Research Article

Indefinite Kernel Network with l^q -Norm Regularization

Zhongfeng Qu and Hongwei Sun

School of Mathematical Science, University of Jinan, Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China

Correspondence should be addressed to Hongwei Sun; ss_sunhw@ujn.edu.cn

Received 29 November 2015; Accepted 13 March 2016

Academic Editor: Bo Yang

Copyright © 2016 Z. Qu and H. Sun. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We study the asymptotical properties of indefinite kernel network with l^q -norm regularization. The framework under investigation is different from classical kernel learning. Positive semidefiniteness is not required by the kernel function. By a new step stone technique, without any interior cone condition for input space $\mathcal{X}$ and L_τ condition for the probability measure $\rho_{\mathcal{X}}$, satisfied error bounds and learning rates are deduced.

1. Introduction

Regression learning has been widely applied in economic decision making, engineering, computer science, and especially statistics. To fix ideas, let the input data space $\mathcal{X}$ be a compact domain of $\mathbb{R}^n$; the output data space $\mathcal{Y}$ is a subset of $\mathbb{R}$, and ρ is a Borel probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, which the random variable (X, Y) is drawn from. Let $\rho(\cdot | x)$ be the conditional distribution for $x \in \mathcal{X}$, and $\rho_{\mathcal{X}}$ is the marginal distribution. The goal of regression learning is to learn the regression function $f_\rho : \mathcal{X} \rightarrow \mathcal{Y}$, which is given by

$$f_\rho(x) = \mathbb{E}(Y | X = x) = \int_{\mathcal{Y}} y d\rho(y | x). \quad (1)$$

As the distribution ρ is unknown, f_ρ can not be calculated directly. In learning theory framework, one can get an approximation of f_ρ from a set of observations $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$ identically and independently drawn according to ρ . Least square scheme is the most popular method for regression problem. The generalization error for any measurable function f ,

$$\mathcal{E}(f) = \int_{\mathcal{X}} (f(x) - y)^2 d\rho, \quad (2)$$

is the mean error that suffered from the use of f as a model for the process producing y from x ; see Cucker and Zhou [1]. A simple computation shows that [1]

$$\begin{aligned} \mathcal{E}(f_\rho) &= \min \int_{\mathcal{X}} (f(x) - y)^2 d\rho, \\ \|f - f_\rho\|_{L^2_{\rho_{\mathcal{X}}}}^2 &= \mathcal{E}(f) - \mathcal{E}(f_\rho). \end{aligned} \quad (3)$$

Thus, $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ can be used to value the learning effect for f to approximate f_ρ .

In order to minimize the expected risk functional, we employ the empirical risk functional

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2. \quad (4)$$

The empirical risk minimization (ERM) principle is to approximate the function f_ρ which minimizes the risk $\mathcal{E}$ by the function f_* which minimizes the empirical risk (4) in some hypothesis space $\mathcal{H}$. Kernel based machines usually take reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_K$ as the hypothesis space, which is associated with a Mercer kernel K . For definitions and properties, see [2]. The well-known regularized least square regression algorithm is

$$f_{\mathbf{z}, \gamma} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma \|f\|_K^2 \right\}. \quad (5)$$

The penalty term $\gamma \|f\|_K^2$ with regularization parameter $\gamma > 0$ is to avoid ill-posed and overfitting. In the literature, a variety of consistence analysis have been done, for instance, the capacity dependent ones (e.g., [3, 4] and references therein) and the capacity independent ones (e.g., [5–8]).

