Learning Adaptive Multiscale Approximations to Data and Functions near Low-Dimensional Sets

Wenjing Liao[§], Mauro Maggioni[§,†,♯], Stefano Vigogna[§]
[§]Department of Mathematics, [†]Electrical and Computer Engineering, [♯]Computer Science, Duke University
Durham, N.C., 27708, U.S.A.
Email: wjliao,mauro,stefano@math.duke.edu

***Abstract*—In the setting where a data set in $\mathbb{R}^D$ consists of samples from a probability measure ρ concentrated on or near an unknown d-dimensional set $\mathcal{M}$, with D large but $d \ll D$, we consider two sets of problems: geometric approximation of $\mathcal{M}$ and regression of a function f on $\mathcal{M}$. In the first case we construct multiscale low-dimensional empirical approximations of $\mathcal{M}$, which are adaptive when $\mathcal{M}$ has geometric regularity that may vary at different locations and scales, and give performance guarantees. In the second case we exploit these empirical geometric approximations to construct multiscale approximations to f on $\mathcal{M}$, which adapt to the unknown regularity of f even when this varies at different scales and locations. We prove guarantees showing that we attain the same learning rates as if f was defined on a Euclidean domain of dimension d, instead of an unknown manifold $\mathcal{M}$. All algorithms have complexity $O(n \log n)$, with constants scaling linearly in D and exponentially in d.**

I. Introduction

We model a data as n i.i.d. samples $X_n := \{x_i\}_{i=1}^n$ from a probability measure ρ in $\mathbb{R}^D$. Broadly speaking, the well-known curse of dimensionality implies that, in many statistical learning and inference tasks, the required sample size n must satisfy $n \gtrsim \varepsilon^{-D}$ in order to achieve accuracy ε, unless further assumptions are made on ρ. The assumptions we make here are that ρ is supported on or near a (somewhat regular) set $\mathcal{M}$ of dimension $d \ll D$. We consider two statistical learning problems, given X_n: (I) **geometric learning** of approximations to the underlying geometric structure $\mathcal{M}$; (II) **regression on $\mathcal{M}$**: if values $Y_n := \{y_i = f(x_i)+\eta_i\}_{i=1}^n$ are also provided (η representing noise), estimate the function f on $\mathcal{M}$. In both cases we seek estimators that have sample requirements $O(\varepsilon^{-d})$ (up to log factors), adapt to a large family of geometric structures $\mathcal{M}$ (not necessarily smooth manifolds) and functions with minimal information about their regularity, are robust to noise in the values y_i's, and are implemented by algorithms with cost $O(n \log n)$.

We will tackle both problems using *multiscale techniques*, that have their roots in geometric measure theory, for (I), and harmonic analysis and approximation theory, for (II). Our main tool for tackling (I) will be an extension of Geometric Multi-Resolution Analysis (GMRA) [1] to adaptively construct approximations $\widehat{\mathcal{M}}$ to $\mathcal{M}$. This is a multiscale geometric approximation scheme for sets of points in high-dimensions that concentrate near low-dimensional sets. We extend here the recent work [2] to a larger class of geometric objects $\mathcal{M}$, which are allowed to have singularities or different regularity at different scales and locations, therefore exploiting the full force of multiscale schemes. We then tackle (II) using GMRA by both learning a low-dimensional approximation $\widehat{\mathcal{M}}$ to $\mathcal{M}$ and simultaneously performing an adaptive multiscale regression scheme on $\widehat{\mathcal{M}}$, inspired by wavelet/multiscale regression [3]. We obtain estimators that adapt to the unknown regularity of f for a large class of f's that may have different, unknown regularity at different scales and locations. We also present numerical evidence on the performance of all the algorithms, as well as evidence showing that not adapting to the manifold leads to worse results as soon as the data is noisy.

The authors are grateful for support from AFOSR FA9550-14-1-0033, NSF ATD-1222567 and ONR N00014-12-1-0601.

II. Multiscale Geometric Approximations

Stated in geometric terms, the simplest and most classical geometric assumption on high-dimensional data is that points are near a single d-dimensional plane, with $d \ll D$. For this model Principal Component Analysis (PCA) is an effective and robust tool to estimate the underlying plane. More generally, one may assume that data lie on a union of several low-dimensional planes, or perhaps on a low-dimensional manifold. In this case, one may use approaches inspired by quantization, for example using K-means to find $\{c_l\}_{l=1}^K \subset \mathbb{R}^D$ that best approximate the data by minimizing $\frac{1}{n}\sum_{i=1}^n \|x_i - c_{l_i}\|^2$, where $l_i := \operatorname{argmin}_{l=1,\dots,K} \|c_l - x_i\|$. Higher-order quantization schemes are also possible, for example one may seek K planes of dimension d that best approximate the data in the sense that they minimize $\min_{S\in\mathcal{F}_{K,d}} \frac{1}{n}\sum_{i=1}^n \operatorname{dist}^2(x_i, S)$, where $\mathcal{F}_{K,d}$ is the collection of sets of K planes of dimension d, and $\operatorname{dist}(x, S) = \min_{y\in S} \|x - y\|$. The global minimizer of these non-convex optimization problems are typically hard to compute, and it is well-known that EM-type algorithms are prone to find only local minima that are significantly worse than the optimal ones (see [2] for a discussion and references).

