Appendix of “Nonparametric Conditional Density Estimation in a High-Dimensional Regression Setting”

Rafael Izbicki* and Ann B. Lee†

A.1 Data Visualization via Spectral Series

As a by-product of the eigenvector approach, the series estimator provides a useful tool for visualizing and organizing complex non-standard data. Figure 1 shows a two-dimensional embedding of the luminous red galaxy data (Section 4, main manuscript) using the first two basis functions $\psi_1(x)$ and $\psi_2(x)$ as coordinates; i.e., we consider the eigenmap $x \mapsto (\psi_1(x), \psi_2(x))$. The eigenfunctions capture the structure of the data and vary smoothly with the response $z$ (redshift). Similar data points are grouped together in the eigenmap; that is, samples with similar covariates are mapped to similar eigenfunction values (Coifman et al., 2005); see, for example, points A and B, or C and D in the figure. As a result, when $f(z|x)$ is smooth as a function of $x$ (Assumption 5 in Section 3 or Assumption 5’ in Appendix A.5), the spectral series estimator yields good results. In addition, because distances in the eigenmap reflect similarity, the eigenvectors themselves are useful for detecting clusters and outliers in the data.

---

*Department of Statistics, Federal University of São Carlos, Brazil.
†Department of Statistics, Carnegie Mellon University, USA.
Figure 1: Top: Embedding of the luminous red galaxies from SDSS when using the first two eigenvectors of the Gaussian kernel operator as coordinates. Bottom: Covariates of 4 selected galaxies. The eigenfunctions capture the connectivity (similarity) structure of the data and vary smoothly with the response (redshift).

A.2 Details on the Galaxy Data

Due to the expansion of the Universe, the wavelength of a photon increases as it travels to us. The redshift of a galaxy is defined as

$$z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} - 1,$$

where $\lambda_{\text{obs}}$ and $\lambda_{\text{emit}}$, respectively, are the wavelengths of a photon when observed by us and when emitted by the galaxy.

In spectroscopy, the flux of a galaxy, i.e., the number of photons emitted per unit area per unit time, is measured as a function of wavelength. The light is collected in bins of width $\sim 1$ Å, where $1$ Å = $10^{-10}$ m. Because transitions of electrons within atoms leads to spikes and troughs in spectra of width $\sim 1$ Å, and because these transitions occur at precisely known wavelengths, one can determine the redshift of a galaxy with great precision via spectroscopy.

On the other hand, in photometry – a low-resolution but less costly alternative to spectroscopy – the photons are collected into a few wavelength bins (also named bands). When performing
photometry, astronomers generally convert photon flux into a magnitude via a logarithmic transformation. Whereas fluxes depend on the details of the observing instrument, (well-calibrated) magnitudes from different instruments can be directly compared. Photometry for the Sloan Digital Sky Survey (SDSS) is carried out in five wavelength bands, denoted $u$, $g$, $r$, $i$, and $z$, that span the visible part of the electromagnetic spectrum. Each band is $\sim 1000 \text{ Å}$ wide. The differences between contiguous magnitudes (or colors; e.g., $g - r$) are useful predictors for the redshift of the galaxy. Our objective is to estimate the conditional density $f(z|x)$, where $x$’s are the observed colors of a galaxy. We train our model using redshifts obtained via spectroscopy and test our method on the following sets of galaxy data:

**Luminous Red Galaxies from SDSS.** This data set consists of 3,000 luminous red galaxies (LRGs) from the SDSS.\(^1\) We use information about 3 colors ($g - r$, $r - i$, and $i - z$) in psf, fiber, petrosian and model magnitudes;\(^2\) that is, there are $3 \cdot 4 = 12$ covariates. For more details about the data, see Freeman et al. (2009).

**Galaxies from Multiple Surveys.** This data set, used by Sheldon et al. (2012), includes information on model and cmodel magnitudes of $u$, $g$, $r$, $i$, and $z$ bands for galaxies from multiple surveys, including SDSS, DEEP2\(^3\) (Lin et al., 2004) and others. As in Sheldon et al. (2012), we base our estimates on the color and raw $r$-band magnitude. Hence, there are $2 \cdot 4 + 2 = 10$ covariates. We use random subsets of different sizes of these data for the various experiments.

**COSMOS.** This data set consists of magnitudes for 752 galaxies that are measured in 42 different photometric bands from 7 magnitude systems (T. Dahlen 2013, private communication).\(^4\) For each system, we compute the differences between adjacent bands. This yields a total of $d = 37$ colors.

### A.3 Details on Simulated Data

The following schemes were used to generate data plotted in Figure 3:

**Data on Manifold.** Data is generated according to $Z|x \sim N(\mathbf{x}, 0.5)$, where $\mathbf{X} = (X_1, \ldots, X_{20})$.

---

\(^1\)http://www.sdss.org/

\(^2\)Multiple estimators of magnitude exist that differ in algorithmic details.

\(^3\)http://deep.ps.uci.edu/

\(^4\)http://cosmos.astro.caltech.edu/
Few Relevant Covariates. Let $Z|X \sim N\left(\frac{1}{p} \sum_{i=1}^{p} x_i, 0.5\right)$, where $X = (X_1, \ldots, X_d) \sim N(0, I_d)$. Here only the first $p$ covariates influences the response (i.e., the conditional density is sparse) but there is no sparse (low-dimensional) structure in $\mathcal{X}$.

\section{A.4 Variations on the Kernel Operator}

Several variants of the operator from Equation 1 in the main manuscript exist in the spectral methods literature; see e.g. von Luxburg (2007) for normalization schemes common in spectral clustering. In this section, we describe a non-symmetric variation of the kernel operator, referred to as the \textit{diffusion operator} (e.g., Lee and Wasserman, 2010). Although the diffusion operator leads to similar empirical performance as the kernel operator of Equation 1, the former operator has better understood theoretical properties, especially in the limit of the bandwidth $\varepsilon \to 0$ (Coifman and Lafon, 2006, Belkin and Niyogi, 2008). This in turn allows us to analyze the dependence of the spectral series estimator on tuning parameters, which ultimately leads to tighter bounds on the loss compared to the kernel PCA operator.

