A spectral series approach to high-dimensional nonparametric regression

Ann B. Lee[*]

Department of Statistics, Carnegie Mellon University, USA
e-mail: annlee@cmu.edu

and

Rafael Izbicki

Department of Statistics, Federal University of São Carlos, Brazil
e-mail: rafaelizbicki@gmail.com

Abstract: A key question in modern statistics is how to make fast and reliable inferences for complex, high-dimensional data. While there has been much interest in sparse techniques, current methods do not generalize well to data with nonlinear structure. In this work, we present an orthogonal series estimator for predictors that are complex aggregate objects, such as natural images, galaxy spectra, trajectories, and movies. Our series approach ties together ideas from manifold learning, kernel machine learning, and Fourier methods. We expand the unknown regression on the data in terms of the eigenfunctions of a kernel-based operator, and we take advantage of orthogonality of the basis with respect to the underlying data distribution, P, to speed up computations and tuning of parameters. If the kernel is appropriately chosen, then the eigenfunctions adapt to the intrinsic geometry and dimension of the data. We provide theoretical guarantees for a radial kernel with varying bandwidth, and we relate smoothness of the regression function with respect to P to sparsity in the eigenbasis. Finally, using simulated and real-world data, we systematically compare the performance of the spectral series approach with classical kernel smoothing, k-nearest neighbors regression, kernel ridge regression, and state-of-the-art manifold and local regression methods.

1. Introduction

A challenging problem in modern statistics is how to handle complex, high-dimensional data. Sparsity has emerged as a major tool for making efficient inferences and predictions for multidimensional data. Generally speaking, sparsity

---

[*]Corresponding author

refers to a situation where the data, despite their apparent high dimensionality, are highly redundant with a low intrinsic dimensionality. In our paper, we use the term "sparse structure" to refer to cases where the underlying distribution P places most of its mass on a subset $\mathcal{X}$ of $\mathbb{R}^d$ of small Lebesgue measure. This scenario includes, but is not limited to, Riemannian submanifolds of $\mathbb{R}^d$, and high-density clusters separated by low-density regions. In applications of interest, observable data often have (complex) sparse structure due to the nature of the underlying physical systems. For example, in astronomy, raw galaxy spectra are of dimension equal to the number of wavelength measurements d, but inspection of a sample of such spectra will reveal clear, low-dimensional features and structure resulting from the shared physical system that generated these galaxies. While the real dimensionality of data is much smaller than d, the challenge remains to exploit this when predicting, for example, the age, composition, and star formation history of a galaxy.

In its simplest form, low-dimensional structure is apparent in the original coordinate system. Indeed, in regression, much research on "large p, small n" problems concerns *variable selection* and the problem of recovering a "sparse" coefficient vector (i.e., a vector with mostly zeros) with respect to the given variables. Such approaches include, for example, lasso-type regularization [55], the Dantzig selector [10], and RODEO [32]. There are also various extensions that incorporate lower-order interactions and groupings of covariates [61, 65, 41] but, like lasso-type estimators, they are not directly applicable to the more intricate structures observed in, e.g., natural images, spectra, and hurricane tracks.

At the same time, there has been a growing interest in statistical methods that explicitly consider *geometric* structure in the data themselves. Most traditional dimension-reducing regression techniques, e.g., principal component regression (PCR; [30]) partial least squares (PLS; [59]) and sparse coding [39], are based on linear data transformations and enforce sparsity (with respect to the L^1 or L^2 norm) of the regression in a rotated space. More recently, several authors [8, 2, 11] have studied local polynomial regression methods on *non-linear* manifolds. For example, Aswani et al. [2] propose a geometry-based regularization scheme that estimates the local covariance matrix at a point and then penalizes regression coefficients perpendicular to the estimated manifold direction. In the same spirit, Cheng and Wu [11] suggest first reducing the dimensionality to the estimated intrinsic dimension of the manifold, and then applying local linear regression to a tangent plane estimate. Local regression and manifold-based methods tend to perform well when there is a clear submanifold but these approaches are not practical in higher dimensions or when the local dimension varies from point to point in the sample space. Hence, existing nonparametric models still suffer when estimating unknown functions (e.g., density and regression functions) on complex objects $\mathbf{x} \in \mathcal{X} \subset \Re^d$, where d is large.

Much statistical research has revolved around adapting classical methods, such as linear, kernel-weighted, and additive models to high dimensions. On the other hand, statisticians have paid little attention to the potential of *orthogonal series* approaches. In low dimensions, orthogonal series is a powerful nonpara-

metric technique for estimating densities and regression functions. Such methods are fast to implement with easily interpretable results, they have sharp optimality properties, and a wide variety of bases allows the data analyst to model multiscale structure and any challenging shape of the target function [17]. As a result, Fourier series approaches have dominated research in signal processing and mathematical physics. This success, however, has not translated to more powerful nonparametric tools in dimensions of the order of $d \sim 100$ or 1000; in fact, extensions via tensor products (as well as more sophisticated adaptive grid or triangulation methods; see [35] and references within) quickly become unpractical in dimensions $d > 3$.

In this work, we will build on ideas from harmonic analysis and spectral methods to construct nonparametric methods for estimating unknown functions in high-dimensional spaces with non-standard data objects (such as images, spectra, and distributions) that possess *sparse nonlinear structure*. We derive a Fourier-like basis $\{\psi_j(\mathbf{x})\}_{j \in \mathbb{N}}$ of $L^2(\mathcal{X})$ that adapts to the intrinsic geometry of the underlying data distribution P, and which is orthonormal with respect to P rather than the Lebesgue measure of the ambient space. The empirical basis functions are then used to estimate functions on complex data $\mathbf{x} \in \mathcal{X}$; such as, for example, the regression function $r(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ of a response variable Y on an object $\mathbf{x}$. Because of the adaptiveness of the basis, there is no need for high-dimensional tensor products. Moreover, we take advantage of the orthogonality property of the basis for fast computation and model selection. We refer to our approach as *spectral series* as it is based on spectral methods (in particular, diffusion maps [14, 12, 33] and spectral connectivity analysis [34]) and Fourier series. Sections 2.1–2.3 describe the main idea of the series method in a regression setting.

Our work generalizes and ties together ideas in classical smoothing, kernel machine learning [48, 49, 15], support vector machines (SVMs; [53]) and manifold regularization [6] *without* the many restrictive assumptions (fixed kernel, exact manifold, infinite unlabeled data and so on) seen in other works. There is a large literature on SVMs and kernel machine learning that use similar approximation spaces as us, but it is unclear whether and how those procedures adapt to the structure of the data distribution. Generally, there is a discrepancy between theoretical work on SVMs, which assume a *fixed* RKHS (e.g., a fixed kernel bandwidth), and applied SVM work, where the RKHS is chosen in a data-dependent way (by, e.g., decreasing the kernel bandwidth ε_n for larger sample sizes n). Indeed, issues concerning the choice of tuning parameters, and their relation to the data distribution P, are considered to be open problems in the mainstream RKHS literature. The manifold regularization work by Belkin et al. [6] addresses adaptivity to sparse structure but under restrictive assumptions, such as the existence of a well-defined submanifold and the presence of infinite unlabeled data.

Another key difference between our work and kernel machine learning is that we *explicitly* compute the eigenfunctions of a kernel-based operator and then use an orthogonal series approach to nonparametric curve estimation. Neither SVMs nor manifold regularizers exploit *orthogonality relative to P*. In our pa-
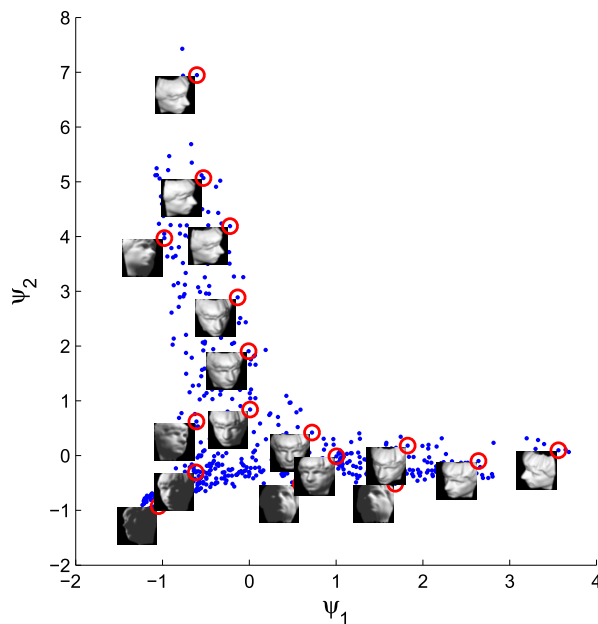
FIG 1. *Embedding or so-called "eigenmap" of the Isomap face data using the first two non-trivial eigenvectors of the Gaussian diffusion kernel. The eigenvectors capture the pose Y and other continuous variations of an image $\mathbf{x}$ fairly well, and the regression $\mathbb{E}(Y|\mathbf{x})$ appears to vary smoothly in sample space.*

per, we point out the advantages of an orthogonal series approach in terms of computational efficiency (such as fast cross-validation and tuning of parameters), visualization, and interpretation. SVMs can sometimes have a "black box feel," whereas the spectral series method allows the user to directly link the data-driven Fourier-like eigenfunctions to the function of interest and the sample space. Indeed, there is a dual interpretation of the computed eigenfunctions: (i) They define *new coordinates of the data* which are useful for nonlinear dimensionality reduction, manifold learning, and data visualization. (ii) They form an orthogonal Hilbert basis for functions on the data and are a means to *nonparametric curve estimation* via the classical orthogonal series method, even when there is no clearly defined manifold structure. There is a large body of work in the machine learning literature addressing the first perspective; see, e.g., Laplacian maps [3], Hessian maps [16], diffusion maps [14], Euclidean Commute Time maps [45], and spectral clustering [50]. In this paper, we are mainly concerned with the second view, i.e., that of estimating unknown functions on complex data objects and understanding the statistical properties of such estimators.

Figure 1, for example, shows a 2D visualization of the Isomap face data using the eigenvectors of a renormalized Gaussian kernel as coordinates (Eq. 4). Assume we want to estimate the pose Y of the faces. How does one solve a regression problem where the predictor $\mathbf{x}$ is an *entire* image? Traditional methods

do not cope well with this task while our spectral series approach (Eq. 1 with estimated eigenfunctions as a basis) can use complex *aggregate* objects $\mathbf{x}$ (e.g., images, spectra, trajectories, and text data) as predictors, without an explicit dimension reduction step. Note that the eigenvectors capture the pose Y and other continuous variations of an image $\mathbf{x}$ fairly well, and that the regression $\mathbb{E}(Y|\mathbf{x})$ appears to vary smoothly in sample space. We will return to the face pose estimation problem in Sec. 6.1. We will also discuss the theoretical properties of a spectral series estimator of the regression function $f(\mathbf{x}) = \mathbb{E}(Y|\mathbf{x})$ in Sec. 5, including the connection between smoothness and efficient estimators.

Our paper has the following aims:

(i) *Unifying.* To generalize and connect ideas in kernel machine learning, manifold learning, spectral methods and classical smoothing, without the many restrictive assumptions (fixed kernel, exact manifold, infinite unlabeled data, low dimension) seen in other works.

(ii) *Theoretical.* To present new theoretical results *in the limit of the kernel bandwidth* $\varepsilon_n \to 0$ that shed light on why RKHS/SVM methods often are so successful for complex data with sparse structure (Theorem 14 and Corollary 16), and to link smoothness of the regression with respect to P to the approximation error of spectral series (Theorem 10).

(iii) *Experimental.* To systematically compare the statistical as well as the computational performance of spectral series and other methods using simulated and real-world data. Competing estimators include classical kernel smoothing, k-nearest neighbors (kNN) regression, regularization in RKHS, and recent state-of-the-art manifold and local regression methods. We ask questions such as: Do the methods scale well with increasing dimension d and increasing sample size n? What is the estimated loss and what is the computational time?

The paper is organized as follows. In Sec. 2, we describe the construction of the spectral series method, including details on how to estimate relevant quantities from empirical data and how to tune model parameters. Sec. 3 discusses the connection to related work in machine learning and statistics. In Sections 4 and 5, we discuss the choice of kernel, and provide theoretical guarantees on the spectral series method. Finally, in Sec. 6, we compare the performance of spectral series and other nonparametric estimators for a wide variety of data sets.

2. Orthogonal series regression

2.1. General formulation

In low dimensions, orthogonal series has proved to be a powerful technique for nonparametric curve estimation [17]. In higher dimensions, there is the question of whether one can find an appropriate basis and actually construct a series estimator that performs well. The general set-up of an orthogonal series regression

is otherwise simple: Let $X_1, \ldots, X_n$ be an iid sample from a distribution P with compact support $\mathcal{X} \subset \mathbb{R}^d$. Suppose we have a real-valued response

$$Y_i = f(X_i) + \epsilon_i,$$

where f is an unknown function, and ϵ_i denotes iid random noise with mean zero and variance σ^2. Our goal is to estimate the regression function $f(x) = \mathbb{E}(Y|X = x)$ in situations where d is large and the data have sparse (i.e., low-dimensional) structure.

Let $\{\psi_j\}_{j \in \mathbb{N}}$ be an orthonormal basis of some appropriate Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$. We consider estimators of the form

$$\widehat{f}(x) = \sum_{j=0}^{J} \widehat{\beta}_j \widehat{\psi}_j(x), \tag{1}$$

where J is a smoothing parameter, and $\widehat{\psi}_j$ and $\widehat{\beta}_j$, in the general case, are data-based estimators of the basis functions ψ_j and the expansion coefficients $\beta_j = \langle f, \psi_j \rangle_{\mathcal{H}}$.

2.2. What basis?

A challenging problem is how to choose a good basis. The standard approach in nonparametric curve estimation is to choose a fixed known basis $\{\psi_j\}_{j \in \mathbb{N}}$ for, say, $L^2([0, 1])$, such as a Fourier or wavelet basis. There is then no need to estimate basis functions. In theory, such an approach can be extended to, eg., $L^2([0, 1]^d)$ by a tensor product,[1] but tensor-product bases, as well as more sophisticated adaptive grid or triangulation methods (see [35] and references within), quickly become unusable for even as few as $d = 5$ dimensions.

What basis should one then choose when the dimension d is large, say, $d \sim 1000$? Ideally, the basis should be able to adapt to the underlying structure of the data distribution. This means: The basis should be orthogonal with respective to the distribution P that generates the data, as opposed to the standard Lebesgue measure of the ambient space; the basis vectors should be concentrated around high-density regions where most of the "action" takes place; and the performance of the final series estimator should depend on the intrinsic rather than the ambient dimension of the data. In what follows, we present a spectral series approach where the unknown function is expanded into the estimated eigenfunctions of a kernel-based integral operator. As we shall see, the proposed estimator has many of the properties listed above.

---

[1]Traditional orthogonal series estimators require $d-1$ tensor products in $\mathbb{R}^d$. For instance, if $d = 2$, then it is common to choose a basis of the form

$$\{\psi_{i,j}(\mathbf{x}) = \psi_i(x_1)\psi_j(x_2) : i, j \in \mathbb{N}\},$$

where $\mathbf{x} = (x_1, x_2)$, and $\{\psi_i(x_1)\}_i$ and $\{\psi_j(x_2)\}_j$ are bases for functions in $L^2(\mathbb{R})$.

