Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 139318, 7 pages
http://dx.doi.org/10.1155/2013/139318

Research Article

Regularized Least Square Regression with Unbounded and Dependent Sampling

Xiaorong Chu and Hongwei Sun

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

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

Received 29 October 2012; Revised 22 March 2013; Accepted 22 March 2013

Academic Editor: Changbum Chun

This paper mainly focuses on the least square regression problem for the α-mixing and ϕ-mixing processes. The standard bound assumption for output data is abandoned and the learning algorithm is implemented with samples drawn from dependent sampling process with a more general output data condition. Capacity independent error bounds and learning rates are deduced by means of the integral operator technique.

1. Introduction and Main Results

The aim of this paper is to study the least square regularized regression learning algorithm. The main novelty of this problem here is the unboundedness and dependence of the sampling process. Let X be a compact metric space (usually a subset of $\mathbb{R}^n$) and $Y = \mathbb{R}$. Suppose that ρ is a probability distribution defined on $Z = X \times Y$. In regression learning, one wants to learn or approximate the regression function $f_\rho : X \to Y$ given by

$$f_\rho(x) = \mathbb{E}(y \mid x) = \int_Y y \, d\rho(y \mid x), \quad x \in X, \tag{1}$$

where $\rho(y \mid x)$ is the conditional distribution of y for given x. f_ρ is not directly computable because ρ is unknown in fact. Instead we learn a good approximation of f_ρ from a set of observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ drawn according to ρ.

The learning algorithm studied here is based on a Mercer kernel $K : X \times X \to \mathbb{R}$ which is a continuous, symmetric, and positive semidefinite function. The RKHS $\mathscr{H}_K$ associated with the Mercer kernel K is the completion of span $\{K_x = K(\cdot, x) : x \in X\}$ with the inner product satisfying $\langle K(x,\cdot), K(x',\cdot)\rangle_K = K(x,x')$. The learning algorithm is a regularization scheme in $\mathscr{H}_K$ given by

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

where $\lambda > 0$ is a regularization parameter.

Error analysis for learning algorithm (2) has been studied in a lot of literatures [1–4], which focused on independent samples. In recent years, there are some studies relaxing the independent restriction and turning to the dependent sampling learning [5–8]. In [8] the learning performance of regularized least square regression was studied with the mixing sequences, and the result for this setting was refined by an operator monotone inequality in [7].

For a stationary real-valued sequence $\{z_i\}_{i\geq 1}$, the σ-algebra generated by the random variables $z_a, z_{a+1}, \ldots, z_b$ is denoted by $\mathscr{M}_a^b$. The uniformly mixing condition (or ϕ-mixing condition) and the strongly mixing condition (or α-mixing condition) are defined as follows.

Definition 1 (ϕ-mixing). The lth ϕ-mixing coefficient for the sequence is defined as

$$\phi_l = \sup_{k\geq 1} \sup_{A \in \mathscr{M}_1^k, B \in \mathscr{M}_{k+l}^\infty} |P(A \mid B) - P(A)|. \tag{3}$$

The process $\{z_i\}_{i\geq 1}$ is said to satisfy a uniformly mixing condition (or ϕ-mixing condition) if $\phi_l \to 0$, as $l \to \infty$.

Definition 2 (α-mixing). The lth α-mixing coefficient for random sequence $\{z_i\}_{i\geq 1}$ is defined as

$$\alpha_l = \sup_{k\geq 1} \sup_{A\in\mathcal{M}_1^k, B\in\mathcal{M}_{k+l}^\infty} |P(A\cap B) - P(A)P(B)|. \tag{4}$$

The random process $\{z_i\}_{i\geq 1}$ is said to satisfy a strongly mixing condition (or α-mixing condition) if $\alpha_l \to 0$, as $l \to \infty$.

By the fact $P(A\cap B) = P(A \mid B)P(B)$, α-mixing condition is weaker than ϕ-mixing condition. Many random processes satisfy the strongly mixing condition, for example, the stationary Markov process which is uniformly pure nondeterministic, the stationary Gaussian sequence with a continuous spectral density that is bounded away from 0, certain ARMA processes, and some aperiodic, Harris-recurrent Markov processes; see [5, 9] and the references therein.

In this paper we follow [7, 8] to consider α-mixing and ϕ-mixing processes, estimate the error bounds, and derive the learning rates of algorithm (2), where the output data satisfy the following unbounded condition.

