Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 458459, 7 pages
http://dx.doi.org/10.1155/2014/458459

Research Article

Constructive Analysis for Least Squares Regression with Generalized K-Norm Regularization

Cheng Wang and Weilin Nie

Department of Mathematics, Huizhou University, Huizhou, Guangdong 516007, China

Correspondence should be addressed to Cheng Wang; math.cwang@gmail.com

Received 18 February 2014; Accepted 11 July 2014; Published 22 July 2014

Academic Editor: Feliz Minhós

We introduce a constructive approach for the least squares algorithms with generalized K-norm regularization. Different from the previous studies, a stepping-stone function is constructed with some adjustable parameters in error decomposition. It makes the analysis flexible and may be extended to other algorithms. Based on projection technique for sample error and spectral theorem for integral operator in regularization error, we finally derive a learning rate.

1. Introduction

In learning theory, we are always given a sample set $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$, which is drawn from a joint distribution ρ on the sample space $Z := X \times Y$. Here, the input space X is a compact metric space and $Y = \mathbb{R}$ for a regression problem. For a function f obtained via some algorithm, a loss functional $L(f(x), y)$ is defined to measure its performance on a sample point (x, y). In regression problem, least square loss $L(f(x), y) = (f(x) - y)^2$ is most widely used. Then, we can use the generalization error to evaluate f over the whole sample space:

$$\mathscr{E}(f) = \int_Z (f(x) - y)^2 d\rho. \tag{1}$$

From [1], we know the goal function is $f_\rho = \int_Z y d\rho(y \mid x)$, which is called the regression function, minimizing the generalization error. Since ρ is always unknown in practice, we have to find another function close to f_ρ based on the sample. The famous empirical risk minimization (ERM) algorithm is introduced in [2, 3]. To avoid overfitting, a penalty term $\Omega(f)$ related to f is added into this algorithm, which is usually called regularization. While the squared K-norm regularization term is extensively studied in [4], and so forth, in this paper, we consider a more general model:

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathscr{H}_K} \frac{1}{m}(f(x_i) - y_i)^2 + \lambda \|f\|_K^p, \tag{2}$$

with some $p > 0$. In this algorithm, minimization is restricted to a hypothesis space $\mathscr{H}_K$ which is a reproducing kernel Hilbert space (RKHS) on X. The RKHS [5] is defined as $\mathscr{H}_K := \overline{\text{span}\{K_x : x \in X\}}$ with $K_x(y) = K(x, y)$, associated with a Mercer Kernel $K : X \times X \to \mathbb{R}$ which is continuous, symmetric, and positive definite. Since X is a compact metric space, Kernel K is bounded and we denote $\kappa = \sup_{x,t \in X} K(x, t)$ in the following.

2. Main Result

Though uniform bounded assumption was abandoned in previous work [6], we still assume

$$|y| \le M \tag{3}$$

almost surely for some constant $M > 0$ throughout this paper for simplicity, since our analysis can be extended to the unbounded situation by choosing some different probability inequality.

For the hypothesis space, a polynomial decay condition is given to control the capacity. To state this condition, we have to firstly recall covering number.

Definition 1. Let $(\mathscr{M}, d)$ be a pseudometric space and $S \subset \mathscr{M}$. For $\varepsilon > 0$, the covering number $\mathscr{N}(S, \varepsilon, d)$ of the set S with

respect to d is defined to be the minimal number of balls of radius ε whose union covers S. That is,

$$\mathcal{N}(S,\varepsilon,d) = \min\left\{n\in\mathbb{N}: \exists\{f_i\}_{i=1}^n \subset \mathcal{M} \text{ such that } S\subset\bigcup_{i=1}^n B(f_i,\varepsilon)\right\}, \tag{4}$$

where $B(f_i,\varepsilon)=\{f\in\mathcal{M}: d(f,f_i)\le\varepsilon\}$.

When metric d is chosen to be $\|\cdot\|_\infty$, that is, $d(f,g)=\|f-g\|_\infty$, it is the classical uniform covering number. It is widely used in [4, 6–8], and so forth, and more detailed analysis can be found in [9, 10]. More recent references [11–14] use ℓ^2-empirical covering number to obtain a sharper upper bound for the excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)$.

