Adv Comput Math
DOI 10.1007/s10444-008-9099-y

Regularized least square regression with dependent samples

Hongwei Sun · Qiang Wu

Received: 14 September 2007 / Accepted: 11 August 2008

Abstract In this paper we study the learning performance of regularized least square regression with α-mixing and ϕ-mixing inputs. The capacity independent error bounds and learning rates are derived by means of an integral operator technique. Even for independent samples our learning rates improve those in the literature. The results are sharp in the sense that when the mixing conditions are strong enough the rates are shown to be close to or the same as those for learning with independent samples. They also reveal interesting phenomena of learning with dependent samples: (i) dependent samples contain less information and lead to worse error bounds than independent samples; (ii) the influence of the dependence between samples to the learning process decreases as the smoothness of the target function increases.

Communicated by Yuesheng Xu.

H. Sun
School of Science, Jinan University, Jinan 250022,
People's Republic of China
e-mail: ss_sunhw@ujn.edu.cn

H. Sun
School of Mathematical Science, Beijing Normal University,
Beijing 100875, People's Republic of China

Q. Wu (✉)
Department of Statistical Science, Institute for Genome Sciences & Policy,
Duke University, Durham, NC 27708, USA
e-mail: qiang@stat.duke.edu

Keywords Regularized least square regression · Integral operator · Strong mixing condition · Capacity independent error bounds

Mathematics Subject Classifications (2000) Primary 68T05 · 68P05 · Secondary 42B10

1 Introduction

Least square regression problem has a long history in machine learning and statistics. In the machine learning terminology, this problem can be stated as follows: Let X be a compact metric space (usually a subset of $\mathbb{R}^n$) and $Y \subset \mathbb{R}$. Assume $x \in X$ and $y \in Y$ are random variables with their dependence described by a joint probability measure ρ on $X \times Y$. Here x is called the input variable and y the output (or response) variable. The objective is the regression function

$$f_\rho(x) = \mathbb{E}(y|x) = \int_Y y d\rho(y|x)$$

which describes how the output variable depends on the input variable. In practice, we do not know the probability measure ρ and f_ρ is not directly computable. Instead, we have in hand a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn according to ρ and f_ρ should be learned from this set of samples.

In this paper, we study the regularized least square regression which, given a Mercer kernel K, learns an empirical regressor by

$$f_{\mathbf{z},\gamma} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma \| f \|_K^2 \right\}, \tag{1.1}$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) associated to the kernel K (for definitions and properties see [1]) and $\gamma > 0$ is a regularization parameter. The resurgence of the kernel methods and the study of learning theory from an approximation point of view [6] have motivated a lot of research on this RKHS based regularization scheme [9, 13, 14, 17] (Sun and Wu, unpublished manuscript). These works have focused on independent samples. However, independence is a restrictive assumption and may be violated in many real data analysis. In recent years the learning with dependent samples starts to attract attentions [10, 11, 15, 16, 18] (Smale and Zhou, unpublished manuscript). The aim of this paper is to study the learning performance of (1.1) for dependent samples.

According to *no free lunch* principle, it is necessary to make some assumptions that naturally measure the dependence between the samples. In the literature, several different notions have been considered in the regression setting. Examples include the mixing sequences [11, 16] and samples sampling from Markov chain distributions generated by dynamic operators (Smale and Zhou, unpublished manuscript). In this paper we follow [11, 16] and adopt the notion of mixing condition due to its ubiquitousness in stationary stochastic processes. Our main results will be the capacity independent error bounds and learning rates for two classes of mixing sequences. The results coincide with the intuition that the dependent data contain less information than independent data by worse error bounds and learning rates. Moreover, they show that the influence of the dependence will be decreased if the target regression function is smoother. It can even disappear if the function is smooth enough, say, when $f_\rho \in \mathcal{H}_K$. The details are given in Section 2.

