Journal of Machine Learning Research 20 (2019) 1-63 Submitted 5/17; Revised 1/19; Published 4/19

Adaptive Geometric Multiscale Approximations for Intrinsically Low-dimensional Data

Wenjing Liao WLIAO60@GATECH.EDU
School of Mathematics
Georgia Institute of Technology, Atlanta, GA, 30313, USA

Mauro Maggioni MAUROMAGGIONIJHU@ICLOUD.COM
Department of Mathematics, Department of Applied Mathematics and Statistics,
Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

Editor: Sujay Sanghavi

Abstract

We consider the problem of efficiently approximating and encoding high-dimensional data sampled from a probability distribution ρ in $\mathbb{R}^D$, that is nearly supported on a d-dimensional set $\mathcal{M}$ - for example supported on a d-dimensional manifold. Geometric Multi-Resolution Analysis (GMRA) provides a robust and computationally efficient procedure to construct low-dimensional geometric approximations of $\mathcal{M}$ at varying resolutions. We introduce GMRA approximations that adapt to the unknown regularity of $\mathcal{M}$, by introducing a thresholding algorithm on the geometric wavelet coefficients. We show that these data-driven, empirical geometric approximations perform well, when the threshold is chosen as a suitable universal function of the number of samples n, on a large class of measures ρ, that are allowed to exhibit different regularity at different scales and locations, thereby efficiently encoding data from more complex measures than those supported on manifolds. These GMRA approximations are associated to a dictionary, together with a fast transform mapping data to d-dimensional coefficients, and an inverse of such a map, all of which are data-driven. The algorithms for both the dictionary construction and the transforms have complexity $CDn \log n$ with the constant C exponential in d. Our work therefore establishes Adaptive GMRA as a fast dictionary learning algorithm, with approximation guarantees, for intrinsically low-dimensional data. We include several numerical experiments on both synthetic and real data, confirming our theoretical results and demonstrating the effectiveness of Adaptive GMRA.

Keywords: Dictionary Learning, Multi-Resolution Analysis, Adaptive Approximation, Manifold Learning, Compression

1. Introduction

We model a data set as n i.i.d. samples $\mathcal{X}_n := \{x_i\}_{i=1}^n$ from a probability measure ρ in $\mathbb{R}^D$. We make the assumption that ρ is supported on or near a set $\mathcal{M}$ of dimension $d \ll D$, and consider the problem, given $\mathcal{X}_n$, of learning a data-dependent dictionary that enables

efficient encoding of (future) data sampled from ρ, together with fast forward and inverse transforms between $\mathbb{R}^D$ and the space of encodings.

In order to circumvent the curse of dimensionality, a popular model for data is sparsity: we say that the data is k-sparse on a suitable dictionary (i.e. a collection of vectors) $\Phi = \{\varphi_i\}_{i=1}^m \subset \mathbb{R}^D$ if each data point $x \in \mathbb{R}^d$ may be expressed as a linear combination of at most k elements of Φ. Clearly the case of interest is $k \ll D$. These *sparse representations* have been used in a variety of statistical signal processing tasks, compressed sensing, machine learning (see e.g. Protter and Elad, 2007; Peyré, 2009; Lewicki et al., 1998; Kreutz-Delgado et al., 2003; Maurer and Pontil, 2010; Chen et al., 1998; Donoho, 2006; Aharon et al., 2005; Candes and Tao, 2007, among many others), and spurred much research about how to learn data-driven dictionaries (see Gribonval et al., 2015; Vainsencher et al., 2011; Maurer and Pontil, 2010, and references therein). The algorithms used in dictionary learning are often computationally demanding, and based on high-dimensional non-convex optimization (Mairal et al., 2010). These approaches have the strength of being very general, with minimal assumptions on the geometry of the dictionary or on the distribution from which the samples are generated. This "worst-case" approach incurs bounds depending upon the ambient dimension D in general (even in the standard case of data lying on one hyperplane).

It is possible to tackle the dictionary learning problem under geometric assumptions on data sets (Maggioni et al., 2016), namely that data lie on or near a low-dimensional set $\mathcal{M}$. There are of course various possible geometric assumptions, the simplest one being that $\mathcal{M}$ is a single d-dimensional subspace, in which case Principal Component Analysis (PCA) (see Pearson, 1901; Hotelling, 1933, 1936) suffices for estimating the subspace. More generally, one may assume that data lie on a union of several low-dimensional planes instead of a single one. The problem of estimating multiple planes, called subspace clustering, is more challenging (see Fischler and Bolles, 1981; Ho et al., 2003; Vidal et al., 2005; Yan and Pollefeys, 2006; Ma et al., 2007, 2008; Chen and Lerman, 2009; Elhamifar and Vidal, 2009; Zhang et al., 2010; Liu et al., 2010; Chen and Maggioni, 2011). This model was shown effective in various applications, including image processing (Fischler and Bolles, 1981), computer vision (Ho et al., 2003) and motion segmentation (Yan and Pollefeys, 2006). Yet another type of geometric model gives rise to manifold learning, where $\mathcal{M}$ is assumed to be a d-dimensional manifold isometrically embedded in $\mathbb{R}^D$, see (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Donoho and Grimes, 2003; Coifman et al., 2005a,b; Zhang and Zha, 2004) and many others. It is of interest to move beyond this model to even more general geometric models, for example where the regularity of the manifold is reduced, and data are not forced to lie exactly on a manifold, but only close to it.

Geometric Multi-Resolution Analysis (GMRA) was proposed in Chen and Maggioni (2010), refined in Allard et al. (2012). In GMRA, geometric approximations of $\mathcal{M}$ are constructed with multiscale techniques that have their roots in geometric measure theory, harmonic analysis and approximation theory. GMRA performs a multiscale tree decomposition of data and builds multiscale low-dimensional geometric approximations to $\mathcal{M}$. Given data, the cover tree algorithm (Beygelzimer et al., 2006) is run to obtain a multiscale tree in which every node is a subset of $\mathcal{M}$, called a dyadic cell, and all dyadic cells at a fixed scale form a partition of $\mathcal{M}$. After the tree is constructed, PCA is performed on the data in each cell to locally approximate $\mathcal{M}$ by the d-dimensional principal subspace, so that every point in that cell may be encoded by the d coefficients for the corresponding principal

directions. At a fixed scale $\mathcal{M}$ is thereby approximated by a piecewise linear set. In Allard et al. (2012) the performance of GMRA for volume measures on a $\mathcal{C}^s, s \in (1, 2]$ manifold was analyzed in the continuous case (i.e. with no sampling), albeit the effectiveness of GMRA was demonstrated empirically on simulated and real-world data, but for a fixed data set, and without out-of-sample extension. In Maggioni et al. (2016), the approximation error of $\mathcal{M}$ was estimated in the non-asymptotic regime with n i.i.d. samples from a measure ρ, satisfying certain technical assumptions, supported on a thin tube of a $\mathcal{C}^2$ manifold of dimension d isometrically embedded in $\mathbb{R}^D$. The concentration bounds in Maggioni et al. (2016) depend on n and d, but not on D, successfully avoiding the curse of dimensionality caused by the ambient dimension. The assumption that ρ is supported in a tube around a manifold can account for noise and does not force the data to lie exactly on a smooth low-dimensional manifold.

In both Allard et al. (2012) and Maggioni et al. (2016), GMRA approximations are constructed on uniform partitions, at a fixed scale, in which all the cells have similar diameters. However, when the regularity of $\mathcal{M}$, such as smoothness or curvature, weighted by the ρ measure, varies at different scales and locations, uniform partitions do not yield optimal approximations. Inspired by the adaptive methods in classical multi-resolution analysis of functions (see Donoho and Johnstone, 1994, 1995; Cohen et al., 2002; Binev et al., 2005, 2007, among many others, and references therein), we propose an adaptive version of GMRA to construct low-dimensional geometric approximations of $\mathcal{M}$ on an adaptive partition, and provide finite sample performance guarantees for a larger classes of geometric structures $\mathcal{M}$ than those considered in Maggioni et al. (2016). This truly takes advantage of the multiscale structure of GMRA, and leads to simple yet provably powerful approximations for a large class of geometric objects that are not necessarily equally regular at all scales and locations.

Our main result (Theorem 8) in this paper may be paraphrased as follows: Let ρ be a probability measure supported on or near a compact d-dimensional manifold $\mathcal{M} \hookrightarrow \mathbb{R}^D$, with $d \geq 3$. Assume that ρ admits a(n unknown) multiscale decomposition satisfying the technical assumptions A1-A5 in section 2.1. Given n i.i.d. samples of ρ, the intrinsic dimension d, and a parameter κ large enough, Adaptive GMRA outputs a dictionary $\widehat{\Phi}_n = \{\widehat{\phi}_i\}_{i \in \mathcal{J}_n}$, an encoding operator $\widehat{\mathcal{D}}_n : \mathbb{R}^D \to \mathbb{R}^{(d+1)\mathcal{J}_n}$ and a decoding operator $\widehat{\mathcal{D}}_n^{-1} : \mathbb{R}^{(d+1)\mathcal{J}_n} \to \mathbb{R}^D$ that, with high probability, satisfy the following properties. For every $x \in \mathbb{R}^D$, $\|\widehat{\mathcal{D}}_n x\|_0 \leq d + 1$ (i.e. only $d + 1$ entries of the encoded data are non-zero), and the Mean Squared Error (MSE), over data sampled from ρ, satisfies

$$\text{MSE} := \mathbb{E}_{x \sim \rho}[\|x - \widehat{\mathcal{D}}_n^{-1} \widehat{\mathcal{D}}_n x\|^2] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}}.$$

Here s is a regularity parameter of ρ (as in definition 5), which allows us to consider $\mathcal{M}$'s and ρ's with nonuniform regularity, varying at different locations and scales. The parameter κ is used in choosing the threshold on the geometric wavelet coefficients, and selecting from the GMRA a multiscale partition and set of local approximate tangent planes to use for encoding the data. Note that the algorithm does not need to know s (indeed, κ is independent of s), but it automatically adapts to obtain a rate that depends on s. We believe, but do not prove, that this rate is indeed optimal. As for computational complexity, constructing $\widehat{\Phi}_n$ takes $\mathcal{O}((C^d + d^2)Dn \log n)$ and computing $\widehat{\mathcal{D}}_n x$ only takes $\mathcal{O}(d(D + d^2) \log n)$, which means we have a fast transform mapping data to their sparse encoding on the dictionary.

In Adaptive GMRA, the dictionary is composed of the low-dimensional planes on adaptive partitions and the encoding operator transforms a point to the local affine $d+1$ principal coefficients of the data in a piece of the partition (the first affine principal component here means the local mean). We state this results in terms of encoding and decoding to stress that learning the geometry in fact yields efficient representations of data, which may be used for performing signal processing tasks in a domain where the data admit a sparse representation, e.g. in compressive sensing or estimation problems (see Iwen and Maggioni, 2013; Chen et al., 2012; Eftekhari and Wakin, 2015). Adaptive GMRA is designed towards robustness, both in the sense of tolerance to noise and to model error (i.e. data not lying on a manifold). We assume d is given throughout this paper. If not, we refer to Little et al. (2017, 2009a,b) for the estimation of intrinsic dimensionality.

The paper is organized as follows. Our main results, including the construction of GMRA, Adaptive GMRA and their finite sample analysis, are presented in Section 2. We show numerical experiments in Section 3. The detailed analysis of GMRA and Adaptive GMRA is presented in Section 4. In Section 5, we discuss the computational complexity of Adaptive GMRA and extend our work to adaptive orthogonal GMRA. Proofs are collected in the appendix.

Notation. We will introduce some basic notation here. $f \lesssim g$ means that there exists a constant C independent on any variable upon which f and g depend, such that $f \leq Cg$; similarly for $\gtrsim$. $f \asymp g$ means that $f \lesssim g$ and $f \gtrsim g$. The cardinality of a set A is denoted by $\#A$. For $x \in \mathbb{R}^D$, $\|x\|$ denotes the Euclidean norm and $B_r(x)$ denotes the Euclidean ball of radius r centered at x. Given a subspace $V \in \mathbb{R}^D$, we denote its dimension by $\dim(V)$ and the orthogonal projection onto V by Proj_V. If A is a linear operator on $\mathbb{R}^D$, $||A||$ is its operator norm. The identity operator is denoted by $\mathbb{I}$.

2. Main results

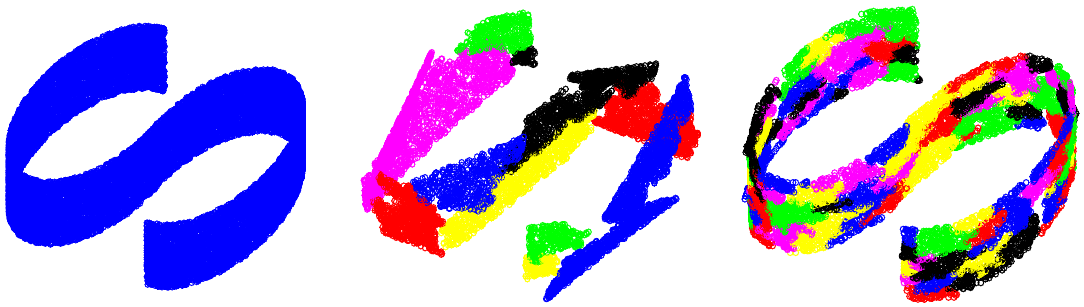
GMRA was proposed in Allard et al. (2012) to efficiently represent points on or near a low-dimensional manifold in high dimensions. We refer the reader to that paper for details of the construction, and we summarize here the main ideas in order to keep the presentation self-contained. The construction of GMRA involves the following steps:

(i) construct a *multiscale tree* $\mathcal{T}$ and the associated decomposition of $\mathcal{M}$ into nested cells $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\in\mathbb{Z}}$ where j represents scale and k location;

(ii) perform *local PCA* on each $C_{j,k}$: let the mean ("center") be $c_{j,k}$ and the d-dim principal subspace $V_{j,k}$. Define $\mathcal{P}_{j,k}(x) := c_{j,k} + \mathrm{Proj}_{V_{j,k}}(x - c_{j,k})$.

(iii) construct a *"difference" subspace* $W_{j+1,k'}$ capturing $\mathcal{P}_{j,k}(C_{j,k}) - \mathcal{P}_{j+1,k'}(C_{j+1,k'})$, for each $C_{j+1,k'} \subseteq C_{j,k}$ (these quantities are associated with the refinement criterion in Adaptive GMRA).

$\mathcal{M}$ may be approximated, at each scale j, by its projection $\mathcal{P}_{\Lambda_j}$ onto the family of linear sets $\Lambda_j := \{\mathcal{P}_{j,k}(C_{j,k})\}_{k\in\mathcal{K}_j}$. For example, linear approximations of the S-manifold at scale 6 and 10 are displayed in Figure 1. In a variety of distances, $\mathcal{P}_{\Lambda_j}(\mathcal{M}) \to \mathcal{M}$. In practice $\mathcal{M}$ is



(a) S-manifold (b) scale 6 (c) scale 10

Figure 1: (a) S-manifold; (b,c) Linear approximations at scale $6, 10$.

unknown, and the construction above is carried over on training data, and its result is random with respect to the training samples. Naturally we are interested in the performance of the construction on new samples. This is analyzed in a setting of "smooth manifold+noise" in Maggioni et al. (2016). When the regularity (such as smoothness or curvature) of $\mathcal{M}$ varies at different locations and scales, linear approximations on fixed uniform partitions are not optimal. Inspired by adaptive methods in classical multi-resolution analysis (see Cohen et al., 2002; Binev et al., 2005, 2007), we propose an adaptive version of GMRA which learns adaptive and near-optimal approximations.

We will start with the multiscale tree decomposition in Section 2.1 and present GMRA and Adaptive GMRA in Section 2.3 and 2.4 respectively.

2.1. Multiscale partitions and trees

A multiscale set of partitions of $\mathcal{M}$ with respect to the probability measure ρ is a family of sets $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\in\mathbb{Z}}$, called dyadic cells, satisfying Assumptions (A1-A5) below, for all integers $j \geq j_{\min}$:

(A1) for any $k \in \mathcal{K}_j$ and $k' \in \mathcal{K}_{j+1}$, either $C_{j+1,k'} \subseteq C_{j,k}$ or $\rho(C_{j+1,k'} \cap C_{j,k}) = 0$. We denote the children of $C_{j,k}$ by $\mathscr{C}(C_{j,k}) = \{C_{j+1,k'} : C_{j+1,k'} \subseteq C_{j,k}\}$. We assume that $a_{\min} \leq \#\mathscr{C}(C_{j,k}) \leq a_{\max}$. Also for every $C_{j,k}$, there exists a unique $k' \in \mathcal{K}_{j-1}$ such that $C_{j,k} \subseteq C_{j-1,k'}$. We call $C_{j-1,k'}$ the parent of $C_{j,k}$.

(A2) $\rho(\mathcal{M} \setminus \cup_{k\in\mathcal{K}_j} C_{j,k}) = 0$, i.e. $\Lambda_j := \{C_{j,k}\}_{k\in\mathcal{K}_j}$ is a cover for $\mathcal{M}$.

(A3) $\exists \theta_1 > 0 : \#\Lambda_j \leq 2^{jd}/\theta_1$.

(A4) $\exists \theta_2 > 0$ such that, if x is drawn from $\rho_{|C_{j,k}}$, then a.s. $\|x - c_{j,k}\| \leq \theta_2 2^{-j}$.

(A5) Let $\lambda_1^{j,k} \geq \lambda_2^{j,k} \geq \ldots \geq \lambda_D^{j,k}$ be the eigenvalues of the covariance matrix $\Sigma_{j,k}$ of $\rho_{|C_{j,k}}$, defined in Table 1. Then:

(i) $\exists \theta_3 > 0$ such that $\forall j \geq j_{\min}$ and $k \in \mathcal{K}_j$, $\lambda_d^{j,k} \geq \theta_3 2^{-2j}/d$,

(ii) $\exists \theta_4 \in (0,1)$ such that $\lambda_{d+1}^{j,k} \leq \theta_4 \lambda_d^{j,k}$.

(A1) implies that the $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$ are associated with a tree structure, and with some abuse of notation we call the above *tree decompositions*. (A1)-(A5) are natural assumptions, easily satisfied by natural multiscale decompositions when $\mathcal{M}$ is a d-dimensional manifold isometrically embedded in $\mathbb{R}^D$: see the work (Maggioni et al., 2016) for a detailed discussion, where the connections between the constants θ_i's and geometric properties of $\mathcal{M}$ (curvatures, reach, etc...) are also discussed. (A2) guarantees that the cells at scale j form a partition of $\mathcal{M}$; (A3) says that there are at most $2^{jd}/\theta_1$ dyadic cells at scale j. (A4) ensures $\mathrm{diam}(C_{j,k}) \lesssim 2^{-j}$. When $\mathcal{M}$ is a d-dimensional manifold, (A5)(i) is the condition that the best rank d approximation to $\Sigma_{j,k}$ is close to the covariance matrix of a d-dimensional Euclidean ball, while (A5)(ii) imposes that the $(d+1)$-th eigenvalue is smaller that the d-th eigenvalue, i.e. the set has significantly larger variances in d directions than in all the remaining ones. The conditions generalize those in (Allard et al., 2012) (which corresponded to the case when $\mathcal{M}$ is a manifold) and in (Maggioni et al., 2016), for example by not assuming that all sets $\{C_{j,k}\}_k$ (for any fixed j) have roughly the same volume, and also by weakening (A5). These changes enlarge the class of measures ρ and sets $\mathcal{M}$ that we consider here, for exampling allowing for a highly nonuniform measure ρ, and an $\mathcal{M}$ substantially "thickened" in many dimensions.

We will construct such $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$ in Section 2.6. In our construction (A1-A4) is satisfied when ρ is a regular doubling probability measure[1] (see Christ, 1990; Deng and Han, 2008). If we further assume that $\mathcal{M}$ is a d-dimensional $\mathcal{C}^s, s \in (1,2]$ closed manifold isometrically embedded in $\mathbb{R}^D$, then (A5) is satisfied as well (See Proposition 14).

It may happen that at the coarsest scales conditions (A3)-(A5) are satisfied but with very poor constants θ: it will be clear that in all that follows we may discard a few coarse scales (i.e. by enlarging $j_{\min}$), and only work at scales that are fine enough and for which (A3)-(A5) truly capture the local geometry of $\mathcal{M}$.

Some notation: a master tree $\mathcal{T}$ is associated with $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$ (using property (A1)), constructed on $\mathcal{M}$; since $C_{j,k}$'s at scale j have similar diameters, $\Lambda_j := \{C_{j,k}\}_{k\in\mathcal{K}_j}$ is called a *uniform partition at scale* j. A proper subtree $\tilde{\mathcal{T}}$ of $\mathcal{T}$ is a collection of nodes of $\mathcal{T}$ with the properties: (i) the root node is in $\tilde{\mathcal{T}}$, (ii) if $C_{j,k}$ is in $\tilde{\mathcal{T}}$ then the parent of $C_{j,k}$ is also in $\tilde{\mathcal{T}}$. Any finite proper subtree $\tilde{\mathcal{T}}$ is associated with a unique partition $\Lambda = \Lambda(\tilde{\mathcal{T}})$ which consists of its outer leaves, by which we mean those $C_{j,k} \in \mathcal{T}$ such that $C_{j,k} \notin \tilde{\mathcal{T}}$ but its parent is in $\tilde{\mathcal{T}}$.

2.2. Empirical GMRA

In practice the master tree $\mathcal{T}$ is not given, nor can be constructed since $\mathcal{M}$ is not known: we will construct one on samples by running a variation of the cover tree algorithm (see Beygelzimer et al., 2006), which only creates candidate "centers" for the $C_{j,k}$, by adding a multiscale partitioning step. From now on we denote the training data by $\mathcal{X}_{2n}$. We randomly split the data into two disjoint groups such that $\mathcal{X}_{2n} = \mathcal{X}'_n \cup \mathcal{X}_n$ where $\mathcal{X}'_n = \{x'_1, \ldots, x'_n\}$ and $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, apply our variation on cover trees on $\mathcal{X}'_n$ to construct a tree satisfying (A1-A5) (see section 2.6). After the tree is constructed, we assign points in the second

1. ρ is regular doubling if there exists $C_1 > 0$ such that $C_1^{-1} r^d \leq \rho(\mathcal{M} \cap B_r(x)) \leq C_1 r^d$ for any $x \in \mathcal{M}$ and $r > 0$. C_1 is called the doubling constant of ρ.

	GMRA	Empirical GMRA
Linear projection on $C_{j,k}$	$\mathcal{P}_{j,k}(x) := c_{j,k} + \mathrm{Proj}_{V_{j,k}}(x - c_{j,k})$	$\widehat{\mathcal{P}}_{j,k}(x) := \widehat{c}_{j,k} + \mathrm{Proj}_{\widehat{V}_{j,k}}(x - \widehat{c}_{j,k})$
Linear projection at scale j	$\mathcal{P}_j := \sum_{k\in\mathcal{K}_j} \mathcal{P}_{j,k}\mathbf{1}_{j,k}$	$\widehat{\mathcal{P}}_j := \sum_{k\in\mathcal{K}_j} \widehat{\mathcal{P}}_{j,k}\mathbf{1}_{j,k}$
Measure	$\rho(C_{j,k})$	$\widehat{\rho}(C_{j,k}) = \widehat{n}_{j,k}/n$
Center	$c_{j,k} := \mathbb{E}_{j,k}x$	$\widehat{c}_{j,k} := \frac{1}{\widehat{n}_{j,k}} \sum_{x_i\in C_{j,k}} x_i$
Principal subspaces	$V_{j,k}$ minimizes $\mathbb{E}_{j,k}\|x - c_{j,k} - \mathrm{Proj}_V(x - c_{j,k})\|^2$ among d-dim subspaces	$\widehat{V}_{j,k}$ minimizes $\frac{1}{\widehat{n}_{j,k}} \sum_{x_i\in C_{j,k}} \|x - \widehat{c}_{j,k} - \mathrm{Proj}_V(x - \widehat{c}_{j,k})\|^2$ among d-dim subspaces
Covariance matrix	$\Sigma_{j,k} := \mathbb{E}_{j,k}(x - c_{j,k})(x - c_{j,k})^T$	$\widehat{\Sigma}_{j,k} := \frac{1}{\widehat{n}_{j,k}} \sum_{x_i\in C_{j,k}} (x_i - \widehat{c}_{j,k})(x_i - \widehat{c}_{j,k})^T$
Inner product with respect to ρ	$\langle \mathcal{P}X, \mathcal{Q}X\rangle := \int_{\mathcal{M}} \langle \mathcal{P}x, \mathcal{Q}x\rangle d\rho$	$1/n \sum_{x_i\in\mathcal{X}_n} \langle \mathcal{P}x_i, \mathcal{Q}x_i\rangle$
Norm with respect to ρ	$\|\mathcal{P}X\| := \left(\int_{\mathcal{M}} \|\mathcal{P}x\|^2 d\rho\right)^{\frac{1}{2}}$	$\left(1/n \sum_{x_i\in\mathcal{X}_n} \|\mathcal{P}x_i\|^2\right)^{\frac{1}{2}}$

Table 1: This table summarizes GMRA-related quantities and their empirical counterparts (Allard et al., 2012; Maggioni et al., 2016). $\mathbf{1}_{j,k}$ is the indicator function on $C_{j,k}$ (i.e.,$\mathbf{1}_{j,k}(x) = 1$ if $x \in C_{j,k}$ and 0 otherwise). Here $\mathbb{E}_{j,k}$ stands for expectation with respect to the conditional distribution $d\rho_{|C_{j,k}}$. The measure of $C_{j,k}$ is $\rho(C_{j,k})$ and the empirical measure is $\widehat{\rho}(C_{j,k}) = \widehat{n}_{j,k}/n$ where $\widehat{n}_{j,k}$ is the number of points in $C_{j,k}$. $V_{j,k}$ and $\widehat{V}_{j,k}$ are the eigen-spaces associated with the largest d eigenvalues of $\Sigma_{j,k}$ and $\widehat{\Sigma}_{j,k}$ respectively. Here $\mathcal{P}, \mathcal{Q}$: $\mathcal{M} \to \mathbb{R}^D$ are two operators.

half of data $\mathcal{X}_n$, to the appropriate cells. In this way we obtain a family of multiscale partitions for the points in $\mathcal{X}_n$, which we truncate to the largest subtree whose leaves contain at least d points in $\mathcal{X}_n$. This subtree is called the *data master tree*, denoted by $\mathcal{T}^n$. We then use $\mathcal{X}_n$ to perform local PCA to obtain the empirical mean $\widehat{c}_{j,k}$ and the empirical d-dimensional principal subspace $\widehat{V}_{j,k}$ on each $C_{j,k}$. Define the empirical projection $\widehat{\mathcal{P}}_{j,k}(x) := \widehat{c}_{j,k} + \mathrm{Proj}_{\widehat{V}_{j,k}}(x - \widehat{c}_{j,k})$ for $x \in C_{j,k}$. Table 1 summarizes the GMRA-related quantities and their empirical counterparts.

2.3. Geometric Multi-Resolution Analysis: uniform partitions

GMRA with respect to the distribution ρ associated with the multiscale tree $\mathcal{T}$ consists a collection of piecewise affine projectors $\{\mathcal{P}_j : \mathbb{R}^D \to \mathbb{R}^D\}_{j\geq j_{\min}}$ on the multiscale partitions $\{\Lambda_j := \{C_{j,k}\}_{k\in\mathcal{K}_j}\}_{j\geq j_{\min}}$. At scale j, $\mathcal{M}$ is approximated by the piecewise linear sets $\{\mathcal{P}_{j,k}(C_{j,k})\}_{k\in\mathcal{K}_j}$. The approximation error of $\mathcal{M}$ by the empirical linear sets $\{\widehat{\mathcal{P}}_{j,k}(C_{j,k})\}_{k\in\mathcal{K}_j}$ is defined as:

$$\mathbb{E}\|X - \widehat{\mathcal{P}}_j X\|^2 = \mathbb{E}\int_{\mathcal{M}} \|x - \widehat{\mathcal{P}}_j x\|^2 d\rho = \mathbb{E}\sum_{k\in\mathcal{K}_j} \int_{C_{j,k}} \|x - \widehat{\mathcal{P}}_{j,k} x\|^2 d\rho$$

where $\widehat{\mathcal{P}}_j$ and $\widehat{\mathcal{P}}_{j,k}$ are built from random samples $x_i \sim \rho$ (according to the GMRA algorithm), X is a random vector distributed according to ρ, and the expectation is taken over X. The squared approximation error above is also called the Mean Square Error (MSE) of GMRA. In order to understand the error, we split it into a bias term and a variance term:

$$\mathbb{E}\|X - \widehat{\mathcal{P}}_j X\| \le \underbrace{\|X - \mathcal{P}_j X\|}_{\text{bias}} + \mathbb{E}\underbrace{\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|}_{\sqrt{\text{variance}}}. \tag{1}$$

To bound the bias term, we need regularity assumptions on ρ, while for the variance term we prove concentration bounds of the relevant quantities around their expected values.

For a fixed distribution ρ, the approximation error of $\mathcal{M}$ at scale j, measured by $\|X - \mathcal{P}_j X\|$, decays at a rate dependent on the regularity of $\mathcal{M}$ in the ρ-measure (see Allard et al., 2012). We quantify the regularity of ρ as follows:

Definition 1 (Model class $\mathcal{A}_s$) *A probability measure ρ supported on $\mathcal{M}$ is in $\mathcal{A}_s$ if*

$$|\rho|_{\mathcal{A}_s} = \sup_{\mathcal{T}} \inf\{A_0 : \|X - \mathcal{P}_j X\| \le A_0 2^{-js}, \forall\, j \ge j_{\min}\} < \infty, \tag{2}$$

where $\mathcal{T}$ varies over the set, assumed non-empty, of multiscale tree decompositions satisfying Assumptions (A1-A5).

We capture the case where the L^2 approximation error is roughly the same on every cell with the following definition:

Definition 2 (Model class $\mathcal{A}_s^\infty$) *A probability measure ρ supported on $\mathcal{M}$ is in $\mathcal{A}_s^\infty$ if*

$$|\rho|_{\mathcal{A}_s^\infty} = \sup_{\mathcal{T}} \inf\{A_0 : \|(X - \mathcal{P}_{j,k}X)\mathbf{1}_{j,k}\| \le A_0 2^{-js}\sqrt{\rho(C_{j,k})},\ \forall\, k \in \mathcal{K}_j, j \ge j_{\min}\} < \infty \tag{3}$$

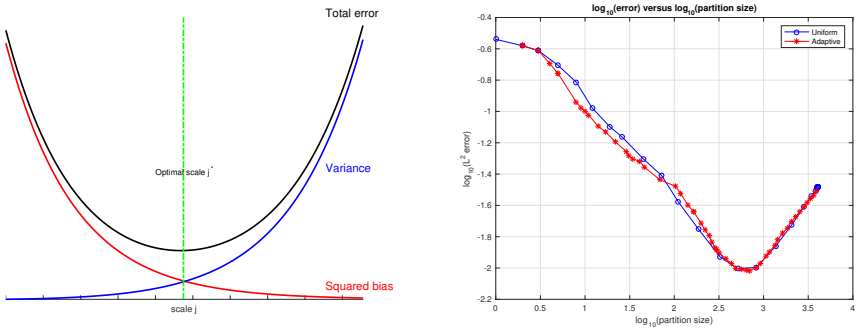
where $\mathcal{T}$ varies over the set, assumed non-empty, of multiscale tree decompositions satisfying Assumptions (A1-A5).