Instead of solving a hard optimization problem, we pursue a multiscale strategy: this not only leads to strong performance guarantees on large classes of geometric structures, but also fast algorithms (besides, we would argue, more insight into the structure of data). Our method is based on an adaptive version of GMRA [1]. Assume the probability measure ρ is supported on a compact d-dimensional Riemannian manifold

$\mathcal{M} \hookrightarrow \mathbb{R}^D$ ($d \geq 3$). Let s be a regularity parameter of ρ to be defined below (see "model class $\boldsymbol{\mathcal{B}_s}$"). For a given accuracy ε, if $n \gtrsim (1/\varepsilon)^{\frac{2s+d-2}{s}} \log(1/\varepsilon)$ samples are available, then adaptive GMRA outputs a dictionary $\widehat{\Phi}_\varepsilon = \{\widehat{\phi}_i\}_{i \in \mathcal{J}_\varepsilon}$ ($\mathcal{J}_\varepsilon$ an index set), an encoding operator $\widehat{\mathcal{D}}_\varepsilon : \mathbb{R}^D \to \mathbb{R}^{\mathcal{J}_\varepsilon}$ and a decoding operator $\widehat{\mathcal{D}}_\varepsilon^{-1} : \mathbb{R}^{\mathcal{J}_\varepsilon} \to \mathbb{R}^D$. With high probability, these objects are s.t.: for every $x \in \mathbb{R}^D$, $\|\widehat{\mathcal{D}}_\varepsilon x\|_0 \leq d+1$ (i.e. only $d+1$ entries are non-zero), and accuracy is guaranteed in the sense that

$$\text{MSE} := \mathbb{E}_{x \sim \rho}[\|x - \widehat{\mathcal{D}}_\varepsilon^{-1} \widehat{\mathcal{D}}_\varepsilon x\|^2] \lesssim \varepsilon^2 \,. \tag{1}$$

As for the computational complexity, constructing the dictionary $\widehat{\Phi}_\varepsilon$ takes $O((C^d + d^2) D \varepsilon^{-\frac{2s+d-2}{s}} \log(1/\varepsilon))$, where C is a constant, and computing $\widehat{\mathcal{D}}_\varepsilon x$ takes $O(d(D+d^2)\log(1/\varepsilon))$. We stated this result in terms of encoding and decoding to stress that learning the geometry in fact yields efficient representations of data, which may be used for transmitting it over a channel with limited capacity, or for performing signal processing or statistical tasks in a domain where the data admits a sparse representation (e.g. in compressed sensing and/or estimation problems [4], [5]). While we stated the results, including the computational complexity, in terms of the requested accuracy ε, there are equivalent formulations in terms of rates of approximation as a function of n, and we will use the latter format for the rest of the paper. The reader may keep in mind that all the results may be interpreted as performing dictionary learning, compression, denoising and providing a sparsifying transform for the data [2].

A. Geometric Multi-Resolution Analysis

GMRA involves a few steps, detailed in [1]:

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

(ii) perform **local PCA** on each $C^{j,k}$: let $c^{j,k}$ be the center and $V^{j,k}$ the d-dim principal subspace of $C^{j,k}$. Define $\mathcal{P}^{j,k}(x) := c^{j,k} + \text{Proj}_{V^{j,k}}(x - c^{j,k})$, where Proj_V is the orthogonal projection onto V;

(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 will not be used here).

$\mathcal{M}$ may be approximated, at each scale j, by its projection onto the family of linear sets $\{\mathcal{P}^{j,k}(C^{j,k})\}_{k \in \mathcal{K}_j}$. Note that, in terms of encoding, for $x \in C^{j,k}$, $\mathcal{P}^{j,k}(x)$ may be represented by a vector in $\mathbb{R}^d$, defining $\mathcal{D}_\varepsilon(x)$, while $\mathcal{D}_\varepsilon^{-1}(\mathcal{D}_\varepsilon(x)) := \mathcal{P}^{j,k}(x)$. Since all the $C^{j,k}$'s at scale j have roughly the same size, we call $\Lambda^j := \{C^{j,k}\}_{k \in \mathcal{K}_j}$ a uniform partition at scale j, and $\{\mathcal{P}^{j,k}(C^{j,k})\}_{k \in \mathcal{K}_j}$ a uniform approximation at scale j. For example, uniform approximations for the S and Z manifold at scale $j = 8$ are shown at the top line of Figure 1.

