Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2012, Article ID 915920, 18 pages
doi:10.1155/2012/915920

Research Article

Error Bounds for l^p-Norm Multiple Kernel Learning with Least Square Loss

Shao-Gao Lv[1] and Jin-De Zhu[2]

[1] *Statistics School, Southwestern University of Finance and Economics, Chengdu 611130, China*
[2] *The 2nd Geological Party of Bureau of Geology and Mineral Resources, Henan, Jiaozuo 450000, China*

Correspondence should be addressed to Shao-Gao Lv, kenan716@mail.ustc.edu.cn

Received 24 April 2012; Accepted 10 July 2012

Academic Editor: Gerd Teschke

The problem of learning the kernel function with linear combinations of multiple kernels has attracted considerable attention recently in machine learning. Specially, by imposing an l^p-norm penalty on the kernel combination coefficient, multiple kernel learning (MKL) was proved useful and effective for theoretical analysis and practical applications (Kloft et al., 2009, 2011). In this paper, we present a theoretical analysis on the approximation error and learning ability of the l^p-norm MKL. Our analysis shows explicit learning rates for l^p-norm MKL and demonstrates some notable advantages compared with traditional kernel-based learning algorithms where the kernel is fixed.

1. Introduction

1.1. Overview of Multiple Kernel Learning

Kernel methods such as Support Vector Machines (SVMs) have been extensively applied to supervised learning tasks such as classification and regression. The performance of a kernel machine largely depends on the data representation via the choice of kernel function. Hence, one central issue in kernel methods is the problem of kernel selection; a great many approaches to selecting the right kernel have been studied in the literature [1–4] and other references therein.

We begin with reviewing the classical supervised learning setup. Let (X, d) be a compact metric space and $Y \subseteq \mathbb{R}$, given a labeled sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{m} \subseteq Z := X \times Y$, sampled *i.i.d.* according to an unknown distribution ρ supported on Z, the goal is to estimate a real-valued function $f_{\mathbf{z}}$ depending on the sample, that generalizes well on new and unseen data. A widely used approach to estimate a function from empirical data consists in minimizing a regularization functional in a Hilbert space $\mathscr{H}$ of real-valued functions: $X \to \mathbb{R}$. Typically, a regularization scheme estimates f as a minimizer of the functional

$$\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega(f), \tag{1.1}$$

where $\mathcal{E}_{\mathbf{z}}(f) = (1/m)\sum_{i=1}^{m} V(f(x_i), y_i)$ is the empirical risk of hypothesis f, measured by a nonnegative loss function $V : \mathbb{R} \times Y \to \mathbb{R}^+$. In addition, $\Omega : \mathscr{H} \to \mathbb{R}$ is a regularizer and $\lambda > 0$ is a trade-off regularization parameter.

In this paper, we assume that $\mathscr{H}$ is a reproducing kernel Hilbert space (RKHS) $\mathscr{H}_K$ with kernel K, see [5]. Every kernel K corresponds to a feature mapping $\Psi_K : X \to \mathscr{H}_K$ satisfying $K(x, y) = \langle \Psi_K(x), \Psi_K(y)\rangle_K$, and each element of $\mathscr{H}_K$ has the following form:

$$f_w(x) = \langle w, \Psi_K(x)\rangle_K, \quad \forall w \in \mathscr{H}_K. \tag{1.2}$$

By restricting the regularization to be the form $\Omega(f) = \|f\|_K^2$, there is a lot of studies from different perspectives such as statistics, optimal recovery and machine learning [6–9], and other references therein. Regularization in an RKHS has a number of attractive features, including the availability of effective error bounds and stability analysis relative to perturbations of the data (see Cucker and Smale [7]; Wu et al. [10]; Bousquet and Elisseeff [6]). Moreover, the optimization problem (1.1) in an RKHS can be reduced to seek for solution in a finite-dimensional space. Although it is simple to prove, this result shows that the variational problem (1.1) can be computational easily.

Because of their simplicity and generality, kernels and associated RKHS play an increasingly important role in Machine Learning, Pattern Recognition and Artificial Intelligence. When the kernel is fixed, an immediate concern is the choice of the regularization parameter λ. This is typically solved by means of cross validation or generalized cross validation [11]. However, the performance of kernel methods critically relies on the choice of the kernel function. A natural question is how to choose the optimal kernel in a collection of candidate kernels.

Kernel learning can range from the width parameter selection of Gaussian kernels [9, 12] to obtaining an optimal linear combination from a set of finite candidate kernels. The latter is often referred to as multiple kernel learning in machine learning and nonparametric group Lasso in statistics [13]. Lanckriet et al. [3] pioneered work on MKL and proposed a semidefinite programming approach to automatically learn a linear combination of candidate kernels for the cases of SVMS. To improve computation efficiency, the multikernel class further is restricted to only convex combinations of kernels [2, 14, 15]. Most learning kernel algorithms are based on considering linear kernel mixtures $K_\theta = \sum \theta_k K_k, (\theta_k \geq 0)$ with a prescribed kernels $K_1, \ldots, K_M$. For notational simplicity, we will frequently use Ψ_k instead of the standard feature Ψ_{K_k}. Compared to (1.1), the primal model for learning with multiple kernels is extended to

$$f_{\tilde{w},\theta}(x) = \sum_{k=1}^{M} \sqrt{\theta_k} \langle w_k, \Psi_k(x)\rangle_{K_k}, \quad \forall \tilde{w} = (w_1, \ldots, w_M), \text{ where } w_k \in \mathscr{H}_{K_k}. \tag{1.3}$$

In this paper, we mainly focus on the l^p-norm MKL, consisting in minimizing the regularized empirical risk with respect to the optimal kernel mixture $\sum_{k=1}^{M}\theta_k K_k$, in addition to l^p-regularizer on θ to avoid overfitting. This leads to the following optimization problem:

$$\inf_{\tilde{w},\theta:\theta\geq 0}\frac{1}{m}\sum_{i=1}^{m}V\left(\sum_{k=1}^{M}\sqrt{\theta_k}\langle w_k,\Psi_k(x_i)\rangle_{K_k},y_i\right)+\lambda\sum_{k=1}^{M}\|w_k\|_{K_k}^{2}+\mu\sum_{k=1}^{M}|\theta_k|^{p}. \tag{1.4}$$

This scheme was introduced in [2] and the existence of its minimum has been discussed in [4].

The optimization problem subsumes state-of-the-art approaches to multiple kernel learning, covering sparse and nonsparse MKL by arbitrary l^p-norm regularization ($1 \leq p \leq \infty$) on the mixing coefficients as well as the incorporation of prior knowledge by allowing for nonisotropic regularizer. Kloft et al. [2] developed two efficient interleaved optimization strategies for the l^p-norm multiple kernel learning, and this interleaved optimization is much faster than the commonly used wrapper approaches, as demonstrated on real-world problems from computational biology. An analysis of this model, based on Rademacher complexities, was first developed by Cortes et al. [1]. Later improved rates of convergence were derived based on the theory of local Rademacher complexities [15]. However, the estimate on local Rademacher complexities with $1 \leq p < 2$ strictly depends on no-correlation assumption of the M different features, which is too strong condition in theory and practice. In this paper, we employ the notion of empirical covering number to present a theoretical analysis of its generalization error. Besides no-correlation condition is not necessary, empirical covering number is one tight upper bound of local Rademacher complexities [16], also independent of the underlying distribution. We will see that some satisfying learning rates are established when the regularization parameter is appropriately chosen. The interaction between the sample error and the approximation error plays an important role in our analysis, and our new methodology mainly depends on the complexity of hypothesis class measured by empirical covering number and the regularity of a target function.

