IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 24, 2018, accepted July 30, 2018, date of publication August 6, 2018, date of current version August 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2863600

Error Analysis of Least-Squares l^q-Regularized Regression Learning Algorithm With the Non-Identical and Dependent Samples

QIN GUO AND PEIXIN YE
School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

Corresponding author: Qin Guo (guoqin_1985@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11671213 and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT The selection of the penalty functional is critical for the performance of a regularized learning algorithm, and thus l^q-regularizer ($1 \leq q \leq 2$) deserves special attention. We consider the regularized least-squares regression learning algorithm for the non-identical and weakly dependent samples. The dependent samples satisfy the polynomially β-mixing condition and the sequence of the non-identical sampling marginal measures converges to a probability measure exponentially in the dual of a Hölder space. We conduct the rigorous unified error analysis and derive the satisfactory learning rates of the algorithm by the stepping stone technique in the error decomposition and the independent-blocks technique in the sample error estimates.

INDEX TERMS Regression function, empirical covering number, drift error, blocking technique, learning rate.

I. INTRODUCTION

In this paper we consider the regularized least-squares regression learning algorithm with l^q-regularizer ($1 \leq q \leq 2$). Let X be a compact metric space and $Y = \mathbb{R}$ be a compact subset of $\mathbb{R}$. Let ρ be a unknown Borel probability distribution on $Z = X \times Y$. A set of random samples $z = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ are drawn according to the measure ρ. We define the generalization error as follows:

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho, \quad \forall f : X \to Y. \tag{1}$$

The regression function f_ρ minimizing $\mathcal{E}(f)$ is defined by

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

where $\rho(\cdot|x)$ is the conditional probability distribution induced by ρ on Y. Regression learning algorithms aim at finding a good approximation f_z of f_ρ based on $\{z_i\}_{i=1}^m$. Our task is to estimate the error

$$\|f_z - f_\rho\|_{\rho_X}^2 = \mathcal{E}(f_z) - \mathcal{E}(f_\rho),$$

where $\|f(\cdot)\|_{\rho_X} = (\int_X |f(\cdot)|^2 d\rho_X)^{\frac{1}{2}}$ and ρ_X is the marginal distribution of ρ on X, see [1], [2]

A kernel K is called a Mercer kernel if it is a continuous, symmetric, and positive semi-definite function on $X \times X$. The hypothesis space $\mathcal{H}_K$ is defined by the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$. It takes the following inner product

$$\left\langle \sum_{i=1}^n \alpha_i K_{x_i}, \sum_{j=1}^m \beta_j K_{y_j} \right\rangle_K := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j K(x_i, y_j).$$

The reproducing property is given by

$$f(x) = \langle f, K_x \rangle_K, \quad \textit{for all } f \in \mathcal{H}_K, \ x \in X. \tag{2}$$

We study the following least-squares regularization algorithm with l^q-regularizer

$$f_{z,\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_z(f) + \lambda \|f\|_K^q \right\}, \quad 0 < q \leq 2, \tag{3}$$

where $\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$ and λ is the positive regularization parameter with $\lim_{m \to \infty} \lambda(m) = 0$.

Smale and Zhou [3], Lv and Feng [4], Caponnetto and Vito [5], Cucker and Smale [6], Vito *et al.* [7], and Mendelson and Neeman [8] carried out the error analysis of the algorithm (3) with $q = 2$ for the independent and identical (i.i.d.) samples. However, the samples are not independent but are not far from being independent in some real data analysis. The mixing conditions can quantify how close to independence a sequence of random samples. Guo *et al.* [10], Sun and Wu [12], Chu and Sun [13], Pan and Xiao [14], and Guo and Ye [16] carried out the regression estimation of the least squares algorithm with the α-mixing and ϕ-mixing samples. Error estimates for classification and coefficient regularized regression learning with the β-mixing samples have been conducted in [10] and [15]. Since the β-mixing is quite easy to establish and covers a more general non-i.i.d. cases such as Gaussian and Markov processes. As pointed out in [15], the β-mixing is"just the right" assumption for extending the analysis of several learning algorithms to the case of weakly dependent samples, since there exists some available results under β-mixing conditions, see [15]. And we can replace the empirical process by another independent empirical process built over an independent block technique for the β-mixing. In this paper, following the approach in [10] and [15], we obtain error bounds of the algorithm (3) for the β-mixing and the sequence of non-identical probability distributions.

