IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2018, accepted March 14, 2018, date of publication March 19, 2018, date of current version April 23, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2817215

Convergence Rate for l^q-Coefficient Regularized Regression With Non-i.i.d. Sampling

QIN GUO[1], PEIXIN YE[1], AND BINLEI CAI[2]
[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
[2]School of Computer Science and Technology, Tianjin University, Tianjin 300071, China

Corresponding author: Peixin Ye (yepx@nankai.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 11271199 and Grant 11671213.

ABSTRACT Many learning algorithms use hypothesis spaces which are trained from samples, but little theoretical work has been devoted to the study of these algorithms. In this paper, we show that mathematical analysis for the kernel-based coefficient least squares for regression with l^q-regularizer, $1 \leq q \leq 2$, which is essentially different from that for algorithms with hypothesis spaces independent of the sample or depending only on the sample size. The error analysis was carried out under the assumption that the samples are drawn from a non-identical sequence of probability measures and satisfy the β-mixing condition. We use the drift error analysis and the independent-blocks technique to deal with the non-identical and dependent setting, respectively. When the sequence of marginal distributions converges exponentially fast in the dual of a Hölder space and the sampling process satisfies polynomially β-mixing, we obtain the capacity dependent error bounds of the algorithm. As a byproduct, we derive a significantly faster learning rate that can be arbitrarily close to the best rate $O(m^{-1})$ for the independent and identical samples.

INDEX TERMS Coefficient-based regularized regression, drift error, learning rate, mixing sequence, uniform concentration inequality.

I. INTRODUCTION

Learning theory aims at finding some relationship between inputs and outputs from observed samples. In this paper we consider the least squares regression problem which is one of the central problems in learning theory and has a variety of applications. It can be formulated as follows.

Let X be a compact metric space and $Y = \mathbb{R}$. Let ρ be a Borel probability measure on $Z = X \times Y$. For a function $f : X \to Y$, the least squares error is defined by

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

For every $x \in X$, let $\rho(\cdot|x)$ be the conditional probability measure induced by ρ on Y. Denote by $L^2_{\rho_X}(X)$ the space of the square integrable functions with respect to ρ_X on X with the norm $\|f(\cdot)\|_{\rho_X} = (\int_X |f(\cdot)|^2 d\rho_X)^{\frac{1}{2}}$, where ρ_X is the marginal distribution of ρ on X. It is well known the regression function f_ρ defined by

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

minimizes the error $\mathcal{E}(f)$ over all $f \in L^2_{\rho_X}(X)$. That is, it is the best one to describe the relation between inputs $x \in X$ and outputs $y \in Y$ in the sense of the least squares error. In regression learning, ρ is unknown and what we have in hand is a set of random samples $z = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ which are drawn independently and identically according to ρ. The task is to find a good approximation f_z of the regression function, which is derived from some learning algorithm, see [1], [2] and the references therein. To measure the approximation ability of f_z, we estimate the excess generalization error

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

In the designation of the learning algorithm, we replace the generalization error $\mathcal{E}(f)$ by the empirical error

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

We expect to find a good approximation of f_ρ by minimizing $\mathcal{E}_z$ in a suitable way.

There is a family of popular learning algorithms which take the form of the regularization schemes in a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel. Such

a kernel K is a continuous, symmetric, and positive semi-definite function on $X \times X$. Let $\mathcal{H}_K$ be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the 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 well-known reproducing property in $\mathcal{H}_K$ takes the form:

$$f(x) = \langle f, K_x \rangle_K, \quad \text{for all } f \in \mathcal{H}_K, x \in X. \tag{1}$$

The least squares regularization scheme with the norm square regularizer is given by

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

where $\lambda > 0$ is the regularization parameter which may depend on the sample size m with $\lim_{m \to \infty} \lambda(m) = 0$. The efficiency of this kind of kernel scheme has been studied in a lot of literatures, see [3]–[5] and the references therein.

Now we consider a different learning scheme, see [6]. In this scheme the data dependent hypothesis space is given by

$$\mathcal{H}_{K,z} = \left\{ f(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i) : \alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m, m \in \mathbb{N} \right\}. \tag{3}$$

We adopt the coefficient-based regularization with l^q-penalization ($1 \leq q \leq 2$) to find the empirical target function

$$f_{z,\eta} = \arg \min_{f \in \mathcal{H}_{K,z}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \eta \Omega_z(f) \right\}, \tag{4}$$

where

$$\Omega_z(f) = \inf \left\{ \sum_{i=1}^{m} |\alpha_i|^q : f = \sum_{i=1}^{m} \alpha_i K_{x_i} \right\}, \quad \eta = \eta(m) > 0. \tag{5}$$

The above algorithm (1.2) can also be rewritten as

$$f_{z,\eta} = \sum_{i=1}^{m} \alpha_i^z K_{x_i},$$

where

$$(\alpha_i^z)_{i=1}^m = \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{k=1}^{m} \left(\sum_{i=1}^{m} \alpha_i K(x_i, x_k) - y_k \right)^2 + \eta \sum_{i=1}^{m} |\alpha_i|^q \right\}, \quad 1 \leq q \leq 2. \tag{6}$$

The application of coefficient-based regularization scheme was first introduced by Vapnik to design linear programming support vector machines, see [7]. In recent years, there has been tremendous interests in studying the error performance of the algorithm (4), see [8]–[18]. Among others, the convergence rates of the algorithm (4) have been obtained in the case of independent samples, see [8], [9], [14]–[18]. However, usually this independent assumption cannot be strictly justified in real-world problems. For example, many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not independent processes, see [19]. Up to now only relatively few results were obtained in the case of dependent samples, see [10], [11], [20]–[22]. Modha and Masry [21] established the minimum complexity regression estimation with m-dependent observations and strongly mixing (α-mixing) observations. Sun and Guo [10], Sun and Wu [20], and Chu and Sun [22] carried out the error analysis of the algorithm (2) and (4) with the strongly and uniformly mixing (ϕ-mixing) samples respectively. Motivated by their work, we consider the following β-mixing sequences. In general, the α-mixing is quite easy to establish but has few consequences. The ϕ-mixing has many nice properties, but few stochastic processes are ϕ-mixing. The β-mixing is neither too weak nor too strong, which is just right. For the details of these mixing conditions and their comparisons, one can refer to [23] and the references therein.

Definition 1: Let $z = \{z_t\}_{t \geq 1}$ be a sequence of random variables. For any $i, j \in \mathbb{N} \cup \{+\infty\}$, σ_i^j denotes the σ-algebra generated by the random variables $\{z_t = (x_t, y_t)\}_{t=i}^j$. Then for any $k \in \mathbb{N}$, the β-mixing coefficients of the stochastic process z are defined as

$$\beta(k) = \sup_{j \geq 1} \mathbb{E} \sup_{A \in \sigma_{j+k}^{\infty}} |P(A|\sigma_1^j) - P(A)|. \tag{7}$$

z is said to be β-mixing, if $\beta(k) \to 0$ as $k \to \infty$. Specifically, it is said to be polynomially β-mixing, if there exists some $\beta_0 > 0$ and $\gamma > 0$ such that, for all $k \geq 1$,

$$\beta(k) \leq \beta_0 k^{-\gamma}. \tag{8}$$

Moreover, identity is a rather restrictive assumption in some real data analysis. Pan and Xiao [8], Smale and Zhou [24], and Guo and Shi [25] considered the non-identical sampling setting for online, classification and least squares regression learning algorithms, respectively. Following their framework, we assume that there is a sequence of Borel probability measures $\{\rho^{(i)}\}_{i=1,2,\ldots}$ on Z. The i-th sample $z_i = (x_i, y_i)$ is drawn according to $\rho^{(i)}$ on Z. Let $\rho_X^{(i)}$ be the marginal distribution of $\rho^{(i)}$. For every $x \in X$, the conditional distribution of $\{\rho^{(i)}\}_{i=1,2,\ldots}$ at x is $\rho(\cdot|x)$, independent of i. It is known from Riesz representation theorem, every probability measure determines a bounded linear functional on $C^s(X)$ via $F(f) = \int_X f d\mu$ for every $f \in C^s(X)$. We make the following assumption about the sequence $\{\rho_X^{(i)}\}$.