In this section, we let the kernel $K$ be a local, radially symmetric function $K_\varepsilon(x, y) = g\left(d(x, y)/\sqrt{\varepsilon}\right)$, where the elements $K_\varepsilon(x, y)$ are positive and bounded for all $x, y \in \mathcal{X}$. We use the notation $K_\varepsilon$ to emphasize the dependence of $K$ on the bandwidth. The first step is to renormalize the kernel according to

$$a_\varepsilon(x, y) = \frac{K_\varepsilon(x, y)}{p_\varepsilon(x)},$$

where $p_\varepsilon(x) = \int K_\varepsilon(x, y)dP(y)$. We refer to $a_\varepsilon(x, y)$ as the \textit{diffusion kernel} (Meila and Shi, 2001). As in Lee and Wasserman (2010), we define a “diffusion operator” $A_\varepsilon$ as

$$A_\varepsilon(h)(x) = \int_{\mathcal{X}} a_\varepsilon(x, y)h(y)dP(y). \quad (1)$$

The operator $A_\varepsilon$ has a discrete set of non-negative eigenvalues $\lambda_{\varepsilon, 0} = 1 \geq \lambda_{\varepsilon, 1} \geq \ldots \geq 0$ with associated eigenfunctions $(\psi_{\varepsilon, i})_i$. The eigenfunctions are orthogonal with respect to the weighted $L^2$ inner product $\langle f, g \rangle_\varepsilon = \int_{\mathcal{X}} f(x)g(x)dS_\varepsilon(x)$, where $S_\varepsilon(A) = \frac{\int_A p_\varepsilon(x)dP(x)}{\int p_\varepsilon(x)dP(x)}$ can be interpreted as a
smoothed version of $P$. See Lee and Wasserman (2010) for additional interpretation and details on how to estimate these quantities.

The construction of both the tensor basis $\{\Psi_{i,j}\}_{i,j}$ and the conditional density estimator proceeds just as in the case when using the operator from Equation 1. The only difference is that as $\{\Psi_{i,j}\}_{i,j}$ are now orthonormal with respect to $\lambda \times S_\epsilon$ instead of $\lambda \times P$, the coefficients of the projection are given by

$$\beta_{i,j} = \int \int f(z|x)\Psi_{i,j}(z,x) dS_\epsilon(x) dz = \int \int f(z|x)\Psi_{i,j}(z,x)s_\epsilon(x) dP(x) dz$$

$$= \int \int \Psi_{i,j}(z,x)s_\epsilon(x) dP(x,z) = \mathbb{E}[\Psi_{i,j}(Z,X)s_\epsilon(X)].$$

### A.5 Bounds for Spectral Series Estimator with Kernel PCA Basis

Smoothness in $x$ can be defined in many ways. The standard approach in the theoretical Support Vector Machine (SVM) literature (Steinwart and Christmann, 2008) is to consider the norm of a fixed Reproducing Kernel Hilbert Space (RKHS). Here we derive rates for a spectral series conditional density estimator with a fixed kernel, using the operator of Equation 1 in the main manuscript. As before, we assume:

**Assumption 1.** $\int f^2(z|x)dP(x)dz < \infty$.

**Assumption 2.** $M_\phi \overset{\text{def}}{=} \sup_z \sup_i \phi_i(z) < \infty$.

**Assumption 3.** $\lambda_1 > \lambda_2 > \ldots > \lambda_J > 0$.

**Assumption 4.** (Smoothness in $z$ direction) $\forall x \in \mathcal{X}$, let

$$h_x : [0,1] \rightarrow \mathbb{R}$$

$$h_x(z) = f(z|x).$$

*Note that $h_x(z)$ is simply $f(z|x)$ viewed as a function of $z$ for a fixed $x$. We assume

$$h_x \in W_\phi(s_x, c_x),$$
where \( s_x \) and \( c_x \) are such that \( \inf_x s_x \overset{\text{def}}{=} \beta > \frac{1}{2} \) and \( \int_X c_x^2 dP(x) < \infty \).

Let \( \mathcal{H}_K \) be the Reproducing Kernel Hilbert Space (RKHS) associated to the kernel \( K \). Instead of Assumption 5, we here assume

**Assumption 5’. Smoothness in \( x \) direction**

$$\forall z \in [0, 1], \ f(z|x) \in \left\{ g \in \mathcal{H}_K : \| g \|^2_{\mathcal{H}_K} \leq c_z^2 \right\},$$

where \( f(z|x) \) is viewed as a function of \( x \), and the \( c_z \) values are such that \( c_K \overset{\text{def}}{=} \int_{[0,1]} c_z^2 dz < \infty \).

In other words, we enforce smoothness in the \( x \) direction by requiring \( f(z|x) \) to be in a RKHS for all \( z \)'s. Smaller values of \( c_K \) indicate smoother functions. The reader is referred to, e.g., Minh et al. (2006) for an account of measuring smoothness through norms in RKHSs.

Let \( n \) be the sample size of the data used to estimate the coefficients \( \beta_{i,j} \), and let \( m \) be the sample size of the data used to estimate the basis functions. In the next section, we prove the following:

**Theorem 1.** Let \( \hat{f}_{I,J}(z|x) \) be the spectral series estimator with cutoffs \( I \) and \( J \), based on the eigenfunctions of the operator in Equation (1). Under Assumptions 1-A.5 and 5’, the loss

$$L(\hat{f}_{I,J}, f) = IJ \times \left[ O_P \left( \frac{1}{n} \right) + O_P \left( \frac{1}{\lambda_J \Delta_J^2 m} \right) \right] + c_K O(\lambda_J) + O \left( \frac{1}{I^2} \right),$$

(2)

where \( \Delta_J = \min_{1 \leq j \leq J} |\lambda_j - \lambda_{j+1}| \).

Some interpretation: The first term of the bound of Theorem 1 corresponds to the sampling error of the estimator. The second and third terms correspond to the approximation error. Note that in \( d+1 \) dimensions, the variance of traditional orthogonal series methods is \( \prod_{i=1}^{d+1} I_i \times O_P \left( \frac{1}{n} \right) \), where \( I_i \) is the number of components used for the \( i \):th variable (Efromovich, 1999). Interestingly, the term \( IJ \times O_P \left( \frac{1}{n} \right) \) for our estimator is the variance term of a traditional series estimator with one tensor product only; i.e., this is the variance for the traditional case with only one explanatory variable. The cost of estimating the basis adds the term \( IJ \times O_P \left( (\lambda_J \Delta_J^2 m)^{-1} \right) \) to our rate.

