An Empirical Feature-based Learning Algorithm Producing Sparse Approximations†

Xin Guo, Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong

83 Tat Chee Avenue, Kowloon, Hong Kong, P. R. China

Email: xinguo2@student.cityu.edu.hk, mazhou@cityu.edu.hk

Abstract

A learning algorithm for regression is studied. It is a modified kernel projection machine [2] in the form of a least square regularization scheme with ℓ^1-regularizer in a data dependent hypothesis space based on empirical features (constructed by a reproducing kernel and the learning data). The algorithm has three advantages. First, it does not involve any optimization process. Second, it produces sparse representations with respect to empirical features under a mild condition, without assuming sparsity in terms of any basis or system. Third, the output function converges to the regression function in the reproducing kernel Hilbert space at a satisfactory rate. Our error analysis does not require any sparsity assumption about the underlying regression function.

Keywords: learning theory, sparsity, reproducing kernel Hilbert space, ℓ^1-regularizer, empirical features

† The work described in this paper was partially supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 103508]. Corresponding author: Ding-Xuan Zhou.

1 Introduction

We propose a *learning algorithm* for regression. It is a modification of the kernel projection machine (KPM) introduced by Blanchard et. al. [2] and analyzed by Zwald [23]. The main advantage of this algorithm is its strong learning ability while producing *sparse approximations* in a very general setting in learning theory, without any hypothesis on sparse representations.

In the regression setting, an input space X is a compact metric space and the output space $Y = \mathbb{R}$. Let $Z = X \times Y$ and ρ be a Borel probability measure on Z with ρ_X the marginal measure on X, and $\rho(\cdot|x)$ the conditional measure at $x \in X$. The *regression function* f_ρ is defined as

$$f_\rho(x) = \int_Y y \,\mathrm{d}\rho(y|x), \qquad x \in X.$$

Our learning algorithm produces approximations of f_ρ in a *reproducing kernel Hilbert space* (RKHS). A symmetric continuous function $K : X \times X \to \mathbb{R}$ is called a Mercer kernel if for any finite subset $\{x_i\}_{i=1}^l$ of X, the $l \times l$ matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semi-definite. For $x \in X$, we denote $K_x = K(\cdot, x)$. The RKHS associated with the Mercer kernel K is a Hilbert space $\mathcal{H}_K$ completed by the span of $\{K_x : x \in X\}$ under the norm $\|\cdot\|_K$ induced by the inner product $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_u \rangle = K(x, u)$. We define an *integral operator* L_K on $\mathcal{H}_K$ by

$$L_K(f) = \int_X K_x f(x) \,\mathrm{d}\rho_X(x), \qquad f \in \mathcal{H}_K.$$

In this paper, we take a general setting in learning theory satisfying

$$f_\rho = L_K^r(g_\rho) \quad \text{for some } r > 0 \text{ and } g_\rho \in \mathcal{H}_K. \tag{1}$$

Since L_K is a compact, self-adjoint positive operator, we can arrange its eigenvalues $\{\lambda_i\}$ (with multiplicity) as a nonincreasing sequence tending to 0 and take an associated sequence of eigenfunctions $\{\phi_i\}$ to be an orthonormal basis of $\mathcal{H}_K$. Then the power L_K^r of L_K can be written by $L_K^r(\sum_i c_i \phi_i) = \sum_i c_i \lambda_i^r \phi_i$ and assumption (1) is equivalent to $f_\rho = \sum_i d_i \lambda_i^r \phi_i$ where $\{d_i\} \in \ell^2$ represents g_ρ as $g_\rho = \sum_i d_i \phi_i$. The exponent r in (1) measures the decay of the coefficients $\{d_i \lambda_i^r\}$ of f_ρ with respect to the orthonormal basis $\{\phi_i\}$ of $\mathcal{H}_K$. It can be regarded as a measurement for the regularity of the regression function f_ρ.

The eigenfunctions $\{\phi_i\}$ can be used to understand feature maps in learning theory. They can be approximated by *empirical features* $\{\phi_i^{\mathbf{x}}\}$ which are eigenfunctions of an *empirical operator* $L_K^{\mathbf{x}}$ associated with a sample $\mathbf{x} \in X^m$. Throughout this paper we assume that $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a sample drawn independently from ρ. We use $\mathbf{x}$ to denote the unlabeled part of the data $\mathbf{x} = \{x_1, \cdots, x_m\}$. The empirical operator $L_K^{\mathbf{x}}$ on $\mathcal{H}_K$ is defined by

$$L_K^{\mathbf{x}}(f) = \frac{1}{m}\sum_{i=1}^m f(x_i) K_{x_i} = \frac{1}{m}\sum_{i=1}^m \langle f, K_{x_i}\rangle K_{x_i}, \qquad f \in \mathcal{H}_K,$$

where we have used the reproducing property of the RKHS that asserts $\langle f, K_x\rangle = f(x)$ for any $f \in \mathcal{H}_K$ and $x \in X$. So $L_K^{\mathbf{x}}$ is a normalized sum of m rank-one operators and it is self-adjoint, positive with rank at most m. Therefore we can write the eigensystem of $L_K^{\mathbf{x}}$ as $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$, with eigenvalues $\lambda_i^{\mathbf{x}}$ arranged in nonincreasing order and $\lambda_i^{\mathbf{x}} = 0$ when $i > m$, and the corresponding eigenfunctions $\{\phi_i^{\mathbf{x}}\}_{i=1}^\infty$ to form an orthonormal basis of $\mathcal{H}_K$. The first m eigenfunctions $\{\phi_i^{\mathbf{x}}\}_{i=1}^m$ can be used as empirical features for learning by regularization schemes in a data dependent hypothesis space $\mathrm{span}\{\phi_i^{\mathbf{x}}\}_{i=1}^m$. The data dependence nature is reflected by the empirical features $\{\phi_i^{\mathbf{x}}\}_{i=1}^m$ obtained from the data $\mathbf{x}$. This idea was used in [2] to introduce the KPM outputting $\sum_{i=1}^m c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{z}}$ where the coefficient vector $c_\gamma^{\mathbf{z}} = (c_{\gamma,1}^{\mathbf{z}}, \cdots, c_{\gamma,m}^{\mathbf{z}})$ is given with a regularization parameter $\gamma > 0$ by

$$c_\gamma^{\mathbf{z}} = \arg\min_{c \in \mathbb{R}^m} \left\{ \frac{1}{m}\sum_{i=1}^m V\left(\sum_{j=1}^m c_j \phi_j^{\mathbf{x}}(x_i), y_i\right) + \gamma \|c\|_0 \right\}.$$

Here $V : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function and $\|c\|_0$ is the number of nonzero entries of the vector $c = (c_1, \cdots, c_m) \in \mathbb{R}^m$. The KPM was analyzed in [23] for classification with $V(f, y) = \max\{1 - yf, 0\}$ and for regression with $V(f, y) = (f - y)^2$ in a Gaussian white noise model.

In this paper we modify the KPM in the least square regression setting by using the ℓ^1-regularizer $\|c\|_1 = \sum_{i=1}^m |c_i|$ instead of the ℓ^0-penalty. Our learning algorithm now takes the form

$$c_\gamma^{\mathbf{z}} = \arg\min_{c \in \mathbb{R}^m} \left\{ \frac{1}{m}\sum_{i=1}^m \left(\left(\sum_{j=1}^m c_j \phi_j^{\mathbf{x}}\right)(x_i) - y_i\right)^2 + \gamma \|c\|_1 \right\}, \tag{2}$$

and the output function is

$$f_\gamma^{\mathbf{z}} = \sum_{i=1}^m c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}}. \tag{3}$$

We use $f_\gamma^{\mathbf{z}}$ to approximate the regression function f_ρ in $\mathcal{H}_K$.

The following Theorem 1, to be proved in Section 3, represents the solution to problem (2) explicitly, and thus shows computational efficiency of our algorithm.

Theorem 1. *For $i \in \mathbb{N}$, denote*

$$S_i^{\mathbf{z}} = \begin{cases} \frac{1}{m\lambda_i^{\mathbf{x}}} \sum_{j=1}^m y_j \phi_i^{\mathbf{x}}(x_j), & \text{if } \lambda_i^{\mathbf{x}} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then the solution to problem (2) is given with $i = 1, \ldots, m$ by

$$c_{\gamma,i}^{\mathbf{z}} = \begin{cases} 0, & \text{if } 2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| \le \gamma, \\ S_i^{\mathbf{z}} - \frac{\gamma}{2\lambda_i^{\mathbf{x}}}, & \text{if } 2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| > \gamma \text{ and } S_i^{\mathbf{z}} > \frac{\gamma}{2\lambda_i^{\mathbf{x}}}, \\ S_i^{\mathbf{z}} + \frac{\gamma}{2\lambda_i^{\mathbf{x}}}, & \text{if } 2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| > \gamma \text{ and } S_i^{\mathbf{z}} < -\frac{\gamma}{2\lambda_i^{\mathbf{x}}}. \end{cases} \tag{4}$$

In particular, $c_{\gamma,i}^{\mathbf{z}} = 0$ if $\lambda_i^{\mathbf{x}} = 0$.