In recent years, coefficient based regularization kernel network (CRKN) attracts more attentions:

$$f_z = \arg \min_{f_\alpha \in \mathcal{H}_{K,x}} \frac{1}{m} \sum_{i=1}^m (y_i - f_\alpha(x_i))^2 + \lambda \Omega(\alpha), \quad (6)$$

where $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$ and $f_\alpha = \sum_{i=1}^m \alpha_i K(x, x_i)$. In this setting, the hypothesis space $\mathcal{H}_K$ is replaced by a finite dimension function space:

$$\mathcal{H}_{K,x} = \left\{ f_\alpha(x) = \sum_{i=1}^m \alpha_i K(x, x_i) : \alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m, m \in \mathbb{N} \right\}, \quad (7)$$

where $K : X \times X \rightarrow \mathbb{R}$ is only asked to be continuous and uniform bounded on $X \times X$ and is called a general kernel. The penalty $\Omega(\alpha)$ is imposed on the coefficients of function $f \in \mathcal{H}_{K,x}$.

Some researches have been done for the mathematical foundation of CRKN. In [9], the frame of error analysis for this coefficient based regularization is proposed. CRKN with l^2 -norm regularizer and indefinite kernel is studied; error bound and learning rate are derived thorough explicit expression of solution f_z and integral operator techniques, not only for independently sampling, but also for weak dependent sampling; see [10, 11].

CRKN with l^1 -norm regularizer attracts more attention, for it often leads to sparsity of the coefficient $\{\alpha_{z,i}\}_{i=1}^m$ of f_z with properly chosen regularization parameter λ . In [12], Xiao and Zhou analyzed the coefficient regularization using the ℓ_1 norm. Under a Lipschitz condition of the kernel K , the rate of order $O(m^{-r/(2(1+m))})$ is obtained. It is very slow since n is usually very large. In [13], under an interior cone condition for input space $\mathcal{X}$ and L_τ condition for the probability measure $\rho_{\mathcal{X}}$, satisfied error bound and sharp convergence rate are obtained. The shortage for this analysis is that conditions are too strict, especially for L_τ condition, since it only includes the almost uniform distribution on $\mathcal{X}$. To the best of our knowledge, the probability measure $\rho_{\mathcal{X}}$ has no direct relation with the consistency of algorithms. So we attempt to deduce satisfied error bound under more general conditions. To this end, we introduce a new step stone technique to the consistence analysis of CRKN with l^1 -norm regularization in [14].

In this paper, we adopt the step stone technique to study the ℓ^q norm based CRKN with $1 \leq q \leq 2$, which is defined as

$$f_z = f_{\alpha_z},$$

$$\text{where } \alpha_z = \arg \min_{\alpha \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^m (y_i - f_\alpha(x_i))^2 + \eta \|\alpha\|_{q,m}^q, \quad (8)$$

where $f_\alpha = \sum_{i=1}^m \alpha_i K_{x_i}(\cdot) = \sum_{i=1}^m \alpha_i K(\cdot, x_i)$, and the norm

$$\|\alpha\|_{q,m} := \left\{ m^{q-1} \sum_{i=1}^m |\alpha_i|^q \right\}^{1/q}. \quad (9)$$

2. Assumptions and Main Results

Throughout the paper, we always assume that $|y| \leq M$ almost surely with some constant $M > 0$. By this assumption, we have that $|f_\rho(x)| \leq M$ for any $x \in X$. Also we require the kernel $K \in C^s(\mathcal{X} \times \mathcal{X})$ with $s > 0$, and the norm $\|K\|_{C^s(\mathcal{X})}$ is denoted by κ .

For the bounded indefinite kernel $K(x, y)$, we consider the Mercer kernel

$$\tilde{K}(x, t) = \int_{\mathcal{X}} K(x, u) K(t, u) d\rho_{\mathcal{X}}(u). \quad (10)$$

The integral operator associated with a general kernel $T(x, y)$ is defined by

$$L_T f(x) = \int_{\mathcal{X}} T(x, u) f(u) d\rho_{\mathcal{X}}(u), \quad (11)$$

which is a compact positive operator. For the approximation ability of kernel K based hypothesis space, we assume that $L_{\tilde{K}}^{-\beta/2} f_\rho \in L_{\rho_{\mathcal{X}}}^2(\mathcal{X})$, for some $\beta > 0$.