To provide some insight regarding the effect of estimating the basis on the final estimator, we work out the details of two examples where the eigenvalues follow a polynomial decay \( \lambda_j \overset{\sim}{=} J^{-2\alpha} \) for some \( \alpha > \frac{1}{2} \). See Ji et al. (2012) for some empirical motivations for polynomial decay rates.
and Steinwart et al. (2009) for theory and examples. The constant \( \alpha \) is typically related to the dimensionality of the data.

**Example 1 (Supervised Learning).** For a power-law eigenvalue decay, the eigengap \( \lambda_J - \lambda_{J+1} = O(J^{-2\alpha-1}) \). If we use the same data for estimating the basis and the coefficients of the expansion, i.e., if \( n = m \), then assuming a fixed kernel \( K \), the optimal cutoffs for the bound from Theorem 1 are \( I \asymp n^{\frac{\alpha}{2\alpha + \beta + 3}} \) and \( J \asymp n^{\frac{\beta}{2\alpha + \beta + 3}} \), yielding the rate \( O_P\left(n^{-\frac{2\alpha\beta}{2\alpha + \beta + 3}}\right) \).

**Example 2 (Semi-Supervised Learning).** Suppose that we have additional unlabeled data which can be used to learn the structure of the data distribution. In the limit of infinite unlabeled data, \( m \to \infty \), the optimal cutoffs are \( I \asymp n^{\frac{\alpha}{2\alpha + \beta + 3}} \) and \( J \asymp n^{\frac{\beta}{2\alpha + \beta + 3}} \), yielding the rate \( O_P\left(n^{-\frac{2\alpha\beta}{2\alpha + \beta + 3}}\right) \).

By comparing the rates from Examples 1 and 2, we see that estimating the basis \( (\psi_j)_j \) decreases the rate of convergence. On the other hand, the possibility of using unlabeled data alleviates the problem. As an illustration, let the RKHS \( \mathcal{H} \) in Assumption A.5 be the isotropic Sobolev space with smoothness \( s > d \) and \( \beta = s \); i.e., assume that \( f(z|x) \) belongs to a Sobolev space with the same smoothness in both directions. It is known that under certain conditions on the domain \( \mathcal{X} \), the eigenvalues \( \lambda_J \asymp J^{-2s/d} \) (see, e.g., Steinwart et al., 2009, Koltchinskii and Yuan, 2010). In this setting, the rate when \( m = n \) is \( O_P\left(n^{-\frac{2s}{2s + (1+d)}}\right) \). Notice similar rates (also not minimax optimal) are obtained when learning regression functions via RKHS’s; see, e.g., Ye and Zhou, 2008 and Steinwart et al., 2009. In the limit of infinite unlabeled data, however, we achieve the rate \( O_P\left(n^{-\frac{2s}{2s + (1+d)}}\right) \), which is the standard minimax rate for estimating functions in \( d+1 \) dimensions (recall that the conditional density is defined in \( d+1 \) dimensions) (Stone, 1982, Hoffmann and Lepski, 2002).

**A.5.1 Proofs**

To simplify the proofs, we assume the functions \( \psi_1, \ldots, \psi_J \) are estimated using an unlabeled sample \( \tilde{X}_1, \ldots, \tilde{X}_m \), drawn independently from the sample used to estimate the coefficients \( \beta_{i,j} \). Without loss of generality, this can be achieved by splitting the labeled sample in two. This split is only for theoretical purposes; in practice using all data to estimate the basis leads to better results.
The technique also allows us to derive bounds for the semi-supervised learning setting described in the paper, and better understand the additional cost of estimating the basis. Define the following quantities:

\[
f_{I,J}(z|x) = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{i,j} \phi_i(z) \psi_j(x), \quad \beta_{i,j} = \int \int \phi_i(z) \psi_j(x) f(z,x) dx dz
\]

\[
\hat{f}_{I,J}(z|x) = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{\beta}_{i,j} \phi_i(z) \hat{\psi}_j(x), \quad \hat{\beta}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \phi_i(z_k) \hat{\psi}_j(x_k).
\]

Note that

\[
\int \int \left( \hat{f}_{I,J}(z|x) - f(z|x) \right)^2 dP(x) dz 
\leq \int \int \left( \hat{f}_{I,J}(z|x) - f_{I,J}(z|x) + f_{I,J}(z|x) - f(z|x) \right)^2 dP(x) dz 
\leq 2 \left( VAR(\hat{f}_{I,J}, f_{I,J}) + B(f_{I,J}, f) \right).
\]

where \(B(f_{I,J}, f) := \int \int (f_{I,J}(z|x) - f(z|x))^2 dP(x) dz\) can be interpreted as a bias term (or approximation error) and \(VAR(\hat{f}_{I,J}, f_{I,J}) := \int \int (\hat{f}_{I,J}(z|x) - f_{I,J}(z|x))^2 dP(x) dz\) can be interpreted as a variance term. First we bound the variance.

**Lemma 1.** \(\forall 1 \leq j \leq J\),

\[
\int \left( \hat{\psi}_j(x) - \psi_j(x) \right)^2 dP(x) = O_P \left( \frac{1}{\lambda_j \delta_j^2 m} \right),
\]

where \(\delta_j = \lambda_j - \lambda_{j+1}\).

For a proof of Lemma 1 see for example Sinha and Belkin (2009).