Definition 2. Denote

$$d_2(\mathbf{a},\mathbf{b}) = \left(\frac{1}{m}\sum_{i=1}^m |a_i-b_i|^2\right)^{1/2} \tag{5}$$

for some $a,b\in\mathbb{R}^m$. For a set $\mathcal{F}$ of functions on X and $\varepsilon>0$, with notation $\mathbf{z}=(z_i)_{i=1}^m\subset X^m$ and $\mathcal{F}|_{\mathbf{z}}=\{(f(z_i))_{i=1}^m: f\in\mathcal{F}\}$, the ℓ^2-empirical covering number of $\mathcal{F}$ is given by

$$\mathcal{N}_2(\mathcal{F},\varepsilon) = \sup_{m\in\mathbb{N}}\sup_{\mathbf{z}\in X^m}\mathcal{N}\left(\mathcal{F}|_{\mathbf{z}},\varepsilon,d_2\right). \tag{6}$$

Now, we can describe the capacity condition of the hypothesis space $\mathcal{H}_K$.

Definition 3. We say that $\mathcal{H}_K$ has empirical polynomial complexity with exponent s, $0<s<2$, if there exists a constant $c_s>0$ such that

$$\log\mathcal{N}_2\left(B_1\left(\mathcal{H}_K\right),\varepsilon\right)\le c_s\varepsilon^{-s},\quad \forall\varepsilon>0, \tag{7}$$

where $B_R(\mathcal{H}_K)=\{f\in\mathcal{H}_K: \|f\|_K\le R\}$ is the ball with radius R in $\mathcal{H}_K$.

The integral operator $L_K: L^2_{\rho_X}\to L^2_{\rho_X}$ defined by

$$L_Kf(x)=\int_X f(t)K(x,t)\,d\rho_X(t) \tag{8}$$

is also important in learning theory and has been studied in [15]. In [1], the authors claim that, for a Mercer Kernel K, the associated L_K is a compact operator with nonincreasing positive eigenvalue sequence μ_i. And the induced fractional operator

$$L_K^rf(x)=\sum_{i\ge1}\mu_i^r\phi_i(x) \tag{9}$$

is well defined, for any $f=\sum_{i\ge1}c_i\phi_i\in L^2_{\rho_X}$ with orthogonal basis $\{\phi_i\}_{i\ge1}$ of $L^2_{\rho_X}$. In the following, we will make use of this notion in our construction analysis.

We additionally introduce the projection operator π on the space of measurable function $f: X\to R$:

$$\pi(f(x)) = \begin{cases} M & f(x)>M\\ f(x) & M\ge f(x)\ge -M\\ -M & f(x)<-M.\end{cases} \tag{10}$$

The main result is stated as follows which will be proved in Section 6.

Theorem 4. *Assume* (3), (7) *hold for sample distribution and hypothesis space $\mathcal{H}_K$. The regression function satisfies $f_\rho\in L_K^r(L^2_{\rho_X})$ for some $r>0$. $f_{\mathbf{z},\lambda}$ is obtained from* (2). *Then, by choosing appropriate λ (explicit expression can be found in the proof) with confidence $1-\delta$ for any $0<\delta<1/2$, we have*

$$\left\|\pi(f_{\mathbf{z},\lambda})-f_\rho\right\|_\rho^2\le C_{p,r,s,M}\log\frac{2}{\delta}m^{-2p\xi/(2s+(s+2)p\xi)} \tag{11}$$

for some constant $C_{p,r,s,M}$ not depending on m or δ and

$$\xi=\begin{cases}1 & r>\dfrac{1}{2}\\ \dfrac{4r}{4r+(1-2r)\,p} & r\le\dfrac{1}{2}.\end{cases} \tag{12}$$

3. Error Decomposition

Various error decomposition methods motivate our research, especially [7, 12, 14, 16, 17]. A general idea of error decomposition is to transform the excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)=\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho^2$ (see [1] for details) to two parts, which can be bounded by some concentration inequality and approximation analysis. In our setting, let f_λ be a function to be determined in $\mathcal{H}_K$; it can be expressed as

$$\begin{aligned}\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f_\rho) &\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f_\rho)+\lambda\|f_{\mathbf{z},\lambda}\|_K^p\\ &\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\\ &\quad+\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))+\lambda\|f_{\mathbf{z},\lambda}\|_K^p-\mathcal{E}(f_\rho)\\ &\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\\ &\quad+\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\lambda\|f_{\mathbf{z},\lambda}\|_K^p-\mathcal{E}(f_\rho)\\ &\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\\ &\quad+\mathcal{E}_{\mathbf{z}}(f_\lambda)+\lambda\|f_\lambda\|_K^p-\mathcal{E}(f_\rho)\\ &\le S_1+S_2+D(\lambda),\end{aligned} \tag{13}$$

where

$$\begin{aligned}S_1 &= \left(\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f_\rho)\right)\\ &\quad-\left(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}(f_\rho)\right),\\ S_2 &= \left(\mathcal{E}_{\mathbf{z}}(f_\lambda)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)-\left(\mathcal{E}(f_\lambda)-\mathcal{E}(f_\rho)\right),\\ D(\lambda) &= \mathcal{E}(f_\lambda)-\mathcal{E}(f_\rho)+\lambda\|f_\lambda\|_K^p.\end{aligned} \tag{14}$$

