma, we examine the uniqueness property of intrinsic sparse modes. It is easy to see that permutations and sign flips of a set of intrinsic sparse modes are still a set of intrinsic sparse modes. Specifically, if $\{\psi_k\}_{k=1}^K$ is a set of intrinsic sparse modes and $\sigma : [K] \to [K]$ is a permutation, $\{\pm\psi_{\sigma(k)}\}_{k=1}^K$ is another set of intrinsic sparse modes. Another kind of non-uniqueness comes from the following concept–identifiability.

Definition 5.3.4 (Identifiability)**.** *For two modes $g_1, g_2 \in \mathbb{R}^N$, they are unidentifiable on partition $\mathcal{P}$ if they are supported on the same patches, i.e., $\{P \in \mathcal{P} : g_1|_P \neq \mathbf{0}\} = \{P \in \mathcal{P} : g_2|_P \neq \mathbf{0}\}$. Otherwise, they are identifiable. For a collection of modes $\{g_i\}_{i=1}^k \subset \mathbb{R}^N$, they are unidentifiable if and only if any pair of them are unidentifiable. They are pair-wisely identifiable if and only if any pair of them are identifiable.*

It is important to point out that the identifiability above is based on the resolution of partition $\mathcal{P}$. Unidentifiable modes for partition $\mathcal{P}$ may have different supports and become identifiable on a refined partition. Unidentifiable intrinsic sparse modes lead to another kind of non-uniqueness for intrinsic sparse modes. For instance, when two intrinsic sparse modes ψ_m and ψ_n are unidentifiable, then any rotation of $[\psi_m, \psi_n]$ while keeping other intrinsic sparse modes unchanged is still a set of intrinsic sparse modes.

Local pieces of intrinsic sparse modes inherit this kind of non-uniqueness. Suppose $\Psi_m \equiv [\psi_{m,1}, \ldots, \psi_{m,d_m}]$ are the local pieces of a set of intrinsic sparse

modes Ψ on patch P_m. First, if $\sigma : [d_m] \to [d_m]$ is a permutation, $\{\pm\psi_{m,\sigma(i)}\}_{i=1}^{d_m}$ are local pieces of another set of intrinsic sparse modes. Second, if $\psi_{m,i}$ and $\psi_{m,j}$ are the local pieces of two unidentifiable intrinsic sparse modes, then any rotation of $[\psi_{m,i}, \psi_{m,j}]$ while keeping other local pieces unchanged are local pieces of another set of intrinsic sparse modes. It turns out that this kind of non-uniqueness has a one-to-one correspondence with the non-uniqueness of joint diagonalizers for problem (5.13), which is characterized in Theorem B.2.1. Keeping this correspondence in mind, the proof of Lemma 5.3.2 is quite intuitive.

Proof. **[Proof of Lemma 5.3.2]** Let $\Psi \equiv [\psi_1, \ldots, \psi_K]$ be an arbitrary set of intrinsic sparse modes. We order columns in Ψ such that unidentifiable modes are grouped together, denoted as $\Psi = [\Psi_1, \ldots, \Psi_Q]$, where Q is the number of unidentifiable groups. Accordingly on patch P_m, $\Psi_m = [\Psi_{m,1}, \ldots, \Psi_{m,Q_m}]$ where Q_m is the number of nonzero unidentifiable groups. Denote the number of columns in each group as $n_{m,i}$, i.e., there are $n_{m,i}$ modes in $\{\psi_k\}_{k=1}^K$ that are nonzero and unidentifiable on patch P_m.

Making use of item 2 in Proposition 5.3.2, one can check that $\psi_{m,i}$ and $\psi_{m,j}$ are unidentifiable if and only if $B_{n;m}^{(\psi)}(i,i) = B_{n;m}^{(\psi)}(j,j)$ for all $n \in [M]$. Since unidentifiable pieces in Ψ_m are grouped together, the same diagonal entries in $\{B_{n;m}^{(\psi)}\}_{n=1}^M$ are grouped together as required in Theorem B.2.1. Now we apply Theorem B.2.1 with M_k replaced by $\Sigma_{n;m}$, Λ_k replaced by $B_{n;m}^{(\psi)}$, D replaced by $D_m^{(\psi)}$, the number of distinct eigenvalues m replaced by Q_m, eigenvalue's multiplicity q_i replaced by $n_{m,i}$ and the diagonalizer V replaced by D_m. Therefore, there exists a permutation matrix Π_m and a block diagonal matrix V_m such that

$$D_m \Pi_m = D_m^{(\psi)} V_m \,, \qquad V_m = \mathrm{diag}\{V_{m,1}, \ldots, V_{m,Q_m}\} \,. \tag{5.41}$$

Recall that $G_m = H_m D_m$ and $\Psi_m = H_m D_m^{(\psi)}$, we obtain that

$$G_m \Pi_m = \Psi_m V_m = [\Psi_{m,1} V_{m,1} \,, \ldots \,, \Psi_{m,Q_m} V_{m,Q_m}] \,. \tag{5.42}$$

From Eqn. (5.42), we can see that identifiable pieces are completely separated and the small rotation matrices, $V_{m,i}$, only mix unidentifiable pieces $\Psi_{m,i}$. Π_m merely permutes the columns in G_m. From the non-uniqueness of local pieces of intrinsic sparse modes, we conclude that G_m are local pieces of another set of intrinsic sparse modes. □

We point out that the local pieces $\{G_m\}_{m=1}^M$ constructed by the ISMD on different patches may correspond to different sets of intrinsic sparse modes. Therefore, the final "patch-up" step should further modify and connect them to build a set of intrinsic sparse modes. Fortunately, the pivoted Cholesky decomposition elegantly solves this problem.

Optimal Sparse Recovery and Consistency of The ISMD

As defined in the ISMD, Ω is the correlation matrix of A with basis G_{ext}, see (5.15). If Ω enjoys a block diagonal structure with each block corresponding to a single intrinsic sparse mode, just like $\Omega^{(\psi)} \equiv L^{(\psi)} \left(L^{(\psi)}\right)^T$, the pivoted Cholesky decomposition can be utilized to recover the intrinsic sparse modes.

It is fairly easy to see that Ω indeed enjoys such a block diagonal structure when there is one set of intrinsic sparse modes that are pair-wisely identifiable. Denoting this identifiable set as $\{\psi_k\}_{k=1}^K$ (only its existence is needed), by Eqn. (5.41), we know that on patch P_m there is a permutation matrix Π_m and a diagonal matrix V_m with diagonal entries either 1 or -1 such that $D_m \Pi_m = D_m^{(\psi)} V_m$. Recall that $\Lambda = D\Omega D^T = D^{(\psi)} \Omega^{(\psi)} \left(D^{(\psi)}\right)^T$; see (5.16) and (5.40). We have

$$\Omega = D^T D^{(\psi)} \Omega^{(\psi)} \left(D^{(\psi)}\right)^T D = \Pi V^T \Omega^{(\psi)} V \Pi^T, \tag{5.43}$$

in which $V = \mathrm{diag}\{V_1, \ldots, V_m\}$ is diagonal with diagonal entries either 1 or -1 and $\Pi = \mathrm{diag}\{\Pi_1, \ldots, \Pi_m\}$ is a permutation matrix. Since the action of ΠV^T does not change the block diagonal structure of $\Omega^{(\psi)}$, Ω still has such a structure and the pivoted Cholesky decomposition can be readily applied. In fact, the action of ΠV^T exactly corresponds to the column permutation and sign flips of intrinsic sparse modes, which is the only kind of non-uniqueness of problem (5.3) when the intrinsic sparse modes are pair-wisely identifiable. For the general case when there are unidentifiable intrinsic sparse modes, Ω still has the block diagonal structure with each block corresponding to a group of unidentifiable modes, resulting in the following theorem.

Theorem 5.3.5. *Suppose the domain partition $\mathcal{P}$ is regular-sparse w.r.t. A. Let $A = GG^T$ be the decomposition given by the ISMD* (5.19) *and $\Psi \equiv [\psi_1, \ldots, \psi_K]$ be an arbitrary set of intrinsic sparse modes. Let columns in Ψ be ordered such that unidentifiable modes are grouped together, denoted as $\Psi = [\Psi_1, \ldots, \Psi_Q]$, where Q is the number of unidentifiable groups and n_q is the number of modes in Ψ_q. Then there exists Q rotation matrices $U_q \in \mathbb{R}^{n_q \times n_q}$*

($1 \leq q \leq Q$) such that

$$G = [\Psi_1 U_1, \ldots, \Psi_Q U_Q], \tag{5.44}$$

with reordering of columns in G if necessary. It immediately follows that

- *the ISMD generates one set of intrinsic sparse modes.*
- *the intrinsic sparse modes are unique up to permutations and rotations within unidentifiable modes.*

Proof. By Eqn. (5.41), Eqn. (5.43) still holds true with block diagonal V_m for $m \in [M]$. Without loss of generality, we assume that $\Pi = I$ since permutation does not change the block diagonal structure that we desire. Then from Eqn. (5.43) we have

$$\Omega = V^T \Omega^{(\psi)} V = V^T L^{(\psi)} \left(L^{(\psi)}\right)^T V. \tag{5.45}$$

In terms of block-wise formulation, we get

$$\Omega_{mn} = V_m^T \Omega_{mn}^{(\psi)} V_n = V_m^T L_m^{(\psi)} \left(L_n^{(\psi)}\right)^T V_n. \tag{5.46}$$

Correspondingly, by (5.42) the local pieces satisfy

$$G_m = [G_{m,1}, \ldots, G_{m,Q_m}] = [\Psi_{m,1} V_{m,1}, \ldots, \Psi_{m,Q_m} V_{m,Q_m}].$$

Now, we prove that Ω has the block diagonal structure in which each block corresponds to a group of unidentifiable modes. Specifically, $G_{m,i} = \Psi_{m,i} V_{m,i}$ and $G_{n,j} = \Psi_{n,j} V_{n,j}$ are two identifiable groups, i.e., $\Psi_{m,i}$ and $\Psi_{n,j}$ are from two identifiable groups, and we want to prove that the corresponding block in Ω, denoted as $\Omega_{m,i;n,j}$, is zero. From Eqn. (5.46), one gets $\Omega_{m,i;n,j} = V_{m,i}^T L_{m,i}^{(\psi)} \left(L_{n,j}^{(\psi)}\right)^T V_{n,j}$, where $L_{m,i}^{(\psi)}$ are the rows in $L_m^{(\psi)}$ corresponding to $\Psi_{m,i}$. $L_{n,j}^{(\psi)}$ is defined similarly. Due to identifiability between $\Psi_{m,i}$ and $\Psi_{n,j}$, we know $L_{m,i}^{(\psi)} \left(L_{n,j}^{(\psi)}\right)^T = 0$ and thus we obtain the block diagonal structure of Ω.

In (5.18), the ISMD performs the pivoted Cholesky decomposition $\Omega = PLL^T P^T$ and generates sparse modes $G = G_{ext} PL$. Due to the block diagonal structure in Ω, every column in PL can only have nonzero entries on local pieces that are not identifiable. Therefore, columns in G have identifiable intrinsic sparse modes completely separated and unidentifiable intrinsic sparse modes

rotated (including sign flip) by certain unitary matrices. Therefore, G is a set of intrinsic sparse modes.

Due to the arbitrary choice of Ψ, we know that the intrinsic sparse modes are unique to permutations and rotations within unidentifiable modes. □

Remark 5.3.1. *From the proof above, we can see that it is the block diagonal structure of Ω that leads to the recovery of intrinsic sparse modes. The pivoted Cholesky decomposition is one way to explore this structure. In fact, the pivoted Cholesky decomposition can be replaced by any other matrix decomposition that preserves this block diagonal structure, for instance, the eigendecomposition if there is no degeneracy.*

Despite the fact that the intrinsic sparse modes depend on the partition $\mathcal{P}$, the following theorem guarantees that the solutions to problem (5.3) give consistent results as long as the partition is regular-sparse.

Theorem 5.3.6. *Suppose that $\mathcal{P}_c$ is a partition, $\mathcal{P}_f$ is a refinement of $\mathcal{P}_c$ and that $\mathcal{P}_f$ is regular-sparse. Suppose $\{g_k^{(c)}\}_{k=1}^K$ and $\{g_k^{(f)}\}_{k=1}^K$ (with reordering if necessary) are the intrinsic sparse modes produced by the ISMD on $\mathcal{P}_c$ and $\mathcal{P}_f$, respectively. Then for every $k \in \{1, 2, \ldots, K\}$, in the coarse partition $\mathcal{P}_c$ $g_k^{(c)}$ and $g_k^{(f)}$ are supported on the same patches, while in the fine partition $\mathcal{P}_f$ the support patches of $g_k^{(f)}$ are contained in the support patches of $g_k^{(c)}$, i.e.,*

$$\{P \in \mathcal{P}_c : g_k^{(f)}|_P \neq \mathbf{0}\} = \{P \in \mathcal{P}_c : g_k^{(c)}|_P \neq \mathbf{0}\},$$
$$\{P \in \mathcal{P}_f : g_k^{(f)}|_P \neq \mathbf{0}\} \subset \{P \in \mathcal{P}_f : g_k^{(c)}|_P \neq \mathbf{0}\}.$$

Moreover, if $g_k^{(c)}$ is identifiable on the coarse patch $\mathcal{P}_c$, it remains unchanged when the ISMD is performed on the refined partition $\mathcal{P}_f$, i.e., $g_k^{(f)} = \pm g_k^{(c)}$.

Proof. Given the finer partition $\mathcal{P}_f$ is regular-sparse, it is easy to prove the coarser partition $\mathcal{P}_c$ is also regular-sparse.[2] Notice that if two modes are identifiable on the coarse partition $\mathcal{P}_c$, they must be identifiable on the fine partition $\mathcal{P}_f$. However, the other direction is not true, i.e., unidentifiable modes may become identifiable if the partition is refined. Based on this observation, Theorem 5.3.6 is a simple corollary of Theorem 5.3.5. □

[2]We provide the proof in supplementary materials; see Lemma B.1.1.

Finally, we provide a necessary condition for a partition to be regular-sparse as follows.

Proposition 5.3.7. *If $\mathcal{P}$ is regular-sparse w.r.t. A, all eigenvalues of Λ are integers. Here, Λ is computed in the ISMD by Eqn.* (5.12).

Proof. Let $\{\psi_k\}_{k=1}^K$ be a set of intrinsic sparse modes. Since $\mathcal{P}$ is regular-sparse, $D^{(\psi)}$ in Eqn. (5.39) is unitary. Therefore, Λ and $\Omega^{(\psi)} \equiv L^{(\psi)} \left(L^{(\psi)}\right)^T$ share the same eigenvalues. Due to the block-diagonal structure of $\Omega^{(\psi)}$, one can see that

$$\Omega^{(\psi)} \equiv L^{(\psi)} \left(L^{(\psi)}\right)^T = \sum_{k=1}^{K} l_k^{(\psi)} \left(l_k^{(\psi)}\right)^T$$

is, in fact, the eigendecomposition of $\Omega^{(\psi)}$. The eigenvalue corresponding to the eigenvector $l_k^{(\psi)}$ is $\|l_k^{(\psi)}\|_2^2$, which is also equal to $\|l_k^{(\psi)}\|_1$ because $L^{(\psi)}$ only elements 0 or 1. From item 1 in Proposition 5.3.2, $\|l_k^{(\psi)}\|_1 = s_k$, which is the patchwise sparseness of ψ_k. □

Combining Theorem 5.3.5, Theorem 5.3.6 and Proposition 5.3.7, we can develop a hierarchical process that gradually finds the finest regular-sparse partition and thus obtains the sparsest decomposition using the ISMD. This sparsest decomposition can be viewed as another definition of intrinsic sparse modes, which are independent of partitions. In our numerical examples, our partitions are all uniform but with different patch sizes. We see that even when the partition is not regular-sparse, the ISMD still produces a nearly optimal sparse decomposition.

5.4 Perturbation Analysis and Two Modifications

In real applications, data are often contaminated by noise. For example, when measuring the covariance function of a random field, sample noise is inevitable if a Monte Carlo type sampling method is utilized. A basic requirement for a numerical algorithm is its stability w.r.t. small noise levels. In Section 5.4, under several assumptions, we are able to prove that the ISMD is stable w.r.t. small perturbations in the input A. In Section 5.4, we provide two modified ISMD algorithms that effectively handle noise in different situations.

Perturbation Analysis of The ISMD

We consider the additive perturbation here, i.e., $\widehat{A}$ is an approximately low rank symmetric PSD matrix that satisfies

$$\widehat{A} = A + \epsilon\widetilde{A}, \qquad \|\widetilde{A}\|_2 \leq 1. \tag{5.47}$$

Here, A is the noiseless rank-K symmetric PSD matrix and $\widetilde{A}$ is the symmetric additive perturbation and $\epsilon > 0$ quantifies the noise level. We divide $\widetilde{A}$ into blocks that are conformal with blocks of A in (5.8) and thus $\widehat{A}_{mn} = A_{mn} + \epsilon\widetilde{A}_{mn}$. In this case, we need to apply the truncated local eigendecomposition (5.10) to capture the correct local rank K_m. Suppose the eigendecomposition of $\widehat{A}_{mm}$ is

$$\widehat{A}_{mm} = \sum_{i=1}^{K_m} \widehat{\gamma}_{m,i}\widehat{h}_{n,i}\widehat{h}_{n,i}^T + \sum_{i>K_m} \widehat{\gamma}_{m,i}\widehat{h}_{n,i}\widehat{h}_{n,i}^T.$$

In this subsection, we assume that the noise level is very small with $\epsilon \ll 1$ such that there is an energy gap between $\widehat{\gamma}_{m,K_m}$ and $\widehat{\gamma}_{m,K_m+1}$. Therefore, the truncation (5.10) captures the correct local rank K_m, i.e.,

$$\widehat{A}_{mm} \approx \widehat{A}_{mm}^{(t)} \equiv \sum_{i=1}^{K_m} \widehat{\gamma}_{m,i}\widehat{h}_{n,i}\widehat{h}_{n,i}^T \equiv \widehat{H}_m\widehat{H}_m^T. \tag{5.48}$$

In the rest of the ISMD, the perturbed local eigenvectors $\widehat{H}_m$ is used as H_m in the noiseless case. We expect that our ISMD is stable w.r.t. this small perturbation and generates slightly perturbed intrinsic sparse modes of A.

To carry out this perturbation analysis, we will restrict ourselves to the case when intrinsic sparse modes of A are pair-wisely identifiable and thus it is possible to compare the error between the noisy output $\widehat{g}_k$ with A's intrinsic sparse mode g_k. When there are unidentifiable intrinsic sparse modes of A, it only makes sense to consider the perturbation of the subspace spanned by those unidentifiable modes and we will not consider this case in this chapter. The following lemma is a preliminary result on the perturbation analysis of local pieces G_m.

Lemma 5.4.1. *Suppose that partition $\mathcal{P}$ is regular-sparse w.r.t. A and all intrinsic modes are identifiable with each other. Furthermore, we assume that for all $m \in [M]$ there exists $E_m^{(eig)}$ such that*

$$\widehat{A}_{mm}^{(t)} = (I + \epsilon E_m^{(eig)})A_{mm}\left(I + \epsilon(E_m^{(eig)})^T\right) \quad \textit{and} \quad \|E_m^{(eig)}\|_2 \leq C_{eig}. \tag{5.49}$$

Here C_{eig} is a constant depending on A but not on ϵ or $\widetilde{A}$. Then there exists $E_m^{(jd)} \in \mathbb{R}^{K_m \times K_m}$ such that

$$\widehat{G}_m = (I + \epsilon E_m^{(eig)}) G_m (I + \epsilon E_m^{(jd)} + \mathcal{O}(\epsilon^2)) J_m \quad \textit{and} \quad \|E_m^{(jd)}\|_F \leq C_{jd}, \tag{5.50}$$

where G_m and $\widehat{G}_m$ are local pieces constructed by the ISMD with input A and $\widehat{A}$ respectively, J_m is the product of a permutation matrix with a diagonal matrix having only ± 1 on its diagonal, and C_{jd} is a constant depending on A but not on ϵ or $\widetilde{A}$. Here, $\| \bullet \|_2$ and $\| \bullet \|_F$ are matrix spectral norm and Frobenius norm, respectively.

