Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(3): 1387–1404.
DOI: 10.3934/era.2023071
Received: 24 August 2022
Revised: 07 December 2022
Accepted: 14 December 2022
Published: 11 January 2023

Research article

Learning capability of the rescaled pure greedy algorithm with non-iid sampling

Qin Guo[1,*] **and Binlei Cai**[2,*]

[1] School of Science, Shandong Jianzhu University, Jinan 250101, China

[2] Shandong Computer Science Center (National Supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250101, China

* **Correspondence:** Email: guoqin19@sdjzu.edu.cn, caibl@sdas.org.

Abstract: We consider the rescaled pure greedy learning algorithm (RPGLA) with the dependent samples drawn according to a non-identical sequence of probability distributions. The generalization performance is provided by applying the independent-blocks technique and adding the drift error. We derive the satisfactory learning rate for the algorithm under the assumption that the process satisfies stationary β-mixing, and also find that the optimal rate $O(n^{-1})$ can be obtained for i.i.d. processes.

Keywords: rescaled pure greedy algorithm; β-mixing; non-identical sequences; drift error; covering number; learning rate

1. Introduction and assumptions

Greedy learning algorithms, or more specifically, applying greedy algorithms to tackle supervised learning problems, have triggered enormous recent research activities since they possess the lower computational burden [1–4]. Theoretical attempts of greedy learning have been widely concerned recently in the framework of learning theory [1–3, 5, 6]. We consider the learning capability of the rescaled pure greedy algorithm (RPGA) in a non-i.i.d sampling setting, which was initiated by Petrova in [7].

A fast review of regression learning as well as greedy algorithms will be given as follows, respectively. Let X be a compact metric space and $Y = \mathbb{R}$. Let $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ be a stationary real-valued sequence with unknown Borel probability distribution ρ on $Z = X \times Y$.

The generalization error can be defined by

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho, \quad \forall f : X \to Y. \tag{1.1}$$

Minimizing $\mathcal{E}(f)$, we can obtain the following regression function

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

where $\rho(\cdot|x)$ denotes the conditional probability measure (given x) on Y. The empirical error $\mathcal{E}_{\mathbf{z}}(f)$ which is a good approximation of the generalization error $\mathcal{E}(f)$ for a fixed function f on X can be defined by

$$\mathcal{E}_{\mathbf{z}}(f) = \|y - f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2. \tag{1.2}$$

The regression problem in learning theory aims at a good approximation $f_{\mathbf{z}}$ of f_ρ, constructed by learning algorithms. Denote by $L^2_{\rho_X}(X)$ the Hilbert space of the square integrable functions defined on X with respect to the measure ρ_X, where ρ_X denotes the marginal probability distribution on X and $\|f(\cdot)\|_{\rho_X} = (\int_X |f(\cdot)|^2 d\rho_X)^{\frac{1}{2}}$. It is known that, for any $f \in L^2_{\rho_X}(X)$, it holds that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2, \tag{1.3}$$

where

$$\|u\|^2 = E(|u(x)|^2) = \|u\|^2_{\rho_X}.$$

The learning ability of the regression algorithm can be measured by the excess generalization error

$$\|f_{\mathbf{z}} - f_\rho\|^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho).$$

Let $\mathcal{H}$ be a real, separable Hilbert space endowed with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| := \|\cdot\|_{\mathcal{H}} = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$. A set of functions $\mathcal{D} \subset \mathcal{H}$ is called a dictionary if $\|g\| = 1$ for every $g \in \mathcal{D}$, $g \in \mathcal{D}$ implies $-g \in \mathcal{D}$ and the closure of $span(\mathcal{D})$ is $\mathcal{H}$. We define the RPGA($\mathcal{D}$) as follows:

RPGA($\mathcal{D}$):

Step 0: Define $f_0 := 0$.

Step m ($m \geq 1$):

(1) If $f = f_{m-1}$, stop the algorithm and define $f_k = f_{m-1} = f$, for $k \geq m$.

(2) If $f \neq f_{m-1}$, choose a direction $\varphi_m \in D$ such that

$$|\langle f - f_{m-1}, \varphi_m\rangle| = \sup_{\varphi\in D}|\langle f - f_{m-1}, \varphi\rangle|. \tag{1.4}$$

With

$$\lambda_m := \langle f - f_{m-1}, \varphi_m\rangle, \tag{1.5}$$

$$\hat{f}_m := f_{m-1} + \lambda_m\varphi_m, \tag{1.6}$$

$$s_m := \frac{< f, \hat{f}_m >}{\|\hat{f}_m\|^2}, \tag{1.7}$$

define the next approximant to be

$$f_m = s_m\hat{f}_m, \tag{1.8}$$

and proceed to Step $m+1$.

Note that the RPGA uses the just appropriate scaling of the output of the pure greedy algorithm (PGA) which can boost convergence rate of the PGA to the optimal approximation rate $O(m^{-\frac{1}{2}})$ for functions in $\mathcal{A}_1(\mathcal{D})$, see [7].

Throughout this paper, we derive the error bounds under the assumption that $|y| \leq M$ almost surely for $M \geq 0$, hence $|f_\rho(x)| \leq M$ for any $x \in X$. We also define the following truncation function as in [8–10].

Definition 1. *Fix $M > 0$, we define the truncation function π_M on the space of the measurable functions $f : X \to \mathbb{R}$ as*

$$\pi_M(f)(x) = \begin{cases} M, & if\ f(x) > M, \\ f(x), & if\ |f(x)| \leq M, \\ -M, & if\ f(x) < -M. \end{cases} \tag{1.9}$$

Now we use the RPGA to realize the greedy learning. Here we consider leaning by the indefinite kernel $K : X \times X \longrightarrow \mathbb{R}$ [11–14] and define the following hypothesis space by

$$\mathcal{H}_{K,\mathbf{z}} = \left\{ f = \sum_{i=1}^{n} \alpha_i K_{x_i} = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot) : \alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n, n \in \mathbb{N} \right\}, \tag{1.10}$$

where

$$\|f\|_{l_1} := \inf \left\{ \sum_{i=1}^{n} |\alpha_i| : f = \sum_{i=1}^{n} \alpha_i K_{x_i} \in \mathcal{H}_{K,\mathbf{z}} \right\}. \tag{1.11}$$

We now present the rescaled pure greedy learning algorithm (RPGLA) as follows:

Algorithm 1 RPGLA

Input: Given a data set $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$, K, $T > 0$ and the dictionary $\mathcal{D}_n = \{K_{x_i}, i = 1, ..., n\}$

Step 1. Normalization: $\widetilde{K}_{x_i} = \frac{K_{x_i}}{\|K_{x_i}\|_n}$, $i = 1, ..., n$

Dictionary: $\widetilde{\mathcal{D}}_n = \{\widetilde{K}_{x_i} : i = 1, ..., n\}$