**Lemma 2.** \(\forall 1 \leq j \leq J\), there exists \(C < \infty\) that does not depend on \(m\) such that

\[
E \left[ \left( \hat{\psi}_j(X) - \psi_j(X) \right)^2 \right] < C,
\]

where \(X \sim P(x)\) is independent of the sample used to construct \(\hat{\psi}_j\).
Proof. Let $\delta \in (0, 1)$. From Sinha and Belkin (2009), it follows that

$$\mathbb{P} \left( \int \left( \tilde{\psi}_j(x) - \psi_j(x) \right)^2 dP(x) > \frac{16 \log \left( \frac{2}{\delta} \right)}{\delta^2 m} \right) < \delta,$$

and therefore $\forall \epsilon > 0$,

$$\mathbb{P} \left( \int \left( \tilde{\psi}_j(x) - \psi_j(x) \right)^2 dP(x) > \epsilon \right) < 2e^{-\frac{\delta^2 \epsilon}{16m}}.$$

Hence

$$\mathbb{E} \left[ \left( \tilde{\psi}_j(X) - \psi_j(X) \right)^2 \right] = \mathbb{E} \left[ \int \left( \tilde{\psi}_j(x) - \psi_j(x) \right)^2 dP(x) \right] = \int_0^\infty \mathbb{P} \left( \int \left( \tilde{\psi}_j(x) - \psi_j(x) \right)^2 dP(x) > \epsilon \right) \, d\epsilon \leq \int 2e^{-\frac{\delta^2 \epsilon}{16m}} \, d\epsilon < \int 2e^{-\frac{\delta^2 \epsilon}{16m}} \, d\epsilon < \infty \quad \square$$

Lemma 3. $\forall 1 \leq j \leq J$ and $\forall 1 \leq j \leq J$, there exists $C < \infty$ that does not depend on $m$ such that

$$\mathbb{E} \left[ V \left[ \phi_i(Z) \left( \tilde{\psi}_j(X) - \psi_j(X) \right) \mid \tilde{X}_1, \ldots, \tilde{X}_m \right] \right] < C$$

Proof. Using that $\phi$ is bounded (Assumption 2), it follows that

$$\mathbb{E} \left[ V \left[ \phi_i(Z) \left( \tilde{\psi}_j(X) - \psi_j(X) \right) \mid \tilde{X}_1, \ldots, \tilde{X}_m \right] \right] \leq V \left[ \phi_i(Z) \left( \tilde{\psi}_j(X) - \psi_j(X) \right) \right] \leq \mathbb{E} \left[ \phi_i^2(Z) \left( \tilde{\psi}_j(X) - \psi_j(X) \right)^2 \right] \leq K \mathbb{E} \left[ \left( \tilde{\psi}_j(X) - \psi_j(X) \right)^2 \right]$$

for some $K < \infty$. The result follows from Lemma 2. \quad \square

Lemma 4. $\forall 1 \leq j \leq J$ and $\forall 1 \leq j \leq J$,

$$\left[ \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \left( \tilde{\psi}_j(X_k) - \psi_j(X_k) \right) - \int \int \phi_i(z) \left( \tilde{\psi}_j(x) - \psi_j(x) \right) dP(z, x) \right]^2 = O_P \left( \frac{1}{n} \right)$$
Proof. Let \( A = \iint \phi_i(z) \left( \hat{\psi}_j(x) - \psi_j(x) \right) dP(z, x) \). By Chebyshev’s inequality it holds that \( \forall M > 0 \)

\[
\mathbb{P} \left( \left| \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \left( \hat{\psi}_j(X_k) - \psi_j(X_k) \right) - A \right| > M \mid \bar{X}_1, \ldots, \bar{X}_m \right) \leq \frac{1}{nM} \mathbb{E} \left[ \phi_i(Z) \left( \hat{\psi}_j(X) - \psi_j(X) \right) \right] \left( \bar{X}_1, \ldots, \bar{X}_m \right].
\]

The conclusion follows from taking an expectation with respect to the unlabeled samples on both sides of the equation and using Lemma 3.

Note that the \( \hat{\psi}'s \) are random functions, and therefore the proof of Lemma 4 relies on the fact that these functions are estimated using a different sample than \( X_1, \ldots, X_n \).

Lemma 5. \( \forall 1 \leq j \leq J \) and \( \forall 1 \leq i \leq I \),

\[
\left( \hat{\beta}_{i,j} - \beta_{i,j} \right)^2 = O_P \left( \frac{1}{n} \right) + O_P \left( \frac{1}{\lambda_j \delta_j^2 m} \right).
\]

Proof. It holds that

\[
\frac{1}{2} \left( \hat{\beta}_{i,j} - \beta_{i,j} \right)^2 \leq \left( \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \psi_j(X_k) - \beta_{i,j} \right)^2 + \left( \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \left( \hat{\psi}_j(X_k) - \psi_j(X_k) \right) \right)^2.
\]

The first term is \( O_P \left( \frac{1}{n} \right) \). Let \( A \) be as in the proof of Lemma 4. By using Cauchy-Schwartz’s inequality and Lemma 4, the second term divided by two is bounded by

\[
\frac{1}{2} \left( \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \left( \hat{\psi}_j(X_k) - \psi_j(X_k) \right) - A + A \right)^2
\]

\[
\leq \left( \frac{1}{n} \sum_{k=1}^{n} \phi_i(Z_k) \left( \hat{\psi}_j(X_k) - \psi_j(X_k) \right) - A \right)^2 + A^2.
\]

\[
\leq O_P \left( \frac{1}{n} \right) + \left( \iint \phi_i(z)^2 dP(z, x) \right) \left( \iint \left( \hat{\psi}_j(x) - \psi_j(x) \right)^2 dP(z, x) \right).
\]

The result follows from Lemma 1 and the orthogonality of \( \phi_i \).
Lemma 6. [Sinha and Belkin 2009, Corollary 1] Under the stated assumptions,

\[ \int \hat{\psi}_j^2(x) dP(x) = O_P \left( \frac{1}{\lambda_j \Delta_j^2 m} \right) + 1 \]

and

\[ \int \hat{\psi}_i(x) \hat{\psi}_j(x) dP(x) = O_P \left( \frac{1}{\sqrt{\lambda_i}} + \frac{1}{\sqrt{\lambda_j}} \right) \frac{1}{\Delta_j \sqrt{m}} \]

where \( \Delta_j = \min_{1 \leq j \leq J} \delta_j \).