Lemma 5.4.1 ensures that local pieces of intrinsic sparse modes can be constructed with $\mathcal{O}(\epsilon)$ accuracy up to permutation and sign flips (characterized by J_m in (5.50)) under several assumptions. The identifiability assumption is necessary. Without such an assumption, these local pieces are not uniquely determined up to permutations and sign flips. The assumption (5.49) holds true when eigendecomposition of A_{mm} is well conditioned, i.e., both eigenvalues and eigenvectors are well conditioned. We expect that a stronger perturbation result is still true without making this assumption. The proof of Lemma 5.4.1 is an application of perturbation analysis for the joint diagonalization problem [24], and is presented in supplementary materials B.3.

Finally, $\widehat{\Omega}$ is the correlation matrix of $\widehat{A}$ with basis $\widehat{G}_{ext} = \text{diag}\{\widehat{G}_1, \widehat{G}_2, \ldots, \widehat{G}_M\}$. Specifically, the (m,n)-th block of $\widehat{\Omega}$ is given by

$$\widehat{\Omega}_{mn} = \widehat{G}_m^{\dagger} \widehat{A}_{mn} \left(\widehat{G}_n^{\dagger}\right)^T.$$

Without loss of generality, we can assume that $J_m = \mathbb{I}_{K_m}$ in (5.50).[3] Based on the perturbation analysis of G_m in Lemma 5.4.1 and the standard perturbation analysis of pseudo-inverse (for instance, see Theorem 3.4 in [122]), it is straightforward to get a bound of the perturbations in $\widehat{\Omega}$, i.e.,

$$\|\widehat{\Omega} - \Omega\|_2 \leq C_{ismd} \epsilon. \tag{5.51}$$

Here, C_{ismd} depends on the smallest singular value of G_m and the constants C_{eig} and C_{jd} in Lemma 5.4.1. Notice that when all intrinsic modes are identifiable with each other, the entries of Ω are either 0 or ± 1. Therefore, when

[3] One can check that $\{J_m\}_{m=1}^M$ only affect the sign of recovered intrinsic sparse modes $[\widehat{g}_1, \widehat{g}_2, \ldots, \widehat{g}_K]$ if pivoted Cholesky decomposition is applied on $\widehat{\Omega}$.

$C_{ismd}\epsilon$ is small enough, we can exactly recover Ω from $\widehat{\Omega}$ as below:

$$\Omega_{ij} = \begin{cases} -1, & \text{for } \widehat{\Omega}_{ij} < -0.5, \\ 0, & \text{for } \widehat{\Omega}_{ij} \in [-0.5, 0.5], \\ 1, & \text{for } \widehat{\Omega}_{ij} > 0.5. \end{cases} \tag{5.52}$$

Following Algorithm 2, we get the pivoted Cholesky decomposition $\Omega = PLL^T P^T$ and output the perturbed intrinsic sparse modes

$$\widehat{G} = \widehat{G}_{ext} PL.$$

Notice that the patchwise sparseness information is all coded in L and we can reconstruct L exactly due to the thresholding step (5.52), $\widehat{G}$ has the same patchwise sparse structure as G. Moreover, because the local pieces $\widehat{G}_{ext}$ are constructed with $\mathcal{O}(\epsilon)$ error, we have

$$\|\widehat{G} - G\|_2 \leq C_g \epsilon, \tag{5.53}$$

where the constant C_g only depends on the constants C_{eig} and C_{jd} in Lemma 5.4.1.

Two Modified ISMD Algorithms

In Section 5.4, we have shown that the ISMD is robust to small noise under the assumption of regular sparsity and identifiability. In this subsection, we provide two modified versions of the ISMD to deal with the cases when these two assumptions fail. The first modification aims at constructing intrinsic sparse modes from noisy input $\widehat{A}$ in the small noise level region as before, but it does not require the regular sparsity and identifiability. The second modification aims at constructing a simultaneous low-rank and sparse approximation of $\widehat{A}$ when the noise level is high. Our numerical experiments demonstrate that these modified algorithms are quite effective in practice.

ISMD with thresholding

In the general case where unidentifiable pairs of intrinsic sparse modes exist, the thresholding idea (5.52) is still applicable but the threshold ϵ_{th} should be learned from the data, i.e., the entries in $\widehat{\Omega}$. Specifically, there are $\mathcal{O}(1)$ entries in $\widehat{\Omega}$ corresponding to the slightly perturbed nonzero entries in Ω; there are also many $\mathcal{O}(\epsilon)$ entries that are contributed by the noise $\epsilon\widetilde{A}$. If the noise level ϵ is small enough, we can see a gap between these two group of entries, and

a threshold ϵ_{th} is chosen such that it separates these two groups. A simple 2-cluster algorithm is able to identify the threshold ϵ_{th}. In our numerical examples, we draw the histogram of absolute values of entries in $\widehat{\Omega}$ and it clearly shows the 2-cluster effect; see Figure 5.10. Finally, we set all the entries in $\widehat{\Omega}$ with absolute value less than ϵ_{th} to 0. In this approach we do not need to know the noise level ϵ a priori and we just learn the threshold from the data. To modify Algorithm 2 with this thresholding technique, we just need to add one line between assembling Ω (line 15) and the pivoted Cholesky decomposition (line 16); see Algorithm 3.

Algorithm 3 Intrinsic sparse mode decomposition with thresholding

Require: $A \in \mathbb{R}^{N\times N}$: symmetric and PSD; $\mathcal{P} = \{P_m\}_{m=1}^{M}$: partition of index set $[N]$

Ensure: $G = [g_1, g_2, \cdots, g_K]$: $A \approx GG^T$

1: The same with Algorithm 2 from Line 1 to Line 13
2: ###Assemble Ω, thresholding and its pivoted Cholesky decomposition
3: $\Omega = D^T \Lambda D$
4: Learn a threshold ϵ_{th} from Ω and set all the entries in Ω with absolute value less than ϵ_{th} to 0
5: $\Omega = PLL^T P^T$
6: ###Assemble the intrinsic sparse modes G
7: $G = H_{ext} DPL$

It is important to point out that when the noise level is high, the $\mathcal{O}(1)$ entries and $\mathcal{O}(\epsilon)$ entries mix together. In this case, we cannot identify such a threshold ϵ_{th} to separate them, and the assumption that there is an energy gap between $\widehat{\gamma}_{m,K_m}$ and $\widehat{\gamma}_{m,K_m+1}$ is invalid. In the next subsection, we will present the second modified version to overcome this difficulty.

Low rank approximation with ISMD

In the case where there is no gap between $\widehat{\gamma}_{m,K_m}$ and $\widehat{\gamma}_{m,K_m+1}$ (i.e., no well-defined local ranks), or when the noise level is so high that the threshold ϵ_{th} cannot be identified, we modify our ISMD to give a low-rank approximation of $A \approx GG^T$, in which G is observed to be patchwise sparse from our numerical examples.

In this modification, the normalization (5.17) is applied and thus we have

$$A \approx \bar{G}_{ext} \bar{\Omega} \bar{G}_{ext}^T.$$

It is important to point out that $\bar{\Omega}$ has the same block diagonal structure as Ω but has different eigenvalues. Specifically, for the case when there is no noise and the regular-sparse assumption holds true, $\bar{\Omega}$ has eigenvalues $\{\|g_k\|_2^2\}_{k=1}^K$ for a certain set of intrinsic sparse modes g_k, while Ω has eigenvalues $\{s_k\}_{k=1}^K$ (here s_k is the patchwise sparseness of the intrinsic sparse mode). We first perform eigendecomposition $\bar{\Omega} = \bar{L}\bar{L}^T$ and then assemble the final result by $G = \bar{G}_{ext}\bar{L}$. The modified algorithm is summarized in Algorithm 4.

Algorithm 4 Intrinsic sparse mode decomposition for low rank approximation

Require: $A \in \mathbb{R}^{N\times N}$: symmetric and PSD; $\mathcal{P} = \{P_m\}_{m=1}^M$: partition of index set $[N]$

Ensure: $G = [g_1, g_2, \cdots, g_K]$: $A \approx GG^T$

1: The same with Algorithm 2 from Line 1 to Line 13
2: ###Assemble Ω, normalization and its eigendecomposition
3: $\Omega = D^T \Lambda D$
4: $G_{ext} = \bar{G}_{ext}E, \quad \bar{\Omega} = E\Omega E^T$ as in (5.17)
5: $\bar{\Omega} = \bar{L}\bar{L}^T$
6: ###Assemble the intrinsic sparse modes G
7: $G = \bar{G}_{ext}\bar{L}$

Here we replace the pivoted Cholesky decomposition of Ω in Algorithm 2 by eigendecomposition of $\bar{\Omega}$. From Remark 5.3.1, this modified version generates exactly the same result with Algorithm 2 if all the intrinsic sparse modes have different l_2 norm (there are no repeated eigenvalues in $\bar{\Omega}$). The advantage of the pivoted Cholesky decomposition is its low computational cost and the fact that it always exploits the (unordered) block diagonal structure of Ω. However, it is more sensitive to noise compared to eigendecomposition, which is much more robust to noise. Moreover, eigendecomposition gives the optimal low rank approximation of $\bar{\Omega}$. Thus, Algorithm 4 gives a more accurate low rank approximation for A compared to Algorithm 2 and Algorithm 3 that use the pivoted Cholesky decomposition.

5.5 Numerical Experiments

In this section, we demonstrate the robustness of our intrinsic sparse mode decomposition method and compare its performance with that of the eigendecomposition, the pivoted Cholesky decomposition, and the convex relaxation of SPCA. All our computations are performed using MATLAB R2015a (64-bit) on an Intel Core i7-3770 (3.40 GHz). The pivoted Cholesky decomposition is implemented in MATLAB according to Algorithm 3.1 in [86].

We will use synthetic covariance matrices of a random permeability field, which models some underground porous media, as the symmetric PSD input A. This random permeability model is adapted from the porous media problem [50, 46] where the physical domain D is two dimensional. The basic model has a constant background and several localized features to model the subsurface channels and inclusions, i.e.,

$$\kappa(x,\omega) = \kappa_0 + \sum_{k=1}^{K} \eta_k(\omega) g_k(x), \quad x \in [0,1]^2, \tag{5.54}$$

where κ_0 is the constant background, $\{g_k\}_{k=1}^K$ are characteristic functions of channels and inclusions and η_k are the associated uncorrelated latent variables controlling the permeability of each feature. Here, we have $K = 35$, including 16 channels and 18 inclusions. Among these modes, there is one artificial smiling face mode that has disjoint branches. It is used here to demonstrate that the ISMD is able to capture long range correlation. For this random medium, the covariance function is

$$a(x,y) = \sum_{k=1}^{K} g_k(x) g_k(y), \quad x, y \in [0,1]^2. \tag{5.55}$$

Since the length scales of channels and inclusions are very small, with width about $1/32$, we need a fine grid to resolve these small features. Such a fine grid is also needed when we do further scientific experiments [50, 46, 65]. In this chapter, the physical domain $D = [0,1]^2$ is discretized using a uniform grid with $h_x = h_y = 1/96$, resulting in $A \in \mathbb{R}^{N\times N}$ with $N = 96^2$. One sample of the random field (and the bird's-eye view) and the covariance matrix are plotted in Figure 5.2. It can be seen that the covariance matrix is sparse and concentrates along the diagonal since modes in the ground-truth media are all localized functions.

Note that this example is synthetic because we construct A from a sparse decomposition (5.55). We would like to test whether different matrix factorization methods, like eigendecomposition, the Cholesky decomposition, and the ISMD, are able to recover this sparse decomposition, or even find a sparser decomposition for A.

Numerical Results of ISMD

The partitions we take for this example are all uniform domain partitions with $H_x = H_y = H$. We run the ISMD with patch sizes $H \in \{1, 1/2, 1/3, 1/4, 1/6,$

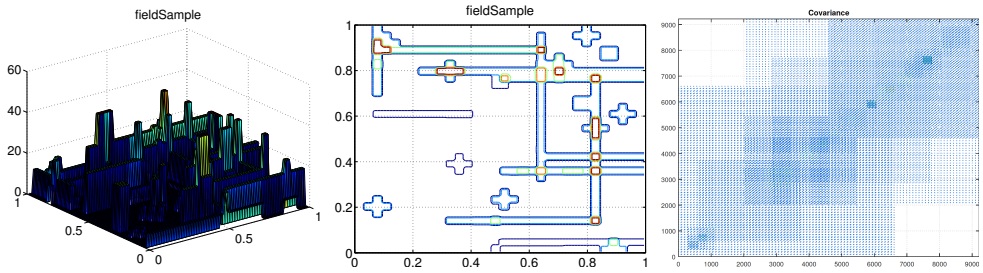


Figure 5.2: One sample and the bird's-eye view. The covariance matrix is plotted on the right.

$1/8, 1/12, 1/16, 1/24, 1/32, 1/48, 1/96\}$ in this section. For the coarsest partition $H = 1$, the ISMD is exactly the eigendecomposition of A. For the finest partition $H = 1/96$, the ISMD is equivalent to the pivoted Cholesky factorization on $\bar{A}$ where $\bar{A}_{ij} = \frac{A_{ij}}{\sqrt{A_{ii}A_{jj}}}$. The pivoted Cholesky factorization on A is also implemented. It is no surprise that all the above methods produce 35 modes. The number of modes is exactly the rank of A. We plot the first 6 modes for each method in Figure 5.3. We can see that both the eigendecomposition (ISMD with $H = 1$) and the pivoted Cholesky factorization on A generate modes which mix different localized feathers together. On the other hand, the ISMD with $H = 1/8$ and $H = 1/32$ recover exactly the localized feathers, including the upside-down smiling face.

We use Lemma 5.3.1 to check when the regular-sparse property fails. It turns out that for $H \geq 1/16$ the regular-sparse property holds and for $H \leq 1/24$ it fails. The eigenvalues of Λ's for $H = 1, 1/8$, and $1/32$ are plotted in Figure 5.4 on the left side. The eigenvalues of Λ when $H = 1$ are all 1's, since every eigenvector has patchwise sparseness 1 in this trivial case. The eigenvalues of Λ when $H = 1/16$ are all integers, corresponding to patchwise sparseness of the intrinsic sparse modes. The eigenvalues of Λ when $H = 1/32$ are not all integers any more, which indicates that this partition is not regular-sparse w.r.t. A according to Lemma 5.3.1.

The consistency of the ISMD (Theorem 5.3.6) manifests itself from $H = 1$ to $H = 1/8$ in Figure 5.3. As Theorem 5.3.6 states, the supports of the intrinsic sparse modes on a coarser partition contain those on a finer partition. In other words, we get sparser modes when we refine the partition as long as the partition is regular-sparse. After checking all the 35 recovered modes, we see that the intrinsic sparse modes get sparser and sparser from $H = 1$ to $H = 1/6$.

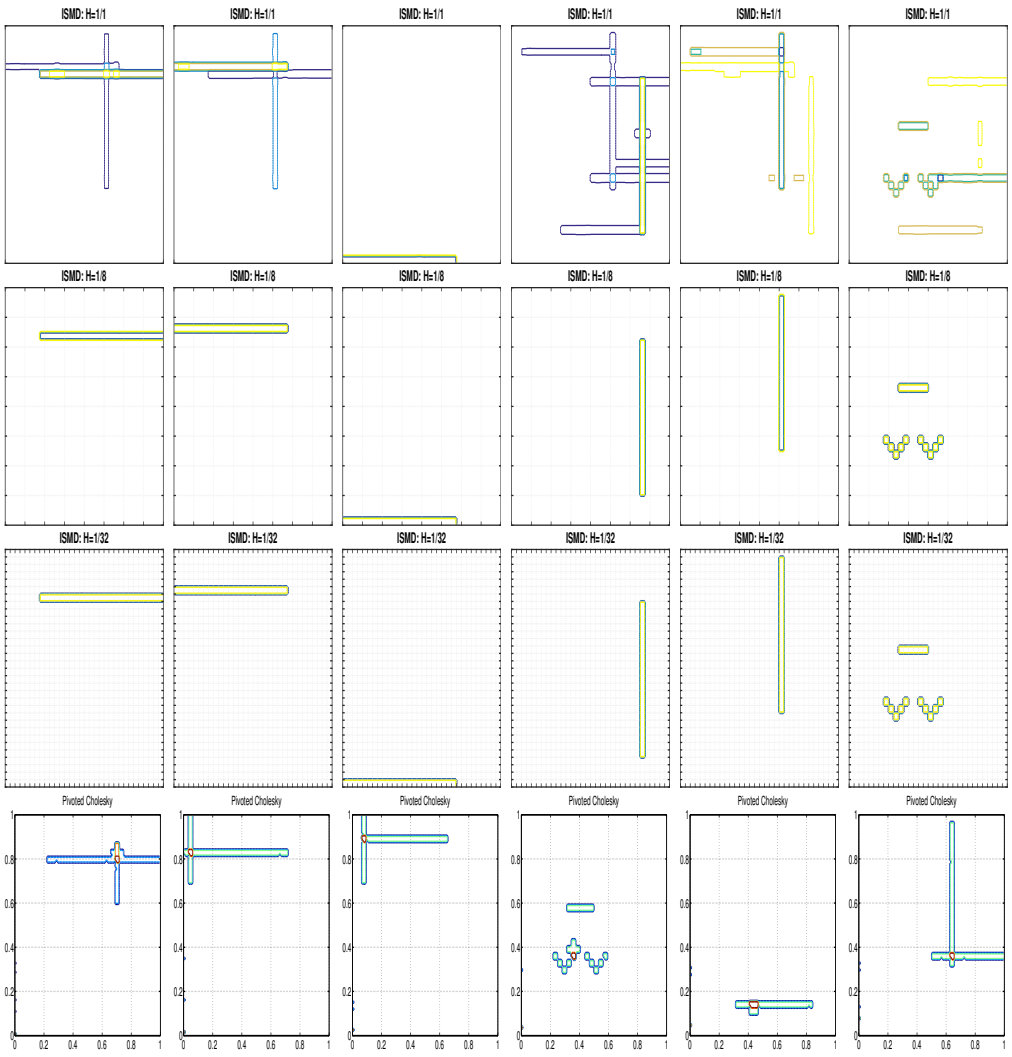


Figure 5.3: First 6 eigenvectors (H=1); First 6 intrinsic sparse modes (H=1/8, regular-sparse); First 6 intrinsic sparse modes (H=1/32; not regular-sparse); First 6 modes from the pivoted Cholesky decomposition of A

When $H \leq 1/6$, all the 35 intrinsic sparse modes are identifiable with each other and these intrinsic modes remain the same for $H = 1/8, 1/12, 1/16$. When $H \leq 1/24$, the regular-sparse property fails, but we still get the sparsest decomposition (the same decomposition with $H = 1/8$). For $H = 1/32$, we recover exactly 33 intrinsic sparse modes but get the other two mixed together. This is not surprising since the partition is not regular-sparse any more. For $H = 1/48$, we exactly recover all of the 35 intrinsic sparse modes again. Table 5.1 lists the cases when we exactly recover the sparse decomposition (5.55) from which we construct A. From Theorem 5.3.5, this decomposition is the optimal sparse decomposition (defined by problem (5.3)) for $H \geq 1/16$. We

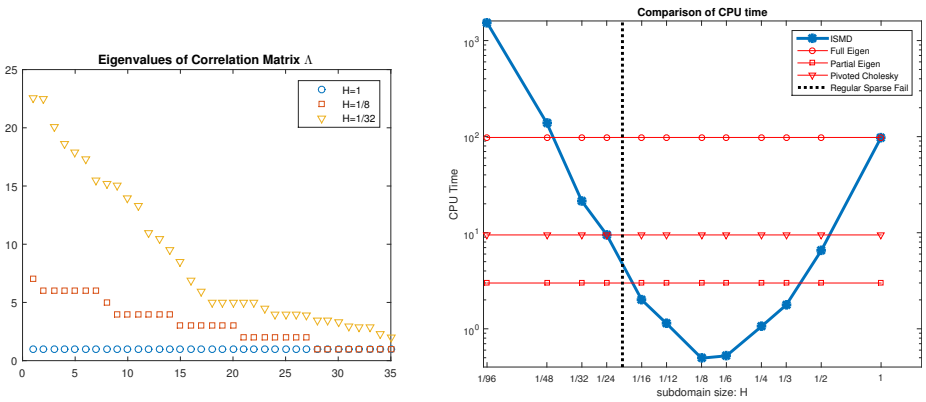


Figure 5.4: Left: Eigenvalues of Λ for $H = 1, 1/8, 1/32$. By Lemma 5.3.1, the partition with $H = 1/32$ is not regular-sparse. Right: CPU time (unit: second) for different partition sizes H.

suspect that this decomposition is also optimal in the l_0 sense (defined by problem (5.2)).