Step 2. Computation: Let $\widetilde{f_0} = 0$

for $k = 1, 2, ...,$ the approximation $\widetilde{f_k}$ is generated by the RPGA($\widetilde{\mathcal{D}}_n$)

if $\|y - \widetilde{f_k}\|_n^2 + \|\widetilde{f_k}\|_{l_1} \leq \|y\|_n^2$ and $k \geq T$ **break**

end

Output: $\pi_M(\widetilde{f_k})$

Many greedy learning schemes were recently successfully used for the i.i.d. sampling [1–6, 15]. For example, Barron et al. [5] have used a complexity regularization principle as the stopping criterion and deduced the best learning rate $O(n/\log n)^{-\frac{1}{2}})$ of various greedy algorithms. Lin et al. [3] have provided the learning capability of the relaxed greedy learning algorithm (RGLA) and proved that the learning rate is faster than the order $O(n^{-\frac{1}{2}})$. Their numerous numerical simulation results have confirmed that the relax greedy algorithm (RGA) is more stable in dealing with noisy machine learning problems than the orthogonal greedy algorithm (OGA). Chen et al. [16] have introduced a sparse semi-supervised method to learn the regression functions from samples using the OGA. They can derive the learning rate

$O(n^{-1})$ under mild assumptions. To reduce the computational burden of the OGA, Fang et al. [1] have considered the applications of the orthogonal super greedy algorithm (OSGA) which selects more than one atoms from a dictionary in each iteration in supervise learning and deduced an almost same learning rate as that of the orthogonal greedy learning algorithm (OGLA) in [5]. Different from the traditional variants RGA and OGA, Xu et al. [4] proposed the truncated greedy algorithm (TGA) which truncates the step size of the PGA at a specified value in each greedy iteration to cut down the model complexity. They also proved that for some specified learning tasks, the truncated greedy learning algorithm (TGLA) can remove the logarithmic factor in the learning rates of the OGLA and the RGLA. All these results show that in the realm of supervised learning, each greedy algorithm possesses its own pros and cons. For instance, compared with the OGA, the PGA and the RGA have benefits in computation but suffer from the low convergence rate. In this paper, we study the learning capability of the RPGA which is the very simple modified version of the PGA. Motivated by the researches of [7], we proceed to deduce the error bound of the RPGLA. Our results will show that the RPGLA furthermore reduce the computational burden without sacrificing the generalization capability when compared with the OGLA and the RGLA. However, usually the independent and identity assumption is rather restrictive. For example, in [17–19], the authors presented the non-i.i.d. sampling setting for different learning algorithms, respectively. We shall study β-mixing and non-identical sampling, see [20] and the references therein for the details.

Definition 2. *Let $\mathbf{z} = \{z_t\}_{t\geq 1}$ 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 $l \in \mathbb{N}$, the β-mixing coefficients of the stochastic process $\mathbf{z}$ are defined as*

$$\beta(l) = \sup_{j\geq 1} \mathbb{E} \sup_{A\in\sigma_{j+l}^{\infty}} |P(A|\sigma_1^j) - P(A)|. \tag{1.12}$$

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

$$\beta(l) \leq \beta_0 l^{-\gamma}. \tag{1.13}$$

The β-mixing condition is "just the right" assumption, which has been adopted in previous studies for learning with weakly dependent samples, see [18, 21] and the references therein. It is quite easy to establish and covers a more general non-i.i.d. cases such as Gaussian and Markov processes. Markov chains appear so often and naturally in applications, especially in marking prediction, biological speech recognition, sequence analysis, content-based web search and character recognition.

We assume that $\{z_i\}_{i=1}^n$ is drawn according to the Borel probability measures $\{\rho^{(i)}\}_{i=1,2,\cdots}$ on Z. Let $\rho_X^{(i)}$ be the marginal distribution of $\rho^{(i)}$. For every $x \in X$, the conditional distribution of $\{\rho^{(i)}\}_{i=1,2,\cdots}$ at x is $\rho(\cdot|x)$.

Definition 3. *We say that $\{\rho_X^{(i)}\}$ converges to ρ_X exponentially in $(C^s(X))^*$, if for $C > 0$ and $0 < \alpha < 1$,*

$$\|\rho_X^{(i)} - \rho_X\|_{(C^s(X))^*} \leq C\alpha^i, \forall i \in \mathbb{N}. \tag{1.14}$$

The above condition (1.14) is also equivalent to

$$\left| \int_X f(x)d\rho_X^{(i)} - \int_X f(x)d\rho_X \right| \leq C\alpha^i(\|f\|_\infty + |f|_{C^s(X)}), \forall f \in C^s(X),\ i \in \mathbb{N}, \tag{1.15}$$

where

$$\|f\|_{C^s(X)} := \|f\|_\infty + |f|_{C^s(X)}, \tag{1.16}$$

and

$$|f|_{C^s(X)} := \sup_{x\neq y\in X} \frac{|f(x)-f(y)|}{(d(x,y))^s}. \tag{1.17}$$

Before giving our key analysis, we firstly need to impose some mild assumptions concerning K, $\mathcal{H}_{K,\mathbf{z}}$ and $\{\rho(y|x) : x \in X\}$ below.

The kernel function K is said to satisfy a Lipschitz condition of order (α,β) with $0 < \alpha, \beta \leq 1$, if for some $C_\alpha, C_\beta > 0$,

$$|K(x,t) - K(x,t')| \leq C_\alpha |t-t'|^\alpha, \forall x,t,t' \in X, \tag{1.18}$$

$$|K(x,t) - K(x',t)| \leq C_\beta |x-x'|^\beta, \forall t,x,x' \in X. \tag{1.19}$$

Let $R > 0$ and B_R be the ball of $\mathcal{H}_{K,\mathbf{z}}$ with radius R:

$$B_R = \left\{ f \in \mathcal{H}_{K,\mathbf{z}} : \|f\|_{l_1} \leq R \right\}. \tag{1.20}$$

As [22], we give the complexity assumption of the unit ball B_1.

Capacity assumption. We say that B_1 has polynomial complexity exponent $0 < p < 2$ if there is some constant $c_p > 0$ such that

$$\log \mathcal{N}_2(B_1,\epsilon) \leq c_p \epsilon^{-p}, \ \forall \epsilon > 0. \tag{1.21}$$

The following concept describes the continuity of $\{\rho(y|x) : x \in X\}$.

Definition 4. *We say that $\{\rho(y|x) : x \in X\}$ satisfies a Lipschitz condition of order s in $(C_s(Y))^*$ if there is some constant $C_\rho \geq 0$ such that*

$$\|\rho(y|x) - \rho(y|u)\|_{(C^s(Y))^*} \leq C_\rho |x-u|^s, \ \forall x,u \in X. \tag{1.22}$$

Throughout this paper, we denote $\kappa^2 = \sup_{t,x\in X} |K(x,t)|$. Since all the constants are independent of δ, n or λ, for simplicity of notation, we denote by C all the constants.

The rest of this paper is organized as follows: in Section 2, we will state the error decomposition of the algorithm (1) and the rate of uniform convergence. In the forthcoming Sections 3–5, we will analyze the drift error, the sample error and the hypothesis error. Finally, we conclude the main results in Section 6.