2.3. Construction of adaptive basis

Our starting point is a symmetric and positive semi-definite (psd) so-called Mercer kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. These kernels include covariance functions and polynomial kernels, but we are in this work primarily interested in *local*, radially symmetric kernels $k_\varepsilon(x, y) = g\left(\frac{\|x-y\|}{\sqrt{\varepsilon}}\right),$[2] where ε is a parameter that defines the scale of the analysis, and the elements $k(x, y)$ are positive and bounded for all $x, y \in \mathcal{X}$. To simplify the theory, we renormalize the kernel according to

$$a_\varepsilon(x, y) = \frac{k_\varepsilon(x, y)}{p_\varepsilon(x)}, \tag{2}$$

where $p_\varepsilon(x) = \int_\mathcal{X} k_\varepsilon(x, y) dP(y)$. This normalization is common in spectral clustering because it yields eigenvectors that act as indicator functions of connected components [57, Section 3.2]. The same normalization is also implicit in traditional Nadaraya-Watson kernel smoothers, which compute the local average at a point x by weighting surrounding points x_i by $\frac{k_\varepsilon(x, x_i)}{\sum_i k_\varepsilon(x, x_i)}$.

We will refer to $a_\varepsilon(x, y)$ in Eq. 2 as the *diffusion kernel*. The term "diffusion" stems from a random walks view over the sample space [36, 34]: One imagines a Markov chain on $\mathcal{X}$ with transition kernel $\Omega_\varepsilon(x, A) = \mathbb{P}(x \to A) = \int_A a_\varepsilon(x, y) dP(y)$. Starting at x, this chain moves to points y close to x, giving preference to points with high density $p(y)$. The chain essentially encodes the "connectivity" of the sample space relative to p, and it has a unique stationary distribution S_ε given by

$$S_\varepsilon(A) = \frac{\int_A p_\varepsilon(x) dP(x)}{\int_\mathcal{X} p_\varepsilon(x) dP(x)},$$

where $S_\varepsilon(A) \to \frac{\int_A p(x) dP(x)}{\int_\mathcal{X} p(x) dP(x)}$ as $\varepsilon \to 0$. For finite ε, the stationary distribution S_ε is a *smoothed* version of P.

In our regression setting, we seek solutions from a Hilbert space associated with the kernel a_ε. Following [34], we define a "diffusion operator" A_ε — which maps a function f to a new function $A_\varepsilon f$ — according to

$$A_\varepsilon f(x) = \int_\mathcal{X} a_\varepsilon(x, y) f(y) dP(y), \text{ for } x \in \mathcal{X}. \tag{3}$$

The operator A_ε has a discrete set of non-negative eigenvalues $\lambda_{\varepsilon,0} = 1 \geq \lambda_{\varepsilon,1} \geq \ldots \geq 0$ with associated eigenfunctions $\psi_{\varepsilon,0}, \psi_{\varepsilon,1}, \ldots$, which we for convenience normalize to have unit norm. These eigenfunctions have two very useful properties: First, they are orthogonal with respect to the density-weighted L^2 inner product

$$\langle f, g \rangle_\varepsilon = \int_\mathcal{X} f(x) g(x) dS_\varepsilon(x);$$

---

[2]Depending on the application, one can replace the Euclidean distance $\|x - y\|$ with a dissimilarity measure $d(x, y)$ that better reflects the distance between two data objects x and y.

that is,

$$\langle \psi_{\varepsilon,i}, \psi_{\varepsilon,j} \rangle_\varepsilon = \delta_{i,j}.$$

Second, they also form a set of oscillatory functions which are concentrated around high-density regions. By construction, $\psi_{\varepsilon,0}$ is a constant function, and the higher-order eigenfunctions are increasingly oscillatory. Generally speaking, $\psi_{\varepsilon,j}$ is the smoothest function relative to P, subject to being orthogonal to $\psi_{\varepsilon,i}$ for $i < j$.

Interpretation. The diffusion operator A_ε and its eigenfunctions contain information about the connectivity structure of the sample space. There are two ways one can view the eigenfunctions $\psi_{\varepsilon,0}, \psi_{\varepsilon,1}, \psi_{\varepsilon,2}, \ldots$:

(i) The eigenfunctions define *new coordinates of the data.* If the data x represent high-dimensional complex objects, there is often no simple way of ordering the data. However, by a so-called "eigenmap"

$$x \mapsto (\psi_{\varepsilon,1}(x), \psi_{\varepsilon,2}(x), \ldots, \psi_{\varepsilon,J}(x)), \tag{4}$$

one can transform the data into an embedded space where points that are highly connected are mapped close to each other [33]. The eigenmap can be used for data visualization as in Fig. 1 and Fig. 6. If we choose $J < d$, then we are effectively reducing the dimensionality of the problem by mapping the data from $\mathbb{R}^d$ to $\mathbb{R}^J$.

(ii) The eigenfunctions form *a Hilbert basis for functions on the data.* More specifically, the set $\psi_{\varepsilon,0}, \psi_{\varepsilon,1}, \ldots$ is an orthogonal basis of $L^2(\mathcal{X}, P)$. The value of this result is that we can express most physical quantities that vary as a function of the data as a series expansion of the form $f(x) = \sum_{j=0}^\infty \beta_{\varepsilon,j} \psi_{\varepsilon,j}(x)$.

In this work, we study the second point of view and its implications on nonparametric estimation in high dimensions.

2.4. Estimating the regression function from data

In practice, of course, we need to estimate the basis $\{\psi_{\varepsilon,j}\}_j$ and the projections $\{\beta_{\varepsilon,j}\}_j$ from data. In this section, we describe the details.

Given $X_1, \ldots, X_n$, we compute a row-stochastic matrix $\mathbb{A}_\varepsilon$, where

$$\mathbb{A}_\varepsilon(i,j) = \frac{k_\varepsilon(X_i, X_j)}{\sum_{l=1}^n k_\varepsilon(X_i, X_l)} \tag{5}$$

for $i, j = 1, \ldots, n$. The elements $\mathbb{A}_\varepsilon(i,j)$ can be interpreted as transition probabilities $\mathbb{A}_\varepsilon(i,j) = \mathbb{P}(x_i \to x_j)$ for a Markov chain over the data points (i.e., this is the discrete analogue of Eq. 2 and a diffusion over $\mathcal{X}$). Let $\widehat{p}_\varepsilon(x) = \frac{1}{n} \sum_{j=1}^n k_\varepsilon(x, X_j)$. The Markov chain has a unique stationary measure given by $(\widehat{s}_\varepsilon(X_1), \ldots, \widehat{s}_\varepsilon(X_n))$, where the ith element

$$\widehat{s}_\varepsilon(X_i) = \frac{\widehat{p}_\varepsilon(X_i)}{\sum_{j=1}^n \widehat{p}_\varepsilon(X_j)} \tag{6}$$

is a kernel-smoothed density estimate at the ith observation.

To estimate the eigenfunctions $\psi_{\varepsilon,1}, \ldots, \psi_{\varepsilon,J}$ of the continuous diffusion operator A_ε in Eq. 3, we first calculate the eigenvalues $\lambda^{\mathbb{A}}_{\varepsilon,1}, \ldots, \lambda^{\mathbb{A}}_{\varepsilon,J}$ and the associated (orthogonal) eigenvectors $\widetilde{\psi}^{\mathbb{A}}_{\varepsilon,1}, \ldots, \widetilde{\psi}^{\mathbb{A}}_{\varepsilon,J}$ of the symmetrized kernel matrix $\widetilde{\mathbb{A}}_\varepsilon$, where

$$\widetilde{\mathbb{A}}_\varepsilon(i,j) = \frac{k_\varepsilon(X_i, X_j)}{\sqrt{\sum_l k_\varepsilon(X_i, X_l)}\sqrt{\sum_l k_\varepsilon(X_l, X_j)}}. \tag{7}$$

We normalize the eigenvectors so that $\frac{1}{n}\sum_{i=1}^{n} \widetilde{\psi}^{\mathbb{A}}_{\varepsilon,j}(i)\widetilde{\psi}^{\mathbb{A}}_{\varepsilon,k}(i) = \delta_{j,k}$, and define the new vectors

$$\psi^{\mathbb{A}}_{\varepsilon,j}(i) = \frac{\widetilde{\psi}^{\mathbb{A}}_{\varepsilon,j}(i)}{\sqrt{\widehat{s}_\varepsilon(X_i)}}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. By construction, it holds that the $\lambda^{\mathbb{A}}_{\varepsilon,j}$'s and $\psi^{\mathbb{A}}_{\varepsilon,j}$'s are eigenvalues and right eigenvectors of the Markov matrix $\mathbb{A}_\varepsilon$:

$$\mathbb{A}_\varepsilon \psi^{\mathbb{A}}_{\varepsilon,j} = \lambda^{\mathbb{A}}_{\varepsilon,j}\psi^{\mathbb{A}}_{\varepsilon,j} \tag{8}$$

where

$$\frac{1}{n}\sum_{i=1}^{n} \psi^{\mathbb{A}}_{\varepsilon,j}(i)\psi^{\mathbb{A}}_{\varepsilon,k}(i)\widehat{s}_\varepsilon(X_i) = \delta_{j,k}. \tag{9}$$

Note that the n-dimensional vector $\psi^{\mathbb{A}}_{\varepsilon,j}$ can be regarded as estimates of $\psi_{\varepsilon,j}(x)$ at the observed values $X_1, \ldots, X_n$. In other words, let

$$\widehat{\lambda}_{\varepsilon,j} \equiv \lambda^{\mathbb{A}}_{\varepsilon,j} \quad \text{and} \quad \widehat{\psi}_{\varepsilon,j}(X_i) \equiv \psi^{\mathbb{A}}_{\varepsilon,j}(i) \tag{10}$$

for $i = 1, \ldots, n$. We estimate the function $\psi_{\varepsilon,j}(x)$ at values of x not corresponding to one of the X_i's using the so-called Nyström method. The idea is to first rearrange the eigenfunction-eigenvalue equation $\lambda_{\varepsilon,j}\psi_{\varepsilon,j} = A_\varepsilon\psi_{\varepsilon,j}$ as

$$\psi_{\varepsilon,j}(x) = \frac{A_\varepsilon\psi_{\varepsilon,j}}{\lambda_{\varepsilon,j}} = \frac{1}{\lambda_{\varepsilon,j}}\int_{\mathcal{X}} \frac{k_\varepsilon(x,y)}{\int_{\mathcal{X}} k_\varepsilon(x,y)dP(y)}\psi_{\varepsilon,j}(y)dP(y),$$

and use the kernel-smoothed estimate

$$\widehat{\psi}_{\varepsilon,j}(x) = \frac{1}{\widehat{\lambda}_{\varepsilon,j}}\sum_{i=1}^{n} \frac{k_\varepsilon(x,X_i)}{\sum_{l=1}^{n} k_\varepsilon(x,X_l)}\widehat{\psi}_{\varepsilon,j}(X_i). \tag{11}$$

for $\widehat{\lambda}_{\varepsilon,j} > 0$.

Our final regression estimator is defined by Eq. 1 with the estimated eigenvectors in Eq. 11 and expansion coefficients computed according to

$$\widehat{\beta}_{\varepsilon,j} = \frac{1}{n}\sum_{i=1}^{n} Y_i\widehat{\psi}_{\varepsilon,j}(X_i)\widehat{s}_\varepsilon(X_i). \tag{12}$$

Remark 1 (**Semi-Supervised Learning, SSL**). *The spectral series frame-work naturally extends to* semi-supervised learning *(SSL)* [64] *where in addition to the "labeled" sample* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *there are additional "unlabeled" data; i.e., data* $X_{n+1}, \ldots, X_{n+m}$ *where the covariates* X_i *but not the labels* Y_i *are known. Typically* $m \gg n$*, as collecting data often is less costly than labeling them. By including unlabeled examples (drawn from the same distribution* P_X*) into the kernel matrix* $\mathbb{A}_\varepsilon$*, we can improve our estimates of* $\lambda_{\varepsilon,j}$*,* $\psi_{\varepsilon,j}$ *and* S_ε*. The summation in Equations* 9 *and* 11 *will then be over all* $n + m$ *observations, while Eq.* 12 *remains the same as before. See e.g.* [38, 62] *for SSL with Laplacian eigenmaps in the limit of infinite unlabeled data, i.e., in the limit* $m \to \infty$*.*

2.5. Loss function and tuning of parameters

We measure the performance of an estimator $\widehat{f}(x)$ via the L^2 loss function

$$L(f, \widehat{f}) = \int_{\mathcal{X}} \left(f(x) - \widehat{f}(x) \right)^2 dP(x). \tag{13}$$

To choose tuning parameters (such as the kernel bandwidth ε and the number of basis functions J), we split the data into a training and a validation set. For each choice of ε and a sufficiently large constant $J_{\max}$, we use the training set and Eqs. 11–12 to estimate the eigenvectors $\psi_{\varepsilon,1}, \ldots, \psi_{\varepsilon,J_{\max}}$ and the expansion co-efficients $\beta_{\varepsilon,0}, \ldots, \beta_{\varepsilon,J_{\max}}$. We then use the validation set $(X_1', Y_1'), \ldots, (X_n', Y_n')$ to compute the estimated loss

$$\widehat{L}(f, \widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i' - \widehat{f}(X_i') \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i' - \sum_{j=0}^{J} \widehat{\beta}_{\varepsilon,j} \widehat{\psi}_{\varepsilon,j}(X_i') \right)^2 \tag{14}$$

for different values of $J \le J_{\max}$. We choose the (ε, J)-model with the lowest estimated loss on the validation set.

The computation for fixed ε and different J is very fast. Due to orthogonality of the basis, the estimates $\widehat{\beta}_{\varepsilon,j}$ and $\widehat{\psi}_{\varepsilon,j}$ depend on ε but *not* on J.

2.6. Scalability

The spectral series estimator is faster than most traditional approaches in high dimensions. Once the kernel matrix has been constructed, the eigendecomposition takes the *same* amount of time for all values of d.

In terms of scalability for large data sets, one can dramatically reduce the computational cost by implementing fast approximate eigendecompositions. For example, the Randomized SVD by Halko et al. [23] cuts down the cost from $O(n^3)$ to roughly $O(n^2)$ with little impact on statistical performance (see Fig. 9). According to Halko et al., these randomized methods are especially well-suited for parallel implementation, which is a topic we will explore in future work.

3. Connection to other work

3.1. Linear regression with transformed data

One can view our series model as a (weighted) linear regression after a data transformation $Z = \Psi(X)$, where $\Psi = (\psi_1, \ldots, \psi_J)$ are the first J eigenvectors of the diffusion operator A_ε. By increasing J, the dimension of the feature space, we achieve more flexible, *fully nonparametric* representations. Decreasing J adds more structure to the regression, as dictated by the eigenstructure of the data.

Eq. 12 is similar to a weighted least squares (WLS) solution to a linear regression in $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ but with an efficient orthogonal series implementation and no issues with collinear variables. Define the $n \times (J+1)$ matrix of predictors,

$$\mathbb{Z} = \begin{pmatrix} 1 & \psi_1(X_1) & \cdots & \psi_J(X_1) \\ 1 & \psi_1(X_2) & \cdots & \psi_J(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \psi_1(X_n) & \cdots & \psi_J(X_n) \end{pmatrix}, \tag{15}$$

and introduce the weight matrix

$$\mathbb{W} = \begin{pmatrix} s(X_1) & 0 & \cdots & 0 \\ 0 & s(X_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s(X_n) \end{pmatrix}, \tag{16}$$

where Ψ_j and s are estimated from data (Equations 6 and 10). Suppose that $Y = \mathbb{Z}\beta + e$, where $Y = (Y_1, \ldots, Y_n)^T$, $\beta = (\beta_1, \ldots, \beta_J)^T$, and the random vector $e = (\epsilon_1, \ldots, \epsilon_n)^T$ represents the errors. By minimizing the weighted residual sum of squares

$$RSS(\beta) = (Y - \mathbb{Z}\beta)^T \mathbb{W}(Y - \mathbb{Z}\beta), \tag{17}$$

we arrive at the WLS estimator

$$\widehat{\beta} = (\mathbb{Z}^T \mathbb{W} \mathbb{Z})^{-1}(\mathbb{Z}^T \mathbb{W} Y) = \frac{1}{n} \mathbb{Z}^T \mathbb{W} Y, \tag{18}$$

where the matrix $\mathbb{W}$ puts more weight on observations in high-density regions. This expression is equivalent to Eq. 12.