Remark 1. *Let us show how the eigenpairs $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$ can be found explicitly. Let $d^{\mathbf{x}} \le m$ be the rank of the Gramian matrix $\mathbb{K} := (K(x_i, x_j))_{i,j=1}^m$. Denote its eigenvalues as $\widehat{\lambda}_1^{\mathbf{x}} \ge \cdots \ge \widehat{\lambda}_{d^{\mathbf{x}}}^{\mathbf{x}} > \widehat{\lambda}_{d^{\mathbf{x}}+1}^{\mathbf{x}} = \cdots \widehat{\lambda}_m^{\mathbf{x}} = 0$, and associated eigenvectors $\{\widehat{\mu}_i\}_{i=1}^m$ to form an orthonormal basis of $\mathbb{R}^m$. We have*

$$\begin{aligned} \lambda_i^{\mathbf{x}} &= \frac{\widehat{\lambda}_i^{\mathbf{x}}}{m} \quad \text{and } \phi_i^{\mathbf{x}} = \frac{1}{\sqrt{\widehat{\lambda}_i^{\mathbf{x}}}} \sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j}, \qquad \text{for } i = 1, \cdots, d^{\mathbf{x}}, \\ \lambda_i^{\mathbf{x}} &= 0, \quad \text{and } \phi_i^{\mathbf{x}}|_{\mathbf{x}} = 0, \qquad \text{for } i = d^{\mathbf{x}} + 1, \ldots, m. \end{aligned} \tag{5}$$

In fact, for $i = 1, \cdots, d^{\mathbf{x}}$, we see that

$$L_K^{\mathbf{x}} \left(\sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j} \right) = \frac{1}{m} \sum_{l=1}^m \sum_{j=1}^m (\widehat{\mu}_i)_j K(x_l, x_j) K_{x_l} = \frac{\widehat{\lambda}_i^{\mathbf{x}}}{m} \sum_{l=1}^m (\widehat{\mu}_i)_l K_{x_l}$$

and $\left\| \sum_{j=1}^m (\widehat{\mu}_i)_j K_{x_j} \right\|_K^2 = \widehat{\mu}_i^T \mathbb{K} \widehat{\mu}_i = \widehat{\lambda}_i^{\mathbf{x}} > 0$.

For $i = d^{\mathbf{x}} + 1, \cdots, m$, $\lambda_i^{\mathbf{x}} > 0$ would imply $\phi_i^{\mathbf{x}} = \frac{1}{\lambda_i^{\mathbf{x}}} L_K^{\mathbf{x}}(\phi_i^{\mathbf{x}}) = \frac{1}{m\lambda_i^{\mathbf{x}}} \sum_{j=1}^m \phi_i^{\mathbf{x}}(x_j) K_{x_j}$ and $\mathbb{K}(\phi_i^{\mathbf{x}}|_{\mathbf{x}}) = m\lambda_i^{\mathbf{x}} \phi_i^{\mathbf{x}}|_{\mathbf{x}}$ where $\phi_i^{\mathbf{x}}|_{\mathbf{x}} = (\phi_i^{\mathbf{x}}(x_j))_{j=1}^m$ is the vector obtained by restricting the function $\phi_i^{\mathbf{x}}$ onto the sampling points. It would then yield $\phi_i^{\mathbf{x}}|_{\mathbf{x}} = 0$ and $\phi_i^{\mathbf{x}} = 0$, a contradiction. So we must have $\lambda_i^{\mathbf{x}} = 0$. It follows that $\langle L_K^{\mathbf{x}}(\phi_i^{\mathbf{x}}), \phi_i^{\mathbf{x}} \rangle = 0$, which means $\frac{1}{m} \sum_{j=1}^m \phi_i^{\mathbf{x}}(x_j) \phi_i^{\mathbf{x}}(x_j) = 0$ and $\phi_i^{\mathbf{x}}|_{\mathbf{x}} = 0$. In this case, $\phi_i^{\mathbf{x}}$ is perpendicular to $\mathrm{span}\{K_{x_i}\}_{i=1}^m$

Note that for $i = d^{\mathbf{x}} + 1, \cdots, m$, $\lambda_i^{\mathbf{x}} = 0$ implies $c_{\gamma,i}^{\mathbf{z}} = 0$. So $\left(\sum_{j=1}^m c_j \phi_j^{\mathbf{x}} \right)(x_i) = \left(\sum_{j=1}^{d^{\mathbf{x}}} c_j \phi_j^{\mathbf{x}} \right)(x_i)$ and optimization problem (2) is the same as $c_{\gamma,i}^{\mathbf{x}} = 0$ for $i = d^{\mathbf{x}} + 1, \cdots, m$, and

$$(c_{\gamma,i}^{\mathbf{x}})_{i=1}^{d^{\mathbf{x}}} = \arg \min_{c \in \mathbb{R}^{d^{\mathbf{x}}}} \left\{ \frac{1}{m} \sum_{i=1}^m \left(\left(\sum_{j=1}^{d^{\mathbf{x}}} c_j \phi_j^{\mathbf{x}} \right)(x_i) - y_i \right)^2 + \gamma \|c\|_1 \right\}.$$

We shall conduct analysis for the error $f_\gamma^{\mathbf{z}} - f_\rho$ in the $\mathcal{H}_K$-metric (stronger than the $L^2_{\rho_X}$-metric, as shown in [14]) and derive learning rate for algorithm (2). Note that learning rates with the metric in $\mathcal{H}_K$ yield those with the metric in $C^s(X)$ when K is C^{2s} with $X \subset \mathbb{R}^n$. See [21].

Let us illustrate our analysis by the following examples when the eigenvalues $\{\lambda_i\}$ have some special asymptotic behaviors. Throughout the paper we assume that $|y| \leq M$ almost surely for some constant $M > 0$. Denote $\kappa = \sup_{x \in X} \sqrt{K(x,x)}$.

Theorem 2. *Assume (1) and for some* $\frac{1}{2r} < \alpha_2 \leq \alpha_1 < (1+r)\alpha_2 - \frac{1}{2}$ *and* $0 < D_1, D_2$, *the eigenvalues* $\{\lambda_i\}$ *decay polynomially as*

$$D_1 i^{-\alpha_1} \leq \lambda_i \leq D_2 i^{-\alpha_2}, \qquad \forall\ i. \tag{6}$$

Let $0 < \delta < 1$. *If we choose*

$$\gamma = \left(2^{1+2r} D_2^{1+r} \|g_\rho\|_K + C_{K,\rho} \left(\log \frac{4}{\delta} \right)^{1+r} \right) / \sqrt{m}, \tag{7}$$

then we have with confidence $1 - \delta$ *that*

$$c_{\gamma,i}^{\mathbf{z}} = 0, \qquad \forall\ m^{\frac{1}{2\alpha_2(1+r)}} + 1 \leq i \leq m, \tag{8}$$

and

$$\|f_\gamma^{\mathbf{z}} - f_\rho\|_K \leq C_1 \left(\log \frac{4}{\delta} \right)^{1+r} m^{-\frac{2\alpha_2 r - 1 - 2(\alpha_1 - \alpha_2)}{4\alpha_2(1+r)}}, \tag{9}$$

where $C_{K,\rho} = 8\kappa^2 \|g_\rho\|_K (\lambda_1^r + 2^{4r}\kappa^{2r}) + 16M\kappa$ *and* C_1 *is a constant independent of* δ *or* m *(which will be specified in the proof).*

Remark 2. *Asymptotic behavior (6) for the eigenvalues* $\{\lambda_i\}$ *of the integral operator is typical for Sobolev smooth kernels on domains in Euclidean spaces, and the power indices* α_1 *and* α_2 *depend on the smoothness of the kernel [12]. When the kernel is smooth enough,* α_2 *can be arbitrarily large and learning rate (9) takes the form* $m^{\epsilon - \frac{r}{2(1+r)}}$ *with an arbitrarily small* $\epsilon > 0$. *When* r *is large enough, it behaves like* $m^{\epsilon - \frac{1}{2}}$ *with an arbitrarily small* $\epsilon > 0$.

