Thresholded Spectral Algorithms for Sparse Approximations

Zheng-Chu Guo

School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China

guozhengchu@zju.edu.cn

Dao-Hong Xiang

Department of Mathematics, Zhejiang Normal University, Jinhua 321004, China

Department of Mathematics, University of Bayreuth, Germany

daohongxiang@zjnu.cn

Xin Guo

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

x.guo@polyu.edu.hk

Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong, Hong Kong

mazhou@cityu.edu.hk

Abstract

Spectral algorithms form a general framework that unifies many regularization schemes in learning theory. In this paper, we propose and analyze a class of thresholded spectral algorithms that are designed based on empirical features. Soft thresholding is adopted to achieve sparse approximations. Our analysis shows that without sparsity assumption of the regression function, the output functions of thresholded spectral algorithms are represented by empirical features with satisfactory sparsity, and the convergence rates are comparable to those of the classical spectral algorithms in the literature.

Key words and phrases. Learning theory; thresholded spectral algorithm; sparsity; learning rate.

AMS Subject Classification Numbers. 68T05, 68Q32, 41A25.

1 Introduction and Main Results

Consider a regression problem modeled by an unknown probability measure ρ on the product space $X \times Y$ of some compact metric (input) space X and output space $Y \subset \mathbb{R}$. Given a training sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn independently from ρ, the regression problem aims at approximating the regression function defined by

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

where $\rho(\cdot|x)$ is the conditional distribution of ρ at $x \in X$. We consider approximations provided by regularization schemes in reproducing kernel Hilbert spaces.

A classical well understood learning algorithm is the regularized least squares, and there is a large literature on error analysis for this algorithm, see e.g. [24, 6, 21, 18, 19, 17]. Recently, optimal learning rates (minimax rates of convergence) for this classical learning algorithm are established in [16]. The regularized least squares algorithm suffers from the so-called saturation phenomenon [10]. Spectral algorithms introduced in [9, 10] with motivations from regularization methods in learning theory and inverse problems can overcome the saturation phenomenon. A spectral algorithm is defined in terms of a Mercer kernel $K : X \times X \to \mathbb{R}$ (which is a continuous, symmetric, and positive semi-definite function) and a filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with a parameter $\lambda > 0$, where $\kappa := \max\left\{1, \sup_{x\in X} \sqrt{K(x,x)}\right\}$. The regularized least squares algorithm is included in the family of spectral algorithms [2, 10] with filter function $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$. Some other members in this family include the Landweber iteration (or gradient descent [23]) with the filter function $g_\lambda(\sigma) = \sum_{i=1}^{t-1}(1-\sigma)^i$ parameterized by $\lambda = \frac{1}{t}$ for an integer $t \in \mathbb{N}$, and spectral cut-off with filter function $g_\lambda(\sigma) = \frac{1}{\sigma}$ if $\sigma \geq \lambda$, and $g_\lambda = 0$ otherwise. Spectral algorithms were well studied for regression in learning theory, see e.g. [10, 2, 7, 4, 8]. Recently, optimal learning rates of spectral algorithms for regression are derived in [14] by a new error decomposition technique and a second order decomposition approach. We refer the reader to [2, 14] and reference therein for more details about the spectral algorithms.

Along a different direction, learning algorithms producing sparse approximations have attracted much attention within the last decade. Besides the spectral cut-off in the family of spectral algorithms, the ℓ^1-regularizer often leads to sparsity, which has been observed in the LASSO algorithm [20, 25, 26] and in compressed sensing. Recently, empirical feature based ℓ^1 regularization was proposed in [11]. Given the sample $\mathbf{z}$, empirical features are a set of functions $\{\phi_i^{\mathbf{x}}\}_{i=1}^\infty$ on X, based on $\mathbf{x} = \{x_i\}_{i=1}^m$ (to be defined below). The output function of the empirical feature based ℓ^1 regularization algorithm is represented by $\sum_{i\in\mathbb{N}} c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{z}}$ with the coefficients $\{c_{\gamma,i}^{\mathbf{z}}\}_{i=1}^\infty$ given by a coefficient based ℓ^1 regularization algorithm. The algorithm is a modification of the kernel projection machine (KPM) introduced in [1] and analyzed in [27]. Sparsity and error analysis of this algorithm for regression are investigated in [11]. It is valuable due to the satisfactory learning rates and strong sparsity under a mild condition and without any sparsity assumptions on the regression function. Empirical feature based algorithms with sparse properties induced by concave regularizers for regression are studied in [12] recently.

Motivated by the spectral algorithms and the empirical feature based ℓ^1 regularization algorithm, we propose a new learning algorithm associated with a filter function g_λ and soft thresholding function G_γ which provides satisfactory learning rates for regression and strong sparsity without any sparsity assumption about the regression function. We first define the filter function g_λ and the soft thresholding function G_γ as follows.

Definition 1.1 (Filter function). *We say that $g_\lambda : [0, \kappa^2] \mapsto \mathbb{R}$, with $0 < \lambda \leq \kappa^2$, is a filter function with qualification $\nu_g \geq \frac{1}{2}$ if there exists a positive constant b independent of λ such that*

$$\sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)| \leq \frac{b}{\lambda}, \qquad \sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)\sigma| \leq b, \qquad \forall 0 < \lambda \leq \kappa^2 \tag{1}$$

and

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leq c_\nu \lambda^\nu, \qquad \forall\ 0 < \nu \leq \nu_g, 0 < \lambda \leq \kappa^2, \tag{2}$$

where $c_\nu > 0$ is a constant depending only on $\nu \in (0, \nu_g]$.

Definition 1.2 (Soft thresholding function). *The soft thresholding function $G_\gamma : \mathbb{R} \mapsto \mathbb{R}$ with a thresholding parameter $\gamma > 0$ is defined as*

$$G_\gamma(\sigma) = \begin{cases} 0, & \text{if } |\sigma| \leq \frac{\gamma}{2}, \\ \sigma - \frac{\gamma}{2} & \text{if } \sigma > \frac{\gamma}{2}, \\ \sigma + \frac{\gamma}{2} & \text{if } \sigma < -\frac{\gamma}{2}. \end{cases} \tag{3}$$

Let $(\mathcal{H}_K, \langle \cdot, \cdot \rangle)$ with the norm $\| \cdot \|_K$ be the reproducing kernel Hilbert space (RKHS) generated by the Mercer kernel. To define empirical features $\{\phi_i^{\mathbf{x}}\}_i$, we need the empirical integral operator $L_K^{\mathbf{x}}$ on $\mathcal{H}_K$ associated with the kernel K and the input data $\mathbf{x} = \{x_1, \cdots, x_m\}$ given by

$$L_K^{\mathbf{x}}(f) = \frac{1}{m}\sum_{i=1}^m f(x_i)K_{x_i}, \qquad f \in \mathcal{H}_K, \tag{4}$$

where K_x is a function in $\mathcal{H}_K$ defined by $K_x(u) = K(x, u)$. Recall the reproducing property $\langle f, K_x \rangle = f(x)$ for any $f \in \mathcal{H}_K$. Let $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ be a set of normalized eigenpairs of $L_K^{\mathbf{x}}$ with the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ forming an orthonormal basis of $\mathcal{H}_K$, then the functions $\{\phi_i^{\mathbf{x}}\}_{i=1}^\infty$ are called empirical features which can be computed explicitly [1, 11] by the eigenpairs of the Gramian matrix $(K(x_i, x_i))_{i,j=1}^m$. Denote

$$f_\rho^{\mathbf{z}} = \frac{1}{m}\sum_{i=1}^m y_i K_{x_i}. \tag{5}$$

Definition 1.3 (Thresholded spectral algorithms)**.** *A thresholded spectral algorithm for sparsity associated with a filter function* $g_\lambda : [0, \kappa^2] \mapsto \mathbb{R}$ *and a soft thresholding function* $G_\gamma : \mathbb{R} \mapsto \mathbb{R}$ *is defined by*

$$f_{\mathbf{z}} = g_\lambda(L_K^{\mathbf{x}}) \sum_{i \in \mathbb{N}} G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}} \rangle) \phi_i^{\mathbf{x}} = \sum_{i \in \mathbb{N}} c_{\lambda,\gamma,\mathbf{z}}^i \phi_i^{\mathbf{x}}, \tag{6}$$

where $g_\lambda(L_K^{\mathbf{x}})$ *is the operator on* $\mathcal{H}_K$ *defined by spectral calculus as* $g_\lambda(L_K^{\mathbf{x}}) = \sum_i g_\lambda(\lambda_i^{\mathbf{x}}) \phi_i^{\mathbf{x}} \otimes \phi_i^{\mathbf{x}} = \sum_i g_\lambda(\lambda_i^{\mathbf{x}}) \langle \cdot, \phi_i^{\mathbf{x}} \rangle \phi_i^{\mathbf{x}}$.