Note that thanks to the orthogonality property $\mathbb{Z}^T \mathbb{W} \mathbb{Z} = n\mathbb{I}$, model search and model selection are feasible even for complex models with very large J. This is in clear contrast with standard multiple regression where one needs to recompute the $\widehat{\beta}_j$ estimates for each model with a different J, invert the matrix $\mathbb{Z}^T \mathbb{W} \mathbb{Z}$, and potentially deal with inputs (columns of the design matrix $\mathbb{Z}$) that are linearly dependent.

Remark 2 (**Heteroscedasticity**). *More generally, let $\sigma(x)$ be a non-negative function rather than a constant, and let ϵ_i be iid realizations of a random variable ϵ with zero mean and unit variance. Consider the regression model $Y_i = f(X_i) + \sigma(X_i)\epsilon_i$. We can handle heteroscedastic errors by applying the same framework as above to a* rescaled *regression function $g(x) = f(x)/\sigma(x)$.*

3.2. Kernel machine learning and regularization in RKHS

Kernel-based regularization methods use similar approximation spaces as us. In kernel machine learning [49, 15], one often considers the variational problem

$$\min_{f \in \mathcal{H}_k} \left[\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \gamma \|f\|_{\mathcal{H}_k}^2 \right], \tag{19}$$

where $L(y_i, f(x_i))$ is a convex loss function, $\gamma > 0$ is a penalty parameter, and $\mathcal{H}_k$ is the Reproducing Kernel Hilbert Space (RKHS) associated with a symmetric positive semi-definite kernel k.[3] Penalizing the RKHS norm $\| \cdot \|_{\mathcal{H}_k}$ imposes smoothness conditions on possible solutions. Now suppose that

$$k(x, y) = \sum_{j=0}^{\infty} \lambda_j \phi_j(x) \phi_j(y),$$

where the RKHS inner product is related to the L^2-inner product according to $\langle \phi_i, \phi_j \rangle_{\mathcal{H}_k} = \frac{1}{\lambda_i} \langle \phi_i, \phi_j \rangle_{L^2(\mathcal{X}, P)} = \frac{1}{\lambda_i} \delta_{i,j}$. Eq. 19 is then equivalent to considering eigen-expansions

$$f(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x),$$

and seeking solutions to $\min_{f \in \mathcal{B}_r} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$, where the hypothesis space

$$B_r = \{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq r \} \tag{20}$$

is a ball of the RKHS $\mathcal{H}_k$ with radius r, and the RKHS norm is given by $\|f\|_{\mathcal{H}_k} = \left(\sum_{j=0}^{\infty} \frac{\beta_j^2}{\lambda_j} \right)^{1/2}$.

Here are some key observations:

(i) The above setting is similar to ours. The regularization in Eq. 19 differentially shrinks contributions from higher-order terms with small λ_j values. In spectral series, we use a projection (i.e., a basis subset selection) method, but the empirical performance is usually similar.

(ii) There are some algorithmic differences, as well as differences in how the two regression estimators are analyzed and interpreted. In our theoretical work, we consider Gaussian kernels with *flexible variances*; that is, we choose the approximation spaces in a data-dependent way (cf. multi-kernel regularization schemes for SVMs [60]) so that the estimator can adapt to sparse structure and the intrinsic dimension of the data. Most theoretical work in kernel machine learning assume a fixed RKHS.

(iii) There are also other differences. Support Vector Machines [53] and other kernel-based regularization methods (such as splines, ridge regression and radial

---

[3]To every continuous, symmetric, and positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is associated a unique RKHS $\mathcal{H}_k$ [1]. This RKHS is defined to be the closure of the linear span of the set of functions $\{k(x, \cdot) : x \in \mathcal{X}\}$ with the inner product satisfying the reproducing property $\langle k(x, \cdot), f \rangle_{\mathcal{H}_k} = f(x)$ for all $x \in \mathcal{X}, f \in \mathcal{H}_k$.

basis functions) never explicitly compute the eigenvectors of the kernel. Instead, these methods rely on the classical Representer Theorem [58] which states that the solution to Eq. 19 is a finite expansion of the form $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$. The original infinite-dimensional variational problem is then reduced to a finite-dimensional optimization of the coefficients α_i. In a naive least-squares implementation, however, one has to recompute these coefficients for each choice of the penalty parameter γ, which can make cross-validation cumbersome. In our spectral series approach, we take advantage of the orthogonality of the basis for fast model selection and computation of the β_j parameters. As in spectral clustering, we also use eigenvectors to *organize* and *visualize* data that can otherwise be hard to interpret.

3.3. Manifold regularization and semi-supervised learning

Our spectral series method is closely related to Laplacian-based regularization: In [6], Belkin et al. extend the kernel-based regularization framework to incorporate additional information about the geometric structure of the marginal P_X. Their idea is to add a data-dependent penalty term to Eq. 19 that controls the complexity as measured by the geometry of the distribution. Suppose that one is given labeled data $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{X,Y}$ as well as unlabeled data $X_{n+1}, \ldots, X_{n+m} \sim P_X$, where in general $m \gg n$. (The limit $m \to \infty$ corresponds to having full knowledge of P_X.) Under the assumption that the support of P_X is a compact submanifold of $\mathbb{R}^d$, the authors propose minimizing a graph-Laplacian regularized least squares function

$$\min_{f \in \mathcal{H}_k} \left[\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \gamma_A \|f\|_{\mathcal{H}_k}^2 + \frac{\gamma_I}{n+m} \sum_{i,j=1}^{n+m} (f(x_i) - f(x_j))^2 W_{i,j} \right],$$
(21)

where $W_{i,j}$ are the edge weights in the graph, and the last Laplacian penalty term favors functions f for which $f(x_i)$ is close to $f(x_j)$ when x_i and x_j are connected with large weights.

Note that the eigenbasis of our row-stochastic matrix $\mathbb{A}_\varepsilon$ minimizes the distortion $\sum_{i,j} (f(x_i) - f(x_j))^2 W_{i,j}$ in Eq. 21 if you regard the entries of $\mathbb{A}_\varepsilon$ as the weights $W_{i,j}$ [3]. Indeed, the eigenvector $\psi_{\varepsilon,j}$ minimizes the term subject to being orthogonal to $\psi_{\varepsilon,i}$ for $i < j$. Hence, including a Laplacian penalty term is comparable to truncating the eigenbasis expansion in spectral series. Moreover, the semi-supervised regularization in Eq. 21 is similar to a semi-supervised version of our spectral series approach, where we first use both labeled and unlabeled data and a kernel with bandwidth ε_{m+n} to compute the eigenbasis, and then extend the eigenfunctions according to Eq. 11 via a (potentially wider) kernel with bandwidth h_n. The main downside of the Laplacian-based framework above is that it is hard to analyze theoretically. As with other kernel-based regularizers, the method also does not explicitly exploit eigenvectors and orthogonal bases.
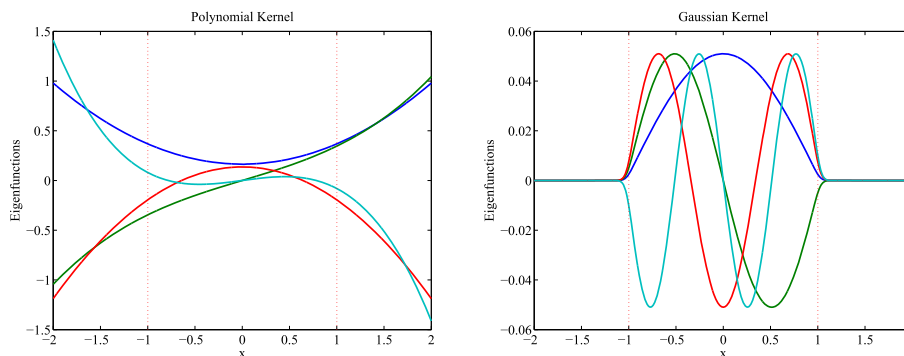
FIG 2. *Uniform distribution U(-1,1). Eigenfunctions of a third-order polynomial kernel,* left, *and of an (un-normalized) Gaussian kernel,* right. *The latter eigenfunctions form an orthogonal Fourier-like basis concentrated on the support of the data distribution.*

4. Choice of kernel

In the RKHS literature, there is a long list of commonly used kernels. These include, e.g., the Gaussian kernel $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{4\varepsilon}\right)$, polynomial kernels $k(x,y) = (\langle x, y\rangle + 1)^q$ [56], and the thin-plate spline kernel $k(x,y) = \|x - y\|^2 \log(\|x - y\|^2)$ [21]. In our numerical experiments (Sec. 6), we will consider both Gaussian and polynomial kernels, but throughout the rest of the paper, we will primarily work with the (row-normalized) Gaussian kernel. There are several reasons for this choice:

(i) The Gaussian kernel can be interpreted as the heat kernel in a manifold setting [3, 22]. We will take advantage of this connection in the theoretical analysis of the spectral series estimator (Sec. 5).

(ii) The eigenfunctions of the Gaussian kernel are simultaneously concentrated in time (i.e., space) and frequency, and are particularly well-suited for estimating functions that are smooth with respect to a low-dimensional data distribution.

The following two examples illustrate some of the differences in the eigenbases of Gaussian and polynomial kernels:

Example 3. *Suppose that P is a uniform distribution $U(-1,1)$ on the real line. Fig. 2, left, shows the eigenfunctions of a third-order polynomial kernel $k(x,y) = (\langle x, y\rangle + 1)^3$. These functions are smooth but have large values outside the support of P. Contrast this eigenbasis with the eigenfunctions in Fig. 2, right, of a Gaussian kernel. The latter functions are concentrated on the support of P and are orthogonal on $(-1, 1)$ as well as on $(-\infty, \infty)$.*

Example 4. *Consider data around a noisy spiral:*

$$\begin{cases} x(u) &= \sqrt{u}\cos(\sqrt{u}) + \epsilon_x \\ y(u) &= \sqrt{u}\sin(\sqrt{u}) + \epsilon_y, \end{cases}$$
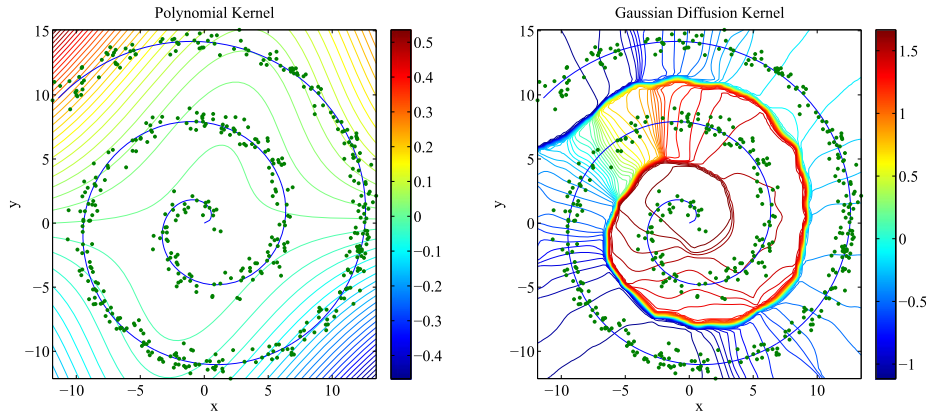
FIG 3. *Spiral data. Contour plots of the fourth eigenvector of a third-order polynomial kernel (*left*) and the fourth eigenvector of a Gaussian diffusion kernel (*right*). The latter eigenvector is localized and varies smoothly along the spiral direction.*

where u is a uniform random variable, and ϵ_x and ϵ_y are normally distributed random variables. The eigenfunctions of a polynomial kernel do not adapt well to the underlying distribution of the data. Fig. 3, left, for example, is a contour plot of the Nyström extension of the fourth empirical eigenvector of a third-order polynomial kernel. In contrast, the eigenfunctions of a Gaussian diffusion kernel vary smoothly along the spiral direction, forming a Fourier-like basis with orthogonal eigenfunctions that concentrate around high-density regions; see Fig. 3, right.

In high dimensions, Gaussian extensions can be seen as a generalization of *prolate spheroidal wave functions* [13]. Prolates were originally introduced by Slepian and Pollack as the solution to the problem of simultaneously and optimally concentrating a function and its Fourier content (see [52] for a fascinating recount of this development in Fourier analysis and modeling). The band-limited functions that maximize their energy content within a space domain $\mathcal{X} \subset \mathbb{R}^n$ are extensions of the eigenfunctions of the integral operator of a Bessel kernel restricted to $\mathcal{X}$ [13, Section 3.1]. In high dimensions, Bessel and Gaussian kernels are equivalent [47], suggesting that the eigenfunctions of the Gaussian kernel are *nearly* optimal.

However, although Gaussian kernels have many advantages, they may not always be the best choice in practice. Ultimately, this is determined by the application and by what the best measure of similarity between two data points would be. Our framework suggests a principled way of selecting the best kernel for regression: Among a set of reasonable candidate kernels, choose the estimator with the *smallest empirical loss* according to Eq. 14. We will, for example, use this approach in Sec. 6 to choose the optimal degree q for a set of polynomial kernels of the form $k(x,y) = (\langle x, y \rangle + 1)^q$.
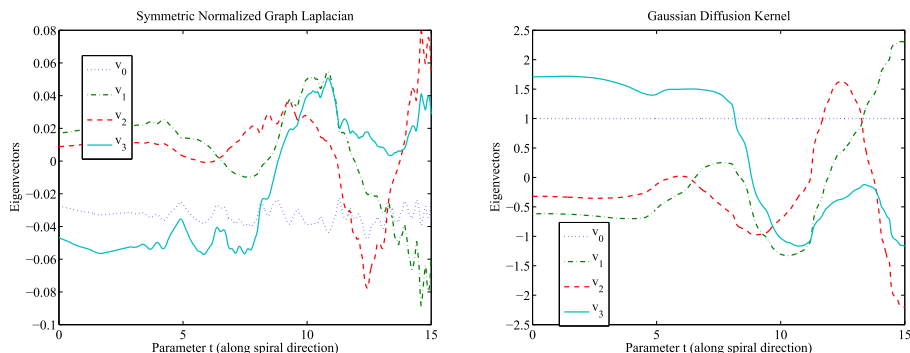
FIG 4. *Projection onto spiral direction* $(t = \sqrt{u})$ *of the eigenvectors of the symmetric graph Laplacian,* left, *and of the Gaussian diffusion kernel,* right. *The latter estimates are less noisy.*

Normalization of local kernels. In the RKHS literature, it is standard to work with "unnormalized" kernels. In spectral clustering [57], on the other hand, researchers often use the "stochastic" and "symmetric" normalization schemes in Eq. 5 and Eq. 7, respectively. We have found (Sec. 6) that the exact normalization often has little effect on the performance in regression. Nevertheless, we choose to use the *row-stochastic* kernel for reasons of interpretation and analysis: First, the limit of the bandwidth $\varepsilon \to 0$ is well-defined, and there is a series of works on the convergence of the graph Laplacian to the Laplace-Beltrami operator on Riemannian manifolds [12, 5, 25, 51, 20]. Fourier functions originate from solving a Laplace eigenvalue problem on a bounded domain; hence, the eigenfunctions of the diffusion operator can be seen as a generalization of Fourier series to manifolds.

Moreover, the row-stochastic kernel yields less variable empirical functions than the unnormalized or symmetric forms. As an illustration, consider the noisy spiral data in Example 4. Fig. 4 shows the estimated projections onto the spiral direction of the eigenfunctions of the *symmetric* and the *stochastic* forms; see the left and right plots, respectively. The eigenfunctions are clearly smoother in the latter case. By construction, the empirical eigenfunctions of the symmetric operator are orthogonal with respect to the empirical distribution $\widehat{P}_n$, whereas the estimated eigenfunctions of the stochastic operator are orthogonal with respect to the *smoothed* data distribution $\widehat{S}_\varepsilon$. The kernel bandwidth ε_n defines the scale of the analysis.

