http://www.aimspress.com/journal/mine

Mathematics in Engineering, 4(4): 1–25.
DOI:10.3934/mine.2022028
Received: 15 March 2021
Accepted: 28 April 2021
Published: 24 August 2021

Research article

Multiscale regression on unknown manifolds†

Wenjing Liao[1], **Mauro Maggioni**[2] **and Stefano Vigogna**[3,*]

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30313, USA

[2] Department of Mathematics, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA

[3] MaLGa Center, Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genova, 16145 Genova, Italy

† **This contribution is part of the Special Issue:** Mathematical aspects of machine learning
Guest Editors: Ernesto De Vito; Lorenzo Rosasco; Silvia Villa
Link: www.aimspress.com/mine/article/6026/special-articles

* **Correspondence:** Email: vigogna@dibris.unige.it.

Abstract: We consider the regression problem of estimating functions on $\mathbb{R}^D$ but supported on a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ with $d \ll D$. Drawing ideas from multi-resolution analysis and nonlinear approximation, we construct low-dimensional coordinates on $\mathcal{M}$ at multiple scales, and perform multiscale regression by local polynomial fitting. We propose a data-driven wavelet thresholding scheme that automatically adapts to the unknown regularity of the function, allowing for efficient estimation of functions exhibiting nonuniform regularity at different locations and scales. We analyze the generalization error of our method by proving finite sample bounds in high probability on rich classes of priors. Our estimator attains optimal learning rates (up to logarithmic factors) as if the function was defined on a known Euclidean domain of dimension d, instead of an unknown manifold embedded in $\mathbb{R}^D$. The implemented algorithm has quasilinear complexity in the sample size, with constants linear in D and exponential in d. Our work therefore establishes a new framework for regression on low-dimensional sets embedded in high dimensions, with fast implementation and strong theoretical guarantees.

Keywords: manifold learning; polynomial regression; partitioning estimates; adaptive methods; optimal rates; multi-resolution analysis; nonlinear approximation; wavelet thresholding

1. Introduction

High-dimensional data challenge classical statistical models and require new understanding of tradeoffs in accuracy and efficiency. The seemingly quantitative fact of the increase of dimension has qualitative consequences in both methodology and implementation, demanding new ways to break what has been called the curse of dimensionality. On the other hand, the presence of inherent nonuniform structure in the data calls into question linear dimension reduction techniques, and motivates a search for intrinsic learning models. In this paper we explore the idea of learning and exploiting the intrinsic geometry and regularity of the data in the context of regression analysis. Our goal is to build low-dimensional representations of high dimensional functions, while ensuring good generalization properties and fast implementation. In view of the complexity of the data, we allow interesting features to change from scale to scale and from location to location. Hence, we will develop multiscale methods, extending classical ideas of multi-resolution analysis beyond regular domains and to the random sample regime.

In regression, the problem is to estimate a function from a finite set of random samples. The minimax mean squared error (MSE) for estimating functions in the Hölder space $C^s([0,1]^D)$, $s > 0$, is $O(n^{-2s/(2s+D)})$, where n is the number of samples. The exponential dependence of the minimax rate on D manifests the curse of dimensionality in statistical learning, as $n = O(\varepsilon^{-(2s+D)/s})$ points are generally needed to achieve accuracy ε. This rate is optimal (in the minimax sense), unless further structural assumptions are made [28, 32]. If the samples concentrate near a d-dimensional set with $d \ll D$, and the function belongs to a nonuniform smoothness space $\mathcal{B}^S$, with $S > s$, we may hope to find estimators converging in $O(n^{-2S/(2S+d)})$. In this quantified sense, we may break the curse of dimensionality by adapting to the intrinsic dimension and regularity of the problem.

A possible approach to this problem is based on first performing dimension reduction, and then regression in the reduced space. Linear dimension reduction methods include principal component analysis (PCA) [24,25,39], for data concentrating on a single subspace, or subspace clustering [8,9,18, 36,47], for a union of subspaces. Going beyond linear models, we encounter isomap [43], locally linear embedding [40], local tangent space alignment [49], Laplacian eigenmaps [2], Hessian eigenmap [15] and diffusion map [12]. Besides the classical Principal Component Regression [26], in [33] diffusion map is used for nonparametric regression expanding the unknown function over the eigenfunctions of a kernel-based operator. It is proved that, when data lie on a d-dimensional manifold, the MSE converges in $O(n^{-1/O(d^2)})$. This rate depends only on the intrinsic dimension, but does not match the minimax rate in the Euclidean space. If infinitely many unlabeled points are sampled, so that the eigenfunctions are exactly computed, the MSE can achieve optimal rates for Sobolev functions with smoothness parameter at least 1. Similar results hold for regression with the Laplacian eigenmaps [50].

Some regression methods have been shown to automatically adapt to the intrinsic dimension and perform as well as if the intrinsic domain was known. Results in this direction have been established for local linear regression [4], k-nearest neighbors [29], and kernel regression [31], where optimal rates depending on the intrinsic dimension were proved for functions in C^2, C^1, and C^s with $s \leq 1$, respectively. Kernel methods such as kernel ridge regression are also known to adapt to the intrinsic dimension [41, 48], while suitable variants of regression trees have been proved to attain intrinsic yet suboptimal learning rates [30]. On the other hand, dyadic partitioning estimates with piecewise polynomial regression can cover the whole scale of spaces C^s, $s > 0$ [21], and be combined with

wavelet thresholding techniques to optimally adapt to broader classes of nonuniform regularity [5, 6]. However, such estimators are cursed by the ambient dimension D, due to the exponential cardinality of a dyadic partition of the D-dimensional hypercube.

This paper aims at generalizing dyadic partitioning estimates [5, 6] to predict functions supported on low-dimensional sets, with optimal performance guarantees and low computational cost. We tie together ideas in classical statistical learning [20, 21, 45], multi-resolution analysis [12, 13, 38], and nonlinear approximation [11,16,17]. Our main tool is geometric multi-resolution analysis (GMRA) [1, 34,35,37], which is a multiscale geometric approximation scheme for point clouds in high dimensions concentrating near low-dimensional sets. Using GMRA we learn low-dimensional local coordinates at multiple scales, on which we perform a multiscale regression estimate by fitting local polynomials. Inspired by wavelet thresholding techniques [5,6,11], we then compute differences between estimators at adjacent scales, and retain the locations where such differences are large enough. This empirically reveals where higher resolution is required to attain a good approximation, generating a data-driven partition which adapts to the local regularity of the function.

Our approach has several distinctive features:

(i) it is *multiscale*, and is therefore well-suited for data sets containing variable structural information at different scales;

(ii) it is *adaptive*, allowing the function to have localized singularities or variable regularity;

(iii) it is entirely *data-driven*, that is, it does not require a priori knowledge about the regularity of the function, and rather learns it automatically from the data;

(iv) it is *provable*, with strong theoretical guarantees of optimal performance on large classes of priors;

(v) it is *efficient*, having straightforward implementation, minor parameter tuning, and low computational cost.

We will prove that, for functions supported on a d-dimensional manifold and belonging to a rich model class characterized by a smoothness parameter S, the MSE of our estimator converges at rate $O((\log n/n)^{2S/(2S+d)})$. This model class contains classical Hölder continuous functions, but further accounts for potential nonuniform regularity. Our results show that, up to a logarithmic factor, we attain the same optimal learning rate as if the function was defined on a known Euclidean domain of dimension d, instead of an unknown manifold embedded in $\mathbb{R}^D$. In particular, the rate of convergence depends on the intrinsic dimension d and not on the ambient dimension D. In terms of computations, all the constructions above can be realized by algorithms of complexity $O(n \log n)$, with constants linear in the ambient dimension D and exponential in the intrinsic dimension d.

The remainder of this paper is organized as follows. We conclude this section by defining some general notation and formalizing the problem setup. In Section 2 we review geometric multi-resolution analysis. In Section 3 we introduce our multiscale regression method, establish the performance guarantees, and discuss the computational complexity of our algorithms. The proofs of our results are collected in Section 4, with some technical details postponed to Appendix A.

Notation. $f \lesssim g$ and $f \gtrsim g$ mean that there exists a positive constant C, independent on any variable upon which f and g depend, such that $f \leq Cg$ and $f \geq Cg$, respectively. $f \asymp g$ means that both $f \lesssim g$ and $f \gtrsim g$ hold. 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 $V \subset \mathbb{R}^D$, we denote its dimension by $\dim(V)$ and the orthogonal projection onto V by Proj_V. Let $f, g : \mathcal{M} \to \mathbb{R}$ be two functions, and let ρ be a probability measure supported on $\mathcal{M}$. We define the inner product of f and g with respect to ρ as $\langle f, g\rangle := \int_{\mathcal{M}} f(x)g(x)d\rho$. The L^2 norm of f with respect to ρ is $\|f\| := (\int_{\mathcal{M}} |f(x)|^2 d\rho)^{\frac{1}{2}}$. Given n i.i.d. samples $\{x_i\}_{i=1}^n$ of ρ, the empirical L^2 norm of f is $\|f\|_n := \frac{1}{n}\sum_{i=1}^n |f(x_i)|^2$. The L^∞ norm of f is $\|f\|_\infty := \sup \mathrm{ess}\, |f|$. We denote probability and expectation by $\mathbb{P}$ and $\mathbb{E}$, respectively. For a fixed $M > 0$, T_M is the truncation operator defined by $T_M(x) := \min(|x|, M)\mathrm{sign}(x)$. We denote by $\mathbf{1}_{j,k}$ the indicator function of an indexed set $C_{j,k}$ (*i.e.*, $\mathbf{1}_{j,k}(x) = 1$ if $x \in C_{j,k}$, and 0 otherwise).

Setup. We consider the problem of estimating a function $f : \mathcal{M} \to \mathbb{R}$ given n samples $\{(x_i, y_i)\}_{i=1}^n$, where

- $\mathcal{M}$ is an unknown Riemannian manifold of dimension d isometrically embedded in $\mathbb{R}^D$, with $d \ll D$;
- ρ is an unknown probability measure supported on $\mathcal{M}$;
- $\{x_i\}_{i=1}^n$ are independently drawn from ρ;
- $y_i = f(x_i) + \zeta_i$;
- f is bounded, with $\|f\|_\infty \le M$;
- $\{\zeta_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random variables with sub-Gaussian norm $\|\zeta_i\|_{\psi_2} \le \sigma^2$, independent of the x_i's.

We wish to construct an estimator $\widehat{f}$ of f minimizing the mean squared error

$$\mathrm{MSE} := \mathbb{E}\|f - \widehat{f}\|^2 = \mathbb{E}\int_{\mathcal{M}} |f(x) - \widehat{f}(x)|^2 d\rho.$$

2. Geometric multi-resolution analysis

Geometric multi-resolution analysis (GMRA) is an efficient tool to build low-dimensional representations of data concentrating on or near a low-dimensional set embedded in high dimensions. To keep the presentation self-contained, we summarize here the main ideas, and refer the reader to [1, 34, 37] for further details. Given a probability measure ρ supported on a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$, GMRA performs the following steps:

(1). Construct a multiscale tree decomposition $\mathcal{T}$ of $\mathcal{M}$ into nested cells $\mathcal{T} := \{C_{j,k}\}_{k\in\mathcal{K}_j, j\in\mathbb{Z}}$, where j represents the scale and k the location. Here $\mathcal{K}_j$ is a location index set.

(2). Compute a local principal component analysis on each $C_{j,k}$. Let $c_{j,k}$ be the mean of x on $C_{j,k}$, and $V_{j,k}$ the d-dimensional principal subspace of $C_{j,k}$. Define $\mathcal{P}_{j,k} := c_{j,k} + \mathrm{Proj}_{V_{j,k}}(x - c_{j,k})$.

An ideal multiscale tree decomposition should satisfy assumptions (A1)÷(A5) below for all integers $j \ge j_{\min}$:

(A1) For every $k \in \mathcal{K}_j$ and $k' \in \mathcal{K}_{j+1}$, either $C_{j+1,k'} \subseteq C_{j,k}$ or $\rho(C_{j+1,k'} \cap C_{j,k}) = 0$. The children of $C_{j,k}$ are the cells $C_{j+1,k'}$ such that $C_{j+1,k'} \subseteq C_{j,k}$. We assume that $1 \le a_{\min} \le \#\{C_{j+1,k'} : C_{j+1,k'} \subseteq C_{j,k}\} \le a_{\max}$ for all $k \in \mathcal{K}_j$ and $j \ge j_{\min}$. 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) $\rho\left(\mathcal{M} \setminus \bigcup_{k\in\mathcal{K}_j} C_{j,k}\right) = 0$, *i.e.,* $\Lambda_j := \{C_{j,k}\}_{k\in\mathcal{K}_j}$ is a partition of $\mathcal{M}$, up to negligible sets.

(A3) There exists $\theta_1 > 0$ such that $\#\Lambda_j \le 2^{jd}/\theta_1$.

(A4) There exists $\theta_2 > 0$ such that, if x is drawn from ρ conditioned on $C_{j,k}$, then $\|x - c_{j,k}\| \le \theta_2 2^{-j}$ almost surely.

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

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

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

These are natural properties for multiscale partitions generalizing dyadic partitions to nonEuclidean domains [10]. (A1) establishes that the cells constitute a tree structure. (A2) says that the cells at scale j form a partition. (A3) guarantees that there are at most $2^{jd}/\theta_1$ cells at scale j. (A4) ensures that the diameter of all cells at scale j is bounded by 2^{-j}, up to a uniform constant. (A5)(i) assumes that the best rank d approximation to the covariance of a cell is close to the covariance matrix of a d-dimensional Euclidean ball, while (A5)(ii) assumes that the cell has significantly larger variance in d directions than in all the remaining ones.