Clearly $\mathcal{A}_s^\infty \subset \mathcal{A}_s$. Also, since $\operatorname{diam}(C_{j,k}) \le 2\theta_2 2^{-j}$, necessarily $\|(\mathbb{I} - \mathcal{P}_{j,k})\mathbf{1}_{j,k}X\| \le \theta_2 2^{-j}\sqrt{\rho(C_{j,k})},\ \forall\, k \in \mathcal{K}_j, j \ge j_{\min}$, and therefore $\rho \in \mathcal{A}_1^\infty$ in any case. Moreover, these classes contain suitable measures supported on manifolds:

Proposition 3 *Let $\mathcal{M}$ be a closed manifold of class $\mathcal{C}^s$, $s \in (1, 2]$ isometrically embedded in $\mathbb{R}^D$, and ρ be a doubing probability measure on $\mathcal{M}$ with the doubling constant C_1. Then our construction of $\{C_{j,k}\}_{k\in\mathcal{K}_j, j\ge j_{\min}}$ in Section 2.6 satisfies (A1-A5), and $\rho \in \mathcal{A}_s^\infty$.*

The proof is postponed to Appendix A.2.

Example 1 *We consider the d-dim S-manifold whose x_1 and x_2 coordinates are on an S-shaped curve and x_i ranges in $[0, 1]$ for $i = 3, \ldots, d+1$. By the Proposition just stated, the volume measure on this S-manifold is in $\mathcal{A}_2^\infty$. Numerically one can identify s from data sampled from $\rho \in \mathcal{A}_s$ as the slope of the line approximating $\log_{10}\|X - \mathcal{P}_j X\|$ as a function of $\log_{10} r_j$ where r_j is the average diameter of $C_{j,k}$'s at scale j. Our numerical experiments in Figure 5 (b) give rise to $s \approx 2.0, 2.1, 2.1$ when $d = 3, 4, 5$ respectively.*



(a) Bias and variance tradeoff (b) Error versus the partition size

Figure 2: (a) Plot of the bias and variance estimates in Eq. (1), with $s = 2, d = 5, n = 100$. (b) shows the approximation error on test data versus the partition size in GMRA and Adaptive GMRA for the 3-dim S-manifold. When the partition size is between 1 and $10^{2.8}$, the bias dominates the error so the error decreases; after that, the variance dominates the error, which becomes increasing.

Example 2 *As a comparison we consider the d-dimensional Z-manifold whose x_1 and x_2 coordinates are on a Z-shaped curve and x_i ranges in $[0, 1]$, for $i = 3, \ldots, d+1$. Volume measure on the Z manifold is in $\mathcal{A}_{1.5}$ (see appendix B.2). Our numerical experiments in Figure 5 (c) give rise to $s \approx 1.5, 1.7, 1.6$ when $d = 3, 4, 5$ respectively.*

The squared bias in (1) satisfies $\|X - \mathcal{P}_j X\|^2 \leq |\rho|^2_{\mathcal{A}_s} 2^{-2js}$ whenever $\rho \in \mathcal{A}_s$ (by definition of $\mathcal{A}_s$). In Proposition 16 we will show that the variance term is estimated in terms of the sample size n and the scale j as follows:

$$\mathbb{E}\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|^2 \leq \frac{d^2 \#\Lambda_j \log[\alpha d \#\Lambda_j]}{\beta 2^{2j} n} = \mathcal{O}\left(\frac{j 2^{j(d-2)}}{n}\right),$$

where α, β are constants depending on θ_2, θ_3. In the case $d = 1$ both the squared bias and the variance decrease as j increases, so choosing the finest scale of the data tree $\mathcal{T}^n$ yields the best rate of convergence. When $d \geq 2$, the squared bias decreases but the variance increases as j gets large as shown Figure 2, as a manifestation of the classical bias-variance tradeoff, except that it arises here in a geometric setting (a related instance of this phenomenon appears in Canas et al. (2012)). By choosing a proper scale j^* to balance these two terms, we obtain the following rate of convergence for empirical GMRA truncated at scale j^*:

Theorem 4 *Suppose $\rho \in \mathcal{A}_s$ for $s \geq 1$. Let $\nu > 0$ be arbitrary and fix $\mu > 0$. Let j^* be chosen such that*

$$2^{-j^*} = \begin{cases} \mu \frac{\log n}{n} & \text{for } d = 1 \\ \mu \left(\frac{\log n}{n}\right)^{\frac{1}{2s+d-2}}, & \text{for } d \geq 2 \end{cases} \tag{4}$$

then there exists $C_1 := C_1(\theta_1, \theta_2, \theta_3, \theta_4, d, \nu, \mu)$ *and* $C_2 := C_2(\theta_1, \theta_2, \theta_3, \theta_4, d, \mu)$ *such that:*

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{j^*} X\| \geq (|\rho|_{\mathcal{A}_s}\mu^s + C_1)\frac{\log n}{n}\right\} \leq C_2 n^{-\nu}, \quad \textit{for } d = 1, \tag{5}$$

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{j^*} X\| \geq (|\rho|_{\mathcal{A}_s}\mu^s + C_1)\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d-2}}\right\} \leq C_2 n^{-\nu}, \quad \textit{for } d \geq 2\,. \tag{6}$$

Theorem 4 is proved in Section 4.2. From the perspective of dictionary learning, it says that GMRA provides a dictionary Φ_{j^*} of cardinality $\asymp dn/\log n$ for $d = 1$ and $\asymp d(n/\log n)^{\frac{d}{2s+d-2}}$ for $d \geq 2$, consisting of the principal directions in each of the $C_{j^*,k}$'s (forming the columns of $\widehat{V}_{j^*,x}$) and the means of the $C_{j^*,k}$'s, so that every x sampled from ρ (and not just samples in the training data) may be encoded with a vector with $d+1$ nonzero entries: one entry encodes the location k of x on the tree, e.g. $(j^*, x) = (j^*, k)$ such that $x \in C_{j^*,k}$, and the other d entries are the coefficients $\widehat{V}_{j^*,x}^T(x - \widehat{c}_{j^*,x})$. We also remind the reader that GMRA automatically constructs a fast transform mapping points x to the vector representing Φ_{j^*} (See Allard et al. (2012); Maggioni et al. (2016) for a discussion). Note that by choosing ν large enough in the Theorem,

$$(6) \implies \text{MSE} = \mathbb{E}\|X - \widehat{\mathcal{P}}_{j^*} X\|^2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}},$$

and (5) implies MSE $\lesssim (\frac{\log n}{n})^2$ for $d = 1$. Clearly, one could fix a desired MSE of size ε^2, and obtain a dictionary of size dependent only on ε and independent of n, for n sufficiently large, thereby obtaining a way of compressing data (for further discussion on this point see Maggioni et al. (2016), where also a special case of Theorem 4 with $s = 2$ was proved).

2.4. Geometric Multi-Resolution Analysis: Adaptive Partitions

The performance guarantee in Theorem 4 is not fully satisfactory for two reasons: (i) the regularity parameter s is required to be known to choose the optimal scale j^*, and this parameter is typically unknown in any practical setting, and (ii) none of the uniform partitions $\{C_{j,k}\}_{k\in\mathcal{K}_j}$ will be optimal if the regularity of ρ (and/or $\mathcal{M}$) varies at different locations and scales. This lack of uniformity in regularity can appear in a wide variety of data sets for various reasons: when clusters exist that have cores denser than the remaining regions of space, when sampled trajectories of a dynamical system linger in certain regions of space for much longer time intervals than others (e.g. metastable states in molecular

	Definition (infinite sample)	Empirical version
Difference operator	$\mathcal{Q}_{j,k} := (\mathcal{P}_j - \mathcal{P}_{j+1})\mathbf{1}_{j,k}$	$\widehat{\mathcal{Q}}_{j,k} := (\widehat{\mathcal{P}}_j - \widehat{\mathcal{P}}_{j+1})\mathbf{1}_{j,k}$
Norm of difference	$\Delta_{j,k}^2 := \int_{C_{j,k}} \|\mathcal{Q}_{j,k}x\|^2 d\rho$	$\widehat{\Delta}_{j,k}^2 := \frac{1}{n}\sum_{x_i\in C_{j,k}} \|\widehat{\mathcal{Q}}_{j,k}x_i\|^2$

Table 2: Refinement criterion and the empirical counterparts

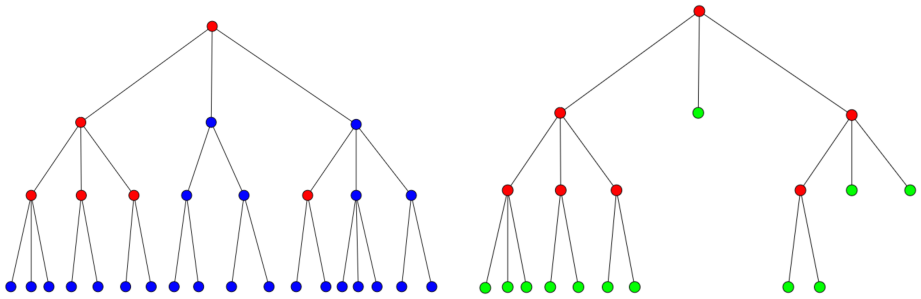


Figure 3: Left: a master tree in which red nodes satisfy $\Delta_{j,k} \geq 2^{-j}\tau_n$ but blue nodes do not. Right: the subtree of the red nodes is the smallest proper subtree that contains all the nodes satisfying $\Delta_{j,k} \geq 2^{-j}\tau_n$, i.e. were red in the figure on the left. Green nodes form the adaptive partition.

dynamics (Rohrdanz et al., 2011; Zheng et al., 2011)), in data sets of images where details exist at different level of resolutions, affecting regularity at different scales in the ambient space, and so on. To fix the ideas we consider again one simplest manifestations of this phenomenon in the examples considered above: uniform partitions work well for the volume measure on the S-manifold but are not optimal for the volume measure on the Z-manifold, for which the ideal partition is coarse on flat regions but finer at and near the corners (see Figure 4). In applications, for example to mesh approximation, it is often the case that the point clouds to be approximated are not uniformly smooth and include different levels of details at different locations and scales (see Figure 9). We therefore propose an adaptive version of GMRA that automatically adapts to the regularity of the data and choose a near-optimal partition.

We expect $\Delta_{j,k}$ defined in Table 2 to be small on approximately flat regions, and large $\Delta_{j,k}$ at many scales at irregular locations. We also expect $\widehat{\Delta}_{j,k}$ to have the same behavior, at least when $\widehat{\Delta}_{j,k}$ is with high confidence close to $\Delta_{j,k}$. We see this phenomenon represented in Figure 4 (a,b): as j increases, for the S-manifold $\|\widehat{\mathcal{P}}_{j+1}x_i - \widehat{\mathcal{P}}_j x_i\|$ decays uniformly at all points, while for the Z-manifold, the same quantity decays rapidly on flat regions but remains large even at fine scales near the corners (where "near" is scale-dependent, decreasing with scale). We wish to include in our approximation the nodes where this quantity is large, since we may expect a large improvement in approximation by including such nodes. However if too few samples exist in a node, then this quantity is not to be trusted, because its variance is large. It turns out that it is enough to consider the following criterion: let $\widehat{\mathcal{T}}_{\tau_n}$ be the smallest proper subtree of $\mathcal{T}^n$ that contains all $C_{j,k} \in \mathcal{T}^n$ for which $\widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n$ where $\tau_n = \kappa\sqrt{(\log n)/n}$. Crucially, κ may be chosen independently of the regularity index (see Theorem 8). Empirical Adaptive GMRA returns piecewise affine projectors on $\widehat{\Lambda}_{\tau_n}$, the partition associated with the outer leaves of $\widehat{\mathcal{T}}_{\tau_n}$. Our algorithm is summarized in Algorithm 1.

Algorithm 1 - Adaptive GMRA

Input: data $\mathcal{X}_{2n} = \mathcal{X}'_n \cup \mathcal{X}_n$, intrinsic dimension d, threshold κ

Output: $\mathcal{T}^n$, $\{C_{j,k}\}$, $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}$: multiscale tree, corresponding cells and adaptive piecewise linear projectors on an adaptive partition.

1: Construct $\mathcal{T}^n$ and $\{C_{j,k}\}$ from $\mathcal{X}'_n$
2: Now use $\mathcal{X}_n$. Compute $\widehat{\mathcal{P}}_{j,k}$ and $\widehat{\Delta}_{j,k}$ on every node $C_{j,k} \in \mathcal{T}^n$.
3: $\widehat{\mathcal{T}}_{\tau_n} \leftarrow$ smallest proper subtree of $\mathcal{T}^n$ containing all $C_{j,k} \in \mathcal{T}^n : \widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n$ where $\tau_n = \kappa\sqrt{(\log n)/n}$.
4: $\widehat{\Lambda}_{\tau_n} \leftarrow$ the partition associated with outer leaves of $\widehat{\mathcal{T}}_{\tau_n}$
5: $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} \leftarrow \sum_{C_{j,k} \in \widehat{\Lambda}_{\tau_n}} \widehat{\mathcal{P}}_{j,k} \mathbf{1}_{j,k}$.

Adaptive partitions may be effectively selected with a criterion that determines whether or not a cell should participate in the adaptive partition. The quantities involved in the selection and their empirical version are summarized in Table 2.

We will provide a finite sample performance guarantee of the empirical Adaptive GMRA for a model class that is more general than $\mathcal{A}_s^\infty$. Given any fixed threshold $\eta > 0$, we let $\mathcal{T}_{(\rho,\eta)}$ be the smallest proper subtree of $\mathcal{T}$ that contains all $C_{j,k} \in \mathcal{T}$ for which $\Delta_{j,k} \geq 2^{-j}\eta$. The corresponding adaptive partition $\Lambda_{(\rho,\eta)}$ consists of the outer leaves of $\mathcal{T}_{(\rho,\eta)}$. We let $\#_j\mathcal{T}_{(\rho,\eta)}$ be the number of cells in $\mathcal{T}_{(\rho,\eta)}$ at scale j.

Definition 5 (Model class $\mathcal{B}_s$) *In the case $d \geq 3$, given $s > 0$, a probability measure ρ supported on $\mathcal{M}$ is in $\mathcal{B}_s$ if ρ satisfies the following regularity condition:*

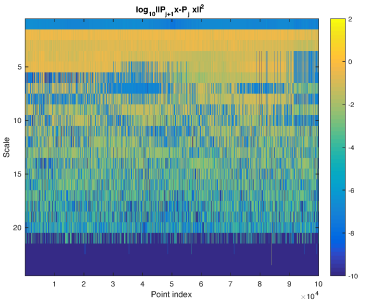
$$|\rho|_{\mathcal{B}_s} := \left(\sup_{\mathcal{T}} \sup_{\eta>0} \eta^p \sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)} \right)^{\frac{1}{p}} < \infty, \quad \text{with } p = \frac{2(d-2)}{2s+d-2} \tag{7}$$

where $\mathcal{T}$ varies over the set, assumed nonempty, of multiscale tree decompositions satisfying Assumptions (A1-A5).

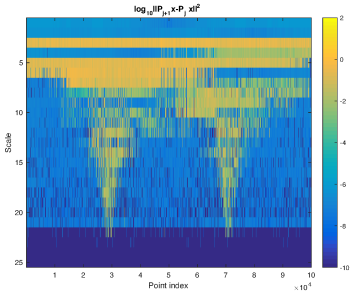
For elements in the model class $\mathcal{B}_s$ we have control on the growth rate of the truncated tree $\mathcal{T}_{(\rho,\eta)}$ as η decreases, namely it is $\mathcal{O}(\eta^{-p})$. Our key estimates on variance and sample complexity in Lemma 15 indicate that the natural measure of the complexity of $\mathcal{T}_{(\rho,\eta)}$ is the weighted tree complexity measure $\sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)}$ in the definition above. First of all, the class $\mathcal{B}_s$ is indeed larger than $\mathcal{A}_s^\infty$ (see appendix A.4 for a proof):

Lemma 6 *$\mathcal{B}_s$ is a more general model class than $\mathcal{A}_s^\infty$: if $\rho \in \mathcal{A}_s^\infty$, then $\rho \in \mathcal{B}_s$ and $|\rho|_{\mathcal{B}_s} \lesssim |\rho|_{\mathcal{A}_s^\infty}$.*

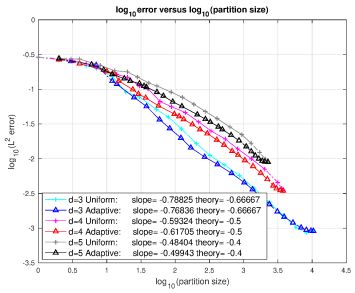
Example 3 *The volume measures on the d-dim ($d \geq 3$) S-manifold and Z-manifold are in $\mathcal{B}_2$ and $\mathcal{B}_{1.5(d-2)/(d-3)}$ respectively (see appendix B). In numerical experiments, s can be approximated by the negative of the slope in the log-log plot of $\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_\eta} X\|^{d-2}$ versus the weighted complexity of the truncated tree according to Eq. (9): see numerical examples in Figure 5.*



(a) S-manifold

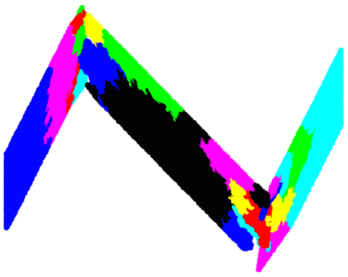


(b) Z-manifold

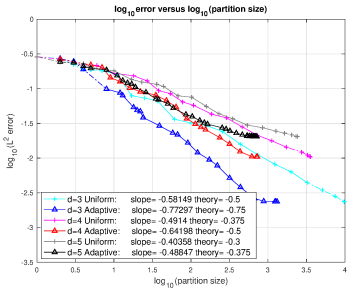


(c) Error vs partition size, S-manifold

(d) Adaptive approx. of S



(e) Adaptive approx. of Z



(f) Error vs partition size, Z-manifold

Figure 4: (a,b): $\log_{10} ||\widehat{\mathcal{P}}_j(x_i) - \widehat{\mathcal{P}}_{j+1}(x_i)||$ from the coarsest scale (top) to the finest scale (bottom), with columns indexed by points, which, for visualization purposes only, are sorted roughly from "left to right" on the manifold. (d,e): adaptive approximations: for the S-manifold the adaptive approximation is close to a uniform approximation, but for the Z-manifold it contains few large pieces near the almost-flat regions, and several small pieces near the "corners". (c,f): log-log plot of the approximation error versus the partition size in GMRA and Adaptive GMRA respectively. Theoretically, the slope is $-2/d$ in both GMRA and Adaptive GMRA for the S-manifold. For the Z-manifold, the slope is $-1.5/d$ in GMRA and $-1.5/(d-1)$ in Adaptive GMRA (see appendix B).

We also need a quasi-orthogonality condition which says that the operators $\{\mathcal{Q}_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$ applied on $\mathcal{M}$ are mostly orthogonal across scales and/or $\|\mathcal{Q}_{j,k}X\|$ quickly decays.

Definition 7 (Quasi-orthogonality) *There exists a constant $B_0 > 0$ such that for any proper subtree $\tilde{\mathcal{T}}$ of any master tree $\mathcal{T}$ satisfying Assumptions (A1-A5),*

$$\| \sum_{C_{j,k}\notin\tilde{\mathcal{T}}} \mathcal{Q}_{j,k}X\|^2 \leq B_0 \sum_{C_{j,k}\notin\tilde{\mathcal{T}}} \|\mathcal{Q}_{j,k}X\|^2. \tag{8}$$

We postpone further discussion of this condition to Section 5.2. One can show (see appendix D) that in the case $d \geq 3$, $\rho \in \mathcal{B}_s$ along with quasi-orthogonality implies a certain rate of

approximation of X by $\mathcal{P}_{\Lambda_{(\rho,\eta)}}X$, as $\eta \to 0^+$:

$$\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}}X\|^2 \le B_{s,d}|\rho|_{\mathcal{B}_s}^p \eta^{2-p} \le B_{s,d}|\rho|_{\mathcal{B}_s}^2 \left(\sum_{j\ge j_{\min}} 2^{-2j}\#_j\mathcal{T}_{(\rho,\eta)}\right)^{-\frac{2s}{d-2}}, \tag{9}$$

where $s = \frac{(d-2)(2-p)}{2p}$ and $B_{s,d} := B_0 2^p/(1-2^{p-2})$.

The main result of this paper is the following performance analysis of empirical Adaptive GMRA (see the proof in Section 4.3).

Theorem 8 *Suppose ρ satisfies quasi-orthogonality and $\mathcal{M}$ is bounded: $\mathcal{M} \subset B_M(0)$. Let $\nu > 0$. There exists $\kappa_0(\theta_2, \theta_3, \theta_4, a_{\max}, d, \nu)$ such that if $\tau_n = \kappa\sqrt{(\log n)/n}$ with $\kappa \ge \kappa_0$, the following holds:*

(i) if $d \ge 3$ and $\rho \in \mathcal{B}_s$ for some $s > 0$, there are c_1 and c_2 such that

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}X\| \ge c_1\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d-2}}\right\} \le c_2 n^{-\nu}. \tag{10}$$

(ii) if $d = 1$, there exist c_1 and c_2 such that

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}X\| \ge c_1\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\right\} \le c_2 n^{-\nu}. \tag{11}$$

(iii) if $d = 2$ and

$$|\rho| := \sup_{\mathcal{T}}\sup_{\eta>0}\frac{1}{\log\frac{1}{\eta}}\sum_{j\ge j_{\min}} 2^{-2j}\#_j\mathcal{T}_{(\rho,\eta)} < +\infty,$$

then there exist c_1 and c_2 such that

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}X\| \ge c_1\left(\frac{\log^2 n}{n}\right)^{\frac{1}{2}}\right\} \le c_2 n^{-\nu}. \tag{12}$$

Notice that by choosing ν large enough, we have

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}X\| \ge c_1\left(\frac{\log^\alpha n}{n}\right)^{\beta}\right\} \le c_2 n^{-\nu} \Rightarrow \mathrm{MSE} \le c_1\left(\frac{\log^\alpha n}{n}\right)^{2\beta},$$

so we also have $\mathrm{MSE} \lesssim (\log n/n)^{\frac{2s}{2s+d-2}}$ for $d \ge 3$ and $\mathrm{MSE} \lesssim \log^d n/n$ for $d = 1, 2$.

The dependencies of the constants in Theorem 8 on the geometric constants are as follows:

$$\begin{array}{lll}
d \ge 3: & c_1 = c_1(\theta_{2,3,4}, a_{\max}, d, s, \kappa, |\rho|_{\mathcal{B}_s}, B_0, \nu), & c_2 = c_2(\theta_{2,3,4}, a_{\min}, a_{\max}, d, s, \kappa, |\rho|_{\mathcal{B}_s}, B_0).\\
d = 2: & c_1 = c_1(\theta_{2,3,4}, a_{\max}, d, \kappa, |\rho|_{\mathcal{B}_s}, B_0, \nu), & c_2 = c_2(\theta_{2,3,4}, a_{\min}, a_{\max}, d, \kappa, |\rho|_{\mathcal{B}_s}, B_0).\\
d = 1: & c_1 = c_1(\theta_{1,2,3,4}, a_{\max}, d, \kappa, B_0, \nu), & c_2 = c_2(\theta_{1,2,3,4}, a_{\min}, a_{\max}, d, \kappa, B_0).
\end{array}$$

Theorem 8 is more satisfactory than Theorem 4 for two reasons: (i) when $d \geq 3$, the same rate $(\log n/n)^{2s/(2s+d-2)}$ is proved for the model class $\mathcal{B}_s$ which is larger than $\mathcal{A}_s^\infty$; (ii) the threshold-based estimator is adaptive: it does not require a priori knowledge of the regularity s, since the choice of κ is independent of s, yet it achieves the rate as if it knew the optimal regularity parameter s.

From the perspective of dictionary learning, when $d \geq 3$, Adaptive GMRA provides a dictionary $\Phi_{\widehat{\Lambda}_{\tau_n}}$ associated with a tree of weighted complexity $(n/\log n)^{d-2/(2s+d-2)}$, so that every x sampled from ρ may be encoded by a vector with $d+1$ nonzero entries, among which one encodes the location of x in the adaptive partition and the other d entries are the local principal coefficients of x.

For a given accuracy ε, in order to achieve MSE $\lesssim \varepsilon^2$, the number of samples we need is $n_\varepsilon \gtrsim (1/\varepsilon)^{(2s+d-2)/s} \log(1/\varepsilon)$. When s is unknown, we can determine s as follows: we fix a small n_0 and run Adaptive GMRA with $n_0, 2n_0, 4n_0, \ldots, Cn_0$ samples. For each sample size, we evenly split data into a training set to construct Adaptive GMRA and a test set to evaluate the MSE. According to Theorem 8, the MSE scales like $[(\log n)/n]^{\frac{2s}{2s+d-2}}$ where n is the sample size. Therefore, the slope in the log-log plot of the MSE versus n gives an approximation of $-2s/(2s+d-2)$.

The threshold τ_n in our adaptive algorithm is independent of s since κ_0 does not depend on s, which means our adaptive algorithm does not require s as a priori information but rather will learn it from data.

Remark 9 *It would also be natural to consider another stopping criterion:* $\mathcal{E}_{j,k}^2 := \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} \|\mathcal{P}_j x - x\|^2 d\rho \leq \eta^2$ *which suggests stopping refinement to finer scales if the approximation error is below certain threshold. The reason why we do not adopt this stopping criterion is that in this case the threshold* η *would have to depend on* s *in order to guarantee the (adaptive) rate* MSE $\lesssim (\log n/n)^{2s/(2s+d-2)}$ *for* $d \geq 3$*. More precisely, for any threshold* $\eta > 0$*, let* $\mathcal{T}_{(\rho,\eta)}^{\mathcal{E}}$ *be the smallest proper subtree of* $\mathcal{T}$ *whose leaves satisfy* $\mathcal{E}_{j,k}^2 \leq \eta^2$*. The corresponding adaptive partition* $\Lambda_{(\rho,\eta)}^{\mathcal{E}}$ *consists of the leaves of* $\mathcal{T}_{(\rho,\eta)}^{\mathcal{E}}$*. This stopping criterion guarantees* $\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}^{\mathcal{E}}} X\| \leq \eta$*. It is natural to define the model class* $\mathcal{F}_s$ *in the case* $d \geq 3$ *to be the set of probability measures* ρ *supported on* $\mathcal{M}$ *such that* $\sup_{\mathcal{T}} \sup_{\eta>0} \eta^{(d-2)/s} \sum_{j \geq j_{\min}} 2^{-2j} \#_j \Lambda_{(\rho,\eta)}^{\mathcal{E}} < \infty$ *where* $\mathcal{T}$ *varies over the set of multiscale tree decompositions satisfying (A1-A5). One can show that* $\mathcal{A}_s^\infty \subsetneq \mathcal{F}_s$*. As an analogue of Theorem 8, we can prove that, there exists* $\kappa_0 > 0$ *such that if our adaptive algorithm adopts the stopping criterion* $\widehat{\mathcal{E}}_{j,k} \leq \tau_n^{\mathcal{E}}$ *where the threshold is chosen as* $\tau_n^{\mathcal{E}} = \kappa (\log n/n)^{\frac{s}{2s+d-2}}$ *with* $\kappa \geq \kappa_0$*, then the empirical approximation on the adaptive partition* $\widehat{\Lambda}_{\tau_n^{\mathcal{E}}}$ *satisfies* MSE $= \|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n^{\mathcal{E}}}} X\|^2 \lesssim (\log n/n)^{2s/(2s+d-2)}$*. With this stopping criterion, the threshold* $\tau_n^{\mathcal{E}}$ *would require knowing* s*, unlike in Theorem 8.*

Theorem 8 is stated when $\mathcal{M}$ is bounded. The assumption of the boundedness of $\mathcal{M}$ is largely irrelevant, and may be replaced by a weaker assumption on the decay of ρ.

Theorem 10 *Let* $d \geq 3$*,* $s, \delta, \lambda, \mu > 0$*. Assume that there exists* C_1 *such that*

$$\int_{B_R(0)^c} ||x||^2 d\rho \leq C_1 R^{-\delta}, \ \forall R \geq R_0.$$

Suppose ρ satisfies quasi-orthogonality. If ρ restricted on $B_R(0)$, denoted by $\rho_{|_{B_R(0)}}$, is in $\mathcal{B}_s$ for every $R \geq R_0$ and $(|\rho_{|_{B_R(0)}}|_{\mathcal{B}_s})^p \leq C_2 R^\lambda$ for some $C_2 > 0$, where $p = \frac{2(d-2)}{2s+d-2}$. Then there exists $\kappa_0(\theta_2, \theta_3, \theta_4, a_{\max}, d, \nu)$ such that if $\tau_n = \kappa\sqrt{\log n/n}$ with $\kappa \geq \kappa_0$, we have

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X\| \geq c_1 \left(\frac{\log n}{n}\right)^{\frac{s}{2s+d-2}\frac{\delta}{\delta+\max(\lambda,2)}}\right\} \leq c_2 n^{-\nu} \tag{13}$$

for some c_1, c_2 independent of n, where the estimator $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X$ is obtained by Adaptive GMRA within $B_{R_n}(0)$ where $R_n = \max(R_0, \mu(n/\log n)^{\frac{2s}{(2s+d-2)(\delta+\max(\lambda,2))}})$, and is equal to 0 for the points outside $B_{R_n}(0)$.

Theorem 10 is proved at the end of Section 4.3. It implies MSE $\lesssim (\log n/n)^{\frac{2s}{2s+d-2}\cdot\frac{\delta}{\delta+\max(\lambda,2)}}$. As δ increases, i.e., $\delta \to +\infty$, the MSE approaches $(\log n/n)^{\frac{2s}{2s+d-2}}$, which is consistent with Theorem 8 for bounded $\mathcal{M}$. Similar results, with similar proofs, would hold under different assumptions on the decay of ρ; for example for ρ decaying at least exponentially, only additional $\log n$ terms in the rate would be lost compared in Theorem 8.