Observe from (8) that the number of nonzero coefficients in the representation $f_\gamma^{\mathbf{z}} = \sum_{i=1}^m c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}}$ *is at most* $m^{\frac{1}{2\alpha_2(1+r)}}$ *which can be much smaller than the sample size* m *when* α_2 *and* r *are large.*

Theorem 3. *Assume (1) and for some $1 < \beta_2 \leq \beta_1 < \beta_2^{1+r}$ and $0 < D_1, D_2$, the eigenvalues $\{\lambda_i\}$ decay exponentially as*

$$D_1 \beta_1^{-i} \leq \lambda_i \leq D_2 \beta_2^{-i}, \qquad \forall\ i. \tag{10}$$

Let $0 < \delta < 1$ and choose

$$\gamma = \left(2^{1+2r} D_2^{1+r} \|g_\rho\|_K + C_{K,\rho} \left(\log \frac{4}{\delta} \right)^{1+r} \right) / \sqrt{m},$$

then we have with confidence $1 - \delta$ that

$$c_{\gamma,i}^{\mathbf{z}} = 0, \qquad \forall\ \frac{\log(m+1)}{2(1+r)\log \beta_2} + 1 \leq i \leq m, \tag{11}$$

and

$$\|f_\gamma^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log \frac{4}{\delta} \right)^{1+r} \sqrt{\log(m+1)}\ m^{-\frac{r - \left(\log \frac{\beta_1}{\beta_2} \big/ \log \beta_2 \right)}{2(1+r)}}, \tag{12}$$

where C_2 is a constant independent of δ or m (which will be specified in the proof).

Remark 3. *Asymptotic behavior (10) for the eigenvalues $\{\lambda_i\}$ of the integral operator is typical for analytic kernels on domains in Euclidean spaces [13]. When r is large enough (meaning that f_ρ has high regularity), learning rate (12) behaves like $m^{\epsilon - \frac{1}{2}}$ with an arbitrarily small $\epsilon > 0$.*

Again we observe from (11) that the number of nonzero coefficients in the representation $f_\gamma^{\mathbf{z}} = \sum_{i=1}^m c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}}$ is at most $\frac{\log(m+1)}{2(1+r)\log\beta_2}$ which is much smaller than the sample size m.

Theorems 2 and 3 will be proved in Section 6.

2 General Analysis

Our general analysis for algorithm (2) is the following theorem to be proved in Section 5.

Theorem 4. *Assume (1). Let $p \in \{1, \ldots, m\}$ and $0 < \delta < 1$. Choose γ to satisfy*

$$2^{1+2r} \|g_\rho\|_K \lambda_p^{1+r} + C_{K,\rho} \frac{(\log \frac{4}{\delta})^{1+r}}{\sqrt{m}} \leq \gamma, \tag{13}$$

then with confidence $1-\delta$ *we have*

$$\|f_\gamma^{\mathbf{z}} - f_\rho\|_K \le \|g_\rho\|_K \lambda_p^r + \frac{\sqrt{2p}\gamma}{\lambda_p} + \frac{C_3 \log\frac{4}{\delta}}{\lambda_p\sqrt{m}} + C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2r,2\}}\right)^{1/2}, \quad (14)$$

where $C_3 = 16\sqrt{2}M\kappa + 2^{3+\max\{2r,1\}}\|g_\rho\|_K \lambda_1^r \kappa^2$ *and* $C_4 = 2^{\max\{r,1\}}\|g_\rho\|_K$.

Let us give a concrete example with $\mathcal{H}_K$ being the Sobolev space $H^s(X)$ of integer index $s > \frac{n}{2}$ and X being the unit ball $X = \{x \in \mathbb{R}^n : |x| \le 1\}$ of $\mathbb{R}^n$. When ρ_X is the normalized Lebesgue measure on X, a classical result in the theory of function spaces (see e.g. [17]) asserts that condition (6) for the eigenvalues $\{\lambda_i\}$ holds with $\alpha_1 = \alpha_2 = \frac{2s}{n}$. Also, if $f_\rho \in H^{(2r+1)s}(X)$ for some $r > \frac{n}{4s}$, we know that condition (1) holds true. Then the following learning rate can be derived from Theorem 4, as in the proof of Theorem 2.

Example 1. *Let* $X = \{x \in \mathbb{R}^n : |x| \le 1\}$ *and* ρ_X *be the normalized Lebesgue measure on* X*. If* K *is the reproducing kernel of the Sobolev space* $H^s(X)$ *of integer index* $s > \frac{n}{2}$ *and* $f_\rho \in H^{(2r+1)s}(X)$ *for some* $r > \frac{n}{4s}$*, then by taking* $\gamma = C_{s,f_\rho}(\log\frac{4}{\delta})^{1+r}/\sqrt{m}$*, we have with confidence* $1-\delta$*,*

$$\|f_\gamma^{\mathbf{z}} - f_\rho\|_K \le C_1' \left(\log\frac{4}{\delta}\right)^{1+r} m^{-\frac{4sr-n}{8s(1+r)}},$$

where C_{s,f_ρ} *and* C_1' *are constants independent of* δ *or* m*.*

3 Explicit Formula for the Coefficients

In this section we prove the representer theorem for algorithm (2). The ℓ^1-regularizer is important in the process. The proof is an immediate consequence of the classical result on soft-thresholding in the context of orthogonal regressors [19], once the orthogonality of $\{\phi_i^{\mathbf{x}}\}$ on the data is derived (see (15) below). We give the proof here for completeness.

Proof of Theorem 1. Let $i \in \mathbb{N}$. Since $(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})$ is an eigenpair of $L_K^{\mathbf{x}}$, we have

$$\lambda_i^{\mathbf{x}}\phi_i^{\mathbf{x}} = L_K^{\mathbf{x}}\phi_i^{\mathbf{x}} = \frac{1}{m}\sum_{j=1}^{m} \phi_i^{\mathbf{x}}(x_j) K_{x_j}.$$

It follows from the reproducing property $\left\langle K_{x_j}, \phi_l^{\mathbf{x}}\right\rangle = \phi_l^{\mathbf{x}}(x_j)$ that

$$\delta_{i,l}\lambda_i^{\mathbf{x}} = \langle \lambda_i^{\mathbf{x}}\phi_i^{\mathbf{x}}, \phi_l^{\mathbf{x}}\rangle = \frac{1}{m}\sum_{j=1}^{m} \phi_i^{\mathbf{x}}(x_j)\phi_l^{\mathbf{x}}(x_j), \qquad i, l \in \mathbb{N}, \quad (15)$$

where $\delta_{i,l} = 1$ if $i = l$ and $\delta_{i,l} = 0$ otherwise. In particular, when $\lambda_i^{\mathbf{x}} = 0$ (which is the case when $i > m$), we have $\phi_i^{\mathbf{x}}(x_j) = 0$ for each $j \in \{1, \ldots, m\}$. Consider the minimization problem (2). Note from the definition of $S_i^{\mathbf{z}}$ that $\frac{1}{m}\sum_{j=1}^m y_j \phi_i^{\mathbf{x}}(x_j) = \lambda_i^{\mathbf{x}} S_i^{\mathbf{z}}$. Apply (15). The empirical error part takes the form

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \left(\left(\sum_{j=1}^m c_j \phi_j^{\mathbf{x}}\right)(x_i) - y_i\right)^2 \\
= &\sum_{p,q=1}^m c_p c_q \frac{1}{m}\sum_{i=1}^m \phi_p^{\mathbf{x}}(x_i)\phi_q^{\mathbf{x}}(x_i) - \frac{2}{m}\sum_{i,j=1}^m y_i c_j \phi_j^{\mathbf{x}}(x_i) + \frac{1}{m}\sum_{i=1}^m y_i^2 \\
= &\sum_{p,q=1}^m c_p c_q \delta_{p,q}\lambda_p^{\mathbf{x}} - 2\sum_{i=1}^m \lambda_i^{\mathbf{x}} S_i^{\mathbf{z}} c_i + \frac{1}{m}\sum_{i=1}^m y_i^2 = \sum_{i=1}^m \lambda_i^{\mathbf{x}} c_i^2 - 2\sum_{i=1}^m \lambda_i^{\mathbf{x}} S_i^{\mathbf{z}} c_i + \frac{1}{m}\sum_{i=1}^m y_i^2.
\end{aligned}$$

Hence we have an equivalent form of (2) as

$$c_\gamma^{\mathbf{z}} = \arg\min_{c \in \mathbb{R}^m} \sum_{i=1}^m \left\{\lambda_i^{\mathbf{x}} \left(c_i - S_i^{\mathbf{z}}\right)^2 + \gamma |c_i|\right\}.$$