The results will be proved in Sections 4–6. We use the integral operator technique introduced in [14]. Recall the integral operator can be used to estimate both the estimation error and the approximation error. We will show how the integral operator techniques can be extended from independent samples to dependent samples. Moreover, our analysis is refined in the sense that it yields a sharper error bound for the independent samples and improves the learning rates in [14]. As for the dependent samples, our results are also sharp in the sense that (i) when the mixing conditions become strong and the samples tend to be "nearly" independent the learning rates can be arbitrarily close to or even the same as those for the independent samples. (ii) They may even outperform some existing capacity dependent results (e.g. [11, 18]) while this is impossible for independent samples.

2 Main results

In this section, we state our main results and discuss their relations to the existing works. For this purpose, we first introduce some notations and concepts.

Let ρ_X denote the marginal distribution of ρ on X. Throughout this paper we assume $f_\rho \in L^2_{\rho_X}(X)$ and $|y| \leq M$ almost surely for some constant $M > 0$. A Mercer kernel K is a continuous, symmetric, positive definite function on $X \times X$. Obviously there holds

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

Recall the reproducing property of the RKHS is given by $f(x) = \langle f, K_x \rangle$ implying that $\|f\|_\infty \leq \kappa \|f\|_K$.

The main purpose of this paper is to study the learning performance of the algorithm (1.1) with dependent samples. We assume the sample sequence $z_i = (x_i, y_i),\ \ i \geq 1$ comes from a strictly stationary process and the dependence will be measured by the strongly mixing condition and uniformly mixing condition.

For two σ-fields $\mathcal{J}$ and $\mathcal{D}$, define the α-coefficient as

$$\alpha(\mathcal{J}, \mathcal{D}) = \sup_{A \in \mathcal{J}, B \in \mathcal{D}} |P(A \cap B) - P(A)P(B)|$$

and ϕ-coefficient

$$\phi(\mathcal{J}, \mathcal{D}) = \sup_{A \in \mathcal{J}, B \in \mathcal{D}} |P(A|B) - P(A)|.$$

Given a sequence of samples $\{z_i\}_{i=1}^{\infty}$, denote by $\mathcal{M}_a^b$ the σ-field generated by random variables $z_a, z_{a+1}, \cdots, z_b$. The strongly mixing condition and uniformly mixing condition are defined as follows.

Definition 2.1 A set of random sequence $z_i,\ i \geq 1$, is said to satisfy a strongly mixing condition (or α-mixing condition) if

$$\alpha_i = \sup_{k\geq 1} \alpha\left(\mathcal{M}_1^k, \mathcal{M}_{k+i}^{\infty}\right) \longrightarrow 0,\ \text{as}\ i \to \infty.$$

It satisfies a uniformly mixing condition (or ϕ-mixing condition) if

$$\phi_i = \sup_{k\geq 1} \phi\left(\mathcal{M}_1^k, \mathcal{M}_{k+i}^{\infty}\right) \longrightarrow 0,\ \text{as}\ i \to \infty.$$

Note that strongly 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 non-deterministic, 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 [2, 11] and the references therein. For the latter two examples, the strongly mixing coefficients even decay exponentially fast, i.e., satisfying a so called exponentially strongly mixing condition. For examples of uniformly mixing sequences see [11]. As a special and trivial example, a sequence of identically and independently distributed samples satisfies both conditions with $\alpha_i = \phi_i = 0$.

In regression learning, we intend to study the approximation ability of $f_{\mathbf{z},\gamma}$ to the true regression function f_ρ. The main results of this paper will be the error bounds and learning rates when the samples (x_i, y_i) are from a strictly stationary process satisfying a strongly mixing condition or a uniformly mixing condition.

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

$$L_K(f)(x) = \int_X K(x,t) f(t) d\rho_X(t). \tag{2.1}$$

Denote the norm in $L^2_{\rho_X}(X)$ as $\|f\|_{\rho_X}$. Our main results can be stated as follows.