Since all cells at scale j have similar diameter, Λ_j is called a uniform partition. A master tree $\mathcal{T}$ is a tree satisfying the properties above. A proper subtree $\tilde{\mathcal{T}}$ of $\mathcal{T}$ is a collection of nodes of $\mathcal{T}$ with the properties: the root node is in $\tilde{\mathcal{T}}$; if a node is in $\tilde{\mathcal{T}}$, then its parent is also in $\tilde{\mathcal{T}}$. Any finite proper subtree $\tilde{\mathcal{T}}$ is associated with a unique partition $\Lambda = \Lambda(\tilde{\mathcal{T}})$ consisting of its outer leaves, by which we mean those nodes that are not in $\tilde{\mathcal{T}}$, but whose parent is.

In practice, the master tree $\mathcal{T}$ is not given. We will construct one by an application of the cover tree algorithm [3] (see [34, Algorithm 3]). In order to make the samples for tree construction and function estimation independent from each other, we split the data in half and use one subset to construct the tree and the other one for local PCA and regression. From now on we index the training data as $\{(x_i, y_i)\}_{i=1}^{2n}$, and split them in $\{(x_i, y_i)\}_{i=1}^{2n} = \{(x_i, y_i)\}_{i=1}^{n} \cup \{(x_i, y_i)\}_{i=n+1}^{2n}$. Running Algorithm [34, Algorithm 3] on $\{x_i\}_{i=n+1}^{2n}$, we construct a family of cells $\{\widehat{C}_{j,k}\}_{k\in\mathcal{K}_j, j_{\min}\le j\le j_{\max}}$ which satisfies (A1)÷(A4) with high probability if ρ is doubling*; furthermore, if $\mathcal{M}$ is a C^s, $s \in (1, \infty)$, d-dimensional closed Riemannian manifold isometrically embedded in $\mathbb{R}^D$, and ρ is the volume measure on $\mathcal{M}$, then (A5) is satisfied as well:

Proposition 1 (Proposition 14 in [34])**.** *Assume ρ is a doubling probability measure on $\mathcal{M}$ with doubling constant C_1. Then, the $\widehat{C}_{j,k}$'s constructed from [34, Algorithm 3] satisfy:*

(a1) (A1) *with $a_{\max} = C_1^2(24)^d$ and $a_{\min} = 1$;*

*ρ is doubling if there exists $C_1 > 1$ such that $C_1^{-1} r^d \le \rho(\mathcal{M} \cap B_r(x)) \le C_1 r^d$ for any $x \in \mathcal{M}$ and $r > 0$; C_1 is called the doubling constant of ρ. See also [10, 14].

(a2) *let* $\widehat{\mathcal{M}} = \bigcup_{j=j_{\min}}^{j_{\max}} \bigcup_{k\in\mathcal{K}_j} \widehat{C}_{j,k}$*; for any* $\nu > 0$*,*

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

(a3) (A3) *with* $\theta_1 = C_1^{-1} 4^{-d}$*;*

(a4) (A4) *with* $\theta_2 = 3$.

If additionally $\mathcal{M}$ *is a* $C^s, s \in (1,\infty)$*, d-dimensional closed Riemannian manifold isometrically embedded in* $\mathbb{R}^D$*, and* ρ *is the volume measure on* $\mathcal{M}$*, then*

(a5) (A5) *is satisfied when j is sufficiently large.*

Since there are finite training points, the constructed master tree has a finite number of nodes. We first build a tree whose leaves contain a single point, and then prune it to the largest subtree whose leaves contain at least d training points. This pruned tree associated with the $\widehat{C}_{j,k}$'s is called the data master tree, and denoted by $\mathcal{T}_n$. The $\widehat{C}_{j,k}$'s cover $\widehat{\mathcal{M}}$, which represents the part of $\mathcal{M}$ that has been explored by the data. Even though assumption (A2) is not exactly satisfied, we claim that (a2) is sufficient for our performance guarantees, for example in the case where $\|f\|_\infty \le M$. Indeed, simply estimating f on $\mathcal{M} \setminus \widehat{\mathcal{M}}$ by 0, for any $\nu > 0$ we have

$$\mathbb{P}\left\{\int_{\mathcal{M}\setminus\widehat{\mathcal{M}}} \|f\|^2 d\rho \ge \frac{28M^2\nu \log n}{3n}\right\} \le 2n^{-\nu} \quad \text{and} \quad \mathbb{E}\int_{\mathcal{M}\setminus\widehat{\mathcal{M}}} \|f\|^2 d\rho \le \frac{56M^2\nu \log n}{3n^{1+\nu}}.$$

In view of these bounds, the rate of convergence on $\mathcal{M} \setminus \widehat{\mathcal{M}}$ is faster than the ones we will obtain on $\widehat{\mathcal{M}}$. We will therefore assume (A2), thanks to (a2). Also, it may happen that conditions (A3)÷(A5) are satisfied at the coarsest scales with very poor constants θ. Nonetheless, it will be clear that in all that follows we may discard a few coarse scales, and only work at scales that are fine enough and for which (A3)÷(A5) truly capture in a quantitative way the local geometry of $\mathcal{M}$. Since regression is performed on an independent subset of data, we can assume, by conditioning, that the $\widehat{C}_{j,k}$'s are given and satisfy the required assumptions. To keep the notation simple, from now on we will use $C_{j,k}$ instead of $\widehat{C}_{j,k}$, and $\mathcal{M}$ in place of $\widehat{\mathcal{M}}$, with a slight abuse of notation.

Besides cover tree, there are other methods that can be applied in practice to obtain multiscale partitions, such as METIS [27], used in [1], iterated PCA [42], and iterated k-means. These methods can be computationally more efficient than cover tree, but lead to partitions where the properties (A1)÷(A5) are not guaranteed to hold.

After constructing the multiscale tree $\mathcal{T}$, GMRA computes a collection of affine projectors $\{\mathcal{P}_j : \mathbb{R}^D \to \mathbb{R}^D\}_{j\ge j_{\min}}$. The main objects of GMRA in their population and sample version are summarized in Table 1. Given a suitable partition $\Lambda \subset \mathcal{T}$, $\mathcal{M}$ can be approximated by the piecewise linear set $\{\mathcal{P}_{j,k}(C_{j,k})\}_{C_{j,k}\in\Lambda}$.

Table 1. Objects of GMRA. $V_{j,k}$ and $\widehat{V}_{j,k}$ are the eigenspaces associated with the largest d eigenvalues of $\Sigma_{j,k}$ and $\widehat{\Sigma}_{j,k}$, respectively.

oracles	empirical counterparts
$\rho(C_{j,k})$	$\widehat{\rho}(C_{j,k}) := \frac{\widehat{n}_{j,k}}{n}, \quad \widehat{n}_{j,k} := \#\{x_i : x_i \in C_{j,k}\}$
$c_{j,k} := \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} x d\rho$	$\widehat{c}_{j,k} := \frac{1}{\widehat{n}_{j,k}} \sum_{x_i \in C_{j,k}} x_i$
$\Sigma_{j,k} := \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} (x - c_{j,k})(x - c_{j,k})^T d\rho$	$\widehat{\Sigma}_{j,k} := \frac{1}{\widehat{n}_{j,k}} \sum_{x_i \in C_{j,k}} (x_i - \widehat{c}_{j,k})(x_i - \widehat{c}_{j,k})^T$
$V_{j,k} := \underset{\dim V = d}{\arg\min} \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} \|x - c_{j,k} - \mathrm{Proj}_V(x - c_{j,k})\|^2 d\rho$	$\widehat{V}_{j,k} := \underset{\dim V = d}{\arg\min} \frac{1}{\widehat{n}_{j,k}} \sum_{x_i \in C_{j,k}} \|x - \widehat{c}_{j,k} - \mathrm{Proj}_V(x - \widehat{c}_{j,k})\|^2$
$\mathcal{P}_{j,k}(x) := c_{j,k} + \mathrm{Proj}_{V_{j,k}}(x - c_{j,k})$	$\widehat{\mathcal{P}}_{j,k}(x) := \widehat{c}_{j,k} + \mathrm{Proj}_{\widehat{V}_{j,k}}(x - \widehat{c}_{j,k})$

3. Multiscale polynomial regression

Given a multiscale tree decomposition $\{C_{j,k}\}_{j,k}$ and training samples $\{(x_i, y_i)\}_{i=1}^n$, we construct a family $\{\widehat{f}_{j,k}\}_{j,k}$ of local estimates of f in two stages: first we compute local coordinates on $C_{j,k}$ using GMRA outlined above, and then we estimate $f_{|C_{j,k}}$ by fitting a polynomial of order ℓ on such coordinates. A global estimator $\widehat{f}_\Lambda^\ell$ is finally obtained by summing the local estimates over a suitable partition Λ. Our regression method is specified in Algorithm 1. The explicit constructions of the constant ($\ell = 0$) and linear ($\ell = 1$) local estimators are detailed in Table 2.

In order to analyze the performance of our method, we introduce the oracle estimator f_Λ^ℓ based on the distribution ρ, defined by

$$\begin{aligned}
&\pi_{j,k} : C_{j,k} \to \mathbb{R}^d, \qquad \pi_{j,k}(x) := V_{j,k}^T (x - c_{j,k}),\\
&p_{j,k}^\ell := \underset{p \in P^\ell}{\arg\min} \int_{C_{j,k}} |y - p \circ \pi_{j,k}(x)|^2 d\rho,\\
&f_{j,k}^\ell := T_M[p_{j,k}^\ell \circ \pi_{j,k}]\\
&f_\Lambda^\ell := \sum_{C_{j,k} \in \Lambda} f_{j,k}^\ell \mathbf{1}_{j,k},
\end{aligned}$$

and split the MSE into a bias and a variance term:

$$\mathbb{E}\|f - \widehat{f}_\Lambda^\ell\|^2 \le 2 \underbrace{\|f - f_\Lambda^\ell\|^2}_{\text{bias}^2} + 2 \underbrace{\mathbb{E}\|f_\Lambda^\ell - \widehat{f}_\Lambda^\ell\|^2}_{\text{variance}}. \tag{1}$$

The bias term is a deterministic approximation error, and will be handled by assuming suitable regularity models for ρ and f (see Definitions 1 and 3). The variance term quantifies the stochastic error arising from finite-sample estimation, and will be bounded using concentration inequalities (see Proposition 2). The role of Λ, encoded in its size $\#\Lambda$, is crucial to balance (1). We will discuss two possible choices: uniform partitions in Section 3.1, and adaptive at multiple scales in Section 3.2.

Algorithm 1 GMRA regression

Input: training data $\{x_i, y_i\}_{i=1}^{2n}$, intrinsic dimension d, bound M, polynomial order ℓ, approximation type (uniform or adaptive).

Output: multiscale tree decomposition $\mathcal{T}_n$, partition Λ, piecewise ℓ-order polynomial estimator $\widehat{f}_\Lambda^\ell$.

1: construct a multiscale tree $\mathcal{T}_n$ by [34, Algorithm 3] on $\{x_i\}_{i=n+1}^{2n}$;

2: compute centers $\widehat{c}_{j,k}$ and subspaces $\widehat{V}_{j,k}$ by empirical GMRA on $\{x_i\}_{i=1}^{n}$;

3: define coordinates $\widehat{\pi}_{j,k}$ on $C_{j,k}$:

$$\widehat{\pi}_{j,k} : C_{j,k} \to \mathbb{R}^d, \qquad \widehat{\pi}_{j,k}(x) := \widehat{V}_{j,k}^T(x - \widehat{c}_{j,k});$$

4: compute local estimators $\widehat{g}_{j,k}^\ell$ by solving the following least squares problems over the space P^ℓ of polynomials of degree $\le \ell$:

$$\widehat{p}_{j,k}^\ell := \arg\min_{p \in P^\ell} \frac{1}{\widehat{n}_{j,k}} \sum_{i=1}^{n} |y_i - p \circ \widehat{\pi}_{j,k}(x_i)|^2 \mathbf{1}_{j,k}(x_i), \qquad \widehat{g}_{j,k}^\ell := \widehat{p}_{j,k}^\ell \circ \widehat{\pi}_{j,k}; \tag{2}$$

5: truncate $\widehat{g}_{j,k}^\ell$ by M:

$$\widehat{f}_{j,k}^\ell := T_M[\widehat{g}_{j,k}^\ell];$$

6: construct a uniform (see Section 3.1) or adaptive (see Section 3.2) partition Λ;

7: define the global estimator $\widehat{f}_\Lambda^\ell$ by summing the local estimators over the partition Λ:

$$\widehat{f}_\Lambda^\ell := \sum_{C_{j,k} \in \Lambda} \widehat{f}_{j,k}^\ell \mathbf{1}_{j,k}.$$

Table 2. Constant and linear local estimators. The truncation in $[\Lambda^{j,k}]_d$ regularizes the least squares problem, which is ill-posed due to the small eigenvalues $\{\lambda_l^{j,k}\}_{l=d+1}^{D}$.