One of the main difficulties for the analysis of CRKN is that the hypothesis space is dependent on the sample.

Definition 1. Define a Banach space

$$\mathcal{H}_1 = \left\{ f : f = \sum_{i=1}^{\infty} \alpha_i K_{t_i}, \{\alpha_i\} \in l^1, \{t_i\} \subset X \right\} \quad (12)$$

with the norm

$$\|f\| = \inf \left\{ \sum_{i=1}^{\infty} |\alpha_i| : f = \sum_{i=1}^{\infty} \alpha_i K_{t_i}, \{\alpha_i\} \in l^1, \{t_i\} \subset X \right\}. \quad (13)$$

The continuity and uniform boundedness of K ensure that $\mathcal{H}_1$ consists of continuous functions. Denote the ball of radius R as $B_R = \{f : f \in \mathcal{H}_1, \|f\| \leq R\}$.

Our estimate of sample error is conducted through a concentration inequality, which is based on the l^2 empirical covering number of B_R . The normalized l^2 metric d_2 on Euclidean space $\mathbb{R}^k$ is defined by

$$(d_2(\alpha, \beta))^2 = \frac{1}{k} \sum_{i=1}^k (a_i - b_i)^2, \quad (14)$$

$$\text{for any } \alpha = (a_i)_{i=1}^k, \beta = (b_i)_{i=1}^k \in \mathbb{R}^k.$$

Under the above metric, the covering number of a subset E of $\mathbb{R}^k$ with $\epsilon > 0$ is denoted by $\mathcal{N}_2(E, \epsilon)$.

Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$, $\mathbf{x} = (x_i)_{i=1}^k \in \mathcal{X}^k$. The sampling operator $S_{\mathbf{x}} : \mathcal{F} \rightarrow \mathbb{R}^k$ is defined by

$S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^k$, for any $f \in \mathcal{F}$. The l^2 empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{k \in \mathbb{N}} \sup_{\mathbf{x} \in \mathcal{X}^k} \mathcal{N}_2(S_{\mathbf{x}}(\mathcal{F}), \epsilon). \quad (15)$$

A bound for l^2 empirical covering number of B_1 was derived in [13]. Suppose that $K \in C^s(\mathcal{X} \times \mathcal{X})$; then,

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_p \left(\frac{1}{\epsilon}\right)^p, \quad \forall \epsilon > 0, \quad (16)$$

where c_p is a constant independent of ϵ , and $0 < p < 2$ is defined by

$$p = \begin{cases} \frac{2n}{n+2s} & \text{when } 0 < s \leq 1, \\ \frac{2n}{n+2} & \text{when } 1 < s \leq 1 + \frac{n}{2}, \\ \frac{n}{s} & \text{when } s > 1 + \frac{n}{2}. \end{cases} \quad (17)$$

Our main conclusion is about the asymptotic convergence rate of kernel network (8).

Theorem 2. Let f_z be the optimal solution of the algorithm (8) with $1 \leq q \leq 2$. Suppose B_1 satisfies the capacity condition (16) with $0 < p < 2$, and $L_{\bar{K}}^{-\beta/2} f_\rho \in L_{\rho_x}^2(\mathcal{X})$ for some $\beta > 0$. Taking $\eta = m^{-t}$ such that $0 < t < q/(p+2)$, we have for any $0 < \delta < 1$, with confidence $1 - \delta$, that there holds

$$\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq \tilde{C} m^{-t \min\{\beta, 1\}}, \quad (18)$$

where $\tilde{C}$ is a constant independent of m .

