J. Math. Anal. Appl. 431 (2015) 1153–1171

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Constructive analysis for coefficient regularization regression algorithms ☆

Weilin Nie, Cheng Wang *

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

ARTICLE INFO

Article history:
Received 30 December 2014
Available online 9 June 2015
Submitted by U. Stadtmueller

Keywords:
Least squares regression
Coefficient regularization
Error decomposition
Constructive stepping-stone function

ABSTRACT

In this paper, we consider the least squares regression algorithm with a generalized coefficient regularization term. A novel error decomposition involving a constructive stepping-stone function is introduced. By choosing appropriate parameters for the constructive function we finally derive a satisfactory learning rate under some condition for the goal function and capacity of the hypothesis space.

1. Introduction

Learning algorithms aim to find some relationship between the input and the output from given sample set. A classical setting [2] for such problem can be described as follows. Denote X and Y as input space and output space, where X is a compact metric space and $Y = \mathbb{R}$ in regression problems. Assume ρ is a joint probability distribution on the sample space $Z := X \times Y$. ρ_X is its marginal distribution on X and $\rho(y|x)$ is the conditional distribution on Y given $X = x$. Then we have

$$\int_Z f(x,y)d\rho = \int_X \int_Y f(x,y)d\rho(y|x)d\rho_X.$$

For any function $f : X \to Y$ we use the least square loss $L_{ls}(f(x), y) = (f(x) - y)^2$ to evaluate the performance. Then the generalization error

☆ This work is supported by NSF of China (Grant Nos. 11326096, 11401247), 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 (Nos. S2013010014601, S2013010015940) and '12.5' Planning Project of Common Construction Subject for Philosophical and Social Sciences in Guangdong (No. GD12XYJ18).

* Corresponding author.
E-mail addresses: niewl@hzu.edu.cn (W. Nie), wangch@hzu.edu.cn (C. Wang).

http://dx.doi.org/10.1016/j.jmaa.2015.06.006
0022-247X/

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho$$

can measure the produced error over the whole sample space for f. Our goal is to find the regression function

$$f_\rho = \int_Y y d\rho(y|x) = \mathbb{E}(y|x),$$

which minimizes $\mathcal{E}(f)$. Since ρ is hard to be obtained in reality, algorithms are usually based on a sample $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ drawn from the distribution ρ. One of such algorithms is ERM (empirical risk minimization) learning scheme [15]

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2,$$

where $\mathcal{H}$ is the hypothesis function space. In learning theory, we often use RKHS $\mathcal{H}_K$ (reproducing kernel Hilbert space) for hypothesis space. That is, for a given Mercer kernel $K : X \times X \to \mathbb{R}$ which is continuous, symmetric and positive semi-definite (the induced matrix $(K(x_i, x_j))_{i=1}^n$ is positive semi-definite), denote $K_x(y) = K(x, y)$ for any $x, y \in X$, then

$$\mathcal{H}_K := \overline{span\{K_x,\ x \in X\}},$$

with inner product $(K_x, K_y)_K = K(x, y)$. Here we recall that the reproducing property is $(f, K_x)_K = f(x)$ for any $f \in \mathcal{H}_K$ and $\|f\|_\infty \le \kappa\|f\|_K$ where $\kappa = \sqrt{\sup_{x,y\in X} K(x, y)}$.

In this paper, we investigate ERM algorithm with a coefficient penalty term,

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_{K,\mathbf{z}}} \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda\Omega(f). \tag{1}$$

We follow the work of [22], in which the hypothesis space is a sample dependent function space

$$\mathcal{H}_{K,\mathbf{z}} = \left\{\sum_{i=1}^m c_i K_{x_i} : c_i \in \mathbb{R},\ i = 1, 2, \cdots, m\right\}.$$

Then the regularization term can be of the form $\Omega(f) = m^{p-1}\sum_{i=1}^m |c_i|^p$ for a function $f = \sum_{i=1}^m c_i K_{x_i}$.

One important advantage of the coefficient regularization algorithm is we can choose a more general kernel, such as indefinite kernels, than that we did in the classical K-norm regularization. Note that $\|f\|_K^2 = \mathbf{c}^T K_{\mathbf{x}}\mathbf{c}$, $\mathbf{c} = (c_1, \cdots, c_m)$ for $f = \sum_{i=1}^m c_i K_{x_i}$, here the kernel matrix $K_{\mathbf{x}} = (K(x_i, x_j))_{i,j=1}^m$ is positive definite while choosing positive definite kernel K. When K is indefinite one, it is not suitable to use such regularization term since $K_{\mathbf{x}}$ may not be positive definite and $\mathbf{c}^T K_{\mathbf{x}}\mathbf{c}$ may be negative. Instead, ℓ^2 norm of $\mathbf{c}$ can still remain positive. We refer to [14] for a detail analysis. It is an interesting work to extend our work in this paper to the algorithms with general indefinite kernels and the MEE (minimum error entropy) algorithms with regularization, which has been studied in [6].

There are already a large number of literature studying such algorithms. [23,9] and [13,8] respectively discussed the case $p = 1$ and $p = 2$. For the algorithms with $p = \frac{1}{2}$, analysis can be found in [24] and etc. In [4,7], the authors did a lot of work for a general $1 \le p \le 2$. In this paper, we extend our previous work in [16], and conduct an error analysis based on a constructive stepping-stone functions for the general $p \in [1, 2]$.

2. Main result

Throughout the paper, we assume

$$|y| \le M \tag{2}$$

for some constant $M > 0$ almost surely. It is natural to apply a projection technique to improve the learning rate [12]. The projection operator π on the space of measurable function $f : X \to R$ is

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

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

$$L_K f(x) = \int_X f(t)K(x,t)d\rho_X(t)$$

is also important in learning theory and our analysis. It has been studied in [10]. In [2], the authors claimed that for a Mercer Kernel K, the associated L_K is a compact operator with non-increasing positive eigenvalue sequence μ_i. And the induced fractional operator

$$L^r_K f(x) = \sum_{i\ge1} \mu_i^r \phi_i(x)$$

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 $L^2_{\rho_X}$.

For the hypothesis space, we will use covering number to describe the capacity.

Definition 1. Let $(\mathcal{M}, d)$ be a pseudo-metric space and $S \subset \mathcal{M}$. For $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ 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,

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

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

When metric d is $\|\cdot\|_\infty$, i.e., $d(f,g) = \|f-g\|_\infty$, it is the classical uniform covering number. It is widely used in [19,23,18,3] and etc., more detailed analysis can be found in [25,26]. More recent references [5,8,7,16] 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

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

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

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

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

$$\log \mathcal{N}_2(B_1(\mathcal{H}_K), \varepsilon) \leq c_s \varepsilon^{-s}, \quad \forall \varepsilon > 0, \tag{3}$$

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

Our main result on learning rate for algorithm (1) is stated as follows.

Theorem 1. *Assume (2), (3) 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 (1). Then by choosing $\lambda = \left(\frac{1}{m}\right)^\alpha$ for*

$$\alpha = \begin{cases} \frac{2pr - p^2 r + p^2}{2pr + 2sr + sp}, & r < 1, \\ \frac{2p}{2s + (2+s)p}, & r \geq 1, \end{cases}$$

with confidence $1 - \delta$ for any $0 < \delta < 1$, we have

$$\|\pi(f_{\mathbf{z},\lambda}) - f_\rho\|_\rho^2 \leq 8\tilde{C} \left(\frac{1}{m}\right)^\eta \log^3 \frac{10}{\delta} \tag{4}$$

for some constant $\tilde{C}$ independent with m or δ and

$$\eta = \begin{cases} \min\left\{r, \frac{2pr}{2pr + 2sr + sp}\right\}, & r < 1, \\ \frac{2p}{2p + 2s + sp}, & r \geq 1. \end{cases}$$

Remark 1. For $r \geq 1$, compared with the classical ℓ^2 empirical learning algorithm, i.e., $p = 2$ in our result, the learning rate is $\frac{1}{1+s}$. This is very close as in [8]. However, in our analysis we abandoned the tedious iteration process. And for general $p \in [1, 2]$, our rate $\frac{2p}{2s+(2+s)p}$ tends to 1 while s tends to 0. In the case of $r < 1$, we still assume $p = 2$ here. Our rate is $\min\left\{r, \frac{2r}{2r+s+sr}\right\}$. This will be better than $\frac{2r}{2+s}$ in the classical analysis such as [17] when r is small, or precisely, $r \leq \frac{2}{2+s}$. As for $p = 1$ and $r \leq 1$, which is considered as a sparsity learning algorithm, [9] proposed that the learning rate can achieve $\frac{4r}{(2+s)(r+3)}$. However, we here obtained a sharper rate $\min\left\{r, \frac{2r}{s+2r+2sr}\right\}$ without iteration.