Remark 1.4. *If we replace the thresholding function* G_γ *by the identity function (which is the limit of* G_γ *as* $\gamma \to 0^+$*), the thresholded spectral algorithm (6) is reduced to the ordinary spectral algorithm* $f_{\mathbf{z}} = g_\lambda(L_K^{\mathbf{x}}) \left(\frac{1}{m} \sum_{i=1}^m y_i K_{x_i} \right)$ *that is studied in [2, 10].*

If we replace the filter function $g_\lambda(\sigma)$ *simply by the reciprocal function* $1/\sigma$ *(which is no longer a filter function, but could be regarded as the limit of* g_λ *if we push* $b \to \infty$ *in (1) and* $c_\nu \to 0^+$ *in (2)), the thresholded spectral algorithm (6) is reduced to the empirical feature-based scheme studied in [11].*

Let ρ_X be the marginal distribution of ρ on X and $(L^2_{\rho_X}, \|\cdot\|_\rho)$ the Hilbert space of ρ_X square integrable functions on X. Define the integral operator L_K on $\mathcal{H}_K$ associated with the Mercer kernel K by

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

Our error analysis is based on the following regularity condition

$$f_\rho = L_K^r(u_\rho) \ \text{ for some } r > 0 \text{ and } u_\rho \in \mathcal{H}_K, \tag{7}$$

where L_K^r denotes the r-th power of L_K on $\mathcal{H}_K$ since $L_K : \mathcal{H}_K \to \mathcal{H}_K$ is a compact and positive operator. The eigenvalues $\{\lambda_i\}$ of L_K are arranged in a nonincreasing order and its corresponding normalized eigenfunctions $\{\phi_i\}_i$ form an orthonormal basis of $\mathcal{H}_K$. Condition (7) is equivalent to $f_\rho = L_K^r u_\rho = \sum_{j \in \mathbb{N}} \lambda_j^r d_j \phi_j$ with $u_\rho = \sum_{j \in \mathbb{N}} d_j \phi_j$ and $\|u_\rho\|_K = \|\{d_j\}\|_{\ell^2}$. Throughout the paper we assume $|y| \leq M$ almost surely for some constant $M > 0$. It would be interesting to extend our analysis to more general situations by assuming exponential decays for $f_\rho(x) - y$ or some moment conditions [3, 13].

To illustrate our general analysis (stated in Theorem 4.1 below) on sparsity and learning rates for the thresholded spectral algorithm (6), we state the following result in a special situation when the eigenvalues $\{\lambda_i\}_{i=1}^\infty$ of L_K decay polynomially.

Theorem 1.5. *Suppose that regularity condition (7) holds with some* $0 < r \leq \nu_g$ *and the eigenvalues* $\{\lambda_i\}$ *of* L_K *decay polynomially as*

$$D_1 i^{-\alpha} \leq \lambda_i \leq D_2 i^{-\alpha}, \quad \forall i \in \mathbb{N} \tag{8}$$

with $\alpha > 1$ and $0 < D_1 \leq D_2$. Let $0 < \delta < 1$. If we choose

$$\gamma = C_0(D_2/\kappa^2 + 1)^{r+1} \left(\log \frac{6m}{\delta}\right)^{r+1} m^{-\frac{r+1}{\max\{2,2r+1\}}} \tag{9}$$

and

$$\lambda = \begin{cases} m^{-\frac{1}{2}+\frac{1}{4\alpha(1+r)}}, & \text{if } 0 < r \leq \frac{1}{2}, \\ m^{-\frac{2\alpha(r+1)-1}{2\alpha(2r+1)(r+1)}}, & \text{if } \frac{1}{2} < r \leq 1, \\ m^{-\frac{3\alpha-1}{4\alpha(1+r)}}, & \text{if } r > 1, \end{cases} \tag{10}$$

then with confidence at least $1-\delta$, we have

$$c^i_{\lambda,\gamma,\mathbf{z}} = 0, \quad \forall i \geq m^{\frac{1}{\alpha \max\{2,2r+1\}}} + 1 \tag{11}$$

and

$$\|f_{\mathbf{z}} - f_\rho\|_\rho \leq C \left(\log \frac{6m}{\delta}\right)^{\max\{r+2,\frac{5}{2}\}} m^{-\theta},$$

where

$$\theta = \begin{cases} \frac{2r+1}{4} - \frac{1}{4\alpha}, & \text{if } 0 < r \leq \frac{1}{2}, \\ \frac{1}{2} - \frac{1}{2\alpha(2r+1)}. & \text{if } r > \frac{1}{2}. \end{cases} \tag{12}$$

and the constants C_0, C are independent of m or δ.

Remark 1.6. *Our learning rate index θ can be of type $\frac{1}{2} - \epsilon$ with an arbitrarily small $\epsilon > 0$ when the regularity parameter r in condition (7) is large enough. Note that condition (8) is satisfied by Sobolev kernels on domains of Euclidean spaces for which α depends on the smoothness index of the Sobolev space. Theorem 1.5 is a corollary of Theorem 4.2 to be given in Section 4.*

If the eigenvalues $\{\lambda_i\}$ decay exponentially, we have the following result.

Theorem 1.7. *Suppose that regularity condition (7) holds with some $0 < r \leq \nu_g$ and the eigenvalues $\{\lambda_i\}$ of L_K decay exponentially as*

$$D_1\beta^{-i} \leq \lambda_i \leq D_2\beta^{-i}, \quad \forall i \in \mathbb{N} \tag{13}$$

with $\beta > 1$ and $0 < D_1 \leq D_2$. Let $0 < \delta < 1$. If we choose

$$\gamma = C_0(D_2/\kappa^2 + 1)^{r+1} \left(\log \frac{6m}{\delta}\right)^{r+1} m^{-\frac{r+1}{\max\{2,2r+1\}}} \tag{14}$$

and $\lambda = D_2 \frac{\log m}{m}$, then with confidence at least $1-\delta$, we have

$$c^i_{\lambda,\gamma,\mathbf{z}} = 0, \quad \forall i \geq \frac{\log(m+1)}{(2r+1)\log\beta} + 1 \tag{15}$$

and

$$\|f_{\mathbf{z}} - f_\rho\|_\rho \leq C' m^{-\min\{r,\frac{1}{2}\}} \left(\log \frac{6m}{\delta}\right)^{r+\frac{5}{2}}, \tag{16}$$

where the constants C_0, C' are independent of m or δ.

Remark 1.8. *We note that when $r \geq 1/2$, the convergence rate in (16), subject to a logarithmic term, is of order $O(m^{-1/2})$. This demonstrates the surprising power of the thresholding methods with respect to empirical features. The proof of Theorem 1.7 will be given in Section 4.*

The remaining part of this paper is organized as follows. In Section 2, some preliminary analysis of sparsity and error for algorithm (6) is carried out. Section 3 develops some technical results which are needed to derive the general results. In Section 4, we first prove the general results of the sparsity and error analysis of algorithm (6), and then we establish the learning rates when the eigenvalues of L_K decay polynomially or exponentially.

2 Error Decomposition

In this section, some preliminary analysis of sparsity and error for algorithm (6) associated with a filter function g_λ and a soft thresholding function G_γ is carried out.

2.1 Representation and decomposition for sparsity

From the definition of the soft thresholding function G_γ, we have $c^i_{\lambda,\gamma,\mathbf{z}} = 0$ when $|\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \rangle| \leq \frac{\gamma}{2}$. To get sparsity, we shall choose some $p = p(m) \in \mathbb{N}$ and $\gamma = \gamma(m, \delta)$ such that with confidence $1 - \delta$,

$$c^i_{\lambda,\gamma,\mathbf{z}} = 0, \quad \forall i \geq p + 1.$$

To this end, we need to estimate the term $|\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \rangle|$. Recall $f^{\mathbf{z}}_\rho = \frac{1}{m} \sum_{i=1}^m y_i K_{x_i}$. The conditional expectation of $f^{\mathbf{z}}_\rho$ given $x_1, \ldots, x_m$ is $\frac{1}{m} \sum_{i=1}^m f_\rho(x_i) K_{x_i} = L^{\mathbf{x}}_K f_\rho$. It shows that $\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \rangle$ is close to $\langle L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle$. So we introduce the following decomposition

$$\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \rangle = \langle f^{\mathbf{z}}_\rho - L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle + \langle L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle.$$

It follows that for $i \in \mathbb{N}$,

$$\left|\left\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \right\rangle\right| \leq \left|\left\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \right\rangle - \langle L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle\right| + |\langle L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle| := J_1 + J_2. \tag{17}$$

Note that the notations J_1, J_2 depend on i. They are used for simplicity.

2.2 Analysis for sparsity

In this subsection, we bound J_1 and J_2 respectively. The first term J_1 can be expressed as

$$J_1 = \left|\left\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \right\rangle - \langle L^{\mathbf{x}}_K f_\rho, \phi^{\mathbf{x}}_i \rangle\right| = \left|\left\langle f^{\mathbf{z}}_\rho, \phi^{\mathbf{x}}_i \right\rangle - \lambda^{\mathbf{x}}_i \left\langle f_\rho, \phi^{\mathbf{x}}_i \right\rangle\right| = |\lambda^{\mathbf{x}}_i| \left|\left\langle \frac{1}{\lambda^{\mathbf{x}}_i} f^{\mathbf{z}}_\rho - f_\rho, \phi^{\mathbf{x}}_i \right\rangle\right|.$$

It can be easily estimated by the following lemma proved in [12]. Denote $Z = X \times Y$.