Corollary 3. Under assumptions of Theorem 2. Suppose that $q = 1$ and $f_\rho \in \mathcal{H}_{\bar{K}}$. Taking $\eta = m^{\epsilon-1/(2+p)}$ with $1/(2+p) > \epsilon > 0$, for any $0 < \delta < 1$, then with confidence $1 - \delta$, there holds

$$\mathcal{E}(f_z) - \mathcal{E}(f_\rho) = O(m^{-1/(2+p)+\epsilon}). \quad (19)$$

The condition of $f_\rho \in \mathcal{H}_{\bar{K}}$ is equal to $\beta \geq 1$. When the kernel $K \in C^\infty(\mathcal{X} \times \mathcal{X})$, the parameter p can be arbitrarily close to 2; thus, our learning rate is almost the optimal rate $1/4$ for CRKN with l^1 -norm regularization; see [13].

3. Rough Error Bounds of l^q -Regularization

We propose the following new error decomposition by employing a stepping-stone approach.

Consider the regularized kernel network with l^2 norm:

$$f_{z,\lambda} = f_{\alpha_{z,\lambda}}, \quad (20)$$

where $\alpha_{z,\lambda} = \arg \min_{\alpha \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^m (y_i - f_\alpha(x_i))^2 + \lambda \|\alpha\|_{2,m}^2$.

Thus, we have that

$$\begin{aligned} \mathcal{E}(f_z) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(f_z) - \mathcal{E}(f_\rho) + \eta \|\alpha_z\|_{q,m}^q \\ &\quad + \eta \|\alpha_{z,\lambda}\|_{q,m}^q \\ &= \{\mathcal{E}(f_z) - \mathcal{E}_z(f_z)\} \\ &\quad + \{\mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}(f_{z,\lambda})\} + \mathcal{E}_z(f_z) \\ &\quad + \eta \|\alpha_z\|_{q,m}^q - \mathcal{E}_z(f_{z,\lambda}) \\ &\quad - \eta \|\alpha_{z,\lambda}\|_{q,m}^q + 2\eta \|\alpha_{z,\lambda}\|_{q,m}^q \\ &\quad + \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \\ &\leq \{\mathcal{E}(f_z) - \mathcal{E}_z(f_z)\} \\ &\quad + \{\mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}(f_{z,\lambda})\} \\ &\quad + 2\eta \|\alpha_{z,\lambda}\|_{2,m}^q + \mathcal{E}(f_{z,\lambda}) \\ &\quad - \mathcal{E}(f_\rho). \end{aligned} \quad (21)$$

The second inequality holds by the definition of f_z and $f_{z,\lambda} \in \mathcal{H}_{K,\mathbf{x}}$.

Let $g_f(z) = (f(x) - y)^2 - (f_\rho(x) - y)^2$ for $f \in \mathcal{H}_1$. For any $f, \tilde{f} \in \mathcal{H}_1$, we have

$$\begin{aligned} &\{\mathcal{E}(f) - \mathcal{E}_z(f)\} + \{\mathcal{E}_z(\tilde{f}) - \mathcal{E}(\tilde{f})\} \\ &= \mathbb{E} g_f - \frac{1}{m} \sum_{i=1}^m g_f(z_i) + \frac{1}{m} \sum_{i=1}^m g_{\tilde{f}}(z_i) - \mathbb{E} g_{\tilde{f}}. \end{aligned} \quad (22)$$

The following concentration inequality is the special case of Lemma 3 in [13]; here we take $\alpha = 1$.

Lemma 4. Let $\mathcal{F}$ be a class of measurable functions on Z . Suppose that there exist constants $b, c > 0$, such that $\|f\|_\infty \leq b$ and $\mathbb{E} f^2 \leq c \mathbb{E} f$ for every $f \in \mathcal{F}$. Moreover, there are some $a > 0$ and $p \in (0, 2)$, such that

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \leq a \left(\frac{1}{\epsilon}\right)^p, \quad \forall \epsilon > 0. \quad (23)$$

Then, for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds, for any $f \in \mathcal{F}$,

$$\begin{aligned} &\left| \mathbb{E} f - \frac{1}{m} \sum_{i=1}^m f(z_i) \right| \\ &\leq \frac{1}{2} \mathbb{E} f + c'_p (\max\{c, b\})^{(2-p)/(2+p)} \left(\frac{a}{m}\right)^{2/(2+p)} \\ &\quad + (2c + 18b) \frac{\log(4/\delta)}{m}, \end{aligned} \quad (24)$$

where c'_p is a constant only depending on p .