Lemma 7. Let \( h(z|x) = \sum_{i=1}^I \sum_{j=1}^J \beta_{i,j} \phi_i(z) \hat{\psi}_j(x) \). Then

\[ \int \int |\tilde{f}_{I,J}(z|x) - h(z|x)|^2 dP(x) dz = IJ \left( O_P \left( \frac{1}{n} \right) + O_P \left( \frac{1}{\lambda J \Delta_j^2 m} \right) \right). \]

Proof.

\[ \int \int |\tilde{f}_{I,J}(z|x) - h(z|x)|^2 dP(x) dz \]

\[ = \sum_{i=1}^I \sum_{j=1}^J \sum_{t=1}^J (\hat{\beta}_{i,j} - \beta_{i,j}) (\hat{\beta}_{i,t} - \beta_{i,t}) \int \hat{\psi}_j(x) \hat{\psi}_t(x) dP(x) \]

\[ \leq \sum_{i=1}^I \sum_{j=1}^J (\hat{\beta}_{i,j} - \beta_{i,j})^2 \int \hat{\psi}_j^2(x) dP(x) + \]

\[ + \sum_{i=1}^I \sum_{j=1}^J \sum_{l=1, l \neq j}^J (\hat{\beta}_{i,j} - \beta_{i,j}) (\hat{\beta}_{i,l} - \beta_{i,l}) \int \hat{\psi}_j(x) \hat{\psi}_l(x) dP(x) \leq \]

\[ \sum_{i=1}^I \sum_{j=1}^J (\hat{\beta}_{i,j} - \beta_{i,j})^2 \int \hat{\psi}_j^2(x) dP(x) + \]

\[ + \left[ \sum_{i=1}^I \sum_{j=1}^J (\hat{\beta}_{i,j} - \beta_{i,j})^2 \right] \left[ \sum_{j=1}^J \sum_{t=1, t \neq j}^J (\int \hat{\psi}_j(x) \hat{\psi}_t(x) dP(x))^2 \right], \]

where the last inequality follows from repeatedly using Cauchy-Schwartz. The result follows from Lemmas 5 and 6.

\[ \square \]

Lemma 8. Let \( h(z|x) \) be as in Lemma 7. Then
\[
\iint |h(z|x) - f_{I,J}(z|x)|^2 dP(x) dz = JOP\left(\frac{1}{\lambda_J \Delta_J m}\right).
\]

**Proof.** Using Cauchy-Schwartz inequality,

\[
\iint |h(z|x) - f_{I,J}(z|x)|^2 dP(x) dz \leq \left(\sum_{j=1}^J \sum_{i=1}^I \beta_{i,j} \phi_i(z) \left(\psi_j(x) - \hat{\psi}_j(x)\right)\right)^2 dP(x) dz
\]

\[
= \left\{\sum_{j=1}^J \left(\sum_{i=1}^I \beta_{i,j} \phi_i(z)\right)^2 \right\} \left\{\sum_{j=1}^J \int \left[\psi_j(x) - \hat{\psi}_j(x)\right]^2 dP(x)\right\}.
\]

The conclusion follows from Lemma 1 and by noticing that \(\sum_{j=1}^J \sum_{i=1}^I \beta_{i,j}^2 \leq ||f(z|x)||^2 < \infty\).

Using the results above, we can now bound the variance term:

**Theorem 2.** Under the stated assumptions,

\[
VAR(\hat{f}_{I,J}, f_{I,J}) = IJ \left(\frac{1}{n} + O_P\left(\frac{1}{\lambda_J \Delta_J^2 m}\right)\right).
\]

**Proof.** Let \(h\) be defined as in Lemma 7. We have

\[
\frac{1}{2} VAR(\hat{f}_{I,J}, f_{I,J}) = \frac{1}{2} \iint \left|\hat{f}_{I,J}(z|x) - h(z|x) + h(z|x) - f_{I,J}(z|x)\right|^2 dP(x) dz
\]

\[
\leq \iint \left|\hat{f}_{I,J}(z|x) - h(z|x)\right|^2 dP(x) dz + \iint \left|h(z|x) - f_{I,J}(z|x)\right|^2 dP(x) dz.
\]

The conclusion follows from Lemmas 7 and 8.

We next bound the bias term.

**Lemma 9.** For each \(z \in [0, 1]\), expand \(g_z(x)\) in the basis \(\psi : g_z(x) = \sum_{j \geq 1} \alpha_j^z \psi_j(x)\), where \(\alpha_j^z = \ldots\)
\[ \int g_z(x)\psi_j(x) dP(x). \] We have

\[ \alpha_j^z = \sum_{i \geq 1} \beta_{i,j} \phi_i(z) \text{ and } \int (\alpha_j^z)^2 dz = \sum_{i \geq 1} \beta_{i,j}^2. \]

**Proof.** The result follows from projecting \( \alpha_j^z \) onto the basis \( \phi \). \( \Box \)

Similarly, we have the following:

**Lemma 10.** For each \( x \in X \), expand \( h_x(z) \) in the basis \( \phi : h_x(z) = \sum_{i \geq 1} \alpha_i^x \phi_i(z), \) where \( \alpha_i^x = \int h_x(z) \phi_i(z) dz \). We have

\[ \alpha_i^x = \sum_{j \geq 1} \beta_{i,j} \psi_j(x) \text{ and } \int (\alpha_i^x)^2 dP(x) = \sum_{j \geq 1} \beta_{i,j}^2. \]

**Lemma 11.** Using the same notation as Lemmas 9 and 10, we have

\[ \beta_{i,j} = \int \alpha_i^x \psi_j(x) dP(x) = \int \alpha_j^z \phi_i(z) dz. \]

**Proof.** The result follows from plugging the definitions of \( \alpha_i^x \) and \( \alpha_j^z \) into the expressions above and recalling the definition of \( \beta_{i,j} \). \( \Box \)

**Lemma 12.** \( \sum_{i \geq I} \int (\alpha_i^x)^2 dP(x) = O \left( \frac{1}{I^2} \right) \).

**Proof.** By Lemma, 10, \( h_x(z) = \sum_{i \geq 1} \alpha_i^x \phi_i(z) \). As by Assumption 4 \( h_x \in W_\phi(s_x, c_x) \),

\[ \sum_{i \geq I} i^{2s_x} (\alpha_i^x)^2 \leq \sum_{i \geq I} i^{2s_x} (\alpha_i^x)^2 \leq c_x^2. \]

Hence

\[ \sum_{i \geq I} \int (\alpha_i^x)^2 dP(x) \leq \int \frac{c_x^2}{I^{2s_x}} dP(x) \leq \frac{1}{I^{2s}} c^2. \]