The first and second terms S_1 and S_2 are called sample error which will be studied in Section 5, while the third term $D(\lambda)$ is regularization error (or approximation error) which is our main work in this paper.

It is known that f_λ can be freely chosen in $\mathscr{H}_K$ which is close to f_ρ in some sense and in previous work f_λ is always naturally chosen to be the one minimizing $D(\lambda)$. However, we will encounter difficulties if the minimizer does not exist or the expression of the minimizer is not explicit. In this paper, we construct a special function in the form

$$f_\lambda = \left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^\alpha f_\rho \tag{15}$$

with some $\alpha, \beta > 0$ to handle this problem.

4. Regularization Error

It is the main contribution of this section to conduct error analysis for the regularization error. Regularization error, also called approximation error, has already been studied in [18]. However, we will analyze this part of error in a different viewpoint. From [1], we know L_K is a compact self-adjoint and positive operator. By applying the spectral theorem for compact operators, we can bound a compact positive operator with its eigenvalues. Firstly, we have to introduce a useful lemma.

Lemma 5. *Let $a, b, c, d > 0$ and $c < d$; one has*

$$\frac{a^c}{a^d + b} \le b^{(c/d)-1}. \tag{16}$$

Proof. By simply taking derivative of the right-hand side with respect to a, we can find that it reaches its maximum when $a = (c/(d-c))^{1/d} b^{1/d}$; that is,

$$\sup_{a>0} \frac{a^c}{a^d + b} = \frac{c}{d}\left(\frac{c}{d-c}\right)^{(c/d)-1} b^{(c/d)-1}. \tag{17}$$

Since $(c/d)(c/(d-c))^{(c/d)-1} = r^r (1-r)^{1-r} < 1$, where $0 < r = c/d < 1$, the lemma is proved. □

Proposition 6. *Assume $f_\rho \in L_K^r(L^2_{\rho_X})$ and (15); there holds*

$$D(\lambda) \le C_{r,p} \lambda^\xi \tag{18}$$

with some constant $C_{r,p}$ depending on r, p and

$$\xi = \begin{cases} 1 & r > \dfrac{1}{2} \\ \dfrac{4r}{4r + (1-2r)\,p} & r \le \dfrac{1}{2} \end{cases} \tag{19}$$

by choosing appropriate α and β.

Proof. Since $D(\lambda) = \mathscr{E}(f_\lambda) - \mathscr{E}(f_\rho) + \lambda \|f_\lambda\|_K^p$, we will analyze the two terms, respectively. Noting that $\mathscr{E}(f) - \mathscr{E}(f_\rho) = \|f - f_\rho\|_\rho^2$ and $L_K^{-r} f_\rho \in L^2_{\rho_X}$, we have

$$\begin{aligned} \mathscr{E}(f_\lambda) - \mathscr{E}(f_\rho) &= \left\|f_\lambda - f_\rho\right\|_\rho^2 \\ &= \left\|\left(\left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^\alpha - I\right) f_\rho\right\|_\rho^2 \\ &= \lambda^{2\beta} \left\|\left(L_K^\alpha + \lambda^\beta I\right)^{-1} f_\rho\right\|_\rho^2 \\ &\le \lambda^{2\beta} \left\|\left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^r\right\|^2 \left\|L_K^{-r} f_\rho\right\|_\rho^2 \\ &= \lambda^{2\beta} \left(\sup_{i\ge 1} \frac{\mu_i^r}{\mu_i^\alpha + \lambda^\beta}\right)^2 \left\|L_K^{-r} f_\rho\right\|_\rho^2. \end{aligned} \tag{20}$$