Remark 2. The power 3 in the term $\log^3 \frac{10}{\delta}$ will lead to a large quantity when δ is small. This always happens in the literature of learning theory, such as [9,8,7] and etc. This will not affect our error bound too much since we usually consider a sufficiently large sample size m. Still, it might be an interesting problem of how to reduce the power in our future work. Secondly, the projection technique makes the infinite norm of $f_{\mathbf{z},\lambda}$ bounded, which reduces the upper bound for sample error. On the other hand, since projection only bounds the infinite norm, it might be possible to additionally use the iteration technique as [19] for the K norm of $f_{\mathbf{z},\lambda}$ to get a sharper learning rate.

3. Error decomposition

Error decomposition can be regarded as the key point in our error analysis. Such technique can be found in [1,11,20] and etc. It involves a stepping-stone function which was firstly introduced in [21]. From the analysis

in [19] we can see that this function can be arbitrarily chosen in $\mathcal{H}_K$ which depends on λ and converges to f_ρ while λ tends to 0. This idea and analysis in [16] motivated our previous work [17] for algorithms with K-norm regularization. Now we apply the same idea to a general coefficient regularization algorithm. Let g_s be a function in $\mathcal{H}_{K,\mathbf{z}}$ to be determined in the following, and denote $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2$ for a function $g : Z \to \mathbb{R}$. Then

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_{K,\mathbf{z}}} \mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega(f)$$

and

$$\begin{aligned}
\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) &\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) + \lambda\Omega(f_{\mathbf{z},\lambda}) \\
&\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \\
&\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \\
&\le \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \mathcal{E}_{\mathbf{z}}(f_s) + \lambda\Omega(f_s) - \mathcal{E}(f_\rho) \\
&\le S_1 + S_2 + D(\lambda)
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \Big(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\Big) - \Big(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho)\Big), \\
S_2 &= \Big(\mathcal{E}_{\mathbf{z}}(f_s) - \mathcal{E}_{\mathbf{z}}(f_\rho)\Big) - \Big(\mathcal{E}(f_s) - \mathcal{E}(f_\rho)\Big), \\
D(\lambda) &= \mathcal{E}(f_s) - \mathcal{E}(f_\rho) + \lambda\Omega(f_s).
\end{aligned}$$

Here S_1 and S_2 are sample error which can be dealt with by some concentration inequality, while $D(\lambda)$ is regularization error. The decomposition is almost the same as classical one except for $\Omega(f_s)$ in $D(\lambda)$. It is constituted by two parts – hypothesis error and drift error in [8] and some related paper. Here we introduce a constructive function f_s and change the form to the classical one for simplicity in expression. In our coefficient regularization algorithm, it may be difficult to find a minimizer of $D(\lambda)$, as the authors did in [19]. However, notice that for any $f_s \in \mathcal{H}_{K,\mathbf{z}}$ satisfying $f_s \to f_\rho$ while $\lambda \to 0$ in some sense, it can be taken as a stepping-stone function in the above error decomposition. This induces our construction approach for f_s.

In the sequel we denote

$$g_s = (L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho, \tag{5}$$

and the stepping stone function

$$f_s = \sum_{i=1}^m \frac{1}{m} g_s(x_i) K_{x_i}. \tag{6}$$

When u here equals to 1, it turns to be the previous stepping stone function $f_s = \sum_{i=1}^m (L_K + \lambda I)^{-1} f_\rho(x_i) K_{x_i}$ as [16]. Compared with the original one, $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$ as in [19,10] and etc., our constructive function includes a parameter u which can be tuned for different conditions, which may lead to a better learning rate.

4. Regularization error

We devote this section to the regularization error $D(\lambda)$. Denote $\|f\|_\rho^2 := \|f\|_{L^2_{\rho_X}}^2 = \int_X |f(x)|^2 d\rho_X$. Since $f_s \in \mathcal{H}_{K,\mathbf{z}}$, we can choose $f_s = \sum_{i=1}^m \frac{1}{m} g_s(x_i) K_{x_i}$ with some $g_s : X \to \mathbb{R}$, then

$$\begin{aligned} D(\lambda) &= \|f_s - f_\rho\|_\rho^2 + \lambda\Omega(f_s) \\ &\le 2\left\| \frac{1}{m}\sum_{i=1}^m g_s(x_i)K_{x_i} - L_K g_s \right\|_\rho^2 + 2\|L_K g_s - f_\rho\|_\rho^2 \\ &\quad + \lambda\left(\frac{1}{m}\sum_{i=1}^m |g_s(x_i)|^p - \int_X |g_s(x)|^p d\rho_X \right) + \lambda \int_X |g_s(x)|^p d\rho_X. \end{aligned}$$

The first and third terms of the right hand side can be bounded by some kind of Bernstein type inequality. And the left two terms should also tend to 0 while $m \to \infty$ or $\lambda \to 0$. We can choose g_s in the form of $\Phi(\lambda, L_K) f_\rho$, where operator $\Phi(\lambda, L_K) \to L_K^{-1}$ while $\lambda \to 0$, with an $L^2_{\rho_X}$ norm upper bound depending on λ.

In the following, we assume $f_\rho = L_K^r g_\rho$ where $g_\rho = \sum_{i\ge1} \rho_i \phi_i$ where $\{\phi_i\}_{i\ge1}$ is the orthogonal basis of $L^2_{\rho_X}$. Then $\|L_K^{-r} f_\rho\|_\rho = \|g_\rho\|_\rho = \sqrt{\sum_{i\ge1} \rho_i^2}$. We recall an inequality proved in our previous work [17].

Lemma 1. *Letting $a, b, c, d > 0$ and $a < A$, we have*

$$\frac{a^c}{a^d + b} \le \begin{cases} b^{\frac{c}{d}-1}, & c < d, \\ A^{c-d}, & c \ge d. \end{cases}$$

Lemma 2. *Let $u, \mu > 0$ and define g_s and f_s by (5) and (6) respectively. Under the condition $f_\rho \in L_K^r(L^2_{\rho_X})$, there holds*

$$\lambda \int_X |g_s(x)|^p d\rho_X \le B_{\lambda,\mu,p,r} \|L_K^{-r} f_\rho\|_\rho^p,$$

where

$$B_{\lambda,\mu,p,r} = \begin{cases} \lambda\mu^{\frac{p(r-1)}{u}}, & r < 1, \\ \lambda\kappa^{2p(r-1)}, & r \ge 1. \end{cases}$$

Proof. From the notations introduced above,

$$\begin{aligned} \lambda \int_X |g_s(x)|^p d\rho_X &\le \lambda \left(\int_X |g_s(x)|^2 d\rho_X \right)^{\frac{p}{2}} \\ &= \lambda \|g_s\|_\rho^p = \lambda \|(L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho\|_\rho^p \\ &= \lambda \Big\| \sum_{i\ge1} \frac{\mu_i^{u+r-1}\rho_i}{\mu_i^u + \mu} \phi_i \Big\|_\rho^p = \lambda \left(\sum_{i\ge1} \frac{\mu_i^{2(u+r-1)}}{(\mu_i^u + \mu)^2} \rho_i^2 \right)^{\frac{p}{2}}. \end{aligned}$$

From Lemma 1 and the fact that $\|L_K\| \le \kappa^2$ [2] we have

$$\frac{\mu_i^{u+r-1}}{\mu_i^u + \mu} \le \begin{cases} \mu^{\frac{r-1}{u}}, & r < 1, \\ \kappa^{2(r-1)}, & r \ge 1, \end{cases}$$

and

$$\lambda \int_X |g_s(x)|^p d\rho_X \le \begin{cases} \lambda \mu^{\frac{p(r-1)}{u}} \|L_K^{-r} f_\rho\|_\rho^p, & r < 1, \\ \lambda \kappa^{2p(r-1)} \|L_K^{-r} f_\rho\|_\rho^p, & r \ge 1. \end{cases}$$

This proves the lemma. □

Next we should deduce the bound for the term $\|L_K g_s - f_\rho\|_\rho^2$.