Thus for $i \in \{1, \ldots, m\}$, when $\lambda_i^{\mathbf{x}} = 0$, we have $c_{\gamma,i}^{\mathbf{z}} = 0$. When $\lambda_i^{\mathbf{x}} > 0$, the component $c_{\gamma,i}^{\mathbf{z}}$ can be found by solving the following optimization problem

$$c_{\gamma,i}^{\mathbf{z}} = \arg\min_{c \in \mathbb{R}} \left\{(c - S_i^{\mathbf{z}})^2 + \frac{\gamma}{\lambda_i^{\mathbf{x}}}|c|\right\}$$

which has the solution given by (4). This proves the theorem. □

Remark 4. *The algorithm can be divided into two parts: computing eigenpairs $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$ and solving the minimization problem (2). So the algorithm can be extended to a semi-supervised learning setting: if other than the labeled data $\{(x_i, y_i)\}_{i=1}^m$, we have some extra unlabeled data $\{x_i\}_{i=m+1}^{m'}$, then we can enhance the learning of the eigenfunctions in the first step by making full use of all the data $\{x_i\}_{i=1}^{m'}$.*

4 Preliminary Analysis for Sparsity

Theorem 1 tells us that $c_{\gamma,i}^{\mathbf{z}} = 0$ whenever $2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| \leq \gamma$. We shall choose suitable $p = p(m)$ with $\frac{p(m)}{m} \to 0$ and γ depending on δ such that with confidence $1 - \delta$,

$$2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| \leq \gamma, \qquad i = p+1, \ldots, m, \tag{16}$$

which would yield the desired sparsity: $c^{\mathbf{z}}_{\gamma,i} = 0$ for $i = p+1, \ldots, m$. The preliminary analysis for sparsity is an important tool for our error analysis.

To achieve the required condition (16), we need to estimate $\lambda_i^{\mathbf{x}}$ and $S_i^{\mathbf{z}}$. The eigenvalue $\lambda_i^{\mathbf{x}}$ is easier to deal with, by the following Hoffman-Wielandt inequality (see [7] for the original inequality for matrices, [8] for the generalization to self-adjoint operators on Hilbert spaces, [9] for an application to approximation of integral operators, and [1] for more general discussion).

Lemma 1. *We have*

$$\sum_{i=1}^{\infty} (\lambda_i - \lambda_i^{\mathbf{x}})^2 \le \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2,$$

where $\|\cdot\|_{\mathrm{HS}}$ *is the Hilbert-Schmidt norm of* $\mathrm{HS}(\mathcal{H}_K)$*, the Hilbert space of all Hilbert-Schmidt operators on* $\mathcal{H}_K$*.*

Recall that $\langle A_1, A_2 \rangle_{\mathrm{HS}} = \sum_j \langle A_1 e_j, A_2 e_j \rangle_K$ for $A_1, A_2 \in \mathrm{HS}(\mathcal{H}_K)$, where $\{e_j\}$ is an orthonormal basis of $\mathcal{H}_K$. The space $\mathrm{HS}(\mathcal{H}_K)$ is a subspace of the space of bounded linear operators on $\mathcal{H}_K$ with norms satisfying $\|A\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \|A\|_{\mathrm{HS}}$.

The quantity $\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}$ has been bounded in the literature [4, 9, 20, 14, 22].

Lemma 2. *For* $0 < \delta < 1$*, we have with confidence* $1 - \delta$*,*

$$\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}} \le \frac{4\kappa^2 \log \frac{2}{\delta}}{\sqrt{m}}. \tag{17}$$

Bounding the coefficients $\{S_i^{\mathbf{z}}\}$ towards (16) is more involved. We first show that $\lambda_i^{\mathbf{x}} S_i^{\mathbf{z}}$ is close to $\lambda_i^{\mathbf{x}} \langle f_\rho, \phi_i^{\mathbf{x}} \rangle$, by means of the following probability inequality in [15] derived from [11, 14].

Lemma 3. *Let* $\{\xi_i\}_{i=1}^m$ *be a set of independent random variables with values in a Hilbert space. If* $\|\xi_i\| \le \widetilde{M} < \infty$ *almost surely for each* $i = 1, \cdots, m$*, then for* $0 < \delta < 1$*, with confidence* $1 - \delta$ *we have*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (\xi_i - \mathbb{E}\xi_i) \right\| \le \frac{4\widetilde{M} \log \frac{2}{\delta}}{\sqrt{m}}.$$

Lemma 4. *For* $0 < \delta < 1$*, with confidence* $1 - \delta$ *we have*

$$\left(\sum_{j \in \mathbb{N}} \left(\lambda_j^{\mathbf{x}} \left(S_j^{\mathbf{z}} - \langle f_\rho, \phi_j^{\mathbf{x}} \rangle \right) \right)^2 \right)^{1/2} \le \frac{8M\kappa \log \frac{2}{\delta}}{\sqrt{m}}. \tag{18}$$

Proof. Consider the set of independent random variables $\{\xi_i = (y_i - f_\rho(x_i))K_{x_i}\}_{i=1}^m$ with values in the Hilbert space $\mathcal{H}_K$. They satisfy $\|\xi_i\| = |y_i - f_\rho(x_i)|\sqrt{K(x_i, x_i)} \le 2M\kappa$ and $\mathbb{E}\xi_i = 0$. So by Lemma 3, we know that for any $0 < \delta < 1$, with confidence $1 - \delta$ we have $\|\frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))K_{x_i}\|_K \le \frac{8M\kappa \log\frac{2}{\delta}}{\sqrt{m}}$.

By the definition of $S_j^{\mathbf{z}}$ and the relation $\lambda_j^{\mathbf{x}}\phi_j^{\mathbf{x}} = L_K^{\mathbf{x}}(\phi_j^{\mathbf{x}}) = \frac{1}{m}\sum_{i=1}^m \phi_j^{\mathbf{x}}(x_i)K_{x_i}$, for each $j \in \mathbb{N}$ we have

$$\lambda_j^{\mathbf{x}}\left(S_j^{\mathbf{z}} - \langle f_\rho, \phi_j^{\mathbf{x}}\rangle\right) = \frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))\,\phi_j^{\mathbf{x}}(x_i) = \left\langle \frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))K_{x_i}, \phi_j^{\mathbf{x}}\right\rangle.$$

But $\{\phi_j^{\mathbf{x}}\}_{j\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_K$, so we have

$$\sum_{j\in\mathbb{N}}\left(\lambda_j^{\mathbf{x}}\left(S_j^{\mathbf{z}} - \langle f_\rho, \phi_j^{\mathbf{x}}\rangle\right)\right)^2 = \left\|\frac{1}{m}\sum_{i=1}^m (y_i - f_\rho(x_i))K_{x_i}\right\|^2,$$

and our conclusion follows. $\square$

Next we need to estimate $\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle$. Since $\{\phi_j\}$ and $\{\phi_i^{\mathbf{x}}\}$ are orthonormal bases of $\mathcal{H}_K$, we observe that

$$(L_K - L_K^{\mathbf{x}})\,\phi_i^{\mathbf{x}} = \sum_{j=1}^\infty \langle \phi_i^{\mathbf{x}}, \phi_j\rangle L_K\phi_j - \lambda_i^{\mathbf{x}}\sum_{j=1}^\infty \langle \phi_i^{\mathbf{x}}, \phi_j\rangle\,\phi_j = \sum_{j=1}^\infty \langle \phi_i^{\mathbf{x}}, \phi_j\rangle\,(\lambda_j - \lambda_i^{\mathbf{x}})\,\phi_j.$$

Then the definition of the Hilbert-Schmidt norm tells us that

$$\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2 = \sum_{i=1}^\infty \|(L_K - L_K^{\mathbf{x}})\,\phi_i^{\mathbf{x}}\|_K^2 = \sum_{i,j=1}^\infty (\lambda_j - \lambda_i^{\mathbf{x}})^2\,(\langle \phi_i^{\mathbf{x}}, \phi_j\rangle)^2. \tag{19}$$

We shall use expression (19) a few times in our analysis for both sparsity and error bounds.