Unbounded Hypothesis. There exist two constants $M > 0$ and $p \geq 2$ such that

$$\mathbb{E}|y|^p \leq M. \tag{5}$$

The error analysis for the algorithm (2) was usually presented under the standard assumption that $|y| \leq M$ almost surely with some constant $M > 0$. This standard assumption was abandoned in [10–14]. In [10] the authors introduced the condition

$$\int_Y \left(\exp\left\{-\frac{|y-f_{\mathcal{H}}|^2}{M}\right\} - \frac{|y-f_{\mathcal{H}}(x)|}{M} - 1\right) d\rho(y\mid x) \leq \frac{\Sigma^2}{2M^2} \tag{6}$$

for almost every $x \in X$ and some constants $M, \Sigma > 0$, where $f_{\mathcal{H}}$ is the orthogonal projection of f_ρ onto the closure of $\mathcal{H}_K$ in $L^2_{\rho_X}(X)$. In [11–13] the error analysis was conducted in another setting satisfying the following moment hypothesis; that is, there exist constants $\widetilde{M} > 0$ and $\widetilde{C} > 0$ such that $\int_Y |y|^l d\rho(y \mid x) \leq \widetilde{C} l! \widetilde{M}^l$ for all $l \in \mathbb{N}$, $x \in X$. Notice that with different constants the moment hypothesis and (6) are equivalent in the case $f_{\mathcal{H}} \in L^\infty(X)$ [13]. Obviously, our unbounded hypothesis is a natural generalization of the moment hypothesis. An example for which unbounded hypothesis (5) is satisfied but moment hypothesis failed has been given in [15]. It mainly studies the half supervised coefficient regularization with indefinite kernels and unbounded sampling, where the unbounded condition is $\int_Z y^2 d\rho \leq \widehat{M}^2$ for some constant $\widehat{M} > 0$.

Since $\mathcal{E}(f_\rho) = \min\mathcal{E}(f)$, where the generalization error $\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho$, the goodness of the approximation of f_ρ by $f_{\mathbf{z},\lambda}$ is usually measured by the excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z},\lambda} - f_\rho\|^2_{\rho_X}$. Denoting

$$\kappa := \sup_{x\in X} \sqrt{K(x,x)} < \infty, \tag{7}$$

the reproducing property in RKHS $\mathcal{H}_K$ yields that $\|f\|_\infty \leq \kappa\|f\|_K$ for any $f \in \mathcal{H}_K$. Thus, the distance between $f_{\mathbf{z},\lambda}$ and f_ρ in $\mathcal{H}_K$ can be applied to measure this approximation as well when $f_\rho \in \mathcal{H}_K$.

The noise-free limit of algorithm (2) takes the form

$$f_\lambda := \arg\min_{f\in\mathcal{H}_K} \left\{\|f - f_\rho\|^2_{\rho_X} + \lambda\|f\|^2_K\right\}, \tag{8}$$

thus the error analysis can be divided into two parts. The difference between $f_{\mathbf{z},\lambda}$ and f_λ is called the sample error, and the distance between f_λ and f_ρ is called the approximation error. We will bound the error in $L^2_{\rho_X}(X)$ and $\mathcal{H}_K$, respectively. Estimate of the sample error is more difficult because $f_{\mathbf{z},\lambda}$ changes with the sample $\mathbf{z}$ and cannot be considered as a fixed function. The approximation error does not depend on the samples, which has been studied in the literature [2, 3, 7, 16, 17].

We mainly devote the next two sections to estimating the sample error with more general sampling processes. Our main results can be stated as follows.

Theorem 3. *Suppose that the unbounded hypothesis holds, $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ for some $r > 0$, and the ϕ-mixing coefficients satisfy a polynomial decay, that is, $\phi_i \leq ai^{-t}$ for some $a > 0$ and $t > 0$. Then, for any $0 < \eta < 1$, one has with confidence $1 - \eta$,*

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\rho_X} = O\left(m^{-\theta\min\{t/2,1\}}(\log m)^{3/4}\right), \tag{9}$$

where θ is given by

$$\theta = \begin{cases} \dfrac{3r}{4(r+1)} & \text{if } 0 < r < \dfrac{1}{2}, \\ \dfrac{r}{2r+1} & \text{if } \dfrac{1}{2} \leq r < 1, \\ \dfrac{1}{3} & \text{if } r \geq 1. \end{cases} \tag{10}$$

Moreover, when $r > 1/2$, one has with confidence $1 - \eta$,

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_K = O\left(m^{-\theta'\min\{t/2,1\}}(\log m)^{1/2}\right), \tag{11}$$

where θ' is given by

$$\theta' = \begin{cases} \dfrac{2r-1}{2(2r+1)} & \text{if } \dfrac{1}{2} < r < \dfrac{3}{2}, \\ \dfrac{1}{4} & \text{if } r \geq \dfrac{3}{2}. \end{cases} \tag{12}$$

Theorem 3 proves the asymptotic convergence of algorithm (2) with the samples satisfying a uniformly mixing condition. Our second main result considers this algorithm with α-mixing process.