Lemma 3. *Let $u > r, \mu > 0$ and define g_s and f_s by (5) and (6) respectively; we have*

$$\|L_K g_s - f_\rho\|_\rho^2 \le \mu^{\frac{2r}{u}} \|L_K^{-r} f_\rho\|_\rho^2.$$

Proof. Since $g_s = (L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho$,

$$\begin{aligned} \|L_K g_s - f_\rho\|_\rho^2 &= \| \sum_{i\ge1} \frac{\mu_i^{u+r}\rho_i}{\mu_i^u + \mu}\phi_i - \sum_{i\ge1} \mu_i^r \rho_i \phi_i\|_\rho^2 \\ &= \| \sum_{i\ge1} \frac{\mu\mu_i^r}{\mu_i^u + \mu I_i}\rho_i\phi_i\|_\rho^2 = \mu^2 \sum_{i\ge1} \frac{\mu_i^{2r}}{(\mu_i^u + \mu)^2}\rho_i^2 \\ &\le \mu^{\frac{2r}{u}} \sum_{i\ge1} \rho_i^2 \le \mu^{\frac{2r}{u}} \|L_K^{-r} f_\rho\|_\rho^2. \end{aligned}$$

The next to last inequality is from $u \ge r$. This is indeed the result of the lemma. □

For the first and third terms in the regularization error, we need Bernstein type inequalities [10] as follows.

Lemma 4. *Let H be a Hilbert space and ξ be a random variable on (Z, ρ) with values in H. Assume $\|\xi\| \le B < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}(\|\xi\|^2)$. Let $\{z_i\}_{i=1}^m$ be independent random drawers of ρ. For any $0 < \delta < 1$, with confidence $1 - \delta$,*

$$\left\| \frac{1}{m} \sum_{i=1}^m (\xi_i - \mathbb{E}\xi_i) \right\| \le \frac{2B}{m} \log \frac{2}{\delta} + \sqrt{\frac{2\sigma^2(\xi)}{m} \log \frac{2}{\delta}}.$$

Lemma 5. *Let $u, \mu > 0$ and define g_s and f_s by (5) and (6) respectively. For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{5}$, we have*

$$\frac{1}{m} \sum_{i=1}^m |g_s(x_i)|^p - \int_X |g_s(x)|^p d\rho_X \le B_{1,\mu} \cdot \left(2M^p + \sqrt{2}(\kappa^{2(r-1)} + 1) M^{p-1} \|L_K^{-r} f_\rho\|_\rho \right) \log \frac{10}{\delta},$$

where

$$B_{1,\mu} = \begin{cases} \left(\frac{1}{\mu^{\frac{p}{u}} m} + \frac{1}{\mu^{\frac{p-r}{u}} \sqrt{m}} \right), & r < 1, \\ \left(\frac{1}{\mu^{\frac{p}{u}} m} + \frac{1}{\mu^{p-1} \sqrt{m}} \right), & r \ge 1. \end{cases}$$

Proof. We firstly apply the above lemma to $\xi = |g_s(x)|^p$ on (X, ρ_X) with values in the Hilbert space $\mathbb{R}$. Then

$$\begin{aligned}\|\xi\|_\infty &= \|(L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho\|_\infty \le \left(\|(L_K^u + \mu I)^{-1} L_K^{u-1}\| \cdot \|f_\rho\|_\infty\right)^p \\ &\le \left(\sup_{i\ge 1} \frac{\mu_i^{u-1}}{\mu_i^u + \mu} M\right)^p \le \frac{1}{\mu^{\frac{p}{u}}} M^p\end{aligned}$$

and

$$\begin{aligned}\sigma^2(\xi) = \mathbb{E}|g_s|^{2p} &\le \|g_s\|_\infty^{2p-2} \|g_s\|_\rho^2 \\ &\le \mu^{\frac{2-2p}{u}} M^{2p-2} \cdot \begin{cases} \mu^{\frac{2(r-1)}{u}} \|L_K^{-r} f_\rho\|_\rho^2, & r < 1, \\ \kappa^{4(r-1)} \|L_K^{-r}\|_\rho^2, & r \ge 1 \end{cases} \\ &\le \begin{cases} \mu^{\frac{2(r-p)}{u}} M^{2p-2} \|L_K^{-r} f_\rho\|_\rho^{2p}, & r < 1, \\ \kappa^{4(r-1)} \mu^{\frac{2-2p}{u}} M^{2p-2} \|L_K^{-r} f_\rho\|_\rho^{2p}, & r \ge 1. \end{cases}\end{aligned}$$

By Bernstein type inequality with these conditions, we can see that

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^m (g_s(x_i) - \mathbb{E}g_s)\right| &= \left|\frac{1}{m}\sum_{i=1}^m g_s(x_i) - \mathbb{E}g_s\right| \\ &\le \begin{cases} \left(\frac{2M^p}{\mu^{\frac{p}{u}} m} + \frac{\sqrt{2} M^{p-1}}{\mu^{\frac{p-r}{u}} \sqrt{m}} \|L_K^{-r} f_\rho\|_\rho\right) \log \frac{10}{\delta}, & r < 1, \\ \left(\frac{2M^p}{\mu^{\frac{p}{u}} m} + \frac{\sqrt{2}\kappa^{2(r-1)} M^{p-1}}{\mu^{\frac{p-1}{u}} \sqrt{m}} \|L_K^{-r} f_\rho\|_\rho\right) \log \frac{10}{\delta}, & r \ge 1. \end{cases}\end{aligned}$$

This proves our result. □

Lemma 6. *Let $u, \mu > 0$ and define g_s and f_s by (5) and (6) respectively. For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{5}$, we have*

$$\left\|\frac{1}{m}\sum_{i=1}^m g_s(x_i) K_{x_i} - L_K g_s\right\|_\rho \le B_{2,\mu} \cdot \left(2\kappa^2 M + \sqrt{2}(\kappa + \kappa^r)\|L_K^{-r} f_\rho\|_\rho\right) \log \frac{10}{\delta},$$

where

$$B_{2,\mu} = \begin{cases} \left(\frac{1}{\mu^{\frac{1}{u}} m} + \frac{1}{\mu^{\frac{1-r}{u}} \sqrt{m}}\right), & r < 1, \\ \left(\frac{1}{\mu^{\frac{1}{u}} m} + \frac{1}{\sqrt{m}}\right), & r \ge 1. \end{cases}$$

Proof. We apply Lemma 4 to $\zeta_i = g_s(x_i) K_{x_i}$ on (X, ρ_X) with values in the Hilbert space $L^2_{\rho_X}$. Since

$$\begin{aligned}\|\zeta_i\|_\rho = \|g_s(x_i) K_{x_i}(x)\|_\rho &\le \kappa^2 \|g_s\|_\infty \\ &= \kappa^2 \|(L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho\|_\infty \le \frac{1}{\mu^{\frac{1}{u}}} \kappa^2 M,\end{aligned}$$

and

$$\sigma^2(\zeta_i) = \mathbb{E}\|\zeta_i\|_\rho^2 = \int_X \left[\int_X g_s^2(x_i)K_{x_i}^2(x)d\rho_X(x)\right] d\rho_X(x_i)$$
$$\leq \kappa^2 \int_X g_s^2(x_i)d\rho_X(x_i) = \kappa^2\|g_s\|_\rho^2 = \kappa^2\|(L_K^u + \mu I)^{-1}L_K^{u-1}f_\rho\|_\rho^2$$
$$\leq \begin{cases} \mu^{\frac{2(r-1)}{u}}\kappa^2\|L_K^{-r}f_\rho\|_\rho^2, & r < 1, \\ \kappa^{2r}\|L_K^{-r}f_\rho\|_\rho^2, & r \geq 1, \end{cases}$$

we have

$$\left\|\frac{1}{m}\sum_{i=1}^m(\zeta_i - \mathbb{E}\zeta_i)\right\| = \left\|\frac{1}{m}\sum_{i=1}^m g_s(x_i)K_{x_i} - L_K g_s\right\|_\rho$$
$$\leq \begin{cases} \left(\frac{2\kappa^2 M}{\mu^{\frac{1}{u}}m} + \frac{\sqrt{2}\kappa}{\mu^{\frac{1-r}{u}}\sqrt{m}}\|L_K^{-r}f_\rho\|_\rho\right)\log\frac{10}{\delta}, & r < 1, \\ \left(\frac{2\kappa^2 M}{\mu^{\frac{1}{u}}m} + \frac{\sqrt{2}\kappa^r}{\sqrt{m}}\|L_K^{-r}f_\rho\|_\rho\right)\log\frac{10}{\delta}, & r \geq 1. \end{cases}$$

This shows that our conclusion holds. □

Combining the lemmas above, we can derive a bound for the regularization error.