5. Theory

In this section, we derive theoretical bounds on the loss (Eq. 13) of a series regression estimator with a radial kernel for a standard fixed RKHS setting (Theorem 13), as well as a setting where the kernel bandwidth ε_n *varies* with the sample size n (Theorem 14). We also further elaborate on the connection

between spectral series and Fourier analysis by generalizing the well-known link between Sobolev differentiable signals and the approximation error in a Fourier basis.

Using the same notation as before, let

$$f(x) = \sum_{j=0}^{\infty} \beta_{\varepsilon,j} \psi_{\varepsilon,j}(x), \qquad f_{\varepsilon,J}(x) = \sum_{j=0}^{J} \beta_{\varepsilon,j} \psi_{\varepsilon,j}(x),$$

$$\widehat{f}_{\varepsilon,J}(x) = \sum_{j=0}^{J} \widehat{\beta}_{\varepsilon,j} \widehat{\psi}_{\varepsilon,j}(x),$$

where $\beta_{\varepsilon,j} = \int_{\mathcal{X}} f(x)\psi_{\varepsilon,j}(x) dS_{\varepsilon}(x)$ and $\widehat{\beta}_{\varepsilon,j} = \frac{1}{n} \sum_{i=1}^{n} Y_i \widehat{\psi}_{\varepsilon,j}(X_i) \widehat{s}_{\varepsilon}(X_i)$. We write

$$|f(x) - \widehat{f}_{\varepsilon,J}(x)|^2 \le 2|f(x) - f_{\varepsilon,J}(x)|^2 + 2|f_{\varepsilon,J}(x) - \widehat{f}_{\varepsilon,J}(x)|^2,$$

and refer to the two terms as "bias" and "variance". Hence, we define the integrated bias and variance

$$L_{\text{bias}} = \int_{\mathcal{X}} |f(x) - f_{\varepsilon,J}(x)|^2 dP(x),$$

and

$$L_{\text{var}} = \int_{\mathcal{X}} |f_{\varepsilon,J}(x) - \widehat{f}_{\varepsilon,J}(x)|^2 dP(x),$$

and bound the two components separately. Our assumptions are:

(A1) P has compact support $\mathcal{X}$ and bounded density $0 < a \le p(x) \le b < \infty$, $\forall x \in \mathcal{X}$.

(A2) The weights are positive and bounded; that is, $\forall x, y \in \mathcal{X}$,

$$0 < m \le k_{\varepsilon}(x, y) \le M,$$

where m and M are constants that do not depend on ε.

(A3) The psd operator A_{ε} has nondegenerate eigenvalues; i.e.,

$$1 \equiv \lambda_{\varepsilon,0} > \lambda_{\varepsilon,1} > \lambda_{\varepsilon,2} > \dots \lambda_{\varepsilon,J} > 0.$$

(A4) For all $0 \le j \le J$ and $X \sim P$, there exists some constant $C < \infty$ (not depending on n) such that

$$\mathbb{E}\left[|\widehat{\varphi}_{\varepsilon,j}(X) - \varphi_{\varepsilon,j}(X)|^2\right] < C,$$

where $\varphi_{\varepsilon,j}(x) = \psi_{\varepsilon,j}(x) s_{\varepsilon}(x)$ and $\widehat{\varphi}_{\varepsilon,j}(x) = \widehat{\psi}_{\varepsilon,j}(x) \widehat{s}_{\varepsilon}(x)$.

Without loss of generality, we assume that the eigenfunctions $\psi_{\varepsilon,j}$ are estimated using an unlabeled sample $\widetilde{X}_1, \dots, \widetilde{X}_n$ that is drawn independently from the data used to estimate the coefficients $\beta_{\varepsilon,j}$. This is to simplify the proofs and can always be achieved by splitting the data in two sets.

5.1. Bias

A key point is that the approximation error of the regression depends on the smoothness of f relative to P. Here we present two different calculations of the bias based on two related notions of smoothness. The first notion is standard in the kernel literature and is based on RKHS norms. The second notion is based on our diffusion framework and can be seen as a generalization of Sobolev differentiability.

Method 1: Smoothness measured by RKHS norm.

Let $\widetilde{a}_\varepsilon(x,y) = \frac{k_\varepsilon(x,y)}{\sqrt{p_\varepsilon(x)}\sqrt{p_\varepsilon(y)}}$, where ε is a strictly positive number. Under previous assumptions, this kernel is symmetric and psd with a unique RKHS which we denote by $\mathcal{H}_\varepsilon$. A standard way to measure smoothness of a function f in a RKHS $\mathcal{H}_\varepsilon$ is through the RKHS norm $\|f\|_{\mathcal{H}_\varepsilon}$ (see, e.g., [37]). One can then define function classes

$$\mathcal{H}_{\varepsilon,M} = \{f \in \mathcal{H}_\varepsilon : \|f\|_{\mathcal{H}_\varepsilon} \leq M\},$$

where M is a positive number dependent on ε.

Proposition 5. *Assume $f \in \mathcal{H}_{\varepsilon,M}$, where $\varepsilon > 0$. Then,*

$$L_{\text{bias}} = O(M\lambda_{\varepsilon,J}).$$

For fixed ε, $\mathcal{H}_{\varepsilon,M}$ contains "smoother" functions for smaller values of M.

Method 2: Smoothness measured by diffusion operator.

Alternatively, let

$$G_\varepsilon = \frac{A_\varepsilon - I}{\varepsilon}, \tag{22}$$

where I is the identity. The operator G_ε has the same eigenvectors $\psi_{\varepsilon,j}$ as the differential operator A_ε. Its eigenvalues are given by $-\nu_{\varepsilon,j}^2 = \frac{\lambda_{\varepsilon,j}-1}{\varepsilon}$, where $\lambda_{\varepsilon,j}$ are the eigenvalues of A_ε. Define the functional

$$\mathcal{J}_\varepsilon(f) = -\langle G_\varepsilon f, f \rangle_\varepsilon \tag{23}$$

which maps a function $f \in L^2(\mathcal{X}, P)$ into a non-negative real number. For small ε, $\mathcal{J}_\varepsilon(f)$ measures the variability of the function f with respect to the distribution P. The expression is a variation of the graph Laplacian regularizers popular in semi-supervised learning [63]. In fact, a Taylor expansion yields $G_\varepsilon f = -\triangle f + \frac{\nabla p}{p} \cdot \nabla f + O(\varepsilon)$ where ∇ is the gradient operator and $\triangle = -\sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$ is the psd Laplace operator in $\mathbb{R}^d$. In kernel regression smoothing, the extra term $\frac{\nabla p}{p} \cdot \nabla f$ is considered an undesirable extra bias, called design

bias. In classical regression, it is removed by using local linear smoothing [18], which is asymptotically equivalent to replacing the original kernel $k_\varepsilon(x, y)$ by the bias-corrected kernel $k_\varepsilon^*(x, y) = \frac{k_\varepsilon(x,y)}{p_\varepsilon(x)p_\varepsilon(y)}$ [12].

The following result bounds the approximation error of an orthogonal series expansion of f. The bound is consistent with Theorem 2 in [62], which applies to the more restrictive setting of SSL with infinite unlabeled data and $\varepsilon \to 0$. Our result holds for *all* ε and J and does *not* assume unlimited data.

Proposition 6. *For* $f \in L^2(\mathcal{X}, P)$,

$$\int_\mathcal{X} |f(x) - f_{\varepsilon,J}(x)|^2 dS_\varepsilon(x) \ \leq \ \frac{\mathcal{J}_\varepsilon(f)}{\nu_{\varepsilon,J+1}^2} \tag{24}$$

$$L_{\text{bias}} = O\left(\frac{\mathcal{J}_\varepsilon(f)}{\nu_{\varepsilon,J+1}^2}\right),$$

where $-\nu_{\varepsilon,J+1}^2$ *is the* $(J+1)^{th}$ *eigenvalue of* G_ε.

Smoothness and sparsity

In the limit $\varepsilon \to 0$, we have several interesting results, including a generalization of the classical connection between Sobolev differentiability and the error decay of Fourier approximations [35, Section 9.1.2] to a setting with adaptive bases and *high-dimensional* data. We denote the quantities derived from the bias-corrected kernel k_ε^* by A_ε^*, G_ε^*, $\mathcal{J}_\varepsilon^*$ and so forth.

Definition 7 (**Smoothness relative to P**). *A function* f *is smooth relative to* P *if*

$$\int_\mathcal{X} \|\nabla f(x)\|^2 dS(x) \leq c^2 < \infty,$$

where $S(A) = \frac{\int_A p(x)dP(x)}{\int p(x)dP(x)}$ *is the stationary distribution of the random walk on the data as* $\varepsilon \to 0$. *The smaller the value of* c, *the smoother the function.*

Lemma 8. *For functions* $f \in C^3(\mathcal{X})$ *whose gradients vanish at the boundary,*

$$\lim_{\varepsilon \to 0} \mathcal{J}_\varepsilon^*(f) = \int_\mathcal{X} \|\nabla f(x)\|^2 dS(x).$$

This is similar to the convergence of the (un-normalized) graph Laplacian regularizer to the density-dependent smoothness functional $\int_\mathcal{X} \|\nabla f(x)\|^2 p^2(x) dx$ [9].

Next we will see that *smoothness* relative to P (Definition 7) and *sparsity* (with respect to the L^2 norm) in the eigenbasis of the diffusion operator (Definition 9 below) are really the same thing. As a result, we can link smoothness and sparsity to the rate of the error decay of the eigenbasis approximation.

Definition 9 (Sparsity in $\mathcal{B}$). *A set of real numbers $\theta_1, \theta_2, \ldots$ lies in a Sobolev ellipsoid $\Theta(s, c)$ if $\sum_{j=1}^{\infty} j^{2s}\theta_{(j)}^2 \leq c^2$ for some number $s > 1/2$. For a given basis $\mathcal{B} = \{\psi_1, \psi_2, \ldots\}$, let*

$$W_{\mathcal{B}}(s, c) = \left\{ f = \sum_j \beta_j \psi_j : \beta_1, \beta_2, \ldots \in \Theta(s, c) \right\}$$

where $s > 1/2$. Functions in $W_{\mathcal{B}}(s, c)$ are sparse in $\mathcal{B}$. The larger the value of s, the sparser the representation.

Theorem 10. *Assume that $\mathcal{B} = \{\psi_1, \psi_2, \ldots\}$ are the eigenvectors of $\triangle$ with eigenvalues $\nu_j^2 = O(j^{2s})$ for some $s > 1/2$. Let $f_J(x) = \sum_{j \leq J} \beta_j \psi_j(x)$. Then, the following two statements are equivalent:*

1. *$\int_{\mathcal{X}} \|\nabla f(x)\|^2 dS(x) \leq c^2$ (smoothness relative to P)*
2. *$f \in W_{\mathcal{B}}(s, c)$ (sparsity in $\mathcal{B}$).*

Furthermore, sparsity in $\mathcal{B}$ (or smoothness relative to P) implies

$$\int_{\mathcal{X}} |f(x) - f_J(x)|^2 dS(x) = o\left(\frac{1}{J^{2s}}\right).$$

The rate s of the error decay depends on the dimension of the data. We will address this issue in Sec. 5.3.

5.2. Variance

The matrix $\mathbb{A}_\varepsilon$ (defined in Eq. 5) can be viewed as a perturbation of the integral operator A_ε due to finite sampling. To estimate the variance, we bound the difference $\psi_{\varepsilon,j} - \widehat{\psi}_{\varepsilon,j}$, where $\psi_{\varepsilon,j}$ are the eigenvectors of A_ε, and $\widehat{\psi}_{\varepsilon,j}$ are the Nyström extensions (Eq. 11) of the eigenvectors of $\mathbb{A}_\varepsilon$. We adopt a strategy from Rosasco et al. [44], which is to introduce two new integral operators that are related to A_ε and $\mathbb{A}_\varepsilon$ but both act on an auxiliary[4] RKHS $\mathcal{H}$ of smooth functions (see Appendix A.2 for details). As before, we write ε_n to indicate that we let the kernel bandwidth ε depend on the sample size n.

Proposition 11. *Let $\varepsilon_n \to 0$ and $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$ as $n \to 0$. Under assumptions (A1)–(A4) and $\forall\, 0 \leq j \leq J$,*

$$\|\psi_{\varepsilon,j} - \widehat{\psi}_{\varepsilon,j}\|_{L^2(\mathcal{X}, P)} = O_P\left(\frac{\gamma_n}{\delta_{\varepsilon,j}}\right),$$

where $\gamma_n = \sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}$ and $\delta_{\varepsilon,j} = \lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1}$.

---

[4]This auxiliary space only enters the intermediate derivations and plays no role in the error analysis of the algorithm itself.

Proposition 12. *Let $\varepsilon_n \to 0$ and $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$. Under (A1)-(A4) and for $f \in L^2(\mathcal{X}, P)$, it holds that*

$$L_{\mathrm{var}} = J\left(O_P\left(\frac{1}{n}\right) + O_P\left(\frac{\gamma_n^2}{\Delta_{\varepsilon,J}^2}\right)\right).$$

where $\Delta_{\varepsilon,J} = \min_{0 \leq j \leq J}(\lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1})$.

5.3. Total loss

Fixed kernel

In kernel machine learning, it is standard to assume a *fixed* RKHS $\mathcal{H}_k$, e.g., with norm $\|\cdot\|_{\mathcal{H}_k}$ and a fixed kernel k with a bandwidth ε not dependent on n. From Propositions 5 and 12 and under assumptions (A1)-(A4), we then have the following result:

Theorem 13. *Assume $f \in \mathcal{H}_k$ with finite norm; i.e., $\|f\|_{\mathcal{H}_k} \leq M$ for some constant $M < \infty$. Then,*

$$L(f, \widehat{f}) = O(\lambda_{\varepsilon,J}) + JO_P\left(\frac{1}{n}\right) + JO_P\left(\frac{1}{n\Delta_{\varepsilon,J}^2}\right), \tag{25}$$

where $\Delta_{\varepsilon,J} = \min_{0 \leq j \leq J}(\lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1})$.

The problem is that $\mathcal{H}_k$, M, and the eigenvalues $\lambda_{\varepsilon,j}$, all depend on ε. This dependence is complicated and poorly understood. Hence, in what follows, we will instead of the RKHS norm use an alternative measure of smoothness based on the diffusion operator (Method 2 in Sec. 5.1). This simplifies the theory and will allow us to analyze the dependence of the series estimator on tuning parameters and sparse structure.

Kernel with decreasing bandwidth

Consider now a Gaussian kernel k_ε with a bandwidth ε_n that *decreases* with increasing n. From Propositions 6 and 12 and under assumptions (A1)-(A4), we have the following results:

Theorem 14. *Let $\varepsilon_n \to 0$ and $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$ as $n \to \infty$. Then, for $f \in L^2(\mathcal{X}, P)$,*

$$L(f, \widehat{f}) = O\left(\frac{\mathcal{J}_\varepsilon(f)}{\nu_{\varepsilon,J+1}^2}\right) + JO_P\left(\frac{1}{n}\right) + JO_P\left(\frac{\gamma_n^2}{\Delta_{\varepsilon,J}^2}\right), \tag{26}$$

where $\mathcal{J}_\varepsilon(f) = -\langle G_\varepsilon f, f \rangle_\varepsilon$, $\nu_{\varepsilon,J+1}^2$ is the $(J+1)^{th}$ eigenvalue of $-G_\varepsilon$, $\gamma_n = \sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}$, and $\Delta_{\varepsilon,J} = \min_{0 \leq j \leq J}(\lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1})$.