Theorem 2.2 *If the sample sequence (x_i, y_i), $i = 1, \ldots, m$ satisfies an α-mixing condition and $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ with $0 < r \leq 1$, then for any $0 < \delta \leq \infty$ and $0 < \eta < 1$, with confidence $1 - \eta$, $\|f_{\mathbf{z},\gamma} - f_\rho\|_{\rho_X}$ is bounded by*

$$\gamma^r \|L_K^{-r} f_\rho\|_{\rho_X} + \frac{C_1}{\eta}\left(\frac{1}{\sqrt{m\gamma}} + \frac{1}{m\gamma^{\frac{3}{2}}}\sqrt{1 + \sum_{i=1}^{m-1}\alpha_i}\right)\left(1 + \gamma^{\frac{(2r-1)\delta}{2(2+\delta)}}\right)\sqrt{1 + \sum_{i=1}^{m-1}\alpha_i^{\frac{\delta}{2+\delta}}}$$

where C_1 is a constant independent of m, γ, η and δ.

Theorem 2.3 *If the sample sequence* (x_i, y_i), $i = 1, \dots, m$ *satisfies a* ϕ*-mixing condition and* $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ *with* $0 < r \le 1$, *then for any* $0 < \eta < 1$, *with confidence* $1 - \eta$

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} \le \gamma^r \| L_K^{-r} f_\rho \|_{\rho_X} + \frac{C_2}{\eta} \left(\frac{1}{\sqrt{m\gamma}} + \frac{1}{m\gamma^{\frac{3}{2}}} \sqrt{1 + \sum_{i=1}^{m-1} \phi_i^{\frac{1}{2}}} \right) \sqrt{1 + \sum_{i=1}^{m-1} \phi_i^{\frac{1}{2}}}$$

where C_2 *is a constant independent of* m, γ *and* η.

In case of independent samples, we have $\alpha_i = \phi_i = 0$. The following bound is an immediate corollary of either result above.

Corollary 2.4 *With independent samples and under the assumption* $L_K^{-r} f_\rho \in L^2_{\rho_X}(X)$ *with* $0 < r \le 1$, *with probability* $1 - \eta$ *there holds*

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} \le \gamma^r \| L_K^{-r} f_\rho \|_{\rho_X} + \frac{C_3}{\eta} \left(\frac{1}{\sqrt{m\gamma}} + \frac{1}{m\gamma^{\frac{3}{2}}} \right).$$

It can be easily checked that this bound leads to the error rate of order $O(m^{-2r/(3+2r)})$ if $0 < r < 1/2$ and $O(m^{-r/(1+2r)})$ if $1/2 \le r \le 1$. In case of $0 < r \le 1/2$, it is better than that given by [13, Corollary 5].[1]

Theorems 2.2 and 2.3 will be proven in next several sections. Before going into the technical proofs, we deduce some learning rates of algorithm (1.1) and discuss their relations to some existing works.

Corollary 2.5 *Under the assumptions of Theorem* 2.2, *if the* α*-mixing coefficients satisfy a polynomial decay, i.e.,* $\alpha_i \le a i^{-t}$ *for some* $a > 0$ *and* $t > 0$, *then by taking* $\gamma = m^{-\theta}$ *we have*

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} = O\left(\left(\frac{1}{m} \right)^{\theta r} \log m \right). \tag{2.2}$$

where θ *is given by*

$$\theta = \begin{cases} \frac{2t}{(2r+3)t+1-2r} & \text{if } 0 < r < 1/2 \text{ and } t \ge 1; \\ \frac{t}{2} & \text{if } 0 < r < 1/2 \text{ and } t < 1; \\ \frac{1}{1+2r} & \text{if } 1/2 \le r \le 1 \text{ and } t \ge 1; \\ \frac{t}{1+2r} & \text{if } 1/2 \le r \le 1 \text{ and } t < 1. \end{cases}$$

[1] The rate of order $O(m^{-r/(1+2r)})$ for the case $0 < r < 1/2$ can be achieved by means of leave one out analysis [19]. There is a gap between the rate in [13] and this rate. Our result decreases this gap but does not close it. People conjecture that integral operator technique with more careful analysis can close this gap. No matter whether this is true or not, in our opinion, the power of integral operator technique lies on its ability of analysis in $\mathcal{H}_K$ or in case of $r \ge 1/2$ and its easy extension to dependent samples.