It should be pointed out that the Tikhonov Regularization in (1.4) has two regularization parameter (λ, μ), which may be hard to deal with in practice. Fortunately, an alternative approach has been studied by Rakotomamonjy et al. [14] and Kloft et al. [2]. More precisely, this approach employs the regularizer $\|\theta\|_{l^p} \leq 1$ as an additional constraint into the optimization problem. By substituting w_k for $\sqrt{\theta}w_k$, they arrive at the following problem:

$$\begin{gathered}\inf_{\tilde{w},\theta:\theta\geq 0}\frac{1}{m}\sum_{i=1}^{m}V\left(\sum_{k=1}^{M}\langle w_k,\Psi_k(x_i)\rangle_{K_k},y_i\right)+\frac{\lambda}{2}\sum_{k=1}^{M}\frac{\|w_k\|_{K_k}^{2}}{\theta_k}\\ \text{subject to } \|\theta\|_{l^p}\leq 1.\end{gathered} \tag{1.5}$$

1.2. Algorithm and Main Consequence

The following Lemma (see [4]) indicates that the above multikernel class can equivalently be represented as a block-norm regularized linear class in the product Hilbert space $\mathscr{H} = \mathscr{H}_{K_1}\times\cdots\times\mathscr{H}_{K_M}$.

Lemma 1.1. *If $p > 0$, $q = 1 + (1/p)$, and $\{a_j, j \in \mathbb{N}_n\} \subseteq \mathbb{R}$, then*

$$\min\left\{\left(\sum_{j\in\mathbb{N}_n}\frac{a_j^2}{\lambda_j}\right)^{1/2} : \lambda_l \geq 0, l \in \mathbb{N}_n, \sum_{j\in\mathbb{N}_n}\lambda_j^p \leq 1\right\} = \left(\sum_{j\in\mathbb{N}_n}\left|a_j\right|^{2/q}\right)^{q/2}, \tag{1.6}$$

and the equality occurs for $\sum_{j\in\mathbb{N}_n}|a_j| > 0$ at

$$\widetilde{\lambda}_j := \frac{\left|a_j\right|^{2/(p+1)}}{\left(\sum_{j\in\mathbb{N}_n}\left|a_j\right|^{2p/(p+1)}\right)^{1/p}}. \tag{1.7}$$

Hence, Lemma 1.1 can be applied to define the feature mapping: $\Psi : x \in X \to (\Psi_1(x), \ldots, \Psi_M(x)) \in \mathcal{H}_{\widetilde{K}}$ associated with a kernel $\widetilde{K}$; the class of functions defined above coincides with

$$H_{p,D,M} = \left\{f_w : x \longrightarrow \langle w, \Psi(x)\rangle_{\mathcal{H}_{\widetilde{K}}} \mid w = (w_1, \ldots, w_M), \|w\|_{2,q} \leq D\right\}, \tag{1.8}$$

when $p \in [1, \infty]$, $q \in [1, 2]$ holds from $q = 2p/(p+1)$. The $l_{2,q}$-norm is defined here as $\|w\|_{2,q} = (\sum_{j=1}^M \|w_j\|_{K_j}^q)^{1/q}$. For simplicity, we write $\|f_w\|_{2,q} = \|w\|_{2,q}$. Clearly learning the complexity of (1.8) will be greater than one that is based on a single kernel only, further it provides greater learning ability while the computational complexity increases accordingly. The sample complexity of the above hypothesis space has been studied by Cortes et al. [1] and Kloft and Blanchard [15]. Thus the primal MKL optimization problem (1.5) is equivalent to the following regularization scheme, which is the primary object of investigation in this paper

$$f_{\mathbf{z}} = \langle w_{\mathbf{z}}, \Psi(x)\rangle_{\mathcal{H}_{\widetilde{K}}}, \quad \text{where } w_{\mathbf{z}} = \arg\min_{w\in\mathcal{H}_{\widetilde{K}}}\frac{1}{m}\sum_{i=1}^m V\left(\langle w, \Psi(x_i)\rangle_{\mathcal{H}_{\widetilde{K}}}, y_i\right) + \lambda\|w\|_{2,q}^2. \tag{1.9}$$

Here we use the symbol "min" instead of "inf," since (1.4) is equivalent to (1.9) and the solution of (1.4) exists and is unique. Remark that the above algorithm is a standard regularized empirical risk minimization; this implies that l^p-norm multiple kernel learning scheme can be free of over-fitting, a phenomenon which occurs when the empirical error is zero but the expected error in far from zero.

In the following, we assume that $\{K_j\}_{j=1,\ldots,M}$ is uniformly bounded, that is,

$$\kappa = \sup_{j\in\{1,\ldots,M\}}\sup_{x\in X}\sqrt{K_j(x,x)} < \infty. \tag{1.10}$$

Also suppose that each K_j is continuous. In other words, each K_j is a Mercer kernel with bound κ; we refer to [17] for more properties and discussions on Mercer kernel.

In this paper, we only focus on the least square loss: $V(f(x),y) = (f(x)-y)^2$. Accordingly, the target function is given by

$$f_\rho(x) = \arg\min\left\{\mathbb{E}\left(f(x)-y\right)^2\right\} = \int_Y y\,d\rho(\cdot \mid x), \quad \forall x \in X, \tag{1.11}$$

where we denote by $\rho(\cdot \mid x)$ the conditional distribution of ρ. Through this paper we assume that $\rho(\cdot \mid x)$ is supported on $[-T,T]$, it follows that $|f_\rho(x)| \le T$ for $x \in X$ almost surely. Since the learner $f_{\mathbf{z}}$ may be much larger than f_ρ, it is natural to apply a projection operator on $f_{\mathbf{z}}$, which was introduced into learning algorithms to improve learning rates.

Definition 1.2. The projection operator π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) = \begin{cases} T, & \text{if } f(x) > T, \\ f(x), & \text{if } \left|f(x)\right| \le T, \\ -T, & \text{if } f(x) < -T, \end{cases} \tag{1.12}$$

where $T > 0$ is called the projection level.

The target of error analysis is to understand how $\pi(f_{\mathbf{z}})$ approximates the regression function f_ρ. More precisely, we aim to estimate the *excess generalization error*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \tag{1.13}$$

for the l^p-norm MKL algorithm (1.4), where $\mathcal{E}(f) = \mathbb{E}(f(x)-y)^2$ denotes the expect error of f.