Corollary 15. *Assume that $f \in C_b^3(\mathcal{X})$ and that the kernel $k = k_\varepsilon^*$ is corrected for bias. Then, for $\varepsilon_n \to 0$ and $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$,*

$$L(f, \widehat{f}) = \frac{\mathcal{J}(f)O(1) + O(\varepsilon_n)}{\nu_{J+1}^2} + JO_P\left(\frac{1}{n}\right) + JO_P\left(\frac{\gamma_n^2}{\varepsilon_n\Delta_J^2}\right), \qquad (27)$$

where ν_{J+1}^2 is the $(J+1)^{th}$ eigenvalue of $\triangle$, $\mathcal{J}(f) = \int_{\mathcal{X}} \|\nabla f(x)\|^2 dS(x)$, and $\Delta_J = \min_{0 \le j \le J}(\nu_{j+1}^2 - \nu_j^2)$.

Some comments on these results: The first term in Eqs. 25–27 corresponds to the approximation error of the estimator and decays with J. The second and third terms correspond to the variance. Note that the variance term $JO_P\left(\frac{1}{n}\right)$ is the same as the variance of a *traditional* orthogonal series estimator in *one* dimension only; in d dimensions, the variance term for a traditional tensor product basis is $O_P\left(\frac{1}{n}\right)\prod_{i=1}^d J_i$ where J_i is the number of components in the ith direction [17]. Hence, there is a considerable gain in using an adaptive bias, but we incur an additional variance term $JO_P\left(\frac{\gamma_n^2}{\varepsilon_n\Delta_J^2}\right)$ from estimating the basis.[5]

If we balance the two ε-terms in Eq. 27, we get a bandwidth of $\varepsilon_n \asymp (1/n)^{2/(d+4)}$. With this choice of ε_n and by ignoring terms of lower order, the rate becomes

$$L(f, \widehat{f}) = O\left(\frac{\mathcal{J}(f)}{\nu_{J+1}^2}\right) + \frac{J}{\Delta_J^2}O_P\left(\frac{\log n}{n}\right)^{\frac{2}{d+4}}. \qquad (28)$$

Finally, if we apply the results in [12, 19, 20, 44] to general Riemannian manifolds (see, for example, [26, 40, 7] for kernel density estimation on manifolds), and use that the eigenvalues of the Laplace-Beltrami operator on an r-dimensional Riemannian manifold are $\nu_j^2 \sim j^{2/r}$ [46], we obtain the following corollary:

Corollary 16. *Suppose the support of the data is on a compact C^∞ submanifold of $\mathbb{R}^d$ with intrinsic dimension r, and suppose that f is smooth relative to P (Definition 7). Under the assumptions of Theorem 14 and Corollary 15, we obtain the rate*

$$L(f, \widehat{f}) = O\left(\frac{1}{J^{2/r}}\right) + J^{2\left(1-\frac{1}{r}\right)}O_P\left(\frac{\log n}{n}\right)^{\frac{2}{r+4}}.$$

It is then optimal to take $J \asymp (n/\log n)^{\frac{1}{r+4}}$, in which case the upper bound becomes

$$\left(\frac{\log n}{n}\right)^{\frac{2}{(r+4)r}}.$$

We make the following observations for a spectral series estimator with flexible kernel bandwidth:

---

[5]In an SSL setting (Remark 1), this extra estimation error vanishes in the limit of infinite unlabeled data.

(i) **Adaptiveness to Low-Dimensional Structure.** If the data in $\mathbb{R}^d$ has intrinsic dimension $r \ll d$, then the rate $n^{-1/O(r^2)}$ above is a significant improvement of the minimax rate $n^{-1/O(d)}$ for a nonparametric regressor in $\mathbb{R}^d$. Our estimator automatically adapts to sparse structure and does not require the knowledge of r or an estimated r in practice. The optimal error rate is achieved when the smoothing parameters J and ε are properly selected for the given r, and the amount of smoothing is in practice chosen by cross-validation as in Sec. 2.5.

(ii) **Minimax Optimality.** In a semi-supervised learning setting, the estimation error of the basis vanishes in the limit of infinite unlabeled data. The loss then reduces to

$$L(f, \widehat{f}) = O\left(\frac{1}{J^{2/r}}\right) + J O_P\left(\frac{1}{n}\right), \tag{29}$$

which is minimized by taking $J \asymp n^{r/(r+2)}$. At the minimum, we achieve the rate

$$\frac{1}{n^{\frac{2}{2+r}}},$$

the minimax rate for a nonparametric estimator of Sobolev smoothness $\beta = 1$ in $\mathbb{R}^D$, where $D = r$. The latter result is also, up to a logarithmic term, in agreement with [62].

6. Numerical examples

Finally, we use data with complex dependencies to compare the spectral series approach with classical kernel smoothing, k-nearest neighbors (kNN) regression, regularization in RKHS, and recent state-of-the-art manifold and local regression methods.

In our experiments, we split the data into three sets for training, validation, and testing, respectively. For the manifold regression estimators from Aswani et al. [2] and Cheng et al. [11], we use the authors' codes with built-in cross-validation. For all other estimators, we tune parameters according to Sec. 2.5. To assess the final models, we compute the estimated loss $\widehat{L}$ and standard error on the test data.[6]

6.1. Estimating pose using images of faces

In our first example, we consider images of artificial faces from the Isomap database [54].[7] There are a total of $n = 698$ 64×64 gray-scale images rendered with different orientation and lighting directions. Fig. 1 shows a visualization of these data where we use the first two non-trivial eigenvectors of the Gaussian

---

[6]The estimated standard error of $\widehat{L}$ is $s/\sqrt{n}$, where s^2 is the empirical variance of $(Y_i - \widehat{f}(X_i))^2$ on the test set.

[7]www.isomap.stanford.edu/datasets.html

diffusion kernel as coordinates (i.e., Eq. 4 with the approximate eigenvectors from Eq. 11).

Our goal is to estimate the horizontal left-right pose of each face. We compare several different approaches to regression:

(i) As a baseline, we choose the classical Nadaraya-Watson estimator with a Gaussian smoothing kernel (*NW*) and the k-nearest neighbors regression estimator (*kNN*). The latter estimator is known to be minimax optimal with respect to local intrinsic dimension [31].

(ii) For the spectral series method (*Series*), we implement the Gaussian kernel (*Series-radial*) and polynomial kernels $k(x, y) = (\langle x, y \rangle + 1)^q$ of different degrees q. We treat q as a tuning parameter and we denote the polynomial kernel with the smallest estimated loss by *Series-polyBest*. Note that choosing $q = 1$ (*Series-poly1*) is equivalent to a linear regression on eigenvectors computed with PCA.

(iii) We also implement the RKHS method in Sec. 3.2 for the same set of kernels as *Series*. For a squared-error loss, Eq. 19 reduces to an infinite-dimensional, generalized ridge regression problem [24, Section 5.8.2]. Hence, we use the term kernel ridge regression (*KRR*) and denote the estimators by *KRR-radial* and *KRR-poly*.

(iv) The last group of estimators include recent manifold and local regression methods [2, 11][8]: *locOLS* is a local ordinary least squares, *locRR* is a local ridge regression, *locEN* is a local elastic net, *locPLS* is a local partial least squares, *locPCR* is a local principal components regression, *NEDE* is the nonparametric exterior derivative estimator, *NALEDE* is the nonparametric adaptive lasso exterior derivative estimator, *NEDEP* is the nonparametric exterior derivative estimator for the "large p, small n" case, and *NALEDEP* is the nonparametric adaptive lasso exterior derivative estimator for the "large p, small n" case. The last 4 regression estimators (*NEDE, NALEDE, NEDEP, NALEDP*) pose the regression as a least-squares problem with a term that penalizes for the regression vector lying in directions perpendicular to an estimated manifold; see [2] for details. In our comparison, we also include *MALLER* [11] which first estimates the local dimension of the data and then performs local linear regression on a tangent plane estimate.

Manifold and local regression methods, unlike *Series*, quickly become computationally intractable in high dimensions. Hence, to be able to compare the different methods, we follow Aswani et al. [2] and rescale the Isomap images from from 64×64 down to 7×7 pixels in size. This reduces the number of covariates from $d = 4096$ to $d = 49$. In other words, we regress the left-right pose Y (our response) on the rescaled image $X \in \mathbb{R}^{49}$ (our predictor). We use 50% of the data for training, 25% for validation and 25% for testing. All covariates are normalized to have mean 0 and standard deviation 1.

Table 1 and Fig. 5 summarize the results of the final (cross-validated) estimators. The approaches that have best performance are *Series-radial* and *KRR-radial*. As expected, *Series* and *KRR* estimators yield similar losses. A

---

[8]For code, go to www.eecs.berkeley.edu/~aaswani/EDE_Code.zip and http://www.math.princeton.edu/~hauwu/regression.zip.

TABLE 1
Estimated loss for Isomap face data

Method	Loss (SE)
NW	1.71 (0.23)
kNN	1.74 (0.21)
Series-poly1	2.96 (0.40)
Series-polyBest (q=2)	0.22 (0.04)
Series-radial	0.16 (0.04)
KRR-poly1	2.95 (0.41)
KRR-polyBest (q=2)	0.22 (0.05)
KRR-radial	0.15 (0.04)

Method	Loss (SE)
locOLS	0.65 (0.17)
locRR	0.46 (0.16)
locEN	0.47 (0.16)
locPLS	0.65 (0.21)
locPCR	0.95 (0.20)
NEDE	0.44 (0.14)
NALEDE	0.46 (0.14)
NEDEP	0.81 (0.31)
NALEDEP	0.85 (0.33)
MALLER	0.24 (0.06)
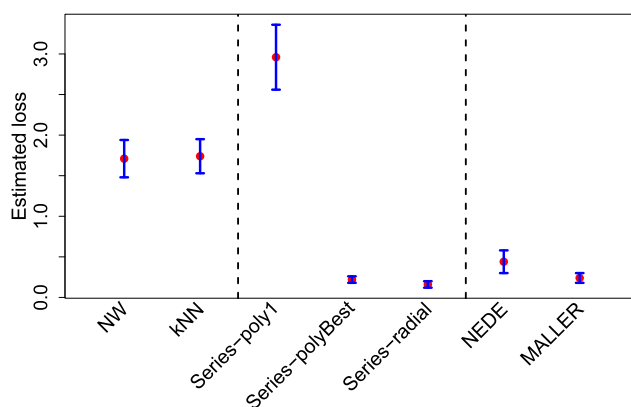
FIG 5. *Estimated loss of estimators for Isomap face data; see Table 1. For visibility, we have divided the estimators into three groups: classical NW kernel and kNN smoothers (*left*), Series/RKHS-type estimators (*center*), and manifold regression estimators, such as* NEDE *and* MALLER *(*right*). Bars represent standard errors.*

first-order polynomial kernel, i.e., a global principal component regression with *Series-* or *KRR-poly1*, performs worse than *NW* and *kNN*. Higher-order polynomial kernels (with degree $q = 2$ resulting in the smallest loss) as well as the manifold and local regression estimators (in particularly, *NEDE* and *MALLER*) improve the *NW* and *kNN* results but *Series-radial* and *KRR-radial* are still the best choices in terms of statistical and computational performance.

6.2. Estimating redshift using SDSS galaxy spectra

In the following (high-dimensional) example, we predict the redshift of galaxies from high-resolution measurements of their emission spectra. Our initial data sample consists of galaxy spectra from ten arbitrarily chosen spectroscopic plates of SDSS DR6.[9] We preprocess and remove spectra according to the three cuts

---

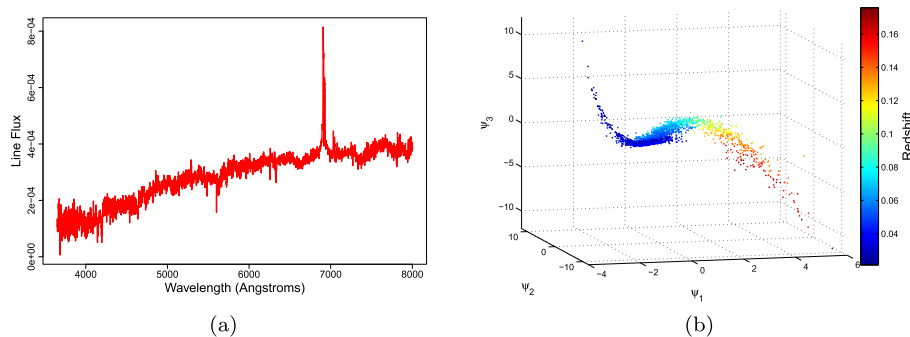[9]http://www.sdss.org/dr6/algorithms/redshift_type.html

FIG 6. *Left: Example of a SDSS galaxy spectrum. Right: Embedding of a sample of SDSS galaxy spectra using the first three non-trivial eigenvectors of the Gaussian diffusion kernel as coordinates; the color codes for the true redshift.*

described in [42]. The final sample consists of $n = 2812$ high-resolution spectra with flux measurements at $d = 3501$ wavelengths. We renormalize each spectra so that it has unit norm. Our goal is to predict a galaxy's redshift Y where the predictor is an *entire* spectrum $\mathbf{x} \in \mathbb{R}^{3501}$. Fig. 6a shows an example of a SDSS spectrum. Fig. 6b shows a low-dimensional visualization of the full data set when using the first few vectors of the diffusion basis as coordinates. Each point in the plot represents a galaxy, and the color codes for the SDSS spectroscopic redshift. The redshift (the response Y) appears to vary smoothly with the eigencoordinates.

For the regression, we use 50% of the data for training, 25% for validation and 25% for testing. Due to the high dimension of the predictor ($d = 3501$), we are unable to implement the computationally intensive manifold and local regression estimators from [2]. Table 2 and Fig. 7 summarize the results for the other approaches to regression. *Series* and *KRR* are essentially equivalent in terms of performance, and as before, the radial kernel (*Series-radial* and *KRR-radial*) yields the smallest estimated loss. For these data, a linear dimensionality reduction with PCA (*series-poly1*) improves upon the *NW* and *kNN* regression results. *MALLER* and higher-order polynomials (with degrees 5 and 6) perform better than PCA, but *Series-radial* still has the smallest estimated loss. Moreover, *MALLER* is much slower than *Series*: the former estimator takes 34 minutes on a 2.70GHz Intel Core i7-4800MQ, whereas *Series* with cross-validation takes less than a minute.

6.3. Scalability

Increasing dimension

In terms of computational speed, the spectral series estimator has a clear competitive edge in high dimensions relative local regression procedures and a least-squares (LS) implementation of Eq. 19 that does not take advantage of orthogo-

TABLE 2
Estimated loss for redshift prediction using SDSS galaxy spectra

Method	Loss (SE) $\times 10^{-5}$
NW	6.13 (1.47)
kNN	6.37 (1.52)
Series-poly1	5.13 (0.47)
Series-polyBest (q=5)	3.22 (0.32)
Series-radial	2.77 (0.33)
KRR-poly1	5.01 (0.49)
KRR-polyBest (q=6)	3.05 (0.33)
KRR-radial	2.84 (0.33)

Method	Loss (SE)
locOLS	–
locRR	–
locEN	–
locPLS	–
locPCR	–
NEDE	–
NALEDE	–
NEDEP	–
NALEDEP	–
MALLER	3.11 (0.38)
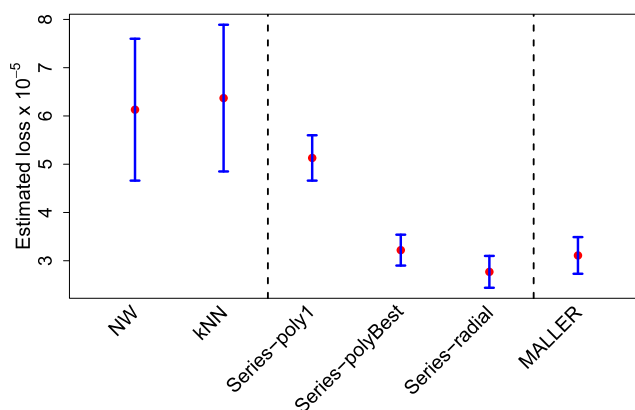
FIG 7. *Estimated loss of different estimators for redshift prediction; see Table 2. Bars represent standard errors. For visibility, we have divided the estimators into 3 groups: classical NW kernel and kNN smoothers (left), Series/RKHS-type estimators (center), and the manifold regression estimator MALLER (right). Bars represent standard errors.*

nal bases (see, e.g., [4, p. 215] for a LS implementation of SSL learning on manifolds). We illustrate the differences with a one-dimensional manifold embedded in d dimensions. Let $Y|\mathbf{x} \sim N(\theta(\mathbf{x}), 0.5)$, where the points $\mathbf{x} = (x_1, \ldots, x_d)$ lie on a unit circle in $\mathbb{R}^d$, and $\theta(\mathbf{x})$ is the angle corresponding to the position of $\mathbf{x}$. For simplicity, we simulate data uniformly on the circle; i.e., we let $\theta(\mathbf{x}) \sim Unif(0, 2\pi)$.