Recall that $\sup_{i\ge1} \mu_i = \|L_K\| \le \kappa$, combining with Lemma 5; there holds

$$\sup_{i\ge 1} \frac{\mu_i^r}{\mu_i^\alpha + \lambda^\beta} \le \begin{cases} \kappa^{r-\alpha} & \alpha \le r \\ \lambda^{\beta((r/\alpha)-1)} & \alpha > r, \end{cases} \tag{21}$$

$$\mathscr{E}(f_\lambda) - \mathscr{E}(f_\rho) \le \begin{cases} \kappa^{2(r-\alpha)} \left\|L_K^{-r} f_\rho\right\|_\rho^2 \lambda^{2\beta} & \alpha \le r \\ \left\|L_K^{-r} f_\rho\right\|_\rho^2 \lambda^{2r(\beta/\alpha)} & \alpha > r. \end{cases} \tag{22}$$

For the term $\lambda \|f_\lambda\|_K^p$, we have the following inequality as $\|f\|_K = \|L_K^{-1/2} f\|_\rho$:

$$\begin{aligned} \|f_\lambda\|_K^p &= \lambda \left\|L_K^{-1/2} f_\lambda\right\|_\rho^p \\ &= \left\|L_K^{-1/2}\left(L_K^\alpha + \lambda^\beta\right)^{-1} L_K^\alpha f_\rho\right\|_\rho^p \\ &\le \left\|\left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^{\alpha+r-1/2}\right\|^p \left\|L_K^{-r} f_\rho\right\|_\rho^p \\ &= \left\|L_K^{-r} f_\rho\right\|_\rho^p \left(\sup_{i\ge 1} \frac{\mu_i^{\alpha+r-1/2}}{\mu_i^\alpha + \lambda^\beta}\right)^p. \end{aligned} \tag{23}$$

This means

$$\lambda \left\|f_\lambda\right\|_K^p \le \begin{cases} \kappa^{(r-1/2)p} \left\|L_K^{-r} f_\rho\right\|_\rho^p \lambda & r > \dfrac{1}{2} \\ \left\|L_K^{-r} f_\rho\right\|_\rho^p \lambda^{(r-1/2)p(\alpha/\beta)+1} & r \le \dfrac{1}{2}. \end{cases} \tag{24}$$

To minimize the sum of upper bounds (22) and (24) is the same to maximize the power of λ. We can choose

$$\begin{aligned} &\alpha = \alpha_{r,p} \le r, \quad \beta = \beta_{r,p} \ge \frac{1}{2}, \quad r > \frac{1}{2}; \\ &\alpha = \alpha_{r,p} = r, \quad \beta = \beta_{r,p} = \frac{2r}{4r + (1-2r)\,p}, \quad r \le \frac{1}{2}. \end{aligned} \tag{25}$$

Then,

$$D(\lambda) \leq \begin{cases} \left(\kappa^{2(r-\alpha)} + \kappa^{(r-1/2)p}\right) \times \left(\left\|L_K^{-r} f_\rho\right\|_\rho^2 + \left\|L_K^{-r} f_\rho\right\|_\rho^p\right) \lambda, & r > \frac{1}{2}; \\ \left(\kappa^{2(r-\alpha)} + 1\right) \times \left(\left\|L_K^{-r} f_\rho\right\|_\rho^2 + \left\|L_K^{-r} f_\rho\right\|_\rho^p\right) \lambda^{4r/(4r+(1-2r)p)}, & r \leq \frac{1}{2}. \end{cases} \tag{26}$$

This proves the result with

$$C_{r,p} = \left(\kappa^{2(r-\alpha)} + \kappa^{(r-1/2)p} + 1\right)\left(\left\|L_K^{-r} f_\rho\right\|_\rho^2 + \left\|L_K^{-r} f_\rho\right\|_\rho^p\right). \tag{27}$$

□

Remark 7. Another choice

$$\begin{aligned} &\alpha_{r,p} > r, \quad \beta_{r,p} > \frac{\alpha_{r,p}}{2r}, \quad r > \frac{1}{2}; \\ &\alpha_{r,p} > r, \quad \beta_{r,p} = \frac{2\alpha_{r,p}}{4r + (1-2r)\,p}, \quad r \leq \frac{1}{2} \end{aligned} \tag{28}$$

can also lead to the same result except for the constants.

Remark 8. In the case $p = 2$, our result turns to $D(\lambda) \leq C_r \lambda^{\min\{2r,1\}}$ which is consistent with the classical one [4]. In fact, for a general $p \leq 2$ of interest, the bound is better than $\min\{2r, 1\}$ since $\xi \geq 2r$, while $r \leq 1/2$.