Lemma 5. *Let $I \subseteq \mathbb{N}$. If $f_\rho = L_K^r(g_\rho)$ for some $r > 0$ and $g_\rho \in \mathcal{H}_K$, then*

$$\left(\sum_{i\in I}|\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle|^2\right)^{1/2} \le \lambda_1^r\|g_\rho\|_K\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}} + 2^r\|g_\rho\|_K\left(\sum_{i\in I}(\lambda_i^{\mathbf{x}})^{2(1+r)}\right)^{1/2}.$$

Proof. Write $g_\rho = \sum_{j=1}^\infty d_j\phi_j$ with $\{d_j\} \in \ell^2$ and $\|\{d_j\}\|_{\ell^2} = \|g_\rho\|_K$. Then $f_\rho = \sum_{j=1}^\infty \lambda_j^r d_j\phi_j$, and for $i \in I$,

$$\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle = \lambda_i^{\mathbf{x}}\sum_{j=1}^\infty \lambda_j^r d_j\langle \phi_j, \phi_i^{\mathbf{x}}\rangle = \lambda_i^{\mathbf{x}}\sum_{j:\lambda_j > 2\lambda_i^{\mathbf{x}}}\lambda_j^r d_j\langle \phi_j, \phi_i^{\mathbf{x}}\rangle + \lambda_i^{\mathbf{x}}\sum_{j:\lambda_j \le 2\lambda_i^{\mathbf{x}}}\lambda_j^r d_j\langle \phi_j, \phi_i^{\mathbf{x}}\rangle.$$

When $\lambda_j > 2\lambda_i^{\mathbf{x}}$, we have $\lambda_i^{\mathbf{x}} \leq \lambda_j - \lambda_i^{\mathbf{x}}$. Hence by the Schwarz inequality,

$$\left| \lambda_i^{\mathbf{x}} \sum_{j:\lambda_j > 2\lambda_i^{\mathbf{x}}} \lambda_j^r d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle \right| \leq \lambda_1^r \|\{d_l\}\|_{\ell^2} \left(\sum_{j:\lambda_j > 2\lambda_i^{\mathbf{x}}} (\lambda_j - \lambda_i^{\mathbf{x}})^2 \left(\langle \phi_j, \phi_i^{\mathbf{x}} \rangle \right)^2 \right)^{1/2}.$$

It follows from (19) that

$$\left(\sum_{i \in I} \left| \lambda_i^{\mathbf{x}} \left\langle f_\rho, \phi_i^{\mathbf{x}} \right\rangle \right|^2 \right)^{1/2} \leq \lambda_1^r \|\{d_j\}\|_{\ell^2} \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}} + 2^r \|\{d_j\}\|_{\ell^2} \left(\sum_{i \in I} (\lambda_i^{\mathbf{x}})^{2(1+r)} \right)^{1/2}.$$

The proof is completed. □

Now we can present our preliminary analysis for sparsity of algorithm (2). The ℓ^1-*regularizer* plays a key role to produce sparse approximations. The phenomenon that the ℓ^1-regularizer can be used to reproduce sparsity has been observed in LASSO [19] and compressed sensing [3, 6], usually under the assumption that the approximated function has a sparse representation with respect to some basis or redundant system. Here we show that sparsity of $f_\gamma^{\mathbf{z}}$ in representation (3) can be *produced* under assumption (1) which does not impose any sparse representation and is a common mild condition in learning theory (e.g. [4, 14, 10]). The choice of the empirical features $\{\phi_i^{\mathbf{x}}\}_{i=1}^m$ is important to ensure the sparsity and convergence rates for the algorithm.

Theorem 5. *Under the same condition as in Theorem 4, with confidence* $1 - \delta$ *we have*

$$c_{\gamma,i}^{\mathbf{z}} = 0, \qquad \forall\ i = p+1, \ldots, m.$$

Proof. By Lemmas 2 and 4, we know that for any $0 < \delta < \frac{1}{2}$ there exists a subset Z_δ of Z^m of measure at least $1 - 2\delta$ such that both (17) and (18) hold for each $\mathbf{z} \in Z_\delta$.

Let $i \in \{1, \ldots, m\}$ and $\mathbf{z} \in Z_\delta$. Then from (18), we see that

$$2\lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}}| \leq 2\lambda_i^{\mathbf{x}} |\langle f_\rho, \phi_i^{\mathbf{x}} \rangle| + 2\lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}} - \langle f_\rho, \phi_i^{\mathbf{x}} \rangle| \leq 2\lambda_i^{\mathbf{x}} |\langle f_\rho, \phi_i^{\mathbf{x}} \rangle| + \frac{16 M \kappa \log \frac{2}{\delta}}{\sqrt{m}}.$$

Applying Lemma 5 to $I = \{i\}$, we have

$$2\lambda_i^{\mathbf{x}} |S_i^{\mathbf{z}}| \leq \lambda_1^r \|g_\rho\|_K \frac{8\kappa^2 \log \frac{2}{\delta}}{\sqrt{m}} + 2^{1+r} \|g_\rho\|_K (\lambda_i^{\mathbf{x}})^{1+r} + \frac{16 M \kappa \log \frac{2}{\delta}}{\sqrt{m}}. \tag{20}$$

By Lemma 1, $|\lambda_i^{\mathbf{x}} - \lambda_i| \leq \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}$, so $(\lambda_i^{\mathbf{x}})^{1+r} \leq (\lambda_i + \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}})^{1+r} \leq 2^r(\lambda_i^{1+r} + \|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}^{1+r})$. It follows that for $i = 1, \cdots, m$, the right-hand side of (20) has an upper bound

$$2^{1+2r} \|g_\rho\|_K \lambda_i^{1+r} + C_{K,\rho} \frac{\left(\log \frac{2}{\delta}\right)^{1+r}}{\sqrt{m}}.$$

Therefore, when

$$2^{1+2r}\|g_\rho\|_K \lambda_p^{1+r} + C_{K,\rho}\frac{\left(\log \frac{2}{\delta}\right)^{1+r}}{\sqrt{m}} \leq \gamma, \tag{21}$$

we know that

$$2\lambda_i^{\mathbf{x}}|S_i^{\mathbf{z}}| \leq \gamma, \qquad \forall\ i = p+1, \ldots, m,$$

which by Theorem 1 yields $c_{\gamma,i}^{\mathbf{z}} = 0$ for $i = p+1, \ldots, m$. Then the conclusion of Theorem 5 follows by scaling 2δ to δ, for which (21) corresponds to (13). □

From Theorem 5 we see immediately that when the eigenvalues $\{\lambda_i\}$ decay polynomially, the sparsity can be explicitly derived by taking p to be $\lceil m^{\frac{1}{2\alpha(1+r)}} \rceil$, the smallest integer greater than or equal to $m^{\frac{1}{2\alpha(1+r)}}$.

Corollary 1. *Assume (1). If for some $D_2 > 0$ and $\alpha > 0$, $\lambda_i \leq D_2 i^{-\alpha}$ holds for each i, then when $\gamma \geq (2^{1+2r}D_2^{1+r}\|g_\rho\|_K + C_{K,\rho}\left(\log \frac{4}{\delta}\right)^{1+r})/\sqrt{m}$, we have with confidence $1-\delta$,*

$$c_{\gamma,i}^{\mathbf{z}} = 0, \qquad \forall\ m^{\frac{1}{2\alpha(1+r)}} + 1 \leq i \leq m.$$

5 Error Analysis

In this section, we prove our error bounds stated in Theorem 4.

Proof of Theorem 4. We follow the proof of Theorem 5 and know that for any $0 < \delta < \frac{1}{2}$ there exists a subset Z_δ of Z^m of measure at least $1 - 2\delta$ such that both (17) and (18) hold for each $\mathbf{z} \in Z_\delta$. Moreover, when (21) is satisfied and $\mathbf{z} \in Z_\delta$, we have $c_{\gamma,i}^{\mathbf{z}} = 0$ for every $i \in \{p+1, \ldots, m\}$ and those $i \in \{1, \ldots, p\}$ with $\lambda_i^{\mathbf{x}} \leq \frac{\lambda_p}{2}$, which follows directly from (20). Hence

$$f_\gamma^{\mathbf{z}} = \sum_{i \in \mathcal{S}} c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}},$$

where $\mathcal{S}$ is defined by $\mathcal{S} = \{i \in \{1, \ldots, p\} : \lambda_i^{\mathbf{x}} > \frac{\lambda_p}{2}\}$. It follows from the orthogonal expansion in terms of the orthonormal basis $\{\phi_i^{\mathbf{x}}\}$ that

$$\|f_\gamma^{\mathbf{z}} - f_\rho\|_K^2 = \sum_{i \in \mathbb{N}\setminus\mathcal{S}} (\langle f_\rho, \phi_i^{\mathbf{x}}\rangle)^2 + \sum_{i \in \mathcal{S}} \left(\langle f_\rho, \phi_i^{\mathbf{x}}\rangle - c_{\gamma,i}^{\mathbf{z}}\right)^2 =: \Delta_1 + \Delta_2. \tag{22}$$

Let $\mathbf{z} \in Z_\delta$ in the following proof.