Proposition 1. *Let $u, \mu > 0$ and define g_s and f_s by (5) and (6) respectively. For any $0 < \delta < 1$, with confidence $1 - \frac{2}{5}\delta$, there holds*

$$D(\lambda) \leq C_{D,u,p,r}B_{\lambda,\mu}\log^2\frac{10}{\delta}$$

where

$$C_{D,u,p,r} = 1 + \kappa^{2(r-1)p} + \|L_K^{-r}f_\rho\|_\rho^2 + 2M^p + \sqrt{2}(\kappa^{2(r-1)} + 1)M^{p-1}\|L_K^{-r}f_\rho\|_\rho$$
$$+ \left(2\kappa^2 M + \sqrt{2}(\kappa + \kappa^r)\|L_K^{-r}f_\rho\|_\rho\right)^2$$

and

$$B_{\lambda,\mu} = \begin{cases} \frac{\lambda}{(\mu I)^{\frac{(1-r)p}{u}}} + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}}m} + \frac{\lambda}{\mu^{\frac{p-r}{u}}\sqrt{m}} + \frac{1}{\mu^{\frac{2}{u}}m^2} + \frac{1}{\mu^{\frac{2(1-r)}{u}}m}, & r < 1, \\ \lambda + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}}m} + \frac{\lambda}{\mu^{\frac{p-1}{u}}\sqrt{m}} + \frac{1}{\mu^{\frac{2}{u}}m^2} + \frac{1}{m}, & r \geq 1. \end{cases}$$

Proof. By the two lemmas above and decomposition of $D(\lambda)$ in the beginning of this section, we get that when $r < 1$,

$$D(\lambda) \leq \lambda\mu^{\frac{(r-1)p}{u}} + \mu^{\frac{2r}{u}}\|L_K^{-r}f_\rho\|_\rho^2$$
$$+ \left(2M^p + \sqrt{2}(\kappa^{2(r-1)} + 1)M^{p-1}\|L_K^{-r}f_\rho\|_\rho\right)\log\frac{10}{\delta}\cdot\left(\frac{\lambda}{\mu^{\frac{p}{u}}m} + \frac{\lambda}{\mu^{\frac{p-r}{u}}\sqrt{m}}\right)$$
$$+ \left(2\kappa^2 M + \sqrt{2}(\kappa + \kappa^r)\|L_K^{-r}f_\rho\|_\rho\right)^2\log^2\frac{10}{\delta}\cdot\left(\frac{1}{\mu^{\frac{2}{u}}m^2} + \frac{1}{\mu^{\frac{2(1-r)}{u}}m}\right)$$

$$\leq \left[1 + \|L_K^{-r} f_\rho\|_\rho^2 + 2M^p + \sqrt{2}(\kappa^{2(r-1)} + 1)M^{p-1}\|L_K^{-r} f_\rho\|_\rho + \left(2\kappa^2 M + \sqrt{2}(\kappa + \kappa^r)\|L_K^{-r} f_\rho\|_\rho\right)^2\right] \log^2 \frac{10}{\delta} \cdot \left(\frac{\lambda}{\mu^{\frac{(1-r)p}{u}}} + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}} m} + \frac{\lambda}{\mu^{\frac{p-r}{u}}\sqrt{m}} + \frac{1}{\mu^{\frac{2}{u}} m^2} + \frac{1}{\mu^{\frac{2(1-r)}{u}} m}\right)$$

and for $r \geq 1$,

$$\begin{aligned} D(\lambda) &\leq \lambda\kappa^{2(r-1)p} + \mu^{\frac{2r}{u}}\|L_K^{-r} f_\rho\|_\rho^2 \\ &\quad + \left(2M^p + \sqrt{2}(\kappa^{2(r-1)} + 1)M^{p-1}\|L_K^{-r} f_\rho\|_\rho\right) \log \frac{10}{\delta} \cdot \left(\frac{\lambda}{\mu^{\frac{p}{u}} m} + \frac{\lambda}{\mu^{\frac{p-1}{u}}\sqrt{m}}\right) \\ &\quad + \left(2\kappa^2 M + \sqrt{2}(\kappa + \kappa^r)\|L_K^{-r} f_\rho\|_\rho\right)^2 \log^2 \frac{10}{\delta} \cdot \left(\frac{1}{\mu^{\frac{2}{u}} m^2} + \frac{1}{m}\right) \\ &\leq \left[\kappa^{2(r-1)p} + \|L_K^{-r} f_\rho\|_\rho^2 + 2M^p + \sqrt{2}(\kappa^{(r-1)p} + 1)\|L_K^{-r} f_\rho\|_\rho + \left(2\kappa M + \sqrt{2}(1 + \kappa^r)\|L_K^{-r} f_\rho\|_\rho\right)^2\right] \log^2 \frac{10}{\delta} \\ &\quad \cdot \left(\lambda + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}} m} + \frac{\lambda}{\mu^{\frac{p-1}{u}}\sqrt{m}} + \frac{1}{\mu^{\frac{2}{u}} m^2} + \frac{1}{m}\right). \end{aligned}$$

This verifies the proposition. □

5. Sample error

A vast amount of literature concentrate on the sample error estimation. Here we will follow the work of [5]. Since the functions f_s and $f_{\mathbf{z},\lambda}$ vary while the sample size m is different, we need a concentration inequality for a set of functions like in [20]. By setting $\tau = 1$ the inequality becomes

Lemma 7. *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 $\|f\|_\infty \leq B_1$ and $\mathbb{E}(f^2) \leq B_2 \mathbb{E}f$. If for some $a > 0$ and $0 < s < 2$,*

$$\log \mathcal{N}_2(\mathcal{F}, \varepsilon) \leq a\varepsilon^{-s}, \quad \forall \varepsilon > 0, \tag{7}$$

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

$$\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathbb{E}f \leq \frac{1}{2}\mathbb{E}f + c_s'\eta'\left(\frac{a}{m}\right)^{\frac{2}{2+s}} + \frac{2B_2 + 18B_1}{m}\log\frac{1}{\delta}, \quad \forall f \in \mathcal{F},$$

where $\eta' := \max\left\{B_2^{\frac{2-s}{2+s}}, B_1^{\frac{2-s}{2+s}}\right\}$.

The result will be used to estimate S_1 and S_2. Firstly we apply this lemma to the function set

$$\mathcal{G} = \left\{g_{\pi,f}(z) = \Big(f_\rho(x) - \pi(f(x))\Big)\Big(\pi(f(x)) + f_\rho(x) - 2y\Big) : f \in B_R(\mathcal{H}_K)\right\},$$

and have the following proposition.

Proposition 2. *Let $\mathcal{G}_1$ be defined as above with some $R \geq 1$. Assume (2) and (3) hold. Then with confidence $1 - \frac{\delta}{10}$, we have*

$$S_1 \leq \frac{1}{2}\Big(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\Big) + C_{s1} R^{\frac{2s}{2+s}} m^{-\frac{2}{2+s}} \log \frac{10}{\delta}$$

where $C_{s1} = c'_s (16 c_s M^s)^{\frac{2}{2+s}} + 176M^2$.