In [7], the authors construct a function based on the generalized Fourier expansion of f_ρ and derive that $D(\lambda) \leq C_1 \lambda^{2r/(r+2)}$ with some constant C_1 for any $0 < r \leq 2$. The rate is always much less than $2r$ and cannot achieve 1 when $1/2 < r \leq 2$. On the other hand, our result is better than $2r$, while $0 < r < 1/2$.

Compared with [19], we get the same rate of upper bound. There, the authors find a connection between $\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^p$ and $\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^p$ with different p, q. However, their analysis needs an existent result, while our method does not.

5. Sample Error

There are a vast number of literatures studying the sample error. Here, we will follow the analysis of [11]. Firstly, we should introduce the Bernstein inequality [20]. Denote $\mathbb{E}g = \int_Z g(z) d\rho$, $\mathbb{E}_{\mathbf{z}} g = (1/m) \sum_{i=1}^m g(z_i)$, and $\sigma^2(g) = \mathbb{E}g^2 - (\mathbb{E}g)^2$ for an integral function g on Z.

Lemma 9. *Assume $|g - \mathbb{E}g| \leq M_g$ for some constant M_g almost surely. Then,*

$$Prob_{\mathbf{z} \in Z^m}\left\{\mathbb{E}_{\mathbf{z}} g - \mathbb{E}g \geq \varepsilon\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2(g) + (1/3) M_g \varepsilon\right)}\right\} \tag{29}$$

for any $\varepsilon > 0$.

Now, we can obtain the sample error bound involving f_λ.

Proposition 10. *Assume (3), for any $0 < \delta < 1$, with confidence $1 - (\delta/2)$; there holds*

$$\begin{aligned} S_2 &= \left(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right) - \left(\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\right) \\ &\leq \frac{32M^2 + 3}{3m} \log\frac{2}{\delta} + 16M^2 \log\frac{2}{\delta}\left(\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\right). \end{aligned} \tag{30}$$

Proof. Let

$$\begin{aligned} g_\lambda(z) &= \left(f_\lambda(x) - y\right)^2 - \left(f_\rho(x) - y\right)^2 \\ &= \left(f_\lambda(x) - f_\rho(x)\right)\left(f_\lambda(x) + f_\rho(x) - 2y\right); \end{aligned} \tag{31}$$

we have $S_2 = \mathbb{E}_{\mathbf{z}} g_\lambda - \mathbb{E}g_\lambda$. Note that $|f_\rho(x)| = |\int_Y y d\rho(y \mid x)| \leq M$ and

$$\begin{aligned} \left\|f_\lambda(x)\right\|_\infty &= \left\|\left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^\alpha f_\rho(x)\right\|_\infty \\ &\leq \left\|\left(L_K^\alpha + \lambda^\beta I\right)^{-1} L_K^\alpha\right\| \cdot \left\|f_\rho\right\|_\infty \\ &= \sup_{i \geq 1} \frac{\mu_i^\alpha}{\mu_i^\alpha + \lambda^\beta} \left\|f_\rho\right\|_\infty \leq \left\|f_\rho\right\|_\infty \leq M. \end{aligned} \tag{32}$$

It is easy to see that

$$\begin{aligned} \left|g_\lambda(z)\right| &= \left|f_\lambda(x) - f_\rho(x)\right| \left|f_\lambda(x) + f_\rho(x) - 2y\right| \\ &\leq 8M^2. \end{aligned} \tag{33}$$

Then, $|g_\lambda - \mathbb{E}g_\lambda| \leq 16M^2$ and

$$\begin{aligned} \sigma^2(g_\lambda) \leq \mathbb{E}g_\lambda^2 &\leq 16M^2 \int_X \left(f_\lambda(x) - f_\rho(x)\right)^2 d\rho_X \\ &\leq 16M^2\left(\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\right). \end{aligned} \tag{34}$$

By Bernstein inequality,

$$\begin{aligned} &\mathrm{Prob}_{\mathbf{z} \in Z^m}\left\{\mathbb{E}_{\mathbf{z}} g_\lambda - \mathbb{E}g_\lambda \geq \varepsilon\right\} \\ &\quad \leq \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2(g_\lambda) + (1/3)\, 16M^2\varepsilon\right)}\right\} \end{aligned} \tag{35}$$

holds with $M_g = 16M^2$. Set the right-hand side to be $\delta/2$; we can solve ε and the following bound

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} g_\lambda - \mathbb{E}g_\lambda &\leq \frac{32M^2}{3m} \log\frac{2}{\delta} + \sqrt{\frac{2\sigma^2(g_\lambda)}{m}} \log\frac{2}{\delta} \\ &\leq \frac{32M^2 + 3}{3m} \log\frac{2}{\delta} \\ &\quad + 16M^2 \log\frac{2}{\delta}\left(\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\right). \end{aligned} \tag{36}$$

This proves the proposition. □

For the sample error term S_1, it is more difficult since it involves the function $f_{\mathbf{z},\lambda}$ which varies, while the sample size m is different. So, we need a concentration inequality for a set of functions as in [21]. By setting $\tau = 1$, the inequality becomes as follows.