Lemma 2.1. *For any $\delta \in (0,1)$, there exists a subset U_1 of Z^m with measure at least $1-\delta/3$ such that*

$$\sqrt{\lambda_i^{\mathbf{x}}}\left|\left\langle \frac{1}{\lambda_i^{\mathbf{x}}} f_\rho^{\mathbf{x}} - f_\rho, \phi_i^{\mathbf{x}}\right\rangle\right| \le \frac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\frac{6m}{\delta}}, \qquad \forall \mathbf{z}\in U_1, i\in\mathbb{N}. \tag{18}$$

Before stating the general result about sparsity, we give the following key lemma [12]. Recall the Hilbert-Schmidt norm of the operator $L_K - L_K^{\mathbf{x}}$ given by

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

Lemma 2.2. *We have*

$$\sum_{i=1}^{\infty}(\lambda_i - \lambda_i^{\mathbf{x}})^2 \le \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}^2. \tag{20}$$

For any $\delta\in(0,1)$, there exists a subset U_2 of Z^m with measure at least $1-\delta/3$ such that

$$\|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}} \le \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}, \qquad \forall \mathbf{z}\in U_2. \tag{21}$$

Proposition 2.3. *For $\mathbf{z}\in U_1\cap U_2$ and $i\in\mathbb{N}$, there holds*

$$J_1 \le \frac{2\sqrt{2}M}{\sqrt{m}}\left(\sqrt{\lambda_i} + \frac{2\kappa}{\sqrt[4]{m}}\sqrt{\log\frac{6}{\delta}}\right)\sqrt{\log\frac{6m}{\delta}}. \tag{22}$$

Proof. By Lemma 2.1, for $\mathbf{z}\in U_1$, we have

$$J_1 = |\lambda_i^{\mathbf{x}}|\left|\left\langle \frac{1}{\lambda_i^{\mathbf{x}}} f_\rho^{\mathbf{z}} - f_\rho, \phi_i^{\mathbf{x}}\right\rangle\right| = \sqrt{\lambda_i^{\mathbf{x}}}\sqrt{\lambda_i^{\mathbf{x}}}\left|\left\langle \frac{1}{\lambda_i^{\mathbf{x}}} f_\rho^{\mathbf{z}} - f_\rho, \phi_i^{\mathbf{x}}\right\rangle\right| \le \sqrt{\lambda_i^{\mathbf{x}}}\frac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\frac{6m}{\delta}}.$$

To bound $\sqrt{\lambda_i^{\mathbf{x}}}$, we have $|\lambda_i^{\mathbf{x}} - \lambda_i| \le \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}$ from (20). It follows that $\sqrt{\lambda_i^{\mathbf{x}}} \le \sqrt{\lambda_i} + \sqrt{\|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}}$, and by Lemma 2.2, for $\mathbf{z}\in U_2$, we have

$$\sqrt{\lambda_i^{\mathbf{x}}} \le \sqrt{\lambda_i} + \frac{2\kappa}{\sqrt[4]{m}}\sqrt{\log\frac{6}{\delta}},$$

Then our desired bound follows by combining the above two estimates. □

Estimating the second term J_2 of (17) is more involved.

Proposition 2.4. *For $\mathbf{z}\in U_2$ and $i\in\mathbb{N}$, there holds*

$$J_2 \le \left\{\frac{8\kappa^{2r+2}}{\sqrt{m}}\left(\lambda_i^{\min\{r,1\}} + \left(\frac{4\kappa^2}{\sqrt{m}}\right)^{\min\{r,1\}}\right) + 2^{2r}\left(\lambda_i^{1+r} + \left(\frac{4\kappa^2}{\sqrt{m}}\right)^{1+r}\right)\right\}\|u_\rho\|_K\left(\log\frac{6}{\delta}\right)^{1+r}.$$

Proof. Recall $f_\rho = L_K^r u_\rho = \sum_{j\in\mathbb{N}}\lambda_j^r d_j\phi_j$ with $u_\rho = \sum_{j\in\mathbb{N}} d_j\phi_j$ and $\|u_\rho\|_K = \|\{d_j\}\|_{\ell^2}$. Now we estimate the term $J_2 = |\langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle| = |\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle|$. To this end, we divide $\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle$ into two parts

$$\lambda_i^{\mathbf{x}}\langle f_\rho, \phi_i^{\mathbf{x}}\rangle = \left(\sum_{j:\lambda_j\ge 2\lambda_i^{\mathbf{x}}} + \sum_{j:\lambda_j< 2\lambda_i^{\mathbf{x}}}\right)\lambda_i^{\mathbf{x}}\lambda_j^r d_j\langle \phi_j, \phi_i^{\mathbf{x}}\rangle. \tag{23}$$

By Lemma 2.2 and (20), since $\lambda_1 \leq \|L_K\| \leq \kappa^2$, the first sum in (23) can be estimated as

$$\begin{aligned}\left|\sum_{j:\lambda_j \geq 2\lambda_i^{\mathbf{x}}} \lambda_i^{\mathbf{x}} \lambda_j^r d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle\right| \leq & (2\lambda_i^{\mathbf{x}})^{\min\{r,1\}} \kappa^{2\max\{0,r-1\}} \sum_j |d_j| \, |\lambda_j - \lambda_i^{\mathbf{x}}| \, |\langle \phi_j, \phi_i^{\mathbf{x}} \rangle| \\ \leq & (2\lambda_i^{\mathbf{x}})^{\min\{r,1\}} \kappa^{2r} \|u_\rho\|_K \frac{4\kappa^2}{\sqrt{m}} \log \frac{6}{\delta} \\ \leq & \frac{8\kappa^{2r+2}\|u_\rho\|_K}{\sqrt{m}} \left(\lambda_i^{\min\{r,1\}} + \left(\frac{4\kappa^2}{\sqrt{m}} \log \frac{6}{\delta} \right)^{\min\{r,1\}} \right) \log \frac{6}{\delta}.\end{aligned}$$

Note that for $t \geq 1$ and $a, b > 0$, one has $(a+b)^t \leq 2^{t-1}(a^t + b^t)$. Hence the second sum in (23) is bounded by

$$\begin{aligned}\left|\sum_{j:\lambda_j < 2\lambda_i^{\mathbf{x}}} \lambda_i^{\mathbf{x}} \lambda_j^r d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle\right| \leq & 2^r (\lambda_i^{\mathbf{x}})^{1+r} \sum_j |d_j \left\langle \phi_j, \phi_i^{\mathbf{x}} \right\rangle| \\ \leq & 2^{2r} \left(\lambda_i^{1+r} + \left(\frac{4\kappa^2}{\sqrt{m}} \log \frac{6}{\delta} \right)^{1+r} \right) \|u_\rho\|_K.\end{aligned}$$

This completes the proof. □

Putting the bounds in Proposition 2.3 and Proposition 2.4 into (17),we obtain the following upper bound for $\left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}} \right\rangle\right|$ with $i \in \mathbb{N}$.

Proposition 2.5. *Under regularity condition (7) with some $r > 0$, for $\mathbf{z} \in U_1 \cap U_2$ and $i \in \mathbb{N}$, we have*

$$\left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}} \right\rangle\right| \leq \frac{C_0}{2} \left(\log \frac{6m}{\delta} \right)^{1+r} \begin{cases} \max\left\{ \frac{\lambda_i}{\kappa^2}, \frac{1}{\sqrt{m}} \right\}^{1+r}, & \text{for } 0 < r \leq \frac{1}{2}, \\ \max\left\{ \left(\frac{\lambda_i}{\kappa^2}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}} \right\} \max\left\{ \frac{\lambda_i}{\kappa^2}, \frac{1}{\sqrt{m}} \right\}^{\frac{1}{2}}, & \text{for } r > \frac{1}{2}, \end{cases}$$

where C_0 is the constant independent of δ, m, or λ_i given by

$$C_0 = 64 \max\{\sqrt{2}\kappa M, 8\kappa^{2r+4}\|u_\rho\|_K, 4^{2r}\kappa^{2r+2}\|u_\rho\|_K\}. \tag{24}$$

Proof. Putting the bounds for J_1 and J_2 in Proposition 2.3 and Proposition 2.4 back into (17), we know that for $\mathbf{z} \in U_1 \cap U_2$ and $i \in \mathbb{N}$,

$$\begin{aligned}\left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}} \right\rangle\right| \leq & 4\sqrt{2}\kappa M \left(\frac{\sqrt{\lambda_i}}{\kappa\sqrt{m}} + \left(\frac{1}{\sqrt{m}} \right)^{\frac{3}{2}} \right) \log \frac{6m}{\delta} \\ & + 32\kappa^{2r+4}\|u_\rho\|_K \left(\frac{(\lambda_i/\kappa^2)^{\min\{r,1\}}}{\sqrt{m}} + \left(\frac{1}{\sqrt{m}} \right)^{\min\{r,1\}+1} \right) \left(\log \frac{6}{\delta} \right)^{r+1} \\ & + 4^{2r+1}\kappa^{2r+2}\|u_\rho\|_K \left(\left(\frac{\lambda_i}{\kappa^2} \right)^{1+r} + \left(\frac{1}{\sqrt{m}} \right)^{1+r} \right) \left(\log \frac{6}{\delta} \right)^{1+r} \\ \leq & \tilde{C}_0 \left(\frac{\sqrt{\lambda_i}}{\kappa\sqrt{m}} + 3 \left(\frac{1}{\sqrt{m}} \right)^{\min\{\frac{3}{2}, r+1\}} + \frac{(\frac{\lambda_i}{\kappa^2})^{\min\{r,1\}}}{\sqrt{m}} + \left(\frac{\lambda_i}{\kappa^2} \right)^{1+r} \right) \left(\log \frac{6}{\delta} \right)^{1+r},\end{aligned}$$

where $\tilde{C}_0 = \max\{4\sqrt{2}\kappa M, 32\kappa^{2r+4}\|u_\rho\|_K, 4^{2r+1}\kappa^{2r+2}\|u_\rho\|_K\}$.