The proof will be given in Section 6 where we show that the log term may even be dropped in some cases. From these rates, we notice the following facts:

1. Stronger dependence between samples implies that they contain less information and hence lead to worse rates. This is a somewhat expected property and our results do reflect it.
2. When the target function f_ρ becomes smoother (i.e., r becomes larger), the influence of the dependence becomes weaker. When $r \leq 1/2$ and $t \geq 1$, the relative gap between the rate indices for mixing sequence and for independent samples, $\big(\frac{2r}{2r+3} - \frac{2r}{(2r+3)+(1-2r)/t}\big)/\big(\frac{2r}{2r+3}\big) = \frac{(1-2r)/t}{3+2r+(1-2r)/t}$, becomes smaller as r increases. When $r > 1/2$ and $t \geq 1$, we see the rate for mixing sequences is even the same as that for independent samples.
3. Though from a rate analysis point of view, it seems the influence of dependence is not so large, we should note that this is only asymptotically true. For finite samples, the influence in fact depends on the mixing coefficients as shown by comparing the bounds in Theorems 2.2 and 2.3 with the bound in Corollary 2.4.

A class of strongly mixing sequences with the exponentially decaying mixing coefficients have caught attentions recently. In this case, our bounds give the following rates:

Corollary 2.6 *Under the assumptions of Theorem* 2.2, *if the* α*-mixing coefficients satisfy an exponential decay, i.e., there are* $a, b, c > 0$ *so that* $\alpha_i \leq ae^{-ci^b}$, *then*

(i) *if* $0 < r < \frac{1}{2}$, *we take* $\gamma = m^{-\frac{2}{2r+3}}$ *and have*

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} = O\left(m^{-\frac{2r}{2r+3}} (\log m)^{\frac{1}{2b}} \right); \tag{2.3}$$

(ii) *if* $\frac{1}{2} \leq r \leq 1$, *we take* $\gamma = m^{-\frac{1}{1+2r}}$ *and have*

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} = O\left(m^{-\frac{r}{2r+1}} \right). \tag{2.4}$$

In [11] the term “effective number of observations” was proposed for mixing sequences in the sense that though we have in hand m observations, but the information they contain is equivalent to that contained by only $m_\alpha < m$ independent samples. This implies that for an algorithm, if we have the rate of $O(m^{-\tau})$ for independent samples, we will have the rate $O(m_\alpha^{-\tau})$ for the mixing sequences. In particular, for the exponentially strongly mixing sequence, it is shown the “effective number of observations” is

$$m_\alpha = \left\lfloor m \left\lceil \left(\frac{8m}{c} \right)^{1/(1+b)} \right\rceil^{-1} \right\rfloor = O\left(m^{b/(1+b)} \right).$$

By this observation, the capacity independent rate should be $O(m^{-2rb/(3+2r)(1+b)})$ by recalling the rate for independent samples is

$O(m^{-2r/(3+2r)})$. However, we notice that our rate is the same as that for independent samples up to a log term. So we conjecture the "effective number of observations" should actually be very close to $O(m)$.

Recently Xu and Chen in [18] studied the learning rates for the exponentially strongly mixing conditional samples based on the observations in [11] and got the capacity dependent rate $m^{-\frac{br}{(b+1)(s+1)}}$ under the assumption of covering number for the unit ball of $\mathcal{H}_K$ decaying as

$$\log \mathcal{N}(\eta) \leq C_0 \eta^{-s}, \quad \forall \eta > 0$$

when $0 < r \leq \frac{1}{2}$. Their rate can be better than ours only when $\frac{2bs+2+2s-b}{2b} < r \leq \frac{1}{2}$. It is a very small region. For a less smooth kernel, say, $s \geq 1$, this is even impossible. Recall the fact that the capacity independent results correspond to the capacity dependent ones for the worst kernel case. We believe the result in [18] is far from optimal and can be improved.