Definition 2: We say that the sequence $\{\rho_X^{(i)}\}$ converges to ρ_X exponentially in $(C^s(X))^*$, if there exist $C > 0$ and $0 < \alpha < 1$, such that

$$\|\rho_X^{(i)} - \rho_X\|_{(C^s(X))^*} \leq C\alpha^i, \quad \forall i \in \mathbb{N}. \tag{9}$$

Recall that the Hölder space $C^s(X)$ with $0 \le s \le 1$, consists of all continuous functions on X with the following norm :

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

By the definition of the dual space $(C^s(X))^*$, the condition (9) is equivalent to

$$\left|\int_X f(x)d\rho_X^{(i)} - \int_X f(x)d\rho_X\right| \le C\alpha^i(\|f\|_\infty + |f|_{C^s(X)}),$$
$$\forall f \in C^s(X), \quad i \in \mathbb{N}. \quad (10)$$

To let the readers have a better understanding of the decay condition (9), we cite two examples of sequences of probability distributions satisfying (9) from [24].

The first one is generated by iterations of a stochastic linear operator acting on an initial probability measure.

Example 3: Let ν be a strictly positive probability distribution on X and $\psi \in C(X\times X)$ be strictly positive satisfying $\int_X \psi(x,u)d\nu(u) = 1$ for each $x \in X$. Define the sequence $\{\rho_X^{(t)}\}$ by

$$\rho_X^{(t+1)}(\Gamma) = \int_\Gamma \left\{\int_X (\psi(x,u))d\rho_X^{(t)}(x)\right\}d\nu(u),$$

where $t \in \mathbb{N}$, and $\Gamma \subseteq X$ is a Borel set. Then $\{\rho_X^{(t)}\}$ converges exponentially to some strictly positive probability distribution ρ_X on X.

The second one is induced by dynamical systems.

Example 4: Let $X = [-\frac{1}{2}, \frac{1}{2}]$ and for each $t \in \mathbb{N}$, the probability distribution $\rho_X^{(t)}$ on X has support $[-2^{-t}, 2^{-t}]$ and uniform density 2^{t-1} on its support. Then with δ_0 being the Dirac distribution at the origin, for each $0 < s \le 1$, we have

$$\left|\int_X f(x)d\rho_X^{(t)} - \int_X f(x)d\delta_0\right| \le 2^{t-1}\int_{-2^{-t}}^{2^{-t}} |f(x)-f(0)|dx \le (2^{-s})^t\|f\|_{C^s(X)}.$$

The rest part of the paper is organized as follows. In Section II, we will state the learning rates and the error decomposition of the algorithm (4). In the forthcoming Section III−V, we will derive the upper bound of the approximation error, the hypothesis error, the drift error and the sample error. The result will be proved in Section VI. Finally, we concludes this paper in Section VII.

II. MAIN RESULT AND ERROR DECOMPOSITION

Our principal goal is to derive the upper bound of the error $\|f_{z,\eta} - f_\rho\|_{\rho_X}^2$ under some mild assumptions of f_ρ and $\mathcal{H}_K$. So we first formulate these assumptions.

Let $L_K : L^2_{\rho_X}(X) \to L^2_{\rho_X}(X)$ be the integral operator defined by

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

Since X is compact and K is continuous, L_K is a compact operator. Its fractional power operator $L_K^r : L^2_{\rho_X}(X) \to L^2_{\rho_X}(X)$, $r > 0$ is defined by

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

where $\{\mu_i\}$ are the eigenvalues of the operator L_K and $\{e_i\}$ are the corresponding eigenfunctions which form an orthonormal basis of $L^2_{\rho_X}(X)$, see [8]. For $r > 0$, the function f_ρ is said to satisfy the regularity condition of order r provided that $L_K^{-r} f_\rho \in L^2_{\rho_X}$.

When $K \in C^{2s}(X\times X)$, K satisfies the following condition, see [15].

Definition 5: We say that the kernel K satisfies the kernel condition of order s, if for some $\kappa_s > 0$

$$|K(x,x) - 2K(x,x') + K(x',x')| \le \kappa_s^2|x-x'|^{2s}, \quad \forall x, \quad x' \in X. \quad (11)$$

We also need the following capacity assumption of the unit ball

$$B_1 = \left\{f \in \mathcal{H}_K : \|f\|_K \le 1\right\}. \quad (12)$$

of $\mathcal{H}_K$ measured by the l^2 empirical covering number, see [16].

Capacity Assumption: There exists an exponent p, with $0 < p < 2$ and a constant $c_p > 0$ such that

$$\mathcal{N}_2(B_1, \epsilon) \le c_p\epsilon^{-p}, \quad \forall \epsilon > 0, \quad (13)$$

where c_p is a constant independent of ϵ.

To estimate $|f_\rho(x)|_{C^s(X)}$ and $\left|\int_Y y^2 d\rho(y|x)\right|_{C^s(X)}$ appearing in the proof, we require the Lipschitz s continuity of conditional distribution sequence $\{\rho(y|x) : x \in X\}$.

Definition 6: We say that the sequence $\{\rho(y|x) : x \in X\}$ is Lipschitz s in $(C_s(Y))^*$ if there exists a constant $C_\rho \ge 0$ such that

$$\|\rho(y|x) - \rho(y|u)\|_{(C^s(X))^*} \le C_\rho|x-u|^s, \quad \forall x,\ u \in X. \quad (14)$$

Throughout this paper, we assume $|y| \le M$ almost surely, it is easy to see $|f_\rho(x)| \le M$ for any $x \in X$. Thus we use the following truncation function to improve learning rates, see [12]–[14].

Definition 6: Fix $M > 0$, the truncation function $\pi_M : X \to [-M, M]$ is defined as

$$\pi_M(x) = \begin{cases} M, & if\ x > M, \\ x, & if\ |x| \le M, \\ -M, & if\ x < -M. \end{cases} \quad (15)$$

For a function $f : X \to \mathbb{R}$ and $M > 0$, we define $\pi_M(f)$, the truncation at level M of f, as $\pi_M(f)(x) = \pi_M(f(x))$ for all $x \in X$.

We also assume all the constants in this paper are independent of δ, m, λ or η.

Now we give our main result of the algorithm (4) by choosing the appropriate parameters λ and η according to m.