Figure 8 summarizes the results. In terms of estimated loss (left panel), *Series* performs better than *MALLER*, and it has a statistical performance similar to the least-squares implementation of kernel ridge regression (*KRR-LS*). As predicted by the theory, the loss of *Series* does not depend on the ambient dimension d. Moreover, the computational time of *Series* is nearly constant as a function of the dimension d (right panel). *KRR-LS* is slower than *Series*,[10] and

---

[10]Cross-validation of *Series* is fast due to the *orthogonality* of the basis. If we compute the expansion coefficients $\widehat{\beta}_{\varepsilon,j}$ (Eq. 12) for all $j \leq J_{\max}$, then we do not need to recompute
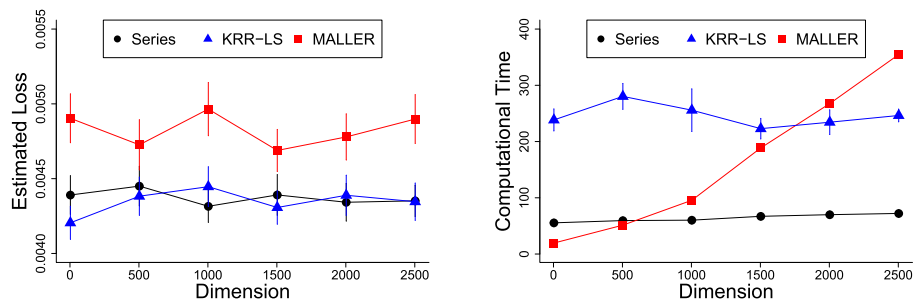
FIG 8. *Increasing the dimension (number of variables) d for a circle embedded in $\mathbb{R}^d$. Estimated loss (left) and computational time (right) as a function of the dimension d for different regression estimators.*
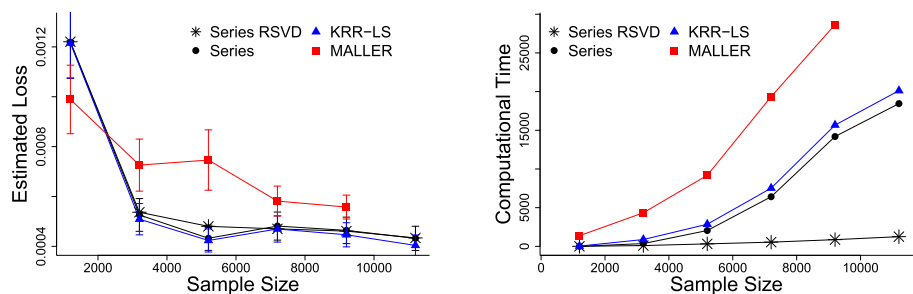

FIG 9. *Increasing the size of the training set for redshift prediction using SDSS galaxy spectra. Estimated loss (left) and computational time (right) as a function of the sample size for different regression estimators. Note that Randomized SVD (Series RSVD) dramatically reduces the computational time, see right, for large sample sizes with little impact on the statistical performance, see left.*

MALLER becomes computationally intractable as d increases. For $d = 2500$ and $n = 2000$, each fit with *MALLER* takes an average of 354 seconds (6 minutes) on an Intel i7-4800MQ CPU 2.70GHz processor, compared to 72 seconds for *Series*.

Increasing sample size

Here we revisit the redshift prediction problem in Sec. 6.2 using galaxy spectra from SDSS DR 12.[11] We increase the size of the training set for a fixed number of 1000 validation spectra and 1000 test spectra. Fig. 9 indicates massive payoffs in implementing Randomized SVD (*Series RSVD*) for large data sets; see

these coefficients for other models with the same kernel and $J \leq J_{\max}$ components in the series expansion. The least squares implementation of Eq. 19, on the other hand, requires recomputing the expansion coefficients for each choice of the smoothing parameter J.

[11]http://www.sdss.org/dr12/

discussion in Sec. 2.6. Even without parallelization, we are able to cut down the computational time with a factor of 15 (right panel) with almost no decrease in statistical performance (left panel). The run time for *SVD* and *KRR-LS* when the sample size n=11200 (and the dimension $d = 3431$) is about 5 hours on an Intel i7-4800MQ CPU 2.70GHz processor. With Randomized SVD, the same regression takes about 20 minutes.

7. Discussion

Our spectral series method can handle complex high-dimensional data objects in many settings where traditional nonparametric methods either perform poorly or are computationally intractable. The series method offers a compression of the data in terms of Fourier coefficients; it is computationally efficient (with regards to the dimension and size of the sample), and it returns orthogonal basis functions that *adapt* to low-dimensional structure in the data distribution. As a result, there is no need for cumbersome tensor products in high dimensions.

Our work shows that for a Gaussian kernel with a flexible bandwidth, the computed eigenfunctions form a Fourier-like orthogonal basis for expressing smoothness relative to the underlying data distribution. More precisely, if a function is smooth with respect to the data distribution, then it is sparse in the eigenbasis with respect to the L^2 norm, and vice versa (Theorem 10). Indeed, in the limit of the sample size $n \to \infty$, spectral series with a Gaussian kernel can be seen as a generalization of Fourier series to high dimensions and sparse structure (Sec. 4).

The two main theorems 13 and 14 provide theoretical bounds on the loss of the final regression estimator for a standard fixed RKHS setting as well as a setting where the kernel bandwidth varies with the sample size n. We show that spectral series regression with a Gaussian kernel is adaptive to *intrinsic* dimension when the bandwidth $\varepsilon_n \to 0$ (Corollary 16). In the case of a submanifold with dimension r embedded in $\mathbb{R}^d$, the convergence rate of the estimator depends on the manifold dimension r rather than the ambient dimension d. The adaption occurs automatically and does not involve manifold estimation. Unlike [8], there is also no need to estimate the dimension of the manifold. We have found that unless the goal is manifold estimation, there is little advantage in using manifold and local linear regression methods. Such methods quickly become computationally intractable in high dimensions without a prior dimension reduction. On the other hand, the computational speed of spectral series does *not* depend on the ambient or intrinsic dimension of the data. Moreover, it is unclear how manifold-based methods behave in more complex settings where there is sparse structure (e.g., high-density regions and clusters) but no well-defined submanifold.

Because of the close connection between spectral series and SVMs, we expect that our new findings (regarding adaptiveness, choice of kernel and the bandwidth) will apply to kernel-based regularized empirical risk minimizers as well. Indeed, our empirical results (Tables 1 and 2) confirm that the performance

of KRR using a Gaussian kernel with a flexible bandwidth is similar to that of spectral series regression. This suggests that one can exploit the advantages of spectral series in terms of interpretation, visualization, and analysis without any real down-sides. In the process of analyzing the performance of the spectral series estimator, we shed light on the empirical success of SVMs for sparse data, and we unify ideas from Fourier analysis, kernel machine learning and spectral clustering.

Future work includes deriving tighter bounds for the convergence rate of spectral series and kernel-based empirical regularizers. We believe that our estimated rates are on the conservative side as our derivations assume that the eigenvectors need to be accurately estimated. Empirical experiments, however, indicate that spectral series with approximate eigenvectors already outperform the k-nearest neighbor estimator which is *minimax optimal* with respect to local intrinsic dimension [31]. In separate papers [27, 28], we will discuss extensions of spectral series to estimating other unknown functions (e.g., conditional densities, density ratios and likelihoods) for high-dimensional complex data and distributions. Another interesting research question is whether one can further improve the performance of spectral series approaches by adaptive basis selection and nonlinear estimators that threshold the series expansion coefficients $|\widehat{\beta}_j|$ as in wavelet thresholding [35].

In the online supplementary materials ([29]), we include sample R code for the spectral series estimator. This code has however not been optimized for speed, as we will leave the large-scale deployment on parallel platforms to future work.

Appendix A: Proofs for bounds on the regression estimator

We start by stating some useful lemmas.

Lemma 17. *[12, Proposition 3] For $f \in C^3(\mathcal{X})$ and $x \in \mathcal{X} \setminus \partial\mathcal{X}$,*

$$-\lim_{\varepsilon \to 0} G_\varepsilon^* = \triangle.$$

If $\mathcal{X}$ is a compact C^∞ submanifold of $\mathbb{R}^d$, then $\triangle$ is the psd Laplace-Beltrami operator of $\mathcal{X}$ defined by $\triangle f(x) = -\sum_{j=1}^{r} \frac{\partial^2 f}{\partial s_j^2}(x)$, where $(s_1, \ldots, s_r)$ are the normal coordinates of the tangent plane at x.

Lemma 18. *$\forall x \in \mathcal{X}$,*

$$\frac{a}{b} \le s_\varepsilon(x) \le \frac{b}{a}$$

Proof. $\forall x \in \mathcal{X}$,

$$\frac{\inf_{x \in \mathcal{X}} p_\varepsilon(x)}{\sup_{x \in \mathcal{X}} p_\varepsilon(x)} \le s_\varepsilon(x) \le \frac{\sup_{x \in \mathcal{X}} p_\varepsilon(x)}{\inf_{x \in \mathcal{X}} p_\varepsilon(x)},$$

where $a \int_{\mathcal{X}} k_\varepsilon(x,y)dy \le p_\varepsilon(x) \le b \int_{\mathcal{X}} k_\varepsilon(x,y)dy$.   □

Lemma 19. *For $f \in L^2(\mathcal{X}, P)$,*

$$L_{\text{bias}} \leq \frac{b}{a} \sum_{j > J} |\beta_{\varepsilon,j}|^2.$$

Proof. From the orthogonality property of the basis functions ψ_j, we have that

$$\int_{\mathcal{X}} |f(x) - f_{\varepsilon,J}(x)|^2 dS_\varepsilon(x) = \sum_{j > J} |\beta_{\varepsilon,j}|^2.$$

The result follows from Lemma 18. $\square$

Lemma 20. *Under the same assumptions as in Proposition 11, it holds that*

$$\|\varphi_{\varepsilon,j} - \widehat{\varphi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)} = O_P\left(\frac{\gamma_n}{\delta_{\varepsilon,j}}\right),$$

where $\gamma_n = \sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}$ and $\delta_{\varepsilon,j} = \lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1}$.

Proof. From [19], $\sup_x |\widehat{p}_\varepsilon(x) - p_\varepsilon(x)| = O_P(\gamma_n)$. Hence,

$$\sup_x |\widehat{s}_\varepsilon(x) - s_\varepsilon(x)| = O_P(\gamma_n).$$

By using Proposition 11, we conclude that

$$
\begin{aligned}
\int_{\mathcal{X}} |\widehat{\psi}_{\varepsilon,j}(x)|^2 dP(x) &\leq 2\int_{\mathcal{X}} |\widehat{\psi}_{\varepsilon,j}(x) - \psi_{\varepsilon,j}(x)|^2 dP(x) + 2\int_{\mathcal{X}} |\psi_{\varepsilon,j}(x)|^2 dP(x) \\
&= O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right) + C,
\end{aligned}
$$

where C is a constant. Write

$$
\begin{aligned}
|\varphi_{\varepsilon,j}(x) - \widehat{\varphi}_{\varepsilon,j}(x)|^2 &= |\psi_{\varepsilon,j}(x)s_\varepsilon(x) - \widehat{\psi}_{\varepsilon,j}(x)\widehat{s}_\varepsilon(x)|^2 \\
&\leq 2|\psi_{\varepsilon,j}(x) - \widehat{\psi}_{\varepsilon,j}(x)|^2|s_\varepsilon(x)|^2 + 2|s_\varepsilon(x) - \widehat{s}_\varepsilon(x)|^2|\widehat{\psi}_{\varepsilon,j}(x)|^2.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\|\varphi_{\varepsilon,j} - \widehat{\varphi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)}^2 &\leq 2\sup_x |s_\varepsilon(x)|^2 \, \|\widehat{\psi}_{\varepsilon,j} - \psi_{\varepsilon,j}\|_{L^2(\mathcal{X},P)}^2 \\
&\quad + 2\sup_x |\widehat{s}_\varepsilon(x) - s_\varepsilon(x)|^2 \, \|\widehat{\psi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)}^2 \\
&= O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right) + O_P(\gamma_n^2)\left(O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right) + C\right) \\
&= O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right).
\end{aligned}
$$

$\square$

Lemma 21. $\forall\, 0 \leq j \leq J$, *it holds that*

$$\left| \frac{1}{n} \sum_{i=1}^{n} Y_i(\widehat{\varphi}_{\varepsilon,j}(X_i) - \varphi_{\varepsilon,j}(X_i)) - \int_{\mathcal{X}} f(x)(\widehat{\varphi}_{\varepsilon,j}(x) - \varphi_{\varepsilon,j}(x))dP(x) \right| = O_P\left(\frac{1}{\sqrt{n}} \right).$$

Proof. Let $S = \frac{1}{n}\sum_{i=1}^{n} Y_i(\widehat{\varphi}_{\varepsilon,j}(X_i) - \varphi_{\varepsilon,j}(X_i))$ and $I = \int_{\mathcal{X}} f(x)(\widehat{\varphi}_{\varepsilon,j}(x) - \varphi_{\varepsilon,j}(x))dP(x)$. According to Chebyshev's inequality, for any $M > 0$,

$$\mathbb{P}\left(|S - I| \geq M \mid \widetilde{X}_1, \ldots, \widetilde{X}_n \right) \quad \leq \quad \frac{\mathbb{V}\left(S - I \mid \widetilde{X}_1, \ldots, \widetilde{X}_n \right)}{M^2}$$

$$\leq \quad \frac{\mathbb{V}\left(Y_1(\widehat{\varphi}_{\varepsilon,j}(X_1) - \varphi_{\varepsilon,j}(X_1)) \mid \widetilde{X}_1, \ldots, \widetilde{X}_n \right)}{nM^2}.$$

Hence, for any $M > 0$,

$$\mathbb{P}\left(|S - I| \geq M \right) \quad \leq \quad \frac{\mathbb{V}\left(Y_1(\widehat{\varphi}_{\varepsilon,j}(X_1) - \varphi_{\varepsilon,j}(X_1)) \right)}{nM^2}$$

$$\leq \quad \frac{\sigma}{nM^2} \left(\mathbb{E}|\widehat{\varphi}_{\varepsilon,j}(X_1) - \varphi_{\varepsilon,j}(X_1)|^2 \right)^{1/2},$$

where we in the last inequality apply the Cauchy-Schwarz inequality. Under assumption (A4), we conclude the result of the lemma. $\qquad\square$