For the uniformly mixing sequences we only remarked that if $\phi_i^{1/2}$ is summable to infinity the rate is the same as that for independent samples. We omit the rate analysis in detail.

3 Preliminaries

In this section, we provide some technical lemmas that will be needed in the proofs.

Our proofs will be based on the integral operator technique introduced in [14]. The following lemma gives the key properties of the integral operator (2.1).

Lemma 3.1 *The integral operator L_K has the following properties:*

(1) *L_K is a positive compact operator from $L^2_{\rho_X}(X)$ to $L^2_{\rho_X}(X)$. Consequently, it has a set of non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ in descendent order. Let e_i be the eigenfunction corresponding to λ_i. Then $\{e_1, e_2, \ldots\}$ is a complete orthonormal system of $L^2_{\rho_X}$.*

(2) *Denote $\Lambda = \{i : \lambda_i > 0\}$. Then $\{\sqrt{\lambda_i} e_i : i \in \Lambda\}$ forms an orthonormal basis of $\mathcal{H}_K$. As a consequence, for each $f \in \overline{\mathcal{H}_K}$, the closure of $\mathcal{H}_K$ in $L^2_{\rho_X}$, there holds*

$$\|f\|_{\rho_X} = \|L_K^{1/2} f\|_K.$$

Lemma 3.1 are well known properties of the integral operator L_K. Its proof can be founded in e.g. Sun and Wu (unpublished manuscript).

Our aim is to bound $\|f_{\mathbf{z},\gamma} - f_\rho\|_{\rho_X}$. We use the usual error decomposition strategy. Define

$$f_\gamma := \arg\min_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|^2_{\rho_X} + \gamma \|f\|^2_K \right\}. \tag{3.1}$$

Then we have

$$\| f_{\mathbf{z},\gamma} - f_\rho \|_{\rho_X} \leq \| f_{\mathbf{z},\gamma} - f_\gamma \|_{\rho_X} + \| f_\gamma - f_\rho \|_{\rho_X}.$$

The first term on the right is called sample error (or estimation error) and the second term is approximation error.

The approximation error does not depend on the samples. Under our assumptions, the approximation error has been studied in [14]. The following result is just [14, Lemma 3].

Lemma 3.2 *Under the assumption $L_K^{-r} f_\rho \in L^2_{\rho_X}$, there holds*

$$\| f_\gamma - f_\rho \|_{\rho_X} \leq \gamma^r \| L_K^{-r} f_\rho \|_{\rho_X} \tag{3.2}$$

In the following sections we will focus on the estimation of the sample error.

In order to deal with the mixing sequences, we use the following two lemmas. For a random variable ξ with values in a Hilbert space $\mathcal{H}$ and $1 \leq u \leq +\infty$, denote the u-th moment as $\|\xi\|_u = (\mathbb{E}\|\xi\|^u_{\mathcal{H}})^{1/u}$ if $1 \leq u < \infty$ and $\|\xi\|_\infty = \sup \|\xi\|_{\mathcal{H}}$.

Lemma 3.3 *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 u-th and v-th moments respectively. If $1 < u, v, t < +\infty$ with $u^{-1} + v^{-1} + t^{-1} = 1$ or $u = v = \infty$, $t = 1$, then*

$$|\mathbb{E}(\xi, \eta) - (\mathbb{E}\xi, \mathbb{E}\eta)| \leq 15\alpha^{\frac{1}{t}}(\mathcal{J}, \mathcal{D})\|\xi\|_u\|\eta\|_v. \tag{3.3}$$

For real-valued random variables, (3.3) is due to Davydov [7]. In the general case it was proved by Dehling and Philipp [8].