Proof. From notations introduced above we know that

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

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 $\mathcal{G}_1$. In the following, we verify the conditions for $\mathcal{G}_1$ in Lemma 7. For any function $g_{\pi,f}(z) \in \mathcal{G}_1$, it holds

$$|g_{\pi,f}(z)| \leq |f_\rho(x) - \pi(f(x))| \cdot |\pi(f(x)) + f_\rho(x) - 2y| \leq 8M^2$$

and

$$\mathbb{E} g_{\pi,f}^2 \leq 16M^2 \int_X (\pi(f(x)) - f_\rho(x))^2 d\rho_X = 16M^2 \mathbb{E} g_{\pi,f}.$$

On the other hand, for any $g_1, g_2 \in \mathcal{G}_1$ depending respectively on $f_1, f_2 \in \mathcal{H}_K$,

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

This means $\mathcal{N}_2(\mathcal{G}_1, \varepsilon) \leq \mathcal{N}_2(B_R(\mathcal{H}_K), \frac{\varepsilon}{4M})$ and

$$\log \mathcal{N}_2(\mathcal{G}_1, \varepsilon) \leq \log \mathcal{N}_2\left(B_R(\mathcal{H}_K), \frac{\varepsilon}{4M}\right) \leq \log \mathcal{N}_2\left(B_1(\mathcal{H}_K), \frac{\varepsilon}{4MR}\right) \leq c_s (4MR)^s \varepsilon^{-s}.$$

Now we can see from Lemma 7 that with confidence $1 - \frac{\delta}{1}0$, there holds

$$\begin{aligned} S_1 &\leq \frac{1}{2}\mathbb{E} g_{\pi,\mathbf{z}} + c'_s (16 c_s M^s)^{\frac{2}{2+s}} R^{\frac{2s}{2+s}} m^{-\frac{2}{2+s}} + \frac{176M^2}{m} \log \frac{10}{\delta} \\ &\leq \frac{1}{2}\Big(\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho)\Big) + \Big(c'_s (16 c_s M^s)^{\frac{2}{2+s}} + 176M^2\Big) R^{\frac{2s}{2+s}} m^{-\frac{2}{2+s}} \log \frac{10}{\delta}. \end{aligned}$$

This proves the proposition. □

Now we will bound the error term S_2. To this end, we have to deduce upper bounds for $\|f_s\|_\infty$ and $\|f_s\|_K$ in probability. From the upper bound for $D(\lambda)$ in Proposition 2, we see that we should choose μ and u such that

$$\mu^{\frac{1}{u}} m \to \infty, \ \mu^{\frac{1-r}{u}} \sqrt{m} \to \infty. \tag{8}$$

Lemma 8. *Let $u, \mu > 0$ satisfying (8), and define g_s and f_s by (5) and (6) respectively. Then for any $0 < \delta < 1$, there exists a subset W_1 of Z^m with measure at least $1 - \frac{\delta}{5}$, when $\mathbf{z} \in W_1$*

$$\|f_s\|_\infty \leq B_{f,s} \log \frac{10}{\delta},$$

where $B_{f,s} := 2\kappa + \sqrt{2}(\kappa^{2r} + 1)\|L_K^{-r} f_\rho\|_\rho + M$.

Proof. Since $f_s = \frac{1}{m} \sum_{i=1}^m g_s(x_i) K_{x_i}$, we have

$$\|f_s\|_\infty \leq \left\| \frac{1}{m} \sum_{i=1}^m g_s(x_i) K_{x_i} - L_K g_s \right\|_\infty + \|L_K g_s\|_\infty.$$

The second term of right hand side is

$$\|(L_K^u + \mu I)^{-1} L_K^u f_\rho\|_\infty \leq M.$$

For the first term, we consider random variables $\xi = g_s(x) K_x$, note that $\|\xi\|_\infty = \|g_s(x) K_x\|_\infty \leq \kappa^2 \|(L_K^u + \mu I)^{-1} L_K^{u-1} f_\rho\|_\infty \leq \frac{\kappa^2}{\mu^{\frac{1}{u}}}$. And

$$\sigma^2(\xi) = \mathbb{E}\|g_s(x) K_x\|_\infty^2 \leq \kappa^4 \int_X g_s^2(x) d\rho_X = \kappa^4 \|g_s\|_\rho^2$$

$$\leq \begin{cases} \kappa^4 \mu^{\frac{2(r-1)}{u}} \|L_K^{-r} f_\rho\|_\rho^2, & r < 1, \\ \kappa^{4r} \|L_K^{-r} f_\rho\|_\rho^2, & r \geq 1. \end{cases}$$

From Lemma 4 we know that with confidence at least $1 - \frac{\delta}{5}$,

$$\left\| \frac{1}{m} \sum_{i=1}^m g_s(x_i) K_{x_i} - L_K g_s \right\|_\infty \leq \begin{cases} \left(\frac{2\kappa^2}{\mu^{\frac{1}{u}} m} + \frac{\sqrt{2}\kappa^2 \|L_K^{-r} f_\rho\|_\rho}{\mu^{\frac{1-r}{u}} \sqrt{m}} \right) \log \frac{10}{\delta}, & r < 1, \\ \left(\frac{2\kappa^2}{\mu^{\frac{1}{u}} m} + \frac{\sqrt{2}\kappa^{2r} \|L_K^{-r} f_\rho\|_\rho}{\sqrt{m}} \right) \log \frac{10}{\delta}, & r \geq 1. \end{cases}$$

So when $\mathbf{z} \in W_1$,

$$\|f_s\|_\infty \leq \begin{cases} \left(2\kappa + \sqrt{2}\|L_K^{-r} f_\rho\|_\rho + M\right) \log \frac{10}{\delta}, & r < 1, \\ \left(2\kappa + \sqrt{2}\kappa^{2r} \|L_K^{-r} f_\rho\|_\rho + M\right) \log \frac{10}{\delta}, & r \geq 1. \end{cases}$$ □

Lemma 9. *Let $u, \mu > 0$ satisfying (8), and define g_s and f_s by (5) and (6) respectively. For any $0 < \delta < 1$, there exists a subset W_2 of Z^m with measure at least $1 - \frac{\delta}{5}$, when $\mathbf{z} \in W_2$*

$$\|f_s\|_K \leq R_\delta := c_R m_\mu \log \frac{10}{\delta}$$

where

$$m_\mu = \mu^{\frac{r-1/2}{u}} + 1$$

and $c_R = (\kappa^{2r-1} + 1)\left(2\kappa + (\sqrt{2} + 1)(\kappa^{2r-1} + 1) + 1\right) \|L_K^{-r} f_\rho\|_\rho$.

Proof. The same as last lemma, we see that

$$\|f_s\|_K \le \left\| \frac{1}{m}\sum_{i=1}^{m} g_s(x_i)K_{x_i} - L_K g_s \right\|_K + \|L_K g_s\|_K.$$

For the term $\|L_K g_s\|_K$, since $\|f\|_K = \|L_K^{-\frac{1}{2}} f\|_{L_{\rho_X}} \le \|L_K^{-\frac{1}{2}} f\|_\rho$ we have

$$\begin{aligned}\|L_K g_s\|_K &= \|(L_K^u + \mu I)^{-1} L_K^u f_\rho\|_K = \|(L_K^u + \mu I)^{-1} L_K^{u-\frac{1}{2}} f_\rho\|_\rho \\ &\le \begin{cases} \mu^{\frac{r-1/2}{u}} \|L_K^{-r} f_\rho\|_\rho, & r < \frac{1}{2}, \\ \kappa^{2r-1} \|L_K^{-r} f_\rho\|_\rho, & r \ge \frac{1}{2}. \end{cases}\end{aligned}$$

Then we apply Lemma 4 to the random variable $\xi = g_s(x)K_x$ with K norm. As

$$\|g_s(x)K_x\|_K \le \kappa \|g_s\|_\infty \le \frac{\kappa}{\mu^{\frac{1}{u}}}$$

and

$$\begin{aligned}\sigma^2(\xi) = \mathbb{E}\|g_s(x)K_x\|_K^2 &\le \kappa^2 \int_X g_s^2(x) d\rho_X \le \kappa^2 \|g_s\|_\rho^2 \\ &\le \begin{cases} \kappa^2 \mu^{\frac{2(r-1)}{u}} \|L_K^{-r} f_\rho\|_\rho^2, & r < 1, \\ \kappa^{4r-2} \|L_K^{-r} f_\rho\|_\rho^2, & r \ge 1. \end{cases}\end{aligned}$$