oracles	empirical counterparts
piecewise constant ($\ell = 0$)	
$g_{j,k}^0(x) := y_{j,k} := \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} y d\rho$	$\widehat{g}_{j,k}^0(x) := \widehat{y}_{j,k} := \frac{1}{\widehat{n}_{j,k}} \sum_{x_i \in C_{j,k}} y_i$
$f_{j,k}^0(x) := T_M[g_{j,k}^0(x)]$	$\widehat{f}_{j,k}^0(x) := T_M[\widehat{g}_{j,k}^0(x)]$
piecewise linear ($\ell = 1$)	
$g_{j,k}^1(x) := [\pi_{j,k}(x)^T \; 2^{-j}]\beta_{j,k}$	$\widehat{g}_{j,k}^1(x) := [\widehat{\pi}_{j,k}(x)^T \; 2^{-j}]\widehat{\beta}_{j,k}$
$\beta_{j,k} := \begin{bmatrix} [\Lambda^{j,k}]_d^{-1} & 0 \\ 0 & 2^{2j} \end{bmatrix} \frac{1}{\rho(C_{j,k})} \int_{C_{j,k}} y \begin{bmatrix} \pi_{j,k}(x) \\ 2^{-j} \end{bmatrix} d\rho$	$\widehat{\beta}_{j,k} := \begin{bmatrix} [\widehat{\Lambda}^{j,k}]_d^{-1} & 0 \\ 0 & 2^{2j} \end{bmatrix} \frac{1}{\widehat{n}_{j,k}} \sum_{x_i \in C_{j,k}} y_i \begin{bmatrix} \widehat{\pi}_{j,k}(x_i) \\ 2^{-j} \end{bmatrix}$
$[\Lambda^{j,k}]_d := \mathrm{diag}(\lambda_1^{j,k}, \ldots, \lambda_d^{j,k})$	$[\widehat{\Lambda}^{j,k}]_d := \mathrm{diag}(\widehat{\lambda}_1^{j,k}, \ldots, \widehat{\lambda}_d^{j,k})$
$f_{j,k}^1(x) := T_M[g_{j,k}^1(x)]$	$\widehat{f}_{j,k}^1(x) := T_M[\widehat{g}_{j,k}^1(x)]$

3.1. Uniform partitions

A first natural choice for Λ is a uniform partition $\Lambda_j := \{C_{j,k}\}_{k\in\mathcal{K}_j}$, $j \geq j_{\min}$. At scale j, f is estimated by $\widehat{f}^{\ell}_{\Lambda_j} = \sum_{k\in\mathcal{K}_j} \widehat{f}^{\ell}_{j,k}\mathbf{1}_{j,k}$. The bias $\|f - f^{\ell}_{\Lambda_j}\|$ decays at a rate depending on the regularity of f, which can be quantified as follows:

Definition 1 (model class $\mathcal{A}^{\ell}_s$)**.** A function $f : \mathcal{M} \to \mathbb{R}$ is in the class $\mathcal{A}^{\ell}_s$ for some $s > 0$ with respect to the measure ρ if

$$|f|_{\mathcal{A}^{\ell}_s} := \sup_{\mathcal{T}} \sup_{j\geq j_{\min}} \frac{\|f - f^{\ell}_{\Lambda_j}\|}{2^{-js}} < \infty,$$

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

We capture the case where the bias is roughly the same on every cell with the following definition:

Definition 2 (model class $\mathcal{A}^{\ell,\infty}_s$)**.** A function $f : \mathcal{M} \to \mathbb{R}$ is in the class $\mathcal{A}^{\ell,\infty}_s$ for some $s > 0$ with respect to the measure ρ if

$$|f|_{\mathcal{A}^{\ell,\infty}_s} := \sup_{\mathcal{T}} \sup_{j\geq j_{\min}} \sup_{k\in\mathcal{K}_j} \frac{\|(f - f^{\ell}_{j,k})\mathbf{1}_{j,k}\|}{2^{-js}\sqrt{\rho(C_{j,k})}} < \infty,$$

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

Clearly $\mathcal{A}^{\ell,\infty}_s \subset \mathcal{A}^{\ell}_s$. These classes contain uniformly regular functions on manifolds, such as Hölder functions.

Example 1. Let $\mathcal{M}$ be a closed smooth d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$, and let ρ be the volume measure on $\mathcal{M}$. Consider a function $f : \mathcal{M} \to \mathbb{R}$ and a smooth chart (U, ϕ) on $\mathcal{M}$. The function $\tilde{f} : \phi(U) \to \mathbb{R}$ defined by $\tilde{f}(v) = f \circ \phi^{-1}(v)$ is called the coordinate representation of f. Let $\lambda = (\lambda_1, \ldots, \lambda_d)$ be a multi-index with $|\lambda| := \lambda_1 + \ldots + \lambda_d = \ell$. The ℓ-order λ-derivative of f is defined as

$$\partial^{\lambda} f(x) := \partial^{\lambda}(f \circ \phi^{-1}).$$

Hölder functions $C^{\ell,\alpha}$ on $\mathcal{M}$ with $\ell \in \mathbb{N}$ and $\alpha \in (0, 1]$ are defined as follows: $f \in C^{\ell,\alpha}$ if the ℓ-order derivatives of f exist, and

$$|f|_{C^{\ell,\alpha}} := \max_{|\lambda|=\ell} \sup_{x\neq z} \frac{|\partial^{\lambda} f(x) - \partial^{\lambda} f(z)|}{d(x, z)^{\alpha}} < \infty,$$

$d(x, z)$ being the geodesic distance between x and z. We will always assume to work at sufficiently fine scales at which $d(x, z) \asymp \|z - x\|_{\mathbb{R}^D}$. Note that $C^{\ell,1}$ is the space of ℓ-times continuously differentiable functions on $\mathcal{M}$ with Lipschitz ℓ-order derivatives. We have $C^{\ell,\alpha} \subset \mathcal{A}^{\ell,\infty}_{\ell+\alpha}$ with $|f|_{\mathcal{A}^{\ell,\infty}_{\ell+\alpha}} \leq \theta_2^{\ell+\alpha} d^{\ell} |f|_{C^{\ell,\alpha}}/\ell!$. The proof is in Appendix A.

Example 2. Let $\mathcal{M}$ be a smooth closed Riemannian manifold isometrically embedded in $\mathbb{R}^D$, and let ρ be the volume measure on $\mathcal{M}$. Let $\Omega \subset \mathcal{M}$ such that $\Gamma := \partial\Omega$ is a smooth and closed d_Γ-dimensional submanifold with finite reach[†]. Let $g = a\mathbf{1}_{\Omega} + b\mathbf{1}_{\Omega^{\complement}}$ for some $a, b \in \mathbb{R}$, where $\mathbf{1}_S$ denotes the indicator

[†]The reach of $\mathcal{M}$ is an important global characteristic of $\mathcal{M}$. Let $D(\mathcal{M}) := \{y \in \mathbb{R}^D : \exists!\ x \in \mathcal{M} \text{ s.t. } \|x - y\| = \inf_{z\in\mathcal{M}} \|z - y\|\}$, $\mathcal{M}_r := \{y \in \mathbb{R}^D : \inf_{x\in\mathcal{M}} \|x - y\| < r\}$. Then $\text{reach}(\mathcal{M}) := \sup\{r \geq 0 : \mathcal{M}_r \subset D(\mathcal{M})\}$. See also [19].

function of a set S. Then $g \in \mathcal{A}^\ell_{(d-d_\Gamma)/2}$ for every $\ell = 0, 1, 2, \ldots$; however, $g \notin \mathcal{A}^{\ell,\infty}_s$ for any $s > 0$. The proof is in Appendix A.

When we take uniform partitions $\Lambda = \Lambda_j$ in (1), the squared bias satisfies

$$\|f - f^\ell_{\Lambda_j}\|^2 \le |f|^2_{\mathcal{A}^\ell_s} 2^{-2js}$$

whenever $f \in \mathcal{A}^\ell_s$, which decreases as j increases. On the other hand, Proposition 2 shows that the variance at the scale j satisfies

$$\mathbb{E}\|f^\ell_{\Lambda_j} - \widehat{f}^\ell_{\Lambda_j}\|^2 \le O\left(\frac{j2^{jd}}{n}\right),$$

which increases as j increases. Choosing the optimal scale $j^\star$ in the bias-variance tradeoff, we obtain the following rate of convergence for uniform estimators:

Theorem 1. *Suppose $\|f\|_\infty \le M$ and $f \in \mathcal{A}^\ell_s$ for $\ell \in \{0, 1\}$ and $s > 0$. Let $j^\star$ be chosen such that*

$$2^{-j^\star} := \mu\left(\frac{\log n}{n}\right)^{\frac{1}{2s+d}}$$

for $\mu > 0$. Then there exist positive constants $c := c(\theta_1, d, \mu)$ and $C := C(\theta_1, d, \mu)$ for $\ell = 0$, or $c := c(\theta_1, \theta_2, \theta_3, d, \mu)$ and $C := C(\theta_1, \theta_2, \theta_3, d, \mu)$ for $\ell = 1$, such that:

(a) *for every $\nu > 0$ there is $c_\nu > 0$ such that*

$$\mathbb{P}\left\{\|f - \widehat{f}^\ell_{\Lambda_{j^\star}}\| > (|f|_{\mathcal{A}^\ell_s}\mu^s + c_\nu)\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}\right\} \le Cn^{-\nu},$$

where $c_\nu := c_\nu(\nu, \theta_1, d, M, \sigma, s, \mu)$ for $\ell = 0$, and $c_\nu := c_\nu(\nu, \theta_1, \theta_2, \theta_3, d, M, \sigma, s, \mu)$ for $\ell = 1$;

(b) $\mathbb{E}\|f - \widehat{f}^\ell_{\Lambda_{j^\star}}\|^2 \le \left(|f|^2_{\mathcal{A}^\ell_s}\mu^s + c\max(M^2, \sigma^2)\right)\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+d}}$.

Theorem 1 is proved in Section 4. Note that the rate depends on the intrinsic dimension d instead of the ambient dimension D. Moreover, the rate is optimal (up to logarithmic factors) at least in the case of $C^{\ell,\alpha}$ functions on $\mathcal{M}$, as discussed in Example 1.

3.2. Adaptive partitions

Theorem 1 is not fully satisfactory for two reasons: (i) the choice of the optimal scale requires knowledge of the regularity of the unknown function; (ii) no uniform scale can be optimal if the regularity of the function varies at different locations and scales. We thus propose an adaptive estimator which learns near-optimal partitions from data, without knowing the possibly nonuniform regularity of the function. Adaptive partitions may be selected by a criterion that determines whether or not a cell should be picked or not. The quantities involved in this selection are summarized in Table 3, along with their empirical versions.

Table 3. Local approximation difference between scales.

oracles	empirical counterparts
$W^\ell_{j,k} := (f^\ell_{\Lambda_j} - f^\ell_{\Lambda_{j+1}})\mathbf{1}_{j,k}$	$\widehat{W}^\ell_{j,k} := (\widehat{f}^\ell_{\Lambda_j} - \widehat{f}^\ell_{\Lambda_{j+1}})\mathbf{1}_{j,k}$
$\Delta^\ell_{j,k} := \|W^\ell_{j,k}\|$	$\widehat{\Delta}^\ell_{j,k} := \|\widehat{W}^\ell_{j,k}\|_n$

$\Delta^\ell_{j,k}$ plays the role of the magnitude of a wavelet coefficient in typical wavelet thresholding constructions, and reduces to it in the case of Haar wavelets on Euclidean domains by dyadic partitioning. It measures the local difference in approximation between two consecutive scales: a large $\Delta^\ell_{j,k}$ suggests a significant reduction of error if we refine $C_{j,k}$ to its children. Intuitively, we should truncate the master tree to the subtree including the nodes where this quantity is large. However, if too few samples exist in a node, then the empirical counterpart $\widehat{\Delta}^\ell_{j,k}$ can not be trusted. We thus proceed as follows. We set a threshold τ_n decreasing in n, and let $\widehat{\mathcal{T}}_n(\tau_n)$ be the smallest proper subtree of $\mathcal{T}_n$ containing all $C_{j,k}$'s for which $\widehat{\Delta}^\ell_{j,k} \geq \tau_n$. Crucially, τ_n may be chosen independently of the regularity of f (see Theorem 2). We finally define our adaptive partition $\widehat{\Lambda}_n(\tau_n)$ as the partition associated with the outer leaves of $\widehat{\mathcal{T}}_n(\tau_n)$. The procedure is summarized in Algorithm 2.

Algorithm 2 Adaptive partition

Input: training data $\{(x_i, y_i)\}_{i=1}^n$, multiscale tree decomposition $\mathcal{T}_n$, local ℓ-order polynomial estimates $\{\widehat{f}^\ell_{j,k}\}_{j,k}$, threshold parameter κ.

Output: adaptive partition $\widehat{\Lambda}_n(\tau_n)$.

1: compute the approximation difference $\widehat{\Delta}^\ell_{j,k}$ on every node $C_{j,k} \in \mathcal{T}_n$;
2: set the threshold $\tau_n := \kappa\sqrt{(\log n)/n}$;
3: select the smallest proper subtree $\widehat{\mathcal{T}}_n(\tau_n)$ of $\mathcal{T}_n$ containing all $C_{j,k}$'s with $\widehat{\Delta}^\ell_{j,k} \geq \tau_n$;
4: define the adaptive partition $\widehat{\Lambda}_n(\tau_n)$ associated with the outer leaves of $\widehat{\mathcal{T}}_n(\tau_n)$.

To provide performance guarantees for our adaptive estimator, we need to define a proper model class based on oracles. Given any master tree $\mathcal{T}$ satisfying assumptions (A1)÷(A5) and a threshold $\tau > 0$, we let $\mathcal{T}(\tau)$ be the smallest subtree of $\mathcal{T}$ consisting of all the cells $C_{j,k}$'s with $\Delta^\ell_{j,k} \geq \tau$. The partition made of the outer leaves of $\mathcal{T}(\tau)$ is denoted by $\Lambda(\tau)$.