When $0 < r \leq 1/2$, one has $\min\{r, 1\} = r$. So

$$\left(\frac{\lambda_i}{\kappa^2}\right)^{\min\{r,1\}} = \left(\frac{\lambda_i}{\kappa^2}\right)^r \geq \frac{\sqrt{\lambda_i}}{\kappa}, \text{ and } \left(\frac{1}{\sqrt{m}}\right)^{\min\{\frac{3}{2},r+1\}} = \left(\frac{1}{\sqrt{m}}\right)^{1+r}.$$

One applies Young's inequality to obtain

$$\left(\frac{\lambda_i}{\kappa^2}\right)^r \frac{1}{\sqrt{m}} \leq \frac{(\lambda_i/\kappa^2)^{1+r}}{1+\frac{1}{r}} + \frac{(1/\sqrt{m})^{1+r}}{1+r}. \tag{25}$$

Therefore, for $\mathbf{z} \in U_1 \cap U_2$ and $i \in \mathbb{N}$, we have

$$\begin{aligned} |\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle| &\leq \tilde{C}_0 \left(5\left(\frac{1}{\sqrt{m}}\right)^{1+r} + 3\left(\frac{\lambda_i}{\kappa^2}\right)^{1+r}\right)\left(\log\frac{6m}{\delta}\right)^{1+r} \\ &\leq 8\tilde{C}_0 \max\left\{\frac{1}{\sqrt{m}}, \frac{\lambda_i}{\kappa^2}\right\}^{1+r}\left(\log\frac{6m}{\delta}\right)^{1+r}. \end{aligned}$$

When $r > 1/2$, one has

$$\left(\frac{1}{\sqrt{m}}\right)^{\min\{\frac{3}{2},r+1\}} = \left(\frac{1}{\sqrt{m}}\right)^{\frac{3}{2}}, \text{ and } \left(\frac{\lambda_i}{\kappa^2}\right)^{\min\{r,1\}} \leq \frac{\sqrt{\lambda_i}}{\kappa},$$

which implies that for $\mathbf{z} \in U_1 \cap U_2$ and $i \in \mathbb{N}$,

$$\begin{aligned} |\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle| &\leq \tilde{C}_0 \left(3\left(\frac{1}{\sqrt{m}}\right)^{3/2} + 2\frac{\sqrt{\lambda_i}}{\kappa\sqrt{m}} + \left(\frac{\lambda_i}{\kappa^2}\right)^{1+r}\right)\left(\log\frac{6m}{\delta}\right)^{1+r} \\ &\leq 6\tilde{C}_0 \max\left\{\left(\frac{\lambda_i}{\kappa^2}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}}\right\}\left(\max\left\{\frac{\lambda_i}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}\right)^{\frac{1}{2}}\left(\log\frac{6m}{\delta}\right)^{1+r}. \end{aligned}$$

The proof is thus completed by letting $C_0 = 16\tilde{C}_0$. □

Proposition 2.6. *Assume regularity condition (7) with some $0 < r \leq \nu_g$. If γ satisfies*

$$\gamma \geq \begin{cases} C_0\left(\log\frac{6m}{\delta}\right)^{1+r}\max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}^{1+r}, & \text{for } 0 < r \leq \frac{1}{2}, \\ C_0\left(\log\frac{6m}{\delta}\right)^{1+r}\max\left\{\left(\frac{\lambda_p}{\kappa^2}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}}\right\}\max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}^{\frac{1}{2}}, & \text{for } r > \frac{1}{2}, \end{cases} \tag{26}$$

where C_0 is the constant given by (24), then we have

$$c^i_{\gamma,\lambda,\mathbf{z}} = 0, \qquad \forall \mathbf{z} \in U_1 \cap U_2, i \geq p+1.$$

Proof. If γ satisfies condition (26), by Proposition 2.5, for $\mathbf{z} \in U_1 \cap U_2$ and $i \geq p+1$, we have

$$|\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle \leq |\frac{\gamma}{2},$$

which implies $c^i_{\gamma,\lambda,\mathbf{z}} = 0$ by the definition of G_γ. □

2.3 Representation and decomposition for error

Now we consider how the output function $f_{\mathbf{z}}$ of algorithm (6) approximates f_ρ. Convergence rates of the approximation in $L^2_{\rho_X}$ will be established. We first make an error decomposition [5, 22, 15] and divide the error $f_{\mathbf{z}} - f_\rho$ into two parts as

$$f_{\mathbf{z}} - f_\rho = \left(f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}}\right) + \left(\tilde{f}_{\mathbf{x}} - f_\rho\right),$$

where

$$\tilde{f}_{\mathbf{x}} = g_\lambda(L_K^{\mathbf{x}})L_K^{\mathbf{x}} f_\rho.$$

For the first part, we see from the identity $\|f\|_\rho = \left\|L_K^{1/2} f\right\|_K$ for $f \in \mathcal{H}_K$ and the fact $\left\|L_K^{1/2}(\lambda I + L_K)^{-1/2}\right\| \leq 1$ that

$$\|f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}}\|_\rho = \left\|L_K^{1/2}(f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}})\right\|_K \leq \left\|(\lambda I + L_K)^{1/2}(f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}})\right\|_K.$$

Then we use the expression $\tilde{f}_{\mathbf{x}} = g_\lambda(L_K^{\mathbf{x}})L_K^{\mathbf{x}} f_\rho = g_\lambda(L_K^{\mathbf{x}}) \sum_{i\in\mathbb{N}} \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle \phi_i^{\mathbf{x}}$ to divide the term $(\lambda I + L_K)^{1/2}(f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}})$ further as

$$\begin{aligned}
(\lambda I + L_K)^{1/2}(f_{\mathbf{z}} - \tilde{f}_{\mathbf{x}}) &= (\lambda I + L_K)^{1/2} g_\lambda(L_K^{\mathbf{x}}) \sum_i \left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right) \phi_i^{\mathbf{x}} \\
&= (\lambda I + L_K)^{1/2}(\lambda I + L_K^{\mathbf{x}})^{-1/2}(\lambda I + L_K^{\mathbf{x}})^{1/2} g_\lambda(L_K^{\mathbf{x}})(\lambda I + L_K^{\mathbf{x}})^{1/2} \\
&\quad (\lambda I + L_K^{\mathbf{x}})^{-1/2} \sum_i \left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right) \phi_i^{\mathbf{x}} \\
&= (\lambda I + L_K)^{1/2}(\lambda I + L_K^{\mathbf{x}})^{-1/2} g_\lambda(L_K^{\mathbf{x}})(\lambda I + L_K^{\mathbf{x}}) \\
&\quad (\lambda I + L_K^{\mathbf{x}})^{-1/2} \sum_i \left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right) \phi_i^{\mathbf{x}}.
\end{aligned}$$

The property (1) of the filter function g_λ tells us that $\|g_\lambda(L_K^{\mathbf{x}})(\lambda I + L_K^{\mathbf{x}})\| \leq 2b$, which implies the following error decomposition.

Proposition 2.7. *We have*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho \leq I_1 + I_2, \tag{27}$$

where

$$I_1 = 2b \left\|(\lambda I + L_K)^{1/2}(\lambda I + L_K^{\mathbf{x}})^{-1/2}\right\| \left\|(\lambda I + L_K^{\mathbf{x}})^{-1/2} \sum_i \left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right) \phi_i^{\mathbf{x}}\right\|_K$$

and

$$I_2 = \|\tilde{f}_{\mathbf{x}} - f_\rho\|_\rho.$$

The term $\left\|(L_K + \lambda I)^{1/2}(L_K^{\mathbf{x}} + \lambda I)^{-1/2}\right\|$ and I_2 are well estimated in [14] as follows. Denote

$$\mathcal{B}_{m,\lambda} = \frac{2\kappa}{\sqrt{m}} \left\{\frac{\kappa}{\sqrt{m\lambda}} + \sqrt{\sum_{i\in\mathbb{N}} \frac{\lambda_i}{\lambda_i + \lambda}}\right\}.$$

Lemma 2.8. *For any $\delta \in (0,1)$, there exists a subset U_3 of Z^m with measure at least $1-\delta/3$ such that*

$$\left\|(L_K+\lambda I)^{1/2}(L_K^{\mathbf{x}}+\lambda I)^{-1/2}\right\| \le \left(\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)\log\frac{6}{\delta}, \quad \forall \mathbf{z}\in U_3. \tag{28}$$

Proposition 2.9. *For $\mathbf{z} \in U_2 \cap U_3$, there holds*

$$I_2 = \|\tilde{f}_{\mathbf{x}} - f_\rho\|_\rho \le \begin{cases} C_r'\left(\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)^{2r+1}\lambda^{r+\frac{1}{2}}\left(\log\frac{6}{\delta}\right)^{2r+1}, & \text{for } 0<r\le 1,\\ C_r\left(\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)\left(4\kappa^2\sqrt{\frac{\lambda}{m}}+\lambda^{r+\frac{1}{2}}\right)\left(\log\frac{6}{\delta}\right)^2, & \text{for } r>1,\end{cases} \tag{29}$$

where C_r, C_r' are constants independent of δ, m, or λ.

Proof. It was shown in [14] that $\|\tilde{f}_{\mathbf{x}} - f_\rho\|_\rho$ can be bounded by

$$\begin{cases} 2^{r+\frac{1}{2}}(b+1+c_{r+\frac{1}{2}})\|u_\rho\|_K\left\|(L_K+\lambda I)^{1/2}(L_K^{\mathbf{x}}+\lambda I)^{-1/2}\right\|^{2r+1}\lambda^{r+\frac{1}{2}}, & \text{for } 0<r\le 1,\\ C_r\left\|(L_K+\lambda I)^{1/2}(L_K^{\mathbf{x}}+\lambda I)^{-1/2}\right\|\left(\lambda^{\frac{1}{2}}\|L_K-L_K^{\mathbf{x}}\|+\lambda^{r+\frac{1}{2}}\right), & \text{for } r>1,\end{cases}$$

where $C_r, c_{r+\frac{1}{2}}$ are constants independent of δ, m, or λ. Then by Lemma 2.2 and Lemma 2.8 we know that for $\mathbf{z} \in U_2 \cap U_3$, there holds

$$\|\tilde{f}_{\mathbf{x}} - f_\rho\|_\rho \le \begin{cases} 2^{r+\frac{1}{2}}(b+1+c_{r+\frac{1}{2}})\|u_\rho\|_K\left(\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)^{2r+1}\lambda^{r+\frac{1}{2}}\left(\log\frac{6}{\delta}\right)^{2r+1}, & \text{for } 0<r\le 1,\\ C_r\left(\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)\left(4\kappa^2\sqrt{\frac{\lambda}{m}}+\lambda^{r+\frac{1}{2}}\right)\left(\log\frac{6}{\delta}\right)^2, & \text{for } r>1,\end{cases}$$

which verifies our desired result by setting $C_r' = 2^{r+\frac{1}{2}}(b+1+c_{r+\frac{1}{2}})\|u_\rho\|_K$. □

It remains to bound I_1 in (27). The core analysis of the error is for the term $\left\|(\lambda I + L_K^{\mathbf{x}})^{-1/2}\sum_i\left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right)\phi_i^{\mathbf{x}}\right\|_K$, which will be given in the next section.