From Lemma 4, there holds

$$\begin{aligned}\left\| \frac{1}{m}\sum_{i=1}^{m} g_s(x_i)K_{x_i} - L_K g_s \right\|_K &\le \begin{cases} \left(\frac{2\kappa}{\mu^{\frac{1}{u}} m} + \frac{\sqrt{2}\kappa \|L_K^{-r} f_\rho\|_\rho}{\mu^{\frac{1-r}{u}} \sqrt{m}} \right) \log \frac{10}{\delta}, & r < 1, \\ \left(\frac{2\kappa}{\mu^{\frac{1}{u}} m} + \frac{\sqrt{2}\kappa^{2r-1} \|L_K^{-r} f_\rho\|_\rho}{\sqrt{m}} \right) \log \frac{10}{\delta}, & r \ge 1 \end{cases} \\ &\le \left(2\kappa + \sqrt{2}(\kappa^{2r-1} + 1) \|L_K^{-r} f_\rho\|_\rho \right) \log \frac{10}{\delta}.\end{aligned}$$

Then we have when $\mathbf{z} \in W_2$,

$$\|f_s\|_K \le \begin{cases} \left(2\kappa + (\sqrt{2}\kappa^{2r-1} + \sqrt{2} + \mu^{\frac{r-1/2}{u}}) \|L_K^{-r} f_\rho\|_\rho \right) \log \frac{10}{\delta}, & r < \frac{1}{2}, \\ \left(2\kappa + ((\sqrt{2}+1)\kappa^{2r-1} + \sqrt{2}) \|L_K^{-r} f_\rho\|_\rho \right) \log \frac{10}{\delta}, & r \ge \frac{1}{2}. \end{cases}$$

And our lemma can be deduced. □

Now we can derive the bound for S_2.

Proposition 3. *Let $u, \mu > 0$ satisfying (8), and define g_s and f_s by (5) and (6) respectively. For any $0 < \delta < 1$, there exists a subset W_3 of Z^m with measure at least $1 - \frac{\delta}{10}$, when $\mathbf{z} \in W_1 \bigcap W_2 \bigcap Z_3$ where Z_2 and Z_3 are defined in Lemmas 8 and 9, there holds*

$$S_2 \le \frac{1}{2}\left(\mathcal{E}(f_s) - \mathcal{E}(f_\rho)\right) + C_{s2} m_\mu^{\frac{2s}{2+s}} m^{-\frac{2}{2+s}} \log^3 \frac{10}{\delta}$$

where $C_{s2} = (20 + c_s' + c_s^{\frac{2}{2+s}})(2B_{f,s} + 3M) C_R^{\frac{2s}{2+s}}$.

Proof. As in Proposition 2, denote

$$g_{f,s}(z) = (f_s(x) - y)^2 - (f_\rho(x) - y)^2 = (f_s(x) - f_\rho(x))(f_s(x) + f_\rho(x) - 2y).$$

Then it is easy to see that

$$\begin{aligned} S_2 &= (\mathcal{E}_{\mathbf{z}}(f_s) - \mathcal{E}_{\mathbf{z}}(f_\rho)) - (\mathcal{E}(f_s) - \mathcal{E}(f_\rho)) \\ &= \frac{1}{m}\sum_{i=1}^{m} g_{f,s}(z_i) - \int_Z g_{f,s}(z)d\rho. \end{aligned}$$

Since $\mathbf{z} \in W_2$ which indicates $\|f_s\|_K \le R_\delta$, we apply Lemma 7 to the function set

$$\mathcal{G}_2 = \{g_{f,s}(z) = (f_s(x) - f_\rho(x))\,(f_s(x) + f_\rho(x) - 2y) : f \in B_{R_\delta}(\mathcal{H}_K)\}.$$

Meanwhile, $\mathbf{z} \in W_1$, so

$$\|g_{f,s}\|_\infty \le (\|f_s\|_\infty + 3M)^2 \le (B_{f,s} + 3M)^2,$$

and

$$\begin{aligned} \mathbb{E}(g_{f,s}^2) &= \mathbb{E}\left((f_s(x) - f_\rho(x))^2(f_s(x) + f_\rho(x) - 2y)^2\right) \\ &\le (\|f_s\|_\infty + 3M)^2\mathbb{E}(f_s(x) - f_\rho(x))^2 = (B_{f,s} + 3M)^2\mathbb{E}g_{f,s}. \end{aligned}$$

For functions $f_{s_1} = \frac{1}{m}\sum_{i=1}^m g_s(x_{1,i})K_{x_i}$ and $f_{s_1} = \frac{1}{m}\sum_{i=1}^m g_s(x_{2,i})K_{x_i}$ satisfying $f_{s_1}, f_{s_2} \in B_{R_\delta}(\mathcal{H}_K)$, we denote $g_1 = (f_{s_1}(x) - y)^2 - (f_\rho(x) - y)^2$ and $g_2 = (f_{s_2}(x) - y)^2 - (f_\rho(x) - y)^2$, then $g_1, g_2 \in \mathcal{G}_2$. We have

$$\begin{aligned} |g_1(z) - g_2(z)| &= |(f_{s_1}(x) - y)^2 - (f_{s_2}(x) - y)^2| \\ &= |f_{s_1}(x) - f_{s_2}(x)| \cdot |f_{s_1}(x) + f_{s_2}(x) - 2y| \\ &\le 2(B_{f,s} + M)|f_{s_1}(x) - f_{s_2}(x)|. \end{aligned}$$

Therefore

$$\begin{aligned} \log \mathcal{N}_2(\mathcal{G}_2, \varepsilon) &\le \log \mathcal{N}_2\left(B_{R_\delta}(\mathcal{H}_K), \frac{\varepsilon}{2(B_{f,s} + M)}\right) \\ &\le \log \mathcal{N}_2\left(B_1(\mathcal{H}_K), \frac{\varepsilon}{2(B_{f,s} + M)R_\delta}\right) \le c_s(2(B_{f,s} + M)R_\delta)^s\varepsilon^{-s}. \end{aligned}$$

Now from Lemma 7 we have

$$\begin{aligned} S_2 &= \frac{1}{m}\sum_{i=1}^{m} g_{f,s}(z_i) - \mathbb{E}g_{f,s} \\ &\le \frac{1}{2}\mathbb{E}g_{f,s} + c_s'(B_{f,s}\log\frac{10}{\delta} + 3M)^{\frac{2(2-s)}{2+s}} c_s^{\frac{2}{2+s}} (2(B_{f,s}\log\frac{10}{\delta} + M)R_\delta)^{\frac{2s}{2+s}} \frac{1}{m^{\frac{2}{2+s}}} \\ &\quad + \frac{20(B_{f,s}\log\frac{10}{\delta} + 3M)^2}{m}\log\frac{10}{\delta} \\ &\le \frac{1}{2}(\mathcal{E}(f_s) - \mathcal{E}(f_\rho)) + c_s' c_s^{\frac{2}{2+s}} (2B_{f,s}\log\frac{10}{\delta} + 3M)^{\frac{4}{2+s}} R_\delta^{\frac{2s}{2+s}} \frac{1}{m^{\frac{2}{2+s}}} \end{aligned}$$

$$+\frac{20(B_{f,s}\log\frac{10}{\delta}+3M)^2}{m}\log\frac{10}{\delta}$$
$$\le\frac{1}{2}\left(\mathcal{E}(f_s)-\mathcal{E}(f_\rho)\right)+(20+c_s'+c_s^{\frac{2}{2+s}})(2B_{f,s}+3M)C_R^{\frac{2s}{2+s}}\dot{m}_\mu^{\frac{2s}{2+s}}m^{-\frac{2}{2+s}}\log^3\frac{10}{\delta}.$$

This proves the proposition. □

6. Total error

From the sections above, we have the bounds for sample error S_1, S_2 and regularization error $D(\lambda)$. Now we can give the proof of the total error bound by combining the three parts.