When only training data X_n is given, and $\mathcal{M}$ is unknown, the construction above is carried over on X_n and its result is random with the samples: the result cited above states that the algorithm will succeed with high probability, providing sparse yet accurate representations for any sample from ρ (i.e. on test data), as measured by MSE (1).

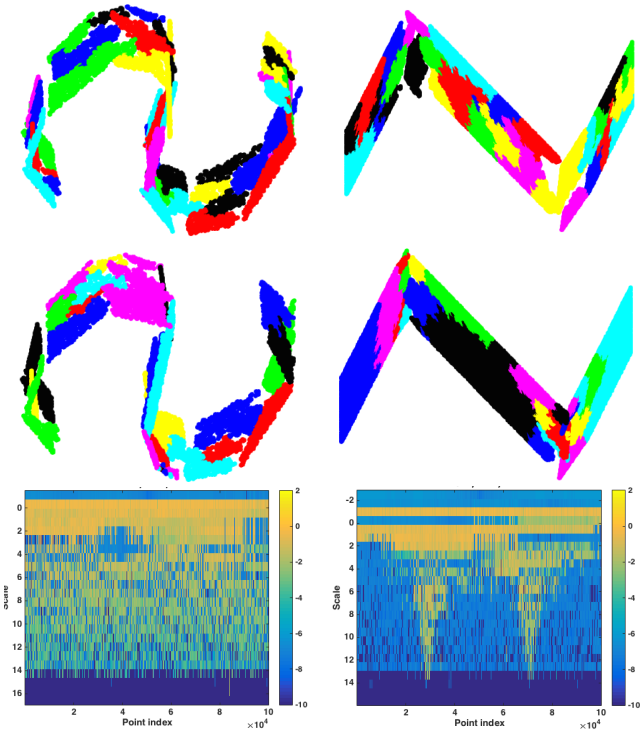

Figure 1: Top line: uniform approximation at scale $j = 8$ for the S and the Z manifold. Center line: adaptive approximation. Bottom line: $\log_{10} \|\mathcal{P}_{\Lambda^j} x_i - \mathcal{P}_{\Lambda^{j+1}} x_i\|$ from coarse scale (top) to finest scale (bottom), with rows indexed by j and columns indexed by points x_i, sorted from left to right on the manifolds.

We write $f \lesssim g$ if $\exists C$ independent of the parameters in f and g such that $f \leq Cg$. $f \asymp g$ means $f \lesssim g$ and $g \lesssim f$.

1. Multiscale tree decomposition of data. A tree decomposition of $\mathcal{M}$ with respect to the distribution ρ is a family $\{C^{j,k}\}_{k \in \mathcal{K}_j, j \in \mathbb{Z}}$, $C^{j,k} \subset \mathcal{M}$, satisfying certain technical conditions that here for brevity we only sketch (see A1-5 in [6]): $\{C^{j,k}\}_{k \in \mathcal{K}_j, j \in \mathbb{Z}}$ forms a **tree structure** $\mathcal{T}$, with j equal to the distance to the root, $\rho(C^{j,k}) \gtrsim 2^{-jd}$, $\text{diam}(C^{j,k}) \lesssim 2^{-j}$, and finally the covariance matrix $\Sigma^{j,k}$ of ρ restricted to $C^{j,k}$ has d singular values comparable to $2^{-2j}/d$, with the others at least a factor smaller. The constants entering in the bounds above will be denoted by $\boldsymbol{\vartheta}$. The tree $\mathcal{T}$ is not given: an empirical tree $\mathcal{T}^n$ whose each leaf contains at least d points and a family of $C_{j,k}$'s are constructed from data using a variation of the cover tree algorithm [7], so that the empirical tree w.h.p. satisfies the properties above, at least if j is not too large. This construction is possible not only when $\mathcal{M}$ is a d-dimensional compact manifold, but also when it is a tube around a manifold (see [1], [2], [6] for details).