Definition 1: The stochastic process $\{z_t\}$ is said to satisfy the β-mixing, if

$$\beta(k) = \sup_{j \geq 1} \mathbb{E} \sup_{A \in \sigma_{j+k}^{\infty}} |P(A|\sigma_1^j) - P(A)| \to 0,$$

as $k \to \infty$, where σ_i^j is the σ-algebra generated by $\{z_t = (x_t, y_t)\}_{t=i}^{j}$, $i, j \in \mathbb{N} \cup \{+\infty\}$. It satisfies a polynomially β-mixing, if for some positive constants $\beta_0 > 0$ and $\gamma > 0$, we have

$$\beta(k) \leq \beta_0 k^{-\gamma}, \forall k \geq 1. \qquad (4)$$

We consider the non-identical setting in [10], [14], and [15]. The probability measure of $z_i = (x_i, y_i)$ is the Borel probability measure $\rho^{(i)}$ on Z. $\rho_X^{(i)}$ is the marginal probability measure of $\rho^{(i)}$ and $\rho(\cdot|x)$ is the conditional probability measure of $\{\rho^{(i)}\}_{i=1,2,\cdots}$ at x, for $x \in X$, independent of i. We assume the sequence $\{\rho_X^{(i)}\}$ converges to ρ_X exponentially fast in the dual $(C^s(X))^*$ of the Hölder space $C^s(X)$, that is,

$$\left| \int_X f(x) d\rho_X^{(i)} - \int_X f(x) d\rho_X \right| \leq C\alpha^i (\|f\|_\infty + |f|_{C^s(X)}), \quad \forall f \in C^s(X),\ i \in \mathbb{N}, \quad (5)$$

where $C^s(X)$ is defined by the set of all continuous functions on X and

$$\|f\|_{C^s(X)} := \|f\|_\infty + |f|_{C^s(X)},$$

where

$$|f|_{C^s(X)} := \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{(d(x, y))^s}.$$

The above assumption of probability measures is reasonable. We can generate the non-identical sequence by iterations of a stochastic linear operator acting on an initial probability measure or induced by dynamical systems, see [15], [17].

The goal of this paper is to obtain the unified error bounds of the algorithm (3) that covers the case $1 \leq q \leq 2$ under the conditions (4) and (5). To the best of our knowledge, this is the first time that a generalization error analysis of the algorithm (3) is extended to the β-mixing and non-identical sampling case. The remainder of this paper is organized as follows. In Section II, we will give our main result and the error decomposition analysis of the total error. The upper bounds of the drift error, the approximation error and the sample error will be obtained in Section III. In Section IV, we will derive the learning rate. Finally, Section V concludes the paper with future research lines.

II. MAIN RESULT AND ERROR DECOMPOSITION

To state the bound of $\|f_{z,\lambda} - f_\rho\|_{\rho_X}^2$, we firstly provide the following assumptions of K, $\mathcal{H}_K$ and f_ρ and some concepts usually used in studying the learning algorithms.

The kernel K is said to satisfy the kernel condition of order s, if for some positive constant κ_s, we have

$$|K(x, x) - 2K(x, x') + K(x', x')| \leq \kappa_s^2 |x - x'|^{2s}, \quad \forall x, x' \in X. \quad (6)$$

When X is a domain of $\mathbb{R}^n$ with smooth boundary and K is C^2, the kernel condition holds true, see [21].

We assume that the unit ball

$$B_1 = \left\{ f \in \mathcal{H}_K : \|f\|_K \leq 1 \right\} \qquad (7)$$

of $\mathcal{H}_K$ has the capacity condition measured by the l^2 empirical covering number, see [18], [19],

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_p \epsilon^{-p}, \quad \forall \epsilon > 0, \qquad (8)$$

where c_p is the positive constant and $0 < p < 2$. In particularly, when $X \subseteq \mathbb{R}^n$ and $K \in C^s(X \times X)$ with some $s > 0$, the condition (8) is valid with

$$p = \begin{cases} 2n/(n+2s), & 0 < s \leq 1, \\ 2n/(n+2), & 1 < s \leq 1 + n/2, \\ n/s, & s > 1 + n/2. \end{cases}$$

Since the Hölder space $C^s(Y)$ and its dual $(C^s(Y))^*$ are well defined for the compact subset Y of $\mathbb{R}$, the sequence $\{\rho(y|x) : x \in X\}$ satisfies Lipschitz s in $(C_s(Y))^*$, that is, for some positive constant C_ρ,

$$\|\rho(y|x) - \rho(y|u)\|_{(C^s(X))^*} \leq C_\rho |x - u|^s, \quad \forall x,\ u \in X. \quad (9)$$

Then we can use (9) to derive the upper bounds of $|f_\rho(x)|_{C^s(X)}$ and $\left| \int_Y y^2 d\rho(y|x) \right|_{C^s(X)}$.