To show some ideas of our error analysis, we first state learning rates of (1.4) in a special case when $f_\rho \in \mathcal{H}_{\widetilde{K}}$ and $\widetilde{K}$ is C^∞ on $X \subset \mathbb{R}^n$.

Theorem 1.3. *Let $f_{\mathbf{z}}$ be defined by* (1.9). *Assume $\widetilde{K}$ is C^∞ on $X \subset \mathbb{R}^n$ and $f_\rho \in \mathcal{H}_{\widetilde{K}}$. For any $0 < \delta < 1$ and $\epsilon > 0$, with confidence $1-\delta$, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) = \left\|\pi(f_{\mathbf{z}}) - f_\rho\right\|_\rho^2 \le \widetilde{C}\log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{1-\epsilon}, \quad \text{with } \lambda = \left(\frac{1}{m}\right)^{1-2\epsilon}, \tag{1.14}$$

where $\widetilde{C}$ is some constant independent of m or δ.

Theorem 1.3 can be viewed as a corollary of our main result presented in Section 5. It can be arbitrary close to $\mathcal{O}(m^{-1})$ by choosing ϵ to be small enough, which is the best convergence rate in learning theory literature.

2. Key Error Analysis

Our main result is about learning rates of (1.4) stated under conditions on the approximation ability of $\mathcal{H}_{\widetilde{K}}$ with respect to f_ρ and capacity of $\mathcal{H}_{\widetilde{K}}$.

The approximation ability of the hypothesis space $\mathscr{H}_{\tilde{K}}$ with respect to f_ρ in the space $L^2_{\rho_X}$ is reflected by regularization error.

Definition 2.1. The regularization error of the triple $(\mathscr{H}_{\tilde{K}}, f_\rho, \rho_X)$ is defined as

$$\mathscr{A}_q(\lambda) = \min_{f \in \mathscr{H}_{\tilde{K}}} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|^2_{2,q} \right\}, \quad \lambda > 0. \tag{2.1}$$

We will assume that for some $0 < \beta \le 1$ and $C_\beta > 0$,

$$\mathscr{A}_q(\lambda) \le C_\beta \lambda^\beta. \tag{2.2}$$

Remark 2.2. Our assumption implies that when f_ρ is replaced by $\mathscr{H}_{\tilde{K}}$, $\mathscr{A}_q(\lambda)$ tends to zero by polynomial order decay as λ goes to zero. Note [7] that $\mathscr{A}_q(\lambda) = o(\lambda)$ would imply $f_\rho = 0$. So $\beta = 1$ in (2.2) is the best we can expect. This case is equivalent to $f_\rho \in \mathscr{H}_{\tilde{K}}$ when $\mathscr{H}_{\tilde{K}}$ is dense in $L^2_{\rho_X}$, see [18]. Assumption (2.2) with $0 < \beta < 1$ can be characterized in terms of interpolation spaces [7].

If ρ_X is the Lebesgue measure on X and the target function $f_\rho \in H^s$, a Sobolev space with power s. When Gaussian kernel $(G_\sigma(x, y) = \exp(-\sigma\|x - y\|^2))$ is taken with a fixed variance σ, a polynomial decay of $\mathscr{A}_q(\lambda)$ is impossible. However, Example 1 of [19] successfully obtains a polynomial decay under the multikernel setting, allowing for varying variances of Gaussian kernels. This shows that multikernel learning can improve the approximation power and learning ability. More interestingly, we will take a special example to show the impact of the multikernel class on the regularization error in Section 5 below. In particular, a proper multikernel class can be applied to improve the regularization error if the regularity of f_ρ is rather high.

Next we define the truncated sample error as

$$\mathcal{S}_{\mathbf{z}}(\lambda, f, T) = \left\{ \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_z(\pi(f_{\mathbf{z}})) \right\} + \left\{ \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \right\}, \tag{2.3}$$

and the sample error as

$$\mathcal{S}_{\mathbf{z}}(\lambda, f) = \left\{ \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) \right\} + \left\{ \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \right\}. \tag{2.4}$$

The function f in the above equation can be arbitrarily chosen; however, only proper choices lead to good estimates of the regularization error. A good choice is $f = f_\lambda$ where

$$f_\lambda = \arg \min_{f \in \mathscr{H}_{\tilde{K}}} \left\{ \mathcal{E}(f) + \lambda \|f\|^2_{2,q} \right\}. \tag{2.5}$$

A useful approach for regularization schemes with sample independent hypothesis spaces such as RKHS is an error decomposition, which decomposes the total error $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)$ into the sum of the truncated sample error and the regularization error stated as follows.

Proposition 2.3. *Let f_λ be defined by* (2.5)*; there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \mathcal{S}_{\mathbf{z}}(\lambda, f_\lambda, T) + \mathcal{A}_q(\lambda). \tag{2.6}$$

Proof. We can decompose $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)$ into

$$\begin{aligned}&\{\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))\} + \left\{\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \lambda\|f_{\mathbf{z}}\|_{2,q}^2 - \left(\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_{2,q}^2\right)\right\}\\ &\quad + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} + \left\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\|_{2,q}^2\right\} - \lambda\|f_{\mathbf{z}}\|_{2,q}^2.\end{aligned} \tag{2.7}$$

To bound the second term, by Definition of $f_{\mathbf{z}}$, $\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \lambda\|f_{\mathbf{z}}\|_{2,q}^2$ can be bounded by

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{2,q}^2 \leq \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_{2,q}^2, \tag{2.8}$$

since $|\pi(f)(x) - y| \leq |f(x) - y|$ holds for any function f on Z. The conclusion follows by combining these two inequalities. □

3. Estimation on Sample Error

We are in a position to estimate the sample error $\mathcal{S}_{\mathbf{z}}(\lambda, f_\lambda, T)$. The sample error $\mathcal{S}_{\mathbf{z}}(\lambda, f_\lambda, T)$ can be written as

$$\begin{aligned}&\{(\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)) - (\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(f_\rho))\}\\ &\quad + \{(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)) - (\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho))\} := \mathcal{S}_1 + \mathcal{S}_2.\end{aligned} \tag{3.1}$$

$\mathcal{S}_2$ can be easily bounded by applying the following one-side Bernstein-type probability inequality.