Theorem 4. *Suppose that the unbounded hypothesis with $p > 2$ holds, $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ for some $r > 0$, and the α-mixing coefficients satisfy a polynomial decay, that is, $\alpha_l \le bl^{-t}$ for some $b > 0$ and $t > 0$. Then, for any $0 < \eta < 1$, one has with confidence $1 - \eta$,*

$$\left\| f_{\mathbf{z},\gamma} - f_\rho \right\|_{\rho_X} = O\left(m^{-\vartheta \min\{(p-2)t/p,1\}} (\log m)^{1/2} \right), \tag{13}$$

where ϑ is given by

$$\vartheta = \begin{cases} \dfrac{pr}{2(2r+p-1)} & \text{if } 0 < r < \dfrac{1}{2},\ 0 < t < \dfrac{p}{p-2}; \\ \dfrac{3pr}{2(4r+3p-2)} & \text{if } 0 < r < \dfrac{1}{2},\ t \ge \dfrac{p}{p-2}; \\ \dfrac{r}{2r+1} & \text{if } \dfrac{1}{2} \le r < 1; \\ \dfrac{1}{3} & \text{if } r \ge 1. \end{cases} \tag{14}$$

Moreover, when $r > 1/2$, with confidence $1 - \eta$,

$$\left\| f_{\mathbf{z},\gamma} - f_\rho \right\|_K = O\left(m^{-\vartheta' \min\{(p-2)t/p,1\}} (\log m)^{1/2} \right), \tag{15}$$

where ϑ' is given by

$$\vartheta' = \begin{cases} \dfrac{2r-1}{4r+2} & \text{if } \dfrac{1}{2} < r < \dfrac{3}{2}, \\ \dfrac{1}{4} & \text{if } r \ge \dfrac{3}{2}. \end{cases} \tag{16}$$

The proof of these two theorems will be given in Sections 2, 3, and 4, and notice that the log term can be dropped when $t \ne 2$. Our error analysis reveals some interesting phenomena for learning with unbounded and dependent sampling.

(i) Smoother target function f_ρ (i.e., r becomes larger) implies better learning rates. Stronger dependence between samples (i.e., t becomes smaller) implies that they contain less information and hence lead to worse rates.

(ii) The learning rates are improved as the dependence between samples becomes weaker and r becomes larger but they are no longer improved after some constant t, r. This phenomenon is called saturation effect, which was discussed in [18–20]. In our setting, saturation effects include saturation for smoothness of function f_ρ mainly relative to the approximation error and saturation for dependence between samples. An interesting phenomenon revealed here is that when α-mixing coefficients satisfy $\alpha_l \le O(l^{-t})$, $l \in \mathbb{N}$ for some $t > 0$, the saturation for dependence between samples is $t = p/(p-2)$ for $p > 2$, which is dependent on the unbounded condition parameter p.

(iii) For ϕ-mixing process, the learning rates have nothing to do with unbound condition parameter p since $\mathbb{E}(y - f_\lambda(x))^2$ is bounded by $\mathbb{E}y^2 < \infty$. But for α-mixing process, to derive the learning rate, we have to estimate $\mathbb{E}|y - f_\lambda(x)|^p$ with $p > 2$.

(iv) Under α-mixing condition, when $t > p/(p-2)$ and $r \ge 1/2$, the influence of the unbounded condition becomes weak. Recall that the learning rate derived in [8] is $O(m^{-r/(1+2r)})$ for $1/2 \le r \le 1, t \ge 1$. It implies that when t is large enough, our learning rate for unbounded samples is as sharp as that for the uniform bounded sampling.

2. Sampling Satisfying ϕ-Mixing Condition

In this section, we would apply the integral operator technique in [7] to handle the sample error with ϕ-mixing condition. However, different from the uniform bounded case the learning performance of the unbounded sampling is not measured directly. Instead, the expectations are estimated first and then the bound for the sample error can be obviously deduced by Markov inequality:

To this end, define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to l^2(\mathbf{x})$ as $S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^m$, where $\mathbf{x}$ is the set of input data $\{x_1, \ldots, x_m\}$. Then its adjoint is $S_{\mathbf{x}}^{\mathbf{T}} c = \sum_{i=1}^m c_i K_{x_i}$ for $c \in l^2(\mathbf{x})$. The analytic expression of optimization solution $f_{\mathbf{z},\lambda}, f_\lambda$ was given in [3],

$$\begin{aligned} f_{\mathbf{z},\lambda} &= \left(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I \right)^{-1} \frac{1}{m} S_{\mathbf{x}}^T y, \\ f_\lambda &= (L_K + \lambda I)^{-1} L_K f_\rho, \end{aligned} \tag{17}$$

where $L_K : L^2_{\rho_X}(X) \to L^2_{\rho_X}(X)$ is the integral operator defined as

$$L_K f(x) = \int_X K(x,t) f(t) \, d\rho_X(t), \quad \text{for any } x \in X. \tag{18}$$

For a random variable ξ with values in a Hilbert space $\mathcal{H}$ and $0 \le u \le +\infty$, denote the uth moment as $\|\xi\|_u = (\mathbb{E}\|\xi\|_{\mathcal{H}}^u)^{1/u}$ if $1 \le u < \infty$ and $\|\xi\|_\infty = \sup \|\xi\|_{\mathcal{H}}$. Lemma 5 is due to Billingsley [21].

Lemma 5. *Let ξ and η be random variables with values in a separable Hilbert space $\mathcal{H}$ measurable σ-field $\mathcal{J}$ and $\mathcal{D}$ and having finite pth and qth moments, respectively, where $p, q \ge 1$ with $p^{-1} + q^{-1} = 1$. Then*

$$\left| \mathbb{E}(\xi, \eta) - (\mathbb{E}\xi, \mathbb{E}\eta) \right| \le 2\phi^{1/p}(\mathcal{J}, \mathcal{D}) \|\xi\|_p \|\eta\|_q. \tag{19}$$

Lemma 6. *For an ϕ-mixing sequence $\{x_i\}$, one has*

$$\mathbb{E}\left\| L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} \right\|^2 \le \frac{\kappa^4}{m} \left(1 + 4 \sum_{i=1}^{m-1} \phi_i^{1/2} \right). \tag{20}$$

Proof. With the definition of the sample operator, we have

$$L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} = L_K - \frac{1}{m} \sum_{i=1}^m K_{x_i} \otimes K_{x_i}. \tag{21}$$

Letting $\eta(x) = K_x \otimes K_x$, then $\eta(x)$ is an $\mathrm{HS}(\mathscr{H}_K)$-valued random variable defined on X. Note that $\mathbb{E}\eta(x) = L_K \in \mathrm{HS}(\mathscr{H}_K)$, and $\|L_K\|_{\mathrm{HS}} \le \kappa^2$, $\|\eta(x)\|_{\mathrm{HS}} \le \kappa^2$. We have

$$\begin{aligned}
&\mathbb{E}\left\|L_K - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2 \\
&\quad\le \mathbb{E}\left\|\mathbb{E}\eta - \frac{1}{m}\sum_{i=1}^{m}\eta(x_i)\right\|_{\mathrm{HS}}^2 \\
&\quad= \frac{1}{m}\|\eta\|_2^2 + \frac{1}{m^2}\sum_{i\neq j}\mathbb{E}\left\langle \eta(x_i), \eta(x_j)\right\rangle_{\mathrm{HS}} - \|L_K\|_{\mathrm{HS}}^2.
\end{aligned} \tag{22}$$

By Lemma 5 with $p = q = 2$, for $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left\langle \eta(x_i), \eta(x_j)\right\rangle_{\mathrm{HS}} &\le \left\langle \mathbb{E}\eta(x_i), \mathbb{E}\eta(x_j)\right\rangle_{\mathrm{HS}} + 2\phi_{|i-j|}^{1/2}\|\eta\|_2^2 \\
&\le \|L_K\|_{\mathrm{HS}}^2 + 2\kappa^4\phi_{|i-j|}^{1/2}.
\end{aligned} \tag{23}$$

Thus the desired estimate can be obtained by plugging (23) into (22). □

Proposition 7. *Suppose that the unbounded hypothesis holds with some $p \ge 2$ and that the sample sequence $\{(x_i, y_i)\}_{i=1}^m$ satisfies an ϕ-mixing condition and $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ with $r > 0$. Then one has*

$$\begin{aligned}
&\mathbb{E}\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\rho_X} \\
&\quad\le C\left(\lambda^{-1/2}m^{-1/2} + \lambda^{-1}m^{-3/4}\left(1 + 4\sum_{i=1}^{m-1}\phi_i^{1/2}\right)^{1/4}\right) \\
&\qquad\times\sqrt{1 + 4\sum_{l=1}^{m-1}\phi_i^{1/2}},
\end{aligned} \tag{24}$$

where C is a constant only dependent on κ, M.