3 Analysis of the Thresholding Error

In this section, we estimate the term $\left\|(\lambda I + L_K^{\mathbf{x}})^{-1/2}\sum_i(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho\rangle)\phi_i^{\mathbf{x}}\right\|_K$. To this end, we first prove that with γ chosen as in (26) and $p \in \mathbb{N}$, we have

$$c^i_{\gamma,\lambda,\mathbf{z}} = 0, \qquad \forall \mathbf{z} \in U_1 \cap U_2, i \in \mathbb{N}\setminus\mathcal{S}, \tag{30}$$

where

$$\mathcal{S} = \mathcal{S}_{p,\mathbf{x}} = \{i = 1,\dots,p : \lambda_i^{\mathbf{x}} > \lambda_p/2\}.$$

To verify this statement, we first notice from Proposition 2.6 that

$$c^i_{\gamma,\lambda,\mathbf{z}} = 0, \qquad \forall \mathbf{z} \in U_1 \cap U_2, i \ge p+1.$$

Then for $i \in \{1, \cdots, p\}$ and $i \notin \mathcal{S}$ (that is, $\lambda_i^{\mathbf{x}} \le \lambda_p/2$), we have from (18) and (21)

$$\begin{aligned}
\left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\right\rangle\right| \le & \left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\right\rangle - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right| + \left|\lambda_i^{\mathbf{x}} \langle f_\rho, \phi_i^{\mathbf{x}}\rangle\right| \\
\le & \frac{2\sqrt{2}M\sqrt{\lambda_i^{\mathbf{x}}}}{\sqrt{m}}\sqrt{\log\frac{6m}{\delta}} + \left|\left(\sum_{j:\lambda_j \ge 2\lambda_i^{\mathbf{x}}} + \sum_{j:\lambda_j < 2\lambda_i^{\mathbf{x}}}\right)\lambda_i^{\mathbf{x}}\lambda_j^r d_j \langle \phi_j, \phi_i^{\mathbf{x}}\rangle\right| \\
\le & \frac{2M\sqrt{\lambda_p}}{\sqrt{m}}\sqrt{\log\frac{6m}{\delta}} + \lambda_p^{\min\{r,1\}}\kappa^{2r}\|u_\rho\|_K\frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta} + \frac{1}{2}\lambda_p^{1+r}\|u_\rho\|_K \\
\le & \frac{C_1^*}{2}\begin{cases} \log\frac{6m}{\delta}\max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}^{1+r}, & \text{if } 0 < r \le \frac{1}{2}, \\ \log\frac{6m}{\delta}\max\left\{\left(\frac{\lambda_p}{\kappa^2}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}}\right\}\left(\frac{\lambda_p}{\kappa^2}\right)^{\frac{1}{2}}, & \text{if } r > \frac{1}{2}, \end{cases}
\end{aligned}$$

where $C_1^* = \max\{16\kappa M, 32\kappa^{3+2r}\|u_\rho\|_K\}$. Obviously, $C_0 \ge C_1^*$. Hence we have $\left|\left\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\right\rangle\right| \le \frac{\gamma}{2}$ and thereby

$$c_{\gamma,\lambda,\mathbf{z}}^i = 0, \qquad \forall \mathbf{z} \in U_1 \cap U_2, i \in \{1, \cdots, p\}, i \notin \mathcal{S}.$$

Combining the above two cases, we know that $c_{\gamma,\lambda,\mathbf{z}}^i = 0$, for $\mathbf{z} \in U_1 \cap U_2$, and $i \in \mathbb{N} \setminus \mathcal{S}$. which means (30) holds true.

Since $\{\phi_i^{\mathbf{x}}\}_i$ form an orthogonal basis of $\mathcal{H}_K$, when (30) holds true, we have for $\mathbf{z} \in U_1 \cap U_2$,

$$\begin{aligned}
& \left\|(\lambda I + L_K^{\mathbf{x}})^{-1/2}\sum_i (G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle)\phi_i^{\mathbf{x}}\right\|_K \\
= & \left\|\sum_i (\lambda + \lambda_i^{\mathbf{x}})^{-1/2}\left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right)\phi_i^{\mathbf{x}}\right\|_K \\
= & \left(\sum_i (\lambda + \lambda_i^{\mathbf{x}})^{-1}\left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right)^2\right)^{1/2} \\
\le & \left(\sum_{i \in \mathbb{N}\setminus\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2}{\lambda + \lambda_i^{\mathbf{x}}}(\langle f_\rho, \phi_i^{\mathbf{x}}\rangle)^2\right)^{1/2} + \left(\sum_{i \in \mathcal{S}} (\lambda + \lambda_i^{\mathbf{x}})^{-1}\left(G_\gamma(\langle f_\rho^{\mathbf{z}}, \phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}} f_\rho, \phi_i^{\mathbf{x}}\rangle\right)^2\right)^{1/2} \\
:= & I_{11} + I_{12},
\end{aligned}$$

where the last inequality holds due to (30).

We estimate I_{11} and I_{12} separately in the following. Denote

$$\Lambda_{p,m} = \max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}. \tag{31}$$

Proposition 3.1. *Let $p \in \mathbb{N}$. Under regularity condition (7) with some $r > 0$, we have for $\mathbf{z} \in U_1 \cap U_2 \cap U_3$,*

$$\begin{aligned}
I_{11} \le & 8\|u_\rho\|_K\kappa^{2r+2}\left\{\min\left\{\frac{\Lambda_{m,p}}{\sqrt{\lambda}}, \Lambda_{m,p}^{1/2}\right\}\lambda_p^r + \frac{\lambda_p^{\min\{1,r\}}\lambda^{\min\{r-\frac{1}{2},0\}}}{(2\lambda+\lambda_p)^{\min\{r,\frac{1}{2}\}}\sqrt{m}}\right. \\
& \left. + 2^r\lambda_p^{\min\{r-1,0\}}\Lambda_{m,p}^{1/2}\left(\left(\sum_{i=p+1}^\infty \lambda_i^{\max\{2,2r\}}\right)^{1/2} + \frac{8}{\sqrt{m}}\right)\right\}\left(\log\frac{6}{\delta}\right)^{\frac{3}{2}}.
\end{aligned}$$

Proof. Recall $f_\rho = L_K^r u_\rho = \sum_{j\in\mathbb{N}} \lambda_j^r d_j \phi_j$ with $u_\rho = \sum_{j\in\mathbb{N}} d_j\phi_j$ and $\|u_\rho\|_K = \|\{d_j\}\|_{\ell^2}$. We first decompose I_{11} as

$$\begin{aligned}
&\left(\sum_{i\in\mathbb{N}\backslash\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2}{\lambda+\lambda_i^{\mathbf{x}}} \langle f_\rho, \phi_i^{\mathbf{x}}\rangle^2\right)^{1/2} \\
\le &\left(\sum_{i\in\mathbb{N}\backslash\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2}{\lambda+\lambda_i^{\mathbf{x}}} \left\langle \sum_{j=p+1}^{\infty} \lambda_j^r d_j\phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2} \qquad (32)\\
&+\left(\left(\sum_{i\in\{1,\ldots,p\}\backslash\mathcal{S}} + \sum_{i=p+1}^{\infty}\right) \frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}} \sum_{j=1}^{p} \lambda_j^{2r} \langle \phi_j, \phi_i^{\mathbf{x}}\rangle^2\right)^{1/2}.
\end{aligned}$$

For $i \in \mathbb{N}\backslash\mathcal{S}$, one has $\lambda_i^{\mathbf{x}} \le \lambda_p/2$ when $i \le p$, and $\lambda_i^{\mathbf{x}} \le \lambda_i + \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}} \le \lambda_p + \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}$ when $i > p$. So the first part in (32) can be bounded as

$$\begin{aligned}
&\left(\sum_{i\in\mathbb{N}\backslash\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2}{\lambda+\lambda_i^{\mathbf{x}}} \left\langle \sum_{j=p+1}^{\infty} \lambda_j^r d_j\phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2} \\
\le &\min\left\{\lambda^{-1/2}\left(\lambda_p + \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}\right), \left(\lambda_p + \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}\right)^{1/2}\right\} \lambda_p^r \|u_\rho\|_K \\
\le &8\kappa^2 \min\left\{\frac{\Lambda_{m,p}}{\sqrt{\lambda}}, \Lambda_{m,p}^{1/2}\right\} \lambda_p^r \log\frac{6}{\delta}\|u_\rho\|_K.
\end{aligned}$$

Then we turn to the second part of (32).