Proof of Theorem 1. Firstly we need an expression for the radius R in the bound of S_1. Recall that $B_R(\mathcal{H}_K)=\{f\in\mathcal{H}_K:\|f\|_K\le R\}$, it requires to find the upper bound for $\|f_{\mathbf{z},\lambda}\|_K$. Since $f_{\mathbf{z},\lambda}\in\mathcal{H}_{K,\mathbf{z}}$, we assume $f_{\mathbf{z},\lambda}=\sum_{i=1}^m c_{i,\mathbf{z}}K_{x_i}$. Then for $1<p\le 2$

$$\begin{aligned}\|f_{\mathbf{z},\lambda}\|_K&=\|\sum_{i=1}^m c_{i,\mathbf{z}}K_{x_i}\|_K\le\sum_{i=1}^m\|c_{i,\mathbf{z}}K_{x_i}\|_K\le\sum_{i=1}^m|c_{i,\mathbf{z}}|\cdot\|K_{x_i}\|_K\\&=\sum_{i=1}^m|c_{i,\mathbf{z}}|\cdot\sqrt{(K_{x_i},K_{x_i})_K}=\sum_{i=1}^m|c_{i,\mathbf{z}}|\cdot\sqrt{K(x_i,x_i)}\le\kappa\sum_{i=1}^m|c_{i,\mathbf{z}}|\\&\le\kappa\left(\sum_{i=1}^m|c_{i,\mathbf{z}}|^p\right)^{\frac{1}{p}}\cdot\left(\sum_{i=1}^m 1^q\right)^{\frac{1}{q}}=\kappa m^{\frac{1}{q}}\left(\sum_{i=1}^m|c_{i,\mathbf{z}}|^p\right)^{\frac{1}{p}}.\end{aligned}$$

Here q satisfies $\frac{1}{p}+\frac{1}{q}=1$, i.e., $\frac{1}{q}=\frac{p-1}{p}$. On the other hand, from the definition of $f_{\mathbf{z},\lambda}$ we can see that

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\lambda m^{p-1}\sum_{i=1}^m|c_{i,\mathbf{z}}|^p\le\mathcal{E}_{\mathbf{z}}(0)+0=\frac{1}{m}\sum_{i=1}^m y_i^2\le M^2.$$

Therefore $\lambda m^{p-1}\sum_{i=1}^m|c_{i,\mathbf{z}}|^p\le M^2$, which leads to $\sum_{i=1}^m|c_{i,\mathbf{z}}|^p\le\frac{M^2}{\lambda m^{p-1}}$. By substituting this upper bound to the above inequality, we have

$$\|f_{\mathbf{z},\lambda}\|_K\le\kappa M^{\frac{2}{p}}\lambda^{-\frac{1}{p}},$$

which indicates $R=\kappa M^{\frac{2}{p}}\lambda^{-\frac{1}{p}}$. It can be verified that the bound for $\|f_{\mathbf{z},\lambda}\|_K$ is in the same form when $p=1$. Note that $\mathcal{E}(f_s)-\mathcal{E}(f_\rho)$ in the bound for S_2 is indeed $\|f_s-f_\rho\|_\rho^2$, part of $D(\lambda)$. From Propositions 1, 2 and 3 we get that with confidence $1-\delta$,

$$\begin{aligned}\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f_\rho)&\le 3C_{D,u,p,r}B_{\lambda,\mu}\log^2\frac{10}{\delta}\\&\quad+2C_{s1}\kappa^{\frac{2}{2+s}}M^{\frac{4s}{(2+s)p}}\frac{1}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}}\log\frac{10}{\delta}+2C_{s2}m_\mu^{\frac{2s}{2+s}}m^{-\frac{2}{2+s}}\log^3\frac{10}{\delta}\\&\le(3C_{D,u,p,r}+2C_{s1}+2C_{s2})\cdot\left(B_{\lambda,\mu}+\frac{1}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}}+m_\mu^{\frac{2s}{2+s}}\frac{1}{m^{\frac{2}{2+s}}}\right)\log^3\frac{10}{\delta}.\end{aligned}$$

Denote

$$\tilde{C}=3C_{D,u,p,r}+2C_{s1}\kappa^{\frac{2}{2+s}}+2C_{s2}.$$

We will find the best learning rate for different r.

Case 1: When $0 < r \leq \frac{1}{2}$, we have

$$\begin{aligned}\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) &\leq \tilde{C}\log^3\frac{10}{\delta}\cdot\left(\frac{\lambda}{\mu^{\frac{(1-r)p}{u}}} + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}}m} + \frac{\lambda}{\mu^{\frac{p-r}{u}}\sqrt{m}}\right.\\ &\quad\left. + \frac{1}{\mu^{\frac{2}{u}}m^2} + \frac{1}{\mu^{\frac{2(1-r)}{u}}m} + \frac{1}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}} + \frac{1}{\mu^{\frac{(1-2r)s}{u(2+s)}}m^{\frac{2}{2+s}}} + \frac{1}{m^{\frac{2}{2+s}}}\right)\\ &\leq \tilde{C}\log^3\frac{10}{\delta}\left(\frac{\lambda}{\gamma^{(1-r)p}} + \mu^{2r} + \frac{\lambda}{\mu^p m} + \frac{\lambda}{\gamma^{p-r}\sqrt{m}} + \frac{1}{\gamma^2 m^2}\right.\\ &\quad\left. + \frac{1}{\gamma^{2(1-r)}m}m^{\frac{2}{2+s}} + \frac{1}{\mu^{\frac{(1-2r)s}{2+s}}m^{\frac{2}{2+s}}} + \frac{1}{\gamma^{\frac{(1-2r)s}{2+s}}m^{\frac{2}{2+s}}}\right)\\ &\leq \tilde{C}\log^3\frac{10}{\delta}\left(\frac{3\lambda}{\gamma^{(1-r)p}} + \gamma^{2r} + \frac{2}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}} + \frac{1}{\mu^{\frac{(1-2r)s}{2+s}}m^{\frac{2}{2+s}}}\right).\end{aligned}$$

Here $\gamma = \mu^{\frac{1}{u}}$. Note that we should choose μ such that $\mu \to 0$ and $\gamma^{2r} \geq \frac{1}{m}$ to maximize the learning rate, i.e., $\gamma \leq 1$ and $\gamma^r\sqrt{m} \geq 1$. Then

$$\frac{\lambda}{\mu^p m} = \frac{\lambda}{\mu^{p-r}\sqrt{m}}\cdot\frac{1}{\mu^r\sqrt{m}} \leq \frac{\lambda}{\mu^{p-r}}\frac{\lambda}{\gamma^{p-r}\sqrt{m}},$$
$$\frac{\lambda}{\gamma^{p-r}\sqrt{m}} = \frac{\lambda}{\gamma^{p-pr}}\cdot\frac{1}{\gamma^r\frac{\lambda}{\mu^{(1-r)p}}},$$

and

$$\frac{1}{\mu^2 m^2} = \frac{1}{\mu^{2(1-r)}m}\cdot\frac{1}{\mu^{2r}m} \leq \frac{1}{\mu^{2(1-r)}m}.$$

This leads to our last inequality above. Let $\lambda = m^{-\alpha}$ and $\gamma = m^{-\beta}$, this bound becomes

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) \leq 9\tilde{C}\log^3\frac{10}{\delta}\left(\frac{1}{m}\right)^\eta$$

where

$$\eta = \min\left\{\alpha - (1-r)p\beta, 2r\beta, 1-2(1-r)\beta, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}, \frac{2}{2+s} - \frac{(1-2r)s\beta}{2+s}\right\}.$$

Now we will choose appropriate α and β to maximize the rate η.

$$\begin{aligned}\eta_{max} &= \max_{\alpha,\beta}\min\left\{\alpha - (1-r)p\beta, 2r\beta, 1-2(1-r)\beta, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}, \frac{2}{2+s} - \frac{(1-2r)s\beta}{2+s}\right\}\\ &= \max_{\alpha}\min\left\{\max_{\beta}\min\{2r\beta, \alpha-(1-r)p\beta\}, \max_{\beta}\min\{2r\beta, 1-2(1-r)\beta\},\right.\\ &\qquad\left.\max_{\beta}\min\left\{2r\beta, \frac{2}{2+s} - \frac{(1-2r)s\beta}{2+s}\right\}, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\}\\ &= \max_{\alpha}\min\left\{\frac{2r\alpha}{(2-p)r+p}, r, \frac{4r}{1+4r-rs}, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\}\end{aligned}$$

$$= \min\left\{r, \max_{\alpha} \min\left\{\frac{2r\alpha}{(2-p)r+p}, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\}\right\}$$
$$= \min\left\{r, \frac{2pr}{2pr+2sr+sp}\right\}.$$