H	1	1/2	1/3	1/4	1/6	1/8
regular-sparse	✔	✔	✔	✔	✔	✔
Exact Recovery	✗	✗	✗	✗	✔	✔
H	1/12	1/16	1/24	1/32	1/48	1/96
regular-sparse	✔	✔	✗	✗	✗	✗
Exact Recovery	✔	✔	✔	✗	✔	✗

Table 5.1: Cases when the ISMD gets exact recovery of the sparse decomposition (5.55)

The CPU time of the ISMD for different H's is shown in Figure 5.4 on the right side. We compare the CPU time for the full eigendecomposition `eig(A)`, the partial eigendecomposition `eigs(A, 35)`, and the pivoted Cholesky decomposition. For $1/16 \leq H \leq 1/3$, the ISMD is even faster than the partial eigendecomposition. Specifically, the ISMD is ten times faster for the case $H = 1/8$. Notice that the ISMD performs the local eigendecomposition by `eig` in Matlab, and thus does not need any prior information about the rank K. If we also assume prior information on the local rank K_m, the ISMD would be even faster. The CPU time curve has a V-shape as predicted by our computational estimation (5.23). The cost first decreases as we refine the mesh because the cost of local eigendecompositions decreases. Then it increases as we refine further because there are M joint diagonalization problem (5.13) to

be solved. When M is very large, i.e., $H = 1/48$ or $H = 1/96$, the 2 layer for-loops from Line 5 to Line 10 in Algorithm 2 become extremely slow in Matlab. When implemented in other languages that have little overhead cost for multiple for-loops, e.g. C or C++, the actual CPU time for $H = 1/96$ would be roughly the same with the CPU time for the pivoted Cholesky decomposition.

Comparison With The Semidefinite Relaxation of SPCA

In comparison, the semidefinite relaxation of SPCA (problem (5.27)) gives poor results in this example. We have tested several values of μ, and found that parameter $\mu = 0.0278$ gives the best performance in the sense that the first 35 eigenvectors of W capture the most variance in A. The first 35 eigenvectors of W, shown in Figure 5.5, explain 95% of the variance, but all of them mix several intrinsic modes like what the eigendecomposition does in Figure 5.3. For this example, it is not clear how to choose the best 35 columns out of all the 9216 columns in W, as proposed in [78]. If columns of W are ordered by the l_2 norm in descending order, the first 35 columns can only explain 31.46% of the total variance, although they are indeed localized. Figure 5.6 shows the first six columns of W with largest norms.

We also compare the CPU time of the ISMD with that of the semidefinite relaxation of SPCA (5.27). The SPCA is computed using the split Bregman iteration. Each split Bregman iteration requires an eigendecomposition of a matrix of size $N \times N$. In comparison, the ISMD is cheaper than a single eigendecomposition, as shown in Figure 5.4. It has been observed that the split Bregman iteration converges linearly. If we set the error tolerance to be $O(\delta)$, the number of iterations needed is about $\mathcal{O}(1/\delta)$. In our implementation, we set the error tolerance to be 10^{-3} and we need to perform 852 iterations. Overall, to solve the convex optimization problem (5.27) with split Bregman iteration takes over 1000 times more CPU time than the ISMD with $H = 1/8$.

It is expected that the ISMD is much faster than SPCA since the SPCA needs to perform many times of partial eigendecomposition to solve problem (5.27), but the ISMD has computational cost comparable to one single partial eigendecomposition. As we discussed in Section 5.1, SPCA is designed and works reasonably well for problem (5.7). When SPCA is applied to our sparse decomposition problem (5.3), it does not work well. However, it is not always the case that the ISMD gives a sparser and more accurate decomposition of

A than SPCA. In subsection 5.5, we will present another example in which SPCA gives a better performance than the ISMD.

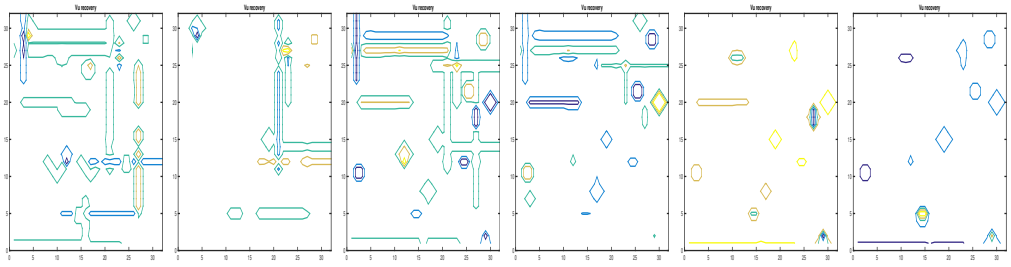


Figure 5.5: Sparse PCA: The first six eigenvectors of W. The first 35 eigenvectors of W explain 95% of the variance.

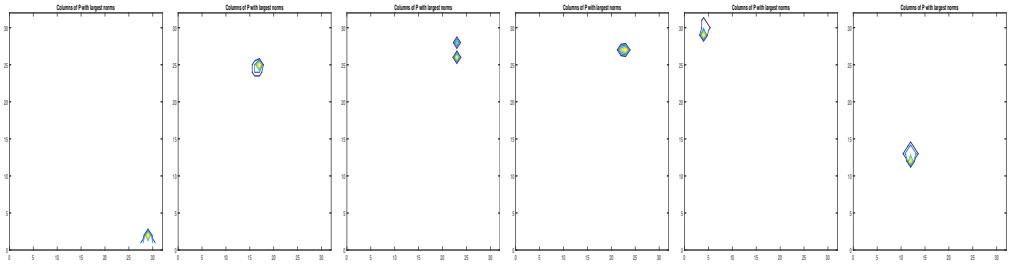


Figure 5.6: Sparse PCA: six columns of W with largest norms. The first 35 columns with largest norms only explain 31.46% of the variance.

We point out that unlike the structured SPCA [72], the ISMD does not take advantage of the specific (rectangular) structure of the physical modes. The "smiling face" mode shows that the ISMD can recover non-convex and non-local sparse modes. Therefore, the ISMD is expected to perform equally well even when there are no such structures known.

ISMD With Small Noise Levels

In this subsection, we report a test on the robustness of the ISMD. In the following test, we perturb the rank-35 covariance matrix $A \in \mathbb{R}^{9216\times 9216}$ with a random matrix:

$$\widehat{A} = A + \epsilon \widetilde{A}\,,$$

where ϵ is the noise level and $\widetilde{A}$ is a random matrix with i.i.d. elements uniformly distributed in $[-1, 1]$. Notice that all elements in A are uniformly bounded by 1, and thus ϵ is a relative noise level. Since all the intrinsic

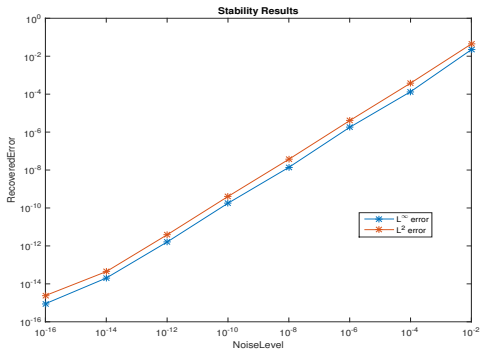


Figure 5.7: L^∞ and l_2 error increases linearly as the noise level increases.

sparse modes are identifiable with each other for the partition with patch size $H = 1/16$, we perform ISMD with simple thresholding (5.52) on $\widehat{A}$ to get the perturbed intrinsic sparse modes $\widehat{G} \equiv [\widehat{g}_1, \ldots, \widehat{g}_K]$. The l^∞ and l_2 error are defined as below:

$$Err_\infty = \max_{k=1,2,\cdots,K} \frac{\|\widehat{g}_k - g_k\|_2}{\|g_k\|_2}, \quad Err_2 = \sqrt{\sum_{k=1}^{K} \frac{\|\widehat{g}_k - g_k\|_2^2}{\|g_k\|_2^2}}.$$

Figure 5.7 shows that Err_∞ and Err_2 depend linearly on the noise level ϵ, which validates our stability analysis in Section 5.4.

Separate Global and Localized Modes with ISMD

In this example, we consider a more sophisticated model in which the media contain several global modes, i.e.,

$$\kappa(x, \omega) = \sum_{k=1}^{K_1} \xi_k(\omega) f_k(x) + \sum_{k=1}^{K_2} \eta_k(\omega) g_k(x), \quad x \in [0, 1]^2, \tag{5.56}$$

where $\{g_k\}_{k=1}^{K_2}$ and η_k models the localized features like channels and inclusions as above, $\{f_k\}_{k=1}^{K_1}$ are functions with support on the entire domain $D = [0, 1]^2$ and ξ_k are the associated latent variables with global influence on the entire domain. Here, we keep the 35 localized features as before, but add two global features with $f_1(x) = \sin(2\pi x_1 + 4\pi x_2)/2$, $f_2(x) = \sin(4\pi x_1 + 2\pi x_2)/2$. ξ_1 and ξ_2 are set to be uncorrelated and have variance 1. For this random medium, the covariance function is

$$a(x, y) = \sum_{k=1}^{K_1} f_k(x) f_k(y) + \sum_{k=1}^{K_2} g_k(x) g_k(y), \quad x, y \in [0, 1]^2. \tag{5.57}$$

As before, we discretize the covariance function with $h_x = h_y = 1/96$ and represent A by a matrix of size 9216×9216. One sample of the random field (and the bird's-eye view) and the covariance matrix are plotted in Figure 5.8. It can be seen that the covariance matrix is dense now because we have two global modes.

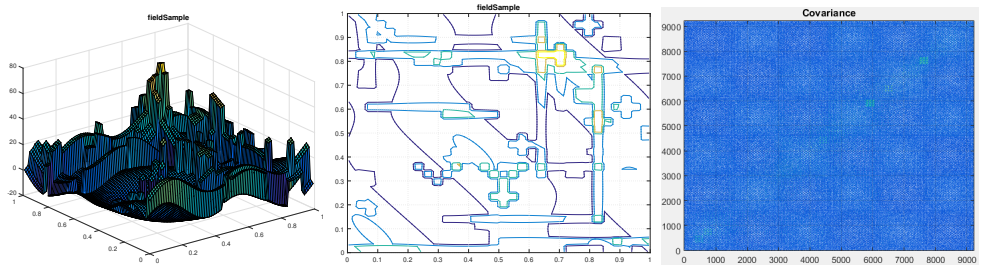


Figure 5.8: One sample and the bird's-eye view. The covariance matrix is plotted on the right.

We apply the ISMD with patch size $H = 1/16$ on A and get 37 intrinsic sparse modes as expected. Moreover, two of them are rotations of $[f_1, f_2]$ and the other 35 are exactly the 35 localized modes in the construction (5.57). We plot the first 6 intrinsic sparse modes in Figure 5.9. As we can see, the ISMD separates the global modes and localized modes in A, or equivalently we separate the low rank dense part and sparse part of A. The reason why we can achieve this separation is that the representation (5.57), in fact, solves the patchwise sparseness minimization problem (5.3). The low-rank-plus-sparse decomposition (also known as Robust PCA, see [26, 22, 87]) can also separate the low rank dense part and the sparse part in A. However, the computational cost of robust PCA is much more expensive than the ISMD.

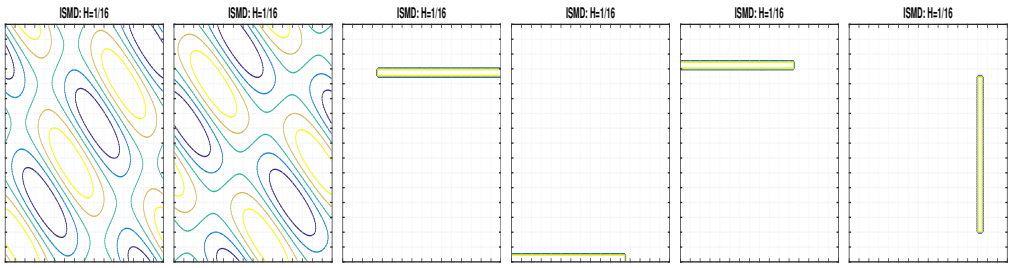


Figure 5.9: First 6 intrinsic sparse modes (H=1/16, regular-sparse)

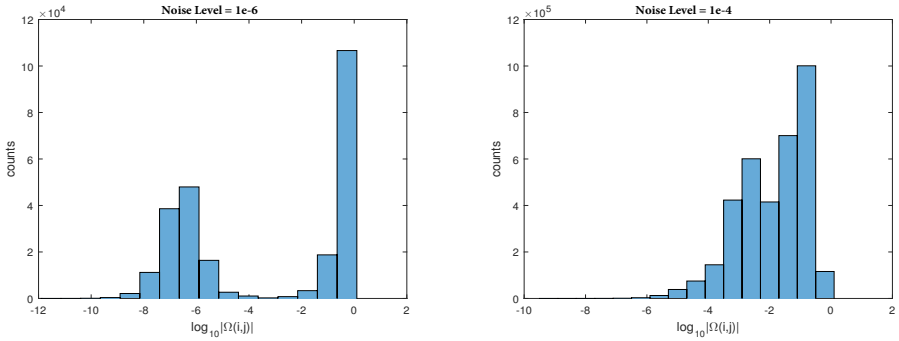


Figure 5.10: Histogram of absolute values of entries in $\widehat{\Omega}$.

Application of Algorithm 3

When A is constructed from model (5.57) but is mixed with small noise as in Section 5.5, we cannot simply apply the thresholding (5.52) any more. In this case, we have unidentifiable modes f_1 and f_2 and thus Ω may contain nonzero values other than ± 1. For the noise level $\epsilon = 10^{-6}$, Figure 5.10 (left) shows the histogram of absolute values of entries in $\widehat{\Omega}$. We can clearly see a gap between $\mathcal{O}(\epsilon)$ entries and $\mathcal{O}(1)$ entries from Figure 5.10(left). Therefore we choose a threshold $\epsilon_{th} = 10^{-3}$ and apply the modified ISMD algorithm 3 on $\widehat{A}$. The first 6 perturbed intrinsic sparse modes $\widehat{g}_k$ are shown in Figure 5.11. We can see that their supports are exactly the same as those of the unperturbed intrinsic sparse modes g_k in Figure 5.9. In fact, the first 37 perturbed intrinsic sparse modes $\{\widehat{g}_k\}_{k=1}^{37}$ exactly capture the supports of the unperturbed intrinsic sparse modes $\{g_k\}_{k=1}^{37}$. However, we have several extra perturbed intrinsic sparse modes with very small l_2 error since $\widehat{\Omega}$ has rank more than 37.

When we raise the noise level ϵ to 10^{-4}, the histogram of the absolute values in $\widehat{\Omega}$ is shown in Figure 5.10(right). In this case, we cannot identify a gap any more. From Figure 5.10(left), we see that the exact Ω has entries in the order of 10^{-3}. Therefore, the noise level $\epsilon = 10^{-4}$ is large enough to mix the true nonzero values and noisy null values in $\widehat{\Omega}$ together. In Figure 5.10 the total counts are different because only values between $10^{-16.5}$ and $10^{0.5}$ are counted.

Application of Algorithm 4

In this section, we consider the one-dimensional Poisson kernel:

$$a(x,y) = e^{-\frac{|x-y|}{l}}, \quad x, y \in [-1,1],$$

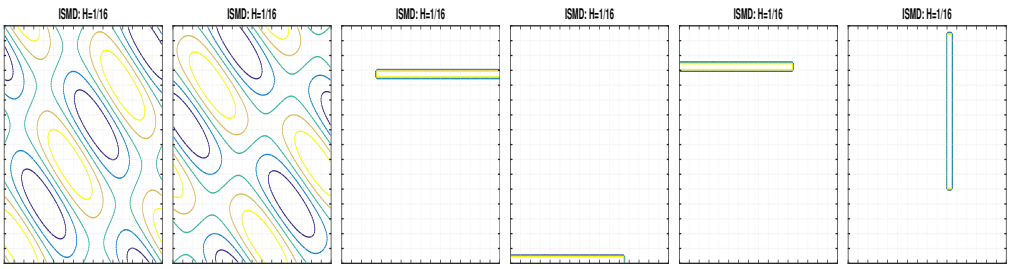


Figure 5.11: Application of Algorithm 3 (H=1/16, approximately regular-sparse): first 6 intrinsic sparse modes

where $l = 1/16$. To refine the small scale, $a(x, y)$ is discretized by a uniform grid with $h = 1/512$, resulting in $A \in \mathbb{R}^{1024\times1024}$. In Figure 5.12 we plot the covariance matrix. By truncating the eigendecomposition with 45 modes, we can approximate A with spectral norm error 5%, and these 45 KL modes are plotted on the right panel of the figure. As one can see, they are all global functions.

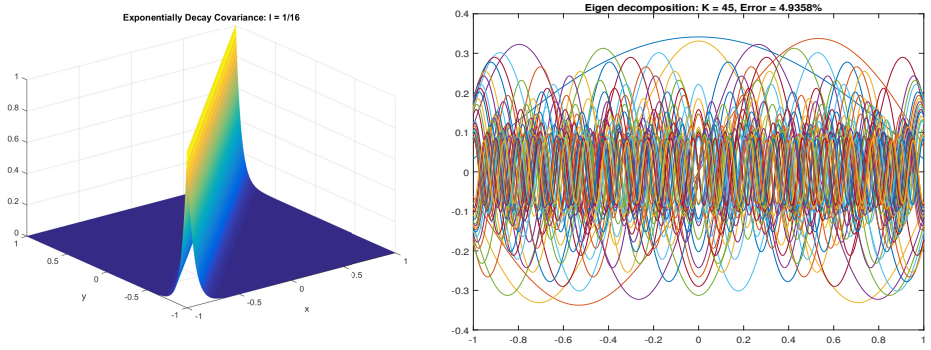


Figure 5.12: eigendecomposition: Covariance function and its first 45 KL modes. Error is 4.936%. Both local and global dimension are 45.

We decompose the domain into 2, 4, and 8 patches respectively and apply the Algorithm 4 with thresholding (5.52) to each case. For all the three cases, every mode has patchwise sparseness either 1 or 2. In Figure 5.13, the left panels show the modes that are nonzero on more than one patch, and the right panels collect the modes that are nonzero on only one patch. To achieve the same accuracy with the eigendecomposition, the numbers of modes needed are 45, 47, and 49 respectively. The total number is slightly larger than the number of eigen modes, but most modes are localized. For the two-patch case, each patch contains 23 nonzero modes, and for the four-patch case, each patch

contains either 12 or 13 nonzero modes, and for the eight-patch case, each patch contains only 7 nonzero modes.