Lemma 3.4 *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 p-th and q-th moments respectively, where $p, q \geq 1$ with $p^{-1} + q^{-1} = 1$. Then*

$$|\mathbb{E}(\xi, \eta) - (\mathbb{E}\xi, \mathbb{E}\eta)| \leq 2\phi^{\frac{1}{p}}(\mathcal{J}, \mathcal{D})\|\xi\|_p\|\eta\|_q \tag{3.4}$$

For real-valued random variables, (3.4) is due to Billingsley [4] whose proof is also valid for $\mathcal{H}$-valued random variables.

4 Rough error bounds

Since $\| f_{\mathbf{z},\gamma} - f_\gamma \|_{\rho_X} \leq \kappa \| f_{\mathbf{z},\gamma} - f_\gamma \|_K$, in this section, we do the error analysis in $\mathcal{H}_K$-norm and give a rough bound.

Proposition 4.1 *Under the assumptions of Theorem 2.2, for* $0 < \delta \le +\infty$ *we have*

$$\mathbb{E}\| f_{\mathbf{z},\gamma} - f_\gamma \|_K \le \frac{\sqrt{30}\kappa M^{\frac{2}{2+\delta}}}{\sqrt{m}\gamma} \left(M + \kappa \| f_\gamma \|_K\right)^{\frac{\delta}{2+\delta}} \sqrt{1 + \sum_{i=1}^{m-1} \alpha_i^{\frac{\delta}{2+\delta}}}.$$

Denote by $\mathbf{x}$ the set of inputs $\{x_1, \ldots, x_m\}$. Define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to l^2(\mathbf{x})$ as $S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^m$. Then its adjoint is $S_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K_{x_i}$ for $c \in l^2(\mathbf{x})$. It is proved in [12–14] that

$$f_{\mathbf{z},\gamma} = \left(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1} \frac{1}{m} S_{\mathbf{x}}^T y.$$

As for f_γ, we have the representation [13]

$$f_\gamma = (L_K + \gamma I)^{-1} L_K f_\rho.$$

This leads to [14, Theorem 1]

$$f_{\mathbf{z},\gamma} - f_\gamma = \left(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1} \left\{ \frac{1}{m} \sum_{i=1}^m \left(y_i - f_\gamma(x_i)\right) K_{x_i} - L_K\left(f_\rho - f_\gamma\right) \right\}. \quad (4.1)$$

So

$$\| f_{\mathbf{z},\gamma} - f_\gamma \|_K \le \frac{1}{\gamma} \left\| \frac{1}{m} \sum_{i=1}^m \xi(z_i) - L_K\left(f_\rho - f_\gamma\right) \right\|_K \quad (4.2)$$

where $\xi(z_i) = (y_i - f_\gamma(x_i)) K_{x_i}$ is a random variable in $\mathcal{H}_K$. Proposition 4.1 is an immediate corollary of the following lemma and Schwartz inequality.

Lemma 4.2 *We have for any* $0 < \delta \le +\infty$

$$\mathbb{E} \left\| \frac{1}{m} \sum_{i=1}^m \xi(z_i) - L_K\left(f_\rho - f_\gamma\right) \right\|_K^2 \le \frac{30\kappa^2 M^{\frac{4}{2+\delta}}}{m} \left(M + \kappa \| f_\gamma \|_K\right)^{\frac{2\delta}{2+\delta}} \left(1 + \sum_{i=1}^{m-1} \alpha_i^{\frac{\delta}{2+\delta}}\right).$$