Proof. By [7, Theorem 3.1], we have

$$\begin{aligned}
&\mathbb{E}\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\rho_X} \\
&\quad\le \left(\lambda^{-1/2} + \lambda^{-1}\left(\mathbb{E}\left\|L_K - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2\right)^{1/4}\right) \\
&\qquad\times\sqrt{\mathbb{E}\left\|\frac{1}{m}\sum_{l=1}^{m-1}\xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2},
\end{aligned} \tag{25}$$

where $\xi(z) = (y - f_\lambda(x))K_x$ is a random variable with values in $\mathscr{H}_K$, and $\mathbb{E}\xi = L_K(f_\rho - f_\lambda)$. A similar computation together with the result of Lemma 6 leads to

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2 \le \frac{1}{m}\left(1 + 4\sum_{i=1}^{m-1}\phi_i^{1/2}\right)\|\xi\|_2^2. \tag{26}$$

It suffices to estimate $\|\xi\|_2$. By Hölder inequality, there is

$$\begin{aligned}
&\mathbb{E}y^2 \le \left(\mathbb{E}|y|^p\right)^{2/p} \le M^{2/p}, \\
&\left\|f_\rho\right\|_{\rho_X}^2 = \int_X f_\rho^2(x)\,d\rho_X \\
&\qquad= \int_X\left(\int_Y y\,d\rho(y\mid x)\right)^2 d\rho_X \le \int_Z y^2 d\rho \le M^{2/p}.
\end{aligned} \tag{27}$$

Thus $\mathbb{E}(y - f_\rho(x))^2 = \mathbb{E}y^2 - \|f_\rho\|_{\rho_X}^2 \le M^{2/p}$ and

$$\mathbb{E}\left(f_\rho(x) - f_\lambda(x)\right)^2 = \left\|\lambda\left(\lambda I + L_K\right)^{-1} f_\rho\right\|_{\rho_X}^2 \le \left\|f_\rho\right\|_{\rho_X}^2 \le M^{2/p}, \tag{28}$$

which implies

$$\begin{aligned}
\|\xi\|_2^2 &= \mathbb{E}\left(\left(y - f_\lambda(x)\right)^2 K(x,x)\right) \le \kappa^2\mathbb{E}\left(y - f_\lambda(x)\right)^2 \\
&= \kappa^2\left(\mathbb{E}\left(y - f_\rho(x)\right)^2 + \mathbb{E}\left(f_\rho(x) - f_\lambda(x)\right)^2\right) \le 2\kappa^2 M^{2/p}.
\end{aligned} \tag{29}$$

Plugging (29) into (26), there holds

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2 \\
&\quad\le 2M^{2/p}k^2 m^{-1}\left(1 + 4\sum_{i=1}^{m-1}\phi_i^{1/2}\right).
\end{aligned} \tag{30}$$

Combining (25), (22), and (30) and taking the constant $C = \sqrt{2}k(k+1)M^{1/p}$, we complete the proof. □

The following proposition provides the bound of the difference between $f_{\mathbf{z},\lambda}$ and f_λ in $\mathscr{H}_K$ with ϕ-mixing process.

Proposition 8. *Under the assumption of Proposition 7, there holds*

$$\mathbb{E}\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \le \sqrt{2}M^{1/p}\kappa\lambda^{-1}m^{-1/2}\sqrt{1 + 4\sum_{i=1}^{m-1}\phi_i^{1/2}}. \tag{31}$$

Proof. The representations of $f_{\mathbf{z},\lambda}$ and f_λ imply that

$$\begin{aligned}
&\mathbb{E}\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \\
&\quad= \mathbb{E}\left\|\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left(\frac{1}{m}\sum_{l=1}^{m-1}\xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right)\right\|_K \\
&\quad\le \lambda^{-1}\sqrt{\mathbb{E}\left\|\frac{1}{m}\sum_{l=1}^{m-1}\xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2}.
\end{aligned} \tag{32}$$

Then the desired bound follows from (30) and (32). □

3. Samples Satisfying α-Mixing Condition

Now we turn to bound the sample error when the sampling process satisfies strongly mixing condition, and unbounded hypothesis holds. In Section 2, the key point is to estimate $\|\xi\|_2$ with the lack of uniform boundedness. For the sampling satisfying α-mixing condition, we have to deal with $\|\xi\|_p$ for some $p > 2$.