2. Low-dimensional, multiscale projections and approximations. Let $\widehat{n}^{j,k}$ be the number of points on $C^{j,k}$, $\widehat{c}^{j,k} := \frac{1}{\widehat{n}^{j,k}} \sum_{i=1}^n x_i \mathbf{1}_{C^{j,k}}(x_i)$ the empirical conditional mean on $C^{j,k}$, and $\widehat{V}^{j,k}$ the eigen-space corresponding to the largest d eigenvalues of the empirical conditional covariance matrix $\widehat{\Sigma}^{j,k} := \frac{1}{\widehat{n}^{j,k}} \sum_{i=1}^n (x_i - \widehat{c}^{j,k})(x_i - \widehat{c}^{j,k})^T \mathbf{1}_{C^{j,k}}(x_i)$. The

empirical affine projectors $\{\widehat{\mathcal{P}}^{j,k}\}_{k\in\mathcal{K}_j, j\in\mathbb{Z}}$ are constructed on the empirical tree $\mathcal{T}^n$ from the empirical centers $\widehat{c}^{j,k}$ and the eigenvectors of $\widehat{\Sigma}^{j,k}$. The empirical counterpart of the uniform approximations of $\mathcal{M}$ is given by a collection of piecewise affine projectors $\{\widehat{\mathcal{P}}_{\Lambda^j} : \mathbb{R}^D \to \mathbb{R}^D\}_{j\in\mathbb{Z}}$, where $\widehat{\mathcal{P}}_{\Lambda^j} := \sum_{k\in\mathcal{K}_j} \widehat{\mathcal{P}}^{j,k}\mathbf{1}_{C^{j,k}}$. Selecting an optimal scale $j^\star$ based on bias-variance tradeoff as in [2], [6] is possible but leads to results that are not fully satisfactory, for two reasons: (i) the knowledge of the regularity of $\mathcal{M}$ is required in order to choose the optimal scale $j^\star$; (ii) none of the uniform partitions $\{\Lambda^j\}_{j\in\mathbb{Z}}$ will be optimal if the regularity and the curvature of ρ vary from location to location. For example, 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 1).

B. Learning adaptive geometric approximations

Inspired by the adaptive methods in classical multiresolution analysis [3], [8], [9], we propose an adaptive version of GMRA which will automatically adapt to the regularity of $\mathcal{M}$ and learn adaptive, near-optimal approximations. Let

$$(\widehat{\Delta}^{j,k})^2 := \frac{1}{n}\sum_{i=1}^n \left\|(\widehat{\mathcal{P}}_{\Lambda^j} - \widehat{\mathcal{P}}_{\Lambda^{j+1}})\mathbf{1}_{C^{j,k}}(x_i)\right\|^2 . \quad (2)$$

$\widehat{\Delta}^{j,k}$ measures the change in approximation in passing from $C^{j,k}$ to its children. In other words, it is an empirical measure of improvement in the quality of approximation. We expect $\widehat{\Delta}^{j,k}$ to be small on approximately flat regions and large at corners. We see this phenomenon represented in Figure 1: as j increases, for the S manifold $\|\widehat{\mathcal{P}}_{\Lambda^j}x_i - \widehat{\mathcal{P}}_{\Lambda^{j+1}}x_i\|$ decays uniformly at all points, while for the Z manifold, the same quantity decays rapidly on flat regions but remains large to fine scales around the corners. We wish to include in our approximation nodes where this quantity is large, since we may expect a large improvement in approximation (bias) by including such nodes. However, if too few samples exist in a node, then this quantity is not to be trusted, as its variance is large. We consider the following criterion for determining a partition for our estimator: 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 s (see Theorem 1). Empirical adaptive GMRA returns piecewise affine projectors on $\widehat{\Lambda}_{\tau_n}$, the partition associated with the outer leaves (the children of the leaves) of $\widehat{\mathcal{T}}_{\tau_n}$. Adaptive partitions of the teapot, armadillo and dragon with a fixed κ are displayed in Figure 2, where every cell is colored by its scale. They match our expectation that cells at irregular locations are selected at finer scales than cells at "flat" locations.

We introduce a large model class $\mathcal{B}_s$ modeling the situation above, for which the performance guarantee of adaptive GMRA will be provided. Let $\Delta^{j,k}$ be the analog of $\widehat{\Delta}^{j,k}$ computed from the distribution ρ (also obtained in the limit of infinite samples). Given any fixed threshold $\eta > 0$ and tree $\mathcal{T}$, 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 $\Delta^{j,k} \geq 2^{-j}\eta$. The model class $\mathcal{B}_s$ imposes on its members a quantitative bound on the size of the truncated tree $\mathcal{T}_{(\rho,\eta)}$ as $\eta \to 0^+$: for $d \geq 3$, given $s > 0$, ρ supported on $\mathcal{M}$ is in the **geometric model class $\mathcal{B}_s$** if

$$|\rho|^p_{\mathcal{B}_s} := \sup_{\mathcal{T}} \sup_{\eta>0} \eta^p \sum_{j\in\mathbb{Z}} 2^{-2j} \#_j \mathcal{T}_{(\rho,\eta)} < \infty\,, \quad p = \tfrac{2(d-2)}{2s+d-2},$$

where $\#_j\mathcal{T}_{(\rho,\eta)}$ is the cardinality of $\mathcal{T}_{(\rho,\eta)}$ at scale j, along with another technical assumption omitted here for lack of space (see [6] for details). Here $\mathcal{T}$ ranges over the set, assumed nonempty, of tree decompositions satisfying the assumptions above. Note that $\mathcal{B}_s \subset \mathcal{B}_{s'}$ for $s \geq s'$.