2. Error decomposition and main results

We use the developed technique for coefficient regularization algorithms for the non-i.i.d. sampling [19,21] to analyze the learning ability of the algorithm (1). We first define the following function space

$$\mathcal{H}_1 = \left\{ f : f = \sum_{j=1}^\infty \alpha_j \overline{K}_{u_j} : \{\alpha_j\} \in l_1, \{u_j\} \subset X, \overline{K}_{u_j} = \frac{K_{u_j}}{\|K_{u_j}\|_{\rho_X}} \right\}, \tag{2.1}$$

with the norm

$$\|f\|_{\mathcal{H}_1} := \inf\Big\{\sum_{j=1}^{\infty}|\alpha_j| : f = \sum_{j=1}^{\infty}\alpha_j\overline{K}_{u_j}\Big\}. \tag{2.2}$$

We define the regularizing function

$$f_\lambda := \arg\min_{f\in\mathcal{H}_1}\{\mathcal{E}(f) + \lambda\|f\|_{\mathcal{H}_1}\}, \lambda > 0. \tag{2.3}$$

In order to describe the error caused by the change of $\{\rho_X^{(i)}\}$, we introduce

$$\mathcal{E}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\int_Z (f(u)-y)^2 d\rho^{(i)}(u,y). \tag{2.4}$$

Now we can give the error decomposition for the algorithm (1).

$$\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho) \le \mathcal{P}(\lambda) + \mathcal{S}(\mathbf{z},\lambda) + \mathcal{H}(\mathbf{z},\lambda) + \mathcal{D}(\lambda), \tag{2.5}$$

where

$$\begin{aligned}
\mathcal{P}(\lambda) &= \{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}_n(\pi_M(\widetilde{f_k}))\} + \{\mathcal{E}_n(f_\lambda) - \mathcal{E}(f_\lambda)\},\\
\mathcal{S}(\mathbf{z},\lambda) &= \{\mathcal{E}_n(\pi_M(\widetilde{f_k})) - \mathcal{E}_{\mathbf{z}}(\pi_M(\widetilde{f_k}))\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_n(f_\lambda)\},\\
\mathcal{H}(\mathbf{z},\lambda) &= \{\mathcal{E}_{\mathbf{z}}(\pi_M(\widetilde{f_k})) - \mathcal{E}_{\mathbf{z}}(f_\lambda)\},\\
\mathcal{D}(\lambda) &= \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\|_{\mathcal{H}_1}.
\end{aligned} \tag{2.6}$$

The drift error $\mathcal{P}(\lambda)$ describes the change of $\rho^{(i)}$ from ρ, and the sample error $\mathcal{S}(\mathbf{z},\lambda)$ connects the estimator $\pi_M(\widetilde{f_k})$ with f_λ. $\mathcal{H}(\mathbf{z},\lambda)$ and $\mathcal{D}(\lambda)$ are known as the hypothesis error and the approximation error, respectively.

To compared with the main results in [16], we shall assume $\mathcal{D}(\lambda)$ satisfies the same decay rate as follows

$$\mathcal{D}(\lambda) \le c_q\lambda^q, \ \forall\ 0 < \lambda \le 1, \tag{2.7}$$

for some exponent $0 < q \le 1$ and a constant $c_q > 0$.

Next we can state the generalization error bound and give the proofs in Sections 3–6.

Theorem 1. *Assume $z_i = (x_i, y_i)_{i=1}^n$ satisfy condition (1.13), the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ satisfies the capacity assumption (1.21) with $0 < p < 2$, the kernel K satisfies a Lipschitz condition of order (α,β) with $0 < k_0 \le K(u,v) \le k_1$ for any $u, v \in X$, the target function f_ρ can be approximated with the exponent $0 < q \le 1$ in $\mathcal{H}_1$, (1.14) for ρ_X and (1.22) for $\rho(y|x)$ hold. Take $k \ge T \ge n$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} \le Ct\{\lambda^q + b_n^{-1}\lambda^{2q-2} + b_n^{-\frac{2}{2+p}}\}, \tag{2.8}$$

where $t = \log\left(\frac{6}{\delta - 6b_n\beta(a_n)}\right)$ with b_n and a_n given explicitly later.

Theorem 2. *Under the assumptions of Theorem 1, if*

$$n \geq \left\{8^{\frac{1}{\zeta}}, \left(\frac{6\beta_0}{\delta}\right)^{\frac{1}{(\gamma+1)(1-\zeta)-1}}\right\},\ \zeta \in \left(0, \frac{\gamma}{\gamma+1}\right), \tag{2.9}$$

then we obtain

$$\|\pi_M(\widetilde{f_k}) - f_\rho\|_{\rho_X}^2 \leq \widetilde{D} n^{-\theta'} \log\left(\frac{12}{\delta}\right), \tag{2.10}$$

where

$$\theta' = \min\left\{\frac{q\zeta}{2-q}, \frac{2\zeta}{2+p}\right\}.$$

Let $\alpha = 0$ and $\zeta = 1$. Then we obtain the following learning rate of the i.i.d. sampling

$$\|\pi_M(\widetilde{f_k}) - f_\rho\|_{\rho_X}^2 \leq \widetilde{C}\left(\frac{1}{n}\right)^{\min\left\{\frac{q}{2-q}, \frac{2}{2+p}\right\}} \log\left(\frac{12}{\delta}\right),$$

which is the same as that in [16]. In particular, as $p \to 0$, $\frac{2}{2+p} \to 1$ which is the optimal convergence rate.

3. Estimates for the drift error

Proposition 3. *Under the assumptions of Theorem 1, the inequality*

$$\mathcal{P}(\lambda) \leq \frac{C\lambda^{2q-2}}{n}, \tag{3.1}$$

holds.

Proof. By (1.1) and (2.4), we get

$$\begin{aligned}
&\left\{(\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\lambda)) - (\mathcal{E}_n(\pi_M(\widetilde{f_k})) - \mathcal{E}_n(f_\lambda))\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\int_Z \left[(\pi_M(\widetilde{f_k})(u) - y)^2 - (f_\lambda(u) - y)^2\right] d(\rho(u,y) - \rho^{(i)}(u,y))\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|\int_X (\pi_M(\widetilde{f_k})(u) - f_\lambda(u))(\pi_M(\widetilde{f_k})(u) + f_\lambda(u) - 2f_\rho(u)) d(\rho_X(u) - \rho_X^{(i)}(u))\right|.
\end{aligned} \tag{3.2}$$

Now (1.15) tells us that

$$\begin{aligned}
&\left\{(\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\lambda)) - (\mathcal{E}_n(\pi_M(\widetilde{f_k})) - \mathcal{E}_n(f_\lambda))\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} C\alpha^i \left\|(\pi_M(\widetilde{f_k})(u) - f_\lambda(u))(\pi_M(\widetilde{f_k})(u) + f_\lambda(u) - 2f_\rho(u))\right\|_{C^s(X)} \\
&\leq \frac{C}{n}\frac{\alpha}{1-\alpha}(3M + \|f_\lambda\|_\infty)\{2|\widetilde{f_k}|_{C^s(X)} + 2|f_\lambda|_{C^s(X)} + 2|f_\rho|_{C^s(X)} + 4M + 2\|f_\lambda\|_\infty\},
\end{aligned} \tag{3.3}$$

where the last inequality holds true since $\|fg\|_{C^s(X)} \leq \|f\|_{C(X)}\|g\|_{C^s(X)} + \|f\|_{C^s(X)}\|g\|_{C(X)}$, see [19].