Proposition 9. *Suppose that the unbounded hypothesis holds with some $p > 2$ and that the sample sequence $\{(x_i, y_i)\}_{i=1}^m$ satisfies an α-mixing condition and $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ with $r > 0$. Then one gets*

$$\mathbb{E}\left\|f_{\mathbf{z},\lambda} - f_\lambda\right\|_{\rho_X} \le \widetilde{C}\lambda^{\min\{(p-2)(2r-1)/2p,0\}} \sqrt{1 + \sum_{l=1}^{m-1} \alpha_l^{(p-2)/p}} \times \left(\lambda^{-1/2} m^{-1/2} + \lambda^{-1} m^{-3/4} \left(1 + \sum_{l=1}^{m-1} \alpha_l\right)^{1/4}\right), \tag{33}$$

where $\widetilde{C}$ is a constant only depending on κ, M and $\|L_K^{-\min\{r,1/2\}} f_\rho\|_{\rho_X}$.

Proof. For the strongly mixing process, by [8, Lemma 5.1],

$$\mathbb{E}\left\|L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2 \le \frac{k^4}{m}\left(1 + 30 \sum_{l=1}^{m-1} \alpha_l\right). \tag{34}$$

Taking $\delta = p - 2$ in [8, Lemma 4.2], we have

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2 \le \frac{1}{m}\|\xi\|_2^2 + \frac{30}{m}\sum_{l=1}^{m-1} \alpha_l^{(p-2)/p} \|\xi\|_p^2. \tag{35}$$

The estimation of $\|\xi\|_2$ has been obtained in Section 2, and now we mainly devote to estimating $\|\xi\|_p$. To get this estimation, the bound of f_λ is needed which can be stated as follows ([3, Lemma 3] or [8, Lemma 4.3]):

$$\left|f_\lambda(x)\right| \le \kappa \left\|f_\lambda\right\|_K \le C_1 \kappa \lambda^{\min\{(2r-1)/2,0\}}, \tag{36}$$

where $C_1 = \|L_K^{-\min\{r,1/2\}} f_\rho\|_{\rho_X}$. Observe that $\|f_\lambda\|_{\rho_X}^2 \le \|f_\rho\|_{\rho_X}^2 \le \mathbb{E}y^2 \le M^{2/p}$. Hence,

$$\left(\mathbb{E}\left|f_\lambda(x)\right|^p\right)^{2/p} \le \left(\left\|f_\lambda\right\|_{\rho_X}^2 C_1^{p-2} \kappa^{p-2} \lambda^{(p-2)\min\{(2r-1)/2,0\}}\right)^{2/p} \le M^{4/p^2}\left(C_1^2\kappa^2 + 1\right)\lambda^{\min\{(p-2)(2r-1)/p,0\}}. \tag{37}$$

Now we can deduce that

$$\begin{aligned}\|\xi\|_p^2 &= \left(\mathbb{E}\left(\left(y - f_\lambda(x)\right)^2 K(x,x)\right)^{p/2}\right)^{2/p} \\ &\le \kappa^2\left(\mathbb{E}\left|y - f_\lambda(x)\right|^p\right)^{2/p} \\ &\le 4\kappa^2\left(\mathbb{E}\max\left\{|y|^p, \left|f_\lambda(x)\right|^p\right\}\right)^{2/p} \\ &\le 2\kappa^2\left(\left(\mathbb{E}|y|^p\right)^{2/p} + \left(\mathbb{E}\left|f_\lambda(x)\right|^p\right)^{2/p}\right) \\ &\le 2\kappa^2\left(M^{2/p} + M^{4/p^2}\left(C_1^2\kappa^2 + 1\right)\right)\lambda^{\min\{(p-2)(2r-1)/p,0\}}.\end{aligned} \tag{38}$$

Plugging this estimate into (35) yields

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - L_K\left(f_\rho - f_\lambda\right)\right\|_K^2 \le C_2 m^{-1} \lambda^{\min\{(p-2)(2r-1)/p,0\}}\left(1 + \sum_{l=1}^{m-1} \alpha_l^{(p-2)/p}\right), \tag{39}$$

where C_2 is a constant only depending on κ, M and $\|L_K^{-\min\{r,1/2\}} f_\rho\|_{\rho_X}$. Then combining (34) and (39) with (25), we complete the proof. □

For α-mixing process we have the following proposition to get the bound of sample error in $\mathscr{H}_K$, and the proof can be directly obtained by the inequality (32).

Proposition 10. *Under assumption of Proposition 9, one has*

$$\mathbb{E}\left\|f_{\mathbf{z},\lambda} - f_\lambda\right\|_K \le C_3 m^{-1/2} \lambda^{-1} \sqrt{1 + \sum_{l=1}^{m-1} \alpha_l^{(p-2)/p}}, \tag{40}$$

where $C_3 = \sqrt{C_2}$.