**Lemma 13.** \( \sum_{j \geq J} \int (\alpha_j^z)^2 dz = c_K O(\lambda_J) \).
Proof. Note that $||h_z(.)||_H^2 = \sum_{j \geq 1} \left( \frac{\alpha_j^z}{\lambda_j} \right)^2$ (Minh, 2010). Using Assumption 5' and that the eigenvalues are decreasing it follows that

$$\sum_{j \geq J} (\alpha_j^z)^2 = \sum_{j \geq J} (\alpha_j^z)^2 \frac{\lambda_j}{\lambda_j} \leq \lambda_J ||h_z(.)||_H^2 \leq \lambda_J c_z^2,$$

and therefore $\sum_{j \geq J} \int (\alpha_j^z)^2 dz \leq \lambda_J \int_z c_z^2 dz = c_K O(\lambda_J)$. \hfill \Box

**Theorem 3.** Under the stated assumptions, the bias is bounded by

$$B(f_{I,J}, f) = c_K O(\lambda_J) + O \left( \frac{1}{T^{2\beta}} \right).$$

*Proof.* By using orthogonality, we have that

$$B(f_{I,J}, f) \overset{\text{def}}{=} \int \int (f(z|x) - f_{I,J}(z|x))^2 dP(x)dz \leq \sum_{i> J} \sum_{i \geq 1} \beta_{i,j}^2 + \sum_{i> I} \sum_{j \geq 1} \beta_{i,j}^2$$

$$= \sum_{j \geq J} \int (\alpha_j^z)^2 dz + \sum_{i \geq I} \int (\alpha_i^x)^2 dP(x),$$

where the last equality comes from Lemmas 9 and 10. The theorem follows from Lemmas 12 and 13. \hfill \Box

By putting together Theorems 2 and 3 according to the bias-variance decomposition of Equation 3, we arrive at Theorem 1 in the appendix.

### A.6 Proofs for the Spectral Series Estimator with Diffusion Basis

The proof of Theorem 1 in the main manuscript follows the same principles as the derivations in the last section. The main differences are:

1. The orthogonality of the diffusion basis is defined with respect to the stationary measure $S$ instead of $P$;

2. The bounds on the eigenfunctions of Lemma 1 have to be adapted for the diffusion basis;
3. The bias term should take into account the new smoothness assumption in the \( x \) direction.

Below we present the results that handle these differences. We assume \( n = m \) unless otherwise stated, and we make the following additional regularity assumptions:

**(RC1)** \( P \) has compact support \( \mathcal{X} \) and bounded density \( 0 < a \leq p(x) \leq b < \infty, \forall x \in \mathcal{X} \).

**(RC2)** The weights are positive and bounded; that is, \( \forall x, y \in \mathcal{X}, 0 < m \leq k(x, y) \leq M \), where \( m \) and \( M \) are constants that do not depend on \( \epsilon \).

**(RC3)** \( \forall 0 \leq j \leq J \) and \( X \sim P \), there exists some constant \( C < \infty \) (not depending on \( n \)) such that

\[
\mathbb{E} \left[ \hat{\varphi}_{\epsilon,j}(X) - \varphi_{\epsilon,j}(X) \right]^2 < C,
\]

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

In the proofs that follow, we handle the first difference by the following lemma:

**Lemma 14.** \( \forall x \in \mathcal{X}, \)

\[
\frac{a}{b} \leq s_{\epsilon}(x) \leq \frac{b}{a}
\]

**Proof.** \( \forall x \in \mathcal{X}, \)

\[
\frac{\inf_{x \in \mathcal{X}} p_{\epsilon}(x)}{\sup_{x \in \mathcal{X}} p_{\epsilon}(x)} \leq s_{\epsilon}(x) \leq \frac{\sup_{x \in \mathcal{X}} p_{\epsilon}(x)}{\inf_{x \in \mathcal{X}} p_{\epsilon}(x)},
\]

where \( a \int_{\mathcal{X}} k_{\epsilon}(x, y) dy \leq p_{\epsilon}(x) \leq b \int_{\mathcal{X}} k_{\epsilon}(x, y) dy. \)

Now, let

\[
G_{\epsilon} = \frac{A_{\epsilon} - I}{\epsilon}, \tag{4}
\]

where \( I \) is the identity operator. The operator \( G_{\epsilon} \) has the same eigenvectors \( \psi_{\epsilon,j} \) as the differential operator \( A_{\epsilon} \). Its eigenvalues are given by \( -\nu_{\epsilon,j}^2 = \frac{2}{\epsilon} \lambda_{\epsilon,j}^{-1} \), where \( \lambda_{\epsilon,j} \) are the eigenvalues of \( A_{\epsilon} \). Define the functional

\[
\mathcal{J}_{\epsilon}(f) = -\langle G_{\epsilon} f, f \rangle_{\epsilon} \tag{5}
\]

which maps a function \( f \in L^2(\mathcal{X}, P) \) into a non-negative real number. For small \( \epsilon \), this functional measures the variability of the function \( f \) with respect to the distribution \( P \).

The following result bounds the approximation error for an orthogonal series expansion of a given function \( f \).
Proposition 1. For $f \in L^2(\mathcal{X}, P)$,

$$
\int_{\mathcal{X}} |f(x) - f_{\epsilon,J}(x)|^2 dP(x) \leq O \left( \frac{\mathcal{J}_\epsilon(f)}{\nu_{\epsilon,J+1}^2} \right)
$$

where $-\nu_{\epsilon,J+1}^2$ is the $(J + 1)^{th}$ eigenvalue of $G_\epsilon$, and $f_{\epsilon,J}$ is the projection of $f$ into the $J$ first eigenfunctions.

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

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

The results follows from Lemma 14. $\square$

The total bias bound, $O \left( \frac{\mathcal{J}_\epsilon(f)}{\nu_{\epsilon,J+1}^2} \right) + O \left( I^{-2\beta} \right)$, is derived in the same fashion as the bound from Theorem 3, where we put together the bound from the bias on the $x$ direction (derived from Proposition 1) with the bias in the $z$ direction (derived in Lemma 12). The following two additional results will be used to derive the bounds in the case $\epsilon \to 0$.