In the following, we estimate $\|f_\lambda\|_\infty$,$|f_\lambda|_{C^s(X)}$,$|\widetilde{f_k}|_{C^s(X)}$ and $|f_\rho|_{C^s(X)}$ separately. Let $f_\lambda(x) = \sum_{j=1}^{\infty} \alpha_{j,\lambda}\overline{K}_{u_j}(x)$, $\{\alpha_{j,\lambda}\} \in l_1$. It follows that

$$\begin{aligned} |f_\lambda(x)| &\leq \frac{\kappa}{\|K_{u_j}\|_{\rho_X}} \sum_{j=1}^{\infty} |\alpha_{j,\lambda}| \\ &\leq \frac{\kappa}{\|K_{u_j}\|_{\rho_X}} \|f_\lambda\|_{\mathcal{H}_1}. \end{aligned} \tag{3.4}$$

Furthermore,

$$\|f_\lambda\|_\infty \leq \frac{\kappa}{\|K_{u_j}\|_{\rho_X}} \frac{\mathcal{D}(\lambda)}{\lambda}. \tag{3.5}$$

The Lipschitz condition (1.18) of the kernel function K yields for any $f \in \mathcal{H}_1$ that

$$|f(x) - f(x')| \leq \frac{C_\alpha |x - x'|^s}{\|K_{u_j}\|_{\rho_X}} \|f\|_{\mathcal{H}_1}, \forall x, x' \in X.$$

Together with (1.17), this implies that

$$\begin{aligned} |f_\lambda|_{C^s(X)} &\leq \frac{C_\alpha \|f_\lambda\|_{\mathcal{H}_1}}{\|K_{u_j}\|_{\rho_X}} \\ &\leq \frac{C_\alpha}{\|K_{u_j}\|_{\rho_X}} \frac{\mathcal{D}(\lambda)}{\lambda}. \end{aligned} \tag{3.6}$$

In the same way, from the definition of $\widetilde{f_k}$, we have

$$\begin{aligned} |\widetilde{f_k}|_{C^s(X)} &\leq C_\alpha \|\widetilde{f_k}\|_{l_1} \\ &\leq C_\alpha \|y\|_n^2 \\ &\leq C_\alpha M^2. \end{aligned} \tag{3.7}$$

In addition, combining (1.17) with (1.22) gives

$$\begin{aligned} |f_\rho|_{C^s(X)} &= \sup_{x,x' \in X} \frac{|\int_Y y d\rho(y|x) - \int_Y y d\rho(y|x')|}{|x - x'|^s} \\ &\leq \frac{\|y\|_{C^s(Y)} C_\rho |x - x'|^s}{|x - x'|^s} \\ &\leq C_\rho (M + (2M)^{1-s}). \end{aligned} \tag{3.8}$$

Plugging (3.5), (3.6), (3.7) and (3.8) into (3.3), the desired estimate (3.1) follows, and the proposition is proved.

4. Estimates for the sample error

In our analysis, we apply the method in [18, 23] to deal with the original weakly dependent sequence. Let (a_n, b_n) be any integer pair with $b_n = [n/2a_n]$. The dependent observations are split into $2b_n$ blocks, each of size a_n. For $1 \le k \le 2b_n$, $Q_k^{a_n}$ denotes the marginal distribution of block $(z_{(k-1)a_n+1}, z_{(k-1)a_n+2}, \cdots, z_{ka_n})$. With the constructed blocks, one can then take a new sequence $(z'_1, \cdots, z'_{2b_n a_n})$ with product distribution $\prod_{k=1}^{2b_n} Q_k^{a_n}$. We further define

$$\begin{aligned} Z_1 &= (z_1, \cdots, z_{a_n}, z_{2a_n+1}, \cdots, z_{3a_n}, \cdots, z_{2(b_n-1)a_n+1}, \cdots, z_{2(b_n-1)a_n}), \\ Z_2 &= (z_{a_n+1}, \cdots, z_{2a_n}, z_{3a_n+1}, \cdots, z_{4a_n}, \cdots, z_{(2b_n-1)a_n+1}, \cdots, z_{2b_n a_n}), \end{aligned}$$

and correspondingly

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

The sample error $\mathcal{S}(\mathbf{z}, \lambda)$ can be rewritten as

$$\begin{aligned} \mathcal{S}(\mathbf{z}, \lambda) &= \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}_n(f_\lambda)) - \mathcal{E}_n(f_\rho)\} \\ &\quad + \{\mathcal{E}_n(\pi_M(\widetilde{f_k})) - \mathcal{E}_n(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi_M(\widetilde{f_k})) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} \\ &:= \mathcal{S}_1(\mathbf{z}, \lambda) + \mathcal{S}_2(\mathbf{z}, k). \end{aligned}$$

We analyze the term $\mathcal{S}_1(\mathbf{z}, \lambda)$ by using the following inequality from [18].

Lemma 4.1. *If g is a measurable function on Z satisfying $\left\|g(z) - \int_Z g d\rho^{(i)}\right\|_\infty \le M$, then for any $\delta > 0$, with confidence $1 - \delta$, there holds*

$$\frac{1}{n}\sum_{i=1}^{n}(g(z_i) - \int_Z g d\rho^{(i)}) \le b_n^{-1}\left\{\frac{8}{3}M\log\left(\frac{2}{\delta - 2b_n\beta(a_n)}\right) + \sqrt{\frac{2}{a_n}\sum_{i=1}^{2a_n b_n}\int_Z g^2 d\rho^{(i)}\log\left(\frac{2}{\delta - 2b_n\beta(a_n)}\right)} + M\right\}.$$

Proposition 4. *Under the assumptions of Theorem 1, for any $0 < \delta < 1$, with confidence $1 - \delta/3$,*

$$\mathcal{S}_2(\mathbf{z}, \lambda) \le C\left\{b_n^{-1}\left(1 + \frac{D(\lambda)^2}{\lambda^2}\right) + D(\lambda)\right\}t. \tag{4.1}$$