Theorem 7: Assume the random samples $z_i = (x_i, y_i)$, $i \geq 1$ satisfy the polynomially β-mixing condition, the marginal distribution ρ_X and conditional distribution $\rho(y|x)$ satisfy (9) and (14), respectively, and the sequence $\{\rho_X^{(i)}\}$ converges to ρ_X exponentially in $(C^s(X))^*$. Suppose K satisfies (11), $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$ and the capacity assumption (13) with $0 < p < 2$ holds. If we take $m \geq \left\{ 8^{\frac{1}{\zeta}}, \left(\frac{4\beta_0}{\delta}\right)^{\frac{1}{(\gamma+1)(1-\zeta)-1}} \right\}$ with $\zeta \in \left(0, \frac{\gamma}{\gamma+1}\right)$, then for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\|\pi_M(f_{z,\eta}) - f_\rho\|^2_{\rho_X} \leq \widetilde{D}\left(\frac{1}{m}\right)^{\theta(r)} \log\left(\frac{8}{\delta}\right), \tag{16}$$

where $\theta(r)$ is defined by

$$\theta(r) = \begin{cases} 2r \min\left\{\frac{2q}{2qr+3q+4r}, \frac{2(\zeta-p)}{4r-p}, \zeta\right\}, & 0 < r < \frac{1}{2}; \\ \min\left\{\frac{q}{1+2q}, \frac{2(\zeta-p)}{2-p}, \zeta\right\}, & r \geq 1/2. \end{cases}$$

We will use error decomposition to analyze the excess generalization error. For regularization schemes with sample independent hypothesis spaces such as RKHSs one decompose the total error into the sum of the sample error involving on the sample z and the approximation error which depends on the approximation ability of the hypothesis space $\mathcal{H}_K$ [4]. For coefficient regularization algorithms, we need a new error decomposition technique, that is, an extra hypothesis error should be introduced. Moreover, in our non-identical setting, the main difficulty is that the measures $\{\rho_X^{(i)}\}$ vary and an essential error is caused by the change of these marginal distributions. To describe this error, we introduce

$$\mathcal{E}_m(f) = \frac{1}{m}\sum_{i=1}^{m} \int_Z (f(u) - y)^2 d\rho^{(i)}(u, y). \tag{17}$$

Let

$$f_\lambda := \arg\min_{f \in \mathcal{H}_K} \{\|f - f_\rho\|^2_{\rho_X} + \lambda \|f\|^2_K\}. \tag{18}$$

Then we have the following error decomposition:

$$\begin{aligned} \mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}(f_\rho) + \eta\Omega_z(f_{z,\eta}) \\ &= \mathcal{P}(z,\eta,\lambda) + \mathcal{S}(z,\eta,\lambda) \\ &\quad + \mathcal{H}(z,\eta,\lambda) + \mathcal{D}(\lambda), \end{aligned} \tag{19}$$

where

$$\begin{aligned} \mathcal{P}(z,\eta,\lambda) &= \{\mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}_m(\pi_M(f_{z,\eta}))\} \\ &\quad + \{\mathcal{E}_m(f_\lambda) - \mathcal{E}(f_\lambda)\}, \\ \mathcal{S}(z,\eta,\lambda) &= \{\mathcal{E}_m(\pi_M(f_{z,\eta})) - \mathcal{E}_z(\pi_M(f_{z,\eta}))\} \\ &\quad + \{\mathcal{E}_z(f_\lambda) - \mathcal{E}_m(f_\lambda)\}, \\ \mathcal{H}(z,\eta,\lambda) &= \{\mathcal{E}_z(\pi_M(f_{z,\eta})) + \eta\Omega_z(f_{z,\eta})\} \\ &\quad - \{\mathcal{E}_z(f_\lambda) + \lambda\|f_\lambda\|^2_K\}, \\ \mathcal{D}(\lambda) &= \|f_\lambda - f_\rho\|^2_{\rho_X} + \lambda\|f_\lambda\|^2_K. \end{aligned} \tag{20}$$

The first term $\mathcal{P}(z,\eta,\lambda)$ of the right hand side is called the drift error caused by the drift of non-identical measure $\rho^{(i)}$ from ρ, and the second term $\mathcal{S}(z,\eta,\lambda)$ is called the sample error which is caused by drawing the sample from each $\rho^{(i)}$. The third term $\mathcal{H}(z,\eta,\lambda)$ is known as the hypothesis error. The last term $\mathcal{D}(\lambda)$ is known as the approximation error.

III. ESTIMATES FOR THE APPROXIMATION AND HYPOTHESIS ERROR

The estimate of approximation error relies on the following proposition from [26], see page 273.

Proposition 8: If A is a positive element of a C^*-algebra $\mathfrak{A}$, $sp(A)$ is the spectral set of A, and denote by $C(sp(A))$ the C^*-algebra of all continuous complex-valued functions on $sp(A)$, the mapping $f \mapsto f(A)$ is a $*$ isomorphism from $C(sp(A))$ onto a C^*-subalgebra $\mathfrak{B}$ of $\mathfrak{A}$, then $f(A)$ is self-adjoint and $sp(f(A)) = \{f(t) : t \in sp(A)\}$.

To estimate $\mathcal{D}(\lambda)$, we firstly establish two lemmas.

Lemma 9: Under the assumption $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$, there holds

$$\|f_\lambda - f_\rho\|^2_{\rho_X} \leq C_1 \lambda^{\min\{2r,2\}}. \tag{21}$$

Proof: It has been proved in [27] that $f_\lambda = (\lambda I + L_K)^{-1} L_K f_\rho$, therefore

$$\begin{aligned} f_\lambda - f_\rho &= -\lambda(\lambda I + L_K)^{-1} L_K^r L_K^{-r} f_\rho \\ &= -\sum_{i=1}^{\infty} \langle L_K^{-r} f_\rho, e_i\rangle_{\rho_X} \lambda(\lambda I + L_K)^{-1} L_K^r e_i. \end{aligned}$$

Since L_K is a positive compact operator, by Proposition 8,

$$\|f_\lambda - f_\rho\|^2_{\rho_X} = \sum_{i=1}^{\infty} \frac{\lambda^2 \lambda_i^{2r}}{(\lambda + \lambda_i)^2} \langle L_K^{-r} f_\rho, \quad e_i\rangle^2_{\rho_X}.$$

When $0 < r < 1$,

$$\begin{aligned} \|f_\lambda - f_\rho\|^2_{\rho_X} &= \lambda^{2r} \sum_{i=1}^{\infty} \frac{\lambda^{2-2r}}{(\lambda + \lambda_i)^{2-2r}} \\ &\quad \times \frac{\lambda_i^{2r}}{(\lambda + \lambda_i)^{2r}} \langle L_K^{-r} f_\rho, e_i\rangle^2_{\rho_X} \\ &\leq \lambda^{2r} \|L_K^{-r} f_\rho\|^2_{\rho_X}, \end{aligned}$$

when $r \geq 1$, it is known from [4] that $\sup_{i \geq 1} \lambda_i = \|L_K\| \leq \kappa^2$,

$$\begin{aligned} \|f_\lambda - f_\rho\|^2_{\rho_X} &= \sum_{i=1}^{\infty} \frac{\lambda_i^2}{(\lambda + \lambda_i)^2} \lambda_i^{2r-2} \lambda^2 \langle L_K^{-r} f_\rho, e_i\rangle^2_{\rho_X} \\ &\leq \lambda^2 \sum_{i=1}^{\infty} \lambda_i^{2r-2} \langle L_K^{-r} f_\rho, e_i\rangle^2_{\rho_X} \\ &\leq \lambda^2 \kappa^{2(2r-2)} \|L_K^{-r} f_\rho\|^2_{\rho_X}, \end{aligned}$$

which implies

$$\|f_\lambda - f_\rho\|^2_{\rho_X} \leq C_1 \lambda^{\min\{2r,2\}}.$$

This completes the proof of Lemma 9. ∎

Lemma 10: Under the assumption $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$, there holds

$$\|f_\lambda\|^2_K \leq C_2 \lambda^{\min\{2r-1,0\}}.$$

Proof: By Proposition 8,

$$f_\lambda = \sum_{i=1}^{\infty} \frac{\lambda_i^{1+r}}{\lambda+\lambda_i} \langle L_K^{-r} f_\rho,\ e_i\rangle_{\rho_X} e_i.$$