Remark 11 *We claim that λ is not large in simple cases. For example, if $\rho \in \mathcal{A}_s^\infty$ and ρ decays in the radial direction in such a way that $\rho(C_{j,k}) \leq C 2^{-jd} \|c_{j,k}\|^{-(d+1+\delta)}$, it is easy to show that $\rho_{|_{B_R(0)}} \in \mathcal{B}_s$ for all $R > 0$ and $|\rho_{|_{B_R(0)}}|_{\mathcal{B}_s}^p \leq R^\lambda$ with $\lambda = d - \frac{(d+1+\delta)(d-2)}{2s+d-2}$ (see the end of Section 4.3).*

Remark 12 *Suppose that ρ was supported in a tube of radius σ around a d-dimensional manifold $\mathcal{M}$, a model that can account both for (bounded) noise and situations where data is not exactly on a manifold, but close to it, as in Maggioni et al. (2016). Then Theorem 8 and Theorem 10 apply in this case, provided one stops the estimator at a scale j such that $2^{-j} \gtrsim \sigma$.*

Remark 13 *In these Theorems we are assuming that d is given because it can be estimated using existing techniques, see Little et al. (2017) and many references therein.*

2.5. Connection to previous works

The works by Allard et al. (2012) and Maggioni et al. (2016) are natural predecessors to this work. In Allard et al. (2012), GMRA and orthogonal GMRA were proposed as data-driven dictionary learning tools to analyze intrinsically low-dimensional point clouds in a high dimensional space. The bias $\|X - \mathcal{P}_j X\|$ were estimated for volume measures on $\mathcal{C}^s, s \in (1, 2]$ manifolds . The performance of GMRA, including sparsity guarantees and computational costs, were systematically studied and tested on both simulated and real data. In Maggioni et al. (2016) the finite sample behavior of empirical GMRA was studied. A non-asymptotic probabilistic bound on the approximation error $\|X - \widehat{\mathcal{P}}_j X\|$ for the model class $\mathcal{A}_2$ (a special case of Theorem 4 with $s = 2$) was established. It was further proved that if the measure ρ is absolutely continuous with respect to the volume measure on a tube of a bounded $\mathcal{C}^2$ manifold with a finite reach, then ρ is in $\mathcal{A}_2$. Running the cover

tree algorithm on data gives rise to a family of multiscale partitions satisfying Assumption (A3-A5). The analysis in Maggioni et al. (2016) robustly accounts for noise and modeling errors as the probability measure is concentrated "near" a manifold. This work extends GMRA by introducing Adaptive GMRA, where low-dimensional linear approximations of $\mathcal{M}$ are built on adaptive partitions at different scales. The finite sample performance of Adaptive GMRA is proved for a large model class. Adaptive GMRA takes full advantage of the multiscale structure of GMRA in order to model data sets of varying complexity across locations and scales. We also generalize the finite sample analysis of empirical GMRA from $\mathcal{A}_2$ to $\mathcal{A}_s$, and analyze the finite sample behavior of orthogonal GMRA and adaptive orthogonal GMRA.

In a different direction, a popular learning algorithm for fitting low-dimensional planes to data is k-flats: let $\mathcal{F}_k$ be the collections of k flats (affine spaces) of dimension d. Given data $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, k-flats solves the optimization problem

$$\min_{S \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} \mathrm{dist}^2(x_i, S) \tag{14}$$

where $\mathrm{dist}(x, S) = \inf_{y \in S} \|x - y\|$. Even though a global minimizer of (14) exists, it is hard to attain due to the non-convexity of the model class $\mathcal{F}_k$, and practitioners are aware that many local minima that are significantly worse than the global minimum exist. While often k is considered given, it may be in fact chosen from the data: for example Theorem 4 in Canas et al. (2012) implies that, given n samples from a probability measure that is absolutely continuous with respect to the volume measure on a smooth d-dimensional manifold $\mathcal{M}$, the expected (out-of-sample) L^2 approximation error of $\mathcal{M}$ by $k_n = C_1(\mathcal{M}, \rho) n^{\frac{d}{2(d+4)}}$ planes is of order $\mathcal{O}(n^{-\frac{2}{d+4}})$. This result is comparable with our Theorem 4 in the case $s = 2$ which says that the L^2 error by empirical GMRA at the scale j such that $2^j \asymp (n/\log n)^{\frac{1}{d+2}}$ achieves a faster rate $\mathcal{O}(n^{-\frac{2}{d+2}})$. So we not only achieve a better rate, but we do so with provable and fast algorithms, that are nonlinear but do not require non-convex optimization.

Multiscale adaptive estimation has been an intensive research area for decades. In the pioneering works by Donoho and Johnstone (see Donoho and Johnstone, 1994, 1995), soft thresholding of wavelet coefficients was proposed as a spatially adaptive method to denoise a function. In machine learning, Binev et al. addressed the regression problem with piecewise constant approximations (see Binev et al., 2005) and piecewise polynomial approximations (see Binev et al., 2007) supported on an adaptive subpartition chosen as the union of data-*independent* cells (e.g. dyadic cubes or recursively split samples). While the works above are in the context of function approximation/learning/denoising, a whole branch of geometric measure theory (following the seminal work by Jones (1990); David and Semmes (1993)) quantifies via multiscale least squares fits the rectifiability of sets and their approximability by multiple images of bi-Lipschitz maps of, say, a d-dimensional square. We can the view the current work as extending those ideas to the setting where data is random, possibly noisy, and guarantees on error on future data become one of the fundamental questions.

Theorem 8 can be viewed as a geometric counterpart of the adaptive function approximation in Binev et al. (2005, 2007). Our results are a "geometric counterpart" of sorts. We would like to point out two main differences between Theorem 8 and Theorem 3 in Binev et al. (2005): (i) In Binev et al. (2005, Theorem 3), there is an extra assumption that

the function is in $\mathcal{A}_\gamma$ with γ arbitrarily small. This assumption takes care of the error at the nodes in $\mathcal{T} \setminus \mathcal{T}^n$ where the thresholding criteria would succeed: these nodes should be added to the adaptive partition but have not been explored by our data. This assumption is removed in our Theorem 8 by observing that the nodes below the data master tree have small measure so their refinement criterion is smaller than $2^{-j}\tau_n$ with high probability. (ii) we consider scale-dependent thresholding criterion $\widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n$ unlike the criterion in Binev et al. (2005, 2007) that is scale-independent. This difference arises because at scale j our linear approximation is built on data within a ball of radius $\lesssim 2^{-j}$ and so the variance of PCA on a fixed cell at scale j is proportional to 2^{-2j}. For the same reason, we measure the complexity of $\mathcal{T}_{(\rho,\eta)}$ in terms of the weighted tree complexity instead of the cardinality since the former one gives an upper bound of the variance in piecewise linear approximation on partition via PCA (see Lemma 15). Using scale-dependent threshold and measuring tree complexity in this way give rise to the best rate of convergence. In contrast, if we use scale-independent threshold and define a model class Γ_s for whose elements $\#\mathcal{T}_{(\rho,\eta)} = \mathcal{O}(\eta^{-\frac{2d}{2s+d}})$ (analogous to the function class in Binev et al. (2005, 2007)), we can still show that $\mathcal{A}_s^\infty \subset \Gamma_s$, but the estimator only achieves MSE $\lesssim ((\log n)/n)^{\frac{2s}{2s+d}}$. However many elements[2] of Γ_s not in $\mathcal{A}_s^\infty$ are in $\mathcal{B}^{s'}$ with $\frac{2(d-2)}{2s'+d-2} = \frac{2d}{2s+d}$, and in Theorem 8 the estimator based on scaled thresholding achieves a better rate, which we believe is optimal.

We refer the reader to Maggioni et al. (2016) for a thorough discussion of further related work related to manifold and dictionary learning.

2.6. Construction of a multiscale tree decomposition

Our multiscale tree decomposition is constructed from a variation of the cover tree algorithm (see Beygelzimer et al., 2006) applied on half of the data denoted by $\mathcal{X}'_n$. In brief the cover tree $T(\mathcal{X}'_n)$ on $\mathcal{X}'_n$ is a leveled tree where each level is a "cover" for the level beneath it. Each level is indexed by j and each node in $T(\mathcal{X}'_n)$ is associated with a point in $\mathcal{X}'_n$. A point can be associated with multiple nodes in the tree but it can appear at most once at every level. Let $T_j(\mathcal{X}'_n) \subset \mathcal{X}'_n$ be the set of nodes of T at level j. The cover tree obeys the following properties for all $j = j_{\min}, \ldots, j_{\max}$:

1. Nesting: $T_j(\mathcal{X}'_n) \subset T_{j+1}(\mathcal{X}'_n)$;
2. Separation: for all distinct $p, q \in T_j(\mathcal{X}'_n)$, $\|p - q\| > 2^{-j}$;
3. Covering: for all $q \in T_{j+1}(\mathcal{X}'_n)$, there is $p \in T_j(\mathcal{X}'_n)$ such that $\|p-q\| < 2^{-j}$. The node at level j associated with p is a parent of the node at level $j+1$ associated with q.

In the third property, q is called a child of p. Each node can potentially have multiple parents satisfying the distance constraint in 3. above, but is only assigned to one of them in the tree. The properties above imply that for any $q \in \mathcal{X}'_n$, there exists $p \in T_j$ such that $\|p - q\| < 2^{-j+1}$. The authors in Beygelzimer et al. (2006) showed that cover tree always exists and that can be constructed in time $\mathcal{O}(C^d D n \log n)$.

2. For these elements, the average cell-wise refinement is monotone in the sense that for every $C_{j,k}$ and $C_{j+1,k'} \subset C_{j,k}$, we have $\Delta_{j+1,k'}/\sqrt{\rho(C_{j+1,k'})} \leq \Delta_{j,k}/\sqrt{\rho(C_{j,k})}$.

We now show that from a set of nets $\{T_j(\mathcal{X}'_n)\}_{j=j_{\min},\ldots,j_{\max}}$ as above we can construct a set of $C_{j,k}$'s with desired properties. (see Appendix A for the construction of $C_{j,k}$'s and the proof of Proposition 14). $\widetilde{\mathcal{M}}$ defined in (31) is equal to the union of the $C_{j,k}$'s up to a set with 0 empirical measure.

Proposition 14 *Assume ρ is a doubling probability measure on $\mathcal{M}$ with doubling constant C_1. Then $\{C_{j,k}\}_{k\in\mathcal{K}_j, j_{\min}\le j\le j_{\max}}$ constructed in Appendix A satisfies the Assumptions*

1. *(A1) with $a_{\max} \le C_1^2(24)^d$ and $a_{\min} = 1$.*
2. *For any $\nu > 0$,*

$$\mathbb{P}\left\{\rho(\mathcal{M}\setminus\widetilde{\mathcal{M}}) > \frac{28\nu\log n}{3n}\right\} \le 2n^{-\nu}; \tag{15}$$

3. *(A3) with $\theta_1 = C_1^{-1}4^{-d}$;*
4. *(A4) with $\theta_2 = 3$.*
5. *Additionally:*
   - 5a. *if ρ satisfies the conditions in (A5) with $B_r(z)$, $z\in\mathcal{M}$, replacing $C_{j,k}$ with constants $\tilde\theta_3, \tilde\theta_4$ such that $\lambda_d(\mathrm{Cov}(\rho_{|_{B_r(z)}})) \ge \tilde\theta_3 r^2/d$ and $\lambda_{d+1}(\mathrm{Cov}(\rho_{|_{B_r(z)}})) \le \tilde\theta_4\lambda_d(\mathrm{Cov}(\rho_{|_{B_r(z)}}))$, then the conditions in (A5) are satisfied by the $C_{j,k}$'s we construct with $\theta_3 := \tilde\theta_3(4C_1)^{-2}12^{-d}$ and $\theta_4 := \tilde\theta_4/\tilde\theta_3 12^{2d+2}C_1^4$.*
   - 5b. *if ρ is the volume measure on a closed $\mathcal{C}^s$ manifold isometrically embedded in $\mathbb{R}^D$, then the conditions in (A5) are satisfied by the $C_{j,k}$'s when j is sufficiently large.*

Even though the $\{C_{j,k}\}$ does not exactly satisfy Assumption (A2), we claim that (15) is sufficient for our performance guarantees in the case that $\mathcal{M}$ is bounded by M and $d\ge 3$, since simply approximating points on $\mathcal{M}\setminus\widetilde{\mathcal{M}}$ by 0 gives the error:

$$\mathbb{P}\left\{\int_{\mathcal{M}\setminus\widetilde{\mathcal{M}}} \|x\|^2 d\rho \ge \frac{28M^2\log n}{3n}\right\} \le 2n^{-\nu}. \tag{16}$$

The constants in Proposition 14 are extremely pessimistic, due to the generality of the assumptions on the space $\mathcal{M}$. Indeed when $\mathcal{M}$ is a nice manifold as in case (5b), the statement in the Proposition says that the constants for the $C_{j,k}$'s we construct are similar to those of the ideal $C_{j,k}$'s. In practice we use a much simpler and more efficient tree construction method and we experimentally obtain the properties above with $a_{\max} = C_1^2 4^d$ and $a_{\min} = 1$, at least for the vast majority of the points, and $\theta_{\{3,4\}} \approxeq \tilde\theta_{\{3,4\}}$. We describe this simpler construction for the multiscale partitions in Appendix A.3, together with experiments suggesting that at least in relatively simple cases one may expect $\theta_{\{3,4\}} \approxeq \tilde\theta_{\{3,4\}}$.

Besides cover trees, there are other methods that can be used in practice for the multiscale partition, such as METIS by Karypis and Kumar (1999) that is used in the numerical

examples in Chen and Maggioni (2010) and Allard et al. (2012), iterated PCA (see some analysis in Szlam (2009)) or iterated k-means. These can be computationally more efficient than cover trees, with the downside being that they may lead to partitions not satisfying our usual assumptions (in theory, and perhaps in practice).

3. Numerical experiments

We conduct numerical experiments on both synthetic and real data to demonstrate the performance of our algorithms. Given $\{x_i\}_{i=1}^n$, we split them to training data for the constructions of empirical GMRA and Adaptive GMRA and test data for the evaluation of the approximation errors:

	L^2 error	L^∞ error
Absolute error	$\left(\frac{1}{n^{\text{test}}} \sum_{x_i \in \text{test set}} \|x_i - \widehat{\mathcal{P}} x_i\|^2\right)^{\frac{1}{2}}$	$\max_{x_i \in \text{test set}} \|x_i - \widehat{\mathcal{P}} x_i\|$
Relative error	$\left(\frac{1}{n^{\text{test}}} \sum_{x_i \in \text{test set}} \|x_i - \widehat{\mathcal{P}} x_i\|^2 / \|x_i\|^2\right)^{\frac{1}{2}}$	$\max_{x_i \in \text{test set}} \|x_i - \widehat{\mathcal{P}} x_i\| / \|x_i\|$

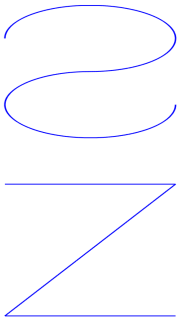
where n^{test} is the cardinality of the test set and $\widehat{\mathcal{P}}$ is the piecewise linear projection given by empirical GMRA or Adaptive GMRA. In our experiments we use absolute error for synthetic data, 3D shape and relative error for the MNIST digit data, natural image patches.

3.1. Synthetic data

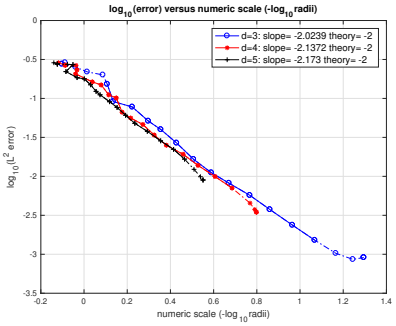
We take samples $\{x_i\}_{i=1}^n$ on the d-dim S and Z-manifolds, whose first two coordinates $x_{i,1}, x_{i,2}$ are on the S and Z curve and other coordinates $x_{i,k} \in [0,1], k = 3, 4, \ldots, d+1$. We evenly split the samples to the training set and the test set. In the noisy case, training data are corrupted by Gaussian noise: $\tilde{x}_i^{\text{train}} = x_i^{\text{train}} + \frac{\sigma}{\sqrt{D}} \xi_i, i = 1, \ldots, \frac{n}{2}$ where $\xi_i \sim \mathcal{N}(0, I_{D \times D})$, but test data are noise-free. Test data error below the noise level implies that we are denoising the data.

3.1.1. Regularity parameter s in the $\mathcal{A}_s$ and $\mathcal{B}_s$ model

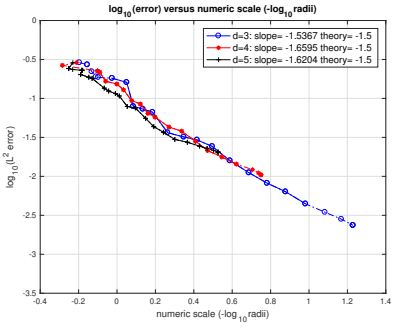
We sample 10^5 training points on the d-dim S- or Z-manifolds, for $d = 3, 4, 5$. The measure on the S-manifold is not exactly the volume measure but is comparable with the volume measure. The log-log plot of the approximation error versus scale in Figure 5 (b) shows that volume measures on the d-dim S-manifold are in $\mathcal{A}_s$ with $s \approx 2.0, 2.1, 2.2$ when $d = 3, 4, 5$, consistent with our theory which gives $s = 2$. Figure 5 (c) shows that volume measures on the d-dim Z-manifold are in $\mathcal{A}_s$ with $s \approx 1.5, 1.7, 1.6$ when $d = 3, 4, 5$, consistent with our theory which gives $s = 1.5$. The log-log plot of the approximation error versus the weighted complexity of the adaptive partition in Figure 5 (d) and (e) gives rises to an approximation of the regularity parameter s in the $\mathcal{B}_s$ model in the table.



(a) Projection of S and Z-manifolds

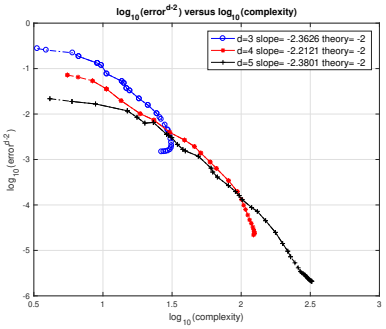


(b) S: error vs. scale

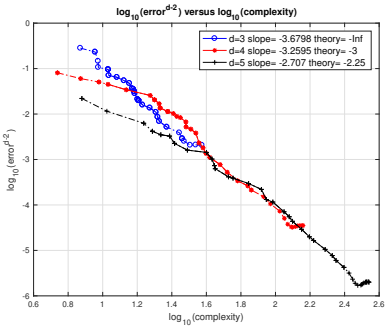


(c) Z: error vs. scale

	the S-manifold	
	theoretical s	numerical s
$d=3$	2	2.1 ± 0.2
$d=4$	2	2.1 ± 0.2
$d=5$	2	2.3 ± 0.1
	the Z-manifold	
	theoretical s	numerical s
$d=3$	$+\infty$	3.4 ± 0.3
$d=4$	3	3.1 ± 0.2
$d=5$	2.25	2.5 ± 0.2



(d) S: error vs. tree complexity

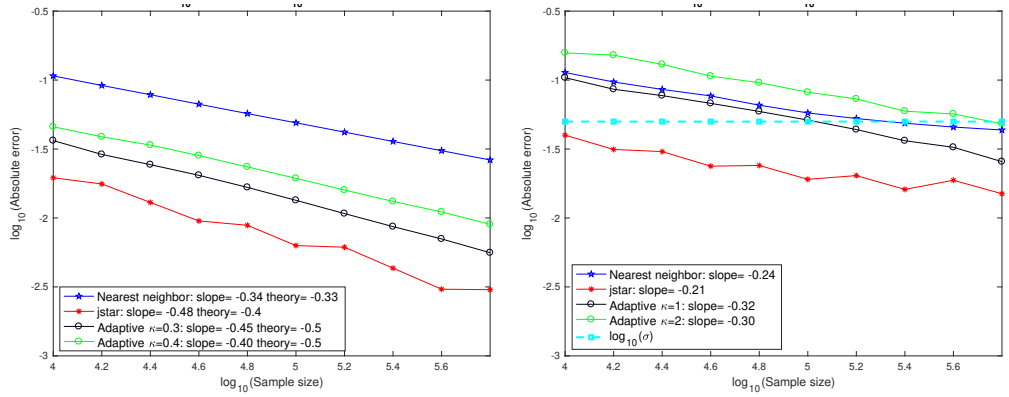


(e) Z: error vs. tree complexity

Figure 5: 10^5 training points are sampled on the d-dimensional S or Z-manifold ($d = 3, 4, 5$). In (b) and (c), we display $\log_{10} \|X - \widehat{\mathcal{P}}_j X\|_n$, versus scale j. The negative of the slope on the solid portion of the line approximates the regularity parameter s in the $\mathcal{A}_s$ model. In (d) and (e), we display the log-log plot of $\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_\eta} X\|_n^{d-2}$ versus the weighted complexity of the adaptive partition for the d-dimensional S and Z-manifold. The negative of the slope on the solid portion of the line approximates the regularity parameter s in the $\mathcal{B}_s$ model. Our five experiments give the s in the table. For the 3-dim Z-manifold, while $s = +\infty$ in the case of infinite samples, we do obtain a large s with 10^5 samples.

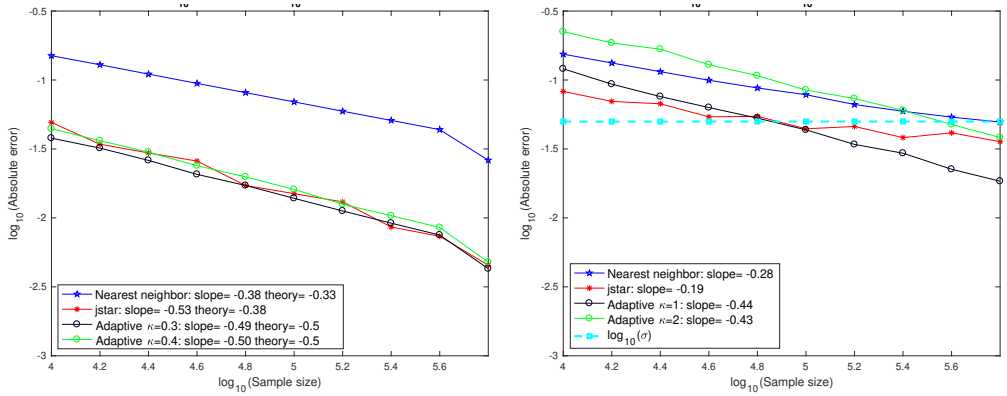
3.1.2. ERROR VERSUS SAMPLE SIZE n

We take n samples on the 3-dim S- and Z-manifolds embedded in $\mathbb{R}^{100}$ ($d = 3, D = 100$). In Figure 6, we set the noise level $\sigma = 0$ (a,c) and $\sigma = 0.05$ (b,d), display the log-log plot of the average approximation error over 10 trails with respect to the sample size n for empirical GMRA at scale j^* which is chosen as per Theorem 4: $2^{-j^*} = [(\log n)/n]^{1/(2s+d-2)}$ with $d = 3$ and $s = 2$ for the S-manifold and $s = 1.5$ for the Z-manifold. For Adaptive GMRA, the ideal κ increases as σ increases. We let $\kappa \in \{0.3, 0.4\}$ when $\sigma = 0$ and $\kappa \in \{1, 2\}$ when $\sigma = 0.05$. We also test the Nearest Neighbor (NN) approximation. The negative of the slope, determined by least squared fit, gives rise to the rate of convergence:



(a) the 3-dim S-manifold, $\sigma = 0$

(b) the 3-dim S-manifold, $\sigma = 0.05$

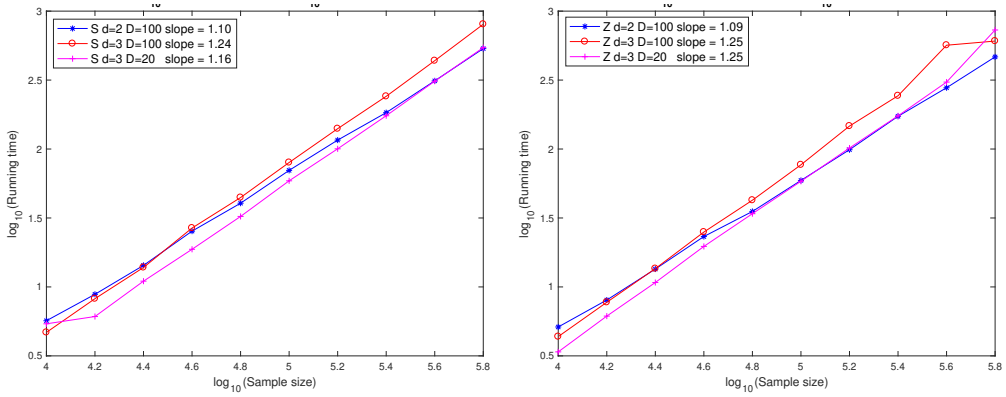


(c) the 3-dim Z-manifold, $\sigma = 0$

(d) the 3-dim Z-manifold, $\sigma = 0.05$

Figure 6: L^2 error versus the sample size n, for the 3-dim S and Z manifolds ($d = 3$), top and bottom rows respectively, of GMRA at the scale j^* chosen as per Theorem 4 (with the constant μ set, arbitrarily, equal to 1), and Adaptive GMRA with varied κ. We let $\kappa \in \{0.3, 0.4\}$ when $\sigma = 0$ (left column) and $\kappa \in \{1, 2\}$ when $\sigma = 0.05$ (right column).

L^2 error $\sim n^{\text{slope}}$. When $\sigma = 0$, the convergence rate for the nearest neighbor approximation should be $1/d = 1/3$. GMRA gives rise to a smaller error and a faster rate of convergence than the nearest neighbor approximation. When $\sigma = 0.05$, Adaptive GMRA yields a faster rate of convergence than GMRA, especially for the Z manifold. We note a de-noising effect when the approximation error falls below σ as n increases. In Adaptive GMRA, different values of κ do yield different errors up to a constant, but the rate of convergence is almost independent of κ, as predicted by Theorem 8.



(a) the S-manifold, $\sigma = 0$ (b) the Z-manifold, $\sigma = 0$

Figure 7: Average running time of GMRA in 10 experiments versus the sample size n for the S (a) and Z (b) manifolds. We set $d = 2, D = 100$ and $d = 3, D = 100$ and $d = 3, D = 20$ respectively.

3.1.3. Running time versus sample size n

The complexity of GMRA is $O(C^d Dn \log n)$. In Figure 7, we display the average running time of GMRA in 10 experiments for the S and Z manifolds when $d = 2, D = 100$ and $d = 3, D = 100$ and $d = 3, D = 20$. The running time of GMRA is almost linear in n. The running time increases as d and D increase since the complexity of GMRA is exponential in d and linear in D.

3.1.4. Robustness of GMRA and Adaptive GMRA

The robustness of the empirical GMRA and Adaptive GMRA is tested on the 3-dim S and Z-manifolds embedded in $\mathbb{R}^{100}$ while σ varies but n is fixed to be 10^5. Figure 8 shows that the average L^2 approximation error in 10 trails increases linearly with respect to σ for both uniform and Adaptive GMRA with $\kappa \in \{0.01, 0.05, 0.5\}$.

3.2. 3D shapes

We run GMRA and Adaptive GMRA on 3D points clouds on the teapot, armadillo and dragon in Figure 9. The teapot data are from the matlab toolbox and others are from the Stanford 3D Scanning Repository `http://graphics.stanford.edu/data/3Dscanrep/`.

Figure 9 shows that the adaptive partitions chosen by Adaptive GMRA matches our expectation that, at irregular locations, cells are selected at finer scales than at "flat" locations.

In Figure 10, we display the absolute L^2/L^∞ approximation error on test data versus scale and partition size. The left column shows the L^2 approximation error versus scale for GMRA and the center approximation. While the GMRA approximation is piecewise

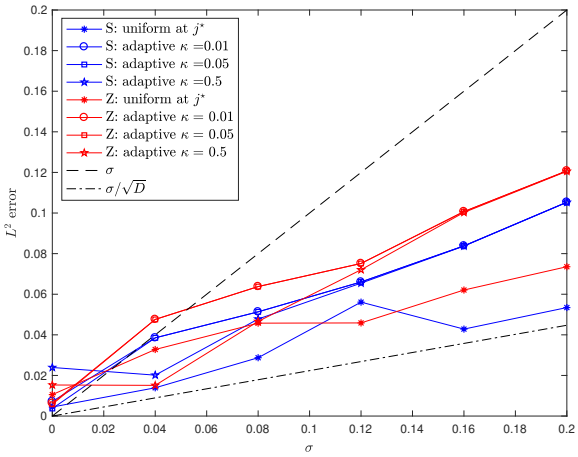


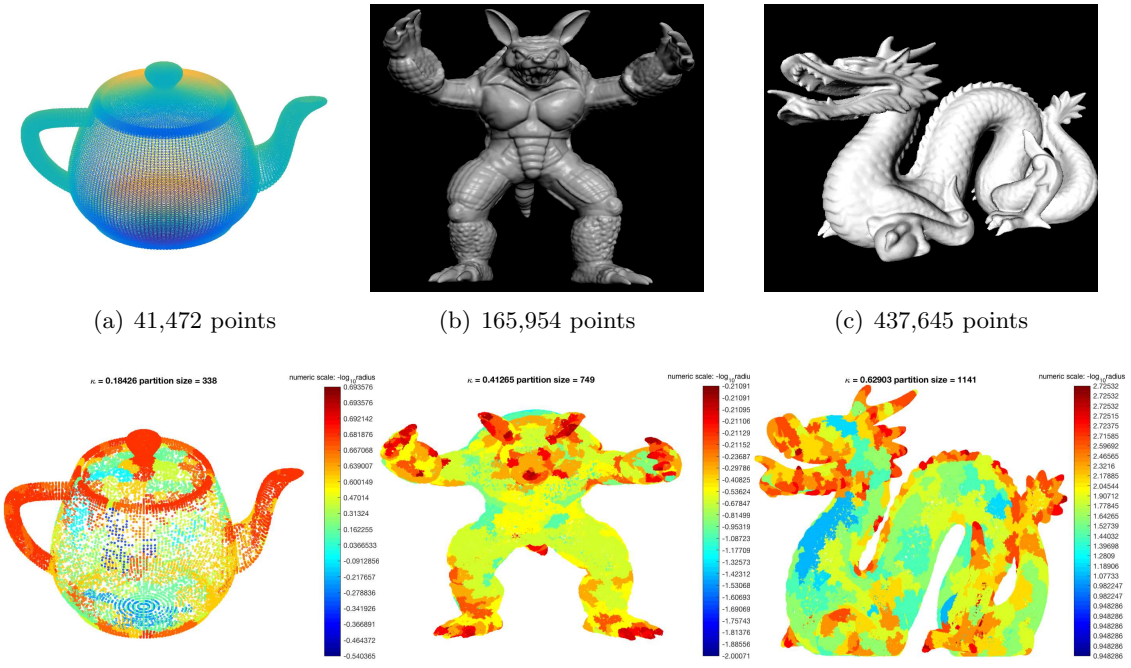
Figure 8: The average L^2 approximation error in 10 trails versus σ for GMRA and Adaptive GMRA with $\kappa \in \{0.01, 0.05, 0.5\}$ on data sampled on the 3-dim S and Z-manifolds. This shows the error of approximation grows linearly with the noise size, suggesting robustness in the construction.

linear, the center approximation is piecewise constant. Both approximation errors decay from coarse to fine scales, but GMRA yields a smaller error than the approximation by local centers. In the middle column, we run GMRA and Adaptive GMRA with the L^2 refinement criterion defined in Table 2 with scale-dependent ($\Delta_{j,k} \geq 2^{-j}\tau_n$) and scale-independent ($\Delta_{j,k} \geq \tau_n$) threshold respectively, and display the log-log plot of the L^2 approximation error versus the partition size. Overall Adaptive GMRA yields the same L^2 approximation error as GMRA with a smaller partition size, but the difference is insignificant in the armadillo and dragon, as these 3D shapes are complicated and the L^2 error simply averages the error at all locations. Then we implement Adaptive GMRA with the L^∞ refinement criterion: $\widehat{\Delta}^\infty_{j,k} = \max_{x_i \in C_{j,k}} \|\widehat{\mathcal{P}}_{j+1}x_i - \widehat{\mathcal{P}}_j x_i\|$ and display the log-log plot of the L^∞ approximation error versus the partition size in the right column. In the L^∞ error, Adaptive GMRA saves a considerable number (about half) of cells in order to achieve the same approximation error as GMRA. In this experiment, scale-independent threshold is slightly better than scale-dependent threshold in terms of saving the partition size.