Lemma 11. *Let $\mathcal{F}$ be a set of measurable functions on Z, and $B_1, B_2 > 0$ is constant such that each function $f \in \mathcal{F}$ satisfies $\|f\|_\infty \le B_1$ and $\mathbb{E}(f^2) \le B_2 \mathbb{E}f$. If for some $a > 0$ and $0 < s < 2$,*

$$\log \mathcal{N}_2(\mathcal{F}, \varepsilon) \le a\varepsilon^{-s}, \quad \forall \varepsilon > 0, \tag{37}$$

then there exists a constant c_s' depending only on s such that, for any $\delta > 0$, with probability at least $1 - \delta$, there holds

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^{m} f(z_i) - \mathbb{E}f \\ &\quad \le \frac{1}{2}\mathbb{E}f + c_s'\eta'\left(\frac{a}{m}\right)^{2/(2+s)} + \frac{2B_2 + 18B_1}{m}\log\frac{1}{\delta} \\ &\qquad\qquad \forall f \in \mathcal{F}, \end{aligned} \tag{38}$$

where $\eta' := \max\{B_2^{(2-s)/(2+s)}, B_1^{(2-s)/(2+s)}\}$.

The result will be used to estimate S_1. We apply this lemma to the function set

$$\begin{aligned} \mathcal{G} = \Big\{ & g_{\pi,f}(z) \\ & = \left(f_\rho(x) - \pi(f(x))\right) \\ & \quad \times \left(\pi(f(x)) + f_\rho(x) - 2y\right) : f \in B_R(\mathcal{H}_K) \Big\} \end{aligned} \tag{39}$$

and have the following proposition.

Proposition 12. *Let $\mathcal{G}$ be defined as above with some $R \ge 1$ satisfying $\|f_{\mathbf{z},\lambda}\|_K \le R$, whose expression will be given in the next section. Assume (3) and (7) hold. Then, we have*

$$\begin{aligned} S_1 \le{}& \frac{1}{2}\left(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\right) \\ &+ \left(c_s'(16c_sM^s)^{2/(2+s)} + 176M^2\right) R^{2s/(2+s)} m^{-2/(2+s)} \log\frac{2}{\delta} \end{aligned} \tag{40}$$

for some constant c_s' depending only on s with confidence $1 - (\delta/2)$.

Proof. From definition, we know that

$$\begin{aligned} S_1 &= \left(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\right) - \left(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} g_{\pi,\mathbf{z}}(z_i) - \int_Z g_{\pi,\mathbf{z}}(z)\, d\rho, \end{aligned} \tag{41}$$

where $g_{\pi,\mathbf{z}}(z) = (f_\rho(x) - \pi(f_{\mathbf{z},\lambda}(x)))(f_\rho(x) + \pi(f_{\mathbf{z},\lambda}(x)) - 2y)$ is an element of $\mathcal{G}$.

In the following, we verify the conditions for $\mathcal{G}$ in Lemma 11. For any function $g_{\pi,f}(z) \in \mathcal{G}$, it holds

$$\begin{aligned} \left|g_{\pi,f}(z)\right| &\le \left|f_\rho(x) - \pi(f(x))\right| \cdot \left|\pi(f(x)) + f_\rho(x) - 2y\right| \\ &\le 8M^2, \\ \mathbb{E}g_{\pi,f}^2 &\le 16M^2 \int_X \left(\pi(f(x)) - f_\rho(x)\right)^2 d\rho_X \\ &= 16M^2 \mathbb{E}g_{\pi,f}. \end{aligned} \tag{42}$$