Example: the volume measure on the d-dim S manifold, where x_1, x_2 are on the S curve and $x_i \in [0,1], i = 3,\ldots,d+1$, is in $\mathcal{B}_2$. The same holds for the volume measure on any smooth compact Riemannian manifold. However, the volume measure on the d-dim Z manifold, $d \geq 3$, is in $\mathcal{B}_s$ with $s = 3(d-2)/2(d-3)$.

We prove the following estimate on the L^2 approximation error of $\mathcal{M}$, $\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}X\|^2 := \int_{\mathcal{M}} \|x - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}x\|^2 d\rho$:

Theorem 1: Let $d \geq 3$ and $\nu > 0$. There exists $\kappa_0(\vartheta, d, \nu)$ such that: if $\kappa \geq \kappa_0$, $\rho \in \mathcal{B}_s$ for some $s > 0$ and $\tau_n = \kappa\sqrt{(\log n)/n}$, then there are $c_1(\vartheta, d, s, |\rho|_{\mathcal{B}_s}, \kappa, \nu)$ and

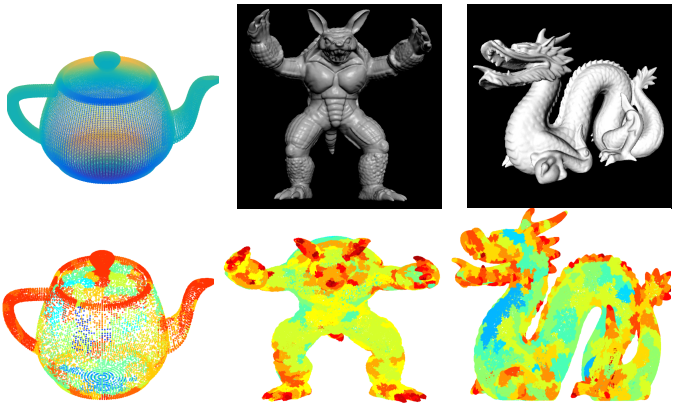

Figure 2: Top line: 3D shapes; bottom line: adaptive partitions in which every cell is colored by its numeric scale ($-\log_{10}$radius). The colors representing coarse to fine scales are ordered as: blue $\to$ green $\to$ yellow $\to$ red. It is noticeable that cells at irregular locations are selected at finer scales than cells at "flat" locations.

Algorithm 1 Adaptive GMRA

Input: X_n : data, κ : threshold
Output: $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}}$: adaptive piecewise linear projectors

1: Construct $\mathcal{T}^n$ and $\{C_{j,k}\}$.
2: 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 $C^{j,k} \in \mathcal{T}^n : \widehat{\Delta}^{j,k} \geq 2^{-j}\tau_n,\ \tau_n = \kappa\sqrt{(\log n)/n}$.
4: $\widehat{\Lambda}_{\tau_n} \leftarrow$ partition associated with outer leaves of $\widehat{\mathcal{T}}_{\tau_n}$
5: $\widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} \leftarrow \sum_{C^{j,k}\in\widehat{\Lambda}_{\tau_n}} \widehat{\mathcal{P}}^{j,k}\mathbf{1}_{C^{j,k}}$.

$c_2(\boldsymbol{\vartheta}, d, s, |\rho|_{\mathcal{B}_s}, \kappa)$ such that

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

and therefore MSE $:= \mathbb{E}\|X - \widehat{\mathcal{P}}_{\widehat{\Lambda}_{\tau_n}} X\|^2 \lesssim (\log n/n)^{\frac{2s}{2s+d-2}}$.

For $d = 1, 2$, we can prove that MSE$\lesssim \log^d n/n$ under weak assumptions [6]. Theorem 1 is satisfactory in two aspects: (i) when $d \geq 3$, the rate for MSE is obtained for a large model class $\mathcal{B}_s$; (ii) the algorithm is adaptive in that it does not require knowledge of the regularity of $\mathcal{M}$, since the choice of κ is independent of s. For the dependency of the subsumed constants on the parameters see [6].

III. Multiscale Regression on $\mathcal{M}$

We now consider the problem of learning a function on $\mathcal{M}$. Let ρ be an unknown probability distribution on the product space $\mathcal{M} \times \mathbb{R}$, where $\mathcal{M}$ is a compact d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. Given n independent samples $\{(x_i, y_i)\}_{i=1}^n$ drawn from ρ, we construct an estimator for the regression function $f_\rho(x) := \int_Y y \ d\rho(y|x)$, where $\rho(y|x)$ is the conditional probability measure on $\mathbb{R}$ with respect to x. Let us assume that $|y| \leq M$ ρ-a.s.. We denote by $\rho_{\mathcal{M}}$ the marginal probability measure on $\mathcal{M}$. Note that this covers the case where $y_i = f(x_i) + \eta_i$, with η_i bounded i.i.d. random variables representing noise, independent of x_i[1].