Lemma 3.1. *Let ξ be a random variable on a probability space Z with variance σ^2 satisfying $|\xi - \mathbb{E}(\xi)| \leq M_\xi$ for some constant M_ξ. Then for any $0 < \delta < 1$, we have*

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}(\xi) \leq \frac{2M_\xi \log(1/\delta)}{3m} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{m}}. \tag{3.2}$$

Proposition 3.2. *Define the random variable $\xi_2(z) = (f_\lambda(x)-y)^2-(f_\rho(x)-y)^2$. For every $0 < \delta < 1$, with confidence at least $1 - \delta/2$, there holds*

$$\mathcal{S}_2 = \frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i) - \mathbb{E}(\xi_2) \leq \mathcal{A}_q(\lambda)\left(1 + \frac{3\kappa^2 M^{2/q^*}\log(2/\delta)}{m\lambda}\right) + \frac{6T^2 + \log(2/\delta)}{m}. \tag{3.3}$$

Proof. From the definition of $\mathcal{A}_q(\lambda)$, we see that

$$\lambda\|f_\lambda\|_{2,q}^2 \le \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\|_{2,q}^2 = \mathcal{A}_q(\lambda). \tag{3.4}$$

Note that $f_\lambda(x) = \langle w_\lambda, \Psi(x)\rangle$ for some $w_\lambda = (w_\lambda^{(1)}, \dots, w_\lambda^{(M)}) \in \mathcal{H}_{\widetilde{K}}$, by the Cauchy-Schwarz inequality $(C. - S.)$, we have for any $x \in X$:

$$\begin{aligned}\langle w_\lambda, \Psi(x)\rangle &= \sum_{j=1}^{M}\left\langle w_\lambda^{(j)}, \Psi_j(x)\right\rangle \overset{C.-S.}{\le} \sum_{j=1}^{M}\left\|w_\lambda^{(j)}\right\|_{K_j}\|\Psi_j(x)\|_{K_j} \\ &\overset{\text{Hölder}}{\le} \|w_\lambda\|_{2,q}\left(\sum_{j=1}^{M}\|\Psi_j(x)\|_{K_j}^{q^*}\right)^{1/q^*},\end{aligned} \tag{3.5}$$

where $(1/q) + (1/q^*) = 1$. Using Assumption (1.10), it follows that

$$\|f_\lambda\|_\infty \le \kappa M^{1/q^*}\|w_\lambda\|_{2,q} \le \kappa M^{1/q^*}\sqrt{\frac{\mathcal{A}_q(\lambda)}{\lambda}}. \tag{3.6}$$

Observe that

$$\xi_2(z) = \left(f_\lambda(x) - f_\rho(x)\right)\left(f_\lambda(x) + f_\rho(x) - 2y\right). \tag{3.7}$$

Since that $|f_\rho(x)| \le T$ almost surely, we have

$$|\xi_2| \le \left(\|f_\lambda\|_\infty + T\right)\left(\|f_\lambda\|_\infty + 3T\right) \le c := \left(\kappa M^{1/q^*}\sqrt{\frac{\mathcal{A}_q(\lambda)}{\lambda}} + 3T\right)^2. \tag{3.8}$$

Hence $|\xi_2 - \mathbb{E}(\xi_2)| \le M_{\xi_2} := 2c$. Moreover, $\mathbb{E}(\xi_2)^2$ equals

$$\mathbb{E}\left(\left(f_\lambda(x) - f_\rho(x)\right)^2\left(f_\lambda(x) + f_\rho(x) - 2y\right)^2\right) \le \left(\|f_\lambda\|_\infty + 3T\right)^2\|f_\lambda - f_\rho\|_\rho^2, \tag{3.9}$$

which implies that $\sigma^2(\xi_2) \le \mathbb{E}(\xi_2^2) \le c\mathcal{A}_q(\lambda)$. The desired result follows from Lemma 3.1. □

Next we estimate the first term $\mathcal{S}_1$. It is more difficult to deal with because it involves a set of random variables $f_{\mathbf{z}}$ varying with $\mathbf{z}$, requiring to consider the functional complexity. For this purpose, we introduce the notion of empirical covering numbers, which often lead to sharp error estimates [16].

Definition 3.3. Let $(\mathcal{U}, d)$ be a pseudometric space and $S \subset \mathcal{U}$. For every $\epsilon > 0$, the covering number $\mathcal{N}(S, \epsilon, d)$ of S with respect to ϵ and d is defined as the minimal number of balls of radius ϵ whose union covers S, that is,

$$\mathcal{N}(S, \epsilon, d) = \min\left\{ l \in \mathbb{N} : S \subset \bigcup_{j=1}^{l} B(s_j, \epsilon) \ \text{for some}\ \{s_j\}_{j=1}^{l} \subset \mathcal{U} \right\}, \tag{3.10}$$

where $B(s_j, \epsilon) = \{s \in \mathcal{U} : d(s, s_j) \leq \epsilon\}$ is a ball in $\mathcal{U}$.

The l^2-empirical covering number of a function set is defined by means of the normalized l^2-metric d_2 on the Euclidian space $\mathbb{R}^k$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{k} \sum_{i=1}^{k} |a_i - b_i|^2 \right)^{1/2} \quad \text{for } \mathbf{a} = (a_i)_{i=1}^{k},\ \mathbf{b} = (b_i)_{i=1}^{k} \in \mathbb{R}^k. \tag{3.11}$$

Definition 3.4. Let $\mathcal{F}$ be a set of function on X, $\mathbf{x} = (x_i)_{i=1}^{k} \subset X^k$ and $\mathcal{F}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^{k} : f \in \mathcal{F}\} \subset \mathbb{R}^k$. Set $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon) = \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \epsilon, d_2)$. The l^2-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{k \in \mathbb{N}} \sup_{\mathbf{x} \in X^k} \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon), \quad \epsilon > 0. \tag{3.12}$$

Denote by B_R the ball of radius R with $R > 0$, $B_R = \{f \in \mathcal{H}_{\tilde{K}} : \ \|f\|_{\tilde{K}} \leq R\}$. We need the following capacity assumption on $\mathcal{H}_{\tilde{K}}$.

Assumption 3.5. There exists an exponent v, with $0 < v < 2$ and a constant $C_{v,\tilde{K}} > 0$ such that

$$\log \mathcal{N}_2(B_1, \epsilon) \leq C_{v,\tilde{K}} \epsilon^{-v}, \quad \forall \epsilon > 0, \tag{3.13}$$

where B_1 is the unit ball of $\mathcal{H}_{\tilde{K}}$ defined as above.

For any function $f_w = \langle w, \Psi(x) \rangle_{\mathcal{H}_{\tilde{K}}}$, by the hölder inequality, we have

$$\|f_w\|_{\tilde{K}} = \left(\sum_{j=1}^{M} \|w_j\|_{K_j}^2 \right)^{1/2} \leq M^{1/q^*} \|f_w\|_{2,q}, \tag{3.14}$$

and it follows from (3.13)

$$\log \mathcal{N}_2\left(B_1^q, \epsilon\right) \leq C_{v,K} M^{v/q^*} \epsilon^{-v}, \tag{3.15}$$

where B_1^q is called the generalized unit ball of $\mathcal{H}_{\tilde{K}}$ associated with q, defined by

$$B_1^q = \left\{ f \in \mathcal{H}_{\tilde{K}}, \|f\|_{2,q} \leq 1 \right\}. \tag{3.16}$$

Note that for any function set $\mathcal{F} \subseteq C(X)$, the empirical covering number $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon)$ is bounded by $\mathcal{N}(\mathcal{F}, \epsilon)$, the (uniform) covering number of $\mathcal{F}$ under the metric $\|\cdot\|_\infty$, since $d_2(f, g) \le \|f - g\|_\infty$. It was shown in [20] that the quantity $\log(\mathcal{N}(B_1, \epsilon)) \le C_0(1/\epsilon)^s$ holds for some $C_0 > 0$ if K is $C^{2n/s}$ on a subset of $\mathbb{R}^n$, hence $\log(\mathcal{N}_2(B_1, \epsilon)) \le C_0(1/\epsilon)^s$ also holds. In particular, s is arbitrarily small for a C^∞ kernel (such as Gaussian kernel). Now we give a concrete example in $\mathbb{R}^n$ to reveal relationship between the regularity of function class and its corresponding empirical covering number.