When $r \le 1/2$, for $j = 1,\ldots,p$, and $i \in \{1,\ldots,p\}\backslash\mathcal{S}$, the fact $2\lambda_i^{\mathbf{x}} \le \lambda_p \le \lambda_j$ implies $\lambda_j \le 2(\lambda_j - \lambda_i^{\mathbf{x}})$, and that $(\lambda_i^{\mathbf{x}})^{2-2r}\lambda_j^{2r} \le (\lambda_p/2)^{2-2r}\lambda_j^{2r} \le 2^{2r-2}\lambda_j^2 \le 2^{2r}(\lambda_j - \lambda_i^{\mathbf{x}})^2$. We also have

$$\frac{(\lambda_i^{\mathbf{x}})^{2r}}{\lambda+\lambda_i^{\mathbf{x}}} = \left(\frac{\lambda_i^{\mathbf{x}}}{\lambda+\lambda_i^{\mathbf{x}}}\right)^{2r}\left(\frac{1}{\lambda+\lambda_i^{\mathbf{x}}}\right)^{1-2r} \le \left(\frac{\lambda_p}{2\lambda+\lambda_p}\right)^{2r}\lambda^{2r-1}.$$

It follows that

$$\begin{aligned}
&\left(\sum_{i\in\{1,\ldots,p\}\backslash\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}} \sum_{j=1}^{p} \lambda_j^{2r}\langle\phi_j,\phi_i^{\mathbf{x}}\rangle^2\right)^{1/2} \\
\le &\|u_\rho\|_K \left(\frac{\lambda_p}{2\lambda+\lambda_p}\right)^r \lambda^{r-\frac{1}{2}} \left(\sum_{i,j}(\lambda_j - \lambda_i^{\mathbf{x}})^2\langle\phi_j,\phi_i^{\mathbf{x}}\rangle^2\right)^{1/2} \\
\le &\|u_\rho\|_K \left(\frac{\lambda_p}{2\lambda+\lambda_p}\right)^r \lambda^{r-\frac{1}{2}} \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}.
\end{aligned}$$

When $1/2 < r \le 1$, one has

$$\frac{(\lambda_i^{\mathbf{x}})^2\lambda_j^{2r}}{\lambda+\lambda_i^{\mathbf{x}}} \le \frac{4^{-1}\lambda_p^2\lambda_j^{2r}}{\lambda+\lambda_p/2} \le \frac{4^{-1}\lambda_p^{2r}\lambda_j^2}{\lambda+\lambda_p/2} \le \frac{\lambda_p^{2r}(\lambda_j-\lambda_i^{\mathbf{x}})^2}{\lambda+\lambda_p/2} = \frac{2\lambda_p^{2r}(\lambda_j-\lambda_i^{\mathbf{x}})^2}{2\lambda+\lambda_p}.$$

Hence

$$\left(\sum_{i\in\{1,\dots,p\}\setminus\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}} \sum_{j=1}^p \lambda_j^{2r} \left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2}$$

$$\leq \|u_\rho\|_K \frac{\sqrt{2}\lambda_p^r}{\sqrt{2\lambda+\lambda_p}} \left(\sum_{i,j} (\lambda_j-\lambda_i^{\mathbf{x}})^2 \left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2} \leq \|u_\rho\|_K \frac{4\sqrt{2}\kappa^2\lambda_p^r}{\sqrt{2\lambda+\lambda_p}\sqrt{m}} \log\frac{6}{\delta}.$$

When $r>1$, we have $\lambda_j^{2r} \leq \lambda_1^{2r-2}\lambda_j^2 \leq 4\lambda_1^{2r-2}(\lambda_j-\lambda_i^{\mathbf{x}})^2$, and then

$$\left(\sum_{i\in\{1,\dots,p\}\setminus\mathcal{S}} \frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}} \sum_{j=1}^p \lambda_j^{2r} \left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2}$$

$$\leq \|u_\rho\|_K \frac{\sqrt{2}\lambda_p\lambda_1^{r-1}}{\sqrt{2\lambda+\lambda_p}} \left(\sum_{i,j} (\lambda_j-\lambda_i^{\mathbf{x}})^2 \left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2} \leq \frac{4\sqrt{2}\kappa^2\|u_\rho\|_K\lambda_1^{r-1}\lambda_p}{\sqrt{2\lambda+\lambda_p}\sqrt{m}} \log\frac{6}{\delta}.$$

Now we estimate the last part of (32) which restricts $i\geq p+1$ and $j\leq p$.

When $r\leq 1$, we have

$$\lambda_i^{\mathbf{x}}\lambda_j^{2r} \leq (\lambda_i + \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}})\lambda_j^{2r} \leq (\lambda_p^{2r-1} + \frac{4\kappa^2}{\sqrt{m}}\lambda_p^{2r-2}\log\frac{2}{\delta})\lambda_j^2,$$

and

$$\lambda_j^2 \leq 3(\lambda_i^2 + (\lambda_i-\lambda_i^{\mathbf{x}})^2 + (\lambda_j-\lambda_i^{\mathbf{x}})^2).$$

Hence

$$\left(\sum_{i=p+1}^{\infty} \frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}} \sum_{j=1}^p \lambda_j^{2r} \left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2}$$

$$\leq \sqrt{\sup_{x>0}\frac{x}{x+\lambda}}\|u_\rho\|_K \left(\lambda_p^{r-\frac{1}{2}} + \frac{2\kappa}{\sqrt[4]{m}}\lambda_p^{r-1}\sqrt{\log\frac{6}{\delta}}\right)$$

$$\times\sqrt{3}\left(\sum_{i=p+1}^{\infty}\sum_{j=1}^{p}\left(\lambda_i^2 + (\lambda_i-\lambda_i^{\mathbf{x}})^2 + (\lambda_j-\lambda_i^{\mathbf{x}})^2\right)\left\langle \phi_j, \phi_i^{\mathbf{x}}\right\rangle^2\right)^{1/2}$$

$$\leq \sqrt{3}\|u_\rho\|_K \left(\lambda_p^{r-\frac{1}{2}} + \frac{2\kappa}{\sqrt[4]{m}}\lambda_p^{r-1}\sqrt{\log\frac{6}{\delta}}\right)\left(\left(\sum_{i=p+1}^{\infty}\lambda_i^2\right)^{1/2} + \frac{4\sqrt{2}\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}\right).$$

When $r>1$, we have

$$\begin{aligned}\lambda_j^{2r} \leq& 2^{2r-1}\left(\lambda_i^{2r} + (\lambda_j-\lambda_i)^{2r}\right)\\ \leq& 2^{2r-1}\left(\lambda_i^{2r} + 2\lambda_j^{2r-1}(\lambda_j-\lambda_i^{\mathbf{x}})^2 + 2\lambda_j^{2r-1}(\lambda_i-\lambda_i^{\mathbf{x}})^2\right),\end{aligned}$$

which yields

$$\lambda_i^{\mathbf{x}}\lambda_j^{2r} \leq \left(\lambda_p + \frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}\right) 2^{2r-1}\left(\lambda_i^{2r} + 2\lambda_1^{2r-1}(\lambda_j - \lambda_i^{\mathbf{x}})^2 + 2\lambda_1^{2r-1}(\lambda_i - \lambda_i^{\mathbf{x}})^2\right).$$

Therefore

$$\begin{aligned}
&\left(\sum_{i=p+1}^{\infty}\frac{(\lambda_i^{\mathbf{x}})^2\|u_\rho\|_K^2}{\lambda+\lambda_i^{\mathbf{x}}}\sum_{j=1}^{p}\lambda_j^{2r}\langle\phi_j,\phi_i^{\mathbf{x}}\rangle^2\right)^{1/2}\\
\leq&\left(\sqrt{\lambda_p} + \frac{2\kappa}{\sqrt[4]{m}}\sqrt{\log\frac{6}{\delta}}\right)2^{r-\frac{1}{2}}\|u_\rho\|_K\left(\sum_{i=p+1}^{\infty}\lambda_i^{2r} + 4\lambda_1^{2r-1}\|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}^2\right)^{1/2}\\
\leq&2^{r-\frac{1}{2}}\|u_\rho\|_K\left(\sqrt{\lambda_p} + \frac{2\kappa}{\sqrt[4]{m}}\sqrt{\log\frac{6}{\delta}}\right)\left(\left(\sum_{i=p+1}^{\infty}\lambda_i^{2r}\right)^{1/2} + 2\lambda_1^{r-\frac{1}{2}}\frac{4\kappa^2}{\sqrt{m}}\log\frac{6}{\delta}\right).
\end{aligned}$$

Combining all the above estimates verifies our desired bound. □

Now we turn to estimate

$$I_{12} = \left(\sum_{i\in\mathcal{S}}(\lambda+\lambda_i^{\mathbf{x}})^{-1}(G_\gamma(\langle f_\rho^{\mathbf{z}} - L_K^{\mathbf{x}}f_\rho, \phi_i^{\mathbf{x}}\rangle))^2\right)^{1/2}.$$

Proposition 3.2. *Let $p \in \mathbb{N}$. Under regularity condition (7) with some $0 < r \leq \nu_g$, for $\mathbf{z} \in U_1 \cap U_2 \cap U_3$, we have*

$$I_{12} \leq \sqrt{p}\left(\frac{\gamma}{\sqrt{4\lambda + 2\lambda_p}} + \frac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\frac{6m}{\delta}}\right). \tag{33}$$

Proof. From the definition (3) of the soft thresholding function G_γ, we find

$$|G_\gamma(\sigma) - \sigma| \leq \frac{\gamma}{2}, \qquad \forall\sigma\in\mathbb{R}.$$

This together with the eigenfunction relation $L_K^{\mathbf{x}}\phi_i^{\mathbf{x}} = \lambda_i^{\mathbf{x}}\phi_i^{\mathbf{x}}$ implies

$$\begin{aligned}
\left|G_\gamma(\langle f_\rho^{\mathbf{z}},\phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}}f_\rho,\phi_i^{\mathbf{x}}\rangle\right| &\leq \frac{\gamma}{2} + \left|\langle f_\rho^{\mathbf{z}},\phi_i^{\mathbf{x}}\rangle - \lambda_i^{\mathbf{x}}\langle f_\rho,\phi_i^{\mathbf{x}}\rangle\right|\\
&= \frac{\gamma}{2} + \sqrt{\lambda_i^{\mathbf{x}}}\sqrt{\lambda_i^{\mathbf{x}}}\left|\langle\frac{1}{\lambda_i^{\mathbf{x}}}f_\rho^{\mathbf{z}},\phi_i^{\mathbf{x}}\rangle - \langle f_\rho,\phi_i^{\mathbf{x}}\rangle\right|.
\end{aligned}$$

Recall that for $i\in\mathcal{S}$, we have $\lambda_i^{\mathbf{x}} \geq \frac{\lambda_p}{2}$. It follows that

$$(\lambda+\lambda_i^{\mathbf{x}})^{-1/2}\left|G_\gamma(\langle f_\rho^{\mathbf{z}},\phi_i^{\mathbf{x}}\rangle) - \langle L_K^{\mathbf{x}}f_\rho,\phi_i^{\mathbf{x}}\rangle\right| \leq \frac{\gamma}{2\sqrt{\lambda+\lambda_p/2}} + \sqrt{\lambda_i^{\mathbf{x}}}\left|\langle\frac{1}{\lambda_i^{\mathbf{x}}}f_\rho^{\mathbf{z}},\phi_i^{\mathbf{x}}\rangle - \langle f_\rho,\phi_i^{\mathbf{x}}\rangle\right|.$$

Then our desired result follows from Lemma 2.1. □

4 Proof of the Main Results

In this section, we present the following general results on sparsity and error analysis of algorithm (6) and then derive the learning rates when the eigenvalues of the integral operator L_K decay polynomially or exponentially.