The integral operator $L_K : L_{\rho_X}^2(X) \to L_{\rho_X}^2(X)$ is defined by

$$(L_K f)(x) = \int_X K(x, t) f(t) d\rho_X(t), \quad x \in X.$$

And

$$L_K^r(f) = \sum_{i=1}^{\infty} \mu_i^r \langle f, e_i \rangle_{L_{\rho_X}^2} e_i, \quad f \in L_{\rho_X}^2(X),$$

where $\{\mu_i\}$ and $\{e_i\}$ are the eigenvalues and eigenfunctions, respectively, see [14].

Throughout this paper, we assume all constants are independent of δ, m, λ or σ and $|y| \leq M$ almost surely. So we can use the truncation function $\pi_M : X \to [-M, M]$, $M > 0$, which is defined by

$$\pi_M(x) = \begin{cases} M, & \textit{if } x > M, \\ x, & \textit{if } |x| \leq M, \\ -M, & \textit{if } x < -M. \end{cases} \tag{10}$$

And $\pi_M(f)(x) = \pi_M(f(x))$ for any $x \in X$, see [20], [22], [23].

Next we can state the learning rates of the algorithm (3).

Theorem 1: Suppose the sampling process satisfies (4), (5) and (9), K satisfies (6), $\mathcal{H}_K$ satisfies (8) and $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$. If $m \geq \left\{ 8^{\frac{1}{\zeta}}, \left(\frac{4\beta_0}{\delta}\right)^{\frac{1}{(\gamma+1)(1-\zeta)-1}} \right\}$, $\zeta \in \left(0, \frac{\gamma}{\gamma+1}\right)$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\|\pi_M(f_{z,\lambda}) - f_\rho\|^2_{\rho_X} \leq \widetilde{D}\left(\frac{1}{m}\right)^{\theta(r)} \log\left(\frac{8}{\delta}\right), \tag{11}$$

where

$$\theta(r) = \begin{cases} 2r \min\left\{\dfrac{q}{2rq - 2r + 2}, \dfrac{\zeta q}{2rq - 2(2r-1)}, \right. \\ \qquad\qquad \left. \dfrac{\zeta q}{rq(2+p) + p}\right\}, \\ 0 < r < \dfrac{1}{2}; \\ \min\left\{\dfrac{q}{1+q}, \dfrac{2\zeta q}{(2+p)q + 2p}, \zeta\right\}, \\ r \geq 1/2. \end{cases}$$

From the above result, we see that, for the case $q = 2$, the learning rate tends to $O\left(m^{-\min\{\frac{2}{3},\,\zeta\}}\right)$ as $p \to 0$. It is faster than $O(m^{-\frac{1}{2}})$ of Theorem 2 in [14] for the α-mixing samples when $\zeta \geq \frac{1}{2}$. This shows that our learning rate has the certain advantage.

Now we are in a position to give the decomposition of the total error. The limit of $f_{z,\lambda}$ is given by

$$f_\lambda := \arg\min_{f \in \mathcal{H}_K} \{\|f - f_\rho\|^2_{\rho_X} + \lambda \|f\|_K^q\}. \tag{12}$$

It plays a stepping stone role between $f_{z,\lambda}$ and the regression function f_ρ. To measure the error generated by the difference of the marginal measures $\{\rho_X^{(i)}\}$, we present

$$\mathcal{E}_m(f) = \frac{1}{m}\sum_{i=1}^m \int_Z (f(u) - y)^2 d\rho^{(i)}(u, y). \tag{13}$$

Then we decompose $\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\rho)$ into several parts:

$$\begin{aligned} \mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\rho) + \lambda \|f_{z,\lambda}\|_K^q \\ &= \mathcal{P}(z, \lambda) + \mathcal{S}(z, \lambda) + \mathcal{D}(\lambda) \\ &\quad + \{(\mathcal{E}_z(\pi_M(f_{z,\lambda})) + \lambda \|f_{z,\lambda}\|_K^q) \\ &\quad - (\mathcal{E}_z(f_\lambda) + \lambda \|f_\lambda\|_K^q)\}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} \mathcal{P}(z, \lambda) &= \{\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}_m(\pi_M(f_{z,\lambda}))\} \\ &\quad + \{\mathcal{E}_m(f_\lambda) - \mathcal{E}(f_\lambda)\}, \\ \mathcal{S}(z, \lambda) &= \{\mathcal{E}_m(\pi_M(f_{z,\lambda})) - \mathcal{E}_z(\pi_M(f_{z,\lambda}))\} \\ &\quad + \{\mathcal{E}_z(f_\lambda) - \mathcal{E}_m(f_\lambda)\}, \\ \mathcal{D}(\lambda) &= \|f_\lambda - f_\rho\|^2_{\rho_X} + \lambda \|f_\lambda\|_K^q. \end{aligned}$$