Using the fact $\{\sqrt{\lambda_i} e_i : i \in \Lambda\}$ forms an orthonormal basis of $\mathcal{H}_K$, see [28], we get

$$\|f_\lambda\|_K^2 = \sum_{i=1}^{\infty} \frac{\lambda_i^{1+2r}}{(\lambda+\lambda_i)^2} \langle L_K^{-r} f_\rho,\ e_i\rangle_{\rho_X}^2.$$

When $0 < r < \frac{1}{2}$,

$$\begin{aligned}\|f_\lambda\|_K^2 &= \sum_{i=1}^{\infty} \frac{\lambda_i^{1+2r}}{(\lambda+\lambda_i)^{1+2r}} \cdot \frac{1}{(\lambda+\lambda_i)^{1-2r}} \langle L_K^{-r} f_\rho,\ e_i\rangle_{\rho_X}^2 \\ &\le \lambda^{2r-1} \|L_K^{-r} f_\rho\|_{\rho_X}^2,\end{aligned}$$

when $r \ge \frac{1}{2}, f_\rho \in \mathcal{H}_K$, thus

$$\|f_\lambda\|_K = \left\|(\lambda I + L_K)^{-1} L_K f_\rho\right\|_K \le \|f_\rho\|_K,$$

which implies

$$\|f_\lambda\|_K^2 \le C_2 \lambda^{\min\{2r-1,0\}}.$$

This proves Lemma 10. ∎

Lemma 9 and Lemma 10 imply the following upper bound for the approximation error.

Proposition 11: Under the assumption $L_K^{-r} f_\rho \in L^2_{\rho_X}$ with $r > 0$, there holds

$$\mathcal{D}(\lambda) \le C_3 \lambda^{\min\{2r,1\}}. \tag{22}$$

For the hypothesis error, we directly invoke the following result on the upper estimate of $\mathcal{H}(z,\eta,\lambda)$ in [12] .

Proposition 12: Under the assumptions of Theorem 7, there holds

$$\mathcal{H}(z,\eta,\lambda) \le \frac{m\eta M^q}{(m\lambda)^q}. \tag{23}$$

IV. ESTIMATES FOR THE DRIFT ERROR

Now we provide the upper bound for $\mathcal{P}(z,\eta,\lambda)$ in the following proposition.

Proposition 13: Under the assumptions of Theorem 7, there holds

$$\begin{aligned}&\left\{\left(\mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}(f_\lambda)\right) - \left(\mathcal{E}_m(\pi_M(f_{z,\eta})) - \mathcal{E}_m(f_\lambda)\right)\right\} \\ &\quad\le \frac{C_4}{m}\left(m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}} + m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}}\sqrt{\frac{D(\lambda)}{\lambda}} + \frac{D(\lambda)}{\lambda}\right).\end{aligned} \tag{24}$$

Proof: By the definitions of $\mathcal{E}(f)$ and $\mathcal{E}_m(f)$, we have

$$\begin{aligned}&\left\{\left(\mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}(f_\lambda)\right) - \left(\mathcal{E}_m(\pi_M(f_{z,\eta})) - \mathcal{E}_m(f_\lambda)\right)\right\} \\ &\quad\le \frac{1}{m}\sum_{i=1}^{m}\left|\int_Z \left[\left(\pi_M(f_{z,\eta})(u) - y\right)^2 - \left(f_\lambda(u) - y\right)^2\right] \right. \\ &\qquad\qquad \left. d\left(\rho(u,y) - \rho^{(i)}(u,y)\right)\right| \\ &\quad= \frac{1}{m}\sum_{i=1}^{m}\left|\int_X \left(\pi_M(f_{z,\eta})(u) - f_\lambda(u)\right)\left(\pi_M(f_{z,\eta})(u)\right.\right. \\ &\qquad\qquad \left.\left. + f_\lambda(u) - 2f_\rho(u)\right) d\left(\rho_X(u) - \rho_X^{(i)}(u)\right)\right|.\end{aligned} \tag{25}$$

By condition (10),

$$\begin{aligned}&\left\{\left(\mathcal{E}(\pi_M(f_{z,\eta})) - \mathcal{E}(f_\lambda)\right) - \left(\mathcal{E}_m(\pi_M(f_{z,\eta})) - \mathcal{E}_m(f_\lambda)\right)\right\} \\ &\quad\le \frac{1}{m}\sum_{i=1}^{m} C\alpha^i \left\|\left(\pi_M(f_{z,\eta})(u) - f_\lambda(u)\right)\left(\pi_M(f_{z,\eta})(u)\right.\right. \\ &\qquad\qquad \left.\left. + f_\lambda(u) - 2f_\rho(u)\right)\right\|_{C^s(X)}.\end{aligned} \tag{26}$$

It is known from [8] that

$$\|fg\|_{C^s(X)} \le \|f\|_{C(X)}\|g\|_{C^s(X)} + \|f\|_{C^s(X)}\|g\|_{C(X)}, \tag{27}$$

therefore,

$$\begin{aligned}&\left\|\left(\pi_M(f_{z,\eta})(u) - f_\lambda(u)\right)\left(\pi_M(f_{z,\eta})(u) + f_\lambda(u) - 2f_\rho(u)\right)\right\|_{C^s(X)} \\ &\quad\le (3M + \kappa\|f_\lambda\|_K)\{2|f_{z,\eta}|_{C^s(X)} + 2|f_\lambda|_{C^s(X)} \\ &\qquad + 2|f_\rho|_{C^s(X)} + 4M + 2\kappa\|f_\lambda\|_K\}.\end{aligned} \tag{28}$$

Next we estimate $|f_\rho|_{C^s(X)}$, $|f_{z,\eta}|_{C^s(X)}$ and $|f_\lambda|_{C^s(X)}$, respectively.

By (14), we have

$$|f_\rho(x)|_{C^s(X)} \le C_\rho (2M)^{1-s}. \tag{29}$$

By (1), for any $f \in \mathcal{H}_K$.

$$\begin{aligned}|f(x) - f(x')| &= |\langle f, K_x - K_{x'}\rangle| \\ &\le \|f\|_K \sqrt{K(x,x) - 2K(x,x') + K(x',x')}.\end{aligned}$$

It follows from (11) that

$$|f|_{C^s(X)} = \sup_{x,x'\in X} \frac{|f(x) - f(x')|}{|x - x'|} \le \kappa_s \|f\|_K. \tag{30}$$

It has been proved in [12] that

$$\|f_\lambda\|_K \le \sqrt{\frac{D(\lambda)}{\lambda}}, \tag{31}$$

$$\|f_{z,\eta}\|_K \le \kappa m^{1-\frac{1}{q}} \left(\frac{M^2}{\eta}\right)^{\frac{1}{q}}. \tag{32}$$

Plugging (31) and (32) into (30), we obtain

$$|f_\lambda|_{C^s(X)} \le \kappa_s \sqrt{\frac{D(\lambda)}{\lambda}}, \tag{33}$$

$$|f_{z,\eta}|_{C^s(X)} \le \kappa_s \kappa m^{1-\frac{1}{q}} \left(\frac{M^2}{\eta}\right)^{\frac{1}{q}}. \tag{34}$$