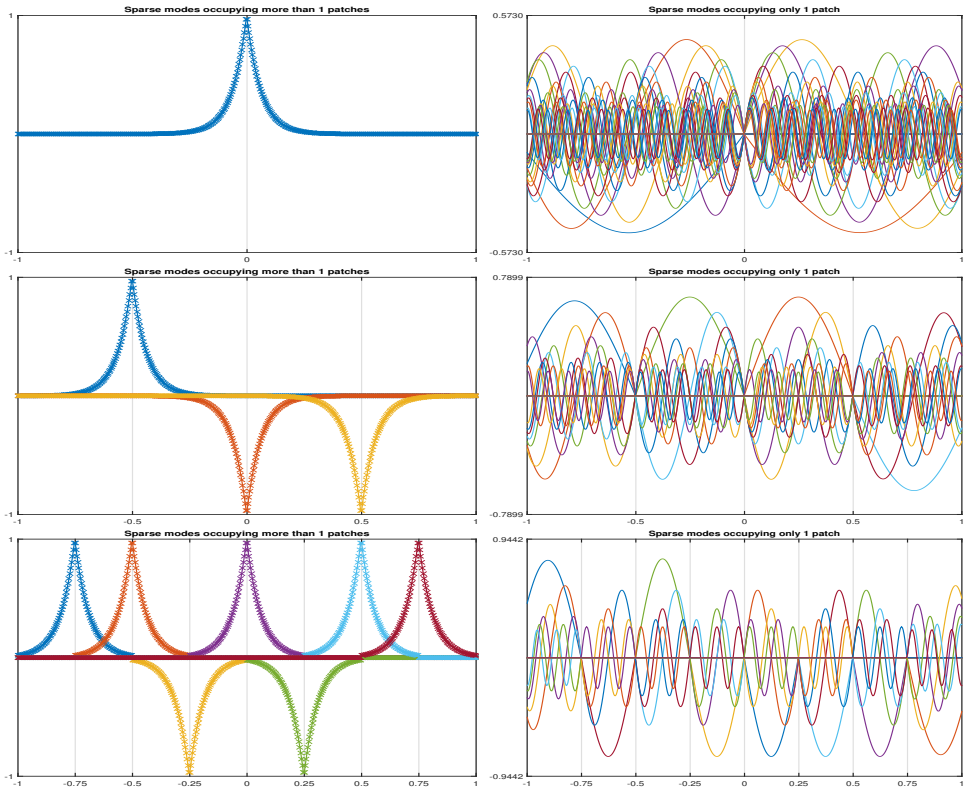


Figure 5.13: Upper: Two patches case. Error is 4.95%. Global dimension is 45 and the local dimension is 23 for both patches. Middle: Four patches case. Error is 4.76%. Global dimension is 47 and the local dimension is 12, 13, 13, and 12 respectively. Bottom: Eight patches case. Error is 4.42%. Global dimension is 49 and the local dimension is 7 for all patches.

For this translational invariant Poisson kernel, the semidefinite relaxation of SPCA (problem (5.27)) also gives satisfactory sparse approximation in the sense of problem (5.26). Numerical tests show that when $\mu < 2$, SPCA tends to put too much weight on the sparsity and it leads to poor approximation to A (over 90% error). In Figure 5.14 we plot 47 physical modes selected out of 513 columns of W, with $\mu = 2.7826$. The error is 4.94%. We also show 5 out of them on the right panel. Note that we have used the translation invariance property in selecting the columns of W.

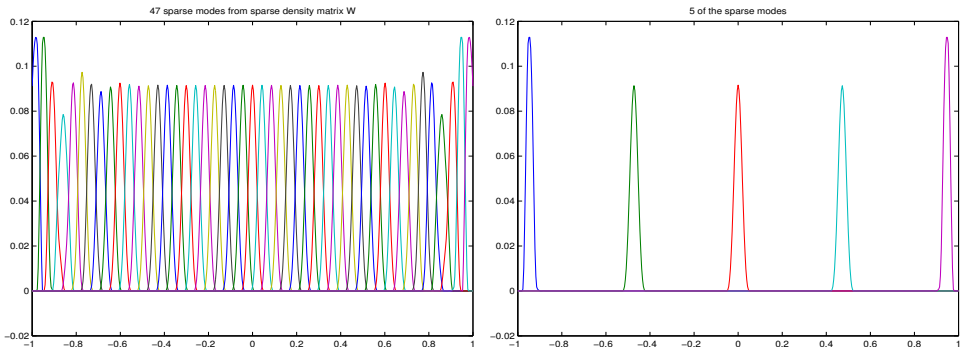


Figure 5.14: Sparse PCA: $\mu = 2.7826$. We specifically choose 47 columns out of W and show all of them on the left side and 5 of them on the right side.

Chapter 6

CONCLUDING DISCUSSIONS

The problem of approximating a positive semidefinite (PSD) operator with sparse/localized basis functions is important in both the physical and data sciences. In this thesis, we present two different methods, the sparse operator compression (Sparse OC) and the intrinsic sparse decomposition (ISMD), to achieve this goal. These two methods work well for different kinds of PSD operators, and they look for different forms of approximations to give the most accurate approximation of a given operator.

Given a PSD operator $\mathcal{K}$ and a positive integer n, the Sparse OC looks for sparse/localized basis functions $\Psi := [\psi_1, \psi_2, \ldots, \psi_n]$ and a PSD matrix $\Sigma \in \mathbb{R}^{n\times n}$ to approximate $\mathcal{K}$, i.e., $\mathcal{K} \approx \Psi\Sigma\Psi^T$. The localized basis functions Ψ are computed in a decoupled way by solving energy minimizing problems on local patches. We have shown that the Sparse OC works well for the solution operator of a large class of elliptic operators with rough coefficients. For strongly elliptic operators of order $2k$ ($k \geq 1$), we have proved that with support size $O(h\log(1/h))$, localized basis functions constructed by the Sparse OC can be used to compress higher order elliptic operators with the optimal compression rate $O(h^{2k})$. For second order elliptic operators with high contrast coefficients, we use the Sparse OC to construct localized basis functions, such that the error (in the energy norm) of the corresponding finite element solution is of order h and is independent of the contrast. Moreover, for the two-phase coefficient model, we have shown that the support diameter of basis functions can be as small as $h\left(\log(1/h) + \log\left(\frac{a_{max}}{a_{min}}\right)\right)$. We have also explored other applications of the Sparse OC. 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. In the application of compressing Hamiltonians in quantum physics, our localized basis functions achieve nearly optimal localization and the optimal operator compression rate simultaneously.

Given a PSD operator $\mathcal{K}$ of rank n, the ISMD looks for a decomposition

$\mathcal{K} = \Psi\Psi^T$ in which the n basis functions $\Psi := [\psi_1, \psi_2, \ldots, \psi_n]$ are required to be as sparse as possible. Instead of minimizing the total number of nonzero entries of the basis functions, the ISMD minimizes the total patchwise sparseness with a prescribed domain partition. The ISMD is equivalent to the eigendecomposition for the coarsest partition and recovers the pivoted Cholesky decomposition for the finest partition. If the partition is regular-sparse with respect to the matrix to be decomposed, we have proved that the ISMD gives the optimal patchwise sparse decomposition. We have also proved that as long as the partition is regular-sparse, the decomposed modes become sparser (in the sense of l^0 norm) as the partition is refined. Finally, we have provided a result on perturbation analysis of the ISMD based on the assumption that the partition is regular-sparse and the intrinsic sparse modes are identifiable with each other.

In our future work, we plan to further explore the topic of operator compression with localized basis functions in several directions.

First of all, we would like to further improve our numerical method for solving elliptic PDEs with high-contrast coefficients. Both the LOD-based methods (i.e. [110, 60]) and our method based on the Sparse OC in Chapter 4 construct localized multiscale finite element basis functions in the offline stage, and the same set of basis functions is used for all right hand sides in $L^2(D)$ or $L^2_a(D)$. However, this purely offline strategy suffers from the drawback that each high-conductivity inclusion/channel should have at least one associated basis function. This drawback originates in the methodology that the same set of basis functions is used for all possible right hand sides. If we can supplement the offline basis functions with some online basis functions that are computed locally within a patch for a new right hand side, we may not need so many basis functions. Specifically, the authors of [3] proposed an offline-online strategy to construct localized multiscale basis functions to achieve any given accuracy $\epsilon > 0$. They construct $\mathcal{O}(\log(1/h) + \log(1/\epsilon))$ basis functions per local patch in the offline stage. For a given right hand side, in addition to those offline prepared basis functions, another basis function is solved based on the right hand side per local patch. In total, there are $\mathcal{O}(\log(1/h) + \log(1/\epsilon)) + 1$ basis functions per patch, and the support size of every basis function is only $2h$. Although it is not proved in their original paper, our recent analysis shows that this offline-online strategy is very robust for high-contrast problems. More pre-

cisely, with $\mathcal{O}(\log(1/h)+\log(\frac{a_{max}}{a_{min}})+\log(1/\epsilon))+1$ localized basis functions per patch, one is able to achieve any prescribed accuracy $\epsilon > 0$. We believe that the offline-online strategy is a promising approach to solve the high-contrast problem, and we are now working to incorporate this idea into our Sparse OC framework.

Secondly, it is interesting to apply the Sparse OC to 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. The domain partition is a nontrivial difficulty when applying the Sparse OC to graph Laplacians. For a continuous elliptic operator on a physical domain D (as we considered in this thesis), the topology of the physical domain gives a natural regular domain partition. For a graph Laplacian, what is a "regular" partition on a graph? Is there an efficient algorithm to compute this "regular" partition? On one hand, a regular partition should cluster points that have similar response to a typical right hand side. The spectral partition (see [30] and references therein) gives such a partition but its computational complexity is not nearly linear. Many nearly linear complexity graph partitioning algorithms have been proposed in the literature; see e.g., [118] and references therein. One can then combine the existing graph-partitioning algorithm and our Sparse OC to design efficient linear system solves for graph Laplacians. On the other hand, in our Sparse OC, the local projection-type approximation property (the Poincare-type inequality on graphs) can serve as a concrete criterion to define and construct the "regular" graph partition. This offers us a seamless combination between the graph partitioning and the Sparse OC. Finally, it is worth mentioning other recent results on nearly linear complexity algorithms to solve linear systems with graph Laplacians, such as the lean algebraic multigrid (LAMG) [84] and the method of Spielman and Teng [119].

Thirdly, we would like to look into the trade-off between the approximation accuracy and basis localization in the operator compression problem. For a large class of elliptic operators, we have proved that one can achieve near optimality on both ends simultaneously in this trade-off. In ISMD, we examine an extreme case where we want full accuracy (decomposition instead of approximating), and we do not consider the accuracy-localization trade-off there. In general, the PSD operator $\mathcal{K}$ of interest may be neither the solution of an ellip-

tic operator nor a nearly low-rank operator, and there is a lot to do to fill in the gap. For the Sparse OC, although the construction of localized basis functions (see Eqn. (1.9)) can be theoretically applied to any PSD operator $\mathcal{K}$, knowledge of $\mathcal{K}^{-1}$ (i.e., the elliptic operator in this thesis) is currently required for an efficient computation of the associated H-norm. We are currently trying to design an efficient algorithm to construct these localized basis functions using only $\mathcal{K}$. For the ISMD, although we provide a heuristic algorithm (e.g. Algorithm 4) to make it work on arbitrary PSD operators, the complete resolution of the accuracy-localization trade-off in the ISMD setting (i.e., approximating in the form of $\mathcal{K} \approx \Psi\Psi^T$) requires a better problem formulation and a more robust algorithm.

Finally, inspired by the recent exciting advances in multiscale finite element methods and numerical homogenization, we are interested in using similar methodologies to solve other problems in the physical and data sciences. In particular, we are interested in applying the Sparse OC to construct localized Wannier functions for a Hamiltonian $\mathcal{H} = -\Delta + V(x)$ in quantum chemistry. There are two specific concerns in this application. First, what is the correct norm to measure the operator compression error? Our Sparse OC looks for localized basis functions Ψ to minimize the following operator compression error

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

which is reasonable in solving elliptic equations and in approximating the covariance operator. However, it is not clear that this is the correct norm for constructing localized Wannier functions. More discussions with domain experts are needed to figure out the correct norm. Second, 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 $V(x)$. We suggest an adaptive partition of the domain with variable local patch size. More precisely, if we pick the piecewise constant functions as the measurement functions, the local projection-type approximation property (see (2.11)) is written as

$$\inf_{c\in\mathbb{R}} \int_{\tau_i} (u(x) - c)^2 \mathrm{d}x \leq k_n \left(\int_{\tau_i} |\nabla u(x)|^2 \mathrm{d}x + \int_{\tau_i} V(x) u(x)^2 \mathrm{d}x \quad \forall u \in H^1(\tau_i) \right).$$

We will adapt the size of the local patch τ_i such that the two terms on the right hand side, i.e., $\int_{\tau_i} |\nabla u(x)|^2 \mathrm{d}x$ and $\int_{\tau_i} V(x) u(x)^2 \mathrm{d}x$, are of the same order.

In this way, we can optimize k_n, which is proportional to the final operator compression error.

BIBLIOGRAPHY

[1] D. Agarwal and B.-C. Chen. "Regression-based Latent Factor Models". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. Paris, France: ACM, 2009, pp. 19–28. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557029. URL: http://doi.acm.org/10.1145/1557019.1557029.

[2] O. Axelsson and V. A. Barker. "7. Iterative Solution of Finite Element Equations". In: *Finite Element Solution of Boundary Value Problems*. SIAM, 2001, pp. 327–421. DOI: 10.1137/1.9780898719253.ch7. eprint: http://epubs.siam.org/doi/pdf/10.1137/1.9780898719253.ch7. URL: http://epubs.siam.org/doi/abs/10.1137/1.9780898719253.ch7.

[3] I. Babuška and R. Lipton. "Optimal Local Approximation Spaces for Generalized Finite Element Methods with Application to Multiscale Problems". en. In: *Multiscale Modeling & Simulation* 9.1 (Jan. 2011), pp. 373–406. ISSN: 1540-3459, 1540-3467. DOI: 10.1137/100791051. URL: http://epubs.siam.org/doi/abs/10.1137/100791051.

[4] I. Babuška, F. Nobile, and R. Tempone. "A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data". en. In: *SIAM Journal on Numerical Analysis* 45.3 (Jan. 2007), pp. 1005–1034. ISSN: 0036-1429, 1095-7170. DOI: 10.1137/050645142. URL: http://epubs.siam.org/doi/abs/10.1137/050645142.

[5] I. Babuška and J. E. Osborn. "Can a Finite Element Method Perform Arbitrarily Badly?" In: *Mathematics of Computation* 69.230 (2000), pp. 443–462. ISSN: 00255718, 10886842. DOI: 10.1090/S0025-5718-99-01085-6. URL: http://www.jstor.org/stable/2584886.

[6] I. Babuška and J. E. Osborn. "Generalized finite element methods: their performance and their relation to mixed methods". In: *SIAM Journal on Numerical Analysis* 20.3 (1983), pp. 510–536.

[7] I. Babuška, R. Tempone, and G. E. Zouraris. "Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations". en. In: *SIAM Journal on Numerical Analysis* 42.2 (Jan. 2004), pp. 800–825. ISSN: 0036-1429, 1095-7170. DOI: 10.1137/S0036142902418680. URL: http://epubs.siam.org/doi/abs/10.1137/S0036142902418680.

[8] M. Bachmayr, A. Cohen, and G. Migliorati. "Representations of Gaussian random fields and approximation of elliptic PDEs with lognormal coefficients". In: *arXiv preprint arXiv:1603.05559* (2016).

[9] A. Barth, C. Schwab, and N. Zollinger. "Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients". In: *Numerische Mathematik* 119.1 (2011), pp. 123–161.

[10] M. Bebendorf. "Low-Rank Approximation of Elliptic Boundary Value Problems with High-Contrast Coefficients". In: *SIAM Journal on Mathematical Analysis* 48.2 (2016), pp. 932–949. DOI: 10.1137/140991030. eprint: http://dx.doi.org/10.1137/140991030. URL: http://dx.doi.org/10.1137/140991030.

[11] M. Bebendorf and W. Hackbusch. "Existence of H-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞ coefficients". In: *Numerische Mathematik* 95.1 (2003), pp. 1–28. ISSN: 0945-3245. DOI: 10.1007/s00211-002-0445-6. URL: http://dx.doi.org/10.1007/s00211-002-0445-6.

[12] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*. Vol. 5. North-Holland Publishing Company Amsterdam, 1978. ISBN: 9780080875262.

[13] L. Berlyand and H. Owhadi. "Flux Norm Approach to Finite Dimensional Homogenization Approximations with Non-Separated Scales and High Contrast". en. In: *Archive for Rational Mechanics and Analysis* 198.2 (Nov. 2010), pp. 677–721. ISSN: 0003-9527, 1432-0673. DOI: 10.1007/s00205-010-0302-1. URL: http://link.springer.com/10.1007/s00205-010-0302-1.

[14] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite optimization and convex algebraic geometry*. Vol. 13. Siam, 2013.

[15] V. I. Bogachev and V. I. Bogachev. *Gaussian measures*. Vol. 62. American Mathematical Society Providence, 1998.

[16] D. Bolin and F. Lindgren. "Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping". In: *The Annals of Applied Statistics* (2011), pp. 523–550.

[17] A. Bunse-Gerstner, R. Byers, and V. Mehrmann. "Numerical Methods for Simultaneous Diagonalization". In: *SIAM Journal on Matrix Analysis and Applications* 14.4 (1993), pp. 927–949. DOI: 10.1137/0614062. eprint: http://dx.doi.org/10.1137/0614062. URL: http://dx.doi.org/10.1137/0614062.

[18] V. M. CALO, Y. EFENDIEV, and J. GALVIS. "ASYMPTOTIC EXPANSIONS FOR HIGH-CONTRAST ELLIPTIC EQUATIONS". In: *Mathematical Models and Methods in Applied Sciences* 24.03 (2014), pp. 465–494. DOI: 10.1142/S0218202513500565. eprint: http://www.worldscientific.com/doi/pdf/10.1142/S0218202513500565. URL: http://www.worldscientific.com/doi/abs/10.1142/S0218202513500565.

[19] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (Feb. 2006), pp. 489–509. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.862083.

[20] E. J. Candès and B. Recht. "Exact Matrix Completion via Convex Optimization". In: *Foundations of Computational Mathematics* 9.6 (2009), p. 717. ISSN: 1615-3383. DOI: 10.1007/s10208-009-9045-5. URL: http://dx.doi.org/10.1007/s10208-009-9045-5.

[21] E. J. Candès et al. "Phase Retrieval via Matrix Completion". In: *SIAM Journal on Imaging Sciences* 6.1 (2013), pp. 199–225. DOI: 10.1137/110848074. eprint: http://dx.doi.org/10.1137/110848074. URL: http://dx.doi.org/10.1137/110848074.

[22] E. J. Candès et al. "Robust Principal Component Analysis?" In: *Journal of the ACM* 58.3 (June 2011), 11:1–11:37. ISSN: 0004-5411. DOI: 10.1145/1970392.1970395. URL: http://doi.acm.org/10.1145/1970392.1970395.

[23] J. F. Cardoso and A. Souloumiac. "Blind beamforming for non-Gaussian signals". In: *IEE Proceedings F - Radar and Signal Processing* 140.6 (Dec. 1993), pp. 362–370. ISSN: 0956-375X. DOI: 10.1049/ip-f-2.1993.0054.

[24] J.-F. Cardoso. *Perturbation of joint diagonalizers*. Tech. rep. 94D023. Paris: Signal Department, Telecom Paris, 1994. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.3&rep=rep1&type=pdf.

[25] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. "Latent variable graphical model selection via convex optimization". en. In: *The Annals of Statistics* 40.4 (Aug. 2012), pp. 1935–1967. ISSN: 0090-5364. DOI: 10.1214/11-AOS949. URL: http://projecteuclid.org/euclid.aos/1351602527.

[26] V. Chandrasekaran et al. "Rank-Sparsity Incoherence for Matrix Decomposition". en. In: *SIAM Journal on Optimization* 21.2 (Apr. 2011), pp. 572–596. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/090761793. URL: http://epubs.siam.org/doi/abs/10.1137/090761793.

[27] Y. Chen et al. "Local Polynomial Chaos Expansion for Linear Differential Equations with High Dimensional Random Inputs". In: *SIAM Journal on Scientific Computing* 37.1 (2015), A79–A102. URL: http://dx.doi.org/10.1137/140970100.