For any $r > 0$, the function set $\mathcal{F}_r$ is defined by $\mathcal{F}_r = \{g_f : f \in B_r\}$. Now we verify that the conditions in Lemma 4 hold. Firstly, $\|g_f\|_\infty \leq (3M + \kappa r)^2$ for any $g_f \in \mathcal{F}_r$, and

$$\mathbb{E}g_f^2 \leq (3M + \kappa r)^2 \mathbb{E}g_f. \quad (25)$$

Secondly, for any $g_{f_1}, g_{f_2} \in \mathcal{F}_r$,

$$\|g_{f_1} - g_{f_2}\|_\infty \leq (\kappa r + 3M) \|f_1 - f_2\|_\infty. \quad (26)$$

Therefore, we have that, for $\forall \epsilon > 0$,

$$\begin{aligned} \mathcal{N}_2(\mathcal{F}_r, \epsilon) &\leq \mathcal{N}_2\left(B_r, \frac{\epsilon}{(\kappa r + 3M)}\right) \\ &\leq \mathcal{N}_2\left(B_1, \frac{\epsilon}{r(\kappa r + 3M)}\right) \\ &\leq c_p \left(\frac{r(\kappa r + 3M)}{\epsilon}\right)^p. \end{aligned} \quad (27)$$

For any vector $\alpha = (a_1, a_2, \dots, a_m)$, there is

$$\sum_{i=1}^m |a_i| = \|\alpha\|_{1,m} \leq \left(\sum_{i=1}^m |a_i|^q\right)^{1/q} \times m^{1-1/q} = \|\alpha\|_{q,m}. \quad (28)$$

Denote

$$\mathcal{W}(r) = \left\{ \mathbf{z} \in \mathcal{Z}^m : \|\alpha_z\|_{q,m} \leq r, \|\alpha_{z,\lambda}\|_{q,m} \leq r \right\}. \quad (29)$$

For any $\mathbf{z} \in \mathcal{W}(r)$, there is $f_z, f_{z,\lambda} \in B_r$.

Applying (22) and Lemma 4 with $b = c = (3M + \kappa r)^2$ and $a = c_p(r(\kappa r + 3M))^p$, it follows that there is a subset $\tilde{V}_r$ of $\mathcal{Z}^m$ with measure at most $\delta/2$ such that

$$\begin{aligned} &\{\mathcal{E}(f_z) - \mathcal{E}_z(f_z)\} + \{\mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}(f_{z,\lambda})\} \\ &\leq \frac{1}{2} \left\{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) + \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \right\} \\ &\quad + 2 \left(1 + \frac{1}{\kappa} \right) c_p' (c_p)^{2/(2+p)} (\kappa r + 3M)^2 \left(\frac{1}{m} \right)^{2/(2+p)} \\ &\quad + 40 (\kappa r + 3M)^2 \frac{\log 4/\delta}{m}, \quad \forall \mathbf{z} \in \mathcal{W}(r) \setminus \tilde{V}_r. \end{aligned} \quad (30)$$

Plugging this estimate into (21) yields that

$$\begin{aligned} &\mathcal{E}(f_z) - \mathcal{E}(f_\rho) + \eta \|\alpha_z\|_{q,m}^q + \eta \|\alpha_{z,\lambda}\|_{q,m}^q \\ &\leq 4 \left(1 + \frac{1}{\kappa} \right) (\kappa + 3)^2 c_p' c_p^{2/(2+p)} \max\{r^2, M^2\} \\ &\quad \cdot \left(\frac{1}{m} \right)^{2/(2+p)} + 3 \left\{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \right\} + 80 (\kappa + 3)^2 \\ &\quad \cdot \max\{r^2, M^2\} \frac{\log 4/\delta}{m} + 3\eta \|\alpha_{z,\lambda}\|_{2,m}^q. \end{aligned} \quad (31)$$

The following excess generalization error of $f_{z,\lambda}$ was proved in [15].