Example 3.6. Let X be a bounded domain in $\mathbb{R}^n$ and H^s the Sobolev space of index s. When $s > n$, the classical Embedding Theorem tells us that $H^s(X)$ is an RKHS and its unit ball B_1 is embedded in a finite ball of the function space $C^{s-(n/2)-\zeta}(X)$ with inclusion bounded where $0 < \zeta < s - n$. From the classical bounds for covering numbers of the unit ball of $C^{s-(n/2)-\zeta}(X)$, we see that

$$\log(\mathcal{N}_2(B_1, \epsilon)) \le C_s \epsilon^{-n/(s-n/2-\zeta)}, \quad \forall \epsilon > 0. \tag{3.17}$$

Hence Assumption (3.13) below holds with $v = n/(s - n/2 - \zeta) < 2$.

Our concentration estimate for the sample error dealing with $\mathcal{S}_1$ is based on the following concentration inequality, which can be found in [12].

Lemma 3.7. *Let $\mathcal{F}$ be a class of measurable functions on Z. Assume that there are constants $B, c > 0$ and $\eta \in [0, 1]$ such that $\|f\|_\infty \le B$ and $\mathbb{E}f^2 \le c(\mathbb{E}f)^\eta$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $v \in (0, 2)$,*

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \le a\epsilon^{-v}, \quad \forall \epsilon > 0, \tag{3.18}$$

then there exists a constant c_v such that for any $t > 0$, with confidence at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \le \frac{1}{2}\gamma^{1-\eta}(\mathbb{E}f)^\eta + c_v\gamma + 2\left(\frac{ct}{m}\right)^{1/(2-\eta)} + \frac{18Bt}{m}, \quad \forall f \in \mathcal{F}, \tag{3.19}$$

where

$$\gamma := \max\left\{c^{(2-v)/(4-2\eta+v\eta)}\left(\frac{a}{m}\right)^{2/(4-2\eta+v\eta)}, B^{(2-v)/(2+v)}\left(\frac{a}{m}\right)^{2/(2+v)}\right\}. \tag{3.20}$$

Denote the set of function $\mathcal{F}_R^q$ with $R > 0$, where

$$\mathcal{F}_R^q = \left\{(\pi(f)(x) - y)^2 - (f_\rho(x) - y)^2 : f \in B_R^q\right\}. \tag{3.21}$$

Proposition 3.8. *If B_1^q satisfies the capacity condition* (3.13) *with some* $0 < \upsilon < 2$, *then for any* $\delta \in (0,1)$, *with confidence* $1 - \delta/2$, *there holds*

$$\mathcal{S}_1 \le \frac{1}{2}\left\|\pi(f_{\mathbf{z}}) - f_\rho\right\|_\rho^2 + \frac{80T^2\log(2/\delta)}{m} + C_{\upsilon,T}\left(\frac{1}{\lambda}\right)^{\upsilon/(2+\upsilon)}\left(\frac{1}{m}\right)^{2/(2+\upsilon)} \tag{3.22}$$

with constant $C_{\upsilon,T} = 4c_\upsilon\left(C_{\upsilon,\tilde{K}}T^2\right)^{\upsilon/(2+\upsilon)} M^{2\upsilon/(2+\upsilon)q^*}$.

Proof. Consider the set $\mathcal{F}_R^q$. Each function $g \in \mathcal{F}_R^q$ can be expressed as $g(z) = (y - \pi(f)(x))^2 - (y - f_\rho(x))^2$ with some $f \in B_R^q$. Then $\mathbb{E}g = \mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho) = \|\pi(f) - f_\rho\|_\rho^2$ and $(1/m)\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho)$. Note that

$$g(z) = \left(\pi(f)(x) - f_\rho(x)\right)\left(\pi(f)(x) + f_\rho(x) - 2y\right). \tag{3.23}$$

Since $|\pi(f)(x)| \le T$ and $|f_\rho(x)| \le T$ for any $x \in X$, we see that for any $z \in Z$,

$$\begin{gathered} \left|g(z)\right| \le 8T^2, \\ \mathbb{E}g^2 = \int_Z \left(\pi(f)(x) - f_\rho(x)\right)^2\left(\pi(f)(x) + f_\rho(x) - 2y\right)^2 \le 4T^2\mathbb{E}g. \end{gathered} \tag{3.24}$$

On the other hand, for any $g_1, g_2 \in \mathcal{F}_R^q$ at point $z = (x, y)$, we have

$$\left|g_1(z) - g_2(z)\right| \le 4T\left|f_1(x) - f_2(x)\right|, \tag{3.25}$$

since the projector operator π is a contractive map. It follows that

$$\mathcal{N}_{2,\mathbf{z}}\left(\mathcal{F}_R^q, \epsilon\right) \le \mathcal{N}_{2,\mathbf{x}}\left(B_R^q, \frac{\epsilon}{4T}\right) \le \mathcal{N}_{2,\mathbf{x}}\left(B_1^q, \frac{\epsilon}{4TR}\right). \tag{3.26}$$

It follows from the capacity condition (3.15)

$$\log\mathcal{N}_2\left(\mathcal{F}_R^q, \epsilon\right) \le C_{\upsilon,\tilde{K}}\left(4TRM^{1/q^*}\right)^{\upsilon}\epsilon^{-\upsilon}. \tag{3.27}$$

Applying Lemma 3.7 with $B = c = 4T^2$, $\eta = 1$, and $a = C_{\upsilon,\tilde{K}}(4TRM^{1/q^*})^{\upsilon}$, we see that for any $\delta \in (0,1)$, with confidence $1 - \delta/2$, there holds

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}\mathbb{E}g + \frac{80T^2\log(2/\delta)}{m} + C_{\upsilon,T}R^{2\upsilon/(2+\upsilon)}\left(\frac{1}{m}\right)^{2/(2+\upsilon)}, \quad \forall f \in \mathcal{F}_R^q. \tag{3.28}$$

Besides, following the definition of $f_{\mathbf{z}}$ (1.9), we have

$$\lambda\left\|f_{\mathbf{z}}\right\|_{2,q}^2 = \lambda\|w_{\mathbf{z}}\|_{2,q}^2 \overset{w=0}{\le} \frac{1}{m}\sum_{i=1}^m y_i^2 \le T^2, \tag{3.29}$$

that is, $\|f_{\mathbf{z}}\|_{2,q} \leq T/\sqrt{\lambda}$. Hence we can replace R with $T/\sqrt{\lambda}$. Thus we derive our desired result. □

4. Total Learning Rates

We are now in a position to obtain the learning rates of projected algorithm (1.9). Main results of this paper will be presented in Theorem 4.1.