[28] Y. Chen et al. "Local Polynomial Chaos Expansion for Linear Differential Equations with High Dimensional Random Inputs". In: *SIAM Journal on Scientific Computing* 37.1 (2015), A79–A102. DOI: 10.1137/

140970100. eprint: http://dx.doi.org/10.1137/140970100. URL: http://dx.doi.org/10.1137/140970100.

[29] E. Chung, Y. Efendiev, and T.-Y. Hou. "Adaptive Multiscale Model Reduction with Generalized Multiscale Finite Element Methods". In: *JCP* 320 (2016), pp. 69–95.

[30] F. R. K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, Dec. 1996. ISBN: 0821803158. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20%5C&path=ASIN/0821803158.

[31] P. G. Ciarlet. *The finite element method for elliptic problems*. Vol. 40. Siam, 2002.

[32] K. A. Cliffe et al. "Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients". In: *Computing and Visualization in Science* 14.1 (2011), pp. 3–15.

[33] A. Cohen, R. DeVore, and C. Schwab. "Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs". In: *Foundations of Computational Mathematics* 10.6 (2010), pp. 615–646. URL: http://link.springer.com/article/10.1007/s10208-010-9072-2.

[34] A. Cohen, R. Devore, and C. Schwab. "Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's". In: *Analysis and Applications* 9.01 (2011), pp. 11–47. URL: http://www.worldscientific.com/doi/abs/10.1142/S0219530511001728.

[35] R. Coifman, V. Rokhlin, and S. Wandzura. "The fast multipole method for the wave equation: a pedestrian prescription". In: *IEEE Antennas and Propagation Magazine* 35 (June 1993), pp. 7–12. DOI: 10.1109/74.250128.

[36] S. Dahlke, E. Novak, and W. Sickel. "Optimal approximation of elliptic problems by linear and nonlinear mappings I". In: *Journal of Complexity* 22.1 (2006), pp. 29–49. ISSN: 0885-064X. DOI: http://dx.doi.org/10.1016/j.jco.2005.06.005. URL: http://www.sciencedirect.com/science/article/pii/S0885064X05000609.

[37] A. d'Aspremont et al. "A Direct Formulation for sparse PCA Using Semidefinite Programming". In: *SIAM Review* 49.3 (2007), pp. 434–448. DOI: 10.1137/050645506.

[38] M. D'Elia and M. Gunzburger. "Coarse-Grid Sampling Interpolatory Methods for Approximating Gaussian Random Fields". In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1 (2013), pp. 270–296.

[39] D. L. Donoho. "Compressed Sensing". In: *IEEE Trans. Inf. Theor.* 52.4 (Apr. 2006), pp. 1289–1306. ISSN: 0018-9448. DOI: 10.1109/TIT.2006.871582. URL: http://dx.doi.org/10.1109/TIT.2006.871582.

[40] A. Doostan and G. Iaccarino. "A least-squares approximation of partial differential equations with high-dimensional random inputs". In: *Journal of Computational Physics* 228.12 (2009), pp. 4332–4345.

[41] A. Doostan and H. Owhadi. "A non-adapted sparse approximation of PDEs with stochastic inputs". In: *Journal of Computational Physics* 230.8 (2011), pp. 3015–3034.

[42] W. E, T. Li, and J. Lu. "Localized bases of eigensubspaces and operator compression". In: *Proceedings of the National Academy of Sciences* 107.4 (2010), pp. 1273–1278. DOI: 10.1073/pnas.0913345107. eprint: http://www.pnas.org/content/107/4/1273.full.pdf. URL: http://www.pnas.org/content/107/4/1273.abstract.

[43] Y. Efendiev, J. Galvis, and T. Y. Hou. "Generalized multiscale finite element methods (GMsFEM)". en. In: *Journal of Computational Physics* 251 (Oct. 2013), pp. 116–135. ISSN: 00219991. DOI: 10.1016/j.jcp.2013.04.045. URL: http://linkinghub.elsevier.com/retrieve/pii/S0021999113003392.

[44] Y. Efendiev, J. Galvis, and X.-H. Wu. "Multiscale finite element methods for high-contrast problems using local spectral basis functions". In: *Journal of Computational Physics* 230.4 (2011), pp. 937–955. ISSN: 0021-9991. DOI: http://dx.doi.org/10.1016/j.jcp.2010.09.026. URL: //www.sciencedirect.com/science/article/pii/S0021999110005292.

[45] Y. Efendiev and T. Y. Hou. *Multiscale Finite Element Methods: Theory and Applications*. Springer, New York, 2009.

[46] J. Galvis and Y. Efendiev. "Domain Decomposition Preconditioners for Multiscale Flows in High Contrast Media: Reduced Dimension Coarse Spaces". en. In: *Multiscale Modeling & Simulation* 8.5 (Jan. 2010), pp. 1621–1644. ISSN: 1540-3459, 1540-3467. DOI: 10.1137/100790112. URL: http://epubs.siam.org/doi/abs/10.1137/100790112.

[47] J. Galvis and Y. Efendiev. "Domain Decomposition Preconditioners for Multiscale Flows in High-Contrast Media". In: *Multiscale Modeling & Simulation* 8.4 (2010), pp. 1461–1483. DOI: 10.1137/090751190. eprint: http://dx.doi.org/10.1137/090751190. URL: http://dx.doi.org/10.1137/090751190.

[48] C. Gao and B. E. Engelhardt. "A sparse factor analysis model for high dimensional latent spaces". In: *NIPS: Workshop on Analysis Operator Learning vs. Dictionary Learning: Fraternal Twins in Sparse Modeling*. 2012.

[49] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. New York, NY, USA: Springer-Verlag New York, Inc., 1991. ISBN: 0-387-97456-3.

[50] M. Ghommem et al. "Mode decomposition methods for flows in high-contrast porous media. Global–local approach". en. In: *Journal of Computational Physics* 253 (Nov. 2013), pp. 226–238. ISSN: 00219991. DOI: 10.1016/j.jcp.2013.06.033. URL: http://linkinghub.elsevier.com/retrieve/pii/S0021999113004634.

[51] M. B. Giles. "Multilevel monte carlo path simulation". In: *Operations Research* 56.3 (2008), pp. 607–617.

[52] C. J. Gittelson. "Representation of Gaussian fields in series with independent coefficients". In: *IMA Journal of Numerical Analysis* 32.1 (2012), pp. 294–319.

[53] T. Gneiting, W. Kleiber, and M. Schlather. "Matérn cross-covariance functions for multivariate random fields". In: *Journal of the American Statistical Association* (2012).

[54] S. Goedecker. "Linear scaling electronic structure methods". In: *Reviews of Modern Physics* 71.4 (1999), p. 1085.

[55] G. H. Golub and C. F. Van Loan. *Matrix computations (3rd ed.)* Baltimore, MD, USA: Johns Hopkins University Press, 1996. ISBN: 0801854148. URL: http://portal.acm.org/citation.cfm?id=248979.

[56] L. Grasedyck, I. Greff, and S. Sauter. "The AL Basis for the Solution of Elliptic Problems in Heterogeneous Media". In: *Multiscale Modeling & Simulation* 10.1 (2012), pp. 245–258. DOI: 10.1137/11082138X. eprint: http://dx.doi.org/10.1137/11082138X. URL: http://dx.doi.org/10.1137/11082138X.

[57] L. Greengard and V. Rokhlin. "A new version of the Fast Multipole Method for the Laplace equation in three dimensions". In: *Acta Numerica* 6 (1997), pp. 229–269. DOI: 10.1017/S0962492900002725.

[58] P. Guttorp and T. Gneiting. "Studies in the history of probability and statistics XLIX On the Matern correlation family". In: *Biometrika* 93.4 (2006), pp. 989–995.

[59] W. Hackbusch. "A Sparse Matrix Arithmetic Based on H-Matrices. Part I: Introduction to H-Matrices". In: *Computing* 62.2 (1999), pp. 89–108. ISSN: 1436-5057. DOI: 10.1007/s006070050015. URL: http://dx.doi.org/10.1007/s006070050015.

[60] F. Hellman and A. Målqvist. "Contrast independent localization of multiscale problems". In: *arXiv preprint arXiv:1610.07398* (2016).

[61] P. Henning and D. Peterseim. "Oversampling for the Multiscale Finite Element Method". In: *Multiscale Modeling & Simulation* 11.4 (2013), pp. 1149–1175. DOI: 10.1137/120900332. eprint: http://dx.doi.org/10.1137/120900332. URL: http://dx.doi.org/10.1137/120900332.

[62] K. Höllig, C. Apprich, and A. Streit. "Introduction to the Web-method and its applications". In: *Advances in Computational Mathematics* 23.1 (2005), pp. 215–237. ISSN: 1572-9044. DOI: 10.1007/s10444-004-1811-y. URL: http://dx.doi.org/10.1007/s10444-004-1811-y.

[63] T. Y. Hou and P. Liu. "Optimal local multi-scale basis functions for linear elliptic equations with rough coefficients". In: *Discrete and Continuous Dynamical Systems, A* 36.8 (2016), pp. 4451–4476. DOI: doi:10.3934/dcds.2016.36.4451.

[64] T. Y. Hou, Q. Li, and P. Zhang. "A Sparse Decomposition of Low Rank Symmetric Positive Semidefinite Matrices". In: *Multiscale Modeling & Simulation* 15.1 (2017), pp. 410–444. DOI: 10.1137/16M107760X. eprint: http://dx.doi.org/10.1137/16M107760X. URL: http://dx.doi.org/10.1137/16M107760X.

[65] T. Y. Hou, Q. Li, and P. Zhang. "Exploring the Locally Low Dimensional Structure in Solving Random Elliptic PDEs". In: *Multiscale Modeling & Simulation* 15.2 (2017), pp. 661–695. DOI: 10.1137/16M1077611. eprint: http://dx.doi.org/10.1137/16M1077611. URL: http://dx.doi.org/10.1137/16M1077611.

[66] T. Y. Hou and X.-H. Wu. "A Multiscale Finite Element Method for Elliptic Problems in Composite Materials and Porous Media". In: *Journal of Computational Physics* 134.1 (1997), pp. 169–189. ISSN: 0021-9991. DOI: http://dx.doi.org/10.1006/jcph.1997.5682. URL: http://www.sciencedirect.com/science/article/pii/S0021999197956825.

[67] T. Y. Hou, X.-H. Wu, and Y. Zhang. "Removing the Cell Resonance Error in the Multiscale Finite Element Method via a Petrov-Galerkin Formulation". In: *Communications in Mathematical Sciences* 2.2 (June 2004), pp. 185–205. URL: http://projecteuclid.org/euclid.cms/1109706534.

[68] T. Y. Hou and P. Zhang. "Sparse operator compression of higher order elliptic operators with rough coefficients". In: *Research in the Mathematical Sciences* (2017). in press.

[69] T. Y. Hou et al. "Wiener Chaos expansions and numerical solutions of randomly forced equations of fluid mechanics". en. In: *Journal of Computational Physics* 216.2 (Aug. 2006), pp. 687–706. ISSN: 00219991. DOI: 10.1016/j.jcp.2006.01.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S0021999106000064.

[70] T. J. Hughes et al. "The variational multiscale method—a paradigm for computational mechanics". In: *Computer methods in applied mechanics and engineering* 166.1 (1998), pp. 3–24.

[71] L. Jacob, G. Obozinski, and J.-P. Vert. "Group lasso with overlap and graph lasso". In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 433–440.

[72] R. Jenatton, G. Obozinski, and F. R. Bach. "Structured Sparse Principal Component Analysis". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*. Ed. by Y. W. Teh and D. M. Titterington. Vol. 9. 2010, pp. 366–373. URL: http://www.jmlr.org/proceedings/papers/v9/jenatton10a/jenatton10a.pdf.

[73] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. "A Modified Principal Component Technique Based on the LASSO". In: *Journal of Computational and Graphical Statistics* 12.3 (2003), pp. 531–547. ISSN: 10618600. URL: http://www.jstor.org/stable/1391037.

[74] I. Jolliffe, M. Uddin, and S. Vines. "Simplified EOFs–Three alternatives to rotation". In: *Climate Research* 20.3 (Apr. 2002), pp. 271–279. DOI: https://doi.org/10.3354/cr020271. URL: http://oro.open.ac.uk/3866/.

[75] K. Karhunen. *Űber lineare Methoden in der Wahrscheinlichkeitsrechnung*. Annales Academiae scientiarum Fennicae: Mathematica - Physica. Universitat Helsinki, 1947. URL: http://books.google.com/books?id=bGUUAQAAIAAJ.

[76] W. Kohn. "Image of the Fermi Surface in the Vibration Spectrum of a Metal". In: *Phys. Rev. Lett.* 2 (9 May 1959), pp. 393–394. DOI: 10.1103/PhysRevLett.2.393. URL: http://link.aps.org/doi/10.1103/PhysRevLett.2.393.

[77] W. Krzanowski and F. Marriott. *Multivariate Analysis: Kendall's Library of Statistics, Volume 2*. Kendall's advanced theory of statistics. Wiley, 1995. ISBN: 9780340593257. URL: https://books.google.com/books?id=jOY0jnXSMvQC.

[78] R. Lai, J. Lu, and S. Osher. "Density matrix minimization with L_1 regularization". In: *Communications in Mathematical Sciences* 13.8 (2015). DOI: http://dx.doi.org/10.4310/CMS.2015.v13.n8.a6.

[79] R. Lai and S. Osher. "A splitting method for orthogonality constrained problems". In: *Journal of Scientific Computing* 58.2 (2014), pp. 431–449.

[80] D. D. Lee and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (Oct. 1999), pp. 788–791. URL: http://dx.doi.org/10.1038/44565.

[81] H. Lee et al. "Efficient Sparse Coding Algorithms". In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS'06. Canada: MIT Press, 2006, pp. 801–808. URL: http://dl.acm.org/citation.cfm?id=2976456.2976557.

[82] R. B. Lehoucq and D. C. Sorensen. "Deflation Techniques for an Implicitly Restarted Arnoldi Iteration". In: *SIAM Journal on Matrix Analysis and Applications* 17.4 (1996), pp. 789–821. DOI: 10.1137/S0895479895281484. eprint: http://dx.doi.org/10.1137/S0895479895281484. URL: http://dx.doi.org/10.1137/S0895479895281484.

[83] F. Lindgren, H. Rue, and J. Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498.

[84] O. E. Livne and A. Brandt. "Lean Algebraic Multigrid (LAMG): Fast Graph Laplacian Linear Solver". In: *SIAM Journal on Scientific Computing* 34.4 (2012), B499–B522. DOI: 10.1137/110843563. eprint: http://dx.doi.org/10.1137/110843563. URL: http://dx.doi.org/10.1137/110843563.

[85] M. Loève. *Probability Theory I, 4th ed.* Graduate Texts in Mathematics. 45. Springer, 1977. ISBN: 9780387902104.

[86] C. Lucas. "LAPACK-style codes for Level 2 and 3 pivoted Cholesky factorizations". In: *LAPACK Working* (2004).

[87] X. Luo. "High dimensional low rank and sparse covariance matrix estimation via convex minimization". In: *arXiv preprint arXiv:1111.1133* (2011). URL: http://www.researchgate.net/publication/51952042_High_Dimensional_Low_Rank_and_Sparse_Covariance_Matrix_Estimation_viaConvex_Minimization/file/e0b4952127dbaebd4b.pdf.

[88] J. Mairal et al. "Online Learning for Matrix Factorization and Sparse Coding". In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 19–60. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1756008.

[89] A. Målqvist and D. Peterseim. "Localization of elliptic multiscale problems". In: *Mathematics of Computation* 83.290 (2014), pp. 2583–2603.

[90] N. Marzari and D. Vanderbilt. "Maximally localized generalized Wannier functions for composite energy bands". In: *Phys. Rev. B* 56 (20 Nov. 1997), pp. 12847–12865. DOI: 10.1103/PhysRevB.56.12847. URL: http://link.aps.org/doi/10.1103/PhysRevB.56.12847.

[91] N. Marzari et al. "Maximally localized Wannier functions: Theory and applications". In: *Rev. Mod. Phys.* 84 (4 Oct. 2012), pp. 1419–1475. DOI: 10.1103/RevModPhys.84.1419. URL: http://link.aps.org/doi/10.1103/RevModPhys.84.1419.

[92] B. Matérn. *Spatial variation*. Vol. 36. Springer Science & Business Media, 2013.

[93] J. Melenk. "On n-Widths for Elliptic Problems". In: *Journal of Mathematical Analysis and Applications* 247.1 (2000), pp. 272–289. ISSN: 0022-247X. DOI: http://dx.doi.org/10.1006/jmaa.2000.6862. URL: http://www.sciencedirect.com/science/article/pii/S0022247X00968628.

[94] P. Ming and X. Yue. "Numerical methods for multiscale elliptic problems". In: *Journal of Computational Physics* 214.1 (2006), pp. 421–445.

[95] S. M. Nikol'skii. "Imbedding Theorems for Different Metrics and Dimensions". In: *Approximation of Functions of Several Variables and Imbedding Theorems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1975, pp. 231–260. ISBN: 978-3-642-65711-5. DOI: 10.1007/978-3-642-65711-5_7. URL: http://dx.doi.org/10.1007/978-3-642-65711-5_7.

[96] F. Nobile, R. Tempone, and C. G. Webster. "A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data". en. In: *SIAM Journal on Numerical Analysis* 46.5 (Jan. 2008), pp. 2309–2345. ISSN: 0036-1429, 1095-7170. DOI: 10.1137/060663660. URL: http://epubs.siam.org/doi/abs/10.1137/060663660.

[97] F. Nobile, R. Tempone, and C. G. Webster. "An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data". en. In: *SIAM Journal on Numerical Analysis* 46.5 (Jan. 2008), pp. 2411–2442. ISSN: 0036-1429, 1095-7170. DOI: 10.1137/070680540. URL: http://epubs.siam.org/doi/abs/10.1137/070680540.

[98] H. Owhadi. "Bayesian Numerical Homogenization". In: *Multiscale Modeling & Simulation* 13.3 (2015), pp. 812–828.

[99] H. Owhadi. "Multigrid with rough coefficients and Multiresolution operator decomposition from Hierarchical Information Games". In: *SIAM Review* 59.1 (2017), pp. 99–149.

[100] H. Owhadi and C. Scovel. "Universal Scalable Robust Solvers from Computational Information Games and fast eigenspace adapted Multiresolution Analysis". In: *arXiv preprint arXiv:1703.10761* (2017).

[101] H. Owhadi and L. Zhang. "Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients". In: *arXiv:1606.07686v1* (2016).

[102] H. Owhadi and L. Zhang. "Homogenization of Parabolic Equations with a Continuum of Space and Time Scales". en. In: *SIAM Journal on Numerical Analysis* 46.1 (Jan. 2008), pp. 1–36. ISSN: 0036-1429, 1095-7170. DOI: 10.1137/060670420. URL: http://epubs.siam.org/doi/abs/10.1137/060670420.

[103] H. Owhadi and L. Zhang. "Localized Bases for Finite-Dimensional Homogenization Approximations with Nonseparated Scales and High Contrast". en. In: *Multiscale Modeling & Simulation* 9.4 (Oct. 2011), pp. 1373–1398. ISSN: 1540-3459, 1540-3467. DOI: 10.1137/100813968. URL: http://epubs.siam.org/doi/abs/10.1137/100813968.

[104] H. Owhadi and L. Zhang. "Metric-based upscaling". In: *Communications on Pure and Applied Mathematics* 60.5 (2007), pp. 675–723. ISSN: 1097-0312. DOI: 10.1002/cpa.20163. URL: http://dx.doi.org/10.1002/cpa.20163.

[105] H. Owhadi, L. Zhang, and L. Berlyand. "Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 48.02 (2014), pp. 517–552. ISSN: 1290-3841. DOI: 10.1051/m2an/2013118. URL: http://www.esaim-m2an.org/article_S0764583X13001180.