Proposition 5. Let $f_{z,\lambda}$ be given by (20) and $K \in C^s(\mathcal{X} \times \mathcal{X})$ with some $s > 0$. Suppose that $L_{\bar{K}}^{-\beta/2} f_\rho \in L_{\rho_x}^2(\mathcal{X})$ for some $\beta > 0$ and $0 < \lambda \leq 1$. Then, for any $\delta \in (0, 1)$, take $\lambda = m^{-\gamma}$ with $0 < \gamma < 2/(2+p)$, where p is defined by (17); then, with probability $1 - \delta/2$ there holds

$$\begin{aligned} \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) &= \|f_z - f_\rho\|_{L_{\rho_x}^2}^2 \\ &\leq \bar{C}_{K,p,\gamma} \left(\log \left(\frac{20\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)} m^{-\min\{\beta, 1\}\gamma}, \end{aligned} \quad (32)$$

$$\begin{aligned} \|\alpha_{z,\lambda}\|_{2,m} &\leq \bar{C}_\gamma \left(\log \left(\frac{20\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)/2} m^{\max\{(1-\beta)\gamma/2, 0\}}, \end{aligned}$$

where $\nu(\gamma, p) = (4 - \gamma(2+p))/(2 - \gamma(2+p))$ and $\bar{C}_{K,p,\gamma}, \bar{C}_\gamma$ are constants independent of δ and m .

Combining this estimate with (31) leads to Proposition 6.

Proposition 6. Let $K \in C^s(\mathcal{X} \times \mathcal{X})$ with some $s > 0$. Assume that $L_{\bar{K}}^{-\beta/2} f_\rho \in L_{\rho_x}^2(\mathcal{X})$, for some $2 \geq \beta > 0$. Taking $\lambda = m^{-\gamma}$ with $0 < \gamma < 2/(2+p)$ and $0 < \eta, \delta < 1$, then for any $r > 0$ there is a subset V_r of Z^m with measure at most δ such that, for any $\mathbf{z} \in \mathcal{W}(r) \setminus V_r$,

$$\begin{aligned} &\mathcal{E}(f_z) - \mathcal{E}(f_\rho) + \eta \|\alpha_z\|_{q,m}^q + \eta \|\alpha_{z,\lambda}\|_{q,m}^q \\ &\leq \tilde{a} \left(m^{-2/(2+p)} + m^{-1} \log \frac{4}{\delta} \right) \max\{r^2, M^2\} \\ &\quad + \tilde{b} \left(\log \left(\frac{20\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)} \\ &\quad \cdot \left(\eta m^{\max\{(1-\beta)\gamma/2, 0\}} + m^{-\min\{\beta, 1\}\gamma} \right), \end{aligned} \quad (33)$$

where $\tilde{a}$ and $\tilde{b}$ are constants independent on δ, m , and η .

4. Refined Error Bound by Iteration

In this section, we apply an iteration technique to deduce an error bound. Proposition 6 ensures that, for $\mathbf{z} \in \mathcal{W}(r) \setminus V_r$,

$$\begin{aligned} &\max \left\{ \|\alpha_z\|_{q,m}, \|\alpha_{z,\lambda}\|_{q,m} \right\} \\ &\leq \tilde{a}^{1/q} \left(\eta^{-1/q} m^{-2/(2+p)q} + \eta^{-1/q} m^{-1/q} \left(\log \frac{4}{\delta} \right)^{1/q} \right) \\ &\quad \cdot \max\{r^{2/q}, M^{2/q}\} \\ &\quad + \tilde{b}^{1/q} \left(\log \left(\frac{20\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)/q} \\ &\quad \cdot \left(m^{\max\{(1-\beta)\gamma/2, 0\}} + \eta^{-1/q} m^{-\min\{\beta, 1\}(\gamma/q)} \right). \end{aligned} \quad (34)$$