Here we choose $\alpha = \frac{2pr-p^2r+p^2}{2pr+2sr+sp}$ and

$$\beta = \begin{cases} \frac{1}{2}, & 0 < r < \min\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\}, \\ \frac{p}{2pr+2sr+sp}, & \min\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\} \le r \le \frac{1}{2}. \end{cases}$$

Case 2: $\frac{1}{2} < r < 1$. In this case, it is easy to see the analysis is almost the same as in Case 1. And we can finally deduce that

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) \le 9\tilde{C}\log^3\frac{10}{\delta}\left(\frac{1}{m}\right)^{\eta}$$

with

$$\eta = \min\left\{r, \frac{2pr}{2pr+2sr+sp}\right\},$$

$\alpha = \frac{2pr-p^2r+p^2}{2pr+2sr+sp}$ and

$$\beta = \begin{cases} \frac{1}{2}, & \frac{1}{2} < r < \max\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\}, \\ \frac{p}{2pr+2sr+sp}, & \max\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\} \le r \le 1. \end{cases}$$

Case 3: As the same analysis, when $r \ge 1$, there holds

$$\begin{aligned} &\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) \\ &\le \tilde{C}\log^3\frac{10}{\delta}\left(\lambda + \mu^{\frac{2r}{u}} + \frac{\lambda}{\mu^{\frac{p}{u}}m} + \frac{\lambda}{\mu}^{\frac{p-1}{u}} + \frac{1}{\mu^{\frac{2}{u}}m^2} + \frac{3}{m^{\frac{2}{2+s}}} + \frac{1}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}}\right) \\ &\le \tilde{C}\log^3\frac{10}{\delta}\left(\lambda + \gamma^{2r} + \frac{\lambda}{\gamma^p m} + \frac{\lambda}{\gamma^{p-1}\sqrt{m}} + \frac{1}{\gamma^2 m^2} + \frac{4}{\lambda^{\frac{2s}{(2+s)p}}m^{\frac{2}{2+s}}}\right) \\ &\le \tilde{C}\log^3\frac{10}{\delta}\left(3\lambda + \frac{1}{\gamma^2 m^2} + \frac{4}{\lambda^{\frac{2s}{(2+s)p}}}\right). \end{aligned}$$

The last inequality is from that fact that

$$\frac{\lambda}{\gamma^p m} = \frac{\lambda}{\gamma^{2r}m}\cdot\gamma^{2r-p} \le \frac{\lambda}{\gamma^{2r}m}$$

and

$$\frac{\lambda}{\gamma^{p-1}\sqrt{m}} = \frac{\lambda}{\gamma^r\sqrt{m}}\cdot\gamma^{r+1-p} \le \frac{\lambda}{\gamma^{2r}m}.$$

Then

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho) \le 9\tilde{C}\log^3\frac{10}{\delta}\left(\frac{1}{m}\right)^{\eta}$$

where

$$\eta = \min\left\{\alpha, 2r\beta, 2-2\beta, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\}.$$

Then

$$\begin{aligned}
\eta_{max} &= \max_{\alpha,\beta} \min\left\{\alpha, 2r\beta, 2-2\beta, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\} \\
&= \max_{\alpha} \min\left\{\max_{\beta} \min\{2r\beta, 2-2\beta\}, \alpha, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\} \\
&= \max_{\alpha} \min\left\{\frac{2r}{1+r}, \alpha, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\} \\
&= \min\left\{\frac{2r}{1+r}, \max_{\alpha}\left\{\alpha, \frac{2}{2+s} - \frac{2s\alpha}{(2+s)p}\right\}\right\} \\
&= \min\left\{\frac{2r}{1+r}, \frac{2p}{2s+(2+s)p}\right\} \\
&= \frac{2p}{2s+(2+s)p}.
\end{aligned}$$

Here $\beta = \frac{1}{1+r}$ and $\alpha = \frac{2p}{2s+(2+s)p}$. □

Remark 3. From the proof we can deduce the final condition for choosing parameters u and μ in our analysis. That is,

$$\mu^{\frac{1}{u}} = m^{-\tilde{\beta}}$$

where

$$\tilde{\beta} = \begin{cases} \frac{1}{2}, & 0 < r < \min\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\}, \\ \frac{p}{2pr+2sr+sp}, & \min\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\} \le r \le \frac{1}{2}, \\ \frac{1}{2}, & \frac{1}{2} < r < \max\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\}, \\ \frac{p}{2pr+2sr+sp}, & \max\left\{\frac{(2-s)p}{2(p+s)}, \frac{1}{2}\right\} \le r \le 1, \\ \frac{1}{1+r}, & r > 1, \end{cases}$$

which depends on the value of r.

References

[1] D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[2] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2002) 1–49.
[3] Y.L. Feng, Least-squares regularized regression with dependent samples and q-penalty, Appl. Anal. 91 (5) (2012) 979–991.
[4] Y.L. Feng, S.G. Lv, Unified approach to coefficient-based regularized regression, Comput. Math. Appl. 62 (2011) 506–515.
[5] Z.C. Guo, D.X. Zhou, Concentration estimates for learning with unbounded sampling, Adv. Comput. Math. 38 (2013) 207–223.
[6] T. Hu, J. Fan, Q. Wu, D.X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl. 13 (2015) 437–455.
[7] S.G. Lv, D.M. Shi, Q.W. Xiao, M.S. Zhang, Sharp learning rates of coefficient-based l^p-regularized regression with indefinite kernels, Sci. China Math. 56 (8) (2013) 1557–1574.

[8] L. Shi, Learning theory estimates for coefficient-based regularized regression, Appl. Comput. Harmon. Anal. 34 (2013) 252–265.
[9] L. Shi, Y.L. Feng, D.X. Zhou, Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces, Appl. Comput. Harmon. Anal. 31 (2011) 286–302.
[10] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their applications, Constr. Approx. 26 (2007) 153–172.
[11] S. Smale, D.X. Zhou, Online learning with Markov sampling, Anal. Appl. 7 (2009) 87–113.
[12] I. Steinwart, D. Hush, Clint Scovel, Optimal rates for regularized least squares regression, in: S. Dasgupta, A. Klivans (Eds.), Proceedings of the 22nd Annual Conference on Learning Theory, 2009, pp. 79–93.
[13] H.W. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, Appl. Comput. Harmon. Anal. 30 (2011) 96–109.
[14] H.W. Sun, Q. Wu, Indefinite kernel network with dependent sampling, Anal. Appl. 11 (5) (2013) 1957–1967.
[15] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.
[16] C. Wang, J. Cai, Convergence analysis of coefficient-based regularization under moment incremental condition, Int. J. Wavelets Multiresolut. Inf. Process. 12 (1) (2014), http://dx.doi.org/10.1142/S0219691314500088.
[17] C. Wang, W.L. Nie, Constructive analysis for least squares regression with generalized k-norm regularization, Abstr. Appl. Anal. 2014 (2014), http://dx.doi.org/10.1155/2014/458459.
[18] C. Wang, D.X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, J. Complexity 27 (2011) 55–67.
[19] Q. Wu, Y. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2006) 171–192.
[20] Q. Wu, Y. Ying, D.X. Zhou, Multi-kernel regularized classifiers, J. Complexity 23 (2007) 108–134.
[21] Q. Wu, D.X. Zhou, Svm soft margin classifier: linear programming versus quadratic programming, Neural Comput. 17 (2005) 1160–1187.
[22] Q. Wu, D.X. Zhou, Learning with sample dependent hypothesis space, Comput. Math. Appl. 56 (2008) 2896–2907.
[23] Q.W. Xiao, D.X. Zhou, Learning by nonsymmetric kernels with data dependent spaces and l^1-regularizer, Taiwanese J. Math. 14 (2010) 1821–1836.
[24] Z.B. Xu, X.Y. Chang, F.M. Xu, L-1/2 regularization: a thresholding representation theory and a fast solver, IEEE Trans. Neural Netw. Learn. Syst. 23 (7) (2012) 1013–1027.
[25] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[26] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.