Definition 3 (model class $\mathcal{B}^\ell_s$)**.** A function $f : \mathcal{M} \to \mathbb{R}$ is in the class $\mathcal{B}^\ell_s$ for some $s > 0$ with respect to the measure ρ if

$$|f|^p_{\mathcal{B}^\ell_s} := \sup_{\mathcal{T}} \sup_{\tau>0} \tau^p \#\mathcal{T}(\tau) < \infty, \qquad p = \frac{2d}{2s+d},$$

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

In general, the truncated tree $\mathcal{T}(\tau)$ grows as the threshold τ decreases. For elements in $\mathcal{B}^\ell_s$, we have control on the growth rate, namely $\#\mathcal{T}(\tau_n) \lesssim \tau^{-p}$. In the classical case of dyadic partitions of the Euclidean space with uniform measure, $\mathcal{B}^\ell_s$ is well understood as a nonlinear approximation space

containing a scale of Besov spaces [11]. The class $\mathcal{B}_s^\ell$ is indeed rich, and contains in particular $\mathcal{A}_s^{\ell,\infty}$, while additionally capturing functions of nonuniform regularity.

Lemma 1. *$\mathcal{A}_s^{\ell,\infty} \subset \mathcal{B}_s^\ell$. If $f \in \mathcal{A}_s^{\ell,\infty}$, then $f \in \mathcal{B}_s^\ell$ and $|f|_{\mathcal{B}_s^\ell} \le (a_{\min}/\theta_1)^{\frac{2s+d}{2d}} |f|_{\mathcal{A}_s^{\ell,\infty}}$.*

The proof is given in Appendix A.

Example 3. Let g be the function in Example 2. Then $g \in \mathcal{B}^\ell_{d(d-d_\Gamma)/(2d_\Gamma)}$ for every $\ell = 0, 1, 2, \ldots$. Notice that $g \in \mathcal{A}^\ell_{(d-d_\Gamma)/2}$, so g has a larger regularity parameter s in the $\mathcal{B}_s^\ell$ model than in the $\mathcal{A}_s^\ell$ model.

We will also need a quasi-orthogonality condition ensuring that the functions $W_{j,k}^\ell$ representing the approximation difference between two scales are almost orthogonal across scales.

Definition 4. We say that f satisfies quasi-orthogonality of order ℓ with respect to the measure ρ if there exists a constant $B_0 > 0$ such that, for any proper subtree $\mathcal{S}$ of any tree $\mathcal{T}$ satisfying assumptions (A1)÷(A5),

$$\Big\| \sum_{C_{j,k} \in \mathcal{T} \setminus \mathcal{S}} W_{j,k}^\ell \Big\|^2 \le B_0 \sum_{C_{j,k} \in \mathcal{T} \setminus \mathcal{S}} \|W_{j,k}^\ell\|^2.$$

The following lemma shows that $f \in \mathcal{B}_s^\ell$, along with quasi-orthogonality, implies a certain approximation rate of f by $f^\ell_{\Lambda(\tau)}$ as $\tau \to 0$. The proof is given in Appendix A.

Lemma 2. *If $f \in \mathcal{B}_s^\ell \cap (L^\infty \cup \mathcal{A}_t^\ell)$ for some $s, t > 0$, and f satisfies quasi-orthogonality of order ℓ, then*

$$\|f - f^\ell_{\Lambda(\tau)}\|^2 \le B_{s,d} |f|^p_{\mathcal{B}_s} \tau^{2-p} \le B_{s,d} |f|^2_{\mathcal{B}_s} \#\Lambda(\tau)^{-\frac{2s}{d}}, \quad p = \frac{2d}{2s+d},$$

with $B_{s,d} := B_0 2^p \sum_{\ell \ge 0} 2^{-\ell(2-p)}$.

The main result of this paper is the following performance analysis of adaptive estimators, which is proved in Section 4.

Theorem 2. *Let $\ell \in \{0, 1\}$ and $M > 0$. Suppose $\|f\| \le M$ and f satisfies quasi-orthogonality of order ℓ. Set $\tau_n := \kappa \sqrt{\log n / n}$. Then:*

(a) *For every $\nu > 0$ there exists $\kappa_\nu := \kappa_\nu(a_{\max}, \theta_2, \theta_3, d, M, \sigma, \nu) > 0$ such that, whenever $f \in \mathcal{B}_s^\ell$ for some $s > 0$ and $\kappa \ge \kappa_\nu$, there are $r, C > 0$ such that*

$$\mathbb{P}\left\{ \|f - \widehat{f}^\ell_{\widehat{\Lambda}_n(\tau_n)}\| > r \left(\frac{\log n}{n} \right)^{\frac{s}{2s+d}} \right\} \le C n^{-\nu}.$$

(b) *There exists $\kappa_0 := \kappa_0(a_{\max}, \theta_2, \theta_3, d, M, \sigma)$ such that, whenever $f \in \mathcal{B}_s^\ell$ for some $s > 0$ and $\kappa \ge \kappa_0$, there is $\bar{C} > 0$ such that*

$$\mathbb{E}\|f - \widehat{f}^\ell_{\widehat{\Lambda}_n(\tau_n)}\|^2 \le \bar{C} \left(\frac{\log n}{n} \right)^{\frac{2s}{2s+d}}.$$

Here r depends on θ_2, θ_3, $a_{\max}$, d, M, s, $|f|_{\mathcal{B}_s^\ell}$, σ, B_0, ν, κ; C depends on θ_2, θ_3, $a_{\max}$, $a_{\min}$, d, s, $|f|_{\mathcal{B}_s^\ell}$, κ; $\bar{C}$ depends on θ_2, θ_3, $a_{\max}$, $a_{\min}$, d, M, s, $|f|_{\mathcal{B}_s^\ell}$, B_0, κ.

Theorem 2 is more satisfactory than Theorem 1 for two reasons: (i) the same rate is achieved for a richer model class; (ii) the estimator does not require a priori knowledge of the regularity of the function, since the choice of κ is independent of s.

For a given accuracy ε, in order to achieve MSE $\lesssim \varepsilon^2$, the number of samples we need is $n_\varepsilon \gtrsim (1/\varepsilon)^{\frac{2s+d}{s}} \log(1/\varepsilon)$. When s is unknown, we can determine s as follows: we fix a small n_0, and run Algorithm 2 with $2n_0, 4n_0, \ldots, 2^j n_0, \ldots$ samples. For each sample size, we evenly split data into a training set to build the adaptive estimator, and a test set to evaluate the MSE. According to Theorem 2, the MSE scales as $(\log n/n)^{\frac{2s}{2s+d}}$. Therefore, the slope in the log-log plot of the MSE versus n gives an approximation of $-2s/(2s+d)$. This could be formalized by a suitable adaptation of Lepski's method.

3.3. Computational complexity

The computational cost of Algorithms 1 and 2 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 [3].

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 [22]. 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 at most $c \log n$ scales where $c > 0$ is a constant, which gives a total cost of $cDdn \log n$.

Multiscale regression. Given $\widehat{n}_{j,k}$ training points on $C_{j,k}$, computing the low-dimensional coordinates $\widehat{\pi}_{j,k}(x_i)$ for all $x_i \in C_{j,k}$ costs $Dd\widehat{n}_{j,k}$, and solving the linear least squares problem (2), where the matrix is of size $\widehat{n}_{j,k} \times d^\ell$, costs at most $\widehat{n}_{j,k} d^{2\ell}$. Hence, constructing the ℓ-order polynomials at scale j takes $\sum_{k\in\mathcal{K}_j} Dd\widehat{n}_{j,k} + d^{2\ell}\widehat{n}_{j,k} = (Dd + d^{2\ell})n$, and there are at most $c \log n$ scales, which sums up to $c(Dd + d^{2\ell})n \log n$.

Adaptive approximation. We need to compute the coefficients $\widehat{\Delta}_{j,k}$ for every $C_{j,k}$, which costs $2(Dd + d^\ell)\widehat{n}_{j,k}$ on $C_{j,k}$, and $2c(Dd + d^\ell)n \log n$ for the whole tree.

In summary, the total cost of constructing GMRA adaptive estimators of order ℓ is

$$C^d Dn \log n + (4Dd + d^{2\ell} + d^\ell)cn \log n.$$

The cost scales linearly with the number of samples n up to a logarithmic factor, and linearly with the ambient dimension D.

4. Proofs

We analyze the error of our estimator by a bias-variance decomposition as in (1). We present the variance estimate in Section 4.1, the proofs for uniform approximations in Section 4.2, and for adaptive approximations in Section 4.3.

4.1. Variance estimate

The following proposition bounds the variance of 0-order (piecewise constant) and 1st-order (piecewise linear) estimators over an arbitrary partition Λ.

Proposition 2. *Suppose $\|f\|_\infty \le M$ and let $\ell \in \{0,1\}$. For any partition Λ, let f_Λ^ℓ and $\widehat{f}_\Lambda^\ell$ be the optimal approximation and the empirical estimators of order ℓ on Λ, respectively. Then, for every $\eta > 0$,*

$$\text{(a) } \mathbb{P}\left\{\|f_\Lambda^\ell - \widehat{f}_\Lambda^\ell\| > \eta\right\} \le \begin{cases} C_0 \#\Lambda \exp\left(-\frac{n\eta^2}{c_0 \max(M^2,\sigma^2)\#\Lambda}\right) & \text{for } \ell = 0 \\ C_1 d \#\Lambda \exp\left(-\frac{n\eta^2}{c_1 \max(d^4M^2,d^2\sigma^2)\#\Lambda}\right) & \text{for } \ell = 1, \end{cases}$$

$$\text{(b) } \mathbb{E}\|f_\Lambda^\ell - \widehat{f}_\Lambda^\ell\|^2 \le \begin{cases} \frac{c_0 \max(M^2,\sigma^2)\#\Lambda \log(C_0\#\Lambda)}{n} & \text{for } \ell = 0 \\ \frac{c_1 \max(d^4M^2,d^2\sigma^2)\#\Lambda \log(C_1 d\#\Lambda)}{n} & \text{for } \ell = 1, \end{cases}$$

for some absolute constants c_0, C_0 and some c_1, C_1 depending on θ_2, θ_3.

Proof. Since f_Λ^ℓ and $\widehat{f}_\Lambda^\ell$ are bounded by M, we define $\Lambda^- := \{C_{j,k} \in \Lambda : \rho(C_{j,k}) \le \frac{\eta^2}{4M^2\#\Lambda}\}$, and observe that

$$\sum_{C_{j,k}\in\Lambda^-} \|(f_\Lambda^\ell - \widehat{f}_\Lambda^\ell)\mathbf{1}_{j,k}\|^2 \le \eta^2.$$

We then restrict our attention to $\Lambda^+ := \Lambda \setminus \Lambda^-$ and apply Lemma 5 with $t = \frac{\eta}{\sqrt{\rho(C_{j,k})\#\Lambda}}$. This leads to (a), while (b) follows from (a) by integrating over $\eta > 0$. □

4.2. Proof of Theorem 1

Notice that $\#\Lambda_j \le 2^{jd}/\theta_1$ by (A3). By choosing $j^\star$ such that $2^{-j^\star} = \mu\left(\frac{\log n}{n}\right)^{\frac{1}{2s+d}}$ for some $\mu > 0$, we have

$$\|f - f_{\Lambda_{j^\star}}^\ell\| \le |f|_{\mathcal{A}_s^\ell} 2^{-j^\star s} \le |f|_{\mathcal{A}_s^\ell} \mu^s \left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}.$$

Moreover, by Proposition 2,

$$\mathbb{P}\left\{\|f_{\Lambda_{j^\star}}^\ell - \widehat{f}_{\Lambda_{j^\star}}^\ell\| \ge c_\nu \left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}\right\} \le \begin{cases} \frac{C_0}{\theta_1\mu^d}(\log n)^{-\frac{d}{2s+d}} n^{-\left(\frac{\theta_1\mu^d c_\nu^2}{c_0\max(M^2,\sigma^2)} - \frac{d}{2s+d}\right)} & (\ell = 0) \\ \frac{C_1 d}{\theta_1\mu^d}(\log n)^{-\frac{d}{2s+d}} n^{-\left(\frac{\theta_1\mu^d c_\nu^2}{c_1\max(d^4M^2,d^2\sigma^2)} - \frac{d}{2s+d}\right)} & (\ell = 1) \end{cases} \le Cn^{-\nu}$$

provided that $\frac{\theta_1\mu^d c_\nu^2}{c_0\max(M^2,\sigma^2)} - \frac{d}{2s+d} > \nu$ for $\ell = 0$ and $\frac{\theta_1\mu^d c_\nu^2}{c_1\max(d^4M^2,d^2\sigma^2)} - \frac{d}{2s+d} > \nu$ for $\ell = 1$. □

4.3. Proof of Theorem 2

We begin by defining several objects of interest:

- $\mathcal{T}_n$: the data master tree whose leaves contain at least d points of training data. It can be viewed as the part of a multiscale tree that our training data have explored. Notice that

$$\#\mathcal{T}_n \le \sum_{j=0}^{\infty} a_{\min}^{-j} \frac{n}{d} = \frac{a_{\min}}{a_{\min}-1}\frac{n}{d} \le a_{\min}\frac{n}{d}.$$

- $\mathcal{T}$: 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.