Following the error decomposition scheme in Proposition 2.3 and combining Propositions 3.2 and 3.8, we derive the following bounds on the total error.

Theorem 4.1. *Suppose that B_1^q satisfies the capacity condition* (3.13) *with some $0 < \upsilon < 2$, and $\mathcal{A}_q(\lambda) \leq C_\beta \lambda^\beta$. For any $\delta \in (0,1)$, with confidence $1-\delta$, there holds*

$$\|\pi(f_{\mathbf{z}}) - f_\rho\|_\rho^2 \leq C' \log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\min\{2\beta/(2\beta+(1+\beta)\upsilon),\beta\}}, \quad \textit{by taking } \lambda = \left(\frac{1}{m}\right)^{\min\{2/(2\beta+(1+\beta)\upsilon),1\}}, \tag{4.1}$$

where $C' = 3\kappa^2 M^{2/q^*} + 86T^2 + 1 + C_{\upsilon,T} 2^{\upsilon/(2+\upsilon)} + C_\beta$ and $C_{\upsilon,T}$ is defined as in Proposition 3.8.

Proof. Following Propositions 3.2 and 3.8, with confidence at least $1-\delta$, $(1/2)\|\pi(f_{\mathbf{z}}) - f_\rho\|_\rho^2$ can be bounded by

$$\frac{3\kappa^2 M^{2/q^*}\mathcal{A}_q(\lambda)}{m\lambda}\log\left(\frac{2}{\delta}\right) + \frac{(86T^2+1)\log(2/\delta)}{m} + C_{\upsilon,T}\left(\frac{2}{\lambda}\right)^{\upsilon/(2+\upsilon)}\left(\frac{1}{m}\right)^{\upsilon/(2+\upsilon)} + 2\mathcal{A}_q(\lambda). \tag{4.2}$$

Firstly, we set

$$\frac{\mathcal{A}_q(\lambda)}{m\lambda} = \mathcal{A}_q(\lambda), \tag{4.3}$$

which implies that $\lambda = (1/m)$. On the other hand, from the assumption $\mathcal{A}_q(\lambda) \leq C_\beta \lambda^\beta$, we set

$$\left(\frac{2}{\lambda}\right)^{\upsilon/(2+\upsilon)}\left(\frac{1}{m}\right)^{2/(2+\upsilon)} = \lambda^\beta, \quad \text{that is}, \; \lambda = m^{-(2/(2\beta+(1+\beta)\upsilon))}. \tag{4.4}$$

Hence our assertion follows by taking $\lambda = (1/m)^{\min\{2/(2\beta+(1+\beta)\upsilon),1\}}$. □

Proof of Theorem 1.3. When $\widetilde{K} \in C^\infty$, it follows that condition (3.13) holds for arbitrary small $\upsilon > 0$. Moreover, $f_\rho \in \mathcal{H}_{\widetilde{K}}$ implies that $\beta = 1$, the conclusion follows easily from Theorem 4.1.

Our learning rates below in Corollary 4.2 will be achieved under the regularity assumption on the regression function that f_ρ lies in the range of $L^r_{\widetilde{K}}$ for some $r > 0$. Given any kernel K, L_K is the integral operator on $L^2_{\rho_X}$ defined by

$$L_K(f)(x) = \int_X K(x,y) f(y) d\rho_X(y), \quad x \in X,\ f \in L^2_{\rho_X}. \tag{4.5}$$

The operator L_K is linear, compact, positive and can be also regarded as a self-adjoint operator on $\mathscr{H}_K$. Hence the fractional power operator $L^r_K : L^2_{\rho_X} \to L^2_{\rho_X}$ is well defined and is given by

$$L^r_K(f) = \sum_k \lambda^r_k \langle f, \varphi_k \rangle_{L^2_{\rho_X}} \varphi_k, \quad f \in L^2_{\rho_X}(X), \tag{4.6}$$

where $\{\lambda_k\}_k$ are eigenvalues of the operator L_K arranged in a decreasing order and $\{\varphi_k\}_k$ are the corresponding eigenfunctions, which form an orthonormal basis of $L^2_{\rho_X}$. In fact, the image of L^r_K is contained in $\mathscr{H}_K$ if $r \geq 1/2$. So $L^{-r}_K f_\rho \in L^2_{\rho_X}$ indicates that f_ρ lies in the range of L^r_K, measuring the regularity of the regression function. □

Corollary 4.2. *Suppose that B^q_1 satisfies the capacity condition* (3.13) *with some* $0 < \upsilon < 2$, *and* $L^{-r}_{\widetilde{K}} f_\rho \in L^2_{\rho_X}$ $(0 < r \leq 1)$. *For any* $0 < \delta < 1$, *with confidence* $1 - \delta$, *there holds*

$$\left\|\pi(f_{\mathbf{z}}) - f_\rho\right\|^2_\rho \leq \widetilde{C} \log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\min\{2\beta/(2\beta+(1+\beta)\upsilon),\beta\}}, \quad \text{with } \beta = \min\{2r, 1\}, \tag{4.7}$$

where $\widetilde{C}$ is some constant independent of m or δ.

Proof. Recall a result from [18], if $L^{-r}_{\widetilde{K}} f_\rho \in L^2_{\rho_X}$ $(0 < r \leq 1/2)$, there holds

$$\mathscr{A}_2(\lambda) = \left\|f_\lambda - f_\rho\right\|^2_\rho + \lambda \left\|f_\lambda\right\|^2_{\widetilde{K}} \leq \lambda^{2r} \left\|L^{-r}_{\widetilde{K}} f_\rho\right\|_{L^2_{\rho_X}}. \tag{4.8}$$

If $r \geq 1/2$, this shows that $f_\rho \in \mathscr{H}_{\widetilde{K}}$ as mentioned above, then we have $f_\lambda = f_\rho$ and

$$\mathscr{A}_2(\lambda) \leq \lambda \left\|f_\rho\right\|^2_{\widetilde{K}}. \tag{4.9}$$

Hence for any $r \in [0, 1]$, we have

$$A_2(\lambda) \leq \lambda^\beta, \quad \text{with } \beta = \min\{2r, 1\}. \tag{4.10}$$

On the other hand, observe Lemma 5.1 below, and we have

$$\mathscr{A}_q(\lambda) \leq C^* \lambda^\beta, \quad \text{with } \beta = \min\{2r, 1\}, \tag{4.11}$$

where C^* is some constant and $q \in [1, 2]$. The conclusion follows immediately from Theorem 4.1. □

Let us compare our learning rates with the existing results.