From the definition of $f_{z,\lambda}$, it is easy to see that the last term in the second equality of (14) is not exceeding to zero, thus

$$\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\rho) \leq \mathcal{P}(z, \lambda) + \mathcal{S}(z, \lambda) + \mathcal{D}(\lambda), \tag{15}$$

and $\mathcal{D}(\lambda)$, $\mathcal{P}(z, \lambda)$, $\mathcal{S}(z, \lambda)$ are known as the approximation error, the drift error, the sample error, respectively. Compared with the error analysis for the regularization scheme with the sample dependent hypothesis spaces in [10], the hypothesis space in our algorithm (3) is independent of the sample. Hence we do not need to introduce an extra hypothesis error, which is caused by the difference between norms of two different Hilbert spaces.

III. ESTIMATES FOR THE ERROR BOUNDS

We mainly derive the upper bounds for $\mathcal{D}(\lambda)$, $\mathcal{P}(z, \lambda)$ and $\mathcal{S}(z, \lambda)$, respectively.

A. ESTIMATES FOR THE APPROXIMATION ERROR AND THE DRIFT ERROR

For the approximation error, we have the same result as [10, Proposition 3.2].

Proposition 1: Assume $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$, there holds

$$\mathcal{D}(\lambda) \leq C_1 \lambda^{\min\{2r, 1\}}. \tag{16}$$

For the drift error, the estimation is similar to that in [10] with the main differences are the bounds on f_λ and $f_{z,\lambda}$.

Proposition 2: Suppose the sampling process satisfies (5) and (9), K satisfies (6), there holds

$$\mathcal{P}(z, \lambda) \leq \frac{C_2}{m}\left(\lambda^{-\frac{1}{q}} + \lambda^{-\frac{1}{q}}\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}} + \left(\frac{D(\lambda)}{\lambda}\right)^{\frac{2}{q}}\right). \tag{17}$$

Proof: It is known from Proposition 4.1 in [10] that

$$\begin{aligned} &\left\{\left(\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\lambda)\right) - \left(\mathcal{E}_m(\pi_M(f_{z,\lambda})) - \mathcal{E}_m(f_\lambda)\right)\right\} \\ &\leq \frac{1}{m}\sum_{i=1}^m C\alpha^i (3M + \kappa \|f_\lambda\|_K)\{2|f_{z,\lambda}|_{C^s(X)} \\ &\qquad + 2|f_\lambda|_{C^s(X)} + 2|f_\rho|_{C^s(X)} + 4M + 2\kappa \|f_\lambda\|_K\}. \end{aligned} \tag{18}$$

By (9), we have

$$|f_\rho(x)|_{C^s(X)} \leq C_\rho (2M)^{1-s}. \tag{19}$$

By (2) and (6), for any $f \in \mathcal{H}_K$.

$$\begin{aligned} |f|_{C^s(X)} &= \sup_{x,x' \in X} \frac{|f(x) - f(x')|}{|x - x'|} \\ &\leq \sup_{x,x' \in X} \frac{\|f\|_K \sqrt{K(x,x) - 2K(x,x') + K(x',x')}}{|x - x'|} \\ &\leq \kappa_s \|f\|_K. \end{aligned} \tag{20}$$

We directly invoke the following error bound of $\|f_\lambda\|_K$ and $\|f_{z,\lambda}\|_K$ in [24].

$$\|f_\lambda\|_K \leq \left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}, \tag{21}$$

$$\|f_{z,\lambda}\|_K \leq \left(\frac{M^2}{\lambda}\right)^{\frac{1}{q}}. \tag{22}$$

Combining (21) and (22) with (20), we have

$$|f_\lambda|_{C^s(X)} \leq \kappa_s\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}, \tag{23}$$

$$|f_{z,\lambda}|_{C^s(X)} \leq \kappa_s\left(\frac{M^2}{\lambda}\right)^{\frac{1}{q}}. \tag{24}$$

Which implies

$$\begin{aligned}
&\Big[\big(\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\lambda)\big) - \big(\mathcal{E}_m(\pi_M(f_{z,\lambda})) - \mathcal{E}_m(f_\lambda)\big)\Big] \\
&\quad \leq \frac{C_3 C\alpha}{1-\alpha}\frac{1}{m}\left(\lambda^{-\frac{1}{q}} + \lambda^{-\frac{1}{q}}\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}} + \left(\frac{D(\lambda)}{\lambda}\right)^{\frac{2}{q}}\right).
\end{aligned}$$