Proof. Let $g(z) = (y - f_\lambda(u))^2 - (y - f_\rho(u))^2$, $z = (u, y) \in Z$. Thus

$$\left\|g(z) - \int_Z g d\rho^{(i)}\right\|_\infty \le 2\left(3M + \frac{\kappa}{\|K_{u_j}\|_{\rho_X}}\frac{\mathcal{D}(\lambda)}{\lambda}\right)^2 := 2B_\lambda$$

and

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

Using Lemma 4.1, with confidence $1 - \delta/3$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(g(z_i) - \int_Z g d\rho^{(i)}\right)$$

$$\leq \Big(\frac{19t}{3}+2\Big)B_\lambda b_n^{-1} + \frac{1}{2a_nb_n}\sum_{i=1}^{2a_nb_n}\int_Z g d\rho^{(i)}$$
$$\leq \Big(\frac{19t}{3}+2\Big)B_\lambda b_n^{-1} + 2(\mathcal{E}_n(f_\lambda)-\mathcal{E}_n(f_\rho)). \tag{4.2}$$

Observe that

$$\mathcal{E}_n(f_\lambda)-\mathcal{E}_n(f_\rho) \leq (\mathcal{E}_n(f_\lambda)-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\rho)-\mathcal{E}_n(f_\rho)) + D(\lambda). \tag{4.3}$$

By (1.15), we have

$$\begin{aligned}
&\mathcal{E}_n(f_\lambda)-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\rho)-\mathcal{E}_n(f_\rho)\\
&\leq \frac{1}{n}\sum_{i=1}^n \Big|\int_X (f_\lambda(u)-f_\rho(u))^2 d(\rho_X^{(i)}(u)-\rho_X(u))\Big|\\
&\leq \frac{1}{n}\sum_{i=1}^n C\alpha^i \Big\|(f_\lambda(u)-f_\rho(u))^2\Big\|_{C^s(X)}\\
&\leq \frac{C\alpha}{n(1-\alpha)}\Big(1+\frac{\mathcal{D}(\lambda)}{\lambda}\Big)^2,
\end{aligned} \tag{4.4}$$

where the last inequality follows from (3.6) and (3.8).

Combining (4.2), (4.3) and (4.4), we get the desired error bound (4.1) of $\mathcal{S}_1(\mathbf{z},\lambda)$. Proposition 4 is proved.

We continue to analyze $\mathcal{S}_2(\mathbf{z},k)$ by applying the following probability inequality for the β-mixing sequences from [18].

Lemma 4.2. *Let $\mathscr{G}$ be a class of measurable functions on Z. Moreover, assume that $\big\|g-\int_Z g d^{(i)}\big\|_\infty \leq M$ for all $g\in\mathscr{G}$. Then*

$$Prob\Big\{\sup_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^n\Big(g(z_i)-\int_Z g(z)d\rho^{(i)}\Big) > \epsilon+\frac{M}{b_n}\Big\} \leq \prod_1 + \prod_2 + 2b_n\beta(a_n),$$

where

$$\prod_1 = Prob\Big\{\sup_{g\in\mathscr{G}}\frac{1}{b_n}\sum_{j=1}^{b_n}\Big(\frac{2b_n}{n}\sum_{i=2(j-1)a_n+1}^{(2j-1)a_n}\Big(g(z_i')-\int_Z g(z)d\rho^{(i)}\Big)\Big)\geq\epsilon\Big\},$$

$$\prod_2 = Prob\Big\{\sup_{g\in\mathscr{G}}\frac{1}{b_n}\sum_{j=1}^{b_n}\Big(\frac{2b_n}{n}\sum_{i=(2j-1)a_n+1}^{2ja_n}\Big(g(z_i')-\int_Z g(z)d\rho^{(i)}\Big)\Big)\geq\epsilon\Big\}.$$

To get the upper bounds of the terms $\prod_1$ and $\prod_2$, we need to invoke the following inequality for the non-identical sequence of probability distributions.

Proposition 5. *Assume $\{X_i\}_{i=1}^n$ is a random sequence in the measurable space $(\mathfrak{X}^n, \prod_{i=1}^n Q_i)$. 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 $\|f\|_\infty \leq B$. Suppose there exists a nonnegative functional w on $\mathscr{F}$ and some positive constants $(\Delta_i)_{i=1}^n$ such that*

$$\mathbb{E}f^2(X_i) \leq w(f)+\Delta_i, \forall f\in\mathscr{F}. \tag{4.5}$$

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

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

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

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

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

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

The above inequalities imply the estimate of $\mathcal{S}_2(\mathbf{z}, k)$.

Proposition 6. *Under the assumptions of Theorem 1, for any $0 < \delta < 1$, with confidence $1 - \delta/3$,*

$$\mathcal{S}_2(\mathbf{z}, k) \le \frac{1}{2}\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} + C_{p,\Phi,\rho}\eta_R + \frac{(192M^2 + 2)t}{b_n}, \tag{4.6}$$

where

$$\eta_R := \left(\frac{R^p}{b_n}\right)^{\frac{2}{2+p}} + \frac{\alpha}{1-\alpha}\frac{1}{n}\max\{R, 1\}. \tag{4.7}$$

Proof. Define the function set $\widetilde{\mathscr{G}}$ on Z^{a_n} by

$$\widetilde{\mathscr{G}} = \left\{G(t_1, \cdots, t_{a_n}) = \frac{2b_n}{n}\sum_{k=1}^{a_n} g(t_k) : g \in \mathscr{G}, \mathscr{G} = \left\{g(z) = g(u, y) = (y - \pi_M(f)(u))^2 - (y - f_\rho(u))^2 : f \in B_R\right\}\right\}$$

and

$$\begin{aligned} w(G) :&= \int_{Z^{a_n}} G^2(t_1, \cdots, t_{a_n})d\rho(t_1)d\rho(t_2)\cdots d\rho(t_{a_n}) \\ &= \frac{4a_n^2b_n^2}{n^2}\int_Z g^2 d\rho. \end{aligned}$$

It follows that

$$\begin{aligned} &\mathbb{E}G^2(z'_{(k-1)a_n+1}, z'_{(k-1)a_n+2}, \cdots, z'_{ka_n}) \\ &\le \frac{4b_n^2a_n}{n^2}\sum_{i=(k-1)a_n+1}^{ka_n}\int_Z g^2 d\rho^{(i)} \\ &\le w(G) + \frac{4b_n^2a_n}{n^2}\sum_{i=(k-1)a_n+1}^{ka_n}\left|\int_Z g^2 d(\rho^{(i)} - \rho)\right|. \end{aligned} \tag{4.8}$$

We see from (1.15) and (1.22) that

$$\left|\int_Z g^2 d(\rho^{(i)}-\rho)\right| \le C\alpha^i \left\|(f_\rho(u)-\pi_M(f)(u))^2 \int_Y (2y-\pi_M(f)(u)-f_\rho(u))^2 d\rho(y|u)\right\|_{C^s(X)}$$
$$\le C\alpha^i(1+R). \tag{4.9}$$

By (4.8) and (4.9), we know that Δ_k in (4.5) satisfies

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

Let $w = \{\vec{t_j} = (t_1^j, \cdots, t_{a_n}^j)\}_{j=1}^d \subset (Z^{a_n})^d$, $d \in \mathbb{N}$. We know that for any functions $G_1 = \frac{2b_n}{n}\sum_{k=1}^{a_n} g_1(t_k)$ and $G_2 = \frac{2b_n}{n}\sum_{k=1}^{a_n} g_2(t_k)$ in $\widetilde{\mathscr{G}}$,