On the other hand, for any $g_1, g_2 \in \mathcal{G}$ depending, respectively, on $f_1, f_2 \in \mathcal{H}_K$,

$$\begin{aligned} \left|g_1(z) - g_2(z)\right| &= \left|\left(\pi(f_2(x)) - y\right)^2 - \left(\pi(f_1(x)) - y\right)^2\right| \\ &= \left|\pi(f_2(x)) - \pi(f_1(x))\right| \\ &\quad \cdot \left|\pi(f_2(x)) + \pi(f_1(x)) - 2y\right| \\ &\le 4M\left|\pi(f_2(x)) - \pi(f_1(x))\right| \\ &\le 4M\left|f_2(x) - f_1(x)\right|. \end{aligned} \tag{43}$$

This means $\mathcal{N}_2(\mathcal{G}, \varepsilon) \le \mathcal{N}_2(B_R(\mathcal{H}_K), \varepsilon/4M)$ and

$$\begin{aligned} \log \mathcal{N}_2(\mathcal{G}, \varepsilon) &\le \log \mathcal{N}_2\left(B_R(\mathcal{H}_K), \frac{\varepsilon}{4M}\right) \\ &\le \log \mathcal{N}_2\left(B_1(\mathcal{H}_K), \frac{\varepsilon}{4MR}\right) \\ &\le c_s(4MR)^s \varepsilon^{-s}. \end{aligned} \tag{44}$$

Now, we can see from Lemma 11 that, with confidence $1 - (\delta/2)$, there holds

$$\begin{aligned} S_1 &\le \frac{1}{2}\mathbb{E}g_{\pi,\mathbf{z}} + c_s'(16c_sM^s)^{2/(2+s)} R^{2s/(2+s)} m^{-2/(2+s)} \\ &\quad + \frac{176M^2}{m}\log\frac{2}{\delta} \\ &\le \frac{1}{2}\left(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\right) \\ &\quad + \left(c_s'(16c_sM^s)^{2/(2+s)} + 176M^2\right) R^{2s/(2+s)} m^{-2/(2+s)} \log\frac{2}{\delta}. \end{aligned} \tag{45}$$

This proves the proposition. □

6. Total Error

Combining the regularization and sample error bounds, we can prove the main result as follows.

Proof of Theorem 4. By substituting the regularization error and sample error (in the error decomposition formula) with obtained bounds in the above two sections, we have

$$\begin{aligned}
&\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho\right)\\
&\quad\le\frac{1}{2}\left(\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho\right)\right)\\
&\qquad+\left(c_s'\left(16c_sM^s\right)^{2/(2+s)}+176M^2\right)R^{2s/(2+s)}m^{-2/(2+s)}\log\frac{2}{\delta}\\
&\qquad+\frac{32M^2+3}{3m}\log\frac{2}{\delta}\\
&\qquad+16M^2\log\frac{2}{\delta}\left(\mathscr{E}\left(f_\lambda\right)-\mathscr{E}\left(f_\rho\right)\right)+D\left(\lambda\right).
\end{aligned}\tag{46}$$

Note that $D(\lambda)=\mathscr{E}(f_\lambda)-\mathscr{E}(f_\rho)+\lambda\|f_\lambda\|_K^p$ and radius R is always larger than 1; the bound becomes

$$\begin{aligned}
&\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho\right)\\
&\le 2\left(c_s'\left(16c_sM^s\right)^{2/(2+s)}+187M^2+1\right)R^{2s/(s+2)}m^{-2/(2+s)}\log\frac{2}{\delta}\\
&\quad+2\left(16M^2+1\right)\log\frac{2}{\delta}C_{r,p}\lambda^{\xi},
\end{aligned}\tag{47}$$

where

$$\xi=\begin{cases}1 & r>\dfrac{1}{2}\\[2mm] \dfrac{4r}{4r+(1-2r)\,p} & r\le\dfrac{1}{2}.\end{cases}\tag{48}$$

From $\lambda\|f_{\mathbf{z},\lambda}\|_K^p\le\mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\lambda\|f_{\mathbf{z},\lambda}\|_K^p\le(1/m)\sum_{i=1}^m y_i^2\le M$, we have the bound for the radius: $R\le M^{1/p}\lambda^{-1/p}$, and the above inequality is now

$$\begin{aligned}
&\mathscr{E}\left(\pi\left(f_{\mathbf{z},\lambda}\right)\right)-\mathscr{E}\left(f_\rho\right)\\
&\quad\le 2\left(c_s'\left(16c_sM^s\right)^{2/(2+s)}+187M^2+1\right)\\
&\qquad\times M^{2s/(s+2)p}\lambda^{-2s/(s+2)p}m^{-2/(2+s)}\log\frac{2}{\delta}\\
&\qquad+2\left(16M^2+1\right)\log\frac{2}{\delta}C_{r,p}\lambda^{\xi}.
\end{aligned}\tag{49}$$