We bound the first term Δ_1 on the right-hand side of (22) by decomposing it further into two parts with $f_\rho = \sum_{j=1}^{\infty} \lambda_j^r d_j \phi_j = \sum_{j=p+1}^{\infty} \lambda_j^r d_j \phi_j + \sum_{j=1}^{p} \lambda_j^r d_j \phi_j$. Here we have written $g_\rho = \sum_{j=1}^{\infty} d_j \phi_j$ with $\{d_j\} \in \ell^2$ and $\|\{d_j\}\|_{\ell^2} = \|g_\rho\|_K$.

The part with $\sum_{j=p+1}^{\infty}$ is easy to deal with: since $\{\phi_i^{\mathbf{x}}\}$ is an orthonormal basis, we have

$$\left(\sum_{i=1}^{\infty} \left\langle \sum_{j=p+1}^{\infty} \lambda_j^r d_j \phi_j, \phi_i^{\mathbf{x}} \right\rangle^2\right)^{1/2} = \left\| \sum_{j=p+1}^{\infty} \lambda_j^r d_j \phi_j \right\|_K \leq \|g_\rho\|_K \lambda_{p+1}^r. \tag{23}$$

The part with $\sum_{j=1}^{p}$ can be estimated by the Schwarz inequality as

$$\left(\sum_{i \in \mathbb{N} \setminus \mathcal{S}} \left\langle \sum_{j=1}^{p} \lambda_j^r d_j \phi_j, \phi_i^{\mathbf{x}} \right\rangle^2\right)^{1/2} \leq \left(\sum_{i \in \mathbb{N} \setminus \mathcal{S}} \|\{d_l\}\|_{\ell^2}^2 \sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2\right)^{1/2}.$$

We continue to bound $\sum_{i \in \mathbb{N} \setminus \mathcal{S}} \sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2$ in two cases.

Case 1: $r \geq 1$. For $i \geq p+1$, we observe that $\lambda_j^{2r} \leq 2^{2r-1}(\lambda_i^{2r} + (\lambda_j - \lambda_i)^{2r})$ and

$$(\lambda_j - \lambda_i)^{2r} \leq \lambda_1^{2r-2}(\lambda_j - \lambda_i)^2 \leq 2\lambda_1^{2r-2}\left(|\lambda_j - \lambda_i^{\mathbf{x}}|^2 + |\lambda_i - \lambda_i^{\mathbf{x}}|^2\right).$$

It follows that

$$\sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2 \leq 2^{2r-1} \sum_{j=1}^{p} \left(\lambda_i^{2r} + 2\lambda_1^{2r-2}|\lambda_i - \lambda_i^{\mathbf{x}}|^2 + 2\lambda_1^{2r-2}|\lambda_j - \lambda_i^{\mathbf{x}}|^2\right) \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2,$$

which in connection with Lemma 1 and (19) yields

$$\begin{aligned}
&\sum_{i=p+1}^{\infty} \sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2 \\
\leq\ & 2^{2r-1} \sum_{i=p+1}^{\infty} \lambda_i^{2r} + 2^{2r}\lambda_1^{2r-2}\left(\sum_{i=1}^{\infty} |\lambda_i - \lambda_i^{\mathbf{x}}|^2 + \sum_{i,j=1}^{\infty} |\lambda_j - \lambda_i^{\mathbf{x}}|^2 \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2\right) \\
\leq\ & 2^{2r-1} \sum_{i=p+1}^{\infty} \lambda_i^{2r} + 2^{1+2r}\lambda_1^{2r-2}\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2.
\end{aligned}$$

For $i \in \{1, \ldots, p\} \setminus \mathcal{S}$ and $j \leq p$, we have $|\lambda_j - \lambda_i^{\mathbf{x}}| \geq \frac{\lambda_j}{2}$ and hence $\lambda_j^{2r} \leq 4\lambda_1^{2r-2}|\lambda_j - \lambda_i^{\mathbf{x}}|^2$. So by (19),

$$\sum_{i \in \{1,\ldots,p\} \setminus \mathcal{S}} \sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2 \leq 4\lambda_1^{2r-2} \sum_{i,j=1}^{\infty} |\lambda_j - \lambda_i^{\mathbf{x}}|^2 \langle \phi_j, \phi_i^{\mathbf{x}} \rangle^2 \leq 4\lambda_1^{2r-2}\|L_K - L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2.$$

Thus in the first case we have

$$\sum_{i\in\mathbb{N}\setminus\mathcal{S}}\sum_{j=1}^{p}\lambda_j^{2r}\left\langle\phi_j,\phi_i^{\mathbf{x}}\right\rangle^2\le 2^{2r-1}\sum_{i=p+1}^{\infty}\lambda_i^{2r}+4\lambda_1^{2r-2}(2^{2r-1}+1)\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2.$$

Case 2: $r<1$. we notice that $\lambda_j^{2r}\le\lambda_p^{2r-2}\lambda_j^2$ and obtain from the above estimate

$$\sum_{i\in\mathbb{N}\setminus\mathcal{S}}\sum_{j=1}^{p}\lambda_j^{2r}\left\langle\phi_j,\phi_i^{\mathbf{x}}\right\rangle^2\le 2\lambda_p^{2r-2}\sum_{i=p+1}^{\infty}\lambda_i^{2}+12\lambda_p^{2r-2}\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}^2.$$

The bounds for the two cases together with (23) give a bound for Δ_1 as

$$\sqrt{\Delta_1}\le\begin{cases}\|g_\rho\|_K\lambda_{p+1}^r+2^r\|\{d_j\}\|_{\ell^2}\left((\sum_{i=p+1}^{\infty}\lambda_i^{2r})^{1/2}+2^{1+r}\lambda_1^{r-1}\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}\right), & \text{if } r\ge1,\\ \|g_\rho\|_K\lambda_{p+1}^r+2\|\{d_j\}\|_{\ell^2}\lambda_p^{r-1}\left((\sum_{i=p+1}^{\infty}\lambda_i^{2})^{1/2}+2\|L_K-L_K^{\mathbf{x}}\|_{\mathrm{HS}}\right), & \text{if } r<1.\end{cases}$$

Now we turn to the second term Δ_2 on the right-hand side of (22). Observe that the case $c_{\gamma,i}^{\mathbf{z}}=0$ corresponds to $|S_i^{\mathbf{z}}|\le\frac{\gamma}{2\lambda_i^{\mathbf{x}}}$. So for either $c_{\gamma,i}^{\mathbf{z}}=0$ or $c_{\gamma,i}^{\mathbf{z}}=S_i^{\mathbf{z}}\pm\frac{\gamma}{2\lambda_i^{\mathbf{x}}}$, we always have

$$|\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle-c_{\gamma,i}^{\mathbf{z}}|\le\frac{\gamma}{2\lambda_i^{\mathbf{x}}}+|S_i^{\mathbf{z}}-\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle|\le\frac{1}{2\lambda_i^{\mathbf{x}}}\left(\gamma+2\lambda_i^{\mathbf{x}}|\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle-S_i^{\mathbf{z}}|\right).$$

But for each $i\in\mathcal{S}$, there holds $2\lambda_i^{\mathbf{x}}\ge\lambda_p$. Hence

$$\sqrt{\Delta_2}=\left(\sum_{i\in\mathcal{S}}\left(\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle-c_{\gamma,i}^{\mathbf{z}}\right)^2\right)^{1/2}\le\frac{\sqrt{2p}\gamma}{\lambda_p}+\frac{2\sqrt{2}}{\lambda_p}\left(\sum_{i\in\mathcal{S}}\left(\lambda_i^{\mathbf{x}}\left(S_i^{\mathbf{z}}-\left\langle f_\rho,\phi_i^{\mathbf{x}}\right\rangle\right)\right)^2\right)^{1/2}.$$

By Lemma 4, this implies

$$\sqrt{\Delta_2}\le\frac{\sqrt{2p}\gamma}{\lambda_p}+\frac{16\sqrt{2}M\kappa\log\frac{2}{\delta}}{\lambda_p\sqrt{m}}.$$

Putting the bounds for Δ_1 and Δ_2 into (22), we know that for $\mathbf{z}\in Z_\delta$, $\|f_\gamma^{\mathbf{z}}-f_\rho\|_K$ is bounded by

$$\|g_\rho\|_K\lambda_p^r+\frac{\sqrt{2p}\gamma}{\lambda_p}+\frac{16\sqrt{2}M\kappa\log\frac{2}{\delta}}{\lambda_p\sqrt{m}}+\begin{cases}2^r\|g_\rho\|_K\left((\sum_{i=p+1}^{\infty}\lambda_i^{2r})^{1/2}+\frac{2^{3+r}\lambda_1^{r-1}\kappa^2\log\frac{2}{\delta}}{\sqrt{m}}\right), & \text{if } r\ge1,\\ 2\|g_\rho\|_K\lambda_p^{r-1}\left((\sum_{i=p+1}^{\infty}\lambda_i^{2})^{1/2}+\frac{8\kappa^2\log\frac{2}{\delta}}{\sqrt{m}}\right), & \text{if } r<1.\end{cases}$$

Then the conclusion of Theorem 4 follows by scaling 2δ to δ. $\square$

6 Achieving Both Sparsity and Learning Rates

We are in a position to derive both sparsity and learning rates in two special situations, based on our general analysis.