- $\mathcal{T}(\tau)$: the smallest subtree of $\mathcal{T}$ which contains $\{C_{j,k} \in \mathcal{T} \,:\, \Delta_{j,k}^{\ell} \geq \tau\}$.
- $\mathcal{T}_n(\tau) := \mathcal{T}(\tau) \cap \mathcal{T}_n$.
- $\widehat{\mathcal{T}}_n(\tau)$: the smallest subtree of $\mathcal{T}_n$ which contains $\{C_{j,k} \in \mathcal{T}_n \,:\, \widehat{\Delta}_{j,k}^{\ell} \geq \tau\}$.
- $\Lambda(\tau)$: the adaptive partition associated with $\mathcal{T}(\tau)$.
- $\Lambda_n(\tau)$: the adaptive partition associated with $\mathcal{T}_n(\tau)$.
- $\widehat{\Lambda}_n(\tau)$: the adaptive partition associated with $\widehat{\mathcal{T}}_n(\tau)$.
- 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.
- Let $b = 2a_{\max} + 5$ where $a_{\max}$ is the maximal number of children that a node has in $\mathcal{T}_n$.

Inspired by the analysis of wavelet thresholding procedures [5,6], we split the error into four terms,

$$\|f - \widehat{f}^{\ell}_{\widehat{\Lambda}_n(\tau_n)}\| \leq e_1 + e_2 + e_3 + e_4,$$

where

$$e_1 := \|f - f^{\ell}_{\widehat{\Lambda}_n(\tau_n)\vee\Lambda_n(b\tau_n)}\| \qquad e_2 := \|f^{\ell}_{\widehat{\Lambda}_n(\tau_n)\vee\Lambda_n(b\tau_n)} - f^{\ell}_{\widehat{\Lambda}_n(\tau_n)\wedge\Lambda_n(\tau_n/b)}\|$$
$$e_3 := \|f^{\ell}_{\widehat{\Lambda}_n(\tau_n)\wedge\Lambda_n(\tau_n/b)} - \widehat{f}^{\ell}_{\widehat{\Lambda}_n(\tau_n)\wedge\Lambda_n(\tau_n/b)}\| \qquad e_4 := \|\widehat{f}^{\ell}_{\widehat{\Lambda}_n(\tau_n)\wedge\Lambda_n(\tau_n/b)} - \widehat{f}^{\ell}_{\widehat{\Lambda}_n(\tau_n)}\|.$$

The goal of the splitting above is to handle the bias and variance separately, as well as to deal with the fact the partition built from those $C_{j,k}$ such that $\widehat{\Delta}_{j,k}^{\ell} \geq \tau_n$ does not coincide with the partition which would be chosen by an oracle based on those $C_{j,k}$ such that $\Delta_{j,k}^{\ell} \geq \tau_n$. This is accounted by the terms e_2 and e_4 which correspond to those $C_{j,k}$ such that $\widehat{\Delta}_{j,k}^{\ell}$ is significantly larger or smaller than $\Delta_{j,k}^{\ell}$ respectively, and which will be proved to be small in probability. The e_1 and e_3 terms correspond to the bias and variance of oracle estimators based on partitions obtained by thresholding the unknown oracle change in approximation $\Delta_{j,k}^{\ell}$.

Since $\widehat{\Lambda}_n(\tau_n) \vee \Lambda_n(b\tau_n)$ is a finer partition than $\Lambda_n(b\tau_n)$, we have

$$e_1 \leq \|f - f^{\ell}_{\Lambda_n(b\tau_n)}\| \leq \|f - f^{\ell}_{\Lambda(b\tau_n)}\| + \|f^{\ell}_{\Lambda(b\tau_n)} - f^{\ell}_{\Lambda_n(b\tau_n)}\| =: e_{11} + e_{12}\,.$$

The e_{11} term is treated by a deterministic estimate based on the model class $\mathcal{B}_s^{\ell}$: by Lemma 2 we have

$$e_{11}^2 \leq B_{s,d}|f|^p_{\mathcal{B}_s^{\ell}}(b\kappa)^{2-p}(\log n/n)^{\frac{2s}{2s+d}},$$

The term e_{12} accounts for the error on the cells that have not been explored by our training data, which is small:

$$\begin{aligned}\mathbb{P}\{e_{12} > 0\} &\leq \mathbb{P}\{\exists\, C_{j,k} \in \mathcal{T}(b\tau_n) \setminus \mathcal{T}_n(b\tau_n)\} \\ &= \mathbb{P}\{\exists\, C_{j,k} \in \mathcal{T}(b\tau_n) : \Delta_{j,k}^{\ell} \geq b\tau_n \text{ and } \widehat{\rho}(C_{j,k}) < d/n\}\end{aligned}$$

$$\leq \sum_{C_{j,k}\in\mathcal{T}(b\tau_n)} \mathbb{P}\{\Delta^\ell_{j,k} \geq b\tau_n \text{ and } \widehat{\rho}(C_{j,k}) < d/n\}.$$

According to (4), we have $(\Delta^\ell_{j,k})^2 \leq 4\|f\|^2_\infty \rho(C_{j,k})$. Then every $C_{j,k}$ with $\Delta^\ell_{j,k} \geq b\tau_n$ satisfies $\rho(C_{j,k}) \gtrsim \frac{b^2\kappa^2}{\|f\|^2_\infty}(\log n/n)$. Hence, provided that n satisfies $\frac{b^2\kappa^2}{\|f\|^2_\infty}\log n \gtrsim 2d$, we have

$$\mathbb{P}\{\Delta^\ell_{j,k} \geq b\tau_n \text{ and } \widehat{\rho}(C_{j,k}) < d/n\} \leq \mathbb{P}\left\{|\rho(C_{j,k}) - \widehat{\rho}(C_{j,k})| \geq \frac{1}{2}\rho(C_{j,k}) \text{ and } \rho(C_{j,k}) \gtrsim \frac{b^2\kappa^2}{\|f\|^2_\infty}\frac{\log n}{n}\right\}$$
$$\leq 2n^{-\frac{3b^2\kappa^2}{28\|f\|^2_\infty}},$$

where the last inequality follows from Lemma 3(b). Therefore, by Definition 3 we obtain

$$\mathbb{P}\{e_{12} > 0\} \lesssim \#\mathcal{T}(b\tau_n)n^{-\frac{3b^2\kappa^2}{28\|f\|^2_\infty}} \leq |f|^p_{\mathcal{B}^\ell_s}(b\tau_n)^{-p}n^{-\frac{3b^2\kappa^2}{28\|f\|^2_\infty}} \leq |f|^p_{\mathcal{B}^\ell_s}(b\kappa)^{-p}n^{-\left(\frac{3b^2\kappa^2}{28\|f\|^2_\infty}-1\right)} \leq |f|^p_{\mathcal{B}^\ell_s}(b\kappa)^{-p}n^{-\nu}$$

as long as $\frac{3b^2\kappa^2}{28\|f\|^2_\infty} - 1 > \nu$. To estimate $\mathbb{E}e^2_{12}$, we observe that, thanks to Lemma 6,

$$e^2_{12} = \sum_{C_{j,k}\in\Lambda_n(b\tau_n)\setminus\Lambda(b\tau_n)} \sum_{\substack{C_{j',k'}\in\Lambda(b\tau_n)\\ C_{j',k'}\subset C_{j,k}}} \|(f^\ell_{j,k} - f^\ell_{j',k'})\mathbf{1}_{j',k'}\|^2 \lesssim M^2.$$

Hence, by choosing $\nu = 1 > \frac{2s}{2s+d}$ we get

$$\mathbb{E}e^2_{12} \lesssim \|f\|^2_\infty \mathbb{P}\{e_{12} > 0\} \lesssim \|f\|^2_\infty |f|^p_{\mathcal{B}^\ell_s}(b\kappa)^{-p}(\log n/n)^{\frac{2s}{2s+d}}.$$

The term e_3 is the variance term which can be estimated by Proposition 2 with $\Lambda = \widehat{\Lambda}_n(\tau_n) \wedge \Lambda_n(\tau_n/b)$. We plug in $\eta = r(\log n/n)^{\frac{s}{2s+d}}$. Bounding $\#\Lambda$ by $\#\Lambda_n(\tau_n/b) \leq \#\Lambda_n \leq n/d$ (as our data master tree has at d points in each leaf) outside the exponential, and by $\#\Lambda_n(\tau_n/b) \leq \#\Lambda(\tau_n/b) \leq |f|^p_{\mathcal{B}^\ell_s}(\tau_n/b)^{-p}$ inside the exponential, we get the following estimates for e_3:

$$\mathbb{P}\left\{e_3 > r\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}\right\} \leq \begin{cases} C_0 n^{1-\frac{r^2\kappa^p}{c_0 b^p |f|^p_{\mathcal{B}^\ell_s}\max\{M^2,\sigma^2\}}} & (\ell = 0)\\ C_1 n^{1-\frac{\gamma^2\kappa^p}{c_1 b^p |f|^p_{\mathcal{B}^\ell_s}\max\{d^4M^2,d^2\sigma^2\}}} & (\ell = 1), \end{cases}$$

where $C_0 = C_0(\theta_2, \theta_3, a_{\max}, d, s, |f|_{\mathcal{B}^\ell_s}, \kappa)$ and $C_1 = C_1(\theta_2, \theta_3, a_{\max}, d, s, |f|_{\mathcal{B}^\ell_s}, \kappa)$. We obtain $\mathbb{P}\{e_3 > r(\log n/n)^{\frac{s}{2s+d}}\} \leq Cn^{-\nu}$ as long as r is chosen large enough to make the exponent smaller than $-\nu$.

To estimate $\mathbb{E}e^2_3$, we apply again Propositions 2 and with $\#\Lambda \leq |f|^p_{\mathcal{B}^\ell_s}(b/\kappa)^p(\log n/n)^{-\frac{d}{2s+d}}$, obtaining

$$\mathbb{E}e^2_3 \leq \bar{C}(\log n/n)^{\frac{2s}{2s+d}}.$$

Next we estimate e_2 and e_4. Since $\widehat{\mathcal{T}}_n(\tau_n) \cap \mathcal{T}_n(\tau_n/b) \subseteq \widehat{\mathcal{T}}_n(\tau_n) \cup \mathcal{T}_n(b\tau_n)$ and $\mathcal{T}_n(b\tau_n) \subseteq \mathcal{T}_n(\tau_n/b)$, we have $e_2 > 0$ if and only if there is a $C_{j,k} \in \mathcal{T}_n$ such that either $C_{j,k}$ is in $\widehat{\mathcal{T}}_n(\tau_n)$ but not in $\mathcal{T}_n(\tau_n/b)$,

or $C_{j,k}$ is in $\mathcal{T}_n(b\tau_n)$ but not in $\widehat{\mathcal{T}}_n(\tau_n)$. This means that either $\widehat{\Delta}^\ell_{j,k} \geq \tau_n$ but $\Delta^\ell_{j,k} < \tau_n/b$, or $\Delta^\ell_{j,k} \geq b\tau_n$ but $\widehat{\Delta}^\ell_{j,k} < \tau_n$. As a consequence,

$$\mathbb{P}\{e_2 > 0\} \leq \sum_{C_{j,k}\in\mathcal{T}_n} \mathbb{P}\left\{\widehat{\Delta}^\ell_{j,k} \geq \tau_n \text{ and } \Delta^\ell_{j,k} < \tau_n/b\right\} + \sum_{C_{j,k}\in\mathcal{T}_n} \mathbb{P}\left\{\Delta^\ell_{j,k} \geq b\tau_n \text{ and } \widehat{\Delta}^\ell_{j,k} < \tau_n\right\},$$

and analogously

$$\mathbb{P}\{e_4 > 0\} \leq \sum_{C_{j,k}\in\mathcal{T}_n} \mathbb{P}\left\{\widehat{\Delta}^\ell_{j,k} \geq \tau_n \text{ and } \Delta^\ell_{j,k} < \tau_n/b\right\}.$$

We can now apply Lemma 7: we use (b) with $\eta = \tau_n/b$, and (a) with $\eta = \tau_n$. We obtain that

$$\mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \leq \begin{cases} C(a_{\min}, d)n^{1-\frac{\kappa^2}{c_0 b^2 \max\{M^2,\sigma^2\}}} & (\ell = 0) \\ C(\theta_2, \theta_3, a_{\min}, d)n^{1-\frac{\kappa^2}{c_1 b^2 \max\{d^4M^2,d^2\sigma^2\}}} & (\ell = 1). \end{cases}$$

We have $\mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \leq Cn^{-\nu}$ provided that κ is chosen such that the exponents are smaller than $-\nu$.

We are left to deal with the expectations. As for e_2, Lemma 6 implies $e_2 \lesssim M$, which gives rise to, for $\nu = 1 > \frac{2s}{2s+d}$,

$$\mathbb{E}e_2^2 \lesssim M^2\mathbb{P}\{e_2 > 0\} \leq CM^2(\log n/n)^{\frac{2s}{2s+d}}.$$

The same bound holds for e_4, which concludes the proof of Theorem 1. □

4.4. Basic concentration inequalities

This section contains the main concentration inequalities of the empirical quantities on their oracles. For piecewise linear estimators, some quantities used in Lemma 5 are decomposed in Table 4. All proofs are collected in Appendix A.