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

so

$$\mathscr{N}_2(\widetilde{\mathscr{G}},\epsilon) \le \mathscr{N}_2(\mathscr{G},\epsilon). \tag{4.10}$$

Moreover,

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

This together with (4.10) yields

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

Note that $\|G\|_\infty \le \|g\|_\infty \le 8M^2$. It is also easy to see that

$$\mathbb{E}G(z'_{(k-1)a_n+1}, z'_{(k-1)a_n+2}, \cdots, z'_{ka_n}) \le \frac{2b_n}{n}\sum_{i=(k-1)a_n+1}^{ka_n}\int_Z g d\rho^{(i)},$$

and

$$w(G) = \frac{4a_n^2b_n^2}{n^2}\int_Z g^2 d\rho \le \int_Z g^2 d\rho \le 8M^2\int_Z g d\rho.$$

Now applying Proposition 5 to $\widetilde{\mathscr{G}}$ in $((Z^{a_n})^{b_n}, \prod_{j=1}^{b_n} Q_{2j-1}^{a_n})$. Let $B = 8M^2$ and $a = c_p(4M)^pR^p$. Then for any $D > 0$, $g \in \mathscr{G}$, with confidence at least $1-e^{-t}$, we have

$$\frac{1}{b_n}\sum_{j=1}^{b_n}\left(\frac{2b_n}{n}\sum_{i=2(j-1)a_n+1}^{(2j-1)a_n}\left(\int_Z g(z)d\rho^{(i)} - g(z_i')\right)\right) \le \frac{8M^2}{D}\left(\int_Z g d\rho\right) + c_p'\eta_1 + \frac{(D+144M^2+2)t}{b_n}.$$

Here

$$\eta_1 = \max\left\{D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1\right\}\left\{\frac{c_p(4M)^p R^p}{b_n}\right\}^{\frac{2}{2+p}} + \frac{4b_n a_n}{n^2} C_{\rho,\Phi} \max\{R, 1\} \sum_{j=1}^{b_n} \sum_{i=1}^{a_n} \alpha^{(2j-2)a_n+i}.$$

It follows by taking $\epsilon_1 = c'_p \eta_1 + \frac{(D+144M^2+2)t}{b_n}$ that

$$Prob\left\{ \sup_{g\in\mathscr{G}} \frac{1}{b_n} \sum_{j=1}^{b_n} \left(\frac{2b_n}{n} \sum_{i=2(j-1)a_n+1}^{(2j-1)a_n} \left(\int_Z g(z) d\rho^{(i)} - g(z'_i) \right) \right) - \frac{8M^2}{D} \left(\int_Z g d\rho \right) \geq \epsilon_1 \right\} \leq e^{-t}.$$

Applying Proposition 5 to $\widetilde{\mathscr{G}}$ in $((Z^{a_n})^{b_n}, \prod_{j=1}^{b_n} Q_{2j-1}^{a_n})$ once again, we have

$$Prob\left\{ \sup_{g\in\mathscr{G}} \frac{1}{b_n} \sum_{j=1}^{b_n} \left(\frac{2b_n}{n} \sum_{i=(2j-1)a_n+1}^{2ja_n} \left(\int_Z g(z) d\rho^{(i)} - g(z'_i) \right) \right) - \frac{8M^2}{D} \left(\int_Z g d\rho \right) \geq \epsilon_2 \right\} \leq e^{-t}.$$

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

$$\eta_2 = \max\left\{D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1\right\}\left\{\frac{c_p(4M)^p R^p}{b_n}\right\}^{\frac{2}{2+p}} + \frac{4b_n a_n}{n^2} C_{\rho,\Phi} \max\{R, 1\} \sum_{j=1}^{b_n} \sum_{i=1}^{a_n} \alpha^{(2j-1)a_n+i}.$$

Moreover, we obviously have

$$\begin{aligned} &\frac{4b_n a_n}{n^2} \sum_{j=1}^{b_n} \sum_{i=1}^{a_n} \alpha^{(2j-2)a_n+i} + \frac{4b_n a_n}{n^2} \sum_{j=1}^{b_n} \sum_{i=1}^{a_n} \alpha^{(2j-1)a_n+i} \\ &\leq \frac{2}{n} \frac{\alpha}{1-\alpha}, \end{aligned}$$

and

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

We know from Lemma 4.2 by taking $\varepsilon = c'_p \tilde{\eta} + \frac{(D+144M^2+2)t}{b_n}$ with

$$\tilde{\eta} = \left\{ \max\left\{D^{\frac{2-p}{2+p}}, (8M^2)^{\frac{2-p}{2+p}} + 1\right\}\left\{\frac{c_p(4M)^p R^p}{b_n}\right\}^{\frac{2}{2+p}} + \frac{2}{n} C_{\rho,\Phi} \max\{R, 1\} \frac{\alpha}{1-\alpha} \right\},$$

then

$$Prob\left\{ \sup_{g\in\mathscr{G}} \frac{1}{n} \sum_{j=1}^{n} \left(\int_Z g(z) d\rho^{(i)} - g(z_i) \right) - \frac{16M^2}{D} \left(\int_Z g d\rho \right) > \epsilon + \frac{16M^2}{b_n} \right\} \leq 2e^{-t} + 2b_n \beta(a_n).$$

Then we obtain (4.6) by taking $2e^{-t} + 2b_n\beta(a_n) := \frac{\delta}{3}$ and $D = 32M^2$.

5. Estimates for the hypothesis error

Different from the widely regularized method with data-dependent hypothesis spaces [8,10,21,22], our estimation for the hypothesis error $\mathcal{E}_{\mathbf{z}}(\pi_M(\widetilde{f}_k)) - \mathcal{E}_{\mathbf{z}}(f_\lambda)$ is based on the following lemma, see Theorem 3.3 in [7].