Plugging (29), (33) and (34) into (28), we have

$$\begin{aligned}&\Big\|\big(\pi_M(f_{z,\eta})(u)-f_\lambda(u)\big)\\&\qquad\big(\pi_M(f_{z,\eta})(u)+f_\lambda(u)-2f_\rho(u)\big)\Big\|_{C^s(X)}\\&\le C_4\bigg(m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}}+m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}}\sqrt{\frac{D(\lambda)}{\lambda}}+\frac{D(\lambda)}{\lambda}\bigg),\end{aligned}\tag{35}$$

then combining with (26), we have

$$\begin{aligned}&\Big[\big(\mathcal{E}(\pi_M(f_{z,\eta}))-\mathcal{E}(f_\lambda)\big)-\big(\mathcal{E}_m(\pi_M(f_{z,\eta}))-\mathcal{E}_m(f_\lambda)\big)\Big]\\&\le\frac{C_5C\alpha}{1-\alpha}\frac{1}{m}\bigg(m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}}+m^{1-\frac{1}{q}}\eta^{-\frac{1}{q}}\sqrt{\frac{D(\lambda)}{\lambda}}+\frac{D(\lambda)}{\lambda}\bigg).\end{aligned}$$

We complete the proof of Proposition 13. ■

V. ESTIMATES FOR THE SAMPLE ERROR

To estimate the sample error, we use the blocking technique in [25] and [29] to deal with the original weakly dependent sequence. Given any integer pair (a_m, b_m) with $b_m = [m/2a_m]$, we divide the sequence into $2b_m$ blocks of length a_m and a remainder block of length $m - 2b_m a_m$. For $1 \le k \le 2b_m$, we denote $Q_k^{a_m}$ the marginal distribution of block $(z_{(k-1)a_m+1}, z_{(k-1)a_m+2}, \cdots, z_{ka_m})$ and take $(z'_1, \cdots, z'_{2b_m a_m})$ to be a random sequence with product distribution $\prod_{k=1}^{2b_m} Q_k^{a_m}$. Define

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

and correspondingly

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

The sample error can be written as

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

We firstly estimate the bound of $\mathcal{S}_2(z,\lambda)$. To do this, we recall the following lemma from [25].

Lemma 14: If g is a measurable function on Z satisfying $\|g(z)-\int_Z g\,d\rho^{(i)}\|_\infty \le M$, for any $\delta > 0$, with confidence $1-\delta$, the quantity $\frac{1}{m}\sum_{i=1}^m\big(g(z_i)-\int_Z g\,d\rho^{(i)}\big)$ can be bounded by

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

We obtain the following result on the upper estimate of $\mathcal{S}_2(z,\lambda)$ by using Lemma 14.

Proposition 15: Under the assumptions of Theorem 7, for any $0 < \delta < 1$, with confidence $1-\delta/2$,

$$\mathcal{S}_2(z,\lambda)\le C_6\bigg\{b_m^{-1}\bigg(1+\frac{D(\lambda)}{\lambda}\bigg)+D(\lambda)\bigg\}t,\tag{36}$$

where $t=\log\big(\frac{4}{\delta-4b_m\beta(a_m)}\big)$.

Proof: For any $z=(u,y)\in Z$, define $g(z)=(y-f_\lambda(u))^2-(y-f_\rho(u))^2$, then

$$\bigg\|g(z)-\int_Z g\,d\rho^{(i)}\bigg\|_\infty\le 2\bigg(3M+\kappa\sqrt{\frac{D(\lambda)}{\lambda}}\bigg)^2:=2B_\lambda,$$

and

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

By applying Lemma 14, with confidence $1-\delta/2$, there holds

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\bigg(g(z_i)-\int_Z g\,d\rho^{(i)}\bigg)\\&\le\bigg(\frac{19t}{3}+2\bigg)B_\lambda b_m^{-1}+\frac{1}{2a_mb_m}\sum_{i=1}^{2a_mb_m}\int_Z g\,d\rho^{(i)}\\&\le\bigg(\frac{19t}{3}+2\bigg)B_\lambda b_m^{-1}+2\big(\mathcal{E}_m(f_\lambda)-\mathcal{E}_m(f_\rho)\big).\end{aligned}\tag{37}$$

Then we estimate $\mathcal{E}_m(f_\lambda)-\mathcal{E}_m(f_\rho)$. Note that

$$\begin{aligned}\mathcal{E}_m(f_\lambda)-\mathcal{E}_m(f_\rho)\le\big(&\mathcal{E}_m(f_\lambda)-\mathcal{E}(f_\lambda)\\&+\mathcal{E}(f_\rho)-\mathcal{E}_m(f_\rho)\big)+D(\lambda).\end{aligned}\tag{38}$$

By (10), we have

$$\begin{aligned}&\mathcal{E}_m(f_\lambda)-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\rho)-\mathcal{E}_m(f_\rho)\\&\le\frac{1}{m}\sum_{i=1}^m\bigg|\int_X\big(f_\lambda(u)-f_\rho(u)\big)^2d\big(\rho_X(u)-\rho_X^{(i)}(u)\big)\bigg|\\&\le\frac{1}{m}\sum_{i=1}^m C\alpha^i\Big\|\big(f_\lambda(u)-f_\rho(u)\big)^2\Big\|_{C^s(X)}.\end{aligned}\tag{39}$$

By (27) and (29),

$$\begin{aligned}&\Big\|\big(f_\lambda(u)-f_\rho(u)\big)^2\Big\|_{C^s(X)}\\&\le 2\bigg(M+\kappa\sqrt{\frac{D(\lambda)}{\lambda}}\bigg)\bigg(M+\kappa\sqrt{\frac{D(\lambda)}{\lambda}}\\&\quad+\kappa_s\sqrt{\frac{D(\lambda)}{\lambda}}+C_\rho(2M)^{1-s}\bigg),\end{aligned}\tag{40}$$

which implies

$$\mathcal{E}_m(f_\lambda)-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\rho)-\mathcal{E}_m(f_\rho)\le\frac{CC_7\alpha}{m(1-\alpha)}\left(1+\frac{\mathcal{D}(\lambda)}{\lambda}\right). \tag{41}$$

By substituting (38) and (41) into (37), we complete the proof of Proposition 15. ∎

Next we estimate $\mathcal{S}_1(z,\eta)$. To deal with the β-mixing sequences, we invoke the following lemma from [25].

Lemma 16: Let $\mathscr{G}$ be a class of measurable functions on Z such that for each $g\in\mathscr{G}$, $\left\|g-\int_Z g d^{(i)}\right\|_\infty\le M$, then

$$Prob\left\{\sup_{g\in\mathscr{G}}\frac{1}{m}\sum_{i=1}^{m}\left(g(z_i)-\int_Z g(z)d\rho^{(i)}\right)>\epsilon+\frac{M}{b_m}\right\}\le\prod_1+\prod_2+2b_m\beta(a_m),$$

where

$$\prod_1=Prob\left\{\sup_{g\in\mathscr{G}}\frac{1}{b_m}\sum_{j=1}^{b_m}\left(\frac{2b_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m}\left(g(z_i')-\int_Z g(z)d\rho^{(i)}\right)\right)\ge\epsilon\right\},$$

$$\prod_2=Prob\left\{\sup_{g\in\mathscr{G}}\frac{1}{b_m}\sum_{j=1}^{b_m}\left(\frac{2b_m}{m}\sum_{i=(2j-1)a_m+1}^{2ja_m}\left(g(z_i')-\int_Z g(z)d\rho^{(i)}\right)\right)\ge\epsilon\right\}.$$

The concentration estimation for $\mathcal{S}_1(z,\eta)$ relies on the following uniform concentration inequality for non-identical sampling.