Table 4. Decomposition of piecewise linear estimators into quantities used in Lemma 5.

oracles	empirical counterparts
$f_{j,k}(x) = T_M\left([(x - c_{j,k})^T \;\; 2^{-j}]Q_{j,k}r_{j,k}\right)$	$\widehat{f}_{j,k}(x) = T_M\left([(x - \widehat{c}_{j,k})^T \;\; 2^{-j}]\widehat{Q}_{j,k}\widehat{r}_{j,k}\right)$
$Q_{j,k} := \begin{bmatrix} [\Sigma_{j,k}]^\dagger_d & 0 \\ 0 & 2^{2j} \end{bmatrix}$	$\widehat{Q}_{j,k} := \begin{bmatrix} [\widehat{\Sigma}_{j,k}]^\dagger_d & 0 \\ 0 & 2^{2j} \end{bmatrix}$
$[\Sigma_{j,k}]^\dagger_d := V_{j,k}[\Lambda^{j,k}]^{-1}_d {V_{j,k}}^T$	$[\widehat{\Sigma}_{j,k}]^\dagger_d := \widehat{V}_{j,k}{[\widehat{\Lambda}^{j,k}]_d}^{-1}(\widehat{V}_{j,k})^T$
$r_{j,k} := \frac{1}{\rho(C_{j,k})}\int_{C_{j,k}} y\begin{bmatrix} (x-c_{j,k}) \\ 2^{-j} \end{bmatrix} d\rho$	$\widehat{r}_{j,k} := \frac{1}{\widehat{n}_{j,k}}\sum_{x_i\in C_{j,k}} y_i\begin{bmatrix} (x_i-\widehat{c}_{j,k}) \\ 2^{-j} \end{bmatrix}$

Lemma 3. *For every $t > 0$ we have:*

(a) $\mathbb{P}\left\{\left|\rho(C_{j,k}) - \widehat{\rho}(C_{j,k})\right| > t\right\} \leq 2\exp\left(\frac{-3nt^2}{6\rho(C_{j,k})+2t}\right)$;

(b) *setting $t = \frac{1}{2}\rho(C_{j,k})$ in* (a) *yields* $\mathbb{P}\left\{|\rho(C_{j,k}) - \widehat{\rho}(C_{j,k})| > \frac{1}{2}\rho(C_{j,k})\right\} \leq 2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right)$;

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

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

Lemma 4. *We have:*

(a) $\mathbb{P}\{\|Q_{j,k} - \widehat{Q}_{j,k}\| > \frac{48}{\theta_3^2}d^2 2^{4j}\|\Sigma_{j,k} - \widehat{\Sigma}_{j,k}\|\}$
$\le 2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right) + \left(4\frac{\theta_2^2}{\theta_3}d + 10\right)\exp\left(-\frac{n\rho(C_{j,k})}{512(\theta_2^2/\theta_3)^2 d^2 + \frac{64}{3}(\theta_2^2/\theta_3)d}\right)$.

(b) $\mathbb{P}\{\|\widehat{Q}^{j,k}\| > \frac{2}{\theta_3}d\,2^{2j}\} \le 2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right) + \left(4\frac{\theta_2^2}{\theta_3}d + 10\right)\exp\left(-\frac{n\rho(C_{j,k})}{128(\theta_2^2/\theta_3)^2 d^2 + \frac{32}{3}(\theta_2^2/\theta_3)d}\right)$.

(c) *Suppose f is in L^∞. For every $t > 0$, we have*
$\mathbb{P}\left\{\|r_{j,k} - \widehat{r}_{j,k}\| > t\right\} \le 2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right) + 8\exp\left(\frac{-n\rho(C_{j,k})t^2}{4\langle\theta_2\rangle^2\|f\|_\infty^2 2^{-2j}+2\langle\theta_2\rangle\|f\|_\infty 2^{-j}t}\right) + 2\exp\left(-c\frac{n\rho(C_{j,k})t^2}{\theta_2^2\|\zeta\|_{\psi_2}^2 2^{-2j}}\right)$
where c is an absolute constant.

Lemma 5. *Suppose f is in L^∞. For every $t > 0$, we have*

$$\mathbb{P}\left\{\|f_{j,k}^\ell - \widehat{f}_{j,k}^\ell\|_\infty > t\right\} \le \begin{cases} C_0\left[\exp\left(-\frac{n\rho(C_{j,k})}{c_0}\right) + \exp\left(-\frac{n\rho(C_{j,k})t^2}{c_0(\|f\|_\infty^2+\|f\|_\infty t)}\right) + \exp\left(-\frac{n\rho(C_{j,k})t^2}{c_0\|\zeta\|_{\psi_2}^2}\right)\right] & \text{for } \ell = 0 \\ C_1 d\left[\exp\left(-\frac{n\rho(C_{j,k})}{c_1 d^2}\right) + \exp\left(-\frac{n\rho(C_{j,k})t^2}{c_1 d^4(\|f\|_\infty^2+\|f\|_\infty t)}\right) + \exp\left(-\frac{n\rho(C_{j,k})t^2}{c_1 d^2\|\zeta\|_{\psi_2}^2}\right)\right] & \text{for } \ell = 1, \end{cases}$$

where c_0, C_0 are absolute constants, c_0' depends on θ_2, and c_1, C_1 depend on θ_2, θ_3.

Lemma 6. *Suppose $f \in L^\infty$. For every $C_{j,k} \in \mathcal{T}$ and $C_{j',k'} \subset C_{j,k}$,*

$$\|f_{j,k} - f_{j',k'}\|_\infty \le 2M, \qquad \|\widehat{f}_{j,k} - \widehat{f}_{j',k'}\|_\infty \le 2M.$$

Lemma 7. *Suppose f is in L^∞. For every $\eta > 0$ and any $\gamma > 1$, we have*

(a) $\mathbb{P}\left\{\widehat{\Delta}_{j,k}^\ell < \eta \;\&\; \Delta_{j,k}^\ell \ge (2a_{\max}+5)\eta\right\} \le \begin{cases} C_0\exp\left(-\frac{n\eta^2}{c_0\max\{\|f\|_\infty^2, \|\zeta\|_{\psi_2}^2\}}\right) & \text{for } \ell = 0 \\ C_1 d\exp\left(-\frac{n\eta^2}{c_1\max\{d^4\|f\|_\infty^2, d^2\|\zeta\|_{\psi_2}^2\}}\right) & \text{for } \ell = 1; \end{cases}$

(b) $\mathbb{P}\left\{\Delta_{j,k}^\ell < \eta \;\&\; \widehat{\Delta}_{j,k}^\ell \ge (2a_{\max}+5)\eta\right\} \le \begin{cases} C_0\exp\left(-\frac{n\eta^2}{c_0\max\{\|f\|_\infty^2, \|\zeta\|_{\psi_2}^2\}}\right) & \text{for } \ell = 0 \\ C_1 d\exp\left(-\frac{n\eta^2}{c_1\max\{d^4\|f\|_\infty^2, d^2\|\zeta\|_{\psi_2}^2\}}\right) & \text{for } \ell = 1; \end{cases}$

C_0, c_0 depend on $a_{\max}$; c_0' depends on $a_{\max}, \theta_2$; C_1, c_1 depend on $a_{\max}, \theta_2, \theta_3$.

Acknowledgements

This work was partially supported by NSF-DMS-125012, AFOSR FA9550-17-1-0280, NSF-IIS-1546392. The authors thank Duke University for donating computing equipment used for this project.

Conflict of interest

The authors declare no conflict of interest.

References

1. W. K. Allard, G. Chen, M. Maggioni, Multi-scale geometric methods for data sets II: geometric multi-resolution analysis, *Appl. Comput. Harmon. Anal.*, **32** (2012), 435–462.
2. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.*, **15** (2003), 1373–1396.
3. A. Beygelzimer, S. Kakade, J. Langford, Cover trees for nearest neighbor, In: *Proceedings of the 23rd international conference on Machine learning*, 2006, 97–104.
4. P. J. Bickel, B. Li, Local polynomial regression on unknown manifolds, *Lecture Notes–Monograph Series*, **54** (2007), 177–186.
5. P. Binev, A. Cohen, W. Dahmen, R. A. DeVore, Universal algorithms for learning theory part II: Piecewise polynomial functions, *Constr. Approx.*, **26** (2007), 127–152.
6. P. Binev, A. Cohen, W. Dahmen, R. A. DeVore, V. N. Temlyakov, Universal algorithms for learning theory part I: Piecewise constant functions, *J. Mach. Learn. Res.*, **6** (2005), 1297–1321.
7. V. Buldygin, E. Pechuk, Inequalities for the distributions of functionals of sub-Gaussian vectors, *Theor. Probability and Math. Statist.*, **80** (2010), 25–36.
8. G. Chen, G. Lerman, Spectral Curvature Clustering (SCC), *Int. J. Comput. Vis.*, **81** (2009), 317–330.
9. G. Chen, M. Maggioni, Multiscale geometric and spectral analysis of plane arrangements, In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, 2825–2832.
10. M. Christ, A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral, *Colloq. Math.*, **60/61** (1990), 601–628.
11. A. Cohen, W. Dahmen, I. Daubechies, R. A. DeVore, Tree approximation and optimal encoding, *Appl. Comput. Harmon. Anal.*, **11** (2001), 192–226.
12. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, et al., Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, *PNAS*, **102** (2005), 7426–7431.
13. I. Daubechies, *Ten lectures on wavelets*, SIAM, 1992.
14. D. Deng, Y. Han, *Harmonic analysis on spaces of homogeneous type*, Springer, 2008.
15. D. L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, *PNAS*, **100** (2003), 5591–5596.
16. D. L. Donoho, J. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81** (1994), 425–455.
17. D. L. Donoho, J. M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *J. Am. Stat. Assoc.*, **90** (1995), 1200–1224.
18. E. Elhamifar, R. Vidal, Sparse subspace clustering, In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 2790–2797.
19. H. Federer, Curvature measures, *T. Am. Math. Soc.*, **93** (1959), 418–491.
20. J. Friedman, T. Hastie, R. Tibshirani, *The elements of statistical learning*, Springer, 2001.
21. L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A distribution-free theory of nonparametric regression*, Springer, 2002.

22. N. Halko, P. G. Martinsson, J. A. Tropp, Finding structure with randomness: stochastic algorithms for constructing approximate matrix decompositions, *SIAM Rev.*, **53** (2011), 217–288.
23. P. C. Hansen, The truncated SVD as a method for regularization, *Bit Numer. Math.*, **27** (1987), 534–553.
24. H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24** (1933), 417–441.
25. H. Hotelling, Relations between two sets of variates, *Biometrika*, **28** (1936), 321–377.
26. I. T. Jolliffe, A note on the use of principal components in regression, *J. C. Stat. Soc. C. Appl.*, **31** (1982), 300–303.
27. G. Karypis, V. Kumar, A fast and high quality multilevel scheme for partitioning irregular graphs, *SIAM J. Sci. Comput.*, **20** (1999), 359–392.
28. T. Klock, A. Lanteri, S. Vigogna, Estimating multi-index models with response-conditional least squares, *Electron. J. Stat.*, **15** (2021), 589–629.
29. S. Kpotufe, *k*-NN regression adapts to local intrinsic dimension, In: *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011, 729–737.
30. S. Kpotufe, S. Dasgupta, A tree-based regressor that adapts to intrinsic dimension, *J. Comput. Syst. Sci.*, **78** (2012), 1496–1515.
31. S. Kpotufe, V. K. Garg, Adaptivity to local smoothness and dimension in kernel regression, In: *Advances in Neural Information Processing Systems 26 (NIPS 2011)*, 2013, 3075–3083.
32. A. Lanteri, M. Maggioni, S. Vigogna, Conditional regression for single-index models, 2020 *arXiv:2002.10008*.
33. A. B. Lee, R. Izbicki, A spectral series approach to high-dimensional nonparametric regression, *Electron. J. Stat.*, **10** (2016), 423–463.
34. W. Liao, M. Maggioni, Adaptive geometric multiscale approximations for intrinsically low-dimensional data, *J. Mach. Learn. Res.*, **20** (2019), 1–63.
35. W. Liao, M. Maggioni, S. Vigogna, Learning adaptive multiscale approximations to data and functions near low-dimensional sets, In: *IEEE Information Theory Workshop (ITW)*, 2016, 226–230.
36. G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, In: *Proceedings of the 26 th International Conference on Machine Learning*, 2010, 663–670.
37. M. Maggioni, S. Minsker, N. Strawn, Multiscale dictionary learning: Non-asymptotic bounds and robustness, *J. Mach. Learn. Res.*, **17** (2016), 1–51.
38. S. Mallat, *A wavelet tour of signal processing*, 2 Eds., Academic Press, 1999.
39. K. Pearson, On lines and planes of closest fit to systems of points in space, *Philos. Mag.*, **2** (1901), 559–572.
40. S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, **290** (2000), 2323–2326.
41. I. Steinwart, D. R. Hush, C. Scovel, Optimal rates for regularized least squares regression, In: *The 22nd Annual Conference on Learning Theory*, 2009.
42. A. Szlam, Asymptotic regularity of subdivisions of euclidean domains by iterated PCA and iterated 2-means, *Appl. Comput. Harmon. Anal.*, **27** (2009), 342–350.

43. J. B. Tenenbaum, V. D. Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, **290** (2000), 2319–2323.
44. J. A. Tropp, User-friendly tools for random matrices: An introduction, NIPS version, 2012.
45. A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009.
46. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, In: *Compressed sensing*, Cambridge University Press, 2012, 210–268.
47. R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), *IEEE T. Pattern Anal.*, **27** (2005), 1945–1959.
48. G. B. Ye, D. X. Zhou, Learning and approximation by Gaussians on Riemannian manifolds, *Adv. Comput. Math.*, **29** (2008), 291–310.
49. Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, *SIAM J. Sci. Comput.*, **26** (2002), 313–338.
50. X. Zhou, N. Srebro, Error analysis of Laplacian eigenmaps for semi-supervised learning, In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011, 901–908.