4. Error Bounds and Learning Rates

In this section we derive the learning rates, that is, the convergence rates of $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\rho_X}$ and $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ as $m \to \infty$ by choosing the regularization parameter λ according to m. The following approximation error bound is needed to get the convergence rates.

Proposition 11. *Supposing that $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ for some $r > 0$, there holds*

$$\left\|f_\lambda - f_\rho\right\|_{\rho_X} \le \lambda^{\min\{r,1\}} \left\|L_K^{-\min\{r,1\}} f_\rho\right\|_{\rho_X}. \tag{41}$$

Moreover, when $r \ge 1/2$, that is, $f_\rho \in \mathscr{H}_K$, there holds

$$\left\|f_\lambda - f_\rho\right\|_K \le \lambda^{\min\{r-(1/2),1\}} \left\|L_K^{-\min\{r,3/2\}} f_\rho\right\|_{\rho_X}. \tag{42}$$

The first conclusion in Proposition 11 has been proved in [20], and the second one can be proved in the same way. To derive the learning rates, we need to balance the approximation error and sample error. For this purpose, the following simple facts are necessary:

$$\sum_{l=1}^{m-1} l^{-s} \leq \begin{cases} \dfrac{1}{1-s} m^{1-s} & \text{if } 0 < s < 1, \\ \log m & \text{if } s = 1, \\ \dfrac{1}{s-1} & \text{if } s > 1. \end{cases} \tag{43}$$

Proof of Theorem 3. The estimate of learning rates in $L^2_{\rho_X}(X)$ norm is divided into two cases.

Case 1. For $0 < t < 2$, by (43) and $\phi_i \leq a i^{-t}$, there is

$$1 + 4 \sum_{i=1}^{m-1} \phi_i^{1/2} \leq 1 + 4\sqrt{a} \sum_{i=1}^{m-1} i^{-t/2} \leq \left(1 + \frac{8\sqrt{a}}{2-t}\right) m^{1-(t/2)}. \tag{44}$$

Thus Proposition 7 yields that

$$\mathbb{E}\left\|f_{\mathbf{z},\lambda} - f_\lambda\right\|_{\rho_X} \leq 2C \left(1 + \frac{16\sqrt{a}}{2-t}\right) \left(\lambda^{-1/2} m^{-t/4} + \lambda^{-1} m^{-3t/8}\right). \tag{45}$$

By Proposition 11 and Markov inequality, with confidence $1 - \eta$, there holds

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\rho_X} \leq O\left(\lambda^{\min\{r,1\}} + \eta^{-1}\left(\lambda^{-1/2} m^{-t/4} + \lambda^{-1} m^{-3t/8}\right)\right). \tag{46}$$

For $0 < r < 1/2$, by taking $\lambda = m^{-3t/(8(r+1))}$, we can deduce the learning rate as $O(m^{-3tr/(8(r+1))})$. When $1/2 \leq r < 1$, taking $\lambda = m^{-t/(2(2r+1))}$, the learning rate $O(m^{-rt/(2(2r+1))})$ can be derived. When $r \geq 1$, the desired convergence rate is obtained by taking $\lambda = m^{-t/6}$.

Case 2. $t \geq 2$. With confidence $1 - \eta$, there holds

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\rho_X} = O\left(\lambda^{\min\{r,1\}} + \eta^{-1}\left(\lambda^{-1/2} m^{-1/2} + \lambda^{-1} m^{-3/4}\right)(\log m)^{3/4}\right). \tag{47}$$

For $0 < r < 1/2$, taking $\lambda = m^{-3/(4(r+1))}$, the learning rate $O(m^{-3r/(4(r+1))}(\log m)^{3/4})$ can be derived, and for $1/2 \leq r < 1$, by taking $\lambda = m^{-1/(2r+1)}$, we can deduce the learning rate $O(m^{-r/(2r+1)}(\log m)^{3/4})$. When $r \geq 1$, the desired convergence rate is obtained by taking $\lambda = m^{-1/3}$.