Proposition 17: Assume $\{X_i\}_{i=1}^n$ is a random sequence in the measurable space $\left(\mathfrak{X}^n,\prod_{i=1}^n Q_i\right)$. Let $\mathscr{F}$ be a set of measurable functions on $\mathfrak{X}$ and $B>0$ be a constant such that each $f\in\mathscr{F}$ satisfies $\|f\|_\infty\le B$. Suppose there exists a nonnegative functional w on $\mathscr{F}$ and some positive constants $(\Delta_i)_{i=1}^n$ such that

$$\mathbb{E}f^2(X_i)\le w(f)+\Delta_i,\forall f\in\mathscr{F}. \tag{42}$$

Also assume for some $a>0$ and $p\in(0,2)$,

$$\log\mathcal{N}_2(\mathscr{F},\varepsilon)\le a\varepsilon^{-p},\quad\forall\varepsilon>0.$$

Then for any $x>0$ and any $D>0$, with probability at least $1-e^{-x}$ there holds

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}f(X_i)-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\le D^{-1}w(f)+c_p'\tilde{\eta}+\frac{(D+18B+2)x}{n},\quad\forall f\in\mathscr{F},$$

where c_p' is a constant depending only on p and

$$\tilde{\eta}:=\max\left\{D^{\frac{2-p}{2+p}},B^{\frac{2-p}{2+p}}+1\right\}\left(\frac{a}{n}\right)^{\frac{2}{p+2}}+\frac{1}{n}\sum_{i=1}^{n}\Delta_i.$$

Now we provide the upper bound for $\mathcal{S}_1(z,\eta)$ in the following proposition.

Proposition 18: Under the assumptions of Theorem 7, for any $0<\delta<1$, with confidence $1-\delta/2$,

$$\{\mathcal{E}_m(\pi_M(f_{z,\eta}))-\mathcal{E}_m(f_\rho)\}-\{\mathcal{E}_z(\pi_M(f_{z,\eta}))-\mathcal{E}_z(f_\rho)\}\le\frac{1}{2}\{\mathcal{E}(\pi_M(f_{z,\eta}))-\mathcal{E}(f_\rho)\}+C_{p,\Phi,\rho}\eta_R+\frac{(192M^2+2)t}{b_m}, \tag{43}$$

where

$$\eta_R:=\left(\frac{R_\eta^p}{b_m}\right)^{\frac{2}{2+p}}+\frac{\alpha}{1-\alpha}\frac{1}{m}\max\{R_\eta,1\} \tag{44}$$

and $t=\log\left(\frac{4}{\delta-4b_m\beta(a_m)}\right)$.

Proof: We apply Proposition 17 to the function set

$$\widetilde{\mathscr{G}}=\left\{G(t_1,\cdots,t_{a_m})=\frac{2b_m}{m}\sum_{k=1}^{a_m}g(t_k):g\in\mathscr{G}\right\}$$

defined on Z^{a_m}, where

$$\mathscr{G}=\left\{g(z)=g(u,y)=(y-\pi_M(f)(u))^2-(y-f_\rho(u))^2:f\in B_R\right\}.$$

Define the functional w on $\widetilde{\mathscr{G}}$ as

$$w(G):=\int_{Z^{a_m}}G^2(t_1,\cdots,t_{a_m})d\rho(t_1)d\rho(t_2)\cdots d\rho(t_{a_m})=\frac{4a_m^2b_m^2}{m^2}\int_Z g^2d\rho.$$

Then

$$\mathbb{E}G^2(z'_{(k-1)a_m+1},z'_{(k-1)a_m+2},\cdots,z'_{ka_m})\le\frac{4b_m^2a_m}{m^2}\sum_{i=(k-1)a_m+1}^{ka_m}\int_Z g^2d\rho^{(i)}\le w(G)+\frac{4b_m^2a_m}{m^2}\sum_{i=(k-1)a_m+1}^{ka_m}\left|\int_Z g^2d(\rho^{(i)}-\rho)\right|.$$

From (10), we know that

$$\left|\int_Z g^2d(\rho^{(i)}-\rho)\right|\le C\alpha^i\left\|(f_\rho(u)-\pi(f)(u))^2\int_Y(2y-\pi(f)(u)-f_\rho(u))^2d\rho(y|u)\right\|_{C^s(X)}.$$

By (14) and (27), we have

$$\left\|(f_\rho(u)-\pi(f)(u))^2\int_Y(2y-\pi(f)(u)-f_\rho(u))^2d\rho(y|u)\right\|_{C^s(X)}\le(68M^2C_\rho(2M)^{1-s}+96M^3\kappa_s+160M^4)(1+R).$$

Thus (42) is satisfied with

$$\Delta_k \le \frac{4b_m^2 a_m}{m^2} C_{\rho,\Phi} \max\{R, 1\} \sum_{i=1}^{a_m} \alpha^{(k-1)a_m+i}.$$

In addition given $d \in \mathbb{N}$ and $w = \{\vec{t}_j = (t_1^j, \cdots, t_{a_m}^j)\}_{j=1}^d \subset (Z^{a_m})^d$, for any $G_1 = \frac{2b_m}{m} \sum_{k=1}^{a_m} g_1(t_k)$ and $G_2 = \frac{2b_m}{m} \sum_{k=1}^{a_m} g_2(t_k)$ in $\widetilde{\mathscr{G}}$, we find

$$\begin{aligned} d_{2,w}^2(G_1, G_2) &= \frac{1}{d} \sum_{j=1}^{d} \left(G_1(\vec{t_j}) - G_2(\vec{t_j}) \right)^2 \\ &= \frac{1}{d} \sum_{j=1}^{d} \left(\frac{2b_m}{m} \sum_{k=1}^{a_m} \left(g_1(t_k^j) - g_2(t_k^j) \right) \right)^2 \\ &\le \frac{1}{da_m} \sum_{j=1}^{d} \sum_{k=1}^{a_m} \left(g_1(t_k^j) - g_2(t_k^j) \right)^2 \\ &= d_{2,w}^2(g_1, g_2), \end{aligned}$$

which implies $\mathscr{N}_2(\widetilde{\mathscr{G}}, \epsilon) \le \mathscr{N}_2(\mathscr{G}, \epsilon)$.