We take $\eta = m^{-t}$ with $0 < t \leq q\gamma/2 < q/(2+p)$, which yields that

$$\begin{aligned} \max \left\{ \|\alpha_z\|_{q,m}, \|\alpha_{z,\lambda}\|_{q,m} \right\} &\leq 2 \left(\tilde{a} \log \frac{4}{\delta} \right)^{1/q} \\ &\cdot m^{t/q-2/(2+p)q} \max \left\{ r^{2/q}, M^{2/q} \right\} + \max \left\{ 2\tilde{b}^{1/q}, \frac{M}{2} \right\} \\ &\cdot \left(\log \left(\frac{20\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)/q} m^{(\gamma/2)(1-\min\{\beta, 1\})} := \frac{1}{2} \\ &\cdot a_{m,\eta} \max \left\{ r^{2/q}, M^{2/q} \right\} + \frac{1}{2} b_{m,\eta}. \end{aligned} \quad (35)$$

It follows that, for any $r > 0$,

$$\begin{aligned} \mathcal{W}(r) &\subset \mathcal{W}(\max \{a_{m,\eta} \max \{r^{2/q}, M^{2/q}\}, b_{m,\eta}\}) \\ &\cup V_r. \end{aligned} \quad (36)$$

Let $r_j = \max \{a_{m,\eta} \max \{r_{j-1}^{2/q}, M^{2/q}\}, b_{m,\eta}\}$ and $r_0 = \max \{M, M^{2/q}\} m^{t/q}$. Hence,

$$\begin{aligned} Z^m &= \mathcal{W}(r_0) \subset \mathcal{W}(r_1) \cup V_{r_0} \subset \dots \\ &\subset \mathcal{W}(r_J) \cup \left(\bigcup_{j=0}^{J-1} V_{r_j} \right). \end{aligned} \quad (37)$$

Proof of Theorem 2. By taking $\gamma = 2t/q$ and replacing δ by $\delta/(J+1)$ in Proposition 6 and the above discussion, with confidence $1 - J\delta/(J+1)$ for some $0 < \delta < 1$, the estimate (38) assures that

$$\begin{aligned} r_J &= \max \left\{ \|\alpha_z\|_q^{1/q}, \|\alpha_{z,\lambda}\|_q^{1/q} \right\} \\ &\leq \tilde{c}_1 \mu \left(\gamma, p, \frac{\delta}{J+1} \right)^{(2/q)^J} m^{\tilde{\theta}}, \end{aligned} \quad (42)$$

where $\tilde{c}_1$ is a constant independent of m , and μ , $\tilde{\theta}$, and J are defined by (39), (40), and (41). Proposition 6 ensures that, with confidence $1 - \delta/(J+1)$,

$$\begin{aligned} \mathcal{E}(f_z) - \mathcal{E}(f_p) &\leq \tilde{a} \left(m^{-2/(2+p)} + m^{-1} \log \frac{4(J+1)}{\delta} \right) \\ &\cdot \max \left\{ r_J^2, M^2 \right\} \\ &+ \tilde{b} \left(\log \left(\frac{20(J+1)\gamma(\gamma, p)}{\delta} \right) \right)^{\nu(\gamma, p)} \\ &\cdot \left(\eta m^{\max\{(1-\beta)q\gamma/2, 0\}} + m^{-\min\{\beta, 1\}\gamma} \right). \end{aligned} \quad (43)$$