A.1. Bias

Proof of Proposition 5. Let $\widetilde{\psi}_{\varepsilon,0}, \widetilde{\psi}_{\varepsilon,1}, \ldots$ be the eigenfunctions of the symmetric operator $\widetilde{A}_\varepsilon(f)(x) = \int_{\mathcal{X}} \widetilde{a}_\varepsilon(x,y)f(y)dP(y)$. It follows from Mercer's theorem that

$$\mathcal{H}_{\varepsilon,M} = \left\{ f = \sum_j \gamma_{\varepsilon,j}\widetilde{\psi}_{\varepsilon,j} : \sum_j \frac{|\gamma_{\varepsilon,j}|^2}{\lambda_{\varepsilon,j}} \leq M \right\}.$$

If $f \in \mathcal{H}_{\varepsilon,M}$, then we can bound the bias with respect to the eigenbasis $\widetilde{\psi}$:

$$\sum_{j>J} |\gamma_{\varepsilon,j}|^2 \leq \lambda_{\varepsilon,J} \sum_{j>J} \frac{|\gamma_{\varepsilon,j}|^2}{\lambda_{\varepsilon,j}} \leq \lambda_{\varepsilon,J} \sum_j \frac{|\gamma_{\varepsilon,j}|^2}{\lambda_{\varepsilon,j}} \leq M\lambda_{\varepsilon,J}.$$

By construction,

$$\gamma_{\varepsilon,j} = \int_{\mathcal{X}} f(x)\widetilde{\psi}_{\varepsilon,j}(x)dP(x)$$

and

$$\beta_{\varepsilon,j} = \int_{\mathcal{X}} f(x)\psi_{\varepsilon,j}(x)dS_\varepsilon(x) = \int_{\mathcal{X}} f(x)\widetilde{\psi}_{\varepsilon,j}(x)\sqrt{s_\varepsilon(x)}dP(x) \leq \sqrt{\frac{b}{a}}|\gamma_{\varepsilon,j}|.$$

Thus, $\sum_{j>J} |\beta_{\varepsilon,j}|^2 \leq \frac{b}{a} \sum_{j>J} |\gamma_{\varepsilon,j}|^2 = O(M\lambda_{\varepsilon,J})$. The result follows from Lemma 19. $\qquad\square$

Proof of Proposition 6. Note that $\mathcal{J}_\varepsilon(f) = \sum_j \nu_{\varepsilon,j}^2 |\beta_{\varepsilon,j}|^2$. Hence,

$$\frac{\mathcal{J}_\varepsilon(f)}{\nu_{\varepsilon,J+1}^2} = \sum_j \frac{\nu_{\varepsilon,j}^2}{\nu_{\varepsilon,J+1}^2} |\beta_{\varepsilon,j}|^2 \geq \sum_{j>J} \frac{\nu_{\varepsilon,j}^2}{\nu_{\varepsilon,J+1}^2} |\beta_{\varepsilon,j}|^2 \geq \int_\mathcal{X} |f(x) - f_{\varepsilon,J}(x)|^2 dS_\varepsilon(x).$$

The last result follows from Lemma 19. $\qquad\qquad\qquad\qquad\qquad\qquad\Box$

Proof of Lemma 8. By Green's first identity

$$\int_\mathcal{X} f \, \nabla^2 f dS(x) + \int_\mathcal{X} \nabla f \cdot \nabla f dS(x) = \oint_{\partial\mathcal{X}} f(n \cdot \nabla f) dS(x) = 0,$$

where n is the normal direction to the boundary $\partial\mathcal{X}$, and the last surface integral vanishes due to the Neumann boundary condition. It follows from Lemma 17 that

$$\begin{aligned}
\lim_{\varepsilon \to 0} \mathcal{J}_\varepsilon^*(f) &= -\lim_{\varepsilon \to 0} \int_\mathcal{X} f(x) G_\varepsilon^* f(x) dS_\varepsilon(x) \\
&= \int_\mathcal{X} f(x) \triangle f(x) dS(x) = \int_\mathcal{X} \|\nabla f(x)\|^2 dS(x). \qquad\Box
\end{aligned}$$

Proof of Theorem 10. We have that

$$c^2 \geq \int_\mathcal{X} \|\nabla f(x)\|^2 dS(x) = \int_\mathcal{X} f(x) \triangle f(x) dS(x) = \sum_j \nu_j^2 \beta_j^2,$$

where $\nu_j^2 = O(j^{2s})$. Hence, $f \in W_\mathcal{B}(s, c)$ and by Theorem 9.1 in [35], $\|f - f_J\|^2 = o(J^{-2s})$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\Box$

A.2. Variance

Let $\mathcal{H}$ be an auxiliary RKHS of smooth functions; we use the term "auxiliary" to denote that the space only enters the intermediate derivations and plays no role in the error analysis of the algorithm itself. We define the two integral operators $A_\mathcal{H}, \widehat{A}_\mathcal{H} : \mathcal{H} \to \mathcal{H}$ where

$$\begin{aligned}
A_\mathcal{H} f(x) &= \frac{\int k_\varepsilon(x,y)\langle f, K(\cdot,y)\rangle_\mathcal{H} dP(y)}{\int k_\varepsilon(x,y) dP(y)} = \int a_\varepsilon(x,y)\langle f, K(\cdot,y)\rangle_\mathcal{H} \, dP(y) \\
\widehat{A}_\mathcal{H} f(x) &= \frac{\sum_{i=1}^n k_\varepsilon(x,X_i)\langle f, K(\cdot,X_i)\rangle_\mathcal{H}}{\sum_{i=1}^n k_\varepsilon(x,X_i)} = \int \widehat{a}_\varepsilon(x,y)\langle f, K(\cdot,y)\rangle_\mathcal{H} \, d\widehat{P}_n(y),
\end{aligned}$$

and K is the reproducing kernel of $\mathcal{H}$. Define the operator norm $\|A\|_\mathcal{H} = \sup_{f \in \mathcal{H}} \|Af\|_\mathcal{H}/\|f\|_\mathcal{H}$ where $\|f\|_\mathcal{H}^2 = \langle f, f\rangle_\mathcal{H}$. Now suppose the weight function k_ε is sufficiently smooth with respect to $\mathcal{H}$ (Assumption 1 in [44]); this condition is for example satisfied by a Gaussian kernel on a compact support $\mathcal{X}$. By Propositions 13.3 and 14.3 in [44], we can then relate the functions $\psi_{\varepsilon,j}$

and $\widehat{\psi}_{\varepsilon,j}$, respectively, to the eigenfunctions $u_{\varepsilon,j}$ and $\widehat{u}_{\varepsilon,j}$ of $A_{\mathcal{H}}$ and $\widehat{A}_{\mathcal{H}}$. We have that

$$\|\psi_{\varepsilon,j} - \widehat{\psi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)} = C_1\|u_{\varepsilon,j} - \widehat{u}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)} \leq C_2\|u_{\varepsilon,j} - \widehat{u}_{\varepsilon,j}\|_{\mathcal{H}} \qquad (30)$$

for some constants C_1 and C_2. According to Theorem 6 in [43] for eigenprojections of positive compact operators, it holds that

$$\|u_{\varepsilon,j} - \widehat{u}_{\varepsilon,j}\|_{\mathcal{H}} \leq \frac{\|A_{\mathcal{H}} - \widehat{A}_{\mathcal{H}}\|_{\mathcal{H}}}{\delta_{\varepsilon,j}}, \qquad (31)$$

where $\delta_{\varepsilon,j}$ is proportional to the eigengap $\lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1}$. As a result, we can bound the difference $\|\psi_{\varepsilon,j} - \widehat{\psi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)}$ by controlling the deviation $\|A_{\mathcal{H}} - \widehat{A}_{\mathcal{H}}\|_{\mathcal{H}}$.

We choose the auxiliary RKHS $\mathcal{H}$ to be a Sobolev space with a sufficiently high degree of smoothness (see below for details). Let $\mathcal{H}^s$ denote the Sobolev space of order s with vanishing gradients at the boundary; that is, let

$$\mathcal{H}^s = \{f \in L^2(\mathcal{X}) \mid D^\alpha f \in L^2(\mathcal{X})\ \forall |\alpha| \leq s,\ D^\alpha f|_{\partial \mathcal{X}} = 0\ \forall |\alpha| = 1\},$$

where $D^\alpha f$ is the weak partial derivative of f with respect to the multi-index α, and $L^2(\mathcal{X})$ is the space of square integrable functions with respect to the Lebesgue measure. Let $C_b^3(\mathcal{X})$ be the set of uniformly bounded, three times differentiable functions with uniformly bounded derivatives whose gradients vanish at the boundary. Now consider $\mathcal{H} \subset \mathcal{H}^s$ and choose s large enough so that $D^\alpha f \in C_b^3(\mathcal{X})$ for all $f \in \mathcal{H}$ and $|\alpha| = s$. Under assumptions (A1)-(A4), we derive the following result:

Lemma 22. *Let* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^{d/2}/\log(1/\varepsilon_n) \to \infty$. *Then* $\|A_{\mathcal{H}} - \widehat{A}_{\mathcal{H}}\|_{\mathcal{H}} = O_P(\gamma_n)$, *where* $\gamma_n = \sqrt{\frac{\log(1/\varepsilon_n)}{n\varepsilon_n^{d/2}}}$.

Proof. Uniformly, for all $f \in C_b^3(\mathcal{X})$, and all x in the support of P,

$$|A_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)| \leq |A_\varepsilon f(x) - \widetilde{A}_\varepsilon f(x)| + |\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)|$$

where $\widetilde{A}_\varepsilon f(x) = \int \widehat{a}_\varepsilon(x,y)f(y)dP(y)$. From [19],

$$\sup_x \frac{|\widehat{p}_\varepsilon(x) - p_\varepsilon(x)|}{|\widehat{p}_\varepsilon(x)p_\varepsilon(x)|} = O_P(\gamma_n).$$

Hence,

$$\begin{aligned}
|A_\varepsilon f(x) - \widetilde{A}_\varepsilon f(x)| &\leq& \frac{|\widehat{p}_\varepsilon(x) - p_\varepsilon(x)|}{|\widehat{p}_\varepsilon(x)p_\varepsilon(x)|} \int |f(y)|k_\varepsilon(x,y)dP(y) \\
&=& O_P(\gamma_n) \int |f(y)|k_\varepsilon(x,y)dP(y) \\
&=& O_P(\gamma_n).
\end{aligned}$$

Next, we bound $\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x)$. We have

$$\widetilde{A}_\varepsilon f(x) - \widehat{A}_\varepsilon f(x) = \int f(y)\widehat{a}_\varepsilon(x,y)(d\widehat{P}_n(y) - dP(y))$$

$$= \frac{1}{p(x) + o_P(1)} \int f(y) k_\varepsilon(x, y)(d\widehat{P}_n(y) - dP(y)).$$

Now, expand $f(y) = f(x) + r_n(y)$ where $r_n(y) = (y - x)^T \nabla f(u_y)$ and u_y is between y and x. So,

$$\int f(y) k_\varepsilon(x, y)(d\widehat{P}_n(y) - dP(y))$$

$$= f(x) \int k_\varepsilon(x, y)(d\widehat{P}_n(y) - dP(y)) + \int r_n(y) k_\varepsilon(x, y)(d\widehat{P}_n(y) - dP(y)).$$

By an application of Talagrand's inequality to each term, as in Theorem 5.1 of [20], we have

$$\int f(y) k_\varepsilon(x, y)(d\widehat{P}_n(y) - dP(y)) = O_P(\gamma_n).$$

Thus, $\sup_{f \in C_b^3(\mathcal{X})} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_\infty = O_P(\gamma_n)$.

The Sobolev space $\mathcal{H}$ is a Hilbert space with respect to the scalar product

$$\langle f, g \rangle_{\mathcal{H}} = \langle f, g \rangle_{L^2(\mathcal{X})} + \sum_{|\alpha|=s} \langle D^\alpha f, D^\alpha g \rangle_{L^2(\mathcal{X})}.$$

We have that

$$
\begin{aligned}
\sup_{f \in \mathcal{H}: \|f\|_{\mathcal{H}}=1} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_{\mathcal{H}}^2 &\leq \sup_{f \in \mathcal{H}} \sum_{|\alpha| \leq s} \|D^\alpha(\widehat{A}_\varepsilon f - A_\varepsilon f)\|_{L^2(\mathcal{X})}^2 \\
&= \sum_{|\alpha| \leq s} \sup_{f \in \mathcal{H}} \|\widehat{A}_\varepsilon D^\alpha f - A_\varepsilon D^\alpha f\|_{L^2(\mathcal{X})}^2 \\
&\leq \sum_{|\alpha| \leq s} \sup_{f \in C_b^3(\mathcal{X})} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_{L^2(\mathcal{X})}^2 \\
&\leq C \sup_{f \in C_b^3(\mathcal{X})} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_\infty^2.
\end{aligned}
$$

for some constant C. Hence,

$$
\begin{aligned}
\sup_{f \in \mathcal{H}} \frac{\|\widehat{A}_\varepsilon f - A_\varepsilon f\|_{\mathcal{H}}}{\|f\|_{\mathcal{H}}} &= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}}=1} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_{\mathcal{H}} \\
&\leq C' \sup_{f \in C_b^3(\mathcal{X})} \|\widehat{A}_\varepsilon f - A_\varepsilon f\|_\infty = O_P(\gamma_n). \qquad \square
\end{aligned}
$$

Proof of Proposition 11. From Eqs. 30-31, we have that

$$\|\psi_{\varepsilon,j} - \widehat{\psi}_{\varepsilon,j}\|_{L^2(\mathcal{X}, P)} \leq C \frac{\|A_{\mathcal{H}} - \widehat{A}_{\mathcal{H}}\|_{\mathcal{H}}}{\lambda_{\varepsilon,j} - \lambda_{\varepsilon,j+1}}$$

for some constant C that does not depend on n. The result follows from Lemma 22. $\qquad \square$

Lemma 23. $\forall 0 \leq j \leq J$,

$$|\widehat{\beta}_{\varepsilon,j} - \beta_{\varepsilon,j}|^2 = O_P\left(\frac{1}{n}\right) + O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right).$$

Proof. Note that $\psi_{\varepsilon,j}(x)s_\varepsilon(x) = \varphi_{\varepsilon,j}(x)$ and

$$
\begin{aligned}
\widehat{\beta}_{\varepsilon,j} &= \frac{1}{n}\sum_{i=1}^n Y_i\widehat{\psi}_{\varepsilon,j}(X_i)\widehat{s}_\varepsilon(X_i) \\
&= \frac{1}{n}\sum_{i=1}^n Y_i\varphi_{\varepsilon,j}(X_i) + \frac{1}{n}\sum_{i=1}^n Y_i\left(\widehat{\varphi}_{\varepsilon,j}(X_i) - \varphi_{\varepsilon,j}(X_i)\right) \\
&= \beta_{\varepsilon,j} + O_P\left(\frac{1}{\sqrt{n}}\right) + \frac{1}{n}\sum_{i=1}^n Y_i\left(\widehat{\varphi}_{\varepsilon,j}(X_i) - \varphi_{\varepsilon,j}(X_i)\right).
\end{aligned}
$$

Let $S = \frac{1}{n}\sum_{i=1}^n Y_i\left(\widehat{\varphi}_{\varepsilon,j}(X_i) - \varphi_{\varepsilon,j}(X_i)\right)$ and $I = \int_{\mathcal{X}} f(x)(\widehat{\varphi}_{\varepsilon,j}(x) - \varphi_{\varepsilon,j}(x))dP(x)$. We conclude that

$$
\begin{aligned}
\frac{1}{2}|\widehat{\beta}_{\varepsilon,j} - \beta_{\varepsilon,j}|^2 &\leq O_P\left(\frac{1}{n}\right) + |S - I|^2 + |I|^2 \\
&\leq O_P\left(\frac{1}{n}\right) + |S - I|^2 + \left(\int_{\mathcal{X}} |f(x)|^2 dP(x)\right)\|\varphi_{\varepsilon,j} - \widehat{\varphi}_{\varepsilon,j}\|_{L^2(\mathcal{X},P)}^2 \\
&= O_P\left(\frac{1}{n}\right) + O_P\left(\frac{\gamma_n^2}{\delta_{\varepsilon,j}^2}\right),
\end{aligned}
\tag{32}
$$

where the second inequality follows from the Cauchy-Schwarz inequality, and the last equality is due to Lemmas 21 and 20. □