Denote the quantities derived from the bias-corrected kernel $k_\epsilon^*$ by $A_\epsilon^*$, $G_\epsilon^*$, $\mathcal{J}_\epsilon^*$, etc. In the limit $\epsilon \to 0$, we have the following result:

Lemma 15. (Coifman and Lafon, 2006; Proposition 3) For $f \in C^3(\mathcal{X})$ and $x \in \mathcal{X} \setminus \partial \mathcal{X}$,

$$
-\lim_{\epsilon \to 0} G_\epsilon^* = \Delta.
$$

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

Lemma 16. For functions $f \in C^3(\mathcal{X})$ whose gradients vanish at the boundary,

$$
\lim_{\epsilon \to 0} \mathcal{J}_\epsilon^*(f) = \int_{\mathcal{X}} \|\nabla f(x)\|^2 dS(x).
$$
Proof. By Green’s first identity

\[
\int_X f \nabla^2 f dS(x) + \int_X \nabla f \cdot \nabla f dS(x) = \int_{\partial X} f(\nabla f \cdot n) dS(x) = 0,
\]

where \(n\) is the normal direction to the boundary \(\partial X\), and the last surface integral vanishes due to the Neumann boundary condition. It follows from Lemma 15 that

\[
\lim_{\epsilon \to 0} J_\epsilon^*(f) = -\lim_{\epsilon \to 0} \int_X f(x) G_\epsilon^* f(x) dS_\epsilon(x) = \int_X f(x) \Delta f(x) dS(x) = \int_X \|\nabla f(x)\|^2 dS(x).
\]

Let \(A_\epsilon\) denote the sample version of the integral operator \(A_\epsilon\). To bound the difference \(\psi_{\epsilon,j} - \hat{\psi}_{\epsilon,j}\) we follow the strategy from Rosasco et al. (2010) and introduce two new integral operators that are related to \(A_\epsilon\) and \(A_\epsilon\), but that both act on an auxiliary \(^5\) RKHS \(\mathcal{H}\) of smooth functions. Define \(A_\mathcal{H}, \hat{A}_\mathcal{H} : \mathcal{H} \to \mathcal{H}\) where

\[
A_\mathcal{H} f(x) = \int k_\epsilon(x,y) \langle f, K(\cdot, y) \rangle_{\mathcal{H}} dP(y) = \int a_\epsilon(x,y) \langle f, K(\cdot, y) \rangle_{\mathcal{H}} dP(y)
\]

\[
\hat{A}_\mathcal{H} f(x) = \frac{\sum_{i=1}^n k_\epsilon(x,X_i) \langle f, K(\cdot, X_i) \rangle_{\mathcal{H}}}{\sum_{i=1}^n k_\epsilon(x,X_i)} = \int \hat{a}_\epsilon(x,y) \langle f, K(\cdot, y) \rangle_{\mathcal{H}} d\hat{P}(y),
\]

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 \(\|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}\). Now suppose the weight function \(k_\epsilon\) is sufficiently smooth with respect to \(\mathcal{H}\) (Assumption 1 in Rosasco et al. 2010); this condition is for example satisfied by a Gaussian kernel on a compact support \(\mathcal{X}\). By Propositions 13.3 and 14.3 in Rosasco et al. (2010), we can then relate the functions \(\psi_{\epsilon,j}\) and \(\hat{\psi}_{\epsilon,j}\), respectively, to the eigenfunctions \(u_{\epsilon,j}\) and \(\hat{u}_{\epsilon,j}\) of \(A_\mathcal{H}\) and \(\hat{A}_\mathcal{H}\). We have that

\[
\|\psi_{\epsilon,j} - \hat{\psi}_{\epsilon,j}\|_{L^2(\mathcal{X},P)} = C_1 \|u_{\epsilon,j} - \hat{u}_{\epsilon,j}\|_{L^2(\mathcal{X},P)} \leq C_2 \|u_{\epsilon,j} - \hat{u}_{\epsilon,j}\|_{\mathcal{H}}
\]

\(^5\)This auxiliary space only enters the intermediate derivations and plays no role in the error analysis of the algorithm itself.
for some constants $C_1$ and $C_2$. According to Theorem 6 in Rosasco et al. (2008) for eigenprojections of positive compact operators, it holds that

$$\|u_{\epsilon,j} - \hat{u}_{\epsilon,j}\|_{\mathcal{H}} \leq \frac{\|A_{\mathcal{H}} - \hat{A}_{\mathcal{H}}\|_{\mathcal{H}}}{\delta_{\epsilon,j}},$$

(8)

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

We choose the auxiliary RKHS $\mathcal{H}$ to be a Sobolev space with a sufficiently high degree of smoothness so that certain assumptions ((RC4)-(RC5) below) are fulfilled. 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 $\alpha$, and $L^2(\mathcal{X})$ is the space of square integrable functions with respect to the Lebesgue measure. Let $C^3_b(\mathcal{X})$ be the set of uniformly bounded, three times differentiable functions with uniformly bounded derivatives whose gradients vanish at the boundary. Now suppose that $\mathcal{H} \subset C^3_b(\mathcal{X})$ and that

(RC4) $\forall f \in \mathcal{H}, |\alpha| = s, \ D^\alpha (\hat{A}_{\mathcal{H}} f - A_{\mathcal{H}} f) = \hat{A}_{\mathcal{H}} D^\alpha f - A_{\mathcal{H}} D^\alpha f$,

(RC5) $\forall f \in \mathcal{H}, |\alpha| = s, \ D^\alpha f \in C^3_b(\mathcal{X})$.

Lemma 17. Let $\epsilon_n \to 0$ and $n\epsilon_n^{d/2}/\log(1/\epsilon_n) \to \infty$. Then, under the stated regularity conditions, $\|A_{\mathcal{H}} - \hat{A}_{\mathcal{H}}\|_{\mathcal{H}} = O_P(\gamma_n)$, where $\gamma_n = \sqrt{\log(1/\epsilon_n)/n\epsilon_n^{d/2}}$.