To balance the two terms, we choose

$$\lambda=m^{-2p/(2s+(s+2)p\xi)}\tag{50}$$

and the result is proved with constant

$$\begin{aligned}
C_{p,r,s,M}&=2\left(c_s'\left(16c_sM^s\right)^{2/(2+s)}+187M^2+1\right)M^{2s/(s+2)p}\\
&\quad+2\left(16M^2+1\right)C_{r,p}.
\end{aligned}\tag{51}$$

□

Remark 13. In [11], the authors also use l^2-empirical covering number and derive an optimal rate $(1/m)^{\min\{2r,1\}/(s+2)}$. Compared with their classical rate for squared K-norm regularization, our result also can achieve the best one $O_p(1/m)$, while s tends to 0. Though when $r\ge 1/2$, that is, $f_\rho\in\mathscr{H}_K$, our rate is worse than $m^{-2/(s+2)}$, we will get a better rate than $(1/m)^{2r}$ when $r\le p/2(p+s)$. Moreover, by the iteration technique [4], we can expect that the radius for $f_{\mathbf{z},\lambda}$ is close to the upper bound of f_λ, which leads to a sharper learning rate $m^{-2\xi^2/(2r(1-2r)s+\xi^2(s+2))}$. This is always better than $m^{2r/(s+2)}$ for any $r\le 1/2$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by NSF of China (Grant no. 11326096), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (no. 2013LYM 0089), Doctor Grants of Huizhou University (Grant no. C511.0206), Major Project of Chinese National Statistics Bureau (no. 2013LZ52), NSF of Guangdong Province in China (no. S2013010014601), "12.5" Planning Project of Common Construction Subject for Philosophical and Social Sciences in Guangdong (no. GD12XYJ18), Project of Science and Technology Innovation in Guangdong Education Department (no. 2013KJCX0175), and Planning Fund Project of Humanities and Social Science Research in Chinese Ministry of Education (no. 14YJAZH040).

References

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

[2] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[3] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[4] 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.

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

[6] C. Wang and D. Zhou, "Optimal learning rates for least squares regularized regression with unbounded sampling," *Journal of Complexity*, vol. 27, no. 1, pp. 55–67, 2011.

[7] Q. W. Xiao and D. X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and $l1$-regularizer," *Taiwanese Journal of Mathematics*, vol. 14, pp. 1821–1836, 2010.

[8] Y. Feng, "Least-squares regularized regression with dependent samples and q-penalty," *Applicable Analysis*, vol. 91, no. 5, pp. 979–991, 2012.

[9] D. X. Zhou, "The covering number in learning theory," *Journal of Complexity*, vol. 18, no. 3, pp. 739–767, 2002.

[10] D. X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1743–1752, 2003.

[11] Z. C. Guo and D. X. Zhou, "Concentration estimates for learning with unbounded sampling," *Advances in Computational Mathematics*, vol. 38, no. 1, pp. 207–223, 2013.

[12] L. Shi, "Learning theory estimates for coefficient-based regularized regression," *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 252–265, 2013.

[13] S. G. Lv, D. M. Shi, Q. Xiao, and M. S. Zhang, "Sharp learning rates of coefficient-based l_p-regularized regression with indefinite kernels," *Science China Mathematics*, vol. 56, no. 8, pp. 1557–1574, 2013.

[14] C. Wang and J. Cai, "Convergence analysis of coefficient-based regularization under moment incremental condition," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 12, no. 1, Article ID 1450008, 19 pages, 2014.

[15] 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.

[16] Q. Wu and D.-X. Zhou, "Learning with sample dependent hypothesis spaces," *Computers & Mathematics with Applications*, vol. 56, no. 11, pp. 2896–2907, 2008.

[17] L. Shi, Y. Feng, and D. Zhou, "Conce ntration estimates for learning with l^1-regularizer and data dependent h ypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.

[18] S. Smale and D. Zhou, "Estimating the approximation error in learning theory," *Analysis and Applications*, vol. 1, no. 1, pp. 17–41, 2003.

[19] I. Steinwart, D. Hush, and C. Scovel, "Optimal rates for regularized least squares regression," in *Proceedings of the 22nd Annual Conference on Learning Theory*, S. Dasgupta and A. Klivans, Eds., pp. 79–93, 2009.

[20] G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, vol. 57, pp. 33–45, 1962.

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