Proof of Proposition 12. Let $\widetilde{f}_{\varepsilon,J}(x) = \sum_{j=0}^J \beta_{\varepsilon,j}\widehat{\psi}_{\varepsilon,j}(x)$. Write

$$
\begin{aligned}
|f_{\varepsilon,J}(X_i) - \widehat{f}_{\varepsilon,J}(X_i)|^2 &= |f_{\varepsilon,J}(X_i) - \widetilde{f}_{\varepsilon,J}(X_i) + \widetilde{f}_{\varepsilon,J}(X_i) - \widehat{f}_{\varepsilon,J}(X_i)|^2 \\
&\leq 2|f_{\varepsilon,J}(X_i) - \widetilde{f}_{\varepsilon,J}(X_i)|^2 + 2|\widetilde{f}_{\varepsilon,J}(X_i) - \widehat{f}_{\varepsilon,J}(X_i)|^2.
\end{aligned}
$$

We bound the contribution to L_{var} from each of these two terms separately:

By using Cauchy's inequality and Proposition 11, we have that

$$\int_{\mathcal{X}} |f_{\varepsilon,J}(x) - \widetilde{f}_{\varepsilon,J}(x)|^2 dP(x) = \int_{\mathcal{X}} \left|\sum_{j=0}^J \beta_{\varepsilon,j}(\psi_{\varepsilon,j}(x) - \widehat{\psi}_{\varepsilon,j}(x))\right|^2 dP(x)$$

$$\leq \left(\sum_{j=0}^J |\beta_{\varepsilon,j}|^2\right) \cdot \sum_{j=0}^J \left(\int_{\mathcal{X}} |\psi_{\varepsilon,j}(x) - \widehat{\psi}_{\varepsilon,j}(x)|^2 dP(x)\right) = J\, O_P\left(\frac{\gamma_n^2}{\Delta_{\varepsilon,J}^2}\right).$$

By construction, it holds that $\frac{1}{n}\sum_i \widehat{\psi}_{\varepsilon,j}(\widetilde{X}_i)\widehat{\psi}_{\varepsilon,\ell}(\widetilde{X}_i)\widehat{s}_\varepsilon(\widetilde{X}_i) = \delta_{j,\ell}$. Furthermore,

$$\int_{\mathcal{X}} \widehat{\psi}_{\varepsilon,j}(x)\widehat{\psi}_{\varepsilon,\ell}(x)d\widehat{S}_\varepsilon(x) = \frac{1}{n}\sum_i \widehat{\psi}_{\varepsilon,j}(X_i)\widehat{\psi}_{\varepsilon,\ell}(X_i)\widehat{s}_\varepsilon(X_i) + O_P\left(\frac{1}{\sqrt{n}}\right)$$

$$= \frac{1}{n}\sum_i \widehat{\psi}_{\varepsilon,j}(\widetilde{X}_i)\widehat{\psi}_{\varepsilon,\ell}(\widetilde{X}_i)\widehat{s}_{\varepsilon}(\widetilde{X}_i) + O_P\left(\frac{1}{\sqrt{n}}\right)$$

$$= \delta_{j,\ell} + O_P\left(\frac{1}{\sqrt{n}}\right).$$

for a sample $X_1, \ldots, X_n$ drawn independently from $\widetilde{X}_1, \ldots, \widetilde{X}_n$. Finally, from the orthogonality property of the $\widehat{\psi}_{\varepsilon,j}$'s together with Lemmas 23 and 18, it follows that

$$\int_{\mathcal{X}} |\widetilde{f}_{\varepsilon,J}(x) - \widehat{f}_{\varepsilon,J}(x)|^2 dP(x)$$

$$= \int_{\mathcal{X}} \frac{1}{\widehat{s}_{\varepsilon}(x)} \left| \sum_{j=0}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})\widehat{\psi}_{\varepsilon,j}(x)\sqrt{\widehat{s}_{\varepsilon}(x)} \right|^2 dP(x)$$

$$= \int_{\mathcal{X}} \frac{1}{\widehat{s}_{\varepsilon}(x)} \left(\sum_{j=0}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})^2\widehat{\psi}_{\varepsilon,j}^2(x)d\widehat{S}_{\varepsilon}(x) \right)$$

$$+ \int_{\mathcal{X}} \frac{1}{\widehat{s}_{\varepsilon}(x)} \left(\sum_{j=0}^{J} \sum_{\ell=0,\ell\neq j}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})(\beta_{\varepsilon,\ell} - \widehat{\beta}_{\varepsilon,\ell})\widehat{\psi}_{\varepsilon,j}(x)\widehat{\psi}_{\varepsilon,\ell}(x)d\widehat{S}_{\varepsilon}(x) \right)$$

$$\leq \frac{b}{a} \sum_{j=0}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})^2 \left(\int_{\mathcal{X}} \widehat{\psi}_{\varepsilon,j}^2(x)d\widehat{S}_{\varepsilon}(x) \right)$$

$$+ \frac{b}{a} \sum_{j=0}^{J} \sum_{\ell=0,\ell\neq j}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})(\beta_{\varepsilon,\ell} - \widehat{\beta}_{\varepsilon,\ell}) \left(\int_{\mathcal{X}} \widehat{\psi}_{\varepsilon,j}(x)\widehat{\psi}_{\varepsilon,\ell}(x)d\widehat{S}_{\varepsilon}(x) \right)$$

$$= \frac{b}{a} \sum_{j=0}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})^2 \left(1 + O_P\left(\frac{1}{\sqrt{n}}\right) \right)$$

$$+ \frac{b}{a} \sum_{j=0}^{J} \sum_{\ell=0,\ell\neq j}^{J} (\beta_{\varepsilon,j} - \widehat{\beta}_{\varepsilon,j})(\beta_{\varepsilon,\ell} - \widehat{\beta}_{\varepsilon,\ell}) O_P\left(\frac{1}{\sqrt{n}}\right)$$

$$= J\left(O_P\left(\frac{1}{n}\right) + O_P\left(\frac{\gamma_n^2}{\Delta_{\varepsilon,J}^2}\right) \right).$$

The result follows. $\square$

Acknowledgments

Supplementary Material

R code for the spectral series estimator
(doi: 10.1214/16-EJS1112SUPP; .zip).

References

[1] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*(3), 337–404. MR0051437

[2] ASWANI, A., P. BICKEL, and C. TOMLIN (2011). Regression on manifolds: Estimation of the exterior derivative. *Annals of Statistics 39*(1), 48–81. MR2797840

[3] BELKIN, M. and P. NIYOGI (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation 6*(15), 1373–1396.

[4] BELKIN, M. and P. NIYOGI (2005a). Semi-supervised learning on Riemannian manifolds. *Machine Learning 56*, 209–239.

[5] BELKIN, M. and P. NIYOGI (2005b). Towards a theoretical foundation for Laplacian-based manifold methods. In *Proc. Conf. on Learning Theory*, Volume 3559, pp. 486–500. MR2203282

[6] BELKIN, M., P. NIYOGI, and V. SINDHWANI (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research 7*, 2399–2434. MR2274444

[7] BERRY, T. and T. SAUER (2016). Density estimation on manifolds with boundary. *preprint arXiv:1511.08271v3*.

[8] BICKEL, P. J. and B. LI (2007). Local polynomial regression on unknown manifolds. In *IMS Lecture Notes–Monograph Series, Complex Datasets and Inverse Problems*, Volume 54, pp. 177–186. Institute of Mathematical Statisitcs. MR2459188

[9] BOUSQUET, O., O. CHAPELLE, and M. HEIN (2003). Measure based regularization. In *Adv. in Neural Inf. Processing Systems*.

[10] CANDÈS, E. and T. TAO (2005). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.*.

[11] CHENG, M. Y. and H. T. WU (2013). Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association 108*, 1421–1434. MR3174718

[12] COIFMAN, R. and S. LAFON (2006a). Diffusion maps. *Applied and Computational Harmonic Analysis 21*, 5–30. MR2238665

[13] COIFMAN, R. and S. LAFON (2006b). Geometric harmonics. *Applied and Computational Harmonic Analysis 21*, 31–52. MR2238666

[14] COIFMAN, R., S. LAFON, A. LEE, M. MAGGIONI, B. NADLER, F. WARNER, and S. ZUCKER (2005). Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps. *Proc. of the National Academy of Sciences 102*(21), 7426–7431.

[15] CUCKER, F. and D. ZHOU (2007). *Learning Theory: An Approximation Theory Viewpoint.* Cambridge University Press. MR2354721

[16] DONOHO, D. and C. GRIMES (2003, May). Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proc. of the National Academy of Sciences 100*(10), 5591–5596. MR1981019

[17] EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Application.* Springer. MR1705298

[18] FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics 21*, 196–216. MR1212173

[19] GINÉ, E. and A. GUILLOU (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann Inst. H. Poincaré 38*, 907–921. MR1955344

[20] GINÉ, E. and V. KOLTCHINSKII (2006). Empirical graph Laplacian approximation of Laplace-Beltrami operators: Large sample results. In *High Dimensional Probability: Proceedings of the Fourth International Conference*, IMS Lecture Notes, pp. 1–22. MR2387756

[21] GIROSI, F., M. JONES, and T. POGGIO (1995). Regularization theory and neural network architectures. *Neural Computation 7*, 219–269.

[22] GRIGOR'YAN, A. (2006). Heat kernels on weighted manifolds and applications. *Cont. Math. 398*, 93–191. MR2218016

[23] HALKO, N., P. G. MARTINSSON, and J. A. TROPP (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review 53*(2), 217–288.MR2806637

[24] HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN (2009). *The Elements of Statistical Learning* (Second ed.). Springer. MR2722294

[25] HEIN, M., J.-Y. AUDIBERT, and U. VON LUXBURG (2005). Intrinsic dimensionality estimation of submanifolds in R^d. In *Proc. of the 22nd Int'l Conf on Machine learning*.

[26] HENRY, G. and D. RODRIGUEZ (2009). Kernel density estimation on riemannian manifolds: Asymptotic results. *Journal of Mathematical Imaging and Vision 34*(3), 235–239. MR2515446

[27] IZBICKI, R., A. LEE, and C. SCHAFER (2014). High-dimensional density ratio estimation with extensions to approximate likelihood computation. *Journal of Machine Learning Research (AISTATS Track)*, 420–429.

[28] IZBICKI, R. and A. LEE (2016). Conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics, to appear*.

[29] IZBICKI, R. and A. LEE (2016). Supplement to "A spectral series approach to high-dimensional nonparametric regression." DOI: 10.1214/16-EJS1112SUPP.

[30] JOLLIFFE, I.T. (2002). *Principal Component Analysis.* Springer. MR2036084

[31] KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pp. 729–737. The MIT Press.

[32] LAFFERTY, J. and L. WASSERMAN (2008). Rodeo: Sparse, greedy nonparametric regression. *Annals of Statistics 36*(1), 28–63. MR2387963

[33] LAFON, S. and A. LEE (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans. Pattern Anal. and Mach. Intel. 28*, 1393–1403.

[34] LEE, A. and L. WASSERMAN (2010). Spectral connectivity analysis. *Journal of the American Statistical Association 105*(491), 1241–1255. MR2752618

[35] MALLAT, S. (2009). *A Wavelet Tour of Signal Processing* (3rd ed.). Academic Press. MR2479996

[36] MEILA, M. and J. SHI (2001). A random walks view on spectral segmentation. In *Proc. Eighth International Conference on Artificial Intelligence and Statistics*.

[37] MINH, H. Q., P. NIYOGI, and Y. YAO (2006). Mercer's theorem, feature maps, and smoothing. In *Learning Theory, 19th Annual Conference on Learning Theory*.

[38] NADLER, B., A. SREBRO, and X. ZHOU (2009). Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data.

[39] OLSHAUSEN, B. A. and D. J. FIELD (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 381*(6583), 607–609.

[40] OZAKIN, A. and A. GRAY (2009). Submanifold density estimation. In *Adv. in Neural Inf. Processing Systems*, pp. 1375–1382.

[41] RAVIKUMAR, P., J. LAFFERTY, H. LIU, and L. WASSERMAN (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B 71*(5), 1009–1030. MR2750255

[42] RICHARDS, J. W., P. E. FREEMAN, A. B. LEE, and C. M. SCHAFER (2009). Exploiting low-dimensional structure in astronomical spectra. *Astrophysical Journal 691*, 32–42.

[43] ROSASCO, L., M. BELKIN, and E. D. VITO (2008). A note on perturbation results for learning empirical operators. CSAIL Technical Report TR-2008-052, CBCL-274, Massachusetts Institute of Technology.

[44] ROSASCO, L., M. BELKIN, and E. D. VITO (2010). On learning with integral operators. *Journal of Machine Learning Research 11*, 905–934. MR2600634

[45] SAERENS, M., F. FOUSS, L. YEN, and P. DUPONT (2004). The principal components analysis of a graph, and its relationships to spectral clustering. In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Lecture Notes in Artificial Intelligence*, pp. 371–383. Springer-Verlag.

[46] SAFAROV, Y. and D. VASSILIEV (1996). *The Asymptotic Distribution of Eigenvalues of Partial Differential Operators*, Volume 155 of *Translations of Mathematical Monographs*. American Mathematical Society. MR1414899

[47] SCHOENBERG, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics 39*(4), 811–841. MR1503439

[48] SCHÖLKOPF, B., A. SMOLA, and K. R. MÜLLER (1997). Kernel principal

component analysis. In *Artificial Neural Networks – ICANN'97*, pp. 583–588. Springer.

[49] SCHÖLKOPF, B. and A. J. SMOLA (2001). *Learning with Kernels*. MIT Press.

[50] SHI, T., M. BELKIN, and B. YU (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics 37*(6B), 3960–3984. MR2572449

[51] SINGER, A. (2006). From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis 21*, 128–134. MR2238670

[52] SLEPIAN, D. (1983). Some comments on fourier analysis, uncertainty and modeling. *SIAM Review 25*(3), 379–393. MR0710468

[53] STEINWART, I. and A. CHRISTMANN (2008). *Support Vector Machines*. Springer. MR2450103

[54] TENENBAUM, J. B., V. DE SILVA, and J. C. LANGFORD (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*(5500), 2319–2323.

[55] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological 58*, 267–288. MR1379242

[56] VAPNIK, V. (1996). *Statistical Learning Theory*. Wiley. MR1641250

[57] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*(4), 395–416. MR2409803

[58] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM. MR1045442

[59] WOLD, S., M. SJÖSTRÖM, and L. ERIKSSON (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems 58*, 109–130.

[60] WU, Q., Y. YING, and D.-X. ZHOU (2007). Multi-kernel regularized classifiers. *Journal of Complexity 23*, 108–134. MR2297018

[61] YUAN, M. and Y. LIN (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*(1), 49–67. MR2212574

[62] ZHOU, X. and N. SREBRO (2011). Error analysis of Laplacian eigenmaps for semi-supervised learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Volume 15, pp. 892–900.

[63] ZHU, X., Z. GHAHRAMANI, and J. LAFFERTY (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML-03, 20th International Conference on Machine Learning*.

[64] ZHU, X. and A. B. GOLDBERG (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning 3*(1), 1–130.

[65] ZOU, H. and T. HASTIE (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*(2), 301–320. MR2137327