Next for bounding the generalization error in $\mathscr{H}_K$, Proposition 8 in connection with Proposition 11 tells us that with confidence $1 - \eta$,

$$\left\|f_{\mathbf{z}} - f_\rho\right\|_K \leq \left(\lambda^{\min\{r-(1/2),1\}} + \eta^{-1}\lambda^{-1} m^{-1/2} \sqrt{1 + 4\sum_{i=1}^{m-1} \phi_i^{1/2}}\right). \tag{48}$$

The rest of the proof is analogous to the estimate of $\|f_{\mathbf{z},\lambda} - f_\rho\|_{\rho_X}$ mentioned previously. □

Proof of Theorem 4. For $0 < t < 1$, by (43) and $\alpha_l \leq b l^{-t}$, there is

$$\begin{aligned} 1 + \sum_{l=1}^{m-1} \alpha_l^{(p-2)/p} &\leq 1 + b^{(p-2)/p} \sum_{l=1}^{m-1} l^{-((p-2)/p)t} \\ &\leq \left(1 + \frac{p b^{(p-2)/p}}{p - (p-2)t}\right) m^{1-((p-2)/p)t}, \\ 1 + \sum_{l=1}^{m-1} \alpha_l &\leq 1 + b \sum_{l=1}^{m-1} l^{-t} \leq \left(1 + \frac{b}{1-t}\right) m^{1-t}. \end{aligned} \tag{49}$$

By Propositions 9 and 11 and Markov inequality, with confidence $1 - \eta$, there holds

$$\left\|f_{\mathbf{z},\lambda} - f_\rho\right\|_{\rho_X} = O\left(\lambda^{\min\{r,1\}} + \eta^{-1}\lambda^{\min\{(p-2)(2r-1)/2p,0\}-1/2} \times \left(m^{-(p-2)t/2p} + \lambda^{-1/2} m^{-(p-2)t/2p - t/4}\right)\right). \tag{50}$$

For $0 < r < 1/2$, by taking $\lambda = m^{-(p-2)t/2(2r+p-1)}$, we can deduce the learning rate as $O(m^{-(p-2)tr/2(2r+p-1)})$. When $1/2 \leq r < 1$, taking $\lambda = m^{-(p-2)t/p(2r+1)}$, the learning rate $O(m^{-(p-2)rt/p(2r+1)})$ can be derived. When $r \geq 1$, the desired convergence rate is obtained by taking $\lambda = m^{-(p-2)t/3p}$.

The rest of the analysis is similar; we omit it here. □

Acknowledgment

This paper is supported by the National Nature Science Foundation of China (no. 11071276).

References

[1] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[2] S. Smale and D.-X. Zhou, "Shannon sampling. II. Connections to learning theory," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 285–302, 2005.

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

[4] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.

[5] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2133–2145, 1996.

[6] S. Smale and D.-X. Zhou, "Online learning with Markov sampling," *Analysis and Applications*, vol. 7, no. 1, pp. 87–113, 2009.

[7] H. Sun and Q. Wu, "A note on application of integral operator in learning theory," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 416–421, 2009.

[8] H. Sun and Q. Wu, "Regularized least square regression with dependent samples," *Advances in Computational Mathematics*, vol. 32, no. 2, pp. 175–189, 2010.

[9] K. B. Athreya and S. G. Pantula, "Mixing properties of Harris chains and autoregressive processes," *Journal of Applied Probability*, vol. 23, no. 4, pp. 880–892, 1986.

[10] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[11] Z.-C. Guo and D.-X. Zhou, "Concentration estimates for learning with unbounded sampling," *Advances in Computational Mathematics*, vol. 38, no. 1, pp. 207–223, 2013.

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

[13] C. Wang and D.-X. Zhou, "Optimal learning rates for least squares regularized regression with unbounded sampling," *Journal of Complexity*, vol. 27, no. 1, pp. 55–67, 2011.

[14] C. Wang and Z. C. Guo, "ERM learning with unbounded sampling," *Acta Mathematica Sinica*, vol. 28, no. 1, pp. 97–104, 2012.

[15] X. R. Chu and H. W. Sun, "Half supervised coefficient regularization for regression learning with unbounded sampling," *International Journal of Computer Mathematics*, 2013.

[16] S. Smale and D.-X. Zhou, "Shannon sampling and function reconstruction from point values," *The American Mathematical Society*, vol. 41, no. 3, pp. 279–305, 2004.

[17] H. Sun and Q. Wu, "Application of integral operator for regularized least-square regression," *Mathematical and Computer Modelling*, vol. 49, no. 1-2, pp. 276–285, 2009.

[18] F. Bauer, S. Pereverzev, and L. Rosasco, "On regularization algorithms in learning theory," *Journal of Complexity*, vol. 23, no. 1, pp. 52–72, 2007.

[19] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri, "Spectral algorithms for supervised learning," *Neural Computation*, vol. 20, no. 7, pp. 1873–1897, 2008.

[20] H. Sun and Q. Wu, "Least square regression with indefinite kernels and coefficient regularization," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 96–109, 2011.

[21] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, NY, USA, 1968.