3.3. MNIST digit data

We consider the MNIST data set from `http://yann.lecun.com/exdb/mnist/`, which contains images of $60,000$ handwritten digits, each of size 28×28, grayscale. The intrinsic dimension of this data set varies for different digits and across scales, as it was observed in Little et al. (2017). We run GMRA by setting the diameter of cells at scale j to be $\mathcal{O}(0.9^j)$ in order to slowly zoom into the data at multiple scales.

We evenly split the digits to the training set and the test set. As the intrinsic dimension is not well-defined, we set GMRA to pick the dimension of $\widehat{V}_{j,k}$ adaptively, as the smallest dimension needed to capture 50% of the energy of the data in $C_{j,k}$. As an example, we display the GMRA approximations of the digit $0, 1, 2$ from coarse scales to fine scales in Figure 11. The histogram of the dimensions of the subspaces $\widehat{V}_{j,k}$ is displayed in (a). (b) represents $\log_{10} \|\widehat{\mathcal{P}}_{j+1}x_i - \widehat{\mathcal{P}}_j x_i\|$ from the coarsest scale (top) to the finest scale (bottom), with columns indexed by the digits, sorted from 0 to 9. We observe that 1 has more fine



(a) 41,472 points (b) 165,954 points (c) 437,645 points

(d) $\kappa \approx 0.18$, partition size $= 338$ (e) $\kappa \approx 0.41$, partition size $= 749$ (f) $\kappa \approx 0.63$, partition size $= 1141$

Figure 9: Top line: 3D shapes; bottom line: adaptive partitions selected with refinement criterion $\widehat{\Delta}_{j,k} \geq 2^{-j}\kappa\sqrt{(\log n)/n}$. Every cell is colored by scale. In the adaptive partition, at irregular locations cells are selected at finer scales than at "flat" locations.

scale information than the other digits. In (c), we display the log-log plot of the relative L^2 error versus scale in GMRA and the center approximation. The improvement of GMRA over center approximation is noticeable. Then we compute the relative L^2 error for GMRA and Adaptive GMRA when the partition size varies. Figure 11 (d) shows that Adaptive GMRA achieves the same accuracy as GMRA with fewer cells in the partition. Errors increase when the partition size exceeds 10^3 due to a large variance at fine scales. In this experiment, scale-dependent threshold and scale-independent threshold yield similar performances.

3.4. Natural image patches

It was argued in Peyré (2009) that many sets of patches extracted from natural images can be modeled a low-dimensional manifold. We use the Caltech 101 dataset from `https://www.vision.caltech.edu/Image_Datasets/Caltech101/` (see F. Li and Perona, 2006), take 40 images from four categories: accordion, airplanes, hedgehog and scissors and extract multiscale patches of size 8×8 from these images. Specifically, if the image is of size $m \times m$,

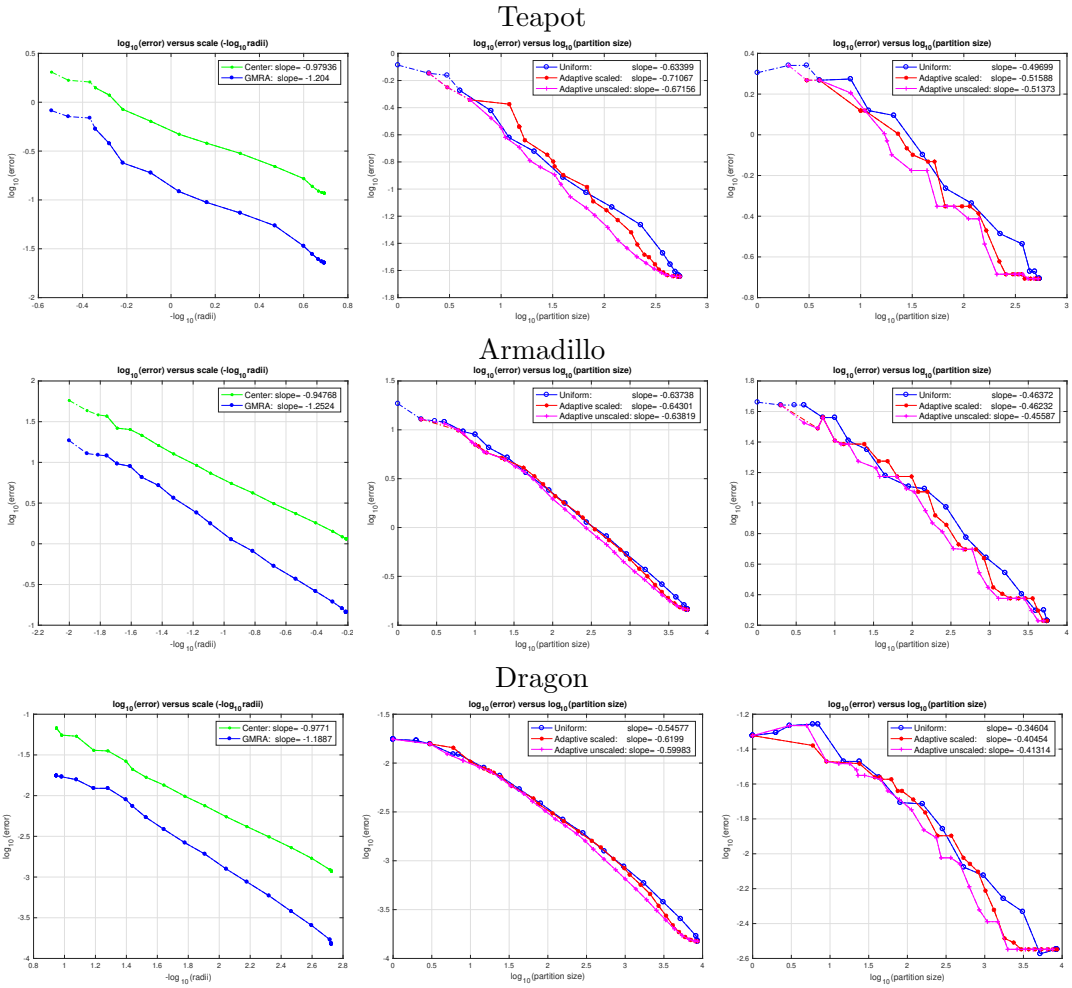
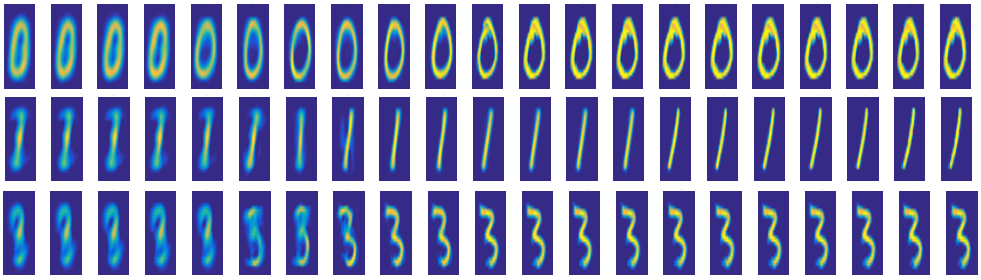
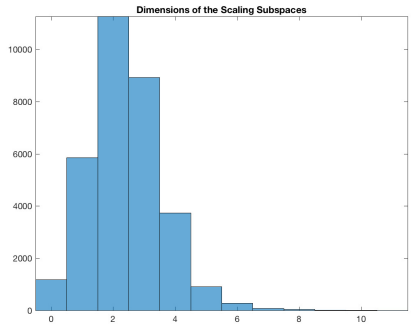


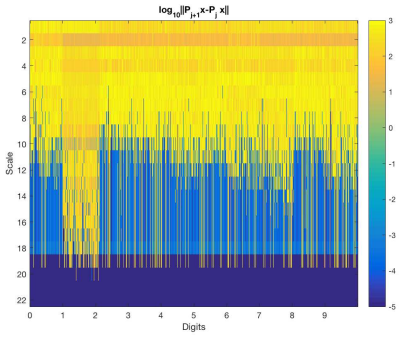
Figure 10: Left column: $\log_{10}(L^2$ error) versus scale for GMRA and center approximation; Middle column: log-log plot of the L^2 error versus partition size for GMRA and Adaptive GMRA with scale-dependent and scale-independent threshold under the L^2 refinement defined in Table 2; Right column: log-log plot of L^∞ error versus partition size for GMRA and Adaptive GMRA with scale-dependent and scale-independent threshold under the L^∞ refinement.

for $\ell = 1, \ldots, \log_2(m/8)$, we collect patches of size $2^\ell 8$, low-pass filter them and downsample them to become patches of size 8×8 (see Gerber and Maggioni (2013) for a discussion about dictionary learning on patches of multiple sizes using multiscale ideas). Then we randomly pick $200,000$ patches, evenly split them to the training set and the test set. In the construction of GMRA, we set the diameter of cells at scale j to be $\mathcal{O}(0.9^j)$ and the dimension of $\widehat{V}_{j,k}$ to be the smallest dimension needed to capture 50% of the energy of the data in $C_{j,k}$. We also run GMRA and Adaptive GMRA on the Fourier magnitudes of these image patches to take advantage of translation-invariance of the Fourier magnitudes. The

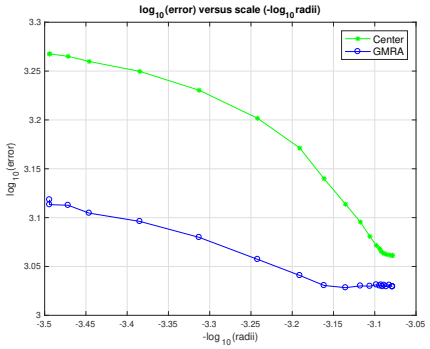




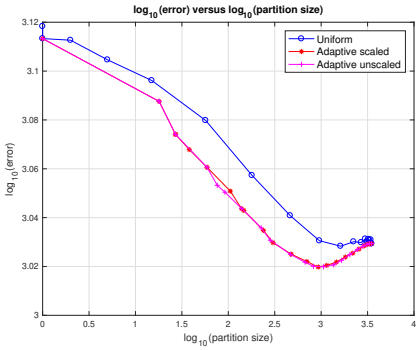
(a) Scaling subspace



(b) $\log_{10} \|\widehat{\mathcal{P}}_{j+1} x_i - \widehat{\mathcal{P}}_j x_i\|$



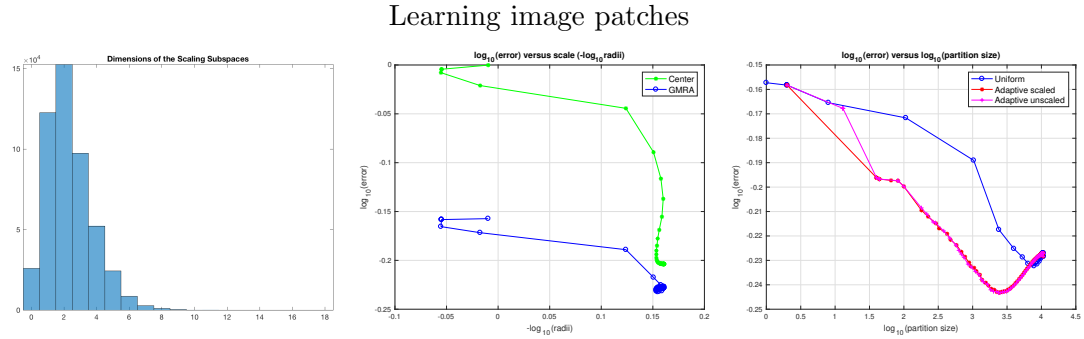
(c) Error versus scale



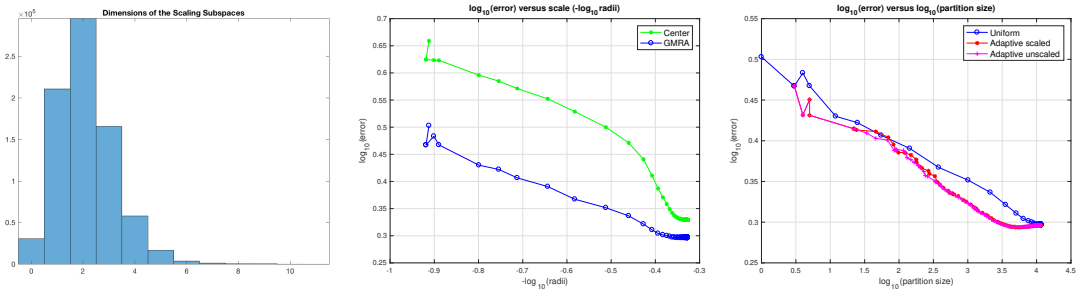
(d) Error versus partition size

Figure 11: The top three rows: multiscale approximations of the digit $0, 1, 2$ in the MNIST data set, from the coarsest scale (left) to the finest scale (right). (a) the histogram of dimensions of the subspaces $\widehat{V}_{j,k}$; (b) $\log_{10} \|\widehat{\mathcal{P}}_{j+1} x_i - \widehat{\mathcal{P}}_j x_i\|$ from the coarsest scale (top) to the finest scale (bottom), with columns indexed by the digits, sorted from 0 to 9; (c) log-log plot of the relative L^2 error versus scale in GMRA and the center approximation; (d) log-log plot of the relative L^2 error versus partition size for GMRA, Adaptive GMRA with scale-dependent and scale-independent threshold.

results are shown in Figure 13. The histograms of the dimensions of the subspaces $\widehat{V}_{j,k}$ are displayed in (a,d). Figure 13 (b) and (e) show the relative L^2 error versus scale for GMRA and the center approximation. We then compute the relative L^2 error for GMRA



(a) Scaling subspace (b) Error versus scale (c) Error versus partition size



(d) Scaling subspace (e) Error versus scale (f) Error versus partition size

Figure 12: Caltech 101 image patches

Figure 13: Top line: learning on $200,000$ image patches; bottom line: results of learning the Fourier magnitudes of the same image patches. (a,d) histograms of the dimensions of the subspaces $\widehat{V}_{j,k}$; (b,e) relative L^2 error versus scale for GMRA and the center approximation; (c,f) relative L^2 error versus the partition size for GMRA, Adaptive GMRA with scale-dependent and scale-independent threshold.

and Adaptive GMRA when the partition size varies and display the log-log plot in (c) and (f). It is noticeable that Adaptive GMRA achieves the same accuracy as GMRA with a smaller partition size. We conducted similar experiments on $200,000$ multiscale patches from CIFAR 10 from `https://www.cs.toronto.edu/~kriz/cifar.html` (see Krizhevsky and Hinton, 2009) with extremely similar results (not shown).

4. Performance analysis of GMRA and Adaptive GMRA

This section is devoted to the performance analysis of empirical GMRA and Adaptive GMRA. We will start with the following stochastic error estimate on any partition.

4.1. Stochastic error on a fixed partition

Suppose $\tilde{\mathcal{T}}$ is a finite proper subtree of the data master tree $\mathcal{T}^n$. Let Λ be the partition consisting the outer leaves of $\tilde{\mathcal{T}}$. The piecewise affine projector on Λ and its empirical version are

$$\mathcal{P}_\Lambda = \sum_{C_{j,k}\in\Lambda} \mathcal{P}_{j,k}\mathbf{1}_{j,k} \quad \text{and} \quad \widehat{\mathcal{P}}_\Lambda = \sum_{C_{j,k}\in\Lambda} \widehat{\mathcal{P}}_{j,k}\mathbf{1}_{j,k}.$$

A non-asymptotic concentration bound on the stochastic error $\|\mathcal{P}_\Lambda X - \widehat{\mathcal{P}}_\Lambda X\|$ is given by:

Lemma 15 *Let Λ be the partition associated a finite proper subtree $\tilde{\mathcal{T}}$ of the data master tree $\mathcal{T}^n$. Suppose Λ contains $\#_j\Lambda$ cells at scale j. Then for any $\eta > 0$,*

$$\mathbb{P}\{\|\mathcal{P}_\Lambda X - \widehat{\mathcal{P}}_\Lambda X\| \geq \eta\} \leq \alpha d \cdot \#\Lambda \cdot e^{-\frac{\beta n\eta^2}{d^2\sum_j 2^{-2j}\#_j\Lambda}} \tag{17}$$

$$\mathbb{E}\|\mathcal{P}_\Lambda X - \widehat{\mathcal{P}}_\Lambda X\|^2 \leq \frac{d^2\log(\alpha d\#\Lambda)\sum_j 2^{-2j}\#_j\Lambda}{\beta n}$$

where $\alpha = \alpha(\theta_2,\theta_3)$ and $\beta = \beta(\theta_2,\theta_3,\theta_4)$.

Lemma 15 and Proposition 16 below are proved in appendix C .

4.2. Performance analysis of empirical GMRA on uniform partitions

According to Eq. (1), the approximation error of empirical GMRA is split into the squared bias and the variance. A corollary of Lemma 15 with $\Lambda = \Lambda_j$ results in an estimate of the variance term.

Proposition 16 *For any $\eta \geq 0$,*

$$\mathbb{P}\{\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\| \geq \eta\} \leq \alpha d\#\Lambda_j e^{-\frac{\beta 2^{2j} n\eta^2}{d^2\#\Lambda_j}} \tag{18}$$

$$\mathbb{E}\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|^2 \leq \frac{d^2\#\Lambda_j\log[\alpha d\#\Lambda_j]}{\beta 2^{2j} n}. \tag{19}$$

In Eq. (1), the squared bias decays like $\mathcal{O}(2^{-2js})$ whenever $\rho \in \mathcal{A}_s$ and the variance scales like $\mathcal{O}(j2^{j(d-2)}/n)$. A proper choice of the scale j gives rise to Theorem 4 whose proof is given below.

Proof of Theorem 4

Proof [Proof of Theorem 4]

$$\begin{aligned}
&\mathbb{E}\|X - \widehat{\mathcal{P}}_j X\|^2 \leq 2\|X - \mathcal{P}_j X\|^2 + 2\mathbb{E}\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|^2 \\
&\leq 2|\rho|^2_{\mathcal{A}_s} 2^{-2sj} + \frac{2d^2\#\Lambda_j\log[\alpha d\#\Lambda_j]}{\beta 2^{2j} n} \leq 2|\rho|^2_{\mathcal{A}_s} 2^{-2sj} + \frac{2d^2 2^{j(d-2)}}{\theta_1\beta n}\log\frac{\alpha d 2^{jd}}{\theta_1}
\end{aligned}$$

as $\#\Lambda_j \leq 2^{jd}/\theta_1$ due to Assumption (A3).

Intrinsic dimension $d = 1$: In this case, both the squared bias and the variance decrease as j increases, so we should choose the scale j^* as large as possible as long as most cells at scale j^* have d points. We will choose j^* such that $2^{-j^*} = \mu\frac{\log n}{n}$ for some $\mu > 0$. After grouping Λ_{j^*} into light and heavy cells whose measure is below or above $\frac{28(\nu+1)\log n}{3n}$, we can show that the error on light cells is upper bounded by $C(\frac{\log n}{n})^2$ and all heavy cells have at least d points with high probability (see Lemma 17).

Lemma 17 *Suppose j^* is chosen such that $2^{-j^*} = \mu\frac{\log n}{n}$ with some $\mu > 0$. Then*

$$\|(X - \mathcal{P}_{j^*}X)\mathbf{1}_{\{C_{j^*,k}:\rho(C_{j^*,k})\le\frac{28(\nu+1)\log n}{3n}\}}\|^2 \le \frac{28(\nu+1)\theta_2^2\mu}{3\theta_1}\left(\frac{\log n}{n}\right)^2,$$
$$\mathbb{P}\left\{\textit{each } C_{j^*,k} \textit{ satisfying } \rho(C_{j^*,k}) > \tfrac{28(\nu+1)\log n}{3n} \textit{ has at least } d \textit{ points}\right\} \ge 1 - n^{-\nu}.$$

Lemma 17 is proved in appendix D. If j^* is chosen as above, The probability estimate in (5) follows from

$$\|X - \mathcal{P}_{j^*}X\| \le |\rho|_{\mathcal{A}_s}2^{-sj^*} \le |\rho|_{\mathcal{A}_s}\mu^s\left(\frac{\log n}{n}\right)^s \le |\rho|_{\mathcal{A}_s}\mu^s\frac{\log n}{n},$$

$$\mathbb{P}\left\{\|\mathcal{P}_{j^*}X - \widehat{\mathcal{P}}_{j^*}X\| \ge C_1\frac{\log n}{n}\right\} \le \tfrac{\alpha}{\theta_1\mu}\left((\log n)/n\right)^{-1}e^{-\frac{\mu\beta\theta_1C_1^2\log n}{d^2}} \le \tfrac{\alpha}{\theta_1\mu}\frac{n\,n^{-\frac{\mu\beta\theta_1C_1^2}{d^2}}}{\log n} \le C_2n^{-\nu}$$

provided that C_1 is chosen such that $\mu\beta\theta_1C_1^2/d^2 - 1 > \nu$.

Intrinsic dimension $d \ge 2$: When $d \ge 2$, the squared bias decreases but the variance increases as j gets large. We choose j^* such that $2^{-j^*} = \mu\left((\log n)/n\right)^{\frac{1}{2s+d-2}}$ to balance these two terms. We use the same technique as $d = 1$ to group Λ_{j^*} into light and heavy cells whose measure is below and above, repectively, $28/3\cdot(\nu+1)(\log n)/n$, we can show that the error on light cells is upper bounded by $C((\log n)/n)^{\frac{2s}{2s+d-2}}$ and all heavy cells have at least d points with high probability (see Lemma 18).

Lemma 18 *Let j^* be chosen such that $2^{-j^*} = \mu\left((\log n)/n\right)^{\frac{1}{2s+d-2}}$ with some $\mu > 0$. Then*

$$\|(X - \mathcal{P}_{j^*}X)\mathbf{1}_{\{C_{j^*,k}:\rho(C_{j^*,k})\le\frac{28(\nu+1)\log n}{3n}\}}\|^2 \le \frac{28(\nu+1)\theta_2^2\mu^{2-d}}{3\theta_1}\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}},$$
$$\mathbb{P}\left\{\forall\, C_{j^*,k} : \rho(C_{j^*,k}) > \frac{28(\nu+1)\log n}{3n},\ C_{j^*,k} \textit{ has at least } d \textit{ points}\right\} \ge 1 - n^{-\nu}.$$

Proof of Lemma 18 is omitted since it is the same as the proof of Lemma 17. The probability estimate in (6) follows from

$$\|X - \mathcal{P}_{j^*}X\| \le |\rho|_{\mathcal{A}_s}2^{-sj^*} \le |\rho|_{\mathcal{A}_s}\mu^s\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d-2}},$$

$$\mathbb{P}\left\{\|\mathcal{P}_{j^*}X - \widehat{\mathcal{P}}_{j^*}X\| \ge C_1\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d-2}}\right\} \le \frac{\alpha d\mu^{-d}}{\theta_1}\left(\frac{\log n}{n}\right)^{-\frac{d}{2s+d-2}}e^{-\frac{\beta\theta_1C_1^2\mu^{d-2}\log n}{d^2}} \le C_2n^{-\nu}$$

provided that $\beta\theta_1C_1^2\mu^{d-2}/d^2 - 1 > \nu$. ∎

4.3. Performance analysis of empirical GMRA on adaptive partitions

Proof [Proof of Theorem 8] In the case that $\mathcal{M}$ is bounded by M, the minimum scale $j_{\min} = \log_2 \frac{\theta_2}{M}$. We first consider the case $d \geq 3$. In our proof C stands for constants that may vary at different locations, but it is independent of n and D. We will begin by defining several objects of interest:

- $\mathcal{T}^n$: the data master tree whose leaf contains at least d points in $\mathcal{X}_n$. It can be viewed as the part of a multiscale tree that our data have explored.
- $\mathcal{T}$: a complete multiscale tree containing $\mathcal{T}^n$. $\mathcal{T}$ can be viewed as the union $\mathcal{T}^n$ and some empty cells, mostly at fine scales with high probability, that our data have not explored.
- $\mathcal{T}_{(\rho,\eta)}$: the smallest subtree of $\mathcal{T}$ which contains $\{C_{j,k} \in \mathcal{T} \,:\, \Delta_{j,k} \geq 2^{-j}\eta\}$.
- $\mathcal{T}_\eta = \mathcal{T}_{(\rho,\eta)} \cap \mathcal{T}^n$.
- $\widehat{\mathcal{T}}_\eta$: the smallest subtree of $\mathcal{T}^n$ which contains $\{C_{j,k} \in \mathcal{T}^n \,:\, \widehat{\Delta}_{j,k} \geq 2^{-j}\eta\}$.
- $\Lambda_{(\rho,\eta)}$: the partition associated with $\mathcal{T}_{(\rho,\eta)}$.
- Λ_η : the partition associated with $\mathcal{T}_\eta$.
- $\widehat{\Lambda}_\eta$: the partition associated with $\widehat{\mathcal{T}}_\eta$.
- Suppose $\mathcal{T}^0$ and $\mathcal{T}^1$ are two subtrees of $\mathcal{T}$. If Λ^0 and Λ^1 are two adaptive partitions associated with $\mathcal{T}^0$ and $\mathcal{T}^1$ respectively, we denote by $\Lambda^0 \vee \Lambda^1$ and $\Lambda^0 \wedge \Lambda^1$ the partitions associated to the trees $\mathcal{T}^0 \cup \mathcal{T}^1$ and $\mathcal{T}^0 \cap \mathcal{T}^1$ respectively.

We also let $b = 2a_{\max} + 5$ where $a_{\max}$ is the maximal number of children that a node has in $\mathcal{T}$; $\kappa_0 = \max(\kappa_1, \kappa_2)$ where $b^2\kappa_1^2/(21\theta_2^2) = \nu + 1$ and $\alpha_2\kappa_2^2/b^2 = \nu + 1$ with α_2 defined in Lemma 20. In order the obtain the MSE bound, one can simply set $\nu = 1$.

The empirical Adaptive GMRA projection is given by $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} = \sum_{C_{j,k} \in \widehat{\Lambda}_{\tau_n}} \widehat{\mathcal{P}}_{j,k}\mathbf{1}_{j,k}$. Using the triangle inequality, we split the error as follows:

$$\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X\| \leq e_1 + e_2 + e_3 + e_4$$

where

$$e_1 := \|X - \mathcal{P}_{\widehat{\Lambda}_{\tau_n} \vee \Lambda_{b\tau_n}} X\|\,, \qquad e_2 := \|\mathcal{P}_{\widehat{\Lambda}_{\tau_n} \vee \Lambda_{b\tau_n}} X - \mathcal{P}_{\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}} X\|$$
$$e_3 := \|\mathcal{P}_{\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}} X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}} X\|\,, \qquad e_4 := \|\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}} X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X\|.$$

A similar split appears in the works of Binev et al. (2005, 2007). The partition built from those $C_{j,k}$'s satisfying $\widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n$ does not exactly coincide with the partition chosen based on those $C_{j,k}$ satisfying $\Delta_{j,k} \geq 2^{-j}\tau_n$. This is accounted by e_2 and e_4, corresponding to those $C_{j,k}$'s whose $\widehat{\Delta}_{j,k}$ is significantly larger or smaller than $\Delta_{j,k}$, which we will prove to be small with high probability. The remaining terms e_1 and e_3 correspond to the bias and variance of the approximations on the partition obtained by thresholding $\Delta_{j,k}$.

Term e_1: The first term e_1 is essentially the bias term. Since $\widehat{\Lambda}_{\tau_n} \vee \Lambda_{b\tau_n} \supseteq \Lambda_{b\tau_n}$,

$$e_1^2 = \|X - \mathcal{P}_{\widehat{\Lambda}_{\tau_n} \vee \Lambda_{b\tau_n}} X\|^2 \leq \|X - \mathcal{P}_{\Lambda_{b\tau_n}} X\|^2 \leq \underbrace{\|X - \mathcal{P}_{\Lambda_{(\rho, b\tau_n)}} X\|^2}_{e_{11}^2} + \underbrace{\|\mathcal{P}_{\Lambda_{(\rho, b\tau_n)}} X - \mathcal{P}_{\Lambda_{b\tau_n}} X\|^2}_{e_{12}^2}.$$

e_{11}^2 may be upper bounded deterministically from Eq. (9):

$$e_{11}^2 \leq B_{s,d} |\rho|_{\mathcal{B}_s}^p (b\tau_n)^{2-p} \leq B_{s,d} |\rho|_{\mathcal{B}_s}^{\frac{2(d-2)}{2s+d-2}} (b\kappa)^{\frac{4s}{2s+d-2}} \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}}. \tag{20}$$

e_{12} encodes the difference between thresholding $\mathcal{T}$ and $\mathcal{T}^n$, but it is 0 with high probability:

Lemma 19 *For any $\nu > 0$, κ such that $\kappa > \kappa_1$, where $b^2\kappa_1^2/(21\theta_2^2) = \nu + 1$,*

$$\mathbb{P}\{e_{12} > 0\} \leq C(\theta_2, a_{\max}, a_{\min}, \kappa) n^{-\nu} \tag{21}$$

The proof is postponed, together with those of the Lemmata that follow, to appendix D). If $\mathcal{M}$ is bounded by M, then $e_{12}^2 \leq 4M^2$ and

$$\mathbb{E}e_{12}^2 \leq 4M^2 \mathbb{P}\{e_{12} > 0\} \leq 4M^2 C n^{-\nu} \leq 4M^2 C \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}} \tag{22}$$

if $\nu > 2s/(2s+d-2)$, for example $\nu = 1$.

Term e_3: e_3 corresponds to the variance on the partition $\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}$. For any $\eta > 0$,

$$\mathbb{P}\{e_3 > \eta\} \leq \alpha d \#(\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}) e^{-\frac{\beta n \eta^2}{d^2 \sum_{j \geq j_{\min}} 2^{-2j} \#_j(\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b})}}$$

according to Lemma 15. Since $\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b} \subset \mathcal{T}_{\tau_n/b}$, for any $j \geq 0$, regardless of $\widehat{\Lambda}_{\tau_n}$, we have $\#_j(\widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}) \leq \#_j \mathcal{T}_{\tau_n/b} \leq \# \mathcal{T}_{\tau_n/b}$. Therefore

$$\mathbb{P}\{e_3 > \eta\} \leq \alpha d \# \mathcal{T}_{\tau_n/b} e^{-\frac{\beta n \eta^2}{d^2 \sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{\tau_n/b}}} \leq \alpha d \# \mathcal{T}_{\tau_n/b} e^{-\frac{\beta n \eta^2}{d^2 |\rho|_{\mathcal{B}_s}^p (\tau_n/b)^{-p}}}, \tag{23}$$

which implies

$$\mathbb{E}e_3^2 = \int_0^{+\infty} \eta \mathbb{P}\{e_3 > \eta\} d\eta = \int_0^{+\infty} \eta \min\left(1, \alpha d \# \mathcal{T}_{\tau_n/b} e^{-\frac{\beta n \eta^2}{d^2 \sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{\tau_n/b}}}\right) d\eta$$

$$\leq \frac{d^2 \log \alpha d \# \mathcal{T}_{\tau_n/b}}{\beta n} \sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{\tau_n/b} \leq C \frac{\log n}{n} \left(\frac{\tau_n}{b}\right)^{-p} \leq C(\theta_2, \theta_3, d, \kappa, s, |\rho|_{\mathcal{B}_s}) \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}}.$$

Term e_2 and e_4: These terms account for the difference of truncating the master tree based on $\Delta_{j,k}$'s and its empirical counterparts $\widehat{\Delta}_{j,k}$'s. We prove that $\widehat{\Delta}_{j,k}$'s concentrate near $\Delta_{j,k}$'s with high probability if there are sufficient samples.