Proof of Theorem 2. We take $p = \lceil m^{1/(2\alpha_2(1+r))} \rceil$ to give $m^{1/(2\alpha_2(1+r))} \le p \le 2m^{1/(2\alpha_2(1+r))}$, so $\lambda_p^{1+r} \le D_2^{1+r}/\sqrt{m}$. Thus the choice of γ in (7) implies condition (13) of Theorem 5. This verifies (8) as well as the condition of Theorem 4. We bound the first three terms of the right-hand side of (14) in Theorem 4 as follows. First,

$$\begin{aligned} & \|g_\rho\|_K \lambda_p^r + \frac{\sqrt{2p}\gamma}{\lambda_p} + \frac{C_3}{\lambda_p \sqrt{m}} \log \frac{4}{\delta} \\ \le\ & \|g_\rho\|_K D_2^r m^{-\frac{\alpha_2 r}{2\alpha_2(1+r)}} + 2D_1^{-1} 2^{\alpha_1} \gamma m^{\left(\frac{1}{2}+\alpha_1\right)\frac{1}{2\alpha_2(1+r)}} + C_3 D_1^{-1} 2^{\alpha_1} \left(\log \frac{4}{\delta}\right) m^{-\frac{1}{2}+\frac{\alpha_1}{2\alpha_2(1+r)}} \\ \le\ & \widetilde{C}_1 \left(\log \frac{4}{\delta}\right)^{1+r} m^{-\frac{2\alpha_2 r - 1 - (\alpha_1 - \alpha_2)}{4\alpha_2(1+r)}}, \end{aligned}$$

where $\widetilde{C}_1 = \|g_\rho\|_K D_2^r + 2^{1+\alpha_1} D_1^{-1}(2^{1+r}\|g_\rho\|_K D_2^{1+r} + C_{K,\rho}) + C_3 D_1^{-1} 2^{\alpha_1}$.

When $r \ge 1$, since $2r\alpha_2 > 1$,

$$\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2r,2\}} \le D_2^{2r} \int_p^\infty x^{-2r\alpha_2}\, \mathrm{d}x = \frac{D_2^{2r} p^{1-2r\alpha_2}}{2r\alpha_2 - 1}.$$

So the last term of the right-hand side of (14) can be bounded as

$$C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2r,2\}} \right)^{1/2} \le \frac{C_4 D_2^r}{\sqrt{2r\alpha_2 - 1}} m^{\frac{1-2r\alpha_2}{4\alpha_2(1+r)}}.$$

Similarly, when $0 < r < 1$, since $\alpha_2 > \frac{1}{2} + (1-r)\alpha_1$, we have

$$\begin{aligned} C_4 \lambda_p^{\min\{r-1,0\}} \left(\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2r,2\}} \right)^{1/2} & \le \frac{C_4 D_1^{r-1} p^{-\alpha_1(r-1)} D_2 p^{(1-2\alpha_2)/2}}{\sqrt{2\alpha_2 - 1}} \\ & \le \frac{C_4 D_1^{r-1} D_2}{\sqrt{2\alpha_2 - 1}} m^{\frac{1+2(1-r)\alpha_1 - 2\alpha_2}{4\alpha_2(1+r)}}. \end{aligned}$$

Now we use Theorem 4 to obtain

$$\|f_\rho - f_\gamma^{\mathbf{z}}\|_K \le C_1 \left(\log \frac{4}{\delta}\right)^{1+r} m^{-\frac{2\alpha_2 r - 1 - 2(\alpha_1 - \alpha_2)}{4\alpha_2(1+r)}}$$

with confidence $1-\delta$, where

$$C_1 = \widetilde{C}_1 + \begin{cases} \frac{C_4 D_2^r}{\sqrt{2r\alpha_2 - 1}}, & \text{when} \quad r \geq 1, \\ \frac{C_4 D_1^{r-1} D_2}{\sqrt{2\alpha_2 - 1}}, & \text{when} \quad 0 < r < 1. \end{cases}$$

The proof of Theorem 2 is complete. □

Proof of Theorem 3. Choosing $p = \lceil \frac{\log(m+1)}{2(1+r)\log\beta_2} \rceil$, we have

$$\frac{\log(m+1)}{2(1+r)\log\beta_2} \leq p \leq 1 + \frac{\log(m+1)}{2(1+r)\log\beta_2}.$$

It follows that

$$m^{\frac{1}{2(1+r)}} \leq \beta_2^p \leq \beta_1^p \leq \beta_1 (2m)^{\frac{\log\beta_1}{2(1+r)\log\beta_2}}.$$

The assumption $\lambda_p \leq D_2 \beta_2^{-p}$ in (10) tells us that

$$\lambda_p^{1+r} \leq \frac{D_2^{1+r}}{\sqrt{m}}.$$

Then

$$2^{1+2r}\|g_\rho\|_K \lambda_p^{1+r} + C_{K,\rho}\frac{\left(\log\frac{4}{\delta}\right)^{1+r}}{\sqrt{m}} \leq 2^{1+2r}\|g_\rho\|_K \frac{D_2^{1+r}}{\sqrt{m}} + C_{K,\rho}\frac{\left(\log\frac{4}{\delta}\right)^{1+r}}{\sqrt{m}} = \gamma.$$

So condition (13) in Theorem 5 holds, and thus we know that with confidence $1-\delta$, $c_{\gamma,i}^{\mathbf{z}} = 0$ for $p+1 \leq i \leq m$. This verifies the desired conclusion (11) for sparsity.

Now we turn to the error analysis. By Theorem 4, bound (14) holds with confidence $1-\delta$. We estimate the first three terms of the right-hand side of (14) as

$$\begin{aligned} & \|g_\rho\|_K \lambda_p^r + \frac{\sqrt{2p}\gamma}{\lambda_p} + C_3 \frac{\log\frac{4}{\delta}}{\lambda_p\sqrt{m}} \\ \leq\ & \|g_\rho\|_K D_2^r m^{-\frac{r}{2(1+r)}} + \widetilde{C}_2 \left(\log\frac{4}{\delta}\right)^{1+r} \sqrt{\log(m+1)}\, m^{-\frac{r-\left(\log\frac{\beta_1}{\beta_2}\big/\log\beta_2\right)}{2(1+r)}} \\ +\ & C_3 D_1^{-1}\left(\log\frac{4}{\delta}\right)\beta_1 2^{\frac{\log\beta_1}{2(1+r)\log\beta_2}} m^{-\frac{r-\left(\log\frac{\beta_1}{\beta_2}\big/\log\beta_2\right)}{2(1+r)}}, \end{aligned} \tag{24}$$

where $\widetilde{C}_2 = \left(\frac{2}{\log 2} + \frac{1}{(1+r)\log\beta_2}\right)^{1/2} (2^{1+2r}\|g_\rho\|_K D_2^{1+r} + C_{K,\rho}) D_1^{-1}\beta_1 2^{\frac{\log\beta_1}{2(1+r)\log\beta_2}}$.