This proves Proposition 2. ■

B. ESTIMATES FOR THE SAMPLE ERROR

For the sample error, we decompose it into two parts:

$$\begin{aligned}
\mathcal{S}(z,\lambda) &= \{\mathcal{E}_m(\pi_M(f_{z,\lambda})) - \mathcal{E}_m(f_\rho)\} \\
&\quad - \{\mathcal{E}_z(\pi_M(f_{z,\lambda})) - \mathcal{E}_z(f_\rho)\} \\
&\quad + \{\mathcal{E}_z(f_\lambda) - \mathcal{E}_z(f_\rho)\} \\
&\quad - \{\mathcal{E}_m(f_\lambda)) - \mathcal{E}_m(f_\rho)\} \\
&:= \mathcal{S}_1(z,\lambda) + \mathcal{S}_2(z,\lambda).
\end{aligned}$$

To estimate them, we use the same blocking technique in [10], [15], and [25]. The sample points are divided into $2b_m$ blocks of length a_m. Let $Q_k^{a_m}$ be the marginal distribution of block $(z_{(k-1)a_m+1}, z_{(k-1)a_m+2}, \cdots, z_{ka_m})$ and $(z'_1, \cdots, z'_{2b_m a_m})$ be a random sequence with product distribution $\prod_{k=1}^{2b_m} Q_k^{a_m}$, for $1 \leq k \leq 2b_m$. Denote

$$\begin{aligned}
Z_1 &= (z_1, \cdots, z_{a_m}, z_{2a_m+1}, \cdots, z_{3a_m}, \cdots, \\
&\qquad z_{2(b_m-1)a_m+1}, \cdots, z_{2(b_m-1)a_m}), \\
Z_2 &= (z_{a_m+1}, \cdots, z_{2a_m}, z_{3a_m+1}, \cdots, z_{4a_m}, \cdots, \\
&\qquad z_{(2b_m-1)a_m+1}, \cdots, z_{2b_m a_m});
\end{aligned}$$

and

$$\begin{aligned}
Z'_1 &= (z'_1, \cdots, z'_{a_m}, z'_{2a_m+1}, \cdots, z'_{3a_m}, \cdots, z'_{2(b_m-1)a_m+1}, \\
&\qquad \cdots, z'_{2(b_m-1)a_m}), \\
Z'_2 &= (z'_{a_m+1}, \cdots, z'_{2a_m}, z'_{3a_m+1}, \cdots, z'_{4a_m}, \cdots, \\
&\qquad z'_{(2b_m-1)a_m+1}, \cdots, z'_{2b_m a_m}).
\end{aligned}$$

Then the following results on the upper estimates of $\mathcal{S}_1(z,\lambda)$ and $\mathcal{S}_2(z,\lambda)$ are obtained by using the same method as employed in [10].

Proposition 3: Suppose the sampling process satisfies (4), (5) and (9), K satisfies (6) and $\mathcal{H}_K$ satisfies (8), then for any $0 < \delta < 1$, with confidence $1 - \delta/2$,

$$\begin{aligned}
\mathcal{S}_1(z,\lambda) &\leq \frac{1}{2}\{\mathcal{E}(\pi_M(f_{z,\lambda})) - \mathcal{E}(f_\rho)\} + C_{p,\Phi,\rho}\eta_R \\
&\quad + \frac{(192M^2+2)t}{b_m}\log\left(\frac{4}{\delta - 4b_m\beta(a_m)}\right),
\end{aligned} \tag{25}$$

with confidence $1 - \delta/2$,

$$\begin{aligned}
\mathcal{S}_2(z,\lambda) &\leq C_4\left\{b_m^{-1}\left(1 + \left(\frac{D(\lambda)}{\lambda}\right)^{\frac{2}{q}}\right) + D(\lambda)\right\} \\
&\quad \times \log\left(\frac{4}{\delta - 4b_m\beta(a_m)}\right),
\end{aligned} \tag{26}$$

where $\eta_R := \left(\frac{R_\lambda^p}{b_m}\right)^{\frac{2}{2+p}} + \frac{\alpha}{1-\alpha}\frac{1}{m}\max\{R_\lambda, 1\}$ and $R_\lambda := \left(\frac{M^2}{\lambda}\right)^{\frac{1}{q}}$.

Proof: We firstly estimate the bound of $\mathcal{S}_2(z,\lambda)$. Consider $g(z) = (y - f_\lambda(x))^2 - (y - f_\rho(x))^2$, $z = (x,y) \in Z$, we have

$$\left\|g(z) - \int_Z g\,d\rho^{(i)}\right\|_\infty \leq 2\left(3M + \kappa\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}\right)^2 := 2B_\lambda,$$

and

$$\int_Z g^2\,d\rho^{(i)} \leq B_\lambda\int_Z g\,d\rho^{(i)}.$$

We need to invoke the following lemma from [15].