Lemma 5.1. *If $f \in \mathcal{H}$, $h \in \mathcal{H}_1^n$, then the output $(f_m)_{m\geq 0}$ of the RPGA satisfies the inequality*

$$\|f - f_m\|^2 - \|f - h\|^2 \leq \frac{4}{m+1}\|h\|^2_{\mathcal{H}_1^n},\ m = 0, 1, 2, \cdots, \tag{5.1}$$

where

$$\mathcal{H}_1^n = \left\{h = \sum_i \alpha_i^n \overline{K}^n_{u_i} : \alpha_i^n = \alpha_i \|\overline{K}_{u_i}\|_n, \overline{K}^n_{u_i} = \frac{\overline{K}_{u_i}}{\|\overline{K}_{u_i}\|_n}, \sum_i \alpha_i \overline{K}_{u_i} \in \mathcal{H}_1\right\} \tag{5.2}$$

with

$$\|f\|_{\mathcal{H}_1^n} := \inf\left\{\sum_i |\alpha_i^n| : f = \sum_i \alpha_i \overline{K}_{u_i}\right\}. \tag{5.3}$$

Proposition 7. *Under the assumptions of Theorem 1, for $k \geq T$ and any $0 < \delta < 1$, with the confidence at least $1 - \delta/3$, there holds*

$$\mathcal{H}(\mathbf{z}, \lambda) \leq 4 \min\left\{\left\{\left(\frac{19t}{3} + 2\right)Mb_n^{-1} + M + \frac{\alpha}{n(1-\alpha)}\left(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\right) + 1\right\}^2, \frac{k_1^2}{k_0^2}\right\}\frac{\mathcal{D}^2(\lambda)}{(k+1)\lambda^2}. \tag{5.4}$$

Proof. By Lemma 5.1, we have

$$\mathcal{H}(\mathbf{z}, \lambda) = \{\mathcal{E}_{\mathbf{z}}(\pi_M(\widetilde{f}_k)) - \mathcal{E}_{\mathbf{z}}(f_\lambda)\} \leq 4\frac{\|f_\lambda\|^2_{\mathcal{H}_1^n}}{k+1}. \tag{5.5}$$

From the definitions of $\|f\|_{\mathcal{H}_1^n}$ and $\|f\|_{\mathcal{H}_1}$, we have

$$\|f_\lambda\|^2_{\mathcal{H}_1^n} \leq \frac{k_1^2}{k_0^2}\|f_\lambda\|^2_{\mathcal{H}_1}. \tag{5.6}$$

Meanwhile, we define the function $g(x) = |\overline{K}_{u_i}(x)|^2$, for any i. Notice that

$$\left\|g(x) - \int_X g d\rho_X^{(j)}\right\|_\infty \leq 2\frac{k_1^2}{k_0^2} := 2M,$$

and

$$\int_Z g^2 d\rho_X^{(j)} \leq M \int_Z g d\rho_X^{(j)}.$$

Using Lemma 4.1, with confidence $1 - \delta/3$, we have

$$\frac{1}{n}\sum_{j=1}^n \left(g(x_j) - \int_X g d\rho_X^{(j)}\right) \leq \left(\frac{19t}{3} + 2\right)Mb_n^{-1} + M. \tag{5.7}$$

By (1.17) and (1.15), we get

$$\frac{1}{n}\sum_{j=1}^{n}\Big(\int_X g d\rho_X^{(j)} - \int_X g d\rho_X\Big) \le \frac{1}{n}\sum_{j=1}^{n} C\alpha^j(\|g\|_\infty + |g|_{C^s(X)}) \le \frac{\alpha}{n(1-\alpha)}\Big(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\Big). \tag{5.8}$$

This in connection with (5.7) tells us that

$$\frac{1}{n}\sum_{j=1}^{n}|\overline{K}(u_i,x_j)|^2 - E\overline{K}_{u_i}^2 = \frac{1}{n}\sum_{j=1}^{n}\Big(g(x_j) - \int_X g d\rho_X\Big) \le \Big(\frac{19t}{3}+2\Big)Mb_n^{-1} + M + \frac{\alpha}{n(1-\alpha)}\Big(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\Big). \tag{5.9}$$

It is easy to see that $\|\overline{K}_{u_j}\|_{\rho_X}^2 = E\overline{K}_{u_j}^2 = 1$. Now (5.9) implies that

$$\begin{aligned}\|\overline{K}_{u_i}\|_n &= \sqrt{\frac{1}{n}\sum_{j=1}^{n}|\overline{K}(u_i,x_j)|^2}\\ &\le \frac{1}{n}\sum_{j=1}^{n}|\overline{K}(u_i,x_j)|^2\\ &\le \Big(\frac{19t}{3}+2\Big)Mb_n^{-1} + M + \frac{\alpha}{n(1-\alpha)}\Big(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\Big) + 1.\end{aligned} \tag{5.10}$$

Therefore,

$$\|f_\lambda\|_{\mathcal{H}_1^n}^2 \le \Big\{\Big(\frac{19t}{3}+2\Big)Mb_n^{-1} + M + \frac{\alpha}{n(1-\alpha)}\Big(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\Big) + 1\Big\}^2 \|f_\lambda\|_{\mathcal{H}_1}^2. \tag{5.11}$$

Combining (5.5), (5.6), (5.11), we obtain (5.4).

6. Proofs of main results

Proof of Theorem 1. Combining the bounds (2.7), (3.1), (4.1), (4.6) and (5.4), with confidence at least $1-\delta$,

$$\begin{aligned}\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} \le{}& c_q\lambda^q + 4\min\Big\{\Big\{\Big(\frac{19t}{3}+2\Big)Mb_n^{-1} + M\\ &+ \frac{\alpha}{1-\alpha}\Big(\frac{k_1^2}{k_0^2} + \frac{2C_\alpha k_1}{k_0^2}\Big) + 1\Big\}^2, \frac{k_1^2}{k_0^2}\Big\}\frac{c_q^2\lambda^{2q}}{(k+1)\lambda^2}\\ &+ \frac{C\lambda^{2q-2}}{n} + C\Big\{b_n^{-1}\Big(1 + \frac{c_q^2\lambda^{2q}}{\lambda^2}\Big) + c_q\lambda^q\Big\}t\\ &+ \frac{1}{2}\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} + C_{p,\Phi,\rho}\eta_R + \frac{(192M^2+2)t}{b_n}.\end{aligned} \tag{6.1}$$

Note that $k \geq T \geq n$. By taking $R = M^2$, then

$$\begin{aligned}\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} &\leq t\Big\{(k+1)^{-1}\lambda^{2q-2} + n^{-1}\lambda^{2q-2} + b_n^{-1}\lambda^{2q-2} \\ &\quad + \lambda^q + b_n^{-\frac{2}{2+p}} + n^{-1} + b_n^{-1}\Big\} \\ &\leq Ct\Big\{\lambda^q + b_n^{-1}\lambda^{2q-2} + b_n^{-\frac{2}{2+p}}\Big\}. \end{aligned} \tag{6.2}$$

This finishes the proof of Theorem 1.