Proof Note the fact $\mathbb{E}\xi = L_K(f_\rho - f_\gamma)$. Using Lemma 3.3 with $u = v = 2 + \delta$, $t = \frac{2+\delta}{\delta}$ with $\delta > 0$ we have for $j < i$

$$\begin{aligned} \mathbb{E}\left\langle \xi(z_i), \xi(z_j) \right\rangle_K &\le \left\langle \mathbb{E}\xi(z_i), \mathbb{E}\xi(z_j) \right\rangle_K + 15 \alpha^{\frac{\delta}{2+\delta}} \left(\mathcal{M}_1^j, \mathcal{M}_i^\infty\right) \|\xi(z_i)\|_{2+\delta} \|\xi(z_j)\|_{2+\delta} \\ &\le \|L_K(f_\gamma - f_\rho)\|_K^2 + 15 \left(\alpha_{i-j}\right)^{\frac{\delta}{2+\delta}} \|\xi\|_{2+\delta}^2. \end{aligned}$$

Then direct computation leads to

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - L_K\left(f_\rho - f_\gamma\right)\right\|_K^2 \le \frac{1}{m}\|\xi\|_2^2 + \frac{30}{m}\sum_{\ell=1}^{m-1}\alpha_\ell^{\frac{\delta}{2+\delta}}\|\xi\|_{2+\delta}^2. \tag{4.3}$$

It suffices to estimate $\|\xi\|_2$ and $\|\xi\|_{2+\delta}$. By the definition of f_γ, we have

$$\mathbb{E}(y - f_\gamma(x))^2 \le \inf_{f\in\mathcal{H}_K}\left\{\mathbb{E}(y - f(x))^2 + \gamma\|f\|_K^2\right\} \le \mathbb{E}y^2 \le M^2$$

which implies

$$\|\xi\|_2^2 = \mathbb{E}\big((y - f_\gamma(x))^2 K(x,x)\big) \le \kappa^2\mathbb{E}(y - f_\gamma(x))^2 \le \kappa^2 M^2. \tag{4.4}$$

For $\|\xi\|_{2+\delta}$ with $\delta > 0$, by $|y - f_\gamma(x)| \le M + \kappa\|f_\gamma\|_K$, we obtain

$$\begin{aligned}\|\xi\|_{2+\delta} &= \left(\mathbb{E}\|\xi\|_K^{2+\delta}\right)^{1/(2+\delta)}\\ &= \left(\mathbb{E}\left((y - f_\gamma(x))^2 K(x,x)\right)^{\frac{2+\delta}{2}}\right)^{1/(2+\delta)}\\ &\le \kappa\left(M + \kappa\|f_\gamma\|_K\right)^{\delta/(2+\delta)}\left(\mathbb{E}(y - f_\gamma(x))^2\right)^{1/(2+\delta)}\\ &\le \kappa\left(M + \kappa\|f_\gamma\|_K\right)^{\delta/(2+\delta)} M^{2/(2+\delta)} \end{aligned} \tag{4.5}$$

and, if $\delta = +\infty$,

$$\|\xi\|_\infty = \|\left(y - f_\gamma(x)\right)\sqrt{K(x,x)}\|_\infty \le \kappa\left(M + \kappa\|f_\gamma\|_K\right).$$

Plugging the estimates (4.4) and (4.5) into (4.3), we finish the proof. □

Notice that $\mathbb{E}\langle\xi(z_i),\xi(z_j)\rangle_K$ are identically $\|L_K(f_\gamma - f_\rho)\|_K^2$ for all $1 \le i, j \le m$ for independent samples while for dependent samples they must be estimated using the mixing coefficients and $\|\xi\|_{2+\delta}$ with some $\delta > 0$. This is the first main difference between two cases. From the proof above we see that the estimation of $\|\xi\|_2$ is rather direct and simple but the estimation of $\|\xi\|_{2+\delta}$ is complicated and needs the upper bound of $\|f_\gamma\|_K$. This is the second difference. The following lemma provides a sharp bound for $\|f_\gamma\|_K$.

Lemma 4.3 *Under the assumption of $L_K^{-r} f_\rho \in L_{\rho_X}^2(X)$ with $0 < r \le 1$, there holds*

$$\|f_\gamma\|_K \le D\gamma^{\min\left(\frac{2r-1}{2},0\right)}$$

for some constant D.