Lemma 1: If g is a measurable function on Z satisfying $\|g(z) - \int_Z g\,d\rho^{(i)}\|_\infty \leq M$, for any $\delta > 0$, with confidence $1 - \delta$, the quantity $\frac{1}{m}\sum_{i=1}^m\left(g(z_i) - \int_Z g\,d\rho^{(i)}\right)$ can be bounded by

$$\begin{aligned}
&\left\{\frac{8}{3}M\log\left(\frac{2}{\delta - 2b_m\beta(a_m)}\right)\right. \\
&\left.+\sqrt{\frac{2}{a_m}\sum_{i=1}^{2a_m b_m}\int_Z g^2\,d\rho^{(i)}\log\left(\frac{2}{\delta - 2b_m\beta(a_m)}\right)} + M\right\}b_m^{-1}.
\end{aligned}$$

Then with confidence $1 - \delta/2$, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m\left(g(z_i) - \int_Z g\,d\rho^{(i)}\right) \\
&\quad \leq \left(\frac{19t}{3} + 2\right)B_\lambda b_m^{-1} + \frac{1}{2a_m b_m}\sum_{i=1}^{2a_m b_m}\int_Z g\,d\rho^{(i)} \\
&\quad \leq \left(\frac{19t}{3} + 2\right)B_\lambda b_m^{-1} + 2\big(\mathcal{E}_m(f_\lambda) - \mathcal{E}_m(f_\rho)\big).
\end{aligned} \tag{27}$$

It has been proved in [10] that

$$\mathcal{E}_m(f_\lambda) - \mathcal{E}_m(f_\rho) \leq \frac{1}{m}\sum_{i=1}^m C\alpha^i\left\|\big(f_\lambda(x) - f_\rho(x)\big)^2\right\|_{C^s(X)} + D(\lambda). \tag{28}$$

By (19),

$$\begin{aligned}\left\|\left(f_\lambda(x)-f_\rho(x)\right)^2\right\|_{C^s(X)} &\leq 2\left(M+\kappa\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}\right)\\ &\quad\times\left(M+(\kappa+\kappa_s)\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}+C_\rho(2M)^{1-s}\right). \qquad (29)\end{aligned}$$

Thus we obtain the upper estimate (26) of $\mathcal{S}_2(z,\lambda)$ by substituting (28) and (29) into (27).

Next we estimate $\mathcal{S}_1(z,\lambda)$. It has been proved in [10] that, with confidence $1-\delta/2$,

$$\begin{aligned}\mathcal{S}_1(z,\lambda) \leq{}& \frac{1}{2}\{\mathcal{E}(\pi_M(f_{z,\lambda}))-\mathcal{E}(f_\rho)\}+C_{p,\Phi,\rho}\eta_R\\ &+\frac{(192M^2+2)}{b_m}\log\left(\frac{4}{\delta-4b_m\beta(a_m)}\right), \qquad (30)\end{aligned}$$

where

$$\eta_R := \left(\frac{R_\lambda^p}{b_m}\right)^{\frac{2}{2+p}}+\frac{\alpha}{1-\alpha}\frac{1}{m}\max\{R_\lambda, 1\}. \qquad (31)$$

Then we can obtain the bound (30) of $\mathcal{S}_1(z,\lambda)$ by only setting $R_\lambda = (\frac{M^2}{\lambda})^{\frac{1}{q}}$. ■

IV. ESTIMATES FOR THE LEARNING RATE

Now we derive the learning rates.

Proof of Theorem 1: Combining the upper bounds of Proposition 1, 2 and 3 with (15), with confidence $1-\delta$, we have

$$\begin{aligned}&\|\pi_M(f_{z,\lambda})-f_\rho\|_{\rho X}^2\\ &\leq D_1 t\left\{\mathcal{D}(\lambda)+\left(m^{-1}\lambda^{-\frac{1}{q}}+m^{-1}\lambda^{-\frac{1}{q}}\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{1}{q}}\right.\right.\\ &\quad\left.\left.+b_m^{-1}\left(\frac{D(\lambda)}{\lambda}\right)^{\frac{2}{q}}\right)+\lambda^{-\frac{1}{q}\frac{2p}{2+p}}b_m^{-\frac{2}{2+p}}+b_m^{-1}\right\}.\end{aligned}$$