By (1.2),

$$|g_1(z) - g_2(z)| \le 4M |f_1(u) - f_2(u)|,$$

which implies

$$\mathscr{N}_2(\mathscr{G}, \epsilon) \le \mathscr{N}_2(B_R, \frac{\epsilon}{4M}).$$

Thus from (2.12), we have

$$\log \mathscr{N}_2(\mathscr{G}, \epsilon) \le c_p (4M)^p R^P \epsilon^{-p}.$$

Observe that $\|G\|_\infty \le \|g\|_\infty \le 8M^2$ and

$$\begin{aligned} &\mathbb{E} G(z'_{(k-1)a_m+1}, z'_{(k-1)a_m+2}, \cdots, z'_{ka_m}) \\ &\qquad \le \frac{2b_m}{m} \sum_{i=(k-1)a_m+1}^{ka_m} \int_Z g d\rho^{(i)}. \end{aligned}$$

We thus apply Proposition 17 to the functional set $\widetilde{\mathscr{G}}$ in the product measurable space $\left((Z^{a_m})^{b_m}, \prod_{j=1}^{b_m} Q_{2j-1}^{a_m}\right)$ with $B = 8M^2$ and $a = c_p (4M)^p R^p$. Also note that

$$w(G) = \frac{4a_m^2 b_m^2}{m^2} \int_Z g^2 d\rho \le \int_Z g^2 d\rho \le 8M^2 \int_Z g d\rho,$$

then for any $D > 0$, $g \in \mathscr{G}$, with probability at least $1 - e^{-t}$, there holds

$$\begin{aligned} &\frac{1}{b_m} \sum_{j=1}^{b_m} \left(\frac{2b_m}{m} \sum_{i=2(j-1)a_m+1}^{(2j-1)a_m} \left(\int_Z g(z) d\rho^{(i)} - g(z'_i) \right) \right) \\ &\qquad \le \frac{8M^2}{D} \left(\int_Z g d\rho \right) + c'_p \eta_1 + \frac{(D + 144M^2 + 2)t}{b_m}, \end{aligned}$$

where

$$\begin{aligned} \eta_1 = \max &\left\{ D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1 \right\} \left\{ \frac{c_p (4M)^p R^p}{b_m} \right\}^{\frac{2}{2+p}} \\ &+ \frac{4b_m a_m}{m^2} C_{\rho,\Phi} \max\{R, 1\} \sum_{j=1}^{b_m} \sum_{i=1}^{a_m} \alpha^{(2j-2)a_m+i}, \end{aligned}$$

which implies

$$\begin{aligned} Prob\Bigg\{ \sup_{g \in \mathscr{G}} &\frac{1}{b_m} \sum_{j=1}^{b_m} \left(\frac{2b_m}{m} \sum_{i=2(j-1)a_m+1}^{(2j-1)a_m} \left(\int_Z g(z) d\rho^{(i)} \right. \right. \\ &\left. \left. - g(z'_i) \right) \right) - \frac{8M^2}{D} \left(\int_Z g d\rho \right) \ge \epsilon_1 \Bigg\} \le e^{-t}, \end{aligned}$$

where $\epsilon_1 = c'_p \eta_1 + \frac{(D+144M^2+2)t}{b_m}$.

In the same way, we apply Proposition 17 to the functional set $\widetilde{\mathscr{G}}$ in the product measurable space $\left((Z^{a_m})^{b_m}, \prod_{j=1}^{b_m} Q_{2j}^{a_m}\right)$, there holds

$$\begin{aligned} Prob\Bigg\{ \sup_{g \in \mathscr{G}} &\frac{1}{b_m} \sum_{j=1}^{b_m} \left(\frac{2b_m}{m} \sum_{i=(2j-1)a_m+1}^{2ja_m} \left(\int_Z g(z) d\rho^{(i)} \right. \right. \\ &\left. \left. - g(z'_i) \right) \right) - \frac{8M^2}{D} \left(\int_Z g d\rho \right) \ge \epsilon_2 \Bigg\} \le e^{-t}, \end{aligned}$$

where $\epsilon_2 = c'_p \eta_2 + \frac{(D+144M^2+2)t}{b_m}$ with

$$\begin{aligned} \eta_2 = \max &\left\{ D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1 \right\} \left\{ \frac{c_p (4M)^p R^p}{b_m} \right\}^{\frac{2}{2+p}} \\ &+ \frac{4b_m a_m}{m^2} C_{\rho,\Phi} \max\{R, 1\} \sum_{j=1}^{b_m} \sum_{i=1}^{a_m} \alpha^{(2j-1)a_m+i}. \end{aligned}$$

Note that

$$\begin{aligned} &\frac{4b_m a_m}{m^2} \sum_{j=1}^{b_m} \sum_{i=1}^{a_m} \alpha^{(2j-2)a_m+i} \\ &\qquad + \frac{4b_m a_m}{m^2} \sum_{j=1}^{b_m} \sum_{i=1}^{a_m} \alpha^{(2j-1)a_m+i} \le \frac{2}{m} \frac{\alpha}{1-\alpha}, \end{aligned}$$

and

$$\left\| g(z) - \int_Z g(z) d\rho^{(i)} \right\|_\infty < 16M^2.$$

From Lemma 16 by taking $\varepsilon = c'_p \tilde{\eta} + \frac{(D+18\cdot 8M^2+2)t}{b_m}$ with

$$\begin{aligned} \tilde{\eta} = \Bigg\{ &\max \left\{ D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1 \right\} \left\{ \frac{c_p (4M)^p R^p}{b_m} \right\}^{\frac{2}{2+p}} \\ &+ \frac{2}{m} C_{\rho,\Phi} \max\{R, 1\} \frac{\alpha}{1-\alpha} \Bigg\}, \end{aligned}$$

we have

$$\begin{aligned} &Prob\left\{ \sup_{g \in \mathscr{G}} \frac{1}{m} \sum_{j=1}^{m} \left(\int_Z g(z) d\rho^{(i)} - g(z_i) \right) \right. \\ &\qquad \left. - \frac{16M^2}{D} \left(\int_Z g d\rho \right) > \epsilon + \frac{16M^2}{b_m} \right\} \\ &\qquad \le 2e^{-t} + 2b_m \beta(a_m). \end{aligned}$$

Finally we derive our result by setting

$$R = \kappa m^{1-\frac{1}{q}} \left(\frac{M^2}{\eta} \right)^{\frac{1}{q}} := R_\eta, \; 2e^{-t} + 2b_m \beta(a_m) := \frac{\delta}{2},$$

$$D = 32M^2.$$

■

VI. ESTIMATES FOR THE TOTAL ERROR

We are in a position to prove the main result.

Proof of Theorem 7: Putting the estimates in Proposition 11, 12, 13, 15 and 18 into (19), with confidence $1-\delta$, we have

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

We take a_m to satisfy $m^{1-\zeta}\le a_m < m^{1-\zeta}+1$, $\zeta\in[0,1]$ and $b_m=[\frac{m}{2a_m}]$. Assume $m\ge 8^{\frac{1}{\zeta}}$, therefore

$$\frac{1}{b_m}\le 8m^{-\zeta}. \tag{45}$$

When $0<r<1/2$,

$$\begin{aligned}\|\pi_M(f_{z,\eta})-f_\rho\|_{\rho_X}^2 &\le D_2 t\Big\{\lambda^{2r}+m^{-\frac{1}{q}}\eta^{-\frac{1}{q}}\lambda^{r-\frac{1}{2}}\\ &\quad+m^{1-q}\eta\lambda^{-q}+m^{-\zeta}\lambda^{2r-1}\\ &\quad+\eta^{-\frac{1}{q}\frac{2p}{2+p}}m^{(1-\frac{1}{q})(\frac{2p}{2+p})-\frac{2\zeta}{2+p}}\Big\}.\end{aligned}$$

Let $\lambda=m^{-\theta_1}$ and $\eta=m^{-\theta_2}$. Then

$$\|\pi_M(f_{z,\eta})-f_\rho\|_{\rho_X}^2\le D_2 t m^{-\theta}, \tag{46}$$

where

$$\begin{aligned}\theta=\min\Big\{&2r\theta_1,\ \frac{1}{q}+\Big(r-\frac{1}{2}\Big)\theta_1-\frac{1}{q}\theta_2,\\ &q-1-q\theta_1+\theta_2,\ \zeta+(2r-1)\theta_1,\\ &\frac{2\zeta}{2+p}-\frac{2p(q-1)}{(2+p)q}-\frac{2p\theta_2}{(2+p)q}\Big\}.\end{aligned} \tag{47}$$