[106] H. Owhadi et al. "Optimal Uncertainty Quantification". In: *SIAM Review* 55.2 (2013), pp. 271–345. DOI: 10.1137/10080782X. eprint: http://dx.doi.org/10.1137/10080782X. URL: http://dx.doi.org/10.1137/10080782X.

[107] V. Ozoliņš et al. "Compressed modes for variational problems in mathematics and physics". In: *Proceedings of the National Academy of Sciences* 110.46 (2013), pp. 18368–18373. URL: http://www.pnas.org/content/110/46/18368.short.

[108] V. Ozoliņš et al. "Compressed plane waves yield a compactly supported multiresolution basis for the Laplace operator". In: *Proceedings of the National Academy of Sciences* 111.5 (2014), pp. 1691–1696. URL: http://www.pnas.org/content/111/5/1691.short.

[109] D. Peterseim. "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*. Ed. by G. R. Barrenechea et al. Cham: Springer International Publishing, 2016, pp. 343–369. ISBN: 978-3-319-41640-3. DOI: 10.1007/978-3-319-41640-3_11. URL: http://dx.doi.org/10.1007/978-3-319-41640-3_11.

[110] D. Peterseim and R. Scheichl. "Robust numerical upscaling of elliptic multiscale problems at high contrast". In: *Computational Methods in Applied Mathematics* 16.4 (2016), pp. 579–603.

[111] E. Prodan and W. Kohn. "Nearsightedness of electronic matter". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.33 (2005), pp. 11635–11638. DOI: 10.1073/pnas.0505436102. eprint: http://www.pnas.org/content/102/33/11635.full.pdf.

[112] B. Recht, M. Fazel, and P. A. Parrilo. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". In: *SIAM Review* 52.3 (2010), pp. 471–501. DOI: 10.1137/070697835. eprint: http://dx.doi.org/10.1137/070697835. URL: http://dx.doi.org/10.1137/070697835.

[113] M. Renardy and R. C. Rogers. *An introduction to partial differential equations*. Vol. 13. Springer Science & Business Media, 2006.

[114] B. Reznick. "On Hilbert's construction of positive polynomials". In: *arXiv preprint arXiv:0707.2156* (2007).

[115] W. Rudin. "Sums of squares of polynomials". In: *The American Mathematical Monthly* 107.9 (2000), pp. 813–821.

[116] H. Schaeffer et al. "Sparse dynamics for partial differential equations". In: *Proceedings of the National Academy of Sciences* 110.17 (2013), pp. 6634–6639. DOI: 10.1073/pnas.1302752110. eprint: http://www.pnas.org/content/110/17/6634.full.pdf. URL: http://www.pnas.org/content/110/17/6634.abstract.

[117] D. C. Sorensen. "Implicit Application of Polynomial Filters in a K-step Arnoldi Method". In: *SIAM J. Matrix Anal. Appl.* 13.1 (Jan. 1992), pp. 357–385. ISSN: 0895-4798. DOI: 10.1137/0613025. URL: http://dx.doi.org/10.1137/0613025.

[118] D. A. Spielman and S.-H. Teng. "A Local Clustering Algorithm for Massive Graphs and Its Application to Nearly Linear Time Graph Partitioning". In: *SIAM Journal on Computing* 42.1 (2013), pp. 1–26. DOI: 10.1137/080744888. eprint: http://dx.doi.org/10.1137/080744888. URL: http://dx.doi.org/10.1137/080744888.

[119] D. A. Spielman and S.-H. Teng. "Nearly Linear Time Algorithms for Preconditioning and Solving Symmetric, Diagonally Dominant Linear Systems". In: *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), pp. 835–885. DOI: 10.1137/090771430. eprint: http://dx.doi.org/10.1137/090771430. URL: http://dx.doi.org/10.1137/090771430.

[120] I. Sraj et al. "Coordinate transformation and Polynomial Chaos for the Bayesian inference of a Gaussian process with parametrized prior covariance function". In: *Computer Methods in Applied Mechanics and Engineering* 298 (2016), pp. 205–228.

[121] M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[122] G. Stewart. "On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems". In: *SIAM Review* 19.4 (Jan. 1977), pp. 634–662. DOI: 10.1137/1019104. URL: http://dx.doi.org/10.1137/1019104.

[123] T. Strouboulis, K. Copps, and I. Babuška. "The generalized finite element method". In: *Computer methods in applied mechanics and engineering* 190.32 (2001), pp. 4081–4193.

[124] R. G. Swan. "Hilbert's theorem on positive ternary quartics". In: *Contemporary Mathematics* 272 (2000), pp. 287–292.

[125] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[126] R. Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.00771.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x.

[127] D. Uherka and A. M. Sergott. "On the continuous dependence of the roots of a polynomial on its coefficients". In: *The American mathematical monthly* 84.5 (1977), pp. 368–370.

[128] V. Q. Vu et al. "Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. Burges et al. 2013, pp. 2670–2678. URL: http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1250.pdf.

[129] A. Wagner et al. "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2 (Feb. 2012), pp. 372–386. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.112.

[130] G. H. Wannier. "The structure of electronic excitation levels in insulating crystals". In: *Physical Review* 52.3 (1937), p. 191.

[131] D. M. Witten, R. Tibshirani, and T. Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis". In: *Biostatistics, 10* (2009), pp. 515–534. URL: https://doi.org/10.1093/biostatistics/kxp008.

[132] D. Xiu and G. Karniadakis. "The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations". In: *SIAM Journal on Scientific Computing* 24.2 (2002), pp. 619–644. DOI: 10.1137/S1064827501387826. URL: http://dx.doi.org/10.1137/S1064827501387826.

[133] D. Xiu and J. S. Hesthaven. "High-Order Collocation Methods for Differential Equations with Random Inputs". In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1118–1139. DOI: 10.1137/040615201. eprint: http://dx.doi.org/10.1137/040615201. URL: http://dx.doi.org/10.1137/040615201.

[134] L. Yan, L. Guo, and D. Xiu. "Stochastic Collocation Algorithms using l1–Minimization". In: *International Journal for Uncertainty Quantification* 2.3 (2012), pp. 279–293.

[135] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[136] Y. Zhang, A. d'Aspremont, and L. El Ghaoui. "Sparse PCA: Convex relaxations, algorithms and applications". In: *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer, New York, 2012, pp. 915–940.

[137] H. Zou, T. Hastie, and R. Tibshirani. "Sparse Principal Component Analysis". In: *Journal of Computational and Graphical Statistics* 15.2 (2006), pp. 265–286. ISSN: 10618600. URL: http://www.jstor.org/stable/27594179.

A p p e n d i x A

SUPPLEMENTARY MATERIALS FOR THE SPARSE OC

A.1 Uniform Ellipticity v.s. Strong Ellipticity

Consider the following partial differential operator of order $2k$ defined in a bounded connected domain $S \subset \mathbb{R}^d$

$$\mathcal{L}(x;S) = (-1)^k \sum_{|\alpha|\le 2k} a_\alpha(x) D^\alpha \equiv (-1)^k \sum_{|\alpha|\le 2k} a_\alpha(x) \prod_{i=1}^{d} \partial_{x_i}^{\alpha_i}. \tag{A.1}$$

Its characteristic polynomial is defined as

$$p(\boldsymbol{\xi}) = \sum_{|\alpha|=2k} a_\alpha(x) \boldsymbol{\xi}^\alpha \equiv \sum_{|\alpha|=2k} a_\alpha(x) \prod_{i=1}^{d} \xi_i^{\alpha_i}. \tag{A.2}$$

The partial differential operator $\mathcal{L}(x;S)$ is called *uniformly elliptic* if there exists $\theta > 0$ such that

$$p(\boldsymbol{\xi}) = \sum_{|\alpha|=2k} a_\alpha(x) \boldsymbol{\xi}^\alpha \ge \theta |\boldsymbol{\xi}|^{2k}, \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, x \in S. \tag{A.3}$$

As defined in Definition 3.4.1, $\mathcal{L}(x;S)$ is *strongly elliptic* if there exists $b_{\sigma\gamma}(x)$ for all $|\sigma|, |\gamma| \le k$ such that

- we have

$$(-1)^k \sum_{|\alpha|\le 2k} a_\alpha(x) D^\alpha u = \sum_{|\sigma|,|\gamma|\le k} (-1)^{|\sigma|} D^\sigma (b_{\sigma\gamma}(x) D^\gamma u), \qquad \forall u \in C^{2k}(S), \tag{A.4}$$

- and there exists a constant $\theta > 0$ such that for any $x \in S$, we have

$$\sum_{|\sigma|,|\gamma|=k} b_{\sigma\gamma}(x) \zeta_\sigma \zeta_\gamma \ge \theta \sum_{|\sigma|=k} \zeta_\sigma^2, \qquad \forall \boldsymbol{\zeta} := [\zeta_\sigma]_{|\sigma|=k} \in \mathbb{R}^{\binom{k+d-1}{k}}. \tag{A.5}$$

A homogeneous polynomial $p(\boldsymbol{\xi})$ of degree $2k$ in the real d-dimensional vector $\boldsymbol{\xi}$ is sum-of-squares (SOS) if and only if there exist a finite number of polynomials, denoted as $g_1(\boldsymbol{\xi}), \ldots, g_n(\boldsymbol{\xi})$, such that

$$p(\boldsymbol{\xi}) = \sum_{i=1}^{n} (g_i(\boldsymbol{\xi}))^2.$$

The following lemma is a simple but useful property of the SOS polynomials.

Lemma A.1.1. *A d-variate homogeneous polynomial of degree $2k$, denoted as $p(\boldsymbol{\xi})$, is SOS if and only if there exists a symmetric positive semidefinite $B \in \mathbb{R}^{Q\times Q}$ such that*

$$p(\boldsymbol{\xi}) = \sum_{|\sigma|=|\gamma|=k} b_{\sigma\gamma}\boldsymbol{\xi}^{\sigma}\boldsymbol{\xi}^{\gamma},$$

where $Q := \binom{k+d-1}{k}$ is the number of the d-variate monomials of degree k.

Lemma A.1.1 can be proved by taking the eigen decomposition of B, and we will not provide the complete proof here. Using Lemma A.1.1, the strong ellipticity condition (A.5) is equivalent to the condition that there exists $\theta > 0$ such that for any $x \in S$

$$\sum_{|\sigma|=|\gamma|=k} b_{\sigma\gamma}(x)\boldsymbol{\xi}^{\sigma}\boldsymbol{\xi}^{\gamma} - \theta \sum_{|\sigma|=k} |\boldsymbol{\xi}|^{2k} = \text{an SOS polynomial.} \tag{A.6}$$

We point out that the characteristic polynomial $p(\boldsymbol{\xi})$, defined in Eqn. (A.2), keeps the same when we rewrite the operator in Eqn. (A.4). Therefore, a necessary condition for $\mathcal{L}$ to be *strongly elliptic* is that there exists $\theta > 0$ such that $p(\boldsymbol{\xi}) - \theta|\boldsymbol{\xi}|^{2k}$ is an SOS polynomial for any $x \in S$. In comparison, the *uniform ellipticity* condition (A.3) requires that the polynomial $p(\boldsymbol{\xi})-\theta|\boldsymbol{\xi}|^{2k}$ be nonnegative. Therefore, a direct application of the Hilbert's theorem (1988) on nonnegative polynomials and SOS polynomials (see e.g. [115, 124, 114]) leads to the following theorem.

Theorem A.1.1. *Suppose that $\mathcal{L} = (-1)^k \sum_{|\alpha|=2k} a_\alpha D^\alpha$ is a uniformly elliptic operator of order $2k$ with constant coefficients. Then in the following three cases, $\mathcal{L}$ is also strongly elliptic:*

- *$d = 1$ or 2 : one or two dimensional physical domain,*
- *$k = 1$: second order partial differential operators,*
- *$(d, k) = (3, 2)$: fourth order partial differential operators in 3 dimensional physical domain.*

For all other cases, i.e., $d \geq 3$ or $k \geq 2$ and $(d, k) \neq (3, 2)$, there exist uniformly elliptic operators with constant coefficients that are not strongly elliptic.

Proof. From the definition of uniform ellipticity, there exists $\theta > 0$ such that the following homogeneous polynomial with order $2k$ is nonnegative, i.e.,

$$\widetilde{p}(\boldsymbol{\xi}) = p(\boldsymbol{\xi}) - \theta|\boldsymbol{\xi}|^{2k} = \sum_{|\alpha|=2k} a_\alpha \boldsymbol{\xi}^\alpha - \theta|\boldsymbol{\xi}|^{2k} \geq 0.$$

Applying the Hilbert's theorem (1988), $\widetilde{p}(\boldsymbol{\xi})$ is also an SOS polynomial for the three cases above. Using Lemma A.1.1, there exists a symmetric positive semidefinite matrix $\widetilde{B} \in \mathbb{R}^{Q\times Q}$ such that $\widetilde{p}(\boldsymbol{\xi}) = \sum_{|\sigma|=|\gamma|=k} \widetilde{b}_{\sigma\gamma}\boldsymbol{\xi}^\sigma\boldsymbol{\xi}^\gamma$. Therefore, we obtain

$$p(\boldsymbol{\xi}) = \sum_{|\sigma|=|\gamma|=k} \left(\widetilde{b}_{\sigma\gamma} + \theta\binom{k}{\sigma}\delta_{\sigma,\gamma}\right)\boldsymbol{\xi}^\sigma\boldsymbol{\xi}^\gamma =: \sum_{|\sigma|=|\gamma|=k} b_{\sigma\gamma}\boldsymbol{\xi}^\sigma\boldsymbol{\xi}^\gamma,$$

where $\binom{k}{\sigma}$ is the multi-index combinatorial number, i.e., $(\sum_{i=1}^d x_i)^k = \sum_{|\sigma|=k}\binom{k}{\sigma}\boldsymbol{x}^\sigma$. Therefore, the elliptic operator can be written as

$$\mathcal{L}u = \sum_{|\sigma|,|\gamma|\leq k} (-1)^{|\sigma|} D^\sigma(b_{\sigma\gamma}D^\gamma u), \qquad \forall u \in C^{2k}(S).$$

Since $\widetilde{p}(\boldsymbol{\xi}) = p(\boldsymbol{\xi}) - \theta|\boldsymbol{\xi}|^{2k}$ is an SOS polynomial, we have proved that $\mathcal{L}$ is also strongly elliptic.

For all other cases, i.e., $d \geq 3$ or $k \geq 2$ and $(d,k) \neq (3,2)$, thanks to the Hilbert's theorem (1988), there is a nonnegative polynomial that is not SOS, denoted as $\widetilde{p}(\boldsymbol{\xi})$. Recall that given the number of variables d and degree $2k$, the set of nonnegative polynomials and the set of SOS polynomials are closed, convex cones. Then there exists $\lambda \in (0,1)$ such that $p_\lambda(\boldsymbol{\xi}) := \lambda\widetilde{p}(\boldsymbol{\xi}) + (1-\lambda)|\boldsymbol{\xi}|^{2k}$ is also nonnegative but not SOS. Finally, the elliptic operator with p_λ as its characteristic polynomial is uniformly elliptic but not strongly elliptic. □

When the coefficients of the elliptic operator $\mathcal{L}$ in Eqn. (A.1) are not constant, the coefficients should be smooth enough such that we can rewrite $\mathcal{L}$ in a divergence form as in Eqn. (A.4). Theorem A.1.2 guarantees that strongly ellipticity and uniformly ellipticity are equivalent for the case $k = 1$ and the case $d = 1$ or 2.

Theorem A.1.2. *Let $a_\alpha \in C^{|\alpha|-k}(\overline{S})$ for $k < |\alpha| \leq 2k$, $a_\alpha \in C(\overline{S})$ for $|\alpha| \leq k$, and $\mathcal{L}u = (-1)^k \sum_{|\alpha|\leq 2k} a_\alpha D^\alpha u$ for all $u \in C^{2k}(S)$. 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.*

Proof. The strategy is to first rewrite the highest order terms in a divergence form, and then to rewrite the lower order terms. For the case $d = 1$, thanks to $a_{2k} \in C^k(\overline{S})$, we can write

$$\mathcal{L}u \equiv (-1)^k \sum_{i=0}^{2k} a_i(x) \frac{\mathrm{d}^i u}{\mathrm{d}x^i} = (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} \left(a_{2k} \frac{\mathrm{d}^k u}{\mathrm{d}x^k} \right) + \widetilde{\mathcal{L}} u,$$

where the residual $\widetilde{\mathcal{L}}$ is an differential operator with order at most $2k-1$. Since the coefficients of $\widetilde{\mathcal{L}}$ are smooth enough, by Lemma 9.7 in [113], we can write $\widetilde{\mathcal{L}}$ in a divergence form as in Eqn. (A.4). The uniform ellipticity condition (A.3) and the strong ellipticity condition (A.5) are the same in this case, i.e., there exists $\theta > 0$ such that $a_{2k}(x) \geq \theta$ for any $x \in S$.

For the case $k = 1$, thanks to $a_\alpha \in C^1(\overline{S})$ for $|\alpha| = 2$, $\mathcal{L}$ can be rewritten in a divergence form as follows:

$$\mathcal{L}u \equiv -\sum_{i,j} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_i b_i(x) \frac{\partial u}{\partial x_i} + c(x)u = -\sum_{i,j} \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + \sum_i \widetilde{b}_i(x) \frac{\partial u}{\partial x_i} + c(x)u,$$

where $\widetilde{b}_i := b_i(x) + \sum_j \frac{\partial a_{ji}}{\partial x_j}$. The uniform ellipticity condition (A.3) and the strong ellipticity condition (A.5) are the same in this case, i.e., there exists $\theta > 0$ such that $\sum_{i,j} a_{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2$ for any $\xi \in \mathbb{R}^d$ and any $x \in S$.

For the case $d = 2$, we need the following lemma, whose proof is provided after the current proof.