A. Additional proofs

Example 1. Let $f \in C^{\ell,\alpha}$. The local estimator $f^{\ell}_{j,k}$ minimizes $\|(f-p)\mathbf{1}_{j,k}\|$ over all possible polynomials p of order less than or equal to ℓ. Thus, in particular, we have $\|(f-f^{\ell}_{j,k})\mathbf{1}_{j,k}\| \le \|(f-p)\mathbf{1}_{j,k}\|$ where p is equal to the ℓ-order Taylor polynomial of f at some $z \in C_{j,k}$. Hence, for $x \in C_{j,k}$ there is $\xi \in \mathcal{M} \cap B_{\theta_2 2^{-j}}(z)$ such that

$$\begin{aligned}|f(x)-p(x)| &\le \sum_{|\lambda|=\ell} \frac{1}{\lambda!}|\partial^{\lambda} f(\xi) - \partial^{\lambda} f(z)||x-z|^{\lambda} \le |f|_{C^{\ell,\alpha}}\|\xi - z\|^{\alpha} \sum_{|\lambda|=\ell} \frac{1}{\lambda!}|x-z|^{\lambda} \\ &\le \frac{d^{\ell}}{\ell!}|f|_{C^{\ell,\alpha}}\|\xi-z\|^{\alpha}\|x-z\|^{\ell} \le \theta_2^{\ell+\alpha}\frac{d^{\ell}}{\ell!}|f|_{C^{\ell,\alpha}} 2^{-j(\ell+\alpha)}.\end{aligned}$$

Therefore, for every j and $k \in \mathcal{K}_j$, we have

$$\|(f-f^{\ell}_{j,k})\mathbf{1}_{j,k}\|^2 \le \theta_2^{2(\ell+\alpha)}(\frac{d^{\ell}}{\ell!})^2 |f|^2_{C^{\ell,\alpha}} 2^{-2j(\ell+\alpha)}\rho(C_{j,k}). \qquad \square$$

Examples 2 and 3. For polynomial estimators of any fixed order $\ell = 0, 1, \ldots$, $g^{\ell}_{j,k} - g\mathbf{1}_{j,k} = 0$ when $C_{j,k} \cap \Gamma = \varnothing$, and $g^{\ell}_{j,k} - g\mathbf{1}_{j,k} = O(1)$ when $C_{j,k} \cap \Gamma \neq \varnothing$. At the scale j, $\rho(C_{j,k}) \approx 2^{-jd}$ and $\rho(\cup\{C_{j,k} : C_{j,k} \cap \Gamma \neq \varnothing\}) \approx 2^{-j(d-d_\Gamma)}\rho(\Gamma)$. Therefore,

$$\|g^{\ell}_{\Lambda_j} - g\| \le O(\sqrt{2^{-j(d-d_\Gamma)}}) = O(2^{-j(d-d_\Gamma)/2}),$$

which implies $g \in \mathcal{A}^{\ell}_{(d-d_\Gamma)/2}$.

In adaptive approximations, $\Delta^{\ell}_{j,k} = 0$ when $C_{j,k} \cap \Gamma = \varnothing$. When $C_{j,k} \cap \Gamma \neq \varnothing$, $\Delta^{\ell}_{j,k} = \|g^{\ell}_{j,k} - \sum_{C_{j+1,k'} \subset C_{j,k}} g^{\ell}_{j+1,k'}\| \lesssim \sqrt{\rho(C_{j,k})} \lesssim 2^{-jd/2}$. Given any fixed threshold $\tau > 0$, in the truncated tree $\mathcal{T}(\tau)$, the leaf nodes intersecting with Γ satisfy $2^{-jd/2} \gtrsim \tau$. In other words, around Γ the tree is truncated

at a coarser scale than $j^\star$ such that $2^{-j^\star} = O(\tau^{\frac{2}{d}})$. The cardinality of $\mathcal{T}(\tau)$ is dominated by the nodes intersecting with Γ, so

$$\#\mathcal{T}(\tau) \lesssim \frac{\rho(\Gamma)2^{-j^\star(d-d_\Gamma)}}{2^{-j^\star d}} = \rho(\Gamma)2^{j^\star d_\Gamma} \lesssim \tau^{-\frac{2d_\Gamma}{d}},$$

which implies $p = 2d_\Gamma/d$. We conclude that $g \in \mathcal{B}^\ell_s$ with $s = \frac{d(2-p)}{2p} = \frac{d}{d_\Gamma}(d - d_\Gamma)/2$. □

Lemma 1. By definition, we have $\|(f - f^\ell_{j,k})\mathbf{1}_{j,k}\| \le |f|_{\mathcal{A}^{\ell,\infty}_s} 2^{-js}\sqrt{\rho(C_{j,k})}$ as long as $f \in \mathcal{A}^{\ell,\infty}_s$. By splitting $(\Delta^\ell_{j,k})^2 \le 2\|(f - f^\ell_{j,k})\mathbf{1}_{j,k}\|^2 + 2\sum_{k':C_{j+1,k'}\subset C_{j,k}} \|(f - f^\ell_{j+1,k'})\mathbf{1}_{j+1,k'}\|^2$, we get

$$(\Delta^\ell_{j,k})^2 \le 4|f|^2_{\mathcal{A}^{\ell,\infty}_s} 2^{-2js}\rho(C_{j,k}).$$

In the selection of adaptive partitions, every $C_{j,k}$ with $\Delta^\ell_{j,k} \ge \tau$ must satisfy $\rho(C_{j,k}) \ge 2^{2js}(\tau/|f|_{\mathcal{A}^{\ell,\infty}_s})^2$. With extra assumptions $\rho(C_{j,k}) \le \theta_0 2^{-jd}$ (true when the measure ρ is doubling), we have

$$\Delta^\ell_{j,k} \ge \tau \Longrightarrow 2^{-j} \ge \left(\frac{\tau}{|f|_{\mathcal{A}^{\ell,\infty}_s}}\right)^{\frac{2}{2s+d}}. \tag{3}$$

Therefore, every cell in $\Lambda(\tau)$ will be at a coarser scale than $j^\star$ with $j^\star$ satisfying (3). Using (A3) we thus get

$$\tau^p \#\mathcal{T}(\tau) \le \tau^p a_{\min} \#\Lambda_{j^\star} \le \theta_1^{-1}\tau^p a_{\min} 2^{j^\star d} \le \frac{a_{\min}|f|^{\frac{2d}{2s+d}}_{\mathcal{A}^{\ell,\infty}_s}}{\theta_1}$$

which yields the result. □

Lemma 2. For any partition $\Lambda \subset \mathcal{T}$, denote by Λ^l the l-th generation partition such that $\Lambda^0 = \Lambda$ and Λ^{l+1} consists of the children of Λ^l. We first prove that $\lim_{l\to\infty} f^\ell_{\Lambda^l} = f$ in $L^2(\mathcal{M})$. Suppose $f \in L^\infty$. Notice that $\|f^\ell_{\Lambda^l} - f\| \le \|f^0_{\Lambda^l} - f\|$. As a result of the Lebesgue differentiation theorem, $f^0_{\Lambda^l} \to f$ almost everywhere. Since f is bounded, $f^0_{\Lambda^l}$ is uniformly bounded, hence $f^0_{\Lambda^l} \to f$ in $L^2(\mathcal{M})$ by the dominated convergence theorem. In the case where $f \in \mathcal{A}^\ell_t$, taking the uniform partition $\Lambda_{j(l)}$ at the coarsest scale of Λ^l, denoted by $j(l)$, we have $\|f - f^\ell_{\Lambda^l}\| \le \|f - f^\ell_{\Lambda_{j(l)}}\| \lesssim 2^{-j(l)t}$, and therefore $f^\ell_{\Lambda^l} \to f$ in $L^2(\mathcal{M})$.

Now, setting $\Lambda = \Lambda(\tau)$ and $\mathcal{S} := \mathcal{T}(\tau) \setminus \Lambda$, by Definitions 3 and 4 we get

$$\begin{aligned}
\|f^\ell_\Lambda - f\|^2 &= \left\|\sum_{l=0}^{L-1}\left(f^\ell_{\Lambda^l} - f^\ell_{\Lambda^{l+1}}\right) + f^\ell_{\Lambda^L} - f\right\|^2 = \left\|\sum_{l=0}^{\infty}(f^\ell_{\Lambda^l} - f^\ell_{\Lambda^{l+1}})\right\|^2 \\
&= \left\|\sum_{C_{j,k}\in\mathcal{T}\setminus\mathcal{S}} W^\ell_{j,k}\right\|^2 \le B_0 \sum_{C_{j,k}\in\mathcal{T}\setminus\mathcal{S}} \|W^\ell_{j,k}\|^2 \\
&= B_0 \sum_{l=0}^{\infty} \sum_{\Delta^\ell_{j,k}\in[2^{-(l+1)}\tau, 2^{-l}\tau)} \Delta^2_{j,k} \le B_0 \sum_{l=0}^{\infty} 2^{-2l}\tau^2 \#\mathcal{T}(2^{-(l+1)}\tau) \\
&= B_0 2^p \tau^{2-p} \sum_{l=0}^{\infty} 2^{-(2-p)l}|f|^p_{\mathcal{B}^\ell_s} \le B_0 2^p |f|^p_{\mathcal{B}^\ell_s}\tau^{2-p} \sum_{l=0}^{\infty} 2^{-(2-p)l},
\end{aligned}$$

which yields the first inequality in Lemma 2. The second inequality follows by observing that $2 - p = \frac{2s}{d}p$ and $|f|^p_{\mathcal{B}^\ell_s}\tau^{2-p} = |f|^2_{\mathcal{B}^\ell_s}(|f|^{-p}_{\mathcal{B}^\ell_s}\tau^p)^{\frac{2s}{d}} \le |f|^2_{\mathcal{B}^\ell_s}\#[\mathcal{T}(\tau)]^{-\frac{2s}{d}}$ by Definition 3. □

Lemma 3. See [34]. □

Lemma 4. (a). Thanks to [23, Theorem 3.2] and assumption (A5), we have

$$\|Q_{j,k} - \widehat{Q}_{j,k}\| = \|[\Sigma_{j,k}]_d^\dagger - [\widehat{\Sigma}_{j,k}]_d^\dagger\| \leq 3\frac{\|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\|}{(\lambda_d^{j,k}-\lambda_{d+1}^{j,k}-\|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\|)^2} \leq 3\frac{\|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\|}{\left(\frac{\theta_3}{2d}2^{-2j}-\|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\|\right)^2}.$$

Hence, the bound follows applying Lemma 3(d) with $t = \frac{\theta_3}{4d}2^{-2j}$.

(b). Observe that $\|\widehat{Q}_{j,k}\| \leq \|[\widehat{\Sigma}_{j,k}]_d^\dagger\| = (\widehat{\lambda}_d^{j,k})^{-1}$. Moreover, $\widehat{\lambda}_d^{j,k} \geq \lambda_d^{j,k} - |\lambda_d^{j,k} - \widehat{\lambda}_d^{j,k}| \geq \frac{\theta_3}{d}2^{-2j} - \|\Sigma_{j,k} - \widehat{\Sigma}_{j,k}\|$ by assumption (A5). Thus, using Lemma 3(d) with $t = \frac{\theta_3}{2d}2^{-2j}$ yields the result.

(c). We condition on the event that $\widehat{n}_{j,k} \geq \frac{1}{2}\mathbb{E}\widehat{n}_{j,k} = \frac{1}{2}n\rho(C_{j,k})$, whose complement occurs with probability lower than $2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right)$ by Lemma 3(b). The quantity $\|r_{j,k} - \widehat{r}_{j,k}\|$ is bounded by $A + B + C + D$ with

$$A := \left\|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^n \left(f(x_i)\begin{bmatrix} x_i - c_{j,k} \\ 2^{-j}\end{bmatrix} - \frac{1}{\rho(C_{j,k})}\int_{C_{j,k}} f(x)\begin{bmatrix} x - c_{j,k} \\ 2^{-j}\end{bmatrix} d\rho\right)\mathbf{1}_{j,k}(x_i)\right\|$$

$$B := \left\|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^n f(x_i)\begin{bmatrix} c_{j,k} - \widehat{c}_{j,k} \\ 0\end{bmatrix}\mathbf{1}_{j,k}(x_i)\right\|$$

$$C := \left\|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^n \zeta_i\begin{bmatrix} x_i - c_{j,k} \\ 2^{-j}\end{bmatrix}\mathbf{1}_{j,k}(x_i)\right\|$$

$$D := \left\|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^n \zeta_i\begin{bmatrix} c_{j,k} - \widehat{c}_{j,k} \\ 0\end{bmatrix}\mathbf{1}_{j,k}(x_i)\right\|.$$

Each term of the sum in A has expectation 0 and bound $2\langle\theta_2\rangle\|f\|_\infty 2^{-j}$. Thus, applying the Bernstein inequality [44, Corollary 7.3.2] we obtain

$$\mathbb{P}\{A > t\} \leq 8\exp\left(-c\frac{n\rho(C_{j,k})t^2}{\langle\theta_2\rangle^2\|f\|_\infty^2 2^{-2j} + \langle\theta_2\rangle\|f\|_\infty 2^{-j}t}\right).$$