To get the fastest learning rates, we choose θ as follow:

$$\begin{aligned}\theta_{\max}=\max_{\theta_2}\min\Big\{&\max_{\theta_1}\min\Big\{2r\theta_1,\ \frac{1}{q}+\Big(r-\frac{1}{2}\Big)\theta_1-\frac{1}{q}\theta_2\Big\},\\ &\max_{\theta_1}\min\Big\{2r\theta_1,\ q-1-q\theta_1+\theta_2\Big\},\\ &\max_{\theta_1}\min\Big\{2r\theta_1,\ \zeta+(2r-1)\theta_1\Big\},\\ &\frac{2\zeta}{2+p}-\frac{2p(q-1)}{(2+p)q}-\frac{2p\theta_2}{(2+p)q}\Big\}.\end{aligned} \tag{48}$$

Let

$$\begin{aligned}2r\theta_1&=\frac{1}{q}+\Big(r-\frac{1}{2}\Big)\theta_1-\frac{1}{q}\theta_2,\\ 2r\theta_1&=q-1-q\theta_1+\theta_2,\\ 2r\theta_1&=\zeta+(2r-1)\theta_1.\end{aligned}$$

We have

$$\begin{aligned}\theta_{\max}&=\max_{\theta_2}\min\Big\{\frac{4r(1-\theta_2)}{2rq+q},\ \frac{2r(q-1+\theta_2)}{2r+q},\ 2r\zeta,\\ &\qquad\frac{2\zeta}{2+p}-\frac{2p(q-1)}{(2+p)q}-\frac{2p\theta_2}{(2+p)q}\Big\}\\ &=\min\Big\{\max_{\theta_2}\min\Big\{\frac{4r(1-\theta_2)}{2rq+q},\ \frac{2r(q-1+\theta_2)}{2r+q}\Big\},\\ &\qquad\max_{\theta_2}\min\Big\{\frac{4r(1-\theta_2)}{2rq+q},\ \frac{2\zeta}{2+p}-\frac{2p(q-1)}{(2+p)q}\\ &\qquad-\frac{2p\theta_2}{(2+p)q}\Big\},2r\zeta\Big\}.\end{aligned}$$

Let

$$\begin{aligned}\frac{4r(1-\theta_2)}{2rq+q}&=\frac{2r(q-1+\theta_2)}{2r+q},\\ \frac{4r(1-\theta_2)}{2rq+q}&=\frac{2\zeta}{2+p}-\frac{2p(q-1)}{(2+p)q}-\frac{2p\theta_2}{(2+p)q}.\end{aligned}$$

We have

$$\theta_{\max}=2r\min\Big\{\frac{2q}{2qr+3q+4r},\ \frac{2(\zeta-p)}{4r-p},\ \zeta\Big\}.$$

When $r\ge 1/2$, the inequality (46) holds with

$$\begin{aligned}\theta=\min\Big\{&\theta_1,\ \frac{1}{q}-\frac{1}{q}\theta_2,\quad q-1-q\theta_1+\theta_2,\ \zeta,\\ &\frac{2\zeta}{2+p}-\frac{2p(q-1)}{q(2+p)}-\frac{2p\theta_2}{(2+p)q}\Big\}\end{aligned} \tag{49}$$

Similarly, we choose

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

to minimize the convergence rate.

Finally, we take m large enough to guarantee $\delta-4b_m\beta(a_m)\ge\frac{\delta}{2}$. Since $\beta(a_m)\le\beta_0(a_m)^{-\gamma}$, we require

$$m\ge\Big(\frac{4\beta_0}{\delta}\Big)^{\frac{1}{(\gamma+1)(1-\zeta)-1}},\ \zeta\in\Big(0,\frac{\gamma}{\gamma+1}\Big),$$

thus

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

This completes the proof of Theorem 7. ■

VII. CONCLUSIONS

In this paper, we derive the learning rates for the algorithm (4) with l^q-regularization for the non-identical and dependent samples. To the best of our knowledge, there is no general error analysis of the algorithm (4) that covers the case $1\le q\le 2$ under the conditions (8) and (10). We establish some probability inequalities and use the block technique to estimate the drift error and the sample error. Based on these estimates, we obtain our final results. Comparing with [12], we extend their error analysis to the non-i.i.d. case. In particular, for the i.i.d. case, that is, taking $\alpha=0$ in (9) and

$\zeta = 1$ in (45), we derive the following learning rate by the same method

$$\|\pi_M(f_{z,\eta}) - f_\rho\|_{\rho_X}^2 \leq \tilde{C}t\left(\frac{1}{m}\right)^{\frac{2q}{(2+p)q+2p(1+q)}}.$$

Note that when p tends to 0, the exponent $\frac{2q}{(2+p)q+2p(1+q)}$ tends to 1 which is the best one obtained so far.

Furthermore, Guo and Ye [11] derived the error bounds of the algorithm (4) with $q = 2$ for the strongly and uniformly mixing samples under the generalized moment hypothesis below.

Definition 19: Generalized Moment Hypothesis. There exist two constants $M > 0$ and $p \geq 2$ such that

$$\int_Z |y|^p d\rho \leq M. \tag{50}$$

It may be interesting to extend our above analysis to the case of the non-i.i.d. samples under the hypothesis (50).

REFERENCES

[1] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, Jun. 2017.

[2] X.-W. Chen and X. Lin "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, May 2014.

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

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

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

[6] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Appl.*, vol. 33, no. 1, pp. 82–95, Jan. 1971.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

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

[9] H. Sun and Q. Wu, "Least square regression with indefinite kernels and coefficient regularization," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 1, pp. 96–109, Jan. 2011.

[10] H. Sun and Q. Guo, "Coefficient regularized regression with non-iid sampling," *Int. J. Comput. Math.*, vol. 88, no. 15, pp. 3113–3124, Oct. 2011.

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

[12] 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.

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

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

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

[16] 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, Sep. 2011.

[17] B. H. Sheng, P. X. Ye, and J. L. Wang, "Learning rates for least square regressions with coefficient regularization," *Acta Math. Sin.*, vol. 28, no. 11, pp. 2205–2212, Nov. 2012.

[18] B. Sheng, P. Ye, and W. Yu, "Convergence rate of coefficient regularized kernel-based learning algorithms," *Appl. Math. Inf. Sci.*, vol. 8, no. 2, pp. 885–889, Mar. 2014.

[19] T. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *J. Multivariate Anal.*, vol. 100, no. 1, pp. 175–194, Jan. 2009.

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

[21] 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.

[22] X. Chu and H. Sun, "Regularized least square regression with unbounded and dependent sampling," *Abstract Appl. Anal.*, vol. 2013, pp. 1–7, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1155/2013/139318

[23] R. C. Bradley, "Basic properties of strong mixing conditions. A survey and some open questions," *Probab. Surv.*, vol. 2, pp. 107–144, Nov. 2005, doi: 10.1214/154957805100000104.

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

[25] 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.

[26] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras: Elementary Theory*, vol. 1. San Diego, CA, USA: Academic, 1983.

[27] 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, Oct. 2002.

[28] 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.

[29] 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. Her 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 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.

BINLEI CAI received the M.S. degree from the School of Information Science and Engineering, Yanshan University, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tianjin University, China. His research interests include cloud computing, performance evaluation, and availability modeling.