Lemma 20 *For any $\eta > 0$ and any $C_{j,k} \in \mathcal{T}$*

$$\max\left\{\mathbb{P}\left\{\widehat{\Delta}_{j,k} \leq \eta \ \text{ and } \ \Delta_{j,k} \geq b\eta\right\}, \mathbb{P}\left\{\Delta_{j,k} \leq \eta \ \text{ and } \ \widehat{\Delta}_{j,k} \geq b\eta\right\}\right\} \leq \alpha_1 e^{-\alpha_2 2^{2j} n \eta^2} \tag{24}$$

for some constants $\alpha_1 := \alpha_1(\theta_2, \theta_3, a_{\max}, d)$ and $\alpha_2 := \alpha_2(\theta_2, \theta_3, \theta_4, a_{\max}, d)$.

This Lemma enables one to show that $e_2 = 0$ and $e_4 = 0$ with high probability:

Lemma 21 *Let α_1 and α_2 be the constants in Lemma 20. For any fixed $\nu > 0$,*

$$\mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \leq \alpha_1 a_{\min} n^{-\nu} \tag{25}$$

when κ is chosen such that $\kappa > \kappa_2$, with $\alpha_2 \kappa_2^2 / b^2 = \nu + 1$.

Since $\mathcal{M}$ is bounded by M, we have $e_2^2 \leq 4M^2$ so

$$\mathbb{E}e_2^2 \leq 4M^2 \mathbb{P}\{e_2 > 0\} \leq 4M^2 \alpha_1 a_{\min} n^{-\nu} \leq 4M^2 \alpha_1 a_{\min} \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d-2}}$$

if $\nu > 2s/(2s+d-2)$, for example $\nu = 1$. The same bound holds for e_4.

Finally, we complete the probability estimate (10): let $c_0^2 = B_{s,d} |\rho|_{\mathcal{B}_s}^{\frac{2(d-2)}{2s+d-2}} (b\kappa)^{\frac{4s}{2s+d-2}}$ such that $e_{11} \leq c_0 \left((\log n)/n\right)^{\frac{s}{2s+d-2}}$. We have

$$\begin{aligned}
&\mathbb{P}\left\{\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X\| \geq c_1 \left((\log n)/n\right)^{\frac{s}{2s+d-2}}\right\} \\
&\leq \mathbb{P}\left\{e_3 > (c_1 - c_0)\left((\log n)/n\right)^{\frac{s}{2s+d-2}}\right\} + \mathbb{P}\{e_{12} > 0\} + \mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \\
&\leq \mathbb{P}\left\{e_3 > (c_1 - c_0)\left((\log n)/n\right)^{\frac{s}{2s+d-2}}\right\} + Cn^{-\nu},
\end{aligned}$$

as long as κ is chosen such that $\kappa > \max(\kappa_1, \kappa_2)$ where $b^2\kappa_1^2/(21\theta_2^2) = \nu + 1$ and $\alpha_2\kappa_2^2/b^2 = \nu + 1$ according to (21) and (25). Applying (23) gives rise to

$$\begin{aligned}
&\mathbb{P}\left\{e_3 > (c_1 - c_0)\left((\log n)/n\right)^{\frac{s}{2s+d-2}}\right\} \leq \alpha d \#\mathcal{T}_{\tau_n/b} e^{-\frac{\beta n}{|\rho|_{\mathcal{B}_s}^p (\tau_n/b)^{-p}} (c_1-c_0)^2 ((\log n)/n)^{\frac{2s}{2s+d-2}}} \\
&\leq \alpha d \#\mathcal{T}_{\tau_n/b} n^{-\frac{\beta(c_1-c_0)^2\kappa^p}{b^p |\rho|_{\mathcal{B}_s}^p}} \leq \alpha d a_{\min} n^{-\left(\frac{\beta(c_1-c_0)^2\kappa^p}{b^p|\rho|_{\mathcal{B}_s}^p} - 1\right)} \leq \alpha d a_{\min} n^{-\nu}
\end{aligned}$$

if c_1 is taken large enough such that $\frac{\beta(c_1-c_0)^2\kappa^p}{b^p|\rho|_{\mathcal{B}_s}^p} \geq \nu + 1$.

We are left with the cases $d = 1, 2$. When $d = 1$, for any distribution ρ satisfying quasi-orthogonality (8) and any $\eta > 0$, the tree complexity may be bounded as follows:

$$\sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)} \leq \sum_{j \geq j_{\min}} 2^{-2j} 2^j / \theta_1 = 2/\theta_1 2^{-j_{\min}} = 2M/(\theta_1\theta_2)\,,$$

so $\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}} X\|^2 \leq 8MB_0\eta^2/(3\theta_1\theta_2)$. Hence

$$e_{11}^2 \leq \tfrac{8MB_0}{3\theta_1\theta_2}(b\tau_n)^2 \leq \tfrac{8MB_0 b^2\kappa^2}{3\theta_1\theta_2}(\log n)/n, \quad \mathbb{P}\{e_3 > \eta\} \leq \alpha d \#\mathcal{T}_{\tau_n/b} e^{-\frac{\theta_1\theta_2\beta n\eta^2}{2Md^2}},$$

which yield $\mathbb{E}e_3^2 \leq 2Md^2 \log \alpha d \#\mathcal{T}_{\tau_n/b}/(\theta_1\theta_2\beta n) \leq C(\log n)/n$ and estimate (11).

When $d = 2$, for any distribution satisfying quasi-orthogonality and given any $\eta > 0$, we have $\sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)} \leq -|\rho|^{-1} \log \eta$, whence $\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}} X\|^2 \leq -\frac{4}{3} B_0 |\rho| \eta^2 \log \eta$. Therefore

$$e_{11}^2 \leq -\tfrac{4}{3} B_0 |\rho| (b\tau_n)^2 \log(b\tau_n) \leq C(\log^2 n)/n \quad , \quad \mathbb{P}\{e_3 > \eta\} \leq \alpha d \#\mathcal{T}_{\tau_n/b} e^{-\frac{2\beta n\eta^2}{d^2 |\rho| \log n}},$$

which yield $\mathbb{E}e_3^2 \le d^2|\rho|\log \alpha d \#\mathcal{T}_{\tau_n/b}(\log n)/(2\beta n) \le C(\log^2 n)/n$ and the probability estimate (12). ■

Proof [Proof of Theorem 10] Let $R > 0$. If we run Adaptive GMRA on $B_R(0)$, and approximate points outside $B_R(0)$ by 0, the MSE of the Adaptive GMRA in $B_R(0)$ is

$$\|(\mathbb{I} - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}})\mathbf{1}_{\{\|x\|\le R\}}X\|^2 \lesssim (|\rho|_{B_{\mathbf{0}}(R)}|^p + R^2)\left((\log n)/n\right)^{\frac{2s}{2s+d-2}} \lesssim R^{\max(\lambda,2)}\left((\log n)/n\right)^{\frac{2s}{2s+d-2}}.$$

The squared error outside $B_R(0)$ is

$$\|\mathbf{1}_{\{\|x\|\ge R\}}X\|^2 = \int_{B_R(0)^c} ||x||^2 d\rho \le CR^{-\delta}. \tag{26}$$

The total MSE is

$$\text{MSE} \lesssim R^{\max(\lambda,2)}\left((\log n)/n\right)^{\frac{2s}{2s+d-2}} + R^{-\delta}.$$

Minimizing over R suggests taking $R = R_n = \max(R_0, \mu(\log n/n)^{-\frac{2s}{(2s+d-2)(\delta+\max(2,\lambda))}})$, yielding MSE $\lesssim ((\log n)/n)^{\frac{2s}{2s+d-2}\cdot\frac{\delta}{\delta+\max(\lambda,2)}}$. The probability estimate (13) follows from Eq. (26) and Eq. (10) in Theorem 8. ■

In Remark 11, we claim that λ is not large in simple cases. If $\rho \in \mathcal{A}_s^\infty$ and ρ decays such that $\rho(C_{j,k}) \le 2^{-jd}\|c_{j,k}\|^{-(d+1+\delta)}$, we have $\Delta_{j,k} \le 2^{-js}2^{-jd/2}\|c_{j,k}\|^{-(d+1+\delta)/2}$. Roughly speaking, for any $\eta > 0$, the cells of distance r to 0 satisfying $\Delta_{j,k} \ge 2^{-j}\eta$ will satisfy $2^{-j} \ge (\eta r^{\frac{d+1+\delta}{2}})^{\frac{2}{2s+d-2}}$. In other words, the cells of distance r to 0 are truncated at scale $j_{\max}$ such that $2^{-j_{\max}} = (\eta r^{\frac{d+1+\delta}{2}})^{\frac{2}{2s+d-2}}$, which gives rise to complexity $\le 2^{-2j_{\max}}r^{d-1}2^{j_{\max}d} \le \eta^{-\frac{2(d-2)}{2s+d-2}}r^{d-1-\frac{(d+1+\delta)(d-2)}{2s+d-2}}$. If we run Adaptive GMRA with threshold η on $B_R(0)$, the weighted complexity of the truncated tree is upper bounded by $\eta^{-\frac{2(d-2)}{2s+d-2}}r^{d-\frac{(d+1+\delta)(d-2)}{2s+d-2}}$. Therefore, $\rho_{|_{B_R(0)}} \in \mathcal{B}_s$ for all $R > 0$ and $|\rho_{|_{B_R(0)}}|_{\mathcal{B}_s}^p \le R^\lambda$ with $\lambda = d - \frac{(d+1+\delta)(d-2)}{2s+d-2}$.

5. Discussions and extensions

5.1. Computational complexity

The computational cost in GMRA and Adaptive GMRA may be split as follows:

Tree construction: Cover tree itself is an online algorithm where a single-point insertion or removal takes cost at most $O(\log n)$. The total computational cost of the cover tree algorithm is $C^d Dn\log n$ where $C > 0$ is a constant (Beygelzimer et al., 2006).

Local PCA: At every scale j, we perform local PCA on the training data restricted to the $C_{j,k}$ for every $k \in \mathcal{K}_j$, using the random PCA algorithm (Halko et al., 2009). Recall that $\widehat{n}_{j,k}$ denotes the number of training points in $C_{j,k}$. The cost of local PCA at scale j is in the order of $\sum_{k\in\mathcal{K}_j} Dd\widehat{n}_{j,k} = Ddn$, and there are about $1/d\log n$ scales which gives rise to

the total cost of $Dn \log n$.

Adaptive approximation: To achieve an adaptive approximation, we need to compute the empirical geometric wavelet coefficients $\widehat{\Delta}_{j,k}$ for every $C_{j,k}$, which costs $2Dd\widehat{n}_{j,k}$ on $C_{j,k}$ and $2Dn \log n$ for the whole tree.

The computational costs of GMRA and Adaptive GMRA are summarized in Table 3.

Operations	Computational cost
Multiscale tree construction	$C^d Dn \log n$
Randomized PCA at all nodes	$Dn \log n$
Computing $\Delta_{j,k}$'s	$2Dn \log n$
GMRA	$C^d Dn \log n + Dn \log n$
Adaptive GMRA	$C^d Dn \log n + 3Dn \log n$
Compute $\mathcal{P}_j(x)$ for a test point	$\underbrace{D \log n}_{\text{find } C_{j,k} \text{ containing } x} + \underbrace{Dd}_{\text{compute } \mathcal{P}_{j,k}(x)} = D(\log n + d)$

Table 3: Computational cost

5.2. Quasi-orthogonality

A main difference between GMRA and orthonormal wavelet bases (see Daubechies, 1992; Mallat, 1998) is that $V_{j,x} \nsubseteq V_{j+1,x}$ where $(j,x) = (j,k)$ such that $x \in C_{j,k}$. Therefore the geometric wavelet subspace $\text{Proj}_{V_{j,x}{}^{\perp}} V_{j+1,x}$ which encodes the difference between $V_{j+1,x}$ and $V_{j,x}$ is in general not orthogonal across scales.

Theorem 8 involves a quasi-orthogonality condition (8), which is satisfied if the operators $\{\mathcal{Q}_{j,k}\}$ applied on $\mathcal{M}$ are rapidly decreasing in norm or are orthogonal. When $\rho \in \mathcal{A}_1^\infty$ such that $\|\mathcal{Q}_{j,k}X\| \sim 2^{-j}\sqrt{\rho(C_{j,k})}$, quasi-orthogonality is guaranteed. In this case, for any node $C_{j,k}$ and $C_{j',k'} \subset C_{j,k}$, we have $\|\mathcal{Q}_{j',k'}X\|/\sqrt{\rho(C_{j',k'})} \lesssim 2^{-(j'-j)}\|\mathcal{Q}_{j,k}X\|/\sqrt{\rho(C_{j,k})}$, which implies $\sum_{C_{j',k'} \subset C_{j,k}} \langle \mathcal{Q}_{j,k}X, \mathcal{Q}_{j',k'}X \rangle \lesssim 2\|\mathcal{Q}_{j,k}X\|^2$. Therefore $B_0 \lesssim 2$. Another setting is when $\mathcal{Q}_{j',k'}$ and $\mathcal{Q}_{j,k}$ are orthogonal whenever $C_{j',k'} \subset C_{j,k}$, as guaranteed in orthogonal GMRA in Section 5.3, in which case exact orthogonality is automatically satisfied.

Quasi-orthogonality enters in the proof of Eq. (9). If quasi-orthogonality is violated, we still have a convergence result in Theorem 8 but the convergence rate will be worse: MSE $\lesssim [(\log n)/n]^{\frac{s}{2s+d-2}}$ when $d \geq 3$ and MSE $\lesssim [(\log^d n)/n]^{\frac{1}{2}}$ when $d = 1, 2$.

5.3. Orthogonal GMRA and adaptive orthogonal GMRA

A different construction, called orthogonal geometric multi-resolution analysis in Section 5 of Allard et al. (2012), follows the classical wavelet theory by constructing a sequence

of increasing subspaces and then the corresponding wavelet subspaces exactly encode the orthogonal complement across scales. Exact orthogonality is therefore satisfied.

5.3.1. Orthogonal GMRA

In the construction, we build the sequence of subspaces $\{\widehat{S}_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$ with a coarse-to-fine algorithm in Table 4. For fixed x and j, (j,x) denotes (j,k) such that $x \in C_{j,k}$. In orthogonal GMRA the sequence of subspaces $S_{j,x}$ is increasing such that $S_{0,x} \subset S_{1,x} \subset \cdots S_{j,x} \subset S_{j+1,x} \cdots$ and the subspace $U_{j+1,x}$ exactly encodes the orthogonal complement of $S_{j,x}$ in $S_{j+1,x}$. Orthogonal GMRA with respect to the distribution ρ corresponds to affine projectors onto the subspaces $\{S_{j,k}\}_{k\in\mathcal{K}_j, j\geq j_{\min}}$.

	Orthogonal GMRA	Empirical orthogonal GMRA
Subpaces	$S_{0,x} = V_{0,x}$ $U_{1,x} = \mathrm{Proj}_{S_{0,x}^\perp} V_{1,x},\ S_{1,x} = S_{0,x} \oplus U_{1,x}$ $\cdots$ $U_{j+1,x} = \mathrm{Proj}_{S_{j,x}^\perp} V_{j+1,x}$ $S_{j+1,x} = S_{j,x} \oplus U_{j+1,x}$	$\widehat{S}_{0,x} = \widehat{V}_{0,x}$ $\widehat{U}_{1,x} = \mathrm{Proj}_{\widehat{S}_{0,x}^\perp} \widehat{V}_{1,x},\ \widehat{S}_{1,x} = \widehat{S}_{0,x} \oplus \widehat{U}_{1,x}$ $\cdots$ $\widehat{U}_{j+1,x} = \mathrm{Proj}_{\widehat{S}_{j,x}^\perp} \widehat{V}_{j+1,x}$ $\widehat{S}_{j+1,x} = \widehat{S}_{j,x} \oplus \widehat{U}_{j+1,x}$
Affine projectors	$\mathcal{S}_j := \sum_{k\in\mathcal{K}_j} \mathcal{S}_{j,k}\mathbf{1}_{j,k}$ $\mathcal{S}_{j,k}(x) := c_{j,k} + \mathrm{Proj}_{S_{j,k}}(x - c_{j,k})$	$\widehat{\mathcal{S}}_j := \sum_{k\in\mathcal{K}_j} \widehat{\mathcal{S}}_{j,k}\mathbf{1}_{j,k}$ $\widehat{\mathcal{S}}_{j,k}(x) := \widehat{c}_{j,k} + \mathrm{Proj}_{\widehat{S}_{j,k}}(x - \widehat{c}_{j,k})$

Table 4: Orthogonal GMRA

For a fixed distribution ρ, the approximation error $\|X - \mathcal{S}_j X\|$ decays as j increases. We will consider the model class $\mathcal{A}_s^{\mathrm{o}}$ where $\|X - \mathcal{S}_j X\|$ decays like $\mathcal{O}(2^{-js})$.

Definition 22 *A probability measure ρ supported on $\mathcal{M}$ is in $\mathcal{A}_s^{\mathrm{o}}$ if*

$$|\rho|_{\mathcal{A}_s^{\mathrm{o}}} = \sup_{\mathcal{T}} \inf\{A_0^o : \|X - \mathcal{S}_j X\| \leq A_0^o 2^{-js}, \forall\, j \geq j_{\min}\} < \infty\,, \tag{27}$$

where $\mathcal{T}$ varies over the set, assumed non-empty, of multiscale tree decompositions satisfying Assumptions (A1-A5).

Notice that $\mathcal{A}_s \subset \mathcal{A}_s^{\mathrm{o}}$. We split the MSE into the squared bias and the variance as: $\mathbb{E}\|X - \widehat{\mathcal{S}}_j X\|^2 = \|X - \mathcal{S}_j X\|^2 + \mathbb{E}\|\mathcal{S}_j X - \widehat{\mathcal{S}}_j X\|^2$. The squared bias $\|X - \mathcal{S}_j X\|^2 \leq |\rho|^2_{\mathcal{A}_s^{\mathrm{o}}} 2^{-2js}$ whenever $\rho \in \mathcal{A}_s^{\mathrm{o}}$. In Lemma 34 we show $\mathbb{E}\|\mathcal{S}_j X - \widehat{\mathcal{S}}_j X\|^2 \leq \frac{d^2 j^4 \#\Lambda_j \log[\alpha j \#\Lambda_j]}{\beta 2^{2j} n} = \mathcal{O}\left(\frac{j^5 2^{j(d-2)}}{n}\right)$ where α and β are the constants in Lemma 15. A proper choice of the scale yields the following result:

Theorem 23 *Assume that $\rho \in \mathcal{A}_s^o$, $s \geq 1$. Let $\nu > 0$ be arbitrary and $\mu > 0$. If j^* is properly chosen such that*

$$2^{-j^*} = \begin{cases} \mu \frac{\log n}{n} & \text{for } d = 1 \\ \mu \left(\frac{\log^5 n}{n} \right)^{\frac{1}{2s+d-2}} & \text{for } d \geq 2 \end{cases},$$

then there exists a constant $C_1(\theta_1, \theta_2, \theta_3, \theta_4, d, \nu, \mu, s)$ such that

$$\mathbb{P}\left\{ \|X - \widehat{\mathcal{S}}_{j^*} X\| \geq (|\rho|_{\mathcal{A}_s^o} \mu^s + C_1) \frac{\log^5 n}{n} \right\} \leq C_2(\theta_1, \theta_2, \theta_3, \theta_4, d, \mu) n^{-\nu} \quad \text{for } d = 1,$$

$$\mathbb{P}\left\{ \|X - \widehat{\mathcal{S}}_{j^*} X\| \geq (|\rho|_{\mathcal{A}_s^o} \mu^s + C_1) \left(\frac{\log^5 n}{n} \right)^{\frac{s}{2s+d-2}} \right\} \leq C_2(\theta_1, \theta_2, \theta_3, \theta_4, d, \mu, s) n^{-\nu} \quad \text{for } d \geq 2\,. \tag{28}$$

Theorem 23 is proved in appendix E.1.

5.3.2. Adaptive Orthogonal GMRA

Definition (infinite sample)	Empirical version
$\Delta_{j,k}^o := \|(\mathcal{S}_j - \mathcal{S}_{j+1}) \mathbf{1}_{j,k} X\|$ $= \left(\|(\mathbb{I} - \mathcal{S}_j) \mathbf{1}_{j,k} X\|^2 - \|(\mathbb{I} - \mathcal{S}_{j+1}) \mathbf{1}_{j,k} X\|^2 \right)^{\frac{1}{2}}$	$\widehat{\Delta}_{j,k}^o := \left\| (\widehat{\mathcal{S}}_j - \widehat{\mathcal{S}}_{j+1}) \mathbf{1}_{j,k} X \right\|$ $= \left(\left\| (\mathbb{I} - \widehat{\mathcal{S}}_j) \mathbf{1}_{j,k} X \right\|^2 - \left\| (\mathbb{I} - \widehat{\mathcal{S}}_{j+1}) \mathbf{1}_{j,k} X \right\|^2 \right)^{\frac{1}{2}}$

Table 5: Refinement criterion in adaptive orthogonal GMRA

Orthogonal GMRA can be constructed adaptively to the data with the refinement criterion defined in Table 5. We let $\tau_n^o := \kappa (\log^5 n / n)^{\frac{1}{2}}$ where κ is a constant, truncate the data master tree $\mathcal{T}^n$ to the smallest proper subtree that contains all $C_{j,k} \in \mathcal{T}^n$ satisfying $\widehat{\Delta}_{j,k}^o \geq 2^{-j} \tau_n^o$, denoted by $\widehat{\mathcal{T}}_{\tau_n^o}$. Empirical adaptive orthogonal GMRA returns piecewise affine projectors on the adaptive partition $\widehat{\Lambda}_{\tau_n^o}$ consisting of the outer leaves of $\widehat{\mathcal{T}}_{\tau_n^o}$. Our algorithm is summarized in Algorithm 2.

If ρ is known, given any fixed threshold $\eta > 0$, we let $\mathcal{T}_{(\rho,\eta)}$ be the smallest proper tree of $\mathcal{T}$ that contains all $C_{j,k} \in \mathcal{T}$ for which $\Delta_{j,k}^o \geq 2^{-j}\eta$. This gives rise to an adaptive partition $\Lambda_{(\rho,\eta)}$ consisting the outer leaves of $\mathcal{T}_{(\rho,\eta)}$. We introduce a model class $\mathcal{B}_s^o$ for whose elements we can control the growth rate of the truncated tree $\mathcal{T}_{(\rho,\eta)}$ as η decreases.

Definition 24 *In the case $d \geq 3$, given $s > 0$, a probability measure ρ supported on $\mathcal{M}$ is in $\mathcal{B}_s^o$ if the following quantity is finite*

$$|\rho|_{\mathcal{B}_s^o}^p := \sup_{\mathcal{T}} \sup_{\eta > 0} \eta^p \sum_{j \geq j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)} \text{ with } p = \frac{2(d-2)}{2s+d-2} \tag{29}$$

where $\mathcal{T}$ varies over the set, assumed non-empty, of multiscale tree decompositions satisfying Assumptions (A1-A5).

Algorithm 2 Empirical Adaptive Orthogonal GMRA

Input: data $\mathcal{X}_{2n} = \mathcal{X}'_n \cup \mathcal{X}_n$, intrinsic dimension d, threshold κ
Output: $\widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}}$: adaptive piecewise linear projectors

1: Construct $\mathcal{T}^n$ and $\{C_{j,k}\}$ from $\mathcal{X}'_n$
2: Compute $\widehat{\mathcal{S}}_{j,k}$ and $\widehat{\Delta}^o_{j,k}$ on every node $C_{j,k} \in \mathcal{T}^n$.
3: $\widehat{\mathcal{T}}_{\tau_n^o} \leftarrow$ smallest proper subtree of $\mathcal{T}^n$ containing all $C_{j,k} \in \mathcal{T}^n : \widehat{\Delta}^o_{j,k} \geq 2^{-j}\tau_n^o$ where $\tau_n^o = \kappa\sqrt{(\log^5 n)/n}$.
4: $\widehat{\Lambda}_{\tau_n^o} \leftarrow$ partition associated with the outer leaves of $\widehat{\mathcal{T}}_{\tau_n^o}$
5: $\widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}} \leftarrow \sum_{C_{j,k} \in \widehat{\Lambda}_{\tau_n^o}} \widehat{\mathcal{S}}_{j,k}\mathbf{1}_{j,k}$.

Notice that exact orthogonality is satisfied for orthogonal GMRA. One can show that, as long as $\rho \in \mathcal{B}^o_s$,

$$\|X - \mathcal{S}_{\Lambda_{(\rho,\eta)}} X\|^2 \leq B^o_{s,d}|\rho|^p_{\mathcal{B}^o_s}\eta^{2-p} \leq B^o_{s,d}|\rho|^2_{\mathcal{B}^o_s}\left(\sum_{j\geq j_{\min}} 2^{-2j}\#_j\mathcal{T}_{(\rho,\eta)}\right)^{-\frac{2s}{d-2}},$$

where $B^o_{s,d} := 2^p/(1-2^{p-2})$. We can prove the following performance guarantee of the empirical adaptive orthogonal GMRA (see Appendix E.2):

Theorem 25 *Suppose $\mathcal{M}$ is bounded: $\mathcal{M} \subset B_M(0)$ and the multiscale tree satisfies $\rho(C_{j,k}) \leq \theta_0 2^{-jd}$ for some $\theta_0 > 0$. Let $d \geq 3$ and $\nu > 0$. There exists $\kappa_0(\theta_0, \theta_2, \theta_3, \theta_4, a_{\max}, d, \nu)$ such that if $\rho \in \mathcal{B}^o_s$ for some $s > 0$ and $\tau_n^o = \kappa\left[(\log^5 n)/n\right]^{\frac{1}{2}}$ with $\kappa \geq \kappa_0$, then there is a c_1 and c_2 such that*

$$\mathbb{P}\left\{\|X - \widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}} X\| \geq c_1\left(\frac{\log^5 n}{n}\right)^{\frac{s}{2s+d-2}}\right\} \leq c_2 n^{-\nu}. \tag{30}$$

In Theorem 25, the constants are $c_1 := c_1(\theta_0, \theta_2, \theta_3, \theta_4, a_{\max}, d, s, \kappa, |\rho|_{\mathcal{B}^o_s}, \nu)$ and $c_2 := c_2(\theta_0, \theta_2, \theta_3, \theta_4, a_{\min}, a_{\max}, d, s, \kappa, |\rho|_{\mathcal{B}^o_s})$. Eq. (30) implies that MSE $\lesssim (\frac{\log^5 n}{n})^{\frac{2s}{2s+d-2}}$ for orthogonal Adaptive GMRA when $d \geq 3$. In the case of $d = 1, 2$, we can prove that MSE $\lesssim \frac{\log^{4+d} n}{n}$.

Acknowledgments

This research was partially funded by NSF-DMS-ATD-1222567, NSF-DMS-1708553, 1724979, 1818751, NSF-IIS-1546392 and AFOSR FA9550-14-1-0033. We are grateful to Duke University for donated computing time and equipment used for part of the numerical experiments of this paper.

Algorithm 3 Construction of a multiscale tree decomposition $\{C_{j,k}\}$

Input: data $\mathcal{X}_n'$

Output: A multiscale tree decomposition $\{C_{j,k}\}$

1: Run cover tree on $\mathcal{X}_n'$ to obtain a set of nets $\{T_j(\mathcal{X}_n')\}_{j\in[j_{\min},j_{\max}]}$

2: $j = j_{\min}$: $C_{j_{\min},0} = \widetilde{\mathcal{M}}$ defined in (31)

3: for $j = j_{\min}+1,\ldots,j_{\max}$: For every C_{j-1,k_0} at scale $j-1$, C_{j-1,k_0} has $\#\left(T_j(\mathcal{X}_n')\cap C_{j-1,k_0}\right)$ children indexed by $a_{j,k}\in T_j(\mathcal{X}_n')\cap C_{j-1,k_0}$ with corresponding $C_{j,k}$'s constructed as follows:

$$C_{j,k}^{(j)} = \widetilde{\mathcal{M}} \bigcap \text{Voronoi}(a_{j,k}, T_j(\mathcal{X}_n')\cap C_{j-1,k_0})$$

and for $i = j+1,\ldots,j_{\max}$

$$C_{j,k}^{(i)} = \Big(\bigcup_{a_{i,k'}\in C_{j,k}^{(i-1)}} B_{\frac{1}{4}2^{-i}}(a_{i,k'})\Big)\bigcup C_{j,k}^{(i-1)}$$

Finally, let $C_{j,k} = C_{j,k}^{(j_{\max})}$.

Appendix A. Tree construction, regularity of geometric spaces

A.1. Tree construction

We now show that from a set of nets $\{T_j(\mathcal{X}_n')\}_{j\in[j_{\min},j_{\max}]}$ from the cover tree algorithm we can construct a set of $C_{j,k}$ with desired properties. Similar constructions are classical in harmonic analysis (Christ, 1990). Let $\{a_{j,k}\}_{k=1}^{N(j)}$ be the set of points in $T_j(\mathcal{X}_n')$. Given a set of points $\{z_1,\ldots,z_m\}\subset\mathbb{R}^D$, the Voronoi cell of z_ℓ with respect to $\{z_1,\ldots,z_m\}$ is defined as

$$\text{Voronoi}(z_\ell,\{z_1,\ldots,z_m\}) = \{x\in\mathbb{R}^D : \|x-z_\ell\|\le\|x-z_i\| \text{ for all } i\neq\ell\}.$$

Let

$$\widetilde{\mathcal{M}} = \bigcup_{j=j_{\min}}^{j_{\max}} \bigcup_{a_{j,k}\in T_j(\mathcal{X}_n')} B_{\frac{1}{4}2^{-j}}(a_{j,k})\,. \tag{31}$$

Our $C_{j,k}$'s are constructed in Algorithm 3. These $C_{j,k}$'s form a multiscale tree decomposition of $\widetilde{\mathcal{M}}$. We will prove that $\mathcal{M}\backslash\widetilde{\mathcal{M}}$ has a negligible measure and $\{C_{j,k}\}_{k\in\mathcal{K}_j,j\in[j_{\min},j_{\max}]}$ satisfies Assumptions (A1-A5). The key is that every $C_{j,k}$ is contained in a ball of radius $3\cdot2^{-j}$ and also contains a ball of radius $2^{-j}/4$.