When $r \geq 1$, the last term in the right-hand side of (14) can be bounded as

$$C_4\left(\sum_{i=p+1}^{\infty}\lambda_i^{2r}\right)^{1/2} \leq C_4 D_2^r \left(\sum_{i=p+1}^{\infty}\beta_2^{-2ri}\right)^{1/2} = \frac{C_4 D_2^r \beta_2^{-pr}}{\sqrt{\beta_2^{2r}-1}} \leq \frac{C_4 D_2^r}{\sqrt{\beta_2^{2r}-1}} m^{-\frac{r}{2(1+r)}}. \tag{25}$$

Similarly, when $0 < r < 1$, we have

$$C_4 \lambda_p^{r-1} \left(\sum_{i=p+1}^{\infty} \lambda_i^2 \right)^{1/2} \leq C_4 D_1^{r-1} \beta_1^{1-r} (2m)^{\frac{(1-r)\log \beta_1}{2(1+r)\log \beta_2}} \frac{D_2 m^{-\frac{1}{2(1+r)}}}{\sqrt{\beta_2^2 - 1}}.$$

Putting this estimate in the case $0 < r < 1$ and (25) in the case $r > 1$ and (24) into bound (14) tells us that with confidence $1 - \delta$, the desired bound (12) for the error holds true with the constant C_2 given by

$$\begin{aligned} C_2 &= \frac{1}{\sqrt{\log 2}} \left(\|g_\rho\|_K D_2^r + C_3 D_1^{-1} \beta_1 2^{\frac{\log \beta_1}{2(1+r)\log \beta_2}} \right) + \widetilde{C}_2 \\ &+ \frac{1}{\sqrt{\log 2}} \begin{cases} \frac{C_4 D_2^r}{\sqrt{\beta_2^{2r} - 1}}, & \text{when } r \geq 1, \\ \frac{C_4 D_1^{r-1} \beta_1^{1-r} D_2}{\sqrt{\beta_2^2 - 1}} 2^{\frac{(1-r)\log \beta_1}{2(1+r)\log \beta_2}}, & \text{when } 0 < r < 1. \end{cases} \end{aligned}$$

The proof of Theorem 3 is complete. □

7 Further Remarks and Discussion

We have proposed a modified KPM (2) for regression with ℓ^1-regularizer. Analysis for the error in the $\mathcal{H}_K$-metric has been conducted by means of a priori condition (1) concerning the regularity of the regression with respect to the kernel K and the marginal distribution ρ_X. Our learning rates have been given in terms of special choices of the regularization parameter $\gamma > 0$ which depends on a priori condition (1). Condition (1) is a standard assumption for least square regularized regression with an infinitely dimensional $\mathcal{H}_K$ in the literature of learning theory [4, 14, 16, 18] and almost all theoretical error bounds are based on similar a priori conditions. To the best of our knowledge, the only theoretical error analysis for a learning algorithm with a regularization parameter determined directly by the data was given recently in [5], where a cross-validation approach was rigorously proved.

It is a common practice to choose the regularization parameter by a cross-validation method, which often leads to satisfactory simulation. Here we present an example to show how to choose the regularization parameter γ for algorithm (2). Rigorous theoretical analysis for such a process will be considered in our further study.

Example 2. *We generate the regression function f_ρ on $\mathbb{R}^{10}$ as*

$$f_\rho(x) = \sum_{i=1}^{3} A_i \exp\left(-\frac{|x - P_i|^2}{2v_i^2} \right), \tag{26}$$

where the parameters are prescribed in Table 1. The data set $\{(x_i, y_i)\}_i^m$ *is drawn indepen-*

i	coefficient A_i	variation v_i^2	center P_i
1	2.0	0.62^2	$(0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)$
2	-3.5	0.64^2	$(0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6)$
3	0.7	0.65^2	$\frac{1}{9}(0.9, 1.7, 2.5, 3.3, 4.1, 4.9, 5.7, 6.5, 7.3, 8.1)$

Table 1: Parameters

dently with x_i*'s uniformly distributed on* $[0,1]^{10}$, $y_i = f_\rho(x_i) + \epsilon_i$, *and* ϵ_i*'s being Gaussian noise with* $\mu = 0$, $\sigma^2 = 0.5^2$ *and truncated onto* $[-1.5, 1.5]$. *The Mercer kernel* K *is the Gaussian with variance* 0.60^2. *Table 2 shows the result of the simulation. For comparison, in the last three columns we list the error performance of the least squares regularized regression (LSR) algorithm*

$$f^{\mathbf{z}}_{\mathrm{LSR},\gamma_1} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma_1 \|f\|_K^2 \right\}.$$

The notations γ^* *and* γ_1^* *in the second and sixth columns denote the optimal* γ *and* γ_1

m	γ^*	$\|c^{\mathbf{z}}_{\gamma^*}\|_0$	Error1	Error2	γ_1^*	LSRError1	LSRError2
300	6.261e-3	16	9.708e-2	1.244e-1	6.769e-3	0.3936	0.4951
600	6.769e-3	13	8.472e-2	1.077e-1	5.790e-3	0.3986	0.5042
1200	3.625e-3	16	6.569e-2	9.000e-2	4.582e-3	0.5229	0.6534
1800	2.270e-3	25	5.054e-2	6.467e-2	3.101e-3	0.5500	0.6945
2400	2.099e-3	20	4.289e-2	6.249e-2	2.653e-3	0.5246	0.6764

Table 2: Learning Error

respectively, which are selected from a geometric sequence $\{10^{-4}, \cdots, 10^{-2}\}$ *of length* 60 *by* 5*-fold cross validation. The learning error is estimated empirically by independently drawing another unlabelled sample set* $\{\xi_j\}$ *uniformly on* $[0,1]^{10}$ *of size* $12,000$ *and with* $f^{\mathbf{z}} = f^{\mathbf{z}}_{\gamma^*}$ *or* $f^{\mathbf{z}}_{\mathrm{LSR},\gamma_1^*}$ *computing*

$$Error1 = \frac{1}{12,000} \sum_{j=1}^{12,000} |f_\rho(\xi_j) - f^{\mathbf{z}}(\xi_j)|,$$

$$Error2 \quad = \quad \left(\frac{1}{12,000} \sum_{j=1}^{12,000} (f_\rho(\xi_j) - f^{\mathbf{z}}(\xi_j))^2 \right)^{1/2} .$$

We have observed sparsity for the coefficients in the representation (3) of the output function in our algorithm. This sparsity is different from that for the representation in terms of $\{K_{x_i}\}_{i=1}^m$. It would be interesting to extend our study to a semisupervised learning setting as indicated in Remark 4. Another extension is to take empirical features in different ways by means of efficient numerical methods for the Gramian matrix $\mathbb{K}$. Exploring sparsity in such extended settings would be of much value for applications.

References

[1] R. Bhatia, L. Elsner, The Hoffman-Wielandt inequality in infinite dimensions, Proc. Indian Acad. Sci. (Math. Sci.) 104 (1994), 483-494.

[2] G. Blanchard, P. Massart, R. Vert, L. Zwald, Kernel projection machine: a new tool for pattern recognition, Proc. NIPS (2004), 1649-1656.

[3] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE. Trans. Inform. Theory 52 (2006), 489-509.

[4] A. Caponnetto, E. De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (2007), 331-368.

[5] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Anal. Appl. 8 (2010), 161-183.

[6] D.L. Donoho, Compressed sensing, IEEE. Trans. Inform. Theory 52 (2006), 1289-1306.

[7] A.J. Hoffman, H.W. Wielandt, The variation of the spectrum of a normal matrix, Duke Math. J. 20 (1953), 37-39.

[8] T. Kato, Variation of discrete spectra, Commun. Math. Phys. 111 (1987), 501-504.

[9] V. Koltchinskii, E. Giné, Random matrix approximation of spectra of integral operators, Bernoulli 6 (2000), 113-167.

[10] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, A. Verri, Spectral algorithms for supervised learning, Neural Comput. 20 (2008), 1873-1897.

[11] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. 22 (1994), 1679-1706.

[12] J.B. Reade, Eigenvalues of positive definite kernels II, SIAM J. Math. Anal. 15 (1984), 137-142.

[13] J.B. Reade, Eigenvalues of analytic kernels, SIAM J. Math. Anal. 15 (1984), 133-136.

[14] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007), 153-172.

[15] S. Smale, D.X. Zhou, Geometry on probability spaces, Constr. Approx. 30 (2009), 311-323.

[16] S. Smale and D.X. Zhou, Online learning with Markov sampling, Anal. Appl. 7 (2009), 87-113.

[17] I. Steinwart, D. Hush, C. Scovel, Optimal rates for regularized least-squares regression, in Proceedings of the 22nd Annual Conference on Learning Theory (S. Dasgupta and A. Klivans eds.), 2009, pp. 79-93.

[18] H.W. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, Appl. Comput. Harmon. Anal. 30 (2011), 96-109.

[19] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Statist. Soc. B 58 (1996), 267-288.

[20] U. von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, Ann. Stat. 36 (2008), 555-586.

[21] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003), 1743-1752.

[22] L. Zwald, G. Blanchard, On the convergence of eigenspaces in kernel principal component analysis, In Advances in Neural Information Processing Systems 18 (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pages 1649-1656. MIT Press, Cambridge, MA, 2006.

[23] L. Zwald, Performances statistiques d'algorithmes d'apprentissage: Kernel Projection Machine et analyse en composantes principales à noyau, PhD thesis, Université Paris-Sud 11, 2005.