In [10], a uniform covering number technique was used to derive the expected value of learning schemes (1.1) where $\Omega(f_w) = \|w\|_K^2$. If all the kernels $\{K_i\}$ are the same with some specialized K and $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $0 < r \leq 1/2$. For any $0 < \zeta < 1/(1+\upsilon)$ and any $0 < \delta < 1$, with confidence $1 - \delta$, then

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \leq \log\left(\frac{2}{\delta}\right)\mathcal{O}\left(m^{-2r\zeta}\right). \tag{4.12}$$

Clearly the learning rates derived from Corollary 4.2 are better than that in [10] since $2r\zeta < 2r/(1+\upsilon) \leq 4r/(4r+(1+2r)\upsilon)$.

In [21], an operator monotonic technique was used to improve the kernel independent error bounds in comparison with the result in [17]. If $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $0 < r \leq 1$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \leq \log\left(\frac{2}{\delta}\right)\mathcal{O}\left(m^{-3r/(2r+2)}\right). \tag{4.13}$$

When $r \geq 1/2$ and $\upsilon \leq (2-r)/3r$, the learning rate given by Corollary 4.2 is better than the above result.

As for empirical risk minimization (ERM), classical results on analysis of ERM schemes give error bounds between the empirical target function and the regression function. In particular, learning rates of type $m^{-\varepsilon}$ with ε arbitrarily close to 1 can be achieved by ERM schemes (see [15]). However, the ERM setting is different from the one on Tikhonov regularization. How to choose the regularization parameter $\lambda = \lambda(m)$, depending on the sample size m, is the essential difficulty for the regularization scheme, even when f_ρ lies in $\mathscr{H}_K$. On the other hand, it is obvious that our result is more general than that of [15] since the case for $f_\rho \notin \mathscr{H}_K$ $(r < 1/2)$ is also covered.

5. Discussion on Regularization Error

By our assumptions on M different kernels $\{K_j\}_{j=1}^M$, we see that $\mathscr{H}_{\widetilde{K}}$ is an RKHS generated by the Mercer kernel $\widetilde{K}$. There are several standard approximation results on regularization error $\mathscr{A}_2(\lambda)$ in learning theory (see [17]). Next we establish a tight connection between $\mathscr{A}_2(\lambda)$ and $\mathscr{A}_q(\lambda)$ with $1 \leq q \leq 2$.

Lemma 5.1. *Let $\mathscr{H}_{\widetilde{K}}$ be a separable RKHS over X associated with a bounded measurable kernel, ρ be a distribution on $X \times [-T, T]$, and $1 \leq q \leq 2$. If there exist constants $C > 0$ and $\beta > 0$ such that $\mathscr{A}_2(\lambda) \leq C\lambda^\beta$, then for all $\lambda > 0$ we have*

$$\mathscr{A}_q(\lambda) \leq C\left(M^{(2-q)/2} + 1\right)\lambda^\beta. \tag{5.1}$$

Proof. If there exists a function $\widetilde{f}_\lambda$ satisfying

$$\lambda\left\|\widetilde{f}_\lambda\right\|_{2,2}^2 + \left\|\widetilde{f}_\lambda - f_\rho\right\|_\rho^2 \leq C\lambda^\beta, \tag{5.2}$$

we see that $\lambda\|\tilde{f}_\lambda\|_{2,2}^2 \le C\lambda^\beta$. We write $\tilde{f}_\lambda(x) = \langle \tilde{w}, \Psi(x)\rangle$ with some $\tilde{w} = (\tilde{w}^{(1)}, \ldots, \tilde{w}^{(M)})$, by the Hölder inequality, we have

$$\left\|\tilde{f}_\lambda\right\|_{2,q}^q = \sum_{j=1}^{M} \left\|\tilde{w}^{(j)}\right\|_{K_j}^q \overset{\text{Hölder}}{\le} M^{(2-q)/2}\left\|\tilde{f}_\lambda\right\|_{2,2}^q. \tag{5.3}$$

Then we obtain

$$\mathcal{A}_q(\lambda) \le \lambda\left\|\tilde{f}_\lambda\right\|_{2,q}^2 + \left\|\tilde{f}_\lambda - f_\rho\right\|_\rho^2 \le CM^{(2-q)/2}\lambda^\beta + C\lambda^\beta = C\left(M^{(2-q)/2} + 1\right)\lambda^\beta. \tag{5.4}$$

□

In other words, if $\mathcal{A}_2(\lambda)$ has a polynomial behavior in λ, then this behavior completely determines the behavior of all $\mathcal{A}_q(\lambda)$. Thus it suffices to assume that the standard 2-approximation error function satisfies (2.2).

From statistical effective dimension point of view, we will discuss the impact of the multikernel class $\mathcal{H}_{\tilde{K}}$ on the approximation error $\|f_\lambda - f_\rho\|_{L^2_{\rho_X}}$. To estimate this error, note that the regularizing function of $\mathcal{A}_2(\lambda)$ exists, is unique, and given by [7]

$$f_\lambda = \left(\lambda I + L_{\tilde{K}}\right)^{-1} L_{\tilde{K}} f_\rho. \tag{5.5}$$

For simplicity, let $M = 2$ and take a Mercer kernel K_o as the original one, by the classical Mercer theorem, K_o can be expressed as $K_o(x,y) = \sum_{k=1}^\infty \lambda_k \varphi_k(x)\varphi_k(y)$. Another kernel we take is $K_o^N = \sum_{k=1}^\infty \lambda_k^N \varphi_k(x)\varphi_k(y)$ $(2 \le N \in \mathbb{N}^+)$. In this case, $\tilde{K} = K_o + K_o^N = \sum_{k=1}^\infty (\lambda_k + \lambda_k^N)\varphi_k(x)\varphi_k(y)$. By the fact that $f_\lambda - f_\rho = \lambda(\lambda I + L_{\tilde{K}})^{-1} f_\rho$ and assumption $L_{K_o}^{-r} f_\rho \in L^2_{\rho_X}$, we have

$$\begin{aligned}\|f_\lambda - f_\rho\|_{L^2_{\rho_X}} &= \lambda\left\|\left(\lambda I + L_{\tilde{K}}\right)^{-1} L_{K_o}^r L_{K_o}^{-r} f_\rho\right\|_{L^2_{\rho_X}} = \lambda\left\|\sum_{k=1}^\infty \alpha_k \frac{\lambda_k^r}{\lambda_k + \lambda_k^N + \lambda}\varphi_k\right\|_{L^2_{\rho_X}} \\ &\le \lambda^{\min\{r/N,1\}}\left\|L_{K_o}^{-r} f_\rho\right\|_{L^2_{\rho_X}} \quad \text{for } \|\alpha\|_{l^2} = \left\|L_{K_o}^{-r} f_\rho\right\|_{L^2_{\rho_X}}.\end{aligned} \tag{5.6}$$

Let us compare the multikernel class regularization with Tikhonov regularization in $\mathcal{H}_{K_o}$ when the Mercer kernel K_o is employed. Denote the saturation index as the maximal r so that the approximation error achieves fastest decay rate under the condition $L_{K_o}^{-r} f_\rho \in L^2_{\rho_X}$. Then (5.6) shows the saturation index for multikernel class regularization is N while it is 1 for Tikhonov regularization in $\mathcal{H}_{K_o}$, as shown in [17].