Lemma A.1.2. *Suppose that* $p(\boldsymbol{\xi}) = \sum_{|\alpha|=2k} a_\alpha(x)\boldsymbol{\xi}^\alpha$ *is a 2-variate homogeneous polynomial of degree* $2k$ *for every* $x \in \overline{S}$*, and that the coefficients* $a_\alpha \in C^k(\overline{S})$ *for all* $|\alpha| = 2k$*. If there exists* $\theta > 0$ *such that* $p(\boldsymbol{\xi}) \geq \theta|\boldsymbol{\xi}|^{2k}$ *for any* $\boldsymbol{\xi} \in \mathbb{R}^2$ *and* $x \in \overline{S}$*, then there exists* $b_{\sigma\gamma}(x) \in C^k(\overline{S})$ *for any* $|\sigma| = |\gamma| = k$ *such that*

$$p(\boldsymbol{\xi}) = \sum_{|\sigma|=|\gamma|=k} b_{\sigma\gamma}(x)\boldsymbol{\xi}^\sigma\boldsymbol{\xi}^\gamma \quad \forall x \in \overline{S} \tag{A.7}$$

and that

$$\sum_{|\sigma|,|\gamma|=k} b_{\sigma\gamma}(x)\zeta_\sigma\zeta_\gamma \geq \frac{\theta}{2} \sum_{|\sigma|=k} \zeta_\sigma^2, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^Q, x \in \overline{S}. \tag{A.8}$$

Due to the smoothness of the coefficients $a_\alpha \in C^{|\alpha|-k}(\overline{S})$ for $k < |\alpha| \leq 2k$, Lemma A.1.2 implies that there exists $b_{\sigma\gamma}(x) \in C^k(\overline{S})$ for any $|\sigma| = |\gamma| = k$ such that Eqn. (A.7) and (A.8) hold true. Thanks to Eqn. (A.7), we know that

$$\widetilde{\mathcal{L}}u := \mathcal{L}u - (-1)^k \sum_{|\sigma|,|\gamma|=k} D^\sigma(b_{\sigma\gamma}(x)D^\gamma u)$$

is an differential operator with order at most $2k-1$. Since the coefficients of $\widetilde{\mathcal{L}}$ are smooth enough, by Lemma 9.7 in [113], we can write $\widetilde{\mathcal{L}}$ in a divergence form as in Eqn. (A.4). Thanks to Eqn. (A.8), we know that the operator $\mathcal{L}$, which can be rewritten in a divergence form as

$$\mathcal{L}u = (-1)^k \sum_{|\sigma|,|\gamma|=k} D^\sigma(b_{\sigma\gamma}(x)D^\gamma u) + \widetilde{\mathcal{L}}u,$$

is a strongly elliptic operator. □

Proof of Lemma A.1.2. We first write the 2-variate homogeneous polynomial $p(\boldsymbol{\xi})$ of degree $2k$ as

$$p(\boldsymbol{\xi}) = \sum_{i=0}^{2k} a_i(x)\xi_1^{2k-i}\xi_2^i = \boldsymbol{\psi}^T B_0(x)\boldsymbol{\psi},$$

where

$$\boldsymbol{\psi} := \begin{bmatrix} \xi_1^k \\ \xi_1^{k-1}\xi_2 \\ \vdots \\ \xi_1\xi_2^{k-1} \\ \xi_2^k \end{bmatrix}, \qquad B_0(x) := \begin{bmatrix} a_0(x) & a_1(x)/2 & & & \\ a_1(x)/2 & a_2(x) & a_3(x)/2 & & \\ \ddots & \ddots & \ddots & & \\ & a_{2k-3}(x)/2 & a_{2k-2}(x) & a_{2k-1}(x)/2 \\ & & a_{2k-1}(x)/2 & a_{2k}(x) \end{bmatrix}.$$

Since $a_\alpha \in C^k(\overline{S})$ for all $|\alpha| = 2k$, all entries in $B_0(x)$ are in $C^k(\overline{S})$. As proved in Chapter 3 in [14], all the $B(x) \in R^{Q\times Q}$ that satisfies the equality constraint (A.7) form the following feasible set:

$$\mathcal{F} = \left\{ B(x) \ : \ B(x) = B_0(x) + \sum_{i=1}^{n} \lambda_i(x) L_i \right\}, \tag{A.9}$$

where $n = k(k-1)/2$ for the case $d = 2$, $\boldsymbol{\lambda}(x) := [\lambda_1(x), \ldots, \lambda_n(x)]^T$ can be any mapping from $\overline{S}$ to $\mathbb{R}^n$, and $\{L_i\}_{i=1}^n$ are constant square matrices of size Q-by-Q. In the end of this proof, we construct an entry-wise continuous mapping

$\widetilde{B}(x) \in \mathcal{F}$, which is associated with an entry-wise continuous mapping $\widetilde{\boldsymbol{\lambda}}(x)$, such that

$$\widetilde{B}(x) \equiv B_0(x) + \sum_{i=1}^{n} \widetilde{\lambda}_i(x) L_i \succeq \frac{3\theta}{4} I_Q, \qquad \forall x \in \overline{S}, \tag{A.10}$$

where I_Q is the identity matrix of size Q-by-Q. Thanks to the continuity of $\widetilde{\boldsymbol{\lambda}}(x)$, the Stone-Weierstrass theorem implies that there exists $\boldsymbol{\lambda}(x) \in C^\infty(\overline{S}, \mathbb{R}^n)$ such that

$$\frac{\theta}{4} I_Q + B(x) - \widetilde{B}(x) = \frac{\theta}{4} I_Q + \sum_{i=1}^{n} (\lambda_i(x) - \widetilde{\lambda}_i(x)) L_i \succeq 0, \quad \forall x \in \overline{S}. \tag{A.11}$$

Combining Eqn. (A.10) and (A.11), we have

$$B(x) - \frac{\theta}{2} I_Q = \frac{\theta}{4} I_Q + B(x) - \widetilde{B}(x) + \widetilde{B}(x) - \frac{3\theta}{4} I_Q \succeq 0, \quad \forall x \in \overline{S},$$

which is equivalent to Eqn. (A.8). Since $\boldsymbol{\lambda}(x) \in C^\infty(\overline{S}, \mathbb{R}^n)$, every entry in $B(x)$ belongs to $C^k(\overline{S})$. Therefore, we have proved Lemma A.1.2.

Construction of $\widetilde{B}(x)$. Let's consider the following 2-variate polynomial

$$\widetilde{p}(\boldsymbol{\xi}) \equiv \sum_{i=0}^{2k} \widetilde{a}_i(x) \xi_1^{2k-i} \xi_2^i := p(\boldsymbol{\xi}) - \frac{3\theta}{4} |\boldsymbol{\xi}|^{2k}.$$

Since $a_i \in C^k(\overline{S})$, we have $\widetilde{a}_i \in C^k(\overline{S})$. Since $p(\boldsymbol{\xi}) \geq \theta |\boldsymbol{\xi}|^{2k}$, we have $\widetilde{p}(\boldsymbol{\xi}) \geq \theta |\boldsymbol{\xi}|^{2k}/4$. Therefore, we know that $\widetilde{a}_0(x) \geq \theta/4$ for any $x \in \overline{S}$. Define $\widehat{a}_i(x) = \widetilde{a}_i(x)/\widetilde{a}_0(x)$. Consider the factorization of the following monic polynomial

$$\widehat{p}(\xi_1) = \sum_{i=0}^{2k} \widehat{a}_i(x) \xi_1^{2k-i} = \prod_{j=1}^{k} \left((\xi_1 - g_j)^2 + h_j^2 \right), \tag{A.12}$$

where $\{g_j(x) \pm \mathrm{i} h_j(x)\}_{j=1}^k$ are the complex root pairs of the nonnegative polynomial $\widehat{p}(\xi_1)$. We order $\{g_j(x) \pm \mathrm{i} h_j(x)\}_{j=1}^k$ such that smaller real part $g_j(x)$ comes first and smaller imaginary part $h_j(x)$ comes first if the real parts are the same.

On one hand, combining the continuity of polynomial roots in terms of its coefficients (see e.g. [127]) and the fact that $\{\widehat{a}_i(x)\}_{i=0}^{2k} \subset C^k(\overline{S})$, we know that both $\{g_i(x)\}_{i=1}^k$ and $\{h_i(x)\}_{i=1}^k$ are continuous on the physical domain $\overline{S}$. On the other hand, thanks to Eqn. (A.12), we have

$$\begin{aligned} \widetilde{p}(\boldsymbol{\xi}) = \widetilde{a}_0(x) \prod_{j=1}^{k} \left((\xi_1 - g_j \xi_2)^2 + h_j^2 \xi_2^2 \right) &= \widetilde{a}_0(x) \sum_{[\tau_1, \tau_2, \ldots, \tau_k] \in \{0,1\}^k} \left(\widetilde{p}_{\boldsymbol{\tau}}(\boldsymbol{\xi}) \right)^2 \\ := \widetilde{a}_0(x) \sum_{[\tau_1, \tau_2, \ldots, \tau_k] \in \{0,1\}^k} &\left(\prod_{j=1}^{k} (\xi_1 - g_j \xi_2)^{\tau_j} (h_j \xi_2)^{1-\tau_j} \right)^2. \end{aligned} \tag{A.13}$$

Here, for any $[\tau_1, \tau_2, \ldots, \tau_k] \in \{0,1\}^k$, the polynomial $\widetilde{p}_{\boldsymbol{\tau}}(\boldsymbol{\xi})$ is defined as

$$\widetilde{p}_{\boldsymbol{\tau}}(\boldsymbol{\xi}) \equiv \sum_{j=0}^{k} a_j^{\boldsymbol{\tau}} \xi_1^{k-j} \xi_2^j := \prod_{j=1}^{k} (\xi_1 - g_j \xi_2)^{\tau_j} (h_j \xi_2)^{1-\tau_j}. \tag{A.14}$$

The coefficients $a^{\boldsymbol{\tau}} = [a_0^{\boldsymbol{\tau}}, \ldots, a_k^{\boldsymbol{\tau}}]^T$ are smooth functions of $\{g_i(x)\}_{i=1}^k$ and $\{h_i(x)\}_{i=1}^k$, and thus every entry in $a^{\boldsymbol{\tau}}$ is continuous on $\overline{S}$. Then we can construct $\widetilde{B}(x)$ as

$$\widetilde{B}(x) = \widetilde{a}_0^{1/2} \sum_{[\tau_1,\tau_2,\ldots,\tau_k]\in\{0,1\}^k} a^{\boldsymbol{\tau}} \left(a^{\boldsymbol{\tau}}\right)^T + \frac{3\theta}{4} \mathrm{diag}\left\{ \binom{k}{0}, \binom{k}{1}, \ldots, \binom{k}{k} \right\}, \tag{A.15}$$

where the first part is the square matrix presentation of polynomial $\widetilde{p}(\boldsymbol{\xi})$, and the second diagonal part is that of the polynomial $\frac{3\theta}{4}|\boldsymbol{\xi}|^{2k}$. Since the first part is positive semi-definite, we conclude that $\widetilde{B}(x) \succeq \frac{3\theta}{4} I_Q$ for every $x \in \overline{S}$, as desired in Eqn. (A.10). □

A.2 Derivations Involving I_1

From Eqn. (3.78) **to Eqn.** (3.79) **in the proof of Theorem 3.5.2**

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

$$\sum_{|\sigma|\le k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \le C_1^2 C_\eta^2 \sum_{s=1}^{k} \sum_{s'=1}^{s} (lh)^{-2s'} |\psi_{i,q}|^2_{s-s',2,S^*}. \tag{A.16}$$

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

$$\begin{aligned} LHS &= \sum_{s=1}^{k} \sum_{|\sigma|=s} \int_{S^*} \left| \sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma-\sigma_1}\psi_{i,q} \right|^2 \\ &\le \sum_{s=1}^{k} \sum_{|\sigma|=s} \left(\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \binom{\sigma}{\sigma_1}^2 \right) \left(\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \int_{S^*} |D^{\sigma_1}\eta|^2 |D^{\sigma-\sigma_1}\psi_{i,q}|^2 \right) \\ &\le C_{1,1}^2 C_\eta^2 \sum_{s=1}^{k} \sum_{|\sigma|=s} \sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \int_{S^*} (lh)^{-2|\sigma_1|} |D^{\sigma-\sigma_1}\psi_{i,q}|^2, \end{aligned} \tag{A.17}$$

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 Eqn. (A.17) 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

$$\sum_{|\sigma|=s}\sum_{\sigma_1\le\sigma,|\sigma_1|\ge 1}\int_{S^*}(lh)^{-2|\sigma_1|}|D^{\sigma-\sigma_1}\psi_{i,q}|^2 \le C_{1,2}\sum_{s'=1}^{s}(lh)^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}. \tag{A.18}$$

Combining Eqn. (A.17) and (A.18), and denoting $C_1=C_{1,1}C_{1,2}^{1/2}$, we have proved Eqn. (A.16). □

Remark A.2.1. *If there are no lower order terms, we can obtain*

$$\sum_{|\sigma|=k}\int_{S^*}\left|\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}\right|^2 \le C_1^2C_\eta^2\sum_{s'=1}^{k}(lh)^{-2s'}|\psi_{i,q}|^2_{k-s',2,S^*}. \tag{A.19}$$

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 Eqn.* (A.16). *Eqn.* (A.19) *is used from Eqn.* (3.66) *to Eqn.* (3.67) *in the proof of Theorem 3.5.1.*

Estimation of $\|\eta\psi_{i,q}\|_{H(S^*)}$ in the proof of Theorem 3.6.1

In this subsection, we will prove the following result that is used in in the proof of Theorem 3.6.1: for all $h>0$ such that $\frac{1-h^{2k}}{1-h^2}\le 2$, we have

$$\|\eta\psi_{i,q}\|_{H(S^*)} \le \frac{C}{2}|\psi_{i,q}|_{k,2,S^*}+\sqrt{\frac{C^2}{4}|\psi_{i,q}|^2_{k,2,S^*}+C|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}\|_{H(S^*)}+\|\psi_{i,q}\|^2_{H(S^*)}}, \tag{A.20}$$

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:

$$\sum_{0\le|\sigma|,|\gamma|\le k}\int_{S^*} a_{\sigma\gamma}D^\sigma(\eta\psi_{i,q})D^\gamma(\eta\psi_{i,q}) = \underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\int_{S^*}\eta a_{\sigma\gamma}(x)D^\sigma\psi_{i,q}D^\gamma(\eta\psi_{i,q})}_{I_3}$$
$$+\underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*}a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma(\eta\psi_{i,q})\,.}_{I_4} \tag{A.21}$$

Repeating the same argument from Eqn. (3.77) to Eqn. (3.79), we obtain

$$|I_4|\le C_1C_\eta\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}h^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}\right)^{1/2}\|\eta\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,max}}. \tag{A.22}$$

Since $\psi_{i,q}\perp\mathcal{P}_{k-1}$ locally in L^2, from Eqn. (3.20) we have

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

Repeating the same argument from Eqn. (3.80) to Eqn. (3.82), we conclude

$$\begin{aligned} I_4 &\le C_1C_\eta C_p\sqrt{\theta_{k,max}}\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}|\psi_{i,q}|^2_{s,2,S^*}\right)^{1/2}\|\eta\psi_{i,q}\|_{H(S^*)} &\text{(A.23)}\\ &\le C_1C_\eta C_p\sqrt{\theta_{k,max}}\left(\sum_{s=1}^{k}s|\psi_{i,q}|^2_{s,2,S^*}\right)^{1/2}\|\eta\psi_{i,q}\|_{H(S^*)} &\text{(A.24)}\\ &\le C_1C_\eta C_p\sqrt{2k\theta_{k,max}}|\psi_{i,q}|_{k,2,S^*}\|\eta\psi_{i,q}\|_{H(S^*)}. &\text{(A.25)}\end{aligned}$$

In the last inequality (3.82), we have used the polynomial approximation property (3.20) 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

$$I_3 = \underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\int_{S^*}\eta^2 a_{\sigma\gamma}(x)D^\sigma\psi_{i,q}D^\gamma\psi_{i,q}}_{I_5}$$
$$+\underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*}\eta a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma\psi_{i,q}\,.}_{I_6} \tag{A.26}$$

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

$$I_5 \leq \|\psi_{i,q}\|^2_{H(S^*)} \tag{A.27}$$

Repeating the same argument from Eqn. (3.77) to Eqn. (3.79) again, we obtain

$$\begin{aligned} I_6 &= \sum_{0\leq|\sigma|,|\gamma|\leq k} \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} \int_{S^*} a_{\sigma\gamma}(x)\eta D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^{\gamma}\psi_{i,q} \\ &\leq \left(\sum_{|\sigma|\leq k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\geq 1}} \binom{\sigma}{\sigma_1} \eta D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,max}} \\ &\leq C_1 C_\eta \sqrt{\theta_{k,max}} \left(\sum_{s=1}^{k} \sum_{s'=1}^{s} h^{-2s'} |\psi_{i,q}|^2_{s-s',2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \ . \end{aligned} \tag{A.28}$$

The derivation of Eqn. (A.28) is nearly the same as that of Eqn. (A.16) and the only difference is that we need to use $|\eta D^{\sigma_1}\eta| \leq C_\eta h^{-|\sigma_1|}$ (thanks to $|\eta| \leq 1$) in Eqn. (A.17). Using exactly the same argument from Eqn. (A.23) to Eqn. (A.25), we conclude that for all $h > 0$ such that $\frac{1-h^{2k}}{1-h^2} \leq 2$,

$$I_6 \leq C_1 C_\eta C_p \sqrt{2k\theta_{k,max}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{A.29}$$

Combining Eqn. (A.26), (A.27) and (A.29), we obtain

$$|I_3| \leq \|\psi_{i,q}\|^2_{H(S^*)} + C_1 C_\eta C_p \sqrt{2k\theta_{k,max}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{A.30}$$

Combining Eqn. (A.21), (A.25) and (A.30), we have

$$\|\eta\psi_{i,q}\|^2_{H(S^*)} \leq \|\psi_{i,q}\|^2_{H(S^*)} + C_1 C_\eta C_p \sqrt{2k\theta_{k,max}} |\psi_{i,q}|_{k,2,S^*} (\|\psi_{i,q}\|_{H(S^*)} + \|\eta\psi_{i,q}\|_{H(S^*)}). \tag{A.31}$$

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

A p p e n d i x B

SUPPLEMENTARY MATERIALS FOR THE ISMD

B.1 A Simple Lemma About Regular-sparse Partitions

Lemma B.1.1. *Suppose that $A \in \mathbb{R}^{N\times N}$ is symmetric and PSD. Let $\mathcal{P}_c$ be a partition of $[N]$ and $\mathcal{P}_f$ be a refinement of $\mathcal{P}_c$. If the finer partition $\mathcal{P}_f$ is regular-sparse with respect to A, then the coarser partition $\mathcal{P}_c$ is also regular-sparse with respect to A.*

Proof. By the definition of regular-sparseness, suppose that $A = \sum_{k=1}^{K} g_k^{(f)} \left(g_k^{(f)}\right)^T$ and that on every patch in $\mathcal{P}^{(f)}$ the nontrivial modes $\{g_k^{(f)}\}_{k=1}^{K}$ on this patch are linearly independent. For any $P_m^{(c)} \in \mathcal{P}_c$, assume

$$\sum_{i=1}^{d_m} \alpha_i g_{k_i^m}^{(f)} \equiv 0 \qquad \text{on patch } P_m^{(c)}, \tag{B.1}$$

where d_m is the local dimension of decomposition $A = \sum_{k=1}^{K} g_k^{(f)} \left(g_k^{(f)}\right)^T$ on $P_m^{(c)}$ and $\{g_{k_i^m}^{(f)}\}_{i=1}^{d_m}$ are the modes which are non zero there. Since $\mathcal{P}_f$ is a refinement of $\mathcal{P}_c$, for any $i \in [d_m]$, there exists one patch $P_n^{(f)} \subset P_m^{(c)}$ such that $g_{k_i^m}^{(f)} \neq 0$ on this smaller patch. Restricting Eqn. (B.1) to $P_n^{(f)}$, we get $\alpha_i = 0$ due to regular-sparse property of $\mathcal{P}_f$. Therefore, $\{g_{k_i^m}^{(f)}\}_{i=1}^{d_m}$ are linearly independent on $P_m^{(c)}$. Since the patch $P_m^{(c)}$ is arbitrarily chosen, we conclude that $\mathcal{P}_c$ is regular-sparse. □

B.2 Joint Diagonalization of Matrices

Joint diagonalization is often used in Blind Source Separation (BSS) and Independent Component Analysis (ICA), and it has been well studied. We adopt its algorithm and sensitivity analysis in the ISMD. Suppose a series of n-dimensional symmetric matrices $\{M_k\}_{k=1}^{K}$ can be decomposed into:

$$M_k = D\Lambda_k D^T, \tag{B.2}$$

where D is an n-dimensional unitary matrix that jointly diagonalizes $\{M_k\}_{k=1}^{K}$ and the eigenvalues are stored in diagonal matrices $\Lambda_k = \mathrm{diag}\{\lambda_1(k), \lambda_2(k), \cdots, \lambda_n(k)\}$. Denote $\boldsymbol{\lambda}_i \equiv [\lambda_i(1), \lambda_i(2), \ldots, \lambda_i(K)]^T \in \mathbb{R}^K$. To find the joint eigenvectors D,

we solve the following optimization problem:

$$\min_{V\in\mathbb{O}(n)} \quad \sum_{k=1}^{K}\sum_{i\neq j}|(V^T M_k V)_{i,j}|^2. \tag{B.3}$$

Obviously the minimum of problem (B.3) is 0 and D is an minimizer. However, the minimizer is not unique. The so-called unicity assumption, i.e., $\boldsymbol{\lambda}_i \neq \boldsymbol{\lambda}_j$ for any $i \neq j$, is widely used in existing literatures and guarantees that D is unique up to column permutation and sign flips. In general, we assume that there are m ($m \leq n$) distinct eigenvalues $\{\boldsymbol{\lambda}_i\}_{i=1}^m$ with multiplicity $\{q_i\}_{i=1}^m$ respectively. Minimizers of problem (B.3) are characterized by the following theorem.