We construct estimators for f_ρ in two steps: first we learn an approximation to $\mathcal{M}$ using on the x_i's the geometric learning outlined above, and then we construct a polynomial estimator for $f_\rho|_{C^{j,k}}$ by learning such polynomial on the given data restricted to $x_i \in C^{j,k}$, projected on $V^{j,k}$. For brevity, we consider here only theorems for piecewise linear polynomials.

For a partition Λ compatible with the tree used to construct GMRA, we construct the estimator $\widehat{f}_\Lambda := \sum_{C^{j,k} \in \Lambda} \widehat{f}^{j,k} \mathbf{1}_{C^{j,k}}$, where $\widehat{f}^{j,k}$ is the least squares fit to the data $\{(\widehat{\mathcal{P}}^{j,k}(x_i), y_i)\}_{x_i \in C^{j,k}}$ among all the linear polynomials $V^{j,k} \to \mathbb{R}$. At scale j, f_ρ is approximated by $\widehat{f}_{\Lambda^j}$. While one could choose the optimal scale $j^\star$ judiciously in order to obtain estimates on $\|\widehat{f}_{\Lambda^{j^\star}} - f_\rho\|_{L^2(\rho_{\mathcal{M}})}$ with high probability [10], such choice of $j^\star$ would depend on the knowledge of a suitably-defined smoothness index of f_ρ. Moreover, this estimator would not capture the variability of f_ρ when f_ρ is not uniformly regular.

For functions that exhibit different regularity at different locations/scales, we need to choose adaptive partitions, ideally in such a way that will not require us to know much about the regularity of f_ρ or the variations of regularity. We propose an adaptive multiscale estimator, inspired by [3]. Let

$$(\widehat{W}^{j,k})^2 := 1/n \sum_{i=1}^n \left|(\widehat{f}_{\Lambda^j} - \widehat{f}_{\Lambda^{j+1}})\mathbf{1}_{C^{j,k}}(x_i)\right|^2 .$$

Like its geometric counterpart $\widehat{\Delta}^{j,k}$ in (2), $\widehat{W}^{j,k}$ measures the variation from the estimator on $C_{j,k}$ to that on the children of $C^{j,k}$. Our adaptive partition is selected according to Algorithm 2. Let $W^{j,k}$ be the analog of $\widehat{W}^{j,k}$ computed from ρ. Given any fixed threshold $\eta > 0$, we let $\mathcal{T}_{(f,\eta)}$ be the smallest proper tree of $\mathcal{T}$ that contains all $C^{j,k} \in \mathcal{T}$ satisfying $W^{j,k} \geq \eta$. We define the **functional model class $\mathcal{B}_s$** for $s > 0$ as follows: a function $f : \mathcal{M} \to \mathbb{R}$ is in $\mathcal{B}_s$ if

$$|f|^p_{\mathcal{B}_s} := \sup_{\mathcal{T}} \sup_{\eta > 0} \eta^p \# \mathcal{T}_{(f,\eta)} < \infty \ , \ \ p = 2d/(2s + d) \, ,$$

where $\#\mathcal{T}_{(f,\eta)}$ is the cardinality of $\mathcal{T}_{(f,\eta)}$, along with another technical assumption omitted here. We prove:

Theorem 2: Let $\nu > 0$. There exists $\kappa_0(\boldsymbol{\vartheta}, d, M, \nu)$ such that: if $\kappa \geq \kappa_0$, $f_\rho \in \mathcal{B}_s$ for some $s > 0$ and $\tau_n = \kappa\sqrt{(\log n)/n}$, then there are $c_1(\boldsymbol{\vartheta}, d, M, s, |f_\rho|_{\mathcal{B}_s}, \kappa, \nu)$ and $c_2(\boldsymbol{\vartheta}, d, M, s, |f_\rho|_{\mathcal{B}_s}, \kappa)$ such that

$$\mathbb{P}\left\{\|f_\rho - \widehat{f}_{\widehat{\Lambda}_{\tau_n}}\|_{L^2(\rho_{\mathcal{M}})} \geq c_1 \left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}\right\} \leq c_2 n^{-\nu} \ , \quad (3)$$

and therefore MSE$= \mathbb{E}\|f_\rho - \widehat{f}_{\widehat{\Lambda}_{\tau_n}}\|^2_{L^2(\rho_{\mathcal{M}})} \lesssim (\log n/n)^{\frac{2s}{2s+d}}$.