Assume a_m satisfies $m^{1-\zeta}\leq a_m < m^{1-\zeta}+1$ with $\zeta\in[0,1]$, $b_m = [\frac{m}{2a_m}]$ and $m\geq 8^{\frac{1}{\zeta}}$. Then

$$\begin{aligned}\frac{1}{b_m} \leq \frac{1}{\frac{m}{2a_m}-1} &\leq \frac{2(m^{1-\zeta}+1)}{m-2m^{1-\zeta}}\\ &\leq \frac{4m^{1-\zeta}}{m-2m^{1-\zeta}} = \frac{4m^{-\zeta}}{1-2m^{-\zeta}}\\ &\leq 8m^{-\zeta}. \qquad (32)\end{aligned}$$

For the case $0<r<1/2$,

$$\begin{aligned}&\|\pi_M(f_{z,\lambda})-f_\rho\|_{\rho X}^2\\ &\leq D_2 t\left\{\lambda^{2r}+m^{-1}\lambda^{\frac{2r-2}{q}}+m^{-\zeta}\lambda^{\frac{2(2r-1)}{q}}+\lambda^{-\frac{1}{q}\frac{2p}{2+p}}m^{-\frac{2\zeta}{2+p}}\right\}.\end{aligned}$$

If $\lambda = m^{-\theta_1}$, then we have

$$\|\pi_M(f_{z,\lambda})-f_\rho\|_{\rho X}^2 \leq D_2 t m^{-\theta}, \qquad (33)$$

where

$$\begin{aligned}\theta = \min&\left\{2r\theta_1,\ 1+\frac{2r-2}{q}\theta_1,\ \zeta\right.\\ &\left.+\frac{2(2r-1)}{q}\theta_1,\ \frac{2\zeta}{2+p}-\frac{2p}{(2+p)q}\theta_1\right\}.\end{aligned}$$

To minimize the learning rate, we take θ as below:

$$\begin{aligned}\theta_{\max} = \min&\left\{\max_{\theta_1}\min\left\{2r\theta_1,\ 1+\frac{2r-2}{q}\theta_1\right\},\right.\\ &\max_{\theta_1}\min\left\{2r\theta_1,\ \zeta+\frac{2(2r-1)}{q}\theta_1\right\},\\ &\left.\max_{\theta_1}\min\left\{2r\theta_1,\ \frac{2\zeta}{2+p}-\frac{2p}{(2+p)q}\theta_1\right\}\right\}.\end{aligned}$$

Let

$$\begin{aligned}2r\theta_1 &= 1+\frac{2r-2}{q}\theta_1,\ 2r\theta_1 = \zeta+\frac{2(2r-1)}{q}\theta_1,\\ 2r\theta_1 &= \frac{2\zeta}{2+p}-\frac{2p}{(2+p)q}\theta_1.\end{aligned}$$

We have

$$\begin{aligned}&\theta_{\max}\\ &=2r\min\left\{\frac{q}{2rq-2r+2},\ \frac{\zeta q}{2rq-2(2r-1)},\ \frac{\zeta q}{rq(2+p)+p}\right\}.\end{aligned}$$

For the case $r\geq 1/2$,

$$\theta = \min\left\{\theta_1,\ 1-\tfrac{1}{q}\theta_1,\ \zeta,\ \tfrac{2\zeta}{2+p}-\tfrac{2p\theta_1}{(2+p)q}\right\}.$$

In the same way, we take

$$\theta_{\max} = \min\left\{\frac{q}{1+q},\ \frac{2\zeta q}{(2+p)q+2p},\ \zeta\right\}$$

to obtain the learning rate.

To ensure $\delta-4b_m\beta(a_m)\geq\frac{\delta}{2}$, from $\beta(a_m)\leq\beta_0(a_m)^{-\gamma}$, we take

$$m\geq\left(\frac{4\beta_0}{\delta}\right)^{\frac{1}{(\gamma+1)(1-\zeta)-1}}, \quad \zeta\in\left(0,\frac{\gamma}{\gamma+1}\right),$$

therefore

$$\log\frac{4}{\delta-4b_m\beta(a_m)}\leq\log\frac{8}{\delta}.$$

We complete the proof of Theorem 1. ■

V. CONCLUSION AND FURTHER DISCUSSION

We obtain the upper error bound of the algorithm (3), $1\leq q\leq 2$, based on the β-mixing and non-identical samples. Moreover, for the identical and independent samples, we can obtain the following result by the same method employed in Theorem 1. We only need take $\alpha=0$ and $\zeta=1$.

$$\|\pi_M(f_{z,\lambda})-f_\rho\|_{\rho X}^2\leq\widetilde{D}'\left(\frac{1}{m}\right)^{\theta'(r)}\log\left(\frac{8}{\delta}\right), \qquad (34)$$

where

$$\theta'(r) = \begin{cases} 2r\min\left\{\frac{q}{2rq-2(2r-1)}, \frac{q}{rq(2+p)+p}\right\}, & 0 < r < \frac{1}{2}; \\ \min\left\{\frac{2q}{(2+p)q+2p}, 1\right\}, & r \geq 1/2. \end{cases}$$

We can see that $\frac{2q}{(2+p)q+2p} \to 1$ as $p \to 0$. This is a satisfactory learning rate.