Theorem 4.1. *Suppose that regularity condition (7) holds with some $0 < r \leq \nu_g$. Let $p \in \mathbb{N}$ and $\delta \in (0,1)$. If γ satisfies*

$$\gamma \geq \begin{cases} C_0 \left(\log \frac{6m}{\delta}\right)^{1+r} \max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}^{1+r}, & \text{for } 0 < r \leq \frac{1}{2}, \\ C_0 \left(\log \frac{6m}{\delta}\right)^{1+r} \max\left\{\left(\frac{\lambda_p}{\kappa^2}\right)^{r+\frac{1}{2}}, \frac{1}{\sqrt{m}}\right\} \max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}^{\frac{1}{2}}, & \text{for } r > \frac{1}{2}, \end{cases}$$

where C_0 is the constant given by (24), then with confidence at least $1-\delta$, we have

$$c^i_{\lambda,\gamma,\mathbf{z}} = 0, \quad \forall i \geq p+1$$

and the following error bound holds true: for $0 < r \leq 1$,

$$\begin{aligned} \|f_{\mathbf{z}} - f_\rho\|_\rho &\leq 2b\left(\tfrac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}} + 1\right)\Bigg\{8\|u_\rho\|_K \kappa^{2r+2}\Bigg[\min\left\{\tfrac{\Lambda_{m,p}}{\sqrt{\lambda}}, \Lambda_{m,p}^{1/2}\right\}\lambda_p^r + \tfrac{\lambda_p^r \lambda^{\min\{r-\frac{1}{2},0\}}}{(2\lambda+\lambda_p)^{\min\{r,1/2\}}\sqrt{m}} \\ &\quad + 2^r \lambda_p^{r-1}\Lambda_{m,p}^{1/2}\left(\left(\textstyle\sum_{i=p+1}^\infty \lambda_i^2\right)^{1/2} + \tfrac{8}{\sqrt{m}}\right)\Bigg]\left(\log\tfrac{6}{\delta}\right)^{\frac{3}{2}} \\ &\quad + \sqrt{p}\left(\tfrac{\gamma}{\sqrt{4\lambda+2\lambda_p}} + \tfrac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\tfrac{6m}{\delta}}\right)\Bigg\}\log\tfrac{6}{\delta} + C_r'\left(\tfrac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)^{2r+1}\lambda^{r+\frac{1}{2}}\left(\log\tfrac{6}{\delta}\right)^{2r+1}, \end{aligned} \tag{34}$$

while for $r > 1$,

$$\begin{aligned} \|f_{\mathbf{z}} - f_\rho\|_\rho &\leq 2b\left(\tfrac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}} + 1\right)\Bigg\{8\|u_\rho\|_K \kappa^{2r+2}\Bigg[\min\left\{\tfrac{\Lambda_{m,p}}{\sqrt{\lambda}}, \Lambda_{m,p}^{1/2}\right\}\lambda_p^r + \tfrac{\lambda_p}{\sqrt{2\lambda+\lambda_p}\sqrt{m}} \\ &\quad + 2^r \Lambda_{m,p}^{1/2}\left(\left(\textstyle\sum_{i=p+1}^\infty \lambda_i^{2r}\right)^{1/2} + \tfrac{8}{\sqrt{m}}\right)\Bigg]\left(\log\tfrac{6}{\delta}\right)^{\frac{3}{2}} \\ &\quad + \sqrt{p}\left(\tfrac{\gamma}{\sqrt{4\lambda+2\lambda_p}} + \tfrac{2\sqrt{2}M}{\sqrt{m}}\sqrt{\log\tfrac{6m}{\delta}}\right)\Bigg\}\log\tfrac{6}{\delta} + C_r\left(\tfrac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1\right)\left(4\kappa^2\sqrt{\tfrac{\lambda}{m}} + \lambda^{r+\frac{1}{2}}\right)\left(\log\tfrac{6}{\delta}\right)^2. \end{aligned} \tag{35}$$

Here $\Lambda_{p,m} = \max\left\{\frac{\lambda_p}{\kappa^2}, \frac{1}{\sqrt{m}}\right\}$.

Proof. Under the specified choice of γ, by Proposition 2.6, we know that

$$c^i_{\gamma,\lambda,\mathbf{z}} = 0, \qquad \forall \mathbf{z} \in U_1 \cap U_2, i \geq p+1.$$

Putting the bounds for I_{11} in Proposition 3.1, for I_{12} in Proposition 3.2, and for I_2 in Proposition 2.9 into error decomposition (27) and applying Lemma 2.8, we see that our stated error bounds hold true for $\mathbf{z} \in U_1 \cap U_2 \cap U_3$. But the measure of the set $U_1 \cap U_2 \cap U_3$ is at least $1-\delta$, so our stated conclusion for the sparsity and error bound holds true, and the proof of the theorem is complete. □

Theorem 1.5 stated in the introduction is an immediate consequence of the following more general result stated for the more general case when the eigenvalues $\{\lambda_i\}$ of L_K decay polynomially with different power indices for the upper and lower bounds as

$$D_1 i^{-\alpha_1} \leq \lambda_i \leq D_2 i^{-\alpha_2}, \qquad \forall i \in \mathbb{N}. \tag{36}$$

Theorem 4.2. *Assume regularity condition (7) with some* $0 < r \leq \nu_g$ *and condition (36) with* $1 < \alpha_2 \leq \alpha_1$, $2\alpha_1(r-1) + 3\alpha_2 - 1 > 0$ *and* $D_1, D_2 > 0$. *Let* $0 < \delta < 1$. *If we choose*

$$\gamma = C_0 (D_2/\kappa^2 + 1)^{r+1} \left(\log \frac{6m}{\delta}\right)^{r+1} m^{-\frac{1+r}{\max\{2, 2r+1\}}} \tag{37}$$

and

$$\lambda = \begin{cases} m^{-\frac{1}{2} + \frac{1}{4\alpha_2(1+r)}}, & \text{if } 0 < r \leq 1/2, \\ m^{-\frac{2\alpha_2(r+1)-1}{2\alpha_2(2r+1)(r+1)}}, & \text{if } 1/2 < r \leq 1, \\ m^{-\frac{3\alpha_2 - 1}{4\alpha_2(1+r)}}, & \text{if } r > 1, \end{cases} \tag{38}$$

where C_0 *is the constant given by (24), then with confidence at least* $1 - \delta$ *we have*

$$c^i_{\lambda,\gamma,\mathbf{z}} = 0, \quad \forall m^{\frac{1}{\alpha_2 \max\{2, 2r+1\}}} + 1 \leq i \leq m \tag{39}$$

and

$$\|f_{\mathbf{z}} - f_\rho\|_\rho \leq C^* \left(\log \frac{6m}{\delta}\right)^{\max\{r+2, \frac{5}{2}\}} m^{-\theta},$$

where

$$\theta = \begin{cases} \frac{3}{4} + \frac{2\alpha_1(r-1)-1}{4\alpha_2}, & \text{if } 0 < r \leq \frac{1}{2}, \\ \frac{2\alpha_1(r-1)+3\alpha_2-1}{2\alpha_2(2r+1)}, & \text{if } \frac{1}{2} < r \leq 1, \\ \frac{1}{2} - \frac{1}{2\alpha_2(2r+1)} & \text{if } r > 1, \end{cases} \tag{40}$$

and the constant C^* *is independent of* m *or* δ.

Proof. We take $p = \lceil m^{\frac{1}{\max\{2,2r+1\}\alpha_2}} \rceil$, which implies

$$m^{\frac{1}{\max\{2,2r+1\}\alpha_2}} \leq p \leq 2m^{\frac{1}{\max\{2,2r+1\}\alpha_2}}$$

and by (36) with $\alpha_1 \geq \alpha_2 > 1$,

$$D_1 2^{-\alpha_1} m^{-\frac{\alpha_1}{\alpha_2 \max\{2,2r+1\}}} \leq D_1 p^{-\alpha_1} \leq \lambda_p \leq D_2 p^{-\alpha_2} \leq D_2 m^{-\frac{1}{\max\{2,2r+1\}}}. \tag{41}$$

It follows that the choice (37) of γ ensures condition (26) in Theorem 4.1. Then we can apply Theorem 4.1 and know that with confidence at least $1 - \delta$, (39) and the error bounds stated in Theorem 4.1 hold true. What is left is to derive learning rates from these error bounds by simplifying quantities included in the error bounds.