The estimator is adaptive: no knowledge of the regularity of f_ρ is required, since κ is independent of s.

Computational complexity. Both adaptive GMRA and multiscale regression may be implemented with algorithms with cost at most $O((C_d + d^2)Dn \log n)$ on n points, with C_d exponential in the intrinsic dimension d.

Algorithm 2 Adaptive Regression

Input: (X_n, Y_n) : data and function values, κ : threshold
Output: $\widehat{f}_{\widehat{\Lambda}_{\tau_n}}$: adaptive piecewise polynomial estimator

1: Perform GMRA on X_n to obtain $\mathcal{T}^n$ and $\{C_{j,k}\}$
2: $\widehat{f}^{j,k} \leftarrow$ local least squares polynomial fit on $C^{j,k} \in \mathcal{T}^n$
3: $\widehat{W}^{j,k} \leftarrow$ local refinement criterion on $C^{j,k} \in \mathcal{T}^n$
4: $\widehat{\mathcal{T}}_{\tau_n} \leftarrow$ smallest proper subtree of $\mathcal{T}^n$ containing $C^{j,k} \in \mathcal{T}^n$ s.t. $\widehat{W}^{j,k} \geq \tau_n$ where $\tau_n = \kappa\sqrt{(\log n)/n}$
5: $\widehat{\Lambda}_{\tau_n} \leftarrow$ partition associated to outer leaves of $\widehat{\mathcal{T}}_{\tau_n}$
6: $\widehat{f}_{\widehat{\Lambda}_{\tau_n}} \leftarrow \sum_{C^{j,k} \in \widehat{\Lambda}_{\tau_n}} \widehat{f}^{j,k} \cdot \mathbf{1}_{C^{j,k}}$

IV. Examples

GMRA and adaptive GMRA. The performance of GMRA and adaptive GMRA is tested on the samples $\{x_i\}_{i=1}^n$ on the 4-dim S and Z manifold embedded in $\mathbb{R}^{20}$. We split the n points evenly to training data, which are used for the constructions of GMRA and adaptive GMRA, and test data, for the evaluation of approximation error. 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 $\xi_i \sim \mathcal{N}(0, I_{D\times D})$. Test data are noise-free, so test data error below the noise level implies that we are denoising the data. In the left column of Figure 3, we set the noise level $\sigma = 0.05$ and display the log-log plot of the approximation error (averaged over 5 trails) with respect to the sample size n for empirical GMRA. For uniform GMRA the scale $j^\star$ is chosen optimally (see [10]): $2^{-j^\star} = (\log n/n)^{\frac{1}{2\gamma+d-2}}$, where $d = 4$, $\gamma = 2$ for the S manifold and $\gamma = 1.5$ for the Z manifold. For adaptive GMRA we choose $\kappa \in \{0.5, 1\}$.

[1]Our results extend, up to $\log n$ factors, to unbounded types of noise (e.g. sub-Gaussian) [10].

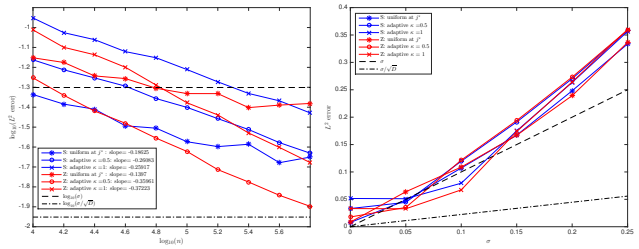

Figure 3: Left: L^2 error versus sample size, for both S and Z manifolds of dimension $d = 4$, and uniform and adaptive GMRA (for $\kappa = 0.5, 1$). Note the denoising effect. Right: L^2 error versus σ, for the same data sets, measuring robustness with respect to noise.

The slope (determined by least squared fit) is the rate of convergence: L^2 error $\sim n^{\text{slope}}$. Adaptive GMRA yields a faster rate of convergence than GMRA for the Z manifold. We note a denoising 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 independent of κ, as predicted by Theorem 1.

Multiscale regression. For multiscale regression we consider the S manifold with $d = 3$, $D = 20$ and $n \in \{20000, 200000\}$, evenly split in training and test data. Let $y = f(x) = 1/\|x - x_0\|$, with the pole x_0 having distance 0.1 from $\mathcal{M}$. We add Gaussian noise to y with $\sigma_Y = 0.05$. We also consider the case, which is beyond the scope of the Theorem, where X_n is also corrupted by Gaussian noise, with standard deviation $\sigma_X = 0.05/\sqrt{D}$. We consider piecewise polynomials of degree $0, 1, 2$, and we look at the performance as a function of the size of the partition picked, with both uniform and adaptive partitions. We compare our methods with standard regression techniques, in particular average of k-NN ($k = 1, ..., 20$, best results on test data reported), CART (Matlab implementation, with default parameters, cross-validation), and NystromCoRE [11] (best results on test data over 10 well-chosen Gaussian kernel width parameters, other parameters are defaults and optimized by cross-validation). We also compare with piecewise polynomial regression with the same partitions, but without projecting the data in $C^{j,k}$ onto $V^{j,k}$. It was suggested in [12] that there is no need to try to adapt to $\mathcal{M}$, but our experiments show otherwise, at least in the multiscale setting: especially when noise is added to the data, not performing the local projections leads to unstable and worse-performing estimators. This does not contradict the results in [12] about the noiseless case (see [10] for further discussion).