Proof of Theorem 2. Under the conditions of Theorem 1, let $n^{1-\zeta} \leq a_n < n^{1-\zeta} + 1$, $\zeta \in [0, 1]$ and $n \geq 8^{\frac{1}{\zeta}}$. Then

$$\begin{aligned}\frac{1}{b_n} &\leq \frac{1}{\frac{n}{2a_n} - 1} \leq \frac{2(n^{1-\zeta} + 1)}{n - 2n^{1-\zeta}} \\ &\leq \frac{4n^{1-\zeta}}{n - 2n^{1-\zeta}} = \frac{4n^{-\zeta}}{1 - 2n^{-\zeta}} \\ &\leq 8n^{-\zeta}. \end{aligned} \tag{6.3}$$

Substitute (6.3) into (6.2), we obtain

$$\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} \leq Ct\{\lambda^q + n^{-\zeta}\lambda^{2q-2} + n^{-\frac{2\zeta}{2+p}}\}. \tag{6.4}$$

By setting $\lambda = n^{-\theta}$, we know that

$$\{\mathcal{E}(\pi_M(\widetilde{f_k})) - \mathcal{E}(f_\rho)\} \leq D_2 t n^{-\theta'}, \tag{6.5}$$

where

$$\theta' = \min\Big\{q\theta,\ \ \zeta - (2-2q)\theta,\ \ \frac{2\zeta}{2+p}\Big\}.$$

To balance the errors in (2.5), we take $\theta = \frac{\zeta}{2-q}$. Then

$$\theta' = \min\left\{\frac{q\zeta}{2-q},\ \frac{2\zeta}{2+p}\right\}.$$

Finally, we choose

$$n \geq \Big(\frac{6\beta_0}{\delta}\Big)^{\frac{1}{(\gamma+1)(1-\zeta)-1}},\ \zeta \in \Big(0, \frac{\gamma}{\gamma+1}\Big),$$

it follows from $\beta(a_n) \leq \beta_0 (a_n)^{-\gamma}$ and $a_n \geq n^{1-\zeta}$ that

$$\frac{12 b_n \beta(a_n)}{\delta} \leq 1,$$

thus

$$t = \log \frac{6}{\delta - 6b_n\beta(a_n)} \leq \log \frac{12}{\delta}.$$

This finishes the proof of Theorem 2.

Acknowledgments

This research is supported by the National Science Foundation for Young Scientists of China (Grant No. 12001328), Doctoral Research Fund of Shandong Jianzhu University (No. XNBS1942), the Development Plan of Youth Innovation Team of University in Shandong Province (No. 2021KJ067) and Shandong Provincial Natural Science Foundation of China (Grant No. ZR2022MF223). All authors contributed substantially to this paper, participated in drafting and checking the manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that they have no competing of interests regarding the publication of this paper.

References

1. J. Fang, S. B. Lin, Z. B. Xu, Learning and approximation capabilities of orthogonal super greedy algorithm, *Knowl-Based Syst.*, **95** (2016), 86–98. https://doi.org/10.1016/j.knosys.2015.12.011

2. H. Chen, L. Q. Li, Z. B. Pan, Learning rates of multi-kernel regression by orthogonal greedy algorithm, *J. Stat. Plan. Infer.*, **143** (2013), 276–282. https://doi.org/10.1016/j.jspi.2012.08.002

3. S. B. Lin, Y. H. Rong, X. P. Sun, Z. B. Xu, Learning capability of relaxed greedy algorithms, *IEEE Trans. Neur. Net. Lear.*, **24** (2013), 1598–1608. https://doi.org/10.1109/TNNLS.2013.2265397

4. L. Xu, S. B. Lin, Z. B. Xu, Learning capability of the truncated greedy algorithm, *Sci. China Inform. Sci.* **59** (2016), 052103. https://doi.org/10.1007/s11432-016-5536-6

5. A. R. Barron, A. Cohen, W. Dahmen, R. A. DeVore, Approximation and learning by greedy algorithms, *Ann. Statist.*, **36** (2008), 64–94. https://doi.org/10.1214/009053607000000631

6. L. Xu, S. B. Lin, J. S. Zeng, X. Liu, Z. B. Xu, Greedy criterion in orthogonal greedy learning, *IEEE Trans. Cybernetics*, **48** (2018), 955–966. https://doi.org/10.1109/TCYB.2017.2669259

7. G. Petrova, Rescaled pure greedy algorithm for Hilbert and Banach spaces, *Appl. Comput. Harmon. Anal.*, **41** (2016), 852–866. https://doi.org/10.1016/j.acha.2015.10.008

8. S. G. Lv, D. M. Shi, Q. W. Xiao, M. S. Zhang, Sharp learning rates of coefficient-based l^q-regularized regression with indefinite kernels, *Sci. China Math.*, **56** (2013), 1557–1574. https://doi.org/10.1007/s11425-013-4688-8

9. Y. L. Feng, S. G. Lv, Unified approach to coefficient-based regularized regression, *Comput. Math. Appl.*, **62** (2011), 506–515. https://doi.org/10.1016/j.camwa.2011.05.034

10. W. L. Nie, C. Wang, Constructive analysis for coefficient regularization regression algorithms, *J. Math. Anal. Appl.*, **431** (2015), 1153–1171. https://doi.org/10.1016/j.jmaa.2015.06.006

11. H. W. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularization, *Appl. Comput. Harmon.*, **30** (2011), 96–109. https://doi.org/10.1016/j.acha.2010.04.001

12. B. Schölkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, 2002.

13. C. J. Liu, Gabor-based kernel pca with fractional power polynomial models for face recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **26** (2004), 572–581. https://doi.org/10.1109/TPAMI.2004.1273927

14. R. Opfer, Multiscale kernels, *Adv. Comput. Math.*, **25** (2006), 357–380. https://doi.org/10.1007/s10444-004-7622-3

15. A. R. Barron, Universal approximation bounds for superposition of a sigmoidal function, *IEEE Trans. Inform. Theory*, **39** (1993), 930–945. https://doi.org/10.1109/18.256500

16. H. Chen, Y. C. Zhou, Y. Y. Tang, Convergence rate of the semi-supervised greedy algorithm, *Neural Networks*, **44** (2013), 44–50. https://doi.org/10.1016/j.neunet.2013.03.001

17. S. Smale, D. X. Zhou, Online learning with markov sampling, *Anal. Appl.*, **7** (2009), 87–113. https://doi.org/10.1142/S0219530509001293

18. Z. C. Guo, L. Shi, Classification with non-i.i.d. sampling, *Math. Comput. Model.*, **54** (2011), 1347–1364. https://doi.org/10.1016/j.mcm.2011.03.042

19. Z. W. Pan, Q. W. Xiao, Least-square regularized regression with non-iid sampling, *J. Stat. Plan. Infer.*, **139** (2009), 3579–3587. https://doi.org/10.1016/j.jspi.2009.04.007

20. R. C. Bradley, Basic properties of strong mixing conditions, *Progr. Probab. Statist.*, **2** (1986), 165–192. https://doi.org/10.1007/978-1-4615-8162-8_8

21. Q. Guo, P. X. Ye, B. L. Cai, Convergence Rate for l^q-Coefficient Regularized Regression With Non-i.i.d. Sampling, *IEEE Access*, **6** (2018), 18804–18813. https://doi.org/10.1109/ACCESS.2018.2817215

22. L. Shi, Y. L. Feng, D. X. Zhou, Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces, *Appl. Comput. Harmon. Anal.*, **31** (2011), 286–302. https://doi.org/10.1016/j.acha.2011.01.001

23. B. Yu, Rates of convergence for empirical processes of stationary mixing sequences, *Ann. Probab.*, **22** (1994), 94–116. https://www.jstor.org/stable/2244496

AIMS Press