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

$$|A_{\epsilon} f(x) - \hat{A}_{\epsilon} f(x)| \leq |A_{\epsilon} f(x) - \hat{A}_{\epsilon} f(x)| + |\hat{A}_{\epsilon} f(x) - \hat{A}_{\epsilon} f(x)|$$

where $\hat{A}_{\epsilon} f(x) = \int \hat{a}_{\epsilon}(x,y) f(y) dP(y)$. From Giné and Guillou (2002),

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

18
Hence,

\[
|A_\epsilon f(x) - \hat{A}_\epsilon f(x)| \leq \left| \frac{\hat{p}_\epsilon(x) - p_\epsilon(x)}{\hat{p}_\epsilon(x) p_\epsilon(x)} \right| \int |f(y)| k_\epsilon(x, y) dP(y)
\]

\[
= \text{O}_P(\gamma_n) \int |f(y)| k_\epsilon(x, y) dP(y)
\]

\[
= \text{O}_P(\gamma_n).
\]

Next, we bound \(\hat{A}_\epsilon f(x) - A_\epsilon f(x)\). We have

\[
\hat{A}_\epsilon f(x) - A_\epsilon f(x) = \int f(y) \hat{a}_\epsilon(x, y)(d\hat{P}_n(y) - dP(y))
\]

\[
= \frac{1}{p(x) + o_P(1)} \int f(y) k_\epsilon(x, y)(d\hat{P}_n(y) - dP(y)).
\]

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,

\[
\int f(y) k_\epsilon(x, y)(d\hat{P}_n(y) - dP(y)) = f(x) \int k_\epsilon(x, y)(d\hat{P}_n(y) - dP(y)) + \int r_n(y) k_\epsilon(x, y)(d\hat{P}_n(y) - dP(y)).
\]

By an application of Talagrand’s inequality to each term, as in Theorem 5.1 of Giné and Koltchinskii (2006), we have

\[
\int f(y) k_\epsilon(x, y)(d\hat{P}_n(y) - dP(y)) = O_P(\gamma_n).
\]

Thus, \(\sup_{f \in C^3_b(\mathcal{X})} \|\hat{A}_\epsilon f - A_\epsilon f\|_\infty = O_P(\gamma_n)\).

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

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

Under regularity conditions (A4)-(A5),

\[
\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} = 1} \|\hat{A}_\epsilon f - A_\epsilon f\|_{\mathcal{H}}^2 
\leq \sup_{f \in \mathcal{H}, |\alpha| \leq s} \|D^\alpha (\hat{A}_\epsilon f - A_\epsilon f)\|_{L^2(\mathcal{X})}^2 
= \sum_{|\alpha| \leq s} \sup_{f \in \mathcal{H}} \|\hat{A}_\epsilon D^\alpha f - A_\epsilon D^\alpha f\|_{L^2(\mathcal{X})}^2 
\leq \sum_{|\alpha| \leq s} \sup_{f \in C^3_b(\mathcal{X})} \|\hat{A}_\epsilon f - A_\epsilon f\|_{L^2(\mathcal{X})}^2 
\leq C \sup_{f \in C^3_b(\mathcal{X})} \|\hat{A}_\epsilon f - A_\epsilon f\|_{L^2(\mathcal{X})}^2.
\]
for some constant $C$. Hence,

$$
\sup_{f \in \mathcal{H}} \frac{\|\hat{A}_f - A_f\|_{\mathcal{H}}}{\|f\|_{\mathcal{H}}} = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} = 1} \|\hat{A}_f - A_f\|_{\mathcal{H}} \leq C \sup_{f \in C^3_b(X)} \|\hat{A}_f - A_f\|_{\infty} = O_P(\gamma_n). \quad \square
$$

For $\epsilon_n \to 0$ and $n \epsilon_n^{d/2} / \log(1/\epsilon_n) \to \infty$, it then holds that:

**Proposition 2.** $\forall 0 \leq j \leq J$, 

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

where $\delta_{\epsilon,j} = \lambda_{\epsilon,j} - \lambda_{\epsilon,j+1}$.

**Proof.** From Lemma 14 and Equation 8, we have that

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

for some constant $C$ that does not depend on $n$. The result follows from Lemma 17. \quad \square

By putting together the bounds on bias and variance, we arrive at our main result.

**Theorem 4.** Under the conditions of Theorem 2 of the paper, if $\epsilon_n \to 0$ and $n \epsilon_n^{d/2} / \log(1/\epsilon_n) \to \infty$, a bound on the loss of the conditional density estimator with diffusion basis is given by

$$
L(f, \hat{f}) = O \left( \mathcal{J}(f) \frac{1}{\nu_{J+1}^2} \right) + O \left( \frac{1}{T^{2\beta}} \right) + IJO_P \left( \frac{1}{n} \right) + IJO_P \left( \frac{\gamma_n^2}{\epsilon_n \Delta_j^2} \right),
$$

where $\mathcal{J}(f) = \int_X \|\nabla f(x)\|^2 dS(x)$, $\Delta_j = \min_{0 \leq j \leq J} (\nu_j^2 - \nu_j^2)$, and $\nu_{J+1}^2$ is the $(J+1)^{th}$ eigenvalue of $\Delta$.

A Taylor expansion yields the following:

**Corollary 1.** Under the conditions of Theorem 2 of the paper, if $\epsilon_n \to 0$ and $n \epsilon_n^{d/2} / \log(1/\epsilon_n) \to \infty$, a bound on the loss of the conditional density estimator with diffusion basis is given by

$$
L(f, \hat{f}) = \frac{\mathcal{J}(f) O(1)}{\nu_{J+1}^2} + O \left( \frac{1}{T^{2\beta}} \right) + IJO_P \left( \frac{1}{n} \right) + IJO_P \left( \frac{\gamma_n^2}{\epsilon_n \Delta_j^2} \right),
$$
Corollary 2. By taking $\epsilon = n^{-\frac{1}{2(d+4)}}$, and ignoring lower order terms,

$$L(f, \hat{f}) = O\left(\frac{J(f)}{\nu_{j+1}^2}\right) + O\left(\frac{1}{I^2}\right) + O_P\left(\frac{\log n}{n}\right)^{\frac{2}{d+4}}. \quad (9)$$

If the support of the data is on a manifold with intrinsic dimension $p$, the eigenvalues of $\Delta$ are $\nu_j^2 \sim j^{2/p}$ (Zhou and Srebro, 2011). Theorem 1 in the main manuscript follows.

References