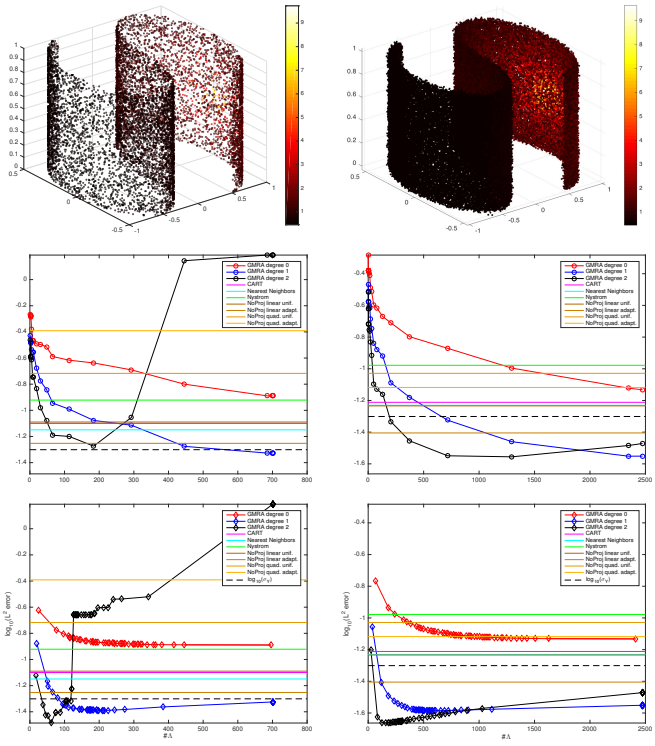

Figure 4: Left column: top: function on S manifold with $d = 3$, $D = 20$, $n = 20,000$, $\sigma_X = 0$, $\sigma_Y = 0.05$; middle and bottom: approximation error with uniform (resp. adaptive) partitions for our estimator, with local polynomials of degree $0, 1, 2$, as a function of the partition size. Adaptive approximation is consistently better and uses smaller partitions. Right column: as in the first column, but with $n = 200,000$ and $\sigma_X = 0.05$. Not projecting onto $V_{j,k}$ is unstable and produces worse estimators (denoted by NoProj, the local linear and quadradic best performing on the test data shown).

References

[1] 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*, vol. 32, no. 3, pp. 435–462, 2012, (submitted:5/2011).

[2] M. Maggioni, S. Minsker, and N. Strawn, "Dictionary learning and non-asymptotic bounds for the Geometric Multi-Resolution Analysis," *to appear in Journal of Machine Learning Research*, 2015.

[3] P. Binev, A. Cohen, W. Dahmen, R. A. Devore, and V. N. Temlyakov, "Universal algorithms for learning theory part i: piecewise constant functions," *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 1297–1321, 2005.

[4] M. A. Iwen and M. Maggioni, "Approximation of points on low-dimensional manifolds via random linear projections," *Inference & Information*, vol. 2, no. 1, pp. 1–31, 2013, arXiv:1204.3337v1, 2012.

[5] 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*, 2012, pp. 1–6.

[6] W. Liao and M. Maggioni, "Robust adaptive Geometric Multi-Resolution Analysis for data in high-dimensions," *in preparation*, 2016.

[7] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," *Proceedings of the 23rd international conference on Machine learning*, 2006.

[8] P. Binev, A. Cohen, W. Dahmen, and R. A. DeVore, "Universal algorithms for learning theory part ii: Piecewise polynomial functions," *Constructive Approximation*, vol. 26, no. 2, pp. 127–152, 2007.

[9] 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*, vol. 48, no. 7, pp. 1895–1921, 2002.

[10] W. Liao, M. Maggioni, and S. Vigogna, "Multiscale regression on manifolds," *in preparation*, 2016.

[11] A. Rudi, R. Camoriano, and L. Rosasco, "Less is more: Nyström computational regularization," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 1648–1656.

[12] P. J. Bickel and B. Li, "Local polynomial regression on unknown manifolds," *Lecture Notes-Monograph Series*, pp. 177–186, 2007.