In this case, our analysis implies that we should use an alternative kernel with faster eigenvalue decay when the spectral coefficients of the target function decay faster: for example, using K_o^N instead of K_o. This has a dimension reduction effect. Essentially, we effectively project the data into the principal components of data. The intuition is also quite clear: if the dimension of the target function is small (spectral coefficient r decays fast), then we should project data to those dimensions by reducing the remaining noisy dimensions (corresponding to fast kernel eigenvalue decay N). In fact, the similar idea under the framework of semisupervised learning has been shown in spectral kernel design methods [22, 23].

In general, for the sample error, there exist rates of convergence which hold independently of the underlying distribution ρ. This is important, as it tells us that we can give convergence guarantees no matter what a kernel is used, even we do not know the underlying distribution. In fact, this is very common in statistical analysis of various machine learning algorithms (see [24]). This decay is usually fast enough for practical use where amounts of samples are available. For the approximation error, however, it is impossible to give rates of convergence which hold for all probability distributions ρ. Hence what determines the learning accuracy is the approximation error. In kernel regression setup, this is determined by the choice of the kernel and enhances the importance of learning kernels [4] and constructing refined kernels [25].

6. Further Study

In the last section, we exclusively discuss sparsity in the case of the square loss regularization functional in (1.1) with the regularizer $\Omega(f) = \|f\|_K^2$ in RKHS. We can derive the explicit expression for this functional from [4], in turn which provides improvement and simplification of our algorithm (1.4).

Lemma 6.1. *For any kernel K and positive constant λ, we have that*

$$\mathcal{S}_\lambda(K) := \min\left\{\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2 + \lambda\|f\|_K^2\right\} = \lambda\left\langle \mathbf{y}, (\lambda I + K(\mathbf{x}))^{-1}\mathbf{y}\right\rangle, \tag{6.1}$$

where the vector $\mathbf{y} = (y_i, \ldots, y_m)^T$ and $K(\mathbf{x})$ denotes the $m \times m$ Gram matrix $(K(x_i, x_j) : i, j = 1, \ldots, m)$ and $\langle\cdot\rangle$ denotes the standard inner product in Euclidean space.

According to Lemma 6.1, the least square algorithm of (1.4) can be rewritten as a one-layer minimization problem

$$\theta_{\mathbf{z}} := \arg\min_{\theta\geq 0}\left\{\langle \mathbf{y}, (\lambda I + K_\theta(\mathbf{x}))^{-1}\mathbf{y}\rangle + \frac{\mu}{\lambda}\sum_{k=1}^{M}|\theta_k|^p\right\}. \tag{6.2}$$

We assume that $f_\rho \in \mathcal{H}_{\theta_0}$ for some $\theta_0 \in \mathbb{R}^M$. Define $J_\theta := \mathrm{supp}(\theta) = \{k : \theta_k \neq 0\}$. We say that f_ρ is sparse relative to θ_0 if the cardinally of J_{θ_0} is far smaller than M.

For $p \in [1, p_M]$ (but p_M is close to 1), it would be interesting to show that the solution $\theta_{\mathbf{z}}$ of (6.2) is approximately sparse following the path of θ_0. In some sense, $\sum_{k\notin J_{\theta_0}}(\theta_{\mathbf{z}})_k$ is very small with a very high probability. A refined analysis of l^p-regularized methods was done by Koltchinskii [26] in the case of combination of M basis functions, mainly taking into account the soft sparsity pattern of the Bayes function and establishing several oracle inequalities in statistical sense. Extending the ideas into the kernels learning setting would be of a great significance, because it can provide theoretical support showing that the l^p-norm MKL can automatically select good kernels, which coincide with the underlying right kernels.

Acknowledgments

The authors would like to thank the two anonymous referees for their valuable comments and suggestions which have substantively improved this paper. The research of S.-G. Lv is supported partially by 211 project for the Southwestern University of Finance and Economics, China [Project no. 211QN2011028] as well as Annual Cultivation of the Fundamental Research Funds for the Central Universities [Project no. JBK120940].

References

[1] C. Cortes, M. Mohri, and A. Rostamizadeh, "Generalization bounds for learning kernels," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 247–254, June 2010.

[2] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "l^p-norm multiple kernel learning," *Journal of Machine Learning Research*, vol. 12, pp. 953–997, 2011.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[4] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[5] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[6] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. 3, pp. 499–526, 2002.

[7] F. Cucker and S. Smale, "On the mathematical foundations of learning," *American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[8] Y. K. Zhu and H. W. Sun, "Consisitency analysis of spectral regularization algorithms," *Abstract and Applied Analysis*, vol. 2012, Article ID 436510, 16 pages, 2012.

[9] Y. Ying and D.-X. Zhou, "Learnability of Gaussians with flexible variances," *Journal of Machine Learning Research*, vol. 8, pp. 249–276, 2007.

[10] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, NY, USA, 2nd edition, 2001.

[12] Q. Wu, Y. Ying, and D.-X. Zhou, "Multi-kernel regularized classifiers," *Journal of Complexity*, vol. 23, no. 1, pp. 108–134, 2007.

[13] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, vol. 68, no. 1, pp. 49–67, 2006.

[14] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 775–782, Corvallis, Ore, USA, June 2007.

[15] M. Kloft and G. Blanchard, "The local rademacher complexity of l^p-norm multiple kernel learning," in *Advances in Neural Information Processing Systems (NIPS '11)*, pp. 2438–2446, MIT Press, 2011.

[16] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer, New York, NY, USA, 1996.

[17] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[18] S. Smale and D.-X. Zhou, "Shannon sampling. II. Connections to learning theory," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 285–302, 2005.

[19] A. Micchelli, M. Pontil, Q. Wu, and D. X. Zhou, "Error bounds for learning the kernel," Research Note 05–09, University of College, London, UK, 2005.

[20] D.-X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *Institute of Electrical and Electronics Engineers*, vol. 49, no. 7, pp. 1743–1752, 2003.

[21] H. Sun and Q. Wu, "A note on application of integral operator in learning theory," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 416–421, 2009.

[22] O. Chapelle, J. Weston, and B. Sch'olkopf, "Cluster kernels for semi-supervised learning15," in *Advances in Neural Information Processing Systems (NIPS '03)*, pp. 585–592, MIT Press, 2003.

[23] R. Johnson and T. Zhang, “Graph-based semi-supervised learning and spectral kernel design,” *Institute of Electrical and Electronics Engineers*, vol. 54, no. 1, pp. 275–288, 2008.

[24] U. von Luxburg and B. Sch’olkopf, “Statistical learning theory: models, concepts, and results”.

[25] Y. Xu and H. Zhang, “Refinement of reproducing kernels,” *Journal of Machine Learning Research*, vol. 10, pp. 107–140, 2009.

[26] V. Koltchinskii, “Sparsity in penalized empirical risk minimization,” *Annales de l’Institut Henri Poincaré Probabilités et Statistiques*, vol. 45, no. 1, pp. 7–57, 2009.