Note that the sum $\sum_{i=p+1}^{\infty} \lambda_i^2$ in (34) for the case $r \leq 1$ and the sum $\sum_{i=p+1}^{\infty} \lambda_i^{2r}$ in (35) for the case $r > 1$ can be unified as $\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2,2r\}}$ which can be estimated as

$$\begin{aligned}\sum_{i=p+1}^{\infty} \lambda_i^{\max\{2,2r\}} \quad \leq \quad & \sum_{i=p+1}^{\infty} D_2^{\max\{2,2r\}} i^{-\alpha_2 \max\{2,2r\}} \\ & \leq D_2^{\max\{2,2r\}} \int_p^{\infty} x^{-\alpha_2 \max\{2,2r\}} dx = \frac{D_2^{\max\{2,2r\}} p^{1-\alpha_2 \max\{2,2r\}}}{\alpha_2 \max\{2,2r\} - 1} \\ & \leq \frac{D_2^{\max\{2,2r\}}}{\alpha_2 \max\{2,2r\} - 1} m^{\frac{1-\alpha_2 \max\{2,2r\}}{\alpha_2 \max\{2,2r+1\}}}.\end{aligned}$$

Concerning the quantity $\mathcal{B}_{m,\lambda} = \frac{2\kappa}{\sqrt{m}} \left\{ \frac{\kappa}{\sqrt{m\lambda}} + \sqrt{\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell}} \right\}$, we apply the polynomial decay (36) of $\{\lambda_i\}_{i=1}^{\infty}$ to get

$$\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell} \leq \sum_{\ell=1}^{\infty} \frac{D_2}{D_2 + \lambda \ell^{\alpha_2}} \leq \lambda^{-1/\alpha_2} \int_0^{\infty} \frac{D_2}{D_2 + x^{\alpha_2}} dx \leq D_2 \frac{\alpha_2}{\alpha_2 - 1} \lambda^{-1/\alpha_2},$$

which implies

$$\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}} = \frac{2\kappa}{\sqrt{m\lambda}} \left(\frac{\kappa}{\sqrt{m\lambda}} + \sqrt{\sum_{i\in\mathbb{N}} \frac{\lambda_i}{\lambda + \lambda_i}} \right) \leq 2\kappa m^{-\frac{1}{2}} \lambda^{-\frac{1}{2} - \frac{1}{2\alpha_2}} \left(\kappa + \sqrt{\frac{D_2 \alpha_2}{\alpha_2 - 1}} \right).$$

The choice (38) of λ guarantees that

$$\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}} + 1 \leq C_{\kappa,\alpha_2,D_2},$$

where C_{κ,α_2,D_2} is a constant independent of δ or m.

Finally we estimate the quantities $\Lambda_{m,p}$ and $\min\left\{ \frac{\Lambda_{m,p}}{\sqrt{\lambda}}, \Lambda_{m,p}^{\frac{1}{2}} \right\}$ by (41) and (38) and find that the stated learning rates hold true with the power index θ given by (40). The proof of the theorem is complete. □

We are in a position to prove our main result stated in the introduction for the case when the eigenvalues $\{\lambda_i\}_i$ decay exponentially.

Proof of Theorem 1.7. It can be easily seen from the exponential decay (13) of $\{\lambda_i\}_{i=1}^{\infty}$ that

$$\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell} \leq \sum_{\ell=1}^{\infty} \frac{D_2}{D_2 + \lambda \beta^\ell} \leq \int_0^{\infty} \frac{D_2}{D_2 + \lambda \beta^x} dx.$$

By setting $\lambda \beta^x = t$, we have $x = \log_\beta(\frac{t}{\lambda})$, and $dx = \frac{1}{\log \beta} \frac{1}{t} dt$. It follows that

$$\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell} \leq \frac{1}{\log \beta} \int_\lambda^{\infty} \frac{D_2}{(t + D_2)t} dt = \frac{1}{\log \beta} \int_\lambda^{\infty} \left(\frac{1}{t} - \frac{1}{t + D_2} \right) dt = \frac{1}{\log \beta} \log \left(1 + \frac{D_2}{\lambda} \right).$$

Take $\lambda = D_2 \frac{\log m}{m}$. We have

$$\sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell} \leq \frac{1}{\log \beta} \log \left(1 + \frac{m}{\log m} \right) \leq \frac{1}{\log \beta} \log m$$

and

$$\frac{\mathcal{B}_{m,\lambda}}{\sqrt{\lambda}}+1=\frac{2\kappa}{\sqrt{m\lambda}}\left(\frac{\kappa}{\sqrt{m\lambda}}+\sqrt{\sum_{i\in\mathbb{N}}\frac{\lambda_i}{\lambda+\lambda_i}}\right)+1\leq\frac{2\kappa}{\sqrt{D_2}}\left(\frac{\kappa}{\sqrt{D_2}}+\frac{1}{\log\beta}\right)+1. \tag{42}$$

We take $p=\lceil\frac{\log(m+1)}{\max\{2,2r+1\}\log\beta}\rceil$ to give

$$\frac{\log(m+1)}{\max\{2,2r+1\}\log\beta}\leq p\leq 1+\frac{\log(m+1)}{\max\{2,2r+1\}\log\beta}. \tag{43}$$

Moreover, by the eigenvalue exponential decay (13), we have

$$\frac{D_1}{\beta}(2m)^{-\frac{1}{\max\{2,2r+1\}}}\leq\lambda_p\leq D_2 m^{-\frac{1}{\max\{2,2r+1\}}}. \tag{44}$$

It follows that

$$\begin{aligned}\sum_{i=p+1}^{\infty}\lambda_i^{\max\{2r,2\}}&\leq D_2^{\max\{2r,2\}}\sum_{i=p+1}^{\infty}\beta^{-i\max\{2r,2\}}=D_2^{\max\{2r,2\}}\frac{\beta^{-(p+1)\max\{2r,2\}}}{1-\beta^{-\max\{2r,2\}}}\\&=D_2^{\max\{2r,2\}}\frac{\beta^{-p\max\{2r,2\}}}{\beta^{\max\{2r,2\}}-1}\leq\frac{D_2^{\max\{2r,2\}}}{\beta^{\max\{2r,2\}}-1}m^{-\frac{\max\{2r,2\}}{\max\{2,2r+1\}}}.\end{aligned} \tag{45}$$

Finally we estimate the quantities $\Lambda_{m,p}$ and $\min\left\{\frac{\Lambda_{m,p}}{\sqrt{\lambda}},\Lambda_{m,p}^{\frac{1}{2}}\right\}$ by (44) and the choice $\lambda=D_2\frac{\log m}{m}$ as

$$\begin{aligned}\Lambda_{m,p}=\max\left\{\frac{\lambda_p}{\kappa^2},\frac{1}{\sqrt{m}}\right\}&\leq(D_2/\kappa^2+1)m^{-\frac{1}{\max\{2,2r+1\}}},\\\min\left\{\frac{\Lambda_{m,p}}{\sqrt{\lambda}},\Lambda_{m,p}^{\frac{1}{2}}\right\}&\leq(D_2/\kappa^2+1)m^{-\frac{1}{2\max\{2,2r+1\}}}.\end{aligned}$$

Combining the above estimates and Theorem 4.1, we find that the stated learning rates hold true. The proof of the theorem is complete. □

Acknowledgement

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104113 and ECS No. PolyU 25301115] and by National Natural Science Foundation of China [Grant No. 11471292, No. 11401524, No. 11531013, and No. 11461161006]. The corresponding author is Dao-Hong Xiang who is also supported by the Alexander von Humboldt Foundation of Germany. This paper was written when the last author visited Shanghai Jiaotong University (SJTU). The hospitality and sponsorships from SJTU and Ministry of Education are greatly appreciated.

References

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

[2] F. Bauer, S. Pereverzev, and L. Rosasco, On regularization algorithms in learning theory, J. Complexity, **23** (2007): 52–72.

[3] A. Caponnetto and E. DeVito, Optimal rates for the regularized least squares algorithm, Found. Comp. Math., **7** (2007): 331–368.

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

[5] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, 2007.

[6] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math., **5** (2005), 59–85.

[7] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone, Learning from examples as an inverse problem, J. Mach. Learn. Res., **6** (2005): 883–904.

[8] L. H. Dicker, D. P. Foster, and D. Hsu, Kernel ridge vs. principal component regression: minimax bounds and adaptability of regularization operators, preprint, 2016. http://www.stat.rutgers.edu /home/ldicker/papers/kernels.pdf.

[9] H. W. Engl, M. Hanke, and A. Neubauer, Regularization of Inverse Problems. Vol. **375** Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, 1996.

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

[11] X. Guo and D. X. Zhou, An empirical feature-based learning algorithm producing sparse approximations, Appl. Comput. Harmonic Anal., **32** (2012): 389–400.

[12] X. Guo, J. Fan, and D. X. Zhou, Sparsity and error analysis of empirical feature-based regularization schemes, J. Mach. Learn. Res., **17** (2016): 1–34.

[13] Z. C. Guo and D. X. Zhou, Concentration estimates for learning with unbounded sampling, Adv. Comput. Math., **38** (2013): 207–223.

[14] Z. C. Guo, S. Lin, and D. X. Zhou. Learning theory of distributed spectral algorithms, Inverse Problems, minor revision under review, 2016.

[15] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl. **13** (2015), 437–455.

[16] S. Lin, X. Guo, and D. X. Zhou, Distributed learning with regularized least squares, submitted, 2015.

[17] S. Mendelson and J. Neeman, Regularization in kernel learning, Ann. Stat., **38**(2010), 526–565.

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

[19] S. Smale and D. X. Zhou, Online learning with markov sampling, Anal. Appl., **7** (2009), 87–113.

[20] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodol.), **58** (1996): 267–288.

[21] Q. Wu, Y. Ying, and D. X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math., **6** (2006), 171–192.

[22] Q. Wu and D. X. Zhou, Learning with sample dependent hypothesis spaces, Computers and Mathematics with Applications **56** (2008), 2896-2907.

[23] Y. Yao, L. Rosasco, and A. Caponnetto, On early stopping in gradient descent learning, Constr. Approx., **26** (2007): 289–315.

[24] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comput., **15** (2003), 1397–1437.

[25] T. Zhang, Some sharp performance bounds for least square regression with L1 regularization, Ann. Statist., **37** (2009): 2109–2114.

[26] P. Zhao and B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res., **7** (2006): 2541–2563.

[27] 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.