Lemma 26 *Every $C_{j,k}$ constructed in Algorithm 3 satisfies $B_{\frac{2^{-j}}{4}}(a_{j,k})\subseteq C_{j,k}\subseteq B_{3\cdot2^{-j}}(a_{j,k})$*

Proof For any $x\in\mathbb{R}^D$ and any set $C\in\mathbb{R}^D$, the diameter of C with respect to x is defined as $\text{diam}(C,x) := \sup_{z\in C}\|z-x\|$. First, we prove that, for every j, $C_{j,k_1}\cap C_{j,k_2}=\emptyset$ whenever $k_1\neq k_2$. Take any $a_{j+1,k_1'}\in C_{j,k_1}$ and $a_{j+1,k_2'}\in C_{j,k_2}$. Our construction guarantees that

$$\text{diam}(C_{j+1,k_1'},a_{j+1,k_1'}) \le \tfrac{1}{4}2^{-(j+1)} + \tfrac{1}{4}2^{-(j+2)} + \ldots < \tfrac{1}{2}2^{-(j+1)}$$

and similarly for $\mathrm{diam}(C_{j+1,k'_2}, a_{j+1,k'_2})$. Since $\|a_{j+1,k'_1} - a_{j+1,k'_2}\| \geq 2^{-(j+1)}$, this implies that $C_{j+1,k'_1} \cap C_{j+1,k'_2} = \emptyset$. In our construction,

$$C_{j,k_1} = \Big(\bigcup_{a_{j+1,k'_1} \in C_{j,k_1}} C_{j+1,k'_1} \Big) \bigcup B_{\frac{2^{-j}}{4}}(a_{j,k_1}),\ C_{j,k_2} = \Big(\bigcup_{a_{j+1,k'_2} \in C_{j,k_2}} C_{j+1,k'_2} \Big) \bigcup B_{\frac{2^{-j}}{4}}(a_{j,k_2}).$$

Since $\|a_{j,k_1} - a_{j,k_2}\| \geq 2^{-j}$, we observe that $B_{\frac{1}{4}2^{-j}}(a_{j,k_1}) \cap B_{\frac{1}{4}2^{-j}}(a_{j,k_2}) = \emptyset$, $C_{j+1,k'_1} \cap B_{\frac{1}{4}2^{-j}}(a_{j,k_2}) = \emptyset$ for every $a_{j+1,k'_1} \in C_{j,k_1}$, and $C_{j+1,k'_2} \cap B_{\frac{1}{4}2^{-j}}(a_{j,k_1}) = \emptyset$ for every $a_{j+1,k'_2} \in C_{j,k_2}$. Therefore $C_{j,k_1} \cap C_{j,k_2} = \emptyset$.

Our construction of $C_{j,k}$'s guarantees that every $C_{j,k}$ contains a ball of radius $\frac{1}{4} \cdot 2^{-j}$. Next we prove that every $C_{j,k}$ is contained in a ball of radius $3 \cdot 2^{-j}$. The cover tree structure guarantees that $\mathcal{X}'_n \subset \cup_{a_{j,k} \in T_j(\mathcal{X}'_n)} B_{2 \cdot 2^{-j}}(a_{j,k})$ for every j. Hence, for every $a_{j,k}$ and every $a_{j+1,k'} \in C_{j,k}$, we obtain $\|a_{j+1,k'} - a_{j,k}\| \leq 2 \cdot 2^{-j}$ and the computation above yields $\mathrm{diam}(C_{j+1,k'}, a_{j+1,k'}) \leq 2^{-j}/4$, and therefore $\mathrm{diam}(C_{j,k}, a_{j,k}) \leq 2 \cdot 2^{-j} + 2^{-j}/4 \leq 3 \cdot 2^{-j}$. In summary $C_{j,k}$ is contained in the ball of radius $3 \cdot 2^{-j}$ centered at $a_{j,k}$. ■

The following Lemma will be useful when comparing comparing covariances of sets:

Lemma 27 *If $B \subseteq A$, then we have $\lambda_d(\mathrm{cov}(\rho|_A)) \geq \frac{\rho(B)}{\rho(A)} \lambda_d(\mathrm{cov}(\rho|_B))$.*

Proof Without loss of generality, we assume both A and B are centered at x_0. Let V be the eigenspace associated with the largest d eigenvalues of $\mathrm{cov}(\rho|_B)$. Then

$$\begin{aligned}
\lambda_d(\mathrm{cov}(\rho|_A)) &= \max_{\dim U = d} \min_{u \in U} \frac{u^T \mathrm{cov}(\rho|_A) u}{u^T u} \geq \min_{v \in V} \frac{v^T \mathrm{cov}(\rho|_A) v}{v^T v} \\
&\geq \min_{v \in V} \frac{v^T \left(\int_A (x - x_0)(x - x_0)^T d\rho \right) v}{\rho(A) v^T v} \\
&= \min_{v \in V} \left(\frac{v^T \left(\int_B (x - x_0)(x - x_0)^T d\rho \right) v}{\rho(A) v^T v} + \frac{v^T \left(\int_{A \setminus B} (x - x_0)(x - x_0)^T d\rho \right) v}{\rho(A) v^T v} \right) \\
&\geq \min_{v \in V} \frac{v^T \left(\int_B (x - x_0)(x - x_0)^T d\rho \right) v}{\rho(A) v^T v} = \frac{\rho(B)}{\rho(A)} \lambda_d(\mathrm{cov}(\rho|_B)).
\end{aligned}$$

■

A.2. Regularity of geometric spaces

To fix the ideas, consider the case where $\mathcal{M}$ is a manifold of class $\mathcal{C}^s$, $s \in \mathbb{R}^+ \setminus \mathbb{Z}$, i.e. around every point x_0 there is a neighborhood U_{x_0} that is parametrized by a function $f : V \to U_{x_0}$, where V is an open connected set of $\mathbb{R}^d$, and $f \in \mathcal{C}^s$, i.e. f is $\lfloor s \rfloor$ times continuously differentiable and the $\lfloor s \rfloor$-th derivative $f^{\lfloor s \rfloor}$ is Hölder continuous of order $s - \lfloor s \rfloor$, i.e. $||f^{\lfloor s \rfloor}(x) - f^{\lfloor s \rfloor}(y)|| \leq ||f^{\lfloor s \rfloor}||_{\mathcal{C}^{s - \lfloor s \rfloor}} ||x - y||^{s - \lfloor s \rfloor}$. In particular, for $s \in (0, 1)$, f is simply a Hölder function of order s. Without loss of generality, up to a (linear) change of coordinates we may assume $x = f(x^d)$ where $x^d \in V$.

If $\mathcal{M}$ is a manifold of class $\mathcal{C}^s, s \in (0,1)$, a constant approximation of f on a set I by the value $x_0 := f(x_0^d)$ on such set yields

$$\frac{1}{\rho(I)}\int_I |f(x^d) - f(x_0^d)|^2 d\rho(x) \leq \frac{1}{\rho(I)}\int_I ||x^d - x_0^d||^{2s}||f||^2_{\mathcal{C}^s} d\rho(x) \leq ||f||^2_{\mathcal{C}^s}\mathrm{diam}(I)^{2s}$$

where we used continuity of f. If I was a ball, we would obtain a bound which would be better by a multiplicative constant no larger than $1/d$. Moreover, the left hand side is minimized by the mean $\frac{1}{\rho(I)}\int_I f(y)d\rho(y)$ of f on I, and so the bound on the right hand side holds a fortiori by replacing $f(x_0^d)$ by the mean.

Next we consider the linear approximation of $\mathcal{M}$ on $I \subset \mathcal{M}$. Suppose there exits θ_0, θ_2 such that I is contained in a ball of radius $\theta_2 r$ and contains a ball of radius $\theta_0 r$. Let $x_0 \in I$ be the closest point on I to the mean. Then I is the graph of a $\mathcal{C}^s$ function f: $P_{T_{x_0}(I)} \to P_{T^\perp_{x_0}(I)}$ where $T_{x_0}(I)$ is the plane tangent to I at x_0 and $T^\perp_{x_0}(I)$ is the orthogonal complement of $T_{x_0}(I)$. Since all the quantities involved are invariant under rotations and translations, up to a change of coordinates, we may assume $x^d = (x_1, \ldots, x_d)$ and $f = (f_1, \ldots, f_{D-d})$ where $f_i := f_i(x^d),\ i = d+1, \ldots, D$. A linear approximation of $f = (f_{d+1}, \ldots, f_D)$ based on Taylor expansion and an application of the mean value theorem yields the error estimates.

- Case 1: $s \in (1,2)$

$$\begin{aligned}
&\frac{1}{\rho(I)}\int_I \left\| f(x^d) - f(x_0^d) - \nabla f(x_0^d)\cdot(x^d - x_0^d)\right\|^2 d\rho \\
=&\sum_{i=d+1}^{D} \frac{1}{\rho(I)} \sup_{\xi_i \in \mathrm{domain}(f_i)} \int_I \left|\nabla f_i(\xi_i)(x^d - x_0^d) - \nabla f_i(x_0^d)\cdot(x^d - x_0^d)\right|^2 d\rho \\
\leq&\sum_{i=d+1}^{D} \frac{1}{\rho(I)} \sup_{\xi_i \in \mathrm{domain}(f_i)} \int_{C_{j,k}} ||x^d - x_0^d||^2 \|\xi_i - x_0^d\|^{2(s-\lfloor s\rfloor)} ||\nabla f_i||^2_{\mathcal{C}^{s-\lfloor s\rfloor}} d\rho \\
\leq& D \max_{i=1,\ldots,D-d} ||\nabla f_i||^2_{\mathcal{C}^{s-\lfloor s\rfloor}} \mathrm{diam}(I)^{2s}\,.
\end{aligned}$$

- Case 2: $s = 2$

$$\begin{aligned}
&\frac{1}{\rho(I)}\int_I \left\| f(x^d) - f(x_0^d) - \nabla f(x_0^d)\cdot(x^d - x_0^d)\right\|^2 d\rho \\
=&\sum_{i=d+1}^{D} \frac{1}{\rho(I)} \int_I \left\| f_i(x^d) - f_i(x_0^d) - \nabla f_i(x_0^d)\cdot(x^d - x_0^d)\right\|^2 d\rho \\
\leq&\sum_{i=d+1}^{D} \frac{1}{\rho(I)} \sup_{\xi_i \in \mathrm{domain}(f_i)} \int_I \left\| \frac{1}{2}(\xi_i - x_0^d)^T D^2 f_i|_{x_0^d}(\xi_i - x_0^d) + o(\|\xi_i - x_0^d\|^2)\right\|^2 d\rho \\
\leq& \frac{D}{2} \max_{i=1,\ldots,D-d} \|D^2 f_i\| \mathrm{diam}(I)^4 + o(2^{-4j}).
\end{aligned}$$

$\mathcal{M}$ does not have boundaries, so the Taylor expansion in the computations above can be performed on the convex hull of $P_{T_{x_0}(I)}$, whose diameter is no larger than $\mathrm{diam}(I)$. Note

that this bound then holds for other linear approximations which are at least as good, in $L^2(\rho|_I)$, as Taylor expansion. One such approximation is, by definition, the linear least square fit of f in $L^2(\rho|_I)$. Let L_I be the least square fit to the function $x \mapsto f(x)$. Then

$$\sum_{i=d+1}^{D} \lambda_i(\mathrm{cov}(\rho|_I))^2 = \frac{1}{\rho(I)} \int_I ||f(x) - L_I(x)||^2 d\rho(x)$$
$$\leq \begin{cases} D \max_{i=1,\ldots,D-d} ||\nabla f_i||^2_{\mathcal{C}^{s-\lfloor s \rfloor}} \mathrm{diam}(I)^{2s}, & s \in (1,2) \\ \frac{D}{2} \max_{i=1,\ldots,D-d} \|D^2 f_i\| \mathrm{diam}(I)^4, & s = 2 \end{cases} . \tag{32}$$

Proof [Proof of Proposition 14] Claim (A1) follows by a simple volume argument: $C_{j,k}$ is contained in a ball of radius $3 \cdot 2^{-j}$, and therefore has volume at most $C_1(3 \cdot 2^{-j})^d$, and each child contains a ball of radius $2^{-(j+1)}/4$, and therefore volume at least $C_1^{-1}(2^{-(j+1)}/4)^d$. It follows that $a_{\max} \leq C_1^2(3 \cdot 2^{-j}/2^{-(j+1)} \cdot 4)^d$. Clearly $a_{\min} \geq 1$ since every $a_{j,k}$ belongs to both $T_j(\mathcal{X}'_n)$ and $T_{j'}(\mathcal{X}'_n)$ with $j' \geq j$. (A1),(A3), (A4) are straightforward consequences of the doubling assumption and Lemma 26. As for (A2), for any $\nu > 0$, we have

$$\mathbb{P}\left\{\rho(\mathcal{M} \setminus \widetilde{\mathcal{M}}) > \frac{28\nu \log n}{3n}\right\} = \mathbb{P}\left\{\widehat{\rho}(\mathcal{M} \setminus \widetilde{\mathcal{M}}) = 0 \text{ and } \rho(\mathcal{M} \setminus \widetilde{\mathcal{M}}) > \frac{28\nu \log n}{3n}\right\}$$
$$\leq \mathbb{P}\left\{|\widehat{\rho}(\mathcal{M} \setminus \widetilde{\mathcal{M}}) - \rho(\mathcal{M} \setminus \widetilde{\mathcal{M}})| > \frac{1}{2}\rho(\mathcal{M} \setminus \widetilde{\mathcal{M}}) \text{ and } \rho(\mathcal{M} \setminus \widetilde{\mathcal{M}}) > \frac{28\nu \log n}{3n}\right\}$$
$$\leq 2e^{-\frac{3}{28} n \rho(\mathcal{M} \setminus \widetilde{\mathcal{M}})} \leq 2n^{-\nu}.$$

In order to prove the last statement about property (A5) in the case of 5a, observe that $B_{2^{-j}/4}(a_{j,k}) \subseteq C_{j,k} \subseteq B_{3 \cdot 2^{-j}}(a_{j,k})$. By Lemma 27 we have

$$\frac{C_1^{-1}(2^{-j}/4)^d}{\rho(C_{j,k})} \lambda_d(\mathrm{cov}(\rho|_{B_{2^{-j}/4}(a_{j,k})}) \leq \lambda_d(\mathrm{cov}(\rho|_{C_{j,k}}) \leq \frac{C_1(3 \cdot 2^{-j})^d}{\rho(C_{j,k})} \lambda_d(\mathrm{cov}(\rho|_{B_{3 \cdot 2^{-j}}(a_{j,k})})$$

and therefore $\lambda_d(\mathrm{cov}(\rho|_{C_{j,k}}) \geq C_1^{-2}(1/12)^d \lambda_d(\mathrm{cov}(\rho|_{B_{2^{-j}/4}(a_{j,k})}) \geq C_1^{-2}(1/12)^d \tilde{\theta}_3 (2^{-j}/4)^2/d$, so that (A5)-(i) holds with $\theta_3 = \tilde{\theta}_3 (4C_1)^{-2}(1/12)^d$. Proceeding similarly for λ_{d+1}, we obtain from the upper bound above that

$$\lambda_{d+1}(\mathrm{cov}(\rho|_{C_{j,k}}) \leq \frac{C_1(3 \cdot 2^{-j})^d}{C_1^{-1}(2^{-j}/4)^d} \lambda_{d+1}(\mathrm{cov}(\rho|_{B_{3 \cdot 2^{-j}}(a_{j,k})}) \leq (12^d)^2 \cdot 144 C_1^4 \tilde{\theta}_4/\tilde{\theta}_3 \lambda_d(\mathrm{cov}(\rho|_{C_{j,k}})$$

so that (A5)-(ii) holds with $\theta_4 = (12^d)^2 \cdot 144 C_1^4 \tilde{\theta}_4/\tilde{\theta}_3$.

In order to prove (A5) in the case of 5b, we use calculations as in Little et al. (2017); Maggioni et al. (2016) where one obtains that the first d eigenvalues of the covariance matrix of $\rho|_{B_r(z)}$ with $z \in \mathcal{M}$, is lower bounded by $\tilde{\theta}_3 r^2/d$ for some $\tilde{\theta}_3 > 0$. Then (A5)-(i) holds for $C_{j,k}$ with $\theta_3 = \tilde{\theta}_3 (4C_1)^{-2}(1/12)^d$. The estimate of $\lambda_{d+1}(\mathrm{cov}(\rho|_{C_{j,k}}))$ follows from (32) such that

$$\sum_{i=d+1}^{D} \lambda_i(\mathrm{cov}(\rho|_{C_{j,k}}))^2 \leq \begin{cases} D \max_{i=1,\ldots,D-d} ||\nabla f_i||^2_{\mathcal{C}^{s-\lfloor s \rfloor}} (6 \cdot 2^{-j})^{2s}, & s \in (1,2) \\ \frac{D}{2} \max_{i=1,\ldots,D-d} \|D^2 f_i\| (6 \cdot 2^{-j})^4, & s = 2 \end{cases} .$$

Therefore, there exists j_0 such that $\lambda_{d+1}(\mathrm{cov}(\rho|_{C_{j,k}})) < \theta_4 \lambda_d(\mathrm{cov}(\rho|_{C_{j,k}}))$ when $j \geq j_0$. The calculation above also implies that $\rho \in \mathcal{A}_s^\infty$ if $\max_{i=1,\ldots,D-d} ||\nabla f_i||^2_{\mathcal{C}^{s-\lfloor s \rfloor}}$ for $s \in (1,2)$ or $\max_{i=1,\ldots,D-d} \|D^2 f_i\|$ for $s = 2$ is uniformly upper bounded.

■

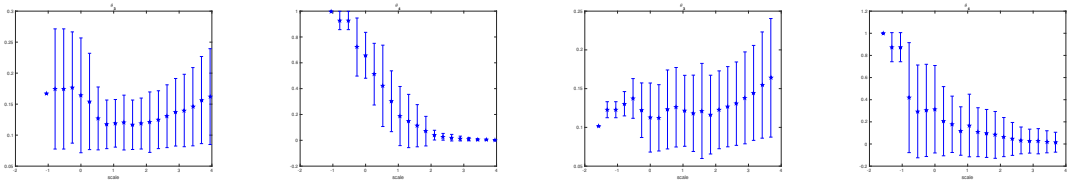
A.3. An alternative tree construction method

The $\{C_{j,k}\}$ constructed by Algorithm 3 is proved to satisfy Assumptions (A1-A5). In numerical experiments, we use a much simpler algorithm to construct $\{C_{j,k}\}$ as follows:

$$C_{j_{\max},k} = \mathrm{Voronoi}(a_{j_{\max},k}, T_{j_{\max}}(\mathcal{X}_n')) \cap B_{2^{-j_{\max}}}(a_{j_{\max},k}),$$

and for any $j < j_{\max}$, we define $C_{j,k} = \bigcup_{a_{j-1,k'} \text{ child of } a_{j,k}} C_{j-1,k'}$.

We observe that the vast majority of $C_{j,k}$'s constructed above satisfy Assumptions (A1-A5) in our numerical experiments. While it is not difficult to construct counterexamples in which the $C_{j,k}$'s thus construct fail to satisfy Assumptions (A1-A5). In Fig. 14, we will show that (A5) is satisfied when we experiment on volume measures on the 3-dim S and Z-manifold. Here we sample 10^5 training data, perform multiscale tree decomposition as stated above, and compute $\theta_3^{j,k}, \theta_4^{j,k}$ at every $C_{j,k}$. In Fig. 14, we display the mean of $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ or $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ versus scale j, with a vertical error bar representing the standard deviation of $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ or $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ at each scale. We observe that $\theta_3 = \min_{j,k} \theta_3^{j,k} \geq 0.05$ at all scales and $\theta_4 = \max_{j,k} \theta_4^{j,k} \leq 1/2$ except at very coarse scales, which demonstrates Assumption (A5) is satisfied here. Indeed θ_4 is not only bounded, but also decreases from coarse scales to fine scales.



(a) $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ versus j of the 3-dim S-manifold (b) $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ versus j of the 3-dim S-manifold (c) $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ versus j of the 3-dim Z-manifold (d) $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ versus j of the 3-dim Z-manifold

Figure 14: The mean of $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ or $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ versus scale j, with a vertical error bar representing the standard deviation of $\{\theta_3^{j,k}\}_{k \in \mathcal{K}_j}$ or $\{\theta_4^{j,k}\}_{k \in \mathcal{K}_j}$ at each scale.

We observe that, every $C_{j,k}$ constructed above is contained in a ball of radius $\theta_2 2^{-j}$ and contains a ball of radius $\theta_0 2^{-j}$, with $\theta_2/\theta_0 \in [1,2]$ for the majority of $C_{j,k}$'s. In Fig. 15, we take the volume measures on the 3-dim S and Z-manifold, and plot $\log_2$ of the outer-radius

and the statistics a lower bound for the in-radius[3] versus the scale of cover tree. Notice that the in-radius is a fraction of the outer-radius at all scales, and $\log_2 \theta_2 - \log_2 \theta_0 \leq 1$ for the majority of cells.

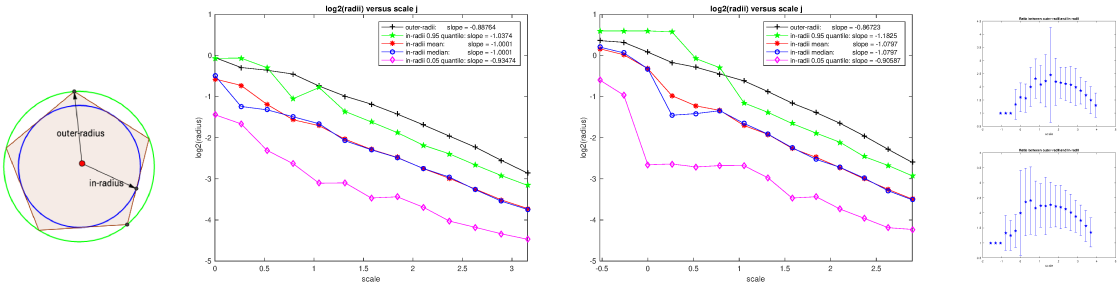


Figure 15: From left to right: the in-radius and outer-radius of a pentagon; $\log_2$ of the outer-radius and the statistics of the in-radius versus the scale of cover tree for the 3-dim S-manifold, and then the same plot for the 3-dim Z-manifold; ratio between outer-radii and in-radii, for the 3-dim S-manifold (top) and the 3-dim Z-manifold (bottom).

A.4. $\mathcal{A}_s^\infty \subset \mathcal{B}_s$

Proof [proof of Lemma 6] Assume $\rho(C_{j,k}) \asymp 2^{-jd}$. According to Definition 2, $\rho \in \mathcal{A}_s^\infty$ if $\|(X - \mathcal{P}_{j,k}X)\mathbf{1}_{j,k}\| \leq |\rho|_{\mathcal{A}_s^\infty} 2^{-js}\sqrt{\rho(C_{j,k})}$, $\forall k \in \mathcal{K}_j, j \geq j_{\min}$, which implies $\Delta_{j,k} \leq 2|\rho|_{\mathcal{A}_s^\infty} 2^{-js}\sqrt{\rho(C_{j,k})} \lesssim |\rho|_{\mathcal{A}_s^\infty} 2^{-j(s+\frac{d}{2})}$.

Let $\eta > 0$ and $\mathcal{T}_{(\rho,\eta)}$ be the smallest proper subtree of $\mathcal{T}$ that contains all $C_{j,k}$ for which $\Delta_{j,k} \geq 2^{-j}\eta$. All the nodes satisfying $\Delta_{j,k} \geq 2^{-j}\eta$ will satisfy $|\rho|_{\mathcal{A}_s^\infty} 2^{-j(s+\frac{d}{2})} \gtrsim 2^{-j}\eta$ which implies $2^{-j} \gtrsim (\eta/|\rho|_{\mathcal{A}_s^\infty})^{\frac{2}{2s+d-2}}$. Therefore, the truncated tree $\mathcal{T}_{(\rho,\eta)}$ is contained in $\mathcal{T}_{j^*} = \cup_{j \leq j^*} \Lambda_j$ with $2^{-j^*} \asymp (\eta/|\rho|_{\mathcal{A}_s^\infty})^{\frac{2}{2s+d-2}}$, so the entropy of $\mathcal{T}_{(\rho,\eta)}$ is upper bounded by the entropy of $\mathcal{T}_{j^*}$, which is $\sum_{j \leq j^*} 2^{-2j} \# \Lambda_j \asymp 2^{j^*(d-2)} \asymp (\eta/|\rho|_{\mathcal{A}_s^\infty})^{-\frac{2(d-2)}{2s+d-2}}$. Then $\rho \in \mathcal{B}_s$ and $|\rho|_{\mathcal{B}_s} \lesssim |\rho|_{\mathcal{A}_s^\infty}$ according to Definition 5. ■

Appendix B. S-manifold and Z-manifold

We consider volume measures on the d dimensional S-manifold and Z-manifold whose x_1 and x_2 coordinates are on the S curve and Z curve in Figure 5 (a) and $x_i, i = 3, \ldots, d+1$ are uniformly distributed in $[0, 1]$.

3. The in-radius of $C_{j,k}$ is approximately computed as follows: we randomly pick a center, and evaluate the largest radius with which the ball contains at least 95% points from $C_{j,k}$. This procedure is repeated for two centers, and then we pick the maximal radius as an approximation of the in-radius.

B.1. S-manifold

Since S-manifold is smooth and has a bounded curvature, the volume measure on the S-manifold is in $\mathcal{A}_2^\infty$. Therefore, the volume measure on the S-manifold is in $\mathcal{A}_2$ and $\mathcal{B}_2$ when $d \geq 3$.

B.2. Z-manifold

B.2.1. The volume on the Z-manifold is in $\mathcal{A}_{1.5}$

The uniform distribution on the d dimensional Z-manifold is in $\mathcal{A}_1$ at two corners and satisfies $\|(X - \mathcal{P}_{j,k}X)\mathbf{1}_{j,k}\| = 0$ when $C_{j,k}$ is away from the corners. There exists $A_0 > 0$ such that $\|(X - \mathcal{P}_{j,k}X)\mathbf{1}_{j,k}\| \leq A_0 2^{-j}\sqrt{\rho(C_{j,k})}$ when $C_{j,k}$ intersects with the corners. At scale j, there are about 2^{jd} cells away from the corners and there are about $2^{j(d-1)}$ cells which intersect with the corners. As a result,

$$\|X - \mathcal{P}_j X\| \leq \mathcal{O}\left(\sqrt{2^{jd} \cdot 0 \cdot 2^{-jd} + 2^{j(d-1)} \cdot 2^{-2j} \cdot 2^{-jd}}\right) = \mathcal{O}(2^{-1.5j}),$$

so the volume measure on Z-manifold is in $\mathcal{A}_{1.5}$.

B.2.2. Model class $\mathcal{B}_s$

Assume $\rho(C_{j,k}) \asymp 2^{-jd}$. We compute the regularity parameter s in the $\mathcal{B}_s$ model class when $d \geq 3$. It is easy to see that $\Delta_{j,k} = 0$ when $C_{j,k}$ is away from the corners and $\Delta_{j,k} \leq 2A_0 2^{-j}\sqrt{\rho(C_{j,k})} \lesssim 2^{-j(\frac{d}{2}+1)}$ when $C_{j,k}$ intersects with the corners. Given any fixed threshold $\eta > 0$, in the truncated tree $\mathcal{T}_{(\rho,\eta)}$, the parent of the leaves intersecting with the corners satisfy $2^{-j(\frac{d}{2}+1)} \gtrsim 2^{-j}\eta$. In other words, at the corners the tree is truncated at a scale coarser than j^* such that $2^{-j^*} = \mathcal{O}(\eta^{\frac{2}{d}})$. Since $\Delta_{j,k} = 0$ when $C_{j,k}$ is away from the corners, the entropy of $\mathcal{T}_{(\rho,\eta)}$ is dominated by the nodes intersecting with the corners whose cardinality is $2^{j(d-1)}$ at scale j. Therefore

$$\text{Entropy of } \mathcal{T}_{(\rho,\eta)} \lesssim \sum_{j \leq j^*} 2^{-2j} 2^{j(d-1)} = \mathcal{O}\left(\eta^{-\frac{2(d-3)}{d}}\right),$$

which implies that $p \leq \frac{2(d-3)}{d}$ and $s \geq \frac{3(d-2)}{2(d-3)} > 1.5$.

Then we study the relation between the error $\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}}X\|$ and the partition size $\#\Lambda_{(\rho,\eta)}$, which is numerically verified in Figure 4. Since all the nodes in $\mathcal{T}_{(\rho,\eta)}$ that intersect with corners are at a scale coarser than j^*, $\#\Lambda_{(\rho,\eta)} \approx 2^{j^*(d-1)} \asymp \eta^{-\frac{2(d-1)}{d}}$. Therefore, $\eta \lesssim [\#\Lambda_{(\rho,\eta)}]^{-\frac{d}{2(d-1)}}$ and

$$\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}}X\| \lesssim \eta^{\frac{2-p}{2}} = \eta^{\frac{2s}{2s+d-2}} \lesssim [\#\Lambda_{(\rho,\eta)}]^{-\frac{2sd}{2(d-1)(2s+d-2)}} = [\#\Lambda_{(\rho,\eta)}]^{-\frac{3}{2(d-1)}}.$$

Appendix C. Proofs of Lemma 15 and Proposition 16

C.1. Concentration inequalities

We first recall a Bernstein inequality from Tropp (2014) which is an exponential inequality to estimate the spectral norm of a sum independent random Hermitian matrices of size

$D \times D$. It features the dependence on an intrinsic dimension parameter which is usually much smaller than the ambient dimension D. For a positive-semidefinite matrix A, the intrinsic dimension is the quantity

$$\mathrm{intdim}(A) = \frac{\mathrm{trace}(A)}{\|A\|}.$$

Proposition 28 (Theorem 7.3.1 in Tropp (2014)) *Let $\xi_1, \ldots, \xi_n$ be $D \times D$ independent random Hermitian matrices that satisfy*

$$\mathbb{E}\xi_i = 0 \text{ and } \|\xi_i\| \le R, \ i = 1, \ldots, n.$$

Form the mean $\xi = \frac{1}{n}\sum_{i=1}^n \xi_i$. Suppose $\mathbb{E}(\xi^2) \preceq \Phi$. Introduce the intrinsic dimension parameter $d_{\mathrm{in}} = \mathrm{intdim}(\Phi)$. Then, for $nt \ge n\|\Phi\|^{1/2} + R/3$,

$$\mathbb{P}\{\|\xi\| \ge t\} \le 4 d_{\mathrm{in}} e^{-\frac{nt^2/2}{n\|\Phi\| + Rt/3}}.$$

We use the above inequalities to estimate the deviation of the empirical mean from the mean and the deviation of the empirical covariance matrix from the covariance matrix when the data $\mathcal{X}_{j,k} = \{x_1, \ldots, x_n\}$ (with a slight abuse of notations) are i.i.d. samples from the distribution $\rho_{|C_{j,k}}$.