Combing the above two inequalities completes the proof. $\square$

For $0 < \eta \leq 1$, we have that $r_0 \geq M$, and $r_j \geq b_{m,\eta} \geq M$; thus, $r_j = \max \{a_{m,\eta} r_{j-1}^{2/q}, b_{m,\eta}\}$. A simple computation (see [13]) gives that

$$\begin{aligned} r_J &\leq \max \left\{ \tilde{a}_{m,\eta}^{(2/q)^J-1} r_0^{(2/q)^J}, b_{m,\eta} (\tilde{a}_{m,\eta} b_{m,\eta})^{(2/q)^{J-1}-1}, b_{m,\eta} \right\} \\ &\leq \mu(\gamma, p, \delta)^{(2/q)^J} \max \left\{ (\tilde{c}_{\kappa, p, M})^{(2/q)^J}, (\tilde{d}_{\kappa, p, M})^{(2/q)^J} \right\}, \quad (38) \\ &2^{2+2/q} \tilde{b}^{1/q} \left\{ m^{\tilde{\theta}}, \right. \end{aligned}$$

where

$$\begin{aligned} \tilde{a}_{m,\eta} &= a_{m,\eta}^{q/(2-q)}, \\ \mu(\gamma, p, \delta) &= \left(\log \frac{20\gamma(\gamma, p)}{\delta} \right)^{\nu(\gamma, p)/q} \left(\log \frac{4}{\delta} \right)^{1/(2-q)}, \end{aligned} \quad (39)$$

and $\tilde{c}_{\kappa, p, M}$ and $\tilde{d}_{\kappa, p, M}$ are both constants independent of m , η , δ , and by taking γ , t as $0 < \gamma < 2/(2+p)$ and $0 < t \leq q\gamma/2$, we can deduce

$$\tilde{\theta} = \frac{\gamma}{2} (1 - \min\{\beta, 1\}), \quad (40)$$

by choosing J to be

$$J = \left\lceil \frac{\ln((2/(2+p) - t - ((2-q)r/2)(1 - \min\{\beta, 1\}))/ (2/(2+p) - t - (2-q)t/q))}{\ln(2/q)} \right\rceil + 1. \quad (41)$$

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Zhongfeng Qu is supported by Shandong Provincial Scientific Research Foundation for Excellent Young Scientists (no. BS2013SF003). Hongwei Sun is supported by the Nature Science Fund of Shandong province China (no. ZR2014AM010).

References

- [1] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.
- [2] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.
- [3] Q. Wu, Y. M. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.
- [4] Y.-L. Xu and D.-R. Chen, "Learning rates of regularized regression for exponentially strongly mixing sequence," *Journal of Statistical Planning and Inference*, vol. 138, no. 7, pp. 2180–2189, 2008.

- [5] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. 3, pp. 499–526, 2002.
- [6] 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.
- [7] H. Sun and Q. Wu, "Regularized least square regression with dependent samples," *Advances in Computational Mathematics*, vol. 32, no. 2, pp. 175–189, 2010.
- [8] T. Zhang, "Leave-one-out bounds for kernel methods," *Neural Computation*, vol. 15, no. 6, pp. 1397–1437, 2003.
- [9] Q. Wu and D.-X. Zhou, "Learning with sample dependent hypothesis spaces," *Computers & Mathematics with Applications*, vol. 56, no. 11, pp. 2896–2907, 2008.
- [10] H. Sun and Q. Wu, "Coefficient regularization in least square kernel regression," *Applied and Computational Harmonic Analysis*, vol. 30, pp. 96–109, 2011.
- [11] H. Sun and Q. Wu, "Indefinite kernel network with dependent sampling," *Analysis and Applications*, vol. 11, no. 5, Article ID 1350020, 2013.
- [12] Q. W. Xiao and D. X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and 1-regularizer," *Taiwanese Journal of Mathematics*, vol. 14, no. 5, pp. 1821–1836, 2010.
- [13] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with l^1 -regularizer and data dependent hypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.
- [14] H. Sun and Q. Wu, "Sparse representation in kernel machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2576–2582, 2015.
- [15] L. Shi, "Learning theory estimates for coefficient-based regularized regression," *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 252–265, 2013.