B is bounded by $\|f\|_\infty\|c_{j,k} - \widehat{c}_{j,k}\|$ so that, using 3(c) with t replaced by $t/\|f\|_\infty$, we get

$$\mathbb{P}\{B > t\} \leq 2\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right) + 8\exp\left(-c\frac{n\rho(C_{j,k})t^2}{\theta_2^2\|f\|_\infty^2 2^{-2j} + \theta_2\|f\|_\infty 2^{-j}t}\right).$$

To estimate C we appeal to [7, Theorem 3.1, Remark 4.2]. For $X \in \mathbb{R}^n$, take $G(X) := \|MX\|$ with $M := \begin{bmatrix} x_1 - c_{j,k} & \dots & x_n - c_{j,k} \\ 2^{-j} & \dots & 2^{-j}\end{bmatrix}$. Then $|\partial_i G(X)| \leq \|x_i - c_{j,k}\| \leq \theta_2 2^{-j}$. Now let $X = (\zeta_1\mathbf{1}_{j,k}(x_1), \dots, \zeta_n\mathbf{1}_{j,k}(x_n))^T$, so that $C = G(X)/\widehat{n}_{j,k}$. Since the ζ_i's are independent, [7, Remark 4.2] applies, and it yields $\mathbb{P}\{G(X) > t\} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$, where $\sigma^2 = \sum_{i=1}^n \|\partial_i G\|_\infty^2\|\zeta_i\|_{\psi_2}^2\mathbf{1}_{j,k}(x_i) \leq \widehat{n}_{j,k}\theta_2^2 2^{-2j}\|\zeta\|_{\psi_2}^2$, and thus

$$\mathbb{P}\{C > t\} \leq 2\exp\left(-\frac{n\rho(C_{j,k})t^2}{2\theta_2^2\|\zeta\|_{\psi_2}^2 2^{-2j}}\right).$$

We are left with D. This term is smaller than $\|c_{j,k} - \widehat{c}_{j,k}\|\left|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^n \zeta_i\mathbf{1}_{j,k}(x_i)\right|$, where, by Lemma 3(c), $\|c_{j,k} - \widehat{c}_{j,k}\| \leq \theta_2 2^{-j}$ with probability higher than $1 - 10\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right)$. Hence, by the standard sub-Gaussian tail inequality [46, Proposition 5.10] we have

$$\mathbb{P}\{D > t\} \leq 10\exp\left(-\frac{3}{28}n\rho(C_{j,k})\right) + e\exp\left(-c\frac{n\rho(C_{j,k})t^2}{\theta_2^2\|\zeta\|_{\psi_2}^2 2^{-2j}}\right).$$

This completes the proof. □

Lemma 5. If $\ell = 0$, then $\|\widehat{f}^{\ell}_{j,k} - \widehat{f}^{\ell}_{j,k}\|_\infty = |y_{j,k} - \widehat{y}_{j,k}|$, which is less than

$$\left|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^{n}\left(f(x_i) - \frac{1}{\rho(C_{j,k})}\int_{C_{j,k}} f(x)d\rho(x)\right)\mathbf{1}_{j,k}(x_i)\right| + \left|\frac{1}{\widehat{n}_{j,k}}\sum_{i=1}^{n}\zeta_i\mathbf{1}_{j,k}(x_i)\right|.$$

Each addend in the first term has expectation 0 and bound $2\|f\|_\infty$, and therefore we can apply the standard Bernstein inequality [44, Theorem 1.6.1]. As for the second term, we use the standard sub-Gaussian tail inequality [46, Proposition 5.10]. This yields the bounds for $\ell = 0$.

For $\ell = 1$, we have

$$\begin{aligned}
&|f_{j,k}(x) - \widehat{f}_{j,k}(x)| \le |[(x - c_{j,k})^T\ 2^{-j}]^T Q_{j,k} r_{j,k} - [(x - \widehat{c}_{j,k})^T\ 2^{-j}]^T \widehat{Q}_{j,k}\widehat{r}_{j,k}| \\
&\le |\begin{bmatrix}(c_{j,k} - \widehat{c}_{j,k})^T & 0\end{bmatrix} Q_{j,k} r_{j,k}| + |\begin{bmatrix}(x - \widehat{c}_{j,k})^T & 2^{-j}\end{bmatrix}(Q_{j,k} r_{j,k} - \widehat{Q}_{j,k}\widehat{r}_{j,k})| \\
&\le \|c_{j,k} - \widehat{c}_{j,k}\|\,\|Q_{j,k}\|\,\|r_{j,k}\| + \|\begin{bmatrix}(x - \widehat{c}_{j,k})^T & 2^{-j}\end{bmatrix}\|\,\|Q_{j,k} r_{j,k} - \widehat{Q}_{j,k}\widehat{r}_{j,k}\| \\
&\le \|c_{j,k} - \widehat{c}_{j,k}\|\,\|Q_{j,k}\|\,\|r_{j,k}\| + \|\begin{bmatrix}(x - \widehat{c}_{j,k})^T & 2^{-j}\end{bmatrix}\|\,\left(\|Q_{j,k} - \widehat{Q}_{j,k}\|\,\|r_{j,k}\| + \|\widehat{Q}_{j,k}\|\|r_{j,k} - \widehat{r}_{j,k}\|\right) \\
&\lesssim \tfrac{\theta_2^2}{\theta_3}\left(d\|f\|_\infty 2^j\|c_{j,k} - \widehat{c}_{j,k}\| + d^2 M 2^{2j}\|\Sigma_{j,k} - \widehat{\Sigma}_{j,k}\| + d2^j\|r_{j,k} - \widehat{r}_{j,k}\|\right),
\end{aligned}$$

where the last inequality holds with high probability thanks to Lemma 4(a)(b). Thus, applying Lemma 3(c)(d) and Lemma 4(c) with t replaced by $\frac{t}{\theta d M 2^j}$, $\frac{t}{\theta d^2 M 2^{2j}}$ and $\frac{t}{\theta d 2^j}$, we obtain the desired result. □

Lemma 6. Follows simply by truncation. □

Lemma 7. We start with (a). Defining $\bar{\Delta}^{\ell}_{j,k} := \|W^{\ell}_{j,k}\|_n$ we have

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\Delta}^{\ell}_{j,k} < \eta \text{ and } \Delta^{\ell}_{j,k} \ge (2a_{\max} + 5)\eta\right\} \\
&\le \mathbb{P}\left\{\widehat{\Delta}^{\ell}_{j,k} < \eta \text{ and } \bar{\Delta}^{\ell}_{j,k} \ge (a_{\max} + 2)\eta\right\} + \mathbb{P}\left\{\bar{\Delta}^{\ell}_{j,k} < (a_{\max} + 2)\eta \text{ and } \Delta^{\ell}_{j,k} \ge (2a_{\max} + 5)\eta\right\} \\
&\le \mathbb{P}\left\{|\bar{\Delta}^{\ell}_{j,k} - \widehat{\Delta}^{\ell}_{j,k}| \ge (1 + a_{\max})\eta\right\} + \mathbb{P}\left\{|\Delta^{\ell}_{j,k} - 2\bar{\Delta}^{\ell}_{j,k}| \ge \eta\right\}.
\end{aligned}$$

The first quantity can be bounded by

$$\begin{aligned}
|\bar{\Delta}^{\ell}_{j,k} - \widehat{\Delta}^{\ell}_{j,k}| \le \|W^{\ell}_{j,k} - \widehat{W}^{\ell}_{j,k}\|_n &\le \|f^{\ell}_{j,k} - \widehat{f}^{\ell}_{j,k}\|_n + \sum_{C_{j+1,k'} \subset C_{j,k}} \|f^{\ell}_{j+1,k'} - \widehat{f}^{\ell}_{j+1,k'}\|_n \\
&\le \|f^{\ell}_{j,k} - \widehat{f}^{\ell}_{j,k}\|_\infty \sqrt{\widehat{\rho}(C_{j,k})} + \sum_{C_{j+1,k'} \subset C_{j,k}} \|f^{\ell}_{j+1,k'} - \widehat{f}^{\ell}_{j+1,k'}\|_\infty \sqrt{\widehat{\rho}(C_{j+1,k'})},
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{P}\left\{|\bar{\Delta}^{\ell}_{j,k} - \widehat{\Delta}^{\ell}_{j,k}| \ge (1 + a_{\max})\eta\right\} \\
&\le \mathbb{P}\left\{\|f^{\ell}_{j,k} - \widehat{f}^{\ell}_{j,k}\|_\infty \sqrt{\widehat{\rho}(C_{j,k})} \ge \eta\right\} + \sum_{C^{j+1,k'} \subset C_{j,k}} \mathbb{P}\left\{\|f^{\ell}_{j+1,k'} - \widehat{f}^{\ell}_{j+1,k'}\|_\infty \sqrt{\widehat{\rho}(C_{j+1,k'})} \ge \eta\right\}.
\end{aligned}$$

We now condition on the event that $|\rho(C_{j,k}) - \widehat{\rho}(C_{j,k})| \le \frac{1}{2}\rho(C_{j,k})$, which entails $\widehat{\rho}(C_{j,k}) \le \frac{3}{2}\rho(C_{j,k})$, and apply Lemma 5 with $t \lesssim \eta/\sqrt{\rho(Cjk)}$. The probability of the complementary event is bounded by

Lemma 3(b). To get rid of the remaining $\rho(C_{j,k})$'s inside the exponentials, we lower bound $\rho(C_{j,k})$ as follows. We have

$$(\Delta^\ell_{j,k})^2 \le 4\|f\|^2_\infty \rho(C_{j,k}). \tag{4}$$

Thus, $\Delta^\ell_{j,k} \ge (2a+5)\eta$ implies $\rho(C_{j,k}) \ge \frac{(2a+5)^2\eta^2}{4\|f\|^2_\infty}$. Therefore, we obtain that

$$\mathbb{P}\{|\bar{\Delta}^\ell_{j,k} - \widehat{\Delta}^\ell_{j,k}| \ge (1+a_{\max})\eta\} \le \begin{cases} C_0 \exp\left(-\frac{n\eta^2}{c_0 \max\{\|f\|^2_\infty, \|\zeta\|^2_{\psi_2}\}}\right) & (\ell = 0) \\ C_1 d \exp\left(-\frac{n\eta^2}{c_1 \max\{d^4\|f\|^2_\infty, d^2\|\zeta\|^2_{\psi_2}\}}\right) & (\ell = 1), \end{cases}$$

where C_0, c_0 depend on a, and C_1, c_1 depend on $a_{\max}, \theta_2, \theta_3$.

Next we estimate $\mathbb{P}\left\{\Delta^\ell_{j,k} - 2\bar{\Delta}^\ell_{j,k} \ge \eta\right\}$ by [21, Theorem 11.2]. Notice that for all $x \in \mathcal{M}$, $|W^\ell_{j,k}(x)| \lesssim \|f\|_\infty$. If $x \notin C_{j,k}$, then $W_{j,k}(x) = 0$, otherwise there is k' such that $x \in C_{j+1,k'} \subset C_{j,k}$. In such a case, $|W_{j,k}(x)| = |f_{j,k}(x) - f_{j+1,k'}(x)|$, and the claim follows from Lemma 6. Thus, [21, Theorem 11.2] gives us

$$\mathbb{P}\left\{\Delta^\ell_{j,k} - 2\bar{\Delta}^\ell_{j,k} \ge \eta\right\} \lesssim \exp\left(-\tfrac{n\eta^2}{c\|f\|^2_\infty}\right),$$

where c is an absolute constant.

Let us turn to (b). We first observe that

$$\widehat{\Delta}^\ell_{j,k} \lesssim M\sqrt{\widehat{\rho}(C_{j,k})}. \tag{5}$$

To see this, note again that $\widehat{W}_{j,k}(x) \ne 0$ only when $x \in C_{j+1,k'} \subset C_{j,k}$ for some k', in which case $|\widehat{W}_{j,k}(x)| = |\widehat{f}_{j,k}(x) - \widehat{f}_{j+1,k'}(x)|$ and we can apply Lemma 6. Now note that $b = 2a_{\max} + 5$. We have

$$\begin{aligned} \mathbb{P}\left\{\Delta^\ell_{j,k} < \eta \text{ and } \widehat{\Delta}^\ell_{j,k} \ge b\eta\right\} \le{}& \mathbb{P}\left\{\Delta^\ell_{j,k} < \eta,\ \widehat{\Delta}^\ell_{j,k} \ge b\eta \text{ and } \rho(C_{j,k}) \ge \frac{b^2\eta^2}{2\|f\|^2_\infty}\right\} \\ &+ \mathbb{P}\left\{\rho(C_{j,k}) < \frac{b^2\eta^2}{2\|f\|^2_\infty} \text{ and } \widehat{\rho}(C_{j,k}) \ge \frac{b^2\eta^2}{\|f\|^2_\infty}\right\} \\ &+ \mathbb{P}\left\{\widehat{\rho}(C_{j,k}) < \frac{b^2\eta^2}{\|f\|^2_\infty} \text{ given } \widehat{\Delta}^\ell_{j,k} \ge b\eta\right\}. \end{aligned}$$

The first probability can be estimated similarly to how we did for (a). Thanks to Lemma 3(a), the second probability is bounded by

$$\mathbb{P}\left\{\rho(C_{j,k}) < \frac{b^2\eta^2}{2\|f\|^2_\infty} \text{ and } |\widehat{\rho}(C_{j,k}) - \rho(C_{j,k})| > \frac{b^2\eta^2}{2\|f\|^2_\infty}\right\} \lesssim \exp\left(-\tfrac{b^2 n\eta^2}{c\|f\|^2_\infty}\right)$$

for an absolute constant c. Finally, the third probability is zero thanks to (5). □