Lemma 29 *Suppose $x_1, \ldots, x_n$ are i.i.d. samples from $\rho_{|C_{j,k}}$. Let the local mean and covariance, and their empirical counterparts, be defined as*

$$c_{j,k} = \int_{C_{j,k}} x d\rho_{|C_{j,k}} \quad , \quad \widehat{c}_{j,k} := \frac{1}{n}\sum_{i=1}^n x_i$$

$$\Sigma_{j,k} = \int_{C_{j,k}} (x - c_{j,k})(x - c_{j,k})^T d\rho_{|C_{j,k}} \quad , \quad \widehat{\Sigma}_{j,k} := \frac{1}{n}\sum_{i=1}^n (x_i - \widehat{c}_{j,k})(x_i - \widehat{c}_{j,k})^T$$

Then

$$\mathbb{P}\{\|\widehat{c}_{j,k} - c_{j,k}\| \ge t\} \le 8 e^{-\frac{3nt^2}{6\theta_2^2 2^{-2j} + 2\theta_2 2^{-j} t}}, \tag{33}$$

$$\mathbb{P}\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge t\} \le \left(\frac{4\theta_2^2}{\theta_3} d + 8\right) e^{-\frac{3nt^2}{24\theta_2^4 2^{-4j} + 8\theta_2^2 2^{-2j} t}}. \tag{34}$$

Proof We start by proving (33). We will apply Bernstein inequality with $\xi_i = x_i - c_{j,k} \in \mathbb{R}^D$. Clearly $\mathbb{E}\xi_i = 0$, and $\|\xi_i\| \le \theta_2 2^{-j}$ due to Assumption (A4). We form the mean $\xi = \frac{1}{n}\sum_{i=1}^n \xi_i = \widehat{c}_{j,k} - c_{j,k}$ and compute the variance

$$\sigma^2 = n^2 \|\mathbb{E}\xi^T \xi\| = \left\| \mathbb{E}\left(\sum_{i=1}^n x_i - c_{j,k}\right)^T \left(\sum_{i=1}^n x_i - c_{j,k}\right) \right\| = \left\| \sum_{i=1}^n \mathbb{E}(x_i - c_{j,k})^T (x_i - c_{j,k}) \right\| \le n\theta_2^2 2^{-2j}.$$

Then for $nt \ge \sigma + \theta_2 2^{-j}/3$,

$$\mathbb{P}\{\|\widehat{c}_{j,k} - c_{j,k}\| \ge t\} \le 8 e^{-\frac{n^2 t^2/3}{\sigma^2 + \theta_2 2^{-j} nt/3}} \le 8 e^{-\frac{3nt^2}{6\theta_2^2 2^{-2j} + 2\theta_2 2^{-j} t}}.$$

We now prove (34). Define the intermediate matrix $\bar{\Sigma}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}(x_i - c_{j,k})(x_i - c_{j,k})^T$. Since $\widehat{\Sigma}_{j,k} - \Sigma_{j,k} = \bar{\Sigma}_{j,k} - \Sigma_{j,k} - (\widehat{c}_{j,k} - c_{j,k})(\widehat{c}_{j,k} - c_{j,k})^T$, we have

$$\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \le \|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| + \|\widehat{c}_{j,k} - c_{j,k}\|^2 \le \|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| + \theta_2 2^{-j}\|\widehat{c}_{j,k} - c_{j,k}\|.$$

A sufficient condition for $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| < t$ is $\|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| < t/2$ and $\|\widehat{c}_{j,k} - c_{j,k}\| < 2^j t/(2\theta_2)$. We apply Proposition 28 to estimate $\mathbb{P}\{\|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge t/2\}$: let $\xi_i = (x_i - c_{j,k})(x_i - c_{j,k})^T - \Sigma_{j,k} \in \mathbb{R}^{D\times D}$. One can verify that $\mathbb{E}\xi_i = 0$ and $\|\xi_i\| \le 2\theta_2^2 2^{-2j}$. We form the mean $\xi = \frac{1}{n}\sum_{i=1}^{n}\xi_i = \bar{\Sigma}_{j,k} - \Sigma_{j,k}$, and then

$$\mathbb{E}\xi^2 = \mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\xi_i\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\xi_i^2 \preceq \frac{1}{n^2}\sum_{i=1}^{n}\theta_2^2 2^{-2j}\Sigma_{j,k} \preceq \frac{\theta_2^2 2^{-2j}}{n}\Sigma_{j,k},$$

which satisfies $\left\|\frac{\theta_2^2 2^{-2j}}{n}\Sigma_{j,k}\right\| \le \theta_2^4 2^{-4j}/n$. Meanwhile

$$d_{\text{in}} = \text{intdim}(\Sigma_{j,k}) = \frac{\text{trace}(\Sigma_{j,k})}{\|\Sigma_{j,k}\|} \le \frac{\theta_2^2 2^{-2j}}{\theta_3 2^{-2j}/d} = \frac{\theta_2^2}{\theta_3}d.$$

Then, Proposition 28 implies

$$\mathbb{P}\{\|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge t/2\} \le \frac{4\theta_2^2}{\theta_3}d e^{\frac{-nt^2/8}{\theta_2^4 2^{-4j} + \frac{\theta_2^2 2^{-2j}t}{3}}} = \frac{4\theta_2^2}{\theta_3}d e^{\frac{-3nt^2}{24\theta_2^4 2^{-4j} + 8\theta_2^2 2^{-2j}t}}.$$

Combining with (33), we obtain

$$\begin{aligned}\mathbb{P}\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge t\} &\le \mathbb{P}\{\|\bar{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge t/2\} + \mathbb{P}\left\{\|\widehat{c}_{j,k} - c_{j,k}\| \ge \frac{2^j t}{2\theta_2}\right\} \\ &\le \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3nt^2}{24\theta_2^4 2^{-4j} + 8\theta_2^2 2^{-2j}t}}.\end{aligned}$$

■

In Lemma 29 data are assumed to be i.i.d. samples from the conditional distribution $\rho_{|C_{j,k}}$. Given $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ which contains i.i.d. samples from ρ, we will show that the empirical measure $\widehat{\rho}(C_{j,k}) = \widehat{n}_{j,k}/n$ is close to $\rho(C_{j,k})$ with high probability.

Lemma 30 *Suppose $x_1, \ldots, x_n$ are i.i.d. samples from ρ. Let $\rho(C_{j,k}) = \int_{C_{j,k}} 1 d\rho$ and $\widehat{\rho}(C_{j,k}) = \widehat{n}_{j,k}/n$ where $\widehat{n}_{j,k}$ is the number of points in $C_{j,k}$. Then*

$$\mathbb{P}\{|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \ge t\} \le 2e^{-\frac{3nt^2}{6\rho(C_{j,k}) + 2t}} \tag{35}$$

for all $t \ge 0$. Setting $t = \frac{1}{2}\rho(C_{j,k})$ gives rise to

$$\mathbb{P}\left\{|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \ge \frac{1}{2}\rho(C_{j,k})\right\} \le 2e^{-\frac{3}{28}n\rho(C_{j,k})}. \tag{36}$$

Combining Lemma 29 and Lemma 30 gives rise to concentration bounds on $\|\widehat{c}_{j,k} - c_{j,k}\|$ and $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|$ where $c_{j,k}$, $\widehat{c}_{j,k}$, $\Sigma_{j,k}$ and $\widehat{\Sigma}_{j,k}$ are the conditional mean, empirical conditional mean, conditional covariance matrix and empirical conditional covariance matrix on $C_{j,k}$, respectively:

Lemma 31 *Suppose $x_1, \ldots, x_n$ are i.i.d. samples from ρ. Define $c_{j,k}, \Sigma_{j,k}$ and $\widehat{c}_{j,k}, \widehat{\Sigma}_{j,k}$ as in Table 1. Then given any $t > 0$,*

$$\mathbb{P}\left\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t\right\} \leq 2e^{-\frac{3}{28}n\rho(C_{j,k})} + 8e^{-\frac{3n\rho(C_{j,k})t^2}{12\theta_2^2 2^{-2j} + 4\theta_2 2^{-j}t}}, \tag{37}$$

$$\mathbb{P}\left\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \geq t\right\} \leq 2e^{-\frac{3}{28}n\rho(C_{j,k})} + \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3n\rho(C_{j,k})t^2}{96\theta_2^4 2^{-4j} + 16\theta_2^2 2^{-2j}t}}. \tag{38}$$

Proof The number of samples on $C_{j,k}$ is $\widehat{n}_{j,k} = \sum_{i=1}^n \mathbf{1}_{j,k}(x_i)$. Clearly $\mathbb{E}[\widehat{n}_{j,k}] = n\rho(C_{j,k})$. Let $\mathcal{I} \subset \{1, \ldots, n\}$ and $|\mathcal{I}| = s$. Conditionally on the event $A_{\mathcal{I}} := \{x_i \in C_{j,k} \text{ for } i \in \mathcal{I} \text{ and } x_i \notin C_{j,k} \text{ for } i \notin \mathcal{I}\}$, the random variables $\{x_i, i \in \mathcal{I}\}$ are i.i.d. samples from $\rho_{|_{C_{j,k}}}$. According to Lemma 30,

$$\begin{aligned}\mathbb{P}\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t \mid \widehat{n}_{j,k} = s\} &= \sum_{\substack{\mathcal{I}\subset\{1,\ldots,n\}\\|\mathcal{I}|=s}} \mathbb{P}\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t \mid A_{\mathcal{I}}\}\frac{1}{\binom{n}{s}} \\ &= \mathbb{P}\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t \mid A_{\{1,\ldots,s\}}\} \leq 8e^{-\frac{3st^2}{6\theta_2^2 2^{-2j} + 2\theta_2 2^{-j}t}},\end{aligned}$$

and

$$\mathbb{P}\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \geq t \mid \widehat{n}_{j,k} = s\} \leq \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3st^2}{24\theta_2^4 2^{-4j} + 8\theta_2^2 2^{-2j}t}}.$$

Furthermore $|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \leq \frac{1}{2}\rho(C_{j,k})$ yields $\widehat{n}_{j,k} \geq \frac{1}{2}n\rho(C_{j,k})$ and then

$$\mathbb{P}\left\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t \;\middle|\; |\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \leq \frac{1}{2}\rho(C_{j,k})\right\} \leq 8e^{-\frac{3n\rho(C_{j,k})t^2}{12\theta_2^2 2^{-2j} + 4\theta_2 2^{-j}t}}, \tag{39}$$

$$\mathbb{P}\left\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \geq t \;\middle|\; |\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \leq \frac{1}{2}\rho(C_{j,k})\right\} \leq \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3n\rho(C_{j,k})t^2}{48\theta_2^4 2^{-4j} + 16\theta_2^2 2^{-2j}t}}. \tag{40}$$

Eq. (39) (40) along with Lemma 30 gives rise to

$$\mathbb{P}\left\{\|\widehat{c}_{j,k} - c_{j,k}\| \geq t\right\} \leq 2e^{-\frac{3}{28}n\rho(C_{j,k})} + 8e^{-\frac{3n\rho(C_{j,k})t^2}{12\theta_2^2 2^{-2j} + 4\theta_2 2^{-j}t}},$$

$$\mathbb{P}\left\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \geq t\right\} \leq 2e^{-\frac{3}{28}n\rho(C_{j,k})} + \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3n\rho(C_{j,k})t^2}{48\theta_2^4 2^{-4j} + 16\theta_2^2 2^{-2j}t}}.$$

■

Given $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|$, we can estimate the angle between the eigenspace of $\widehat{\Sigma}_{j,k}$ and $\Sigma_{j,k}$ with the following proposition.

Proposition 32 (Davis and Kahan (1970) or Theorem 3 in Zwald and Blanchard (2006)) *Let* $\delta_d(\Sigma_{j,k}) = \frac{1}{2}(\lambda_d^{j,k} - \lambda_{d+1}^{j,k})$. *If* $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \le \frac{1}{2}\delta_d(\Sigma_{j,k})$, *then*

$$\left\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\right\| \le \frac{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|}{\delta_d(\Sigma_{j,k})}.$$

According to Assumption (A4) and (A5), $\delta_d(\Sigma_{j,k}) \ge \theta_3 2^{-2j}/(4d)$. An application of Proposition 32 yields

$$\begin{aligned}\mathbb{P}\left\{\left\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\right\| \ge t\right\} &\le \mathbb{P}\left\{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| \ge \frac{\theta_3(1-\theta_4)t}{2d2^{2j}}\right\} \\ &\le 2e^{-\frac{3}{28}n\rho(C_{j,k})} + \left(\frac{4\theta_2^2}{\theta_3}d + 8\right)e^{-\frac{3\theta_3^2(1-\theta_4)^2 n\rho(C_{j,k})t^2}{384\theta_2^4 d^2 + 32\theta_2^2\theta_3(1-\theta_4)dt}}. \end{aligned} \tag{41}$$

Proof [Proof of Lemma 15] Since

$$\|\mathcal{P}_\Lambda X - \widehat{\mathcal{P}}_\Lambda X\|^2 = \sum_{C_{j,k}\in\Lambda}\int_{C_{j,k}} \|\mathcal{P}_{j,k}x - \widehat{\mathcal{P}}_{j,k}x\|^2 d\rho = \sum_j \sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|\mathcal{P}_{j,k}x - \widehat{\mathcal{P}}_{j,k}x\|^2 d\rho,$$

we obtain the estimate

$$\mathbb{P}\left\{\|\mathcal{P}_\Lambda X - \widehat{\mathcal{P}}_\Lambda X\| \ge \eta\right\} \le \sum_j \mathbb{P}\left\{\sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|\mathcal{P}_{j,k}x - \widehat{\mathcal{P}}_{j,k}x\|^2 d\rho \ge \frac{2^{-2j}\#_j\Lambda\eta^2}{\sum_{j\ge j_{\min}} 2^{-2j}\#_j\Lambda}\right\}. \tag{42}$$

Next we prove that, for any fixed scale j,

$$\mathbb{P}\left\{\sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|\mathcal{P}_{j,k}x - \widehat{\mathcal{P}}_{j,k}x\|^2 d\rho \ge t^2\right\} \le \alpha\#_j\Lambda e^{-\frac{\beta 2^{2j} n t^2}{\#_j\Lambda}}. \tag{43}$$

Then Lemma 15 is proved by setting $t^2 = 2^{-2j}\#_j\Lambda\eta^2/(\sum_{j\ge 0} 2^{-2j}\#_j\Lambda)$.

The proof of (43) starts with the following calculation:

$$\begin{aligned}&\sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|\mathcal{P}_{j,k}x - \widehat{\mathcal{P}}_{j,k}x\|^2 d\rho \\ &= \sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|c_{j,k} + \mathrm{Proj}_{V_{j,k}}(x - c_{j,k}) - \widehat{c}_{j,k} - \mathrm{Proj}_{\widehat{V}_{j,k}}(x - \widehat{c}_{j,k})\|^2 d\rho \\ &\le \sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \|(\mathbb{I} - \mathrm{Proj}_{\widehat{V}_{j,k}})(c_{j,k} - \widehat{c}_{j,k}) + (\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}})(x - c_{j,k})\|^2 d\rho \\ &\le 2\sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \left[\|c_{j,k} - \widehat{c}_{j,k}\|^2 + \|(\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}})(x - c_{j,k})\|^2\right] d\rho \\ &\le 2\sum_{k:C_{j,k}\in\Lambda}\int_{C_{j,k}} \left[\|c_{j,k} - \widehat{c}_{j,k}\|^2 + \theta_2^2 2^{-2j}\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|^2\right] d\rho \end{aligned}$$

For any fixed j and given $t > 0$, we divide Λ into light cells $\Lambda_{j,t}^-$ and heavy cells $\Lambda_{j,t}^+$, where

$$\Lambda_{j,t}^- := \left\{ C_{j,k} \in \Lambda : \rho(C_{j,k}) \le \frac{t^2}{20\theta_2^2 2^{-2j} \#_j \Lambda} \right\} \text{ and } \Lambda_{j,t}^+ := \Lambda \setminus \Lambda_{j,t}^- .$$

Since $\int_{C_{j,k}} \left[\|c_{j,k} - \widehat{c}_{j,k}\|^2 + \theta_2^2 2^{-2j} \|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|^2 \right] d\rho \le 5\theta_2^2 2^{-2j} \rho(C_{j,k})$, for light sets we have

$$2 \sum_{k: C_{j,k} \in \Lambda_{j,t}^-} \int_{C_{j,k}} \left[\|c_{j,k} - \widehat{c}_{j,k}\|^2 + \theta_2^2 2^{-2j} \|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|^2 \right] d\rho \le \frac{t^2}{2}. \tag{44}$$

Next we consider $C_{j,k} \in \Lambda_{j,t}^+$. We have

$$\begin{aligned} &\mathbb{P}\left\{ \|\widehat{c}_{j,k} - c_{j,k}\| \ge \frac{t}{\sqrt{8\rho(C_{j,k})\#_j\Lambda}} \right\} \\ &\le 2\exp\left(-\frac{3}{28} n\rho(C_{j,k}) \right) + 8e^{-\frac{\frac{3n\rho(C_{j,k})\frac{t^2}{8\rho(C_{j,k})\#_j\Lambda}}{12\theta_2^2 2^{-2j} + 4\theta_2 2^{-j} \frac{t}{\sqrt{8\rho(C_{j,k})\#_j\Lambda}}}} \le C_1 e^{-C_2 \frac{2^{2j} n t^2}{\#_j \Lambda}}, \end{aligned} \tag{45}$$

and

$$\begin{aligned} &\mathbb{P}\left\{ \|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\| \ge \frac{2^j t}{\theta_2 \sqrt{8\rho(C_{j,k})\#_j\Lambda}} \right\} \\ &\le 2e^{-\frac{3}{28} n\rho(C_{j,k})} + \left(\frac{4\theta_2^2}{\theta_3} d + 8 \right) e^{-\frac{\frac{3\theta_3^2(1-\theta_4)^2 n\rho(C_{j,k}) \frac{2^{2j}t^2}{8\theta_2^2\rho(C_{j,k})\#_j\Lambda}}{384\theta_2^4 d^2 + 32\theta_2^2\theta_3(1-\theta_4)d \frac{2^j t}{\theta_2\sqrt{8\rho(C_{j,k})\#_j\Lambda}}}} \le C_3 d e^{-C_4 \frac{2^{2j} n t^2}{d^2 \#_j \Lambda}} \end{aligned} \tag{46}$$

where positive constants C_1, C_2, C_3, C_4 depend on θ_2 and θ_3. Combining (44), (45) and (46) gives rise to (43) with $\alpha = \max(C_1, C_3)$ and $\beta = \min(C_2, C_4)$.

■

Proof [Proof of Proposition 16] The bound (18) follows directly from Lemma 15 applied to $\Lambda = \Lambda_j$; (19) follows from (18) by integrating the probability over η:

$$\begin{aligned} &\mathbb{E}\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|^2 = \int_0^{+\infty} \eta \mathbb{P}\left\{ \|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\| \ge \eta \right\} d\eta \\ &\le \int_0^{+\infty} \eta \min\left\{ 1, \alpha d \#\Lambda_j e^{-\frac{\beta 2^{2j} n \eta^2}{d^2 \#\Lambda_j}} \right\} d\eta = \int_0^{\eta_0} \eta d\eta + \int_{\eta_0}^{+\infty} \alpha d\eta \#\Lambda_j e^{-\frac{\beta 2^{2j} n \eta^2}{d^2 \#\Lambda_j}} d\eta \end{aligned}$$

where $\alpha d \#\Lambda_j e^{-\frac{\beta 2^{2j} n \eta_0^2}{d^2 \#\Lambda_j}} = 1$. Then

$$\mathbb{E}\|\mathcal{P}_j X - \widehat{\mathcal{P}}_j X\|^2 = \frac{1}{2}\eta_0^2 + \frac{\alpha}{2\beta} \cdot \frac{\#\Lambda_j^2}{2^{2j} n} e^{-\beta \frac{2^{2j} n \eta_0^2}{\#\Lambda_j}} \le \frac{d^2 \#\Lambda_j \log[\alpha d \#\Lambda_j]}{\beta 2^{2j} n}.$$

■

Appendix D. Proof of Eq. (9), Lemma 17, 19, 20, 21

Proof [Proof of Eq. (9)] Let $\Lambda_{(\rho,\eta)}^{+0} = \Lambda_{(\rho,\eta)}$ and $\Lambda_{(\rho,\eta)}^{+n}$ be the partition consisting of the children of $\Lambda_{(\rho,\eta)}^{+(n-1)}$ for $n = 1, 2, \ldots$. Then

$$\begin{aligned}
\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}} X\| &= \|\sum_{\ell=0}^{n-1} (\mathcal{P}_{\Lambda_{(\rho,\eta)}^{+\ell}} X - \mathcal{P}_{\Lambda_{(\rho,\eta)}^{+(\ell+1)}} X) + \mathcal{P}_{\Lambda^{+n}(\rho,\eta)} X - X\| \\
&= \|\sum_{\ell=0}^{\infty} (\mathcal{P}_{\Lambda_{(\rho,\eta)}^{+\ell}} X - \mathcal{P}_{\Lambda_{(\rho,\eta)}^{+(\ell+1)}} X) + \lim_{n\to\infty} \mathcal{P}_{\Lambda_{(\rho,\eta)}^{+n}} X - X\| \\
&\le \|\sum_{C_{j,k} \notin \mathcal{T}_{(\rho,\eta)}} \mathcal{Q}_{j,k} X\| + \|\lim_{n\to\infty} \mathcal{P}_{\Lambda_{(\rho,\eta)}^{+n}} X - X\|.
\end{aligned}$$

We have $\|\lim_{n\to\infty} \mathcal{P}_{\Lambda_{(\rho,\eta)}^{+n}} X - X\| = 0$ due to Assumption (A4). Therefore,

$$\begin{aligned}
&\|X - \mathcal{P}_{\Lambda_{(\rho,\eta)}} X\|^2 \le \|\sum_{C_{j,k} \notin \mathcal{T}_{(\rho,\eta)}} \mathcal{Q}_{j,k} X\|^2 \le \sum_{C_{j,k} \notin \mathcal{T}_{(\rho,\eta)}} B_0 \|\mathcal{Q}_{j,k} X\|^2 = B_0 \sum_{C_{j,k} \notin \mathcal{T}_{(\rho,\eta)}} \Delta_{j,k}^2 \\
&\le B_0 \sum_{\ell \ge 0} \sum_{C_{j,k} \in \mathcal{T}_{(\rho,2^{-(\ell+1)}\eta)} \setminus \mathcal{T}_{(\rho,2^{-\ell}\eta)}} \Delta_{j,k}^2 \le B_0 \sum_{\ell \ge 0} \sum_{j \ge j_{\min}} (2^{-j} 2^{-\ell} \eta)^2 \#_j \mathcal{T}_{(\rho,2^{-(\ell+1)}\eta)} \\
&\le B_0 \sum_{\ell \ge 0} 2^{-2\ell} \eta^2 \sum_{j \ge j_{\min}} 2^{-2j} \#_j \mathcal{T}_{(\rho,2^{-(\ell+1)}\eta)} \le B_0 \sum_{\ell \ge 0} 2^{-2\ell} \eta^2 |\rho|_{\mathcal{B}_s}^p [2^{-(\ell+1)}\eta]^{-p} \\
&\le B_0 2^p \left(\sum_{\ell \ge 0} 2^{-\ell(2-p)} \right) |\rho|_{\mathcal{B}_s}^p \eta^{2-p} \le B_{s,d} |\rho|_{\mathcal{B}_s}^p \eta^{2-p}.
\end{aligned}$$

■

Proof [of Lemma 17]

$$\begin{aligned}
&\left\| (X - \mathcal{P}_{j^*} X) \mathbf{1}_{\left\{ C_{j^*,k} : \rho(C_{j^*,k}) \le \frac{28(\nu+1)\log n}{3n} \right\}} \right\|^2 \le \sum_{\left\{ C_{j^*,k} : \rho(C_{j^*,k}) \le \frac{28(\nu+1)\log n}{3n} \right\}} \int_{C_{j^*,k}} \|x - \mathcal{P}_{j^*,k}\|^2 d\rho \\
&\le \#\left\{ C_{j^*,k} : \rho(C_{j^*,k}) \le \frac{28(\nu+1)\log n}{3n} \right\} \theta_2^2 2^{-2j^*} \frac{28(\nu+1)\log n}{3n} \\
&\le \tfrac{28(\nu+1)\theta_2^2}{3\theta_1} 2^{j^*(d-2)} (\log n)/n \le \tfrac{28(\nu+1)\theta_2^2 \mu}{3\theta_1} \left((\log n)/n \right)^2.
\end{aligned}$$

For every $C_{j^*,k}$, we have

$$\begin{aligned}
&\mathbb{P}\left\{ \rho(C_{j^*,k}) > \tfrac{28}{3}(\nu+1)(\log n)/n \text{ and } \widehat{\rho}(C_{j^*,k}) < d/n \right\} \\
&\le \mathbb{P}\left\{ |\widehat{\rho}(C_{j^*,k}) - \rho(C_{j^*,k})| > \rho(C_{j^*,k})/2 \text{ and } \rho(C_{j^*,k}) > \tfrac{28}{3}(\nu+1)(\log n)/n \right\} \\
&\qquad \text{for } n \text{ so large that } 14(\nu+1)\log n > 3d \\
&\le 2e^{-\frac{3}{28} n \rho(C_{j^*,k})} \le 2n^{-\nu-1}.
\end{aligned}$$

Then

$$\mathbb{P}\left\{\text{each } C_{j^*,k} \text{ satisfying } \rho(C_{j^*,k}) > \tfrac{28}{3}(\nu+1)(\log n)/n \text{ has at most } d \text{ points}\right\}$$
$$\leq \#\left\{C_{j^*,k} : \rho(C_{j^*,k}) < \tfrac{28}{3}(\nu+1)(\log n)/n\right\} 2n^{-\nu-1} \leq \#\Lambda_{j^*} 2n^{-\nu-1} \leq 2n^{-\nu}/(\theta_1 \mu \log n) < n^{-\nu},$$

when n is so large that $\theta_1 \mu \log n > 2$. ■

Proof [of Lemma 19] Since $\mathcal{T}_{b\tau_n} \subset \mathcal{T}_{(\rho,b\tau_n)}$, $\mathbb{P}\{e_{12} > 0\}$ if and only if there exists $C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)} \setminus \mathcal{T}_{b\tau_n}$. In other words, $\mathbb{P}\{e_{12} > 0\}$ if and only if there exists $C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)}$ such that $\widehat{\rho}(C_{j,k}) < d/n$ and $\Delta_{j,k} > 2^{-j} b\tau_n$. Therefore,

$$\begin{aligned}
\mathbb{P}\{e_{12} > 0\} &\leq \sum_{C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)}} \mathbb{P}\{\widehat{\rho}(C_{j,k}) < d/n \text{ and } \Delta_{j,k} > 2^{-j} b\tau_n\} \\
&\leq \sum_{C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)}} \mathbb{P}\left\{\widehat{\rho}(C_{j,k}) < d/n \text{ and } \rho(C_{j,k}) > \tfrac{4b^2\tau_n^2}{9\theta_2^2}\right\} \quad \left(\text{since } \Delta_{j,k} \leq \tfrac{3}{2}\theta_2 2^{-j}\sqrt{\rho(C_{j,k})}\right) \\
&\leq \sum_{C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)}} \mathbb{P}\left\{|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| > \rho(C_{j,k})/2 \text{ and } \rho(C_{j,k}) > \tfrac{4b^2\tau_n^2}{9\theta_2^2}\right\} \\
&\qquad \text{(for } n \text{ large enough so that } 2b^2\kappa^2 \log n > 9\theta_2^2 d) \\
&\leq \sum_{C_{j,k} \in \mathcal{T}_{(\rho,b\tau_n)}} 2e^{-\frac{3}{28} n \cdot \frac{4b^2\kappa^2 \log n}{9\theta_2^2 n}} \leq 2n^{-\frac{b^2\kappa^2}{21\theta_2^2}} \#\mathcal{T}_{(\rho,b\tau_n)}.
\end{aligned}$$

The leaves of $\mathcal{T}_{(\rho,b\tau_n)}$ satisfy $\rho(C_{j,k}) > 4b^2\tau_n^2/(9\theta_2^2)$. Since $\rho(\mathcal{M}) = 1$, there are at most $9\theta_2^2/(4b^2\tau_n^2)$ leaves in $\mathcal{T}_{(\rho,b\tau_n)}$. Meanwhile, since every node in $\mathcal{T}$ has at least $a_{\min}$ children, $\#\mathcal{T}_{(\rho,b\tau_n)} \leq 9\theta_2^2 a_{\min}/(4b^2\tau_n^2)$. Then for a fixed but arbitrary $\nu > 0$,

$$\mathbb{P}\{e_{12} > 0\} \leq \tfrac{18\theta_2^2 a_{\min}}{4b^2\tau_n^2} n^{-\frac{b^2\kappa^2}{21\theta_2^2}} \leq \tfrac{18\theta_2^2 a_{\min}}{4b^2\kappa^2} n^{1-\frac{b^2\kappa^2}{21\theta_2^2}} \leq C(\theta_2, a_{\max}, a_{\min}, \kappa) n^{-\nu},$$

if κ is chosen such that $\kappa > \kappa_1$ where $b^2\kappa_1^2/(21\theta_2^2) = \nu + 1$. ■

Proof [of Lemma 20] We first prove (24). Introduce the intermediate variable

$$\bar{\Delta}_{j,k} := \|\mathcal{Q}_{j,k}\|_n = \|(\mathcal{P}_j - \mathcal{P}_{j+1})\mathbf{1}_{j,k} X\|_n$$

and then observe that

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\Delta}_{j,k} \leq \eta \text{ and } \Delta_{j,k} \geq b\eta\right\} &\leq \mathbb{P}\left\{\widehat{\Delta}_{j,k} \leq \eta \text{ and } \bar{\Delta}_{j,k} \geq (a_{\max}+2)\eta\right\} \\
&\quad + \mathbb{P}\left\{\bar{\Delta}_{j,k} \leq (a_{\max}+2)\eta \text{ and } \Delta_{j,k} \geq (2a_{\max}+5)\eta\right\}. \quad (47)
\end{aligned}$$

The bound in Eq. (24) is proved in the following three steps. In Step One, we show that $\Delta_{j,k} \geq b\eta$ implies $\rho(C_{j,k}) \geq \mathcal{O}(2^{2j}\eta^2)$. Then we estimate $\mathbb{P}\left\{\widehat{\Delta}_{j,k} \leq \eta \text{ and } \bar{\Delta}_{j,k} \geq (a_{\max}+2)\eta\right\}$ in Step Two and $\mathbb{P}\left\{\bar{\Delta}_{j,k} \leq (a_{\max}+2)\eta \text{ and } \bar{\Delta}_{j,k} \geq (2a_{\max}+5)\eta\right\}$ in Step Three.