In some practical applications, we may encounter the other mixing sampling processes such as α-mixing or φ-mixing processes, see [14], [25]. It may be interesting to continue our error analysis for the other weakly dependent samples.

REFERENCES

[1] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2001.

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

[4] S.-G. Lv and Y.-L. Feng, "Integral operator approach to learning theory with unbounded sampling," *Complex Anal. Oper. Theory*, vol. 6, no. 3, pp. 533–548, Jun. 2012.

[5] A. Caponnetto and E. De Vito, "Fast rates for regularized least-squares algorithm," *Found. Comput. Math.*, vol. 7, no. 3, pp. 331–368, Apr. 2005.

[6] F. Cuker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias–variance problem," *Found. Comput. Math.*, vol. 2, no. 4, pp. 413–428, 2002.

[7] E. De Vito, A. Caponnetto, and L. Rosasco, "Model selection for regularized least-squares algorithm in learning theory," *Found. Comput. Math.*, vol. 5, no. 1, pp. 59–85, Feb. 2005.

[8] S. Mendelson and J. Neeman, "Regularization in kernel learning," *Ann. Statist.*, vol. 38, no. 1, pp. 526–565, Jan. 2010.

[9] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, 2006.

[10] Q. Guo, P. Ye, and B. Cai, "Convergence rate for l^q-coefficient regularized regression with non-i.i.d. sampling," *IEEE Access*, vol. 6, pp. 18804–18813, Apr. 2018.

[11] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2133–2145, Nov. 1996.

[12] H. W. Sun and Q. Wu, "Regularized least square regression with dependent samples," *Adv. Comput. Math.*, vol. 32, no. 2, pp. 175–189, Feb. 2010.

[13] X. Chu and H. Sun, "Regularized least square regression with unbounded and dependent sampling," *Abstract Appl. Anal.*, vol. 2013, Mar. 2013, Art. no. 139318.

[14] Z.-W. Pan and Q.-W. Xiao, "Least-square regularized regression with non-iid sampling," *J. Stat. Planning Inference*, vol. 139, no. 10, pp. 3579–3587, Oct. 2009.

[15] Z.-C. Guo and L. Shi, "Classification with non-i.i.d. sampling," *Math. Comput. Model.*, vol. 54, nos. 5–6, pp. 1347–1364, Sep. 2011.

[16] Q. Guo and P. X. Ye, "Coefficient-based regularized regression with dependent and unbounded sampling," *Int. J. Wavelets, Multiresolut. Inf. Process.*, vol. 14, no. 5, pp. 1–14, Sep. 2016.

[17] S. Smale and D. X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, pp. 87–113, Jan. 2009.

[18] L. Shi, "Learning theory estimates for coefficient-based regularized regression," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 252–265, Mar. 2013.

[19] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 2, pp. 286–302, 2011.

[20] S. Lv, D. Shi, Q. Xiao, and M. Zhang, "Sharp learning rates of coefficient-based l^q -regularized regression with indefinite kernels," *Sci. China Math.*, vol. 56, no. 8, pp. 1557–1574, Aug. 2013.

[21] D. X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1743–1752, Jul. 2003.

[22] W. Nie and C. Wang, "Constructive analysis for coefficient regularization regression algorithms," *J. Math. Anal. Appl.*, vol. 431, no. 2, pp. 1153–1171, Nov. 2015.

[23] Y.-L. Feng and S.-G. Lv, "Unified approach to coefficient-based regularized regression," *Comput. Math. Appl.*, vol. 62, no. 1, pp. 506–515, Jul. 2011.

[24] Y. L. Feng, "Least-squares regularized regression with dependent samples and q-penalty," *Appl. Anal.*, vol. 91, no. 5, pp. 979–991, May 2012.

[25] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, Jan. 1994.

QIN GUO received the B.S. degree in mathematics and the M.S. degree in applied mathematics from the University of Jinan, Shandong, China, in 2008 and 2011, respectively. She is currently pursuing the Ph.D. degree with Nankai University, Tianjin, China. His current research interests include approximation theory, machine learning, and data mining.

PEIXIN YE received the M.S. degree in mathematics from Xiamen University, Fujian, China, in 1998, and the Ph.D. degree in mathematics from Beijing Normal University in 2001. He is currently a Full Professor with Nankai University. He has published over 50 journal and conference papers. His current research interests include approximation theory, machine learning, and compressed sensing.