Theorem B.2.1. *Suppose that* $\{M_k\}_{k=1}^K$ *are generated by* (B.2) *and that* V *is a global minimizer of problem* (B.3). *There exists a permutation matrix* $\Pi \in \mathbb{R}^{n\times n}$ *and block diagonal matrix* R *such that*

$$V\Pi = DR\,, \qquad R = diag\{R_1, \ldots, R_m\}\,, \tag{B.4}$$

in which $R_i \in \mathbb{O}(q_i)$.

Theorem B.2.1 is the generalization of eigendecomposition of a single symmetric matrix to the case with multiple matrices. Although it is elementary, we provide the sketch of its proof here for completeness.

Proof. Since V is a global minimizer and thus achieves zero in its objective function, $V^T M_k V$ is diagonal for any $k \in [K]$. Denote $\Gamma \equiv V^T M_k V = \mathrm{diag}\{\gamma_1(k), \gamma_2(k), \cdots, \gamma_n(k)\}$ and $\boldsymbol{\gamma}_i \equiv [\gamma_i(1), \gamma_i(2), \ldots, \gamma_i(K)]^T \in \mathbb{R}^K$. Define $D = [d_1, d_2, \ldots, d_n]$ and $V = [v_1, v_2, \ldots, v_n]$. If $\boldsymbol{\gamma}_i \neq \boldsymbol{\lambda}_j$, then $v_i^T d_j = 0$ since they belong to different eigen spaces for at least one M_k. Both D and V span the full space $\mathbb{R}^n$, and thus there is a one-to-one mapping between $\{\boldsymbol{\gamma}_i\}_{i=1}^n$ to $\{\boldsymbol{\lambda}_i\}_{i=1}^m$ with multiplicity $\{q_i\}_{i=1}^m$. Therefore, there exists a permutation matrix Π such that

$$[\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n]\,\Pi = [\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_n]\,.$$

Correspondingly, denoting $\widetilde{D} = [\widetilde{d}_1, \widetilde{d}_2, \ldots, \widetilde{d}_n] \equiv V\Pi$, we have

$$M_k d_{i,j} = \lambda_i(k) d_{i,j}\,, \qquad M_k \widetilde{d}_{i,j} = \lambda_i(k)\widetilde{d}_{i,j}\,,$$

where $\{d_{i,j}\}_{j=1}^{q_i}$ and $\{\widetilde{d}_{i,j}\}_{j=1}^{q_i}$ are the eigenvectors in D and $\widetilde{D}$ respectively corresponding to the eigenvalue $\boldsymbol{\lambda}_i$. By orthogonality between eigenspaces and completeness of D and $\widetilde{D}$, $\{d_{i,j}\}_{j=1}^{q_i}$ and $\{\widetilde{d}_{i,j}\}_{j=1}^{q_i}$ must span the same q_i-dimensional subspace. Since both $\{d_{i,j}\}_{j=1}^{q_i}$ and $\{\widetilde{d}_{i,j}\}_{j=1}^{q_i}$ are orthonormal, there exists $R_i \in \mathbb{O}(q_i)$ such that $\widetilde{d}_{i,j} = R_i d_{i,j}$ for $j \in [q_i]$. □

The sensitivity analysis of the joint diagonalization problem (B.3) is studied in [24], and we directly quote its main results below.

Proposition B.2.2. *Suppose that $\{\widehat{M}_k\}_{k=1}^K$ are generated as follows:*

$$\widehat{M}_k = M_k + \epsilon\widetilde{M}_k, \quad M_k = D\Lambda_k D^T,$$

where D is unitary, ϵ is a real scalar, matrices $\widetilde{M}_k$ are arbitrary, and matrices Λ_k are diagonal as in (B.2)*. Suppose that the unicity assumption, i.e., $\boldsymbol{\lambda}_i \neq \boldsymbol{\lambda}_j$ for any $i \neq j$, holds true. Then any solution of the joint diagonalization problem* (B.3) *with the perturbed input $\{\widehat{M}_k\}_{k=1}^K$, denoted by $\widehat{D}$, is in the form*

$$\widehat{D} = D(\mathbb{I} + \epsilon E + o(\epsilon))J,$$

where J is the product of a permutation matrix with a diagonal matrix having only ± 1 on its diagonal. Matrix E has a null diagonal and is antisymmetric, i.e., $E + E^T = 0$. Its off-diagonal entries E_{ij} are give by

$$E_{ij} = \frac{1}{2}\sum_{k=1}^{K} f_{ij}(k) d_i^T(\widetilde{M}_k + \widetilde{M}_k^T)d_j\,, \quad \textit{with} \quad f_{ij}(k) = \frac{\lambda_j(k) - \lambda_i(k)}{\sum_{l=1}^K(\lambda_j(l) - \lambda_i(l))^2}\,.$$

In this paper, we solve problem (B.3) using a Jacobi-like algorithm proposed in [23, 17]. The idea is to perform 2-dimensional rotation to reduce the amplitude of the off-diagonal pairs one by one. Denote by $R = R(p, q, c, s)$ the 2-dimensional rotation that deals with (p, q) entries of M_k:

$$R = R(p,q,c,s) = I + (c-1)\boldsymbol{e}_p\boldsymbol{e}_p^T - s\boldsymbol{e}_q\boldsymbol{e}_q^T + s\boldsymbol{e}_q\boldsymbol{e}_p^T + (c-1)\boldsymbol{e}_p\boldsymbol{e}_q^T, \quad \text{(B.5)}$$

where $c^2 + s^2 = 1$ for unitarity. A simple calculation shows that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i\neq j}|(R^T M_k R)_{i,j}|^2 =& \sum_{k=1}^{K}\sum_{i\neq j}|M_k(i,j)|^2 - \sum_{k=1}^{K}\left(|M_k(p,q)|^2 + |M_k(q,p)|^2\right)\\ &+ \sum_{k=1}^{K}\left(sc(M_k(q,q) - M_k(p,p)) + c^2 M_k(p,q) - s^2 M_k(q,p)\right)^2\\ &+ \sum_{k=1}^{K}\left(sc(M_k(q,q) - M_k(p,p)) - s^2 M_k(p,q) + c^2 M_k(q,p)\right)^2\,.\end{aligned} \quad \text{(B.6)}$$

It can be shown that the choice of c and s that minimizes (B.6) also minimizes $\|L_{pq}z\|_2$ where $z = [c^2 - s^2, 2cs]^T$ is a 2×1 vector, and

$$L_{pq} := \begin{bmatrix} M_1(p,q) & \frac{M_1(q,q)-M_1(p,p)}{2} \\ \vdots & \vdots \\ M_K(p,q) & \frac{M_K(q,q)-M_K(p,p)}{2} \end{bmatrix}, \tag{B.7}$$

is a $K \times 2$ matrix. It is apparent that the singular vector corresponding to the smallest singular value does the job. Denote this singular vector by $\boldsymbol{w}$ with $\boldsymbol{w}(1) \geq 0$. The optimizer of Eqn. (B.6) is given by:

$$c = \sqrt{\frac{1 + \boldsymbol{w}(1)}{2}}, \quad s = \frac{\boldsymbol{w}(2)}{2c}. \tag{B.8}$$

We perform such rotation for each pair of (p, q) until the algorithm converges, as shown in Algorithm 5. The algorithm has been shown to have quadratic

Algorithm 5 Jacobi-like Joint Diagonalization

Require: $\epsilon > 0$; $\{M_k\}_{k=1}^K$, which are symmetric and jointly diagonalizable.
Ensure: $V \in \mathbb{O}(n)$ such that $\sum_{k=1}^K \sum_{i \neq j} |(V^T M_k V)_{i,j}|^2 \leq \epsilon \sum_{k=1}^K \|M_k\|_F^2$.

$V \leftarrow I$
while $\sum_{k=1}^K \sum_{i \neq j} |(V^T M_k V)_{i,j}|^2 > \epsilon \sum_{k=1}^K \|M_k\|_F^2$ **do**
  for $p = 1, 2, \cdots, n$ **do**
    for $q = p + 1, p + 2, \cdots, n$ **do**
      define L_{pq} as in (B.7)
      compute $\boldsymbol{w}$, the normalized singular vector corresponding to the smallest singular value
      set $c = \sqrt{\frac{1+\boldsymbol{w}(1)}{2}}$, $s = \frac{\boldsymbol{w}(2)}{2c}$ and $R = R(p, q, c, s)$
      set $V \leftarrow VR$; $M_k \leftarrow V^T M_k V$ for $k = 1, 2, \cdots, K$
    end for
  end for
end while

asymptotic convergence rate and is numerically stable; see [17].

B.3 Proof of Lemma 5.4.1

We point out that for the noiseless case, the ISMD in fact solves the following optimization problem to obtain G_m:

$$\boxed{\begin{aligned} \min_{G_m \in \mathbb{R}^{|P_m| \times K_m}} \quad & \sum_{n=1}^{M} \sum_{i \neq j} |B_{n;m}(i,j)|^2 \\ \text{s.t.} \quad & G_m G_m^T = A_{mm}\,, \\ & B_{n;m} = G_m^\dagger A_{mn} A_{nn}^\dagger A_{mn}^T \left(G_m^\dagger\right)^T, \end{aligned}} \tag{B.9}$$

in which

$$G_m^\dagger = (G_m^T G_m)^{-1} G_m^T, \quad A_{nn}^\dagger = \sum_{i=1}^{K_n} \gamma_{n,i}^{-1} h_{n,i} h_{n,i}^T \tag{B.10}$$

is the (Moore–Penrose) pseudo-inverse of G_m and A_{nn} respectively. The ISMD solves this optimization problem in two steps:

1. Perform eigendecomposition $A_{mm} = H_m H_m^T$. Then the feasible G_m can be written as $H_m D_m$ with unitary matrix D_m.
2. Find the rotation D_m which solves the joint diagonalization problem (5.13).

Similarly, one can check that for the noisy case, the ISMD (with truncated eigendecomposition (5.10)) solves the same optimization problem with perturbed input to obtain $\widehat{G}_m$:

$$\boxed{\begin{aligned} \min_{G_m \in \mathbb{R}^{|P_m| \times K_m}} \quad & \sum_{n=1}^{M} \sum_{i \neq j} |B_{n;m}(i,j)|^2 \\ \text{s.t.} \quad & G_m G_m^T = \widehat{A}_{mm}^{(t)}\,, \\ & B_{n;m} = G_m^\dagger \widehat{A}_{mn} \left(\widehat{A}_{nn}^{(t)}\right)^\dagger \widehat{A}_{mn}^T \left(G_m^\dagger\right)^T, \end{aligned}} \tag{B.11}$$

where $\widehat{A}_{nn}^{(t)}$ is the truncated $\widehat{A}_{nn}$ defined in Eqn. (5.48) and

$$\left(\widehat{A}_{nn}^{(t)}\right)^\dagger = \sum_{i=1}^{K_n} \widehat{\gamma}_{n,i}^{-1} \widehat{h}_{n,i} \widehat{h}_{n,i}^T \tag{B.12}$$

is the pseudo-inverse of $\widehat{A}_{nn}^{(t)}$.

Since G_m is a minimizer of problem (B.9), the identity matrix $\mathbb{I}_{K_m}$ is one minimizer of the following joint diagonalization problem:

$$\boxed{\min_{V\in\mathbb{O}(K_m)} \quad \sum_{n=1}^{M}\sum_{i\neq j}|(V^T B_{n;m}V)_{i,j}|^2,} \tag{B.13}$$

where

$$B_{n;m} = G_m^{\dagger} A_{mn} A_{nn}^{\dagger} A_{mn}^T \left(G_m^{\dagger}\right)^T = D_m^T \Sigma_{n;m} D_m, \tag{B.14}$$

where D_m and $\Sigma_{n;m}$ are defined in the procedure of the ISMD. Let $\{\psi_k\}_{k=1}^K$ be a set of intrinsic sparse modes of A. Combining Lemma 5.3.1 with Lemma 5.3.2, we get

$$B_{n;m} = D_m^T \Sigma_{n;m} D_m = \Pi_m V_m^T \left(D^{(\psi)}\right)^T \Sigma_{n;m} D^{(\psi)} V_m \Pi_m = \Pi_m V_m^T B_{n;m}^{(\psi)} V_m \Pi_m = \Pi_m B_{n;m}^{(\psi)} \Pi_m. \tag{B.15}$$

The last equality is due to the fact that V_m are diagonal matrices with diagonal entries either 1 or -1 in the identifiable case.[1] If Ψ_m is reordered by Π_m, we simply have $B_{n;m} = B_{n;m}^{(\psi)}$ for all $n \in [M]$. Therefore, there exists such a set of intrinsic sparse modes $\{\psi_k\}_{k=1}^K$ that for all $n \in [M]$

$$B_{n;m} = B_{n;m}^{(\psi)}. \tag{B.16}$$

One can easily verify that the unicity assumption holds true for the joint diagonalization problem (B.13) because the intrinsic sparse modes $\{\psi_k\}_{k=1}^K$ are pair-wisely identifiable.

Combining the equality constraints in problem (B.9) and problem (B.11) and the assumption (5.49), we have

$$\widehat{G}_m \widehat{G}_m^T = \left((I + \epsilon E_m^{(eig)})G_m\right)\left((I + \epsilon E_m^{(eig)})G_m\right)^T.$$

Define

$$F_m \equiv (I + \epsilon E_m^{(eig)})G_m. \tag{B.17}$$

Then, there exists $U_m \in \mathbb{O}(K_m)$ such that $\widehat{G}_m = F_m U_m$. Since $\widehat{G}_m$ is a minimizer of problem (B.11), U_m is one minimizer of the following joint diagonalization problem:

$$\boxed{\min_{V\in\mathbb{O}(K_m)} \quad \sum_{n=1}^{M}\sum_{i\neq j}|(V^T \widehat{B}_{n;m}V)_{i,j}|^2,} \tag{B.18}$$

[1] Readers can verify that Eqn. (B.15) is still true in the non-identifiable case.

where

$$\widehat{B}_{n;m} = F_m^{\dagger} \widehat{A}_{mn} \left(\widehat{A}_{nn}^{(t)}\right)^{\dagger} \widehat{A}_{mn}^T \left(F_m^{\dagger}\right)^T. \tag{B.19}$$

From standard perturbation analysis of pseudo-inverse (for instance see Theorem 3.4 in [122]), we have

$$F_m^{\dagger} = G_m^{\dagger} + \epsilon E_m^{(ginv)}, \quad \|E_m^{(ginv)}\|_2 \leq \mu \sigma_{min}^{-2}(G_m) \|E_m^{(eig)} G_m\|_2 \leq \mu C_{eig} \sigma_{min}^{-2}(G_m) \|G_m\|_2 \tag{B.20}$$

and

$$\left(\widehat{A}_{nn}^{(t)}\right)^{\dagger} = A_{nn}^{\dagger} + \epsilon E_n^{(ainv)}, \quad \|E_n^{(ainv)}\|_2 \leq \mu \gamma_{n,K_n}^{-2} \|\widehat{A}_{nn}^{(t)} - A_{nn}\|_2 / \epsilon.$$

Here, $\sigma_{min}(G_m)$ is the smallest nonzero singular value of G_m and γ_{n,K_n} is the K_n-th eigenvalue of A_{nn} as defined in (5.9). Denote the (K_n+1)-th eigenvalue of $\widehat{A}_{nn}$ as $\widehat{\gamma}_{n,K_n+1}$. From Corollary 8.1.6 in [55], we have $\widehat{\gamma}_{n,K_n+1} \leq \epsilon \|\widetilde{A}_{nn}\|_2$. Then, we get

$$\|\widehat{A}_{nn}^{(t)} - A_{nn}\|_2 \leq \|\widehat{A}_{nn}^{(t)} - \widehat{A}_{nn}\|_2 + \|\widehat{A}_{nn} - A_{nn}\|_2 \leq 2\epsilon \|\widetilde{A}_{nn}\|_2 \leq 2\epsilon,$$

where $\|\widetilde{A}\|_2 \leq 1$ has been used in the last inequality. Therefore, we obtain

$$\left(\widehat{A}_{nn}^{(t)}\right)^{\dagger} = A_{nn}^{\dagger} + \epsilon E_n^{(ainv)}, \quad \|E_n^{(ainv)}\|_2 \leq 2\mu \gamma_{n,K_n}^{-2}. \tag{B.21}$$

When $\epsilon \ll 1$, the constant μ can be taken as 2 in both (B.20) and (B.21). Combining (5.47), (B.20), and (B.21), we get

$$\begin{aligned}
\widehat{B}_{n;m} =& B_{n;m} + \epsilon \widetilde{B}_{n;m}\,, \\
\widetilde{B}_{n;m} =& E_m^{(ginv)} A_{mn} A_{nn}^{\dagger} A_{mn}^T \left(G_m^{\dagger}\right)^T + G_m^{\dagger} \widetilde{A}_{mn} A_{nn}^{\dagger} A_{mn}^T \left(G_m^{\dagger}\right)^T + G_m^{\dagger} A_{mn} E_n^{(ainv)} A_{mn}^T \left(G_m^{\dagger}\right)^T \\
&+ G_m^{\dagger} A_{mn} A_{nn}^{\dagger} \widetilde{A}_{mn}^T \left(G_m^{\dagger}\right)^T + G_m^{\dagger} A_{mn} A_{nn}^{\dagger} A_{mn}^T \left(E_m^{(ginv)}\right)^T.
\end{aligned} \tag{B.22}$$

By Proposition B.2.2, there exists $E_m^{(jd)} \in \mathbb{R}^{K_m \times K_m}$ such that

$$U_m = (\mathbb{I}_{K_m} + \epsilon E_m^{(jd)} + o(\epsilon)) J_m,$$

where J_m is the product of a permutation matrix with a diagonal matrix having only ± 1 on its diagonal. Matrix $E_m^{(jd)}$ has a null diagonal and is antisymmetric, i.e., $E_m^{(jd)} + \left(E_m^{(jd)}\right)^T = 0$. Its off-diagonal entries $E_m^{(jd)}(i,j)$ are given by

$$E_m^{(jd)}(i,j) = \sum_{n=1}^{M} f(n) \circ \widetilde{B}_{n;m}\,, \quad \text{with} \quad f_{ij}(n) = \frac{B_{n;m}(j,j) - B_{n;m}(i,i)}{\sum_{n=1}^{M} (B_{n;m}(j,j) - B_{n;m}(i,i))^2}\,.$$

Here, $f(n)$ is the matrix with entries $f_{ij}(n)$ and $f(n) \circ \widetilde{B}_{n;m}$ is the matrix point-wise product (also known as the Hadamard product). Notice that we take advantage of the fact that $\widetilde{B}_{n;m}$ is symmetric to simplify $E_m^{(jd)}(i,j)$. Since $B_{n;m}(j,j) - B_{n;m}(i,i)$ is either ± 1 or 0, $|f_{ij}(n)| \leq 1$ for any i,j and n, and thus we have $\|f(n)\|_F \leq K_m$. Therefore, we conclude

$$\|E_m^{(jd)}\|_F \leq \sum_{n=1}^{M} \|f(n) \circ \widetilde{B}_{n;m}\|_F \leq \sum_{n=1}^{M} \|f(n)\|_F \|\widetilde{B}_{n;m}\|_F \leq K_m^{3/2} \sum_{n=1}^{M} \|\widetilde{B}_{n;m}\|_2, \tag{B.23}$$

where we have used triangle inequality, $\|f(n) \circ \widetilde{B}_{n;m}\|_F \leq \|f(n)\|_F \|\widetilde{B}_{n;m}\|_F$ and $\|\widetilde{B}_{n;m}\|_F \leq K_m^{1/2} \|\widetilde{B}_{n;m}\|_2$ in deriving the above inequalities. Combining (B.22), (5.47), (B.20) and (B.21), we know that $\|\widetilde{B}_{n;m}\|_2$ are bounded by a constant, denoted by C_{jd}, which only depends on A and C_{eig}. From the assumption (5.49), C_{eig} is a constant depending on A but not on ϵ or $\widetilde{A}$. Therefore, C_{jd} depends only on A but not on ϵ or $\widetilde{A}$.