Step One: Notice that $\Delta_{j,k} \leq \frac{3}{2}\theta_2 2^{-j}\sqrt{\rho(C_{j,k})}$. As a result, $\Delta_{j,k} \geq b\eta$ implies

$$\rho(C_{j,k}) \geq \frac{4b^2 2^{2j}\eta^2}{9\theta_2^2}. \tag{48}$$

Step Two:

$$\mathbb{P}\left\{\widehat{\Delta}_{j,k} \leq \eta \ \text{ and } \ \bar{\Delta}_{j,k} \geq (a_{\max}+2)\eta\right\} \leq \mathbb{P}\left\{|\widehat{\Delta}_{j,k} - \bar{\Delta}_{j,k}| \geq (a_{\max}+1)\eta\right\}. \tag{49}$$

We can write

$$\begin{aligned}
|\widehat{\Delta}_{j,k} - \bar{\Delta}_{j,k}| &\leq \left\|(\mathcal{P}_{j,k} - \widehat{\mathcal{P}}_{j,k})\mathbf{1}_{j,k}X\right\|_n + \sum_{C_{j+1,k'}\in\mathscr{C}(C_{j,k})}\left\|(\widehat{\mathcal{P}}_{j+1,k'} - \mathcal{P}_{j+1,k'})\mathbf{1}_{j+1,k'}X\right\|_n \\
&\leq \underbrace{\left(\|c_{j,k} - \widehat{c}_{j,k}\| + \theta_2 2^{-j}\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|\right)\sqrt{\widehat{\rho}(C_{j,k})}}_{e_1} \\
&\quad + \underbrace{\sum_{C_{j+1,k'}\in\mathscr{C}(C_{j,k})}\left(\|c_{j+1,k'} - \widehat{c}_{j+1,k'}\| + \theta_2 2^{-(j+1)}\|\mathrm{Proj}_{V_{j+1,k'}} - \mathrm{Proj}_{\widehat{V}_{j+1,k'}}\|\right)\sqrt{\widehat{\rho}(C_{j+1,k'})}}_{e_2}.
\end{aligned} \tag{50}$$

Term e_1: We will estimate $\mathbb{P}\{e_1 > \eta\}$. Conditional on the event that $\{|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \leq \frac{1}{2}\rho(C_{j,k})\}$, we have $e_1 \leq \frac{3}{2}\left(\|c_{j,k} - \widehat{c}_{j,k}\| + \theta_2 2^{-j}\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|\right)\sqrt{\rho(C_{j,k})}$. A similar argument to the proof of Lemma 15 along with (48) give rise to

$$\mathbb{P}\left\{\frac{3}{2}\left(\|c_{j,k} - \widehat{c}_{j,k}\| + \theta_2 2^{-j}\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|\right)\sqrt{\rho(C_{j,k})} > \eta\right\} \leq \tilde{\gamma}_1 e^{-\tilde{\gamma}_2 2^{2j} n\eta^2}$$

where $\tilde{\gamma}_1 := \tilde{\gamma}_1(\theta_2, \theta_3, d)$ and $\tilde{\gamma}_2 := \tilde{\gamma}_2(\theta_2, \theta_3, \theta_4, d)$; otherwise $\mathbb{P}\left\{|\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| > \frac{1}{2}\rho(C_{j,k})\right\} \leq 2e^{-\frac{3}{28}n\rho(C_{j,k})} \leq 2e^{-\frac{b^2 2^{2j} n\eta^2}{21\theta_2^2}}$. Therefore

$$\mathbb{P}\{e_1 > \eta\} \leq \max(\tilde{\gamma}_1, 2)e^{-\min(\tilde{\gamma}_2, \frac{b^2}{21\theta_2^2})2^{2j}n\eta^2} \tag{51}$$

Term e_2: We will estimate $\mathbb{P}\{e_2 > a_{\max}\eta\}$. Let $\Lambda^- = \left\{C_{j+1,k'} \in \mathscr{C}(C_{j,k}) : \rho(C_{j+1,k'}) \leq \frac{2^{2j}\eta^2}{8\theta_2^2}\right\}$ and $\Lambda^+ = \mathscr{C}(C_{j,k}) \setminus \Lambda^-$. For every $C_{j+1,k'} \in \Lambda^-$, when we condition on the event that $\left\{\rho(C_{j+1,k'}) \leq \frac{2^{2j}\eta^2}{8\theta_2^2} \text{ and } \widehat{\rho}(C_{j+1,k'}) \leq \frac{2^{2j}\eta^2}{4\theta_2^2}\right\}$, we obtain

$$\begin{aligned}
&\sum_{C_{j+1,k'}\in\Lambda^-}\left(\|c_{j+1,k'} - \widehat{c}_{j+1,k'}\| + \theta_2 2^{-(j+1)}\|\mathrm{Proj}_{V_{j+1,k'}} - \mathrm{Proj}_{\widehat{V}_{j+1,k'}}\|\right)\sqrt{\widehat{\rho}(C_{j+1,k'})} \\
&\leq \sum_{C_{j+1,k'}\in\Lambda^-}\theta_2 2^{-j}\sqrt{\widehat{\rho}(C_{j,k})} \leq a_{\max}\eta/2;
\end{aligned} \tag{52}$$

otherwise,

$$\begin{aligned}
&\mathbb{P}\left\{\rho(C_{j+1,k'}) \le \tfrac{2^{2j}\eta^2}{8\theta_2^2} \text{ and } \widehat{\rho}(C_{j+1,k'}) > \tfrac{2^{2j}\eta^2}{4\theta_2^2}\right\} \\
&\le \mathbb{P}\left\{\rho(C_{j+1,k'}) \le \tfrac{2^{2j}\eta^2}{8\theta_2^2} \text{ and } |\widehat{\rho}(C_{j+1,k'}) - \rho(C_{j+1,k'})| \ge \tfrac{2^{2j}\eta^2}{8\theta_2^2}\right\} \\
&\le 2e^{-\left(3n\left(\frac{2^{2j}\eta^2}{8\theta_2^2}\right)^2\right)/\left(6\rho(C_{j+1,k'})+2\frac{2^{2j}\eta^2}{8\theta_2^2}\right)} \le 2e^{-\frac{3\cdot 2^{2j}n\eta^2}{64\theta_2^2}}. \qquad (53)
\end{aligned}$$

For $C_{j+1,k'} \in \Lambda^+$, a similar argument to e_1 gives rise to

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{C_{j+1,k'}\in\Lambda^+}\left(\|c_{j+1,k'} - \widehat{c}_{j+1,k'}\| + \theta_2 2^{-(j+1)}\|\mathrm{Proj}_{V_{j+1,k'}} - \mathrm{Proj}_{\widehat{V}_{j+1,k'}}\|\right)\sqrt{\widehat{\rho}(C_{j+1,k'})} > a_{\max}\eta/2\right\} \\
&\le \sum_{C_{j+1,k'}\in\Lambda^+}\mathbb{P}\left\{\left(\|c_{j+1,k'} - \widehat{c}_{j+1,k'}\| + \theta_2 2^{-(j+1)}\|\mathrm{Proj}_{V_{j+1,k'}} - \mathrm{Proj}_{\widehat{V}_{j+1,k'}}\|\right)\sqrt{\widehat{\rho}(C_{j+1,k'})} \ge \eta/2\right\} \\
&\le \tilde{\gamma}_3 e^{-\tilde{\gamma}_4 2^{2j}n\eta^2} \qquad (54)
\end{aligned}$$

where $\tilde{\gamma}_3 := \tilde{\gamma}_3(\theta_2, \theta_3, a_{\max}, d)$ and $\tilde{\gamma}_4 := \tilde{\gamma}_4(\theta_2, \theta_3, \theta_4, a_{\max}, d)$.

Finally combining (49), (50), (51), (52), (53) and (54) yields

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\Delta}_{j,k} \le \eta \text{ and } \bar{\Delta}_{j,k} \ge (a_{\max}+2)\eta\right\} &\le \mathbb{P}\left\{|\widehat{\Delta}_{j,k} - \bar{\Delta}_{j,k}| \ge (a_{\max}+1)\eta\right\} \\
&\le \mathbb{P}\{e_1 > \eta\} + \mathbb{P}\{e_2 > a_{\max}\eta\} \le \tilde{\gamma}_5 e^{-\tilde{\gamma}_6 2^{2j}n\eta^2} \qquad (55)
\end{aligned}$$

for some constants $\tilde{\gamma}_5 := \tilde{\gamma}_5(\theta_2, \theta_3, a_{\max}, d)$ and $\tilde{\gamma}_6 := \tilde{\gamma}_6(\theta_2, \theta_3, \theta_4, a_{\max}, d)$.

Step Three: The probability $\mathbb{P}\left\{\bar{\Delta}_{j,k} \le (a_{\max}+2)\eta \text{ and } \Delta_{j,k} \ge (2a_{\max}+5)\eta\right\}$ is estimated as follows. For a fixed $C_{j,k}$, we define the function

$$f(x) = \|(\mathcal{P}_j - \mathcal{P}_{j+1})\,\mathbf{1}_{j,k}x\|\,,\ x \in \mathcal{M}.$$

Observe that $|f(x)| \le \frac{3}{2}\theta_2 2^{-j}$ for any $x \in \mathcal{M}$. We define $\|f\|^2 = \int_{\mathcal{M}} f^2(x)d\rho$ and $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(x_i)$. Then

$$\begin{aligned}
&\mathbb{P}\left\{\bar{\Delta}_{j,k} \le (a_{\max}+2)\eta \text{ and } \Delta_{j,k} \ge (2a_{\max}+5)\eta\right\} \\
&\le \mathbb{P}\left\{\Delta_{j,k} - 2\bar{\Delta}_{j,k} \ge \eta\right\} = \mathbb{P}\left\{\|f\| - 2\|f\|_n \ge \eta\right\} \le 3e^{-\frac{2^{2j}n\eta^2}{648\theta_2^2}} \qquad (56)
\end{aligned}$$

where the last inequality follows from Györfi et al. (2002, Theorem 11.2). Combining (47), (55) and (56) yields (24).

Next we turn to the bound in Eq. (24), which corresponds to the case that $\Delta_{j,k} \le \eta$ and $\widehat{\Delta}_{j,k} \ge b\eta$. In this case we have $\widehat{\Delta}_{j,k} \le \frac{3}{2}\theta_2 2^{-j}\sqrt{\widehat{\rho}(C_{j,k})}$ which implies

$$\widehat{\rho}(C_{j,k}) \ge \frac{4b^2 2^{2j}\eta^2}{9\theta_2^2}, \qquad (57)$$

instead of (48). We shall use the fact that $\rho(C_{j,k}) \geq (2b^2 2^{2j}\eta^2)/(9\theta_2^2)$ given (57) with high probability, by writing

$$\begin{aligned}\mathbb{P}\left\{\Delta_{j,k} \leq \eta \text{ and } \widehat{\Delta}_{j,k} \geq b\eta\right\} \leq & \mathbb{P}\left\{\Delta_{j,k} \leq \eta \text{ and } \widehat{\Delta}_{j,k} \geq b\eta \,\middle|\, \rho(C_{j,k}) \geq \tfrac{2b^2 2^{2j}\eta^2}{9\theta_2^2}\right\} \\ & + \mathbb{P}\left\{\rho(C_{j,k}) \leq \tfrac{2b^2 2^{2j}\eta^2}{9\theta_2^2} \text{ and } \widehat{\rho}(C_{j,k}) \geq \tfrac{4b^2 2^{2j}\eta^2}{9\theta_2^2}\right\} \quad (58)\end{aligned}$$

where the first term is estimated as above and the second one is estimated through Eq. (35) in Lemma 30:

$$\begin{aligned}& \mathbb{P}\left\{\rho(C_{j,k}) \leq \tfrac{2b^2 2^{2j}\eta^2}{9\theta_2^2} \text{ and } \widehat{\rho}(C_{j,k}) \geq \tfrac{4b^2 2^{2j}\eta^2}{9\theta_2^2}\right\} \\ \leq & \mathbb{P}\left\{\rho(C_{j,k}) \leq \frac{2b^2 2^{2j}\eta^2}{9\theta_2^2} \text{ and } |\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| \geq \frac{2b^2 2^{2j}\eta^2}{9\theta_2^2}\right\} \\ \leq & 2e^{-\left(3n(\frac{2b^2 2^{2j}\eta^2}{9\theta_2^2})^2\right)\Big/\left(6\rho(C_{j,k})+2\frac{2b^2 2^{2j}\eta^2}{9\theta_2^2}\right)} \leq 2e^{-\frac{3b^2 2^{2j} n\eta^2}{36\theta_2^2}}.\end{aligned}$$

Using the estimate in (58), we obtain the bound (24) which concludes the proof. ■

Proof [Proof of Lemma 21] We will show how Lemma 20 implies Eq. (25). Clearly $e_2 = 0$ if $\widehat{\Lambda}_{\tau_n} \vee \Lambda_{b\tau_n} = \widehat{\Lambda}_{\tau_n} \wedge \Lambda_{\tau_n/b}$, or equivalently $\widehat{\mathcal{T}}_{\tau_n} \cup \mathcal{T}_{b\tau_n} = \widehat{\mathcal{T}}_{\tau_n} \cap \mathcal{T}_{\tau_n/b}$. In the case $e_2 > 0$, the inclusion $\widehat{\mathcal{T}}_{\tau_n} \cap \mathcal{T}_{\tau_n/b} \subset \widehat{\mathcal{T}}_{\tau_n} \cup \mathcal{T}_{b\tau_n}$ is strict, i.e. there exists $C_{j,k} \in \mathcal{T}^n$ such that either $C_{j,k} \in \widehat{\mathcal{T}}_{\tau_n}$ and $C_{j,k} \notin \mathcal{T}_{\tau_n/b}$, or $C_{j,k} \in \mathcal{T}_{b\tau_n}$ and $C_{j,k} \notin \widehat{\mathcal{T}}_{\tau_n}$. In other words, there exists $C_{j,k} \in \mathcal{T}^n$ such that either $\Delta_{j,k} < 2^{-j}\tau_n/b$ and $\widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n$, or $\Delta_{j,k} \geq b2^{-j}\tau_n$ and $\widehat{\Delta}_{j,k} < 2^{-j}\tau_n$. As a result,

$$\begin{aligned}\mathbb{P}\{e_2 > 0\} \leq & \sum_{C_{j,k}\in\mathcal{T}^n} \mathbb{P}\left\{\widehat{\Delta}_{j,k} < 2^{-j}\tau_n \text{ and } \Delta_{j,k} \geq b2^{-j}\tau_n\right\} \quad (59) \\ & + \sum_{C_{j,k}\in\mathcal{T}^n} \mathbb{P}\left\{\Delta_{j,k} < 2^{-j}\tau_n/b \text{ and } \widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n\right\}.\end{aligned}$$

$$\mathbb{P}\{e_4 > 0\} \leq \sum_{C_{j,k}\in\mathcal{T}^n} \mathbb{P}\left\{\Delta_{j,k} < 2^{-j}\tau_n/b \text{ and } \widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n\right\}. \quad (60)$$

We apply (24) in Lemma 20 to estimate the first term in (59):

$$\begin{aligned}& \sum_{C_{j,k}\in\mathcal{T}^n} \mathbb{P}\left\{\widehat{\Delta}_{j,k} < 2^{-j}\tau_n \text{ and } \Delta_{j,k} \geq b2^{-j}\tau_n\right\} \leq \sum_{C_{j,k}\in\mathcal{T}^n} \alpha_1 e^{-\alpha_2 n 2^{2j}\cdot 2^{-2j}\kappa^2\frac{\log n}{n}} \\ & = \alpha_1 \#\mathcal{T}^n n^{-\alpha_2\kappa^2} \leq \alpha_1 a_{\min} n n^{-\alpha_2\kappa^2} \leq \alpha_1 a_{\min} n^{1-\alpha_2\kappa^2} = \alpha_1 a_{\min} n^{-(\alpha_2\kappa^2-1)}.\end{aligned}$$

Using (24), we estimate the second term in (59) and (60) as follows

$$\sum_{C_{j,k}\in\mathcal{T}^n} \mathbb{P}\left\{\Delta_{j,k} < 2^{-j}\tau_n/b \text{ and } \widehat{\Delta}_{j,k} \geq 2^{-j}\tau_n\right\} \leq \sum_{C_{j,k}\in\mathcal{T}^n} \alpha_1 e^{-\alpha_2 n 2^{2j}\cdot\frac{2^{-2j}}{b^2}\kappa^2\frac{\log n}{n}} \leq \alpha_1 a_{\min} n^{-(\alpha_2\kappa^2/b^2-1)}.$$

We therefore obtain (25) by choosing $\kappa > \kappa_2$ with $\alpha_2\kappa_2^2/b^2 = \nu + 1$. ■

Appendix E. Proofs for orthogonal GMRA

E.1. Performance analysis of orthogonal GMRA

The proofs of Theorem 23 and Theorem 25 are resemblant to the proofs of Theorem 4 and Theorem 8. The main difference lies in the variance term, which results in the extra log factors in the convergence rate of orthogonal GMRA. Let Λ be the partition associated with a finite proper subtree $\tilde{\mathcal{T}}$ of the data master tree $\mathcal{T}^n$, and let

$$\mathcal{S}_\Lambda = \sum_{C_{j,k}\in\Lambda} \mathcal{S}_{j,k}\mathbf{1}_{j,k} \quad \text{and} \quad \widehat{\mathcal{S}}_\Lambda = \sum_{C_{j,k}\in\Lambda} \widehat{\mathcal{S}}_{j,k}\mathbf{1}_{j,k}.$$

Lemma 33 *Let Λ be the partition associated with a finite proper subtree $\tilde{\mathcal{T}}$ of the data master tree $\mathcal{T}^n$. Suppose Λ contains $\#_j\Lambda$ cells at scale j. Then for any $\eta > 0$,*

$$\mathbb{P}\{\|\mathcal{S}_\Lambda X - \widehat{\mathcal{S}}_\Lambda X\| \geq \eta\} \leq \alpha d \left(\sum_{j\geq j_{\min}} j\#_j\Lambda\right) e^{-\frac{\beta n\eta^2}{d^2\sum_{j\geq j_{\min}} j^4 2^{-2j}\#_j\Lambda}} \tag{61}$$

where α and β are the constants in Lemma 15.

Proof [Proof of Lemma 33] The increasing subspaces $\{S_{j,x}\}$ in the construction of orthogonal GMRA may be written as

$$\begin{aligned}
S_{0,x} &= V_{0,x}\\
S_{1,x} &= V_{0,x} \oplus V_{0,x}^\perp V_{1,x}\\
S_{2,x} &= V_{0,x} \oplus V_{0,x}^\perp V_{1,x} \oplus V_{1,x}^\perp V_{0,x}^\perp V_{2,x}\\
&\cdots\\
S_{j,x} &= V_{0,x} \oplus V_{0,x}^\perp V_{1,x} \oplus \ldots \oplus V_{j-1,x}^\perp \cdots V_{1,x}^\perp V_{0,x}^\perp V_{j,x}.
\end{aligned}$$

Therefore $\|\mathrm{Proj}_{S_{j,x}} - \mathrm{Proj}_{\widehat{S}_{j,x}}\| \leq \sum_{\ell=0}^j (j+1-\ell)\|\mathrm{Proj}_{V_{\ell,x}} - \mathrm{Proj}_{\widehat{V}_{\ell,x}}\|$, and then

$$\mathbb{P}\left\{\|\mathrm{Proj}_{S_{j,x}} - \mathrm{Proj}_{\widehat{S}_{j,x}}\| \geq t\right\} \leq \sum_{\ell=0}^{j} \mathbb{P}\left\{\|\mathrm{Proj}_{V_{\ell,x}} - \mathrm{Proj}_{\widehat{V}_{\ell,x}}\| \geq t/j^2\right\}. \tag{62}$$

The rest of the proof is almost the same as the proof of Lemma 15 in appendix C with a slight modification of (41) substituted by (62). ■

The corollary of Lemma 33 with $\Lambda = \Lambda_j$ results in the following estimate of the variance in empirical orthogonal GMRA.

Lemma 34 *For any $\eta \geq 0$,*

$$\mathbb{P}\{\|\mathcal{S}_j X - \widehat{\mathcal{S}}_j X\| \geq \eta\} \leq \alpha dj\#\Lambda_j e^{-\frac{\beta 2^{2j} n\eta^2}{d^2 j^4 \#\Lambda_j}}, \tag{63}$$

$$\mathbb{E}\|\mathcal{S}_j X - \widehat{\mathcal{S}}_j X\|^2 \leq \frac{d^2 j^4 \#\Lambda_j \log[\alpha dj\#\Lambda_j]}{\beta 2^{2j} n}. \tag{64}$$

Proof [Proof of Theorem 23]

$$\mathbb{E}\|X - \widehat{\mathcal{S}}_j X\|^2 \le \|X - \mathcal{S}_j X\|^2 + \mathbb{E}\|\mathcal{S}_j X - \widehat{\mathcal{S}}_j X\|^2$$
$$\le |\rho|^2_{\mathcal{A}_s^o} 2^{-2sj} + \frac{d^2 j^4 \#\Lambda_j \log[\alpha d j \#\Lambda_j]}{\beta 2^{2j} n} \le |\rho|^2_{\mathcal{A}_s^o} 2^{-2sj} + \frac{d^2 j^4 2^{j(d-2)}}{\theta_1 \beta n} \log \frac{\alpha d j 2^{jd}}{\theta_1}.$$

When $d \ge 2$, We choose j^* such that $2^{-j^*} = \mu \left((\log^5 n)/n\right)^{\frac{1}{2s+d-2}}$. By grouping Λ_{j^*} into light and heavy cells whose measure is below or above $\frac{28}{3}(\nu+1)\log^5 n/n$, we can show that the error on light cells is upper bounded by $C((\log^5 n)/n)^{\frac{2s}{2s+d-2}}$ and all heavy cells have at least d points with high probability.

Lemma 35 *Suppose j^* is chosen such that $2^{-j^*} = \mu \left(\frac{\log^5 n}{n}\right)^{\frac{1}{2s+d-2}}$ with some $\mu > 0$. Then*

$$\|(X - \mathcal{P}_{j^*} X)\mathbf{1}_{\{C_{j^*,k}:\rho(C_{j^*,k}) \le \frac{28(\nu+1)\log^5 n}{3n}\}}\|^2 \le \frac{28(\nu+1)\theta_2^2 \mu^{2-d}}{3\theta_1} \left(\frac{\log^5 n}{n}\right)^{\frac{2s}{2s+d-2}},$$
$$\mathbb{P}\left\{\forall\, C_{j^*,k} \,:\, \rho(C_{j^*,k}) > \tfrac{28(\nu+1)\log^5 n}{3n},\ C_{j^*,k} \text{ has at least } d \text{ points}\right\} \ge 1 - n^{-\nu}.$$

Proof of Lemma 35 is omitted since it is the same as the proof of Lemma 17. Lemma 35 guarantees that a sufficient amount of cells at scale j^* has at least d points. The probability estimate in (28) follows from

$$\mathbb{P}\left\{\|\mathcal{S}_{j^*} X - \widehat{\mathcal{S}}_{j^*} X\| \ge C_1 \left(\frac{\log^5 n}{n}\right)^{\frac{s}{2s+d-2}}\right\} \le C_2 \log n \left(\frac{\log^5 n}{n}\right)^{-\frac{d}{2s+d-2}} e^{-\beta\theta_1\mu^{d-2}C_1^2(2s+d-2)^4/d^2 \log n}$$
$$\le C_2 (\log n)\, n^{\frac{d}{2s+d-2}} n^{-\beta\theta_1\mu^{d-2}C_1^2(2s+d-2)^4/d^2} \le C_2 n^{1-\beta\theta_1\mu^{d-2}C_1^2(2s+d-2)^4/d^2} \le C_2 n^{-\nu}$$

provided C_1 is chosen such that $\beta\theta_1\mu^{d-2}C_1^2(2s+d-2)^4/d^2 - 1 > \nu$. The proof when $d = 1$ is completely analogous to that of Theorem 4. ∎

E.2. Performance analysis of adaptive orthogonal GMRA

Proof [Proof of Theorem 25] Empirical adaptive orthogonal GMRA is given by $\widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}} = \sum_{C_{j,k} \in \widehat{\Lambda}_{\tau_n^o}} \widehat{\mathcal{S}}_{j,k} \mathbf{1}_{j,k}$. Using triangle inequality, we have

$$\|X - \widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}} X\| \le e_1 + e_2 + e_3 + e_4$$

with each term given by

$$e_1 := \|X - \mathcal{S}_{\widehat{\Lambda}_{\tau_n^o} \vee \Lambda_{b\tau_n^o}} X\| \qquad e_2 := \|\mathcal{S}_{\widehat{\Lambda}_{\tau_n^o} \vee \Lambda_{b\tau_n^o}} X - \mathcal{S}_{\widehat{\Lambda}_{\tau_n^o} \wedge \Lambda_{\tau_n^o/b}} X\|$$
$$e_3 := \|\mathcal{S}_{\widehat{\Lambda}_{\tau_n^o} \wedge \Lambda_{\tau_n^o/b}} X - \widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o} \wedge \Lambda_{\tau_n^o/b}} X\| \qquad e_4 := \|\widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o} \wedge \Lambda_{\tau_n^o/b}} X - \widehat{\mathcal{S}}_{\widehat{\Lambda}_{\tau_n^o}} X\|$$

where $b = 2a_{\max} + 5$. We will prove the case $d \ge 3$. Here one proceeds in the same way as in the proof of Theorem 8. A slight difference lies in the estimates of e_3, e_2 and e_4.

Term e_3: $\mathbb{E}e_3^2$ is the variance. One can verify that $\mathcal{T}_{(\rho,\tau_n^o/b)} \subset \mathcal{T}_{j_0} := \cup_{j\le j_0}\Lambda_j$ where j_0 is the largest integer satisfying $2^{j_0 d} \le 9b^2\theta_0\theta_2^2/(4\tau_n^{o2})$. The reason is that $\Delta_{j,k}^o \le \frac{3}{2}\theta_2 2^{-j}\sqrt{\theta_0 2^{-jd}}$ so $\Delta_{j,k}^o \ge 2^{-j}\tau_n^o/b$ implies $2^{j_0 d} \le 9b^2\theta_0\theta_2^2/(4\tau_n^{o2})$. For any $\eta > 0$,

$$\mathbb{P}\{e_3 > \eta\} \le \alpha d j_0 \#\mathcal{T}_{\tau_n^o/b} e^{-\frac{\beta n\eta^2}{j_0^4 \sum_{j\ge j_{\min}} 2^{-2j}\#_j\mathcal{T}_{\tau_n^o/b}}} \le \alpha d j_0 \#\mathcal{T}_{\tau_n^o/b} e^{-\frac{\beta n\eta^2}{j_0^4|\rho|_{\mathcal{B}_s^o}^p(\tau_n^o/b)^{-p}}}$$

The estimate of $\mathbb{E}e_3^2$ follows from

$$\mathbb{E}e_3^2 = \int_0^{+\infty} \eta\mathbb{P}\left\{e_3 > \eta\right\} d\eta = \int_0^{+\infty} \eta \min\left(1, \alpha d j_0 \#\mathcal{T}_{\tau_n^o/b} e^{-\frac{\beta n\eta^2}{j_0^4 \sum_{j\ge j_{\min}} 2^{-2j}\#_j\mathcal{T}_{\tau_n^o/b}}}\right) d\eta$$

$$\le \frac{j_0^4 \log \alpha j_0 \#\mathcal{T}_{\tau_n^o/b}}{\beta n} \sum_{j\ge j_{\min}} 2^{-2j}\#_j\mathcal{T}_{\tau_n^o/b} \le C\frac{\log^5 n}{n}(\tau_n^o/b)^{-p} \le C(\theta_0,\theta_2,\theta_3,a_{\max},\kappa,d,s)\left(\frac{\log^5 n}{n}\right)^{\frac{2s}{2s+d-2}}.$$

Term e_2 and e_4: These two terms are analyzed with Lemma 36 stated below such that for any fixed but arbitrary $\nu > 0$,

$$\mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \le \beta_1 a_{\min}/d n^{-\nu}$$

if κ is chosen such that $\kappa > \kappa_2$ with $d^4\beta_2\kappa_2^2/b^2 = \nu + 1$.

Lemma 36 *$b = 2a_{\max} + 5$. For any $\eta > 0$ and any $C_{j,k} \in \mathcal{T}$*

$$\max\left(\mathbb{P}\left\{\widehat{\Delta}_{j,k}^o \le \eta \ \text{ and } \ \Delta_{j,k}^o \ge b\eta\right\}, \mathbb{P}\left\{\Delta_{j,k}^o \le \eta \ \text{ and } \ \widehat{\Delta}_{j,k}^o \ge b\eta\right\}\right) \le \beta_1 j e^{-\beta_2 n 2^{2j}\eta^2/j^4},$$

with positive constants $\beta_1 := \beta_1(\theta_2,\theta_3,\theta_4,a_{\max},d)$ and $\beta_2 := \beta_2(\theta_2,\theta_3,\theta_4,a_{\max},d)$.

■

References

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. In *Proceedings OF SPARS*, pages 9–12, 2005.

W. K. Allard, G. Chen, and M. Maggioni. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32 (3):435–462, 2012.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *International Conference on Machine Learning*, pages 97–104, 2006.

P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.

P. Binev, A. Cohen, W. Dahmen, R.A. DeVore, and V. Temlyakov. Universal algorithms for learning theory part ii: piecewise polynomial functions. *Constructive Approximation*, 26(2):127–152, 2007.

G. Canas, T. Poggio, and L. Rosasco. Learning manifolds with k-means and k-flats. In *Advances in Neural Information Processing Systems*, 2012.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, pages 2313–2351, 2007. math.ST/0506081.

G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9:517–558, 2009.

G. Chen and M. Maggioni. Multiscale geometric and spectral analysis of plane arrangements. In *Conference on Computer Vision and Pattern Recognition*, 2011.

G. Chen, M. Iwen, S. Chin, and M. Maggioni. A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements. In *Visual Communications and Image Processing (VCIP), 2012 IEEE*, pages 1–6, 2012.

Guangliang Chen and Mauro Maggioni. Multiscale geometric wavelets for the analysis of point clouds. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2010.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Michael Christ. A t(b) theorem with remarks on analytic capacity and the cauchy integral. *Colloquium Mathematicae*, 60-61(2):601–628, 1990.

A. Cohen, I. Daubechies, O. G. Guleryuz, and M. T. Orchard. On the importance of combining wavelet-based nonlinear approximation with coding strategies. *IEEE Transactions on Information Theory*, 48(7):1895–1921, 2002.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005a.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7432–7438, 2005b.

I. Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

G. David and S. Semmes. *Analysis of and on uniformly rectifiable sets*, volume 38 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1993.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

D. Deng and Y. Han. *Harmonic analysis on spaces of homogeneous type*, volume 19. Springer Science and Business Media, 2008.

D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

D. L Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, pages 5591–5596, March 2003.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

D. L. Donoho and J. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

A. Eftekhari and M. B. Wakin. Sparse subspace clustering. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.

E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Conference on Computer Vision and Pattern Recognition*, pages 2790–2797, June 2009.

R. Fergus F. Li and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

S. Gerber and M. Maggioni. Multiscale dictionaries, transforms, and learning in high-dimensions. In *Proc. SPIE conference Optics and Photonics*, 2013.

R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. New York: Springer, 2002.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. 2009.

J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11–18, 2003.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(4):17–441,498–520, 1933.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

M. A. Iwen and M. Maggioni. Approximation of points on low-dimensional manifolds via random linear projections. *Inference and Information*, 2(1):1–31, 2013. arXiv:1204.3337v1, 2012.

P. W. Jones. Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae*, 102(1):1–15, 1990.

G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1999.

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, Kjersti Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, February 2003.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

M.S. Lewicki, T.J. Sejnowski, and H. Hughes. Learning overcomplete representations. *Neural Computation*, 12:337–365, 1998.

A. V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets i: Multiscale svd, noise and curvature. *Applied and Computational Harmonic Analysis*, 43 (3):504–567, 2017.

A.V. Little, Y.-M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *Proc. AAAI Fall Symposium Series*, 2009a.

A.V. Little, J. Lee, Y.-M. Jung, and M. Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In *Proc. IEEE/SP 15th Workshop on Statistical Signal Processing*, 2009b.

G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1–17, 2007.

Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 50(3):413–458, 2008.

M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: Non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17(1):43–93, January 2016.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

S. G. Mallat. *A wavelet tour in signal processing*. Academic Press, 1998.

A. Maurer and M. Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

G. Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.

G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34 (1):17–31, 2009.

M. Protter and M. Elad. Sparse and redundant representations and motion-estimation-free algorithm for video denoising, 2007.

M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *Journal of Chemical Physics*, (134):124116, 2011.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

A. Szlam. Asymptotic regularity of subdivisions of euclidean domains by iterated PCA and iterated 2-means. *Applied and Computational Harmonic Analysis*, 27(3):342–350, 2009.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

J. A. Tropp. User-friendly tools for random matrices: An introduction. NIPS version, 2014. URL `http://users.cms.caltech.edu/~jtropp/`.

D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. *The Journal of Machine Learning Research*, 12:3259–3281, 2011.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12), 2005.

J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *European conference on computer vision*, volume 4, pages 94–106, 2006.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University*, 8(4):406–424, 2004.

W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi. Polymer reversal rate calculated via locally scaled diffusion map. *Journal of Chemical Physics*, (134):144108, 2011.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, pages 1649–1656, 2006.