Proof Suppose $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$ be the eigenvalues and e_i the corresponding eigenfunctions of the compact operator L_K. First we consider the case $0 < r < \frac{1}{2}$. By the fact that

$$f_\gamma = (\gamma I + L_K)^{-1} L_K f_\rho = (\gamma I + L_K)^{-1} L_K^{1+r} L_K^{-r} f_\rho = \sum_{i=1}^{\infty}\frac{\lambda_i^{1+r}}{\gamma + \lambda_i}\left\langle L_K^{-r} f_\rho, e_i\right\rangle_{\rho_X} e_i,$$

we have

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

In the case of $\frac{1}{2} \leq r \leq 1$, $f_\rho \in \mathcal{H}_K$,

$$\| f_\gamma \|_K = \| (\gamma I + L_K)^{-1} L_K f_\rho \|_K \leq \| f_\rho \|_K.$$

This proves the lemma. □

By Proposition 4.1 and Lemma 4.3 we obtain a rough error bound for learning with α-mixing sequences. Though better result is possible as will be given in next section, it is useful by providing the analysis of approximation in $\mathcal{H}_K$.

For ϕ-mixing sequences, we have the following conclusion.

Proposition 4.4 *Under the assumptions of Theorem* 2.3 *we have*

$$\mathbb{E} \| f_{\mathbf{z},\gamma} - f_\gamma \|_K \leq \frac{2\kappa M}{\sqrt{m}\gamma} \sqrt{1 + \sum_{i=1}^{m-1} \phi_i^{1/2}}. \tag{4.6}$$

Proof The proof is analogous to that for Proposition 4.1 except that we need Lemma 3.4 with $p = q = 2$ to estimate $\mathbb{E}\langle \xi(z_i), \xi(z_j) \rangle_K$ to obtain

$$\mathbb{E}\langle \xi(z_i), \xi(z_j) \rangle_K \leq \| L_K(f_\gamma - f_\rho) \|_K^2 + 2 \left(\phi_{|i-j|} \right)^{1/2} \| \xi \|_2^2.$$

□

5 Refined error bounds

In this section we provide refined bounds and prove our main theorems. To this end, we need the following concepts.

Let $HS(\mathcal{H}_K)$ be the class of all the Hilbert Schmidt operators on $\mathcal{H}_K$. It forms a Hilbert space with inner product

$$\langle T, S \rangle_{HS} := \sum_{i=1}^{\infty} \langle T\varphi_i, S\varphi_i \rangle_K$$

where φ_i is an orthonormal basis of $\mathcal{H}_K$ and this definition does not depend on the choice of the basis. We need the following properties of the Hilbert Schmidt operators.

- For any $h \in \mathcal{H}_K$ the operator defined by $h \otimes h(f) = \langle h, f \rangle_K h$ is a Hilbert Schmidt operator and

$$\| h \otimes h \|_{HS} = \| h \|_K^2.$$

- The integral operator L_K, as an operator on $\mathcal{H}_K$, belongs to $HS(\mathcal{H}_K)$ and

$$\|L_K\|_{HS}^2 = \sum_{i=1}^{\infty} \|L_K\big(\sqrt{\lambda_i} e_i\big)\|_K^2 = \sum_{i=1}^{\infty} \lambda_i^2 \leq \left(\sum_{i=1}^{\infty} \lambda_i\right)^2 \leq \kappa^4.$$

- For a Hilbert Schmidt operator T there holds $\|T\| \leq \|T\|_{HS}$.

The following lemma will be the key for our refined bounds.

Lemma 5.1 *For an α-mixing sequence $\{x_i\}$, we have*

$$\mathbb{E}\left\|L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2 \leq \mathbb{E}\left\|L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|_{HS}^2 \leq \frac{\kappa^4}{m}\left(1 + 30 \sum_{\ell=1}^{m-1} \alpha_\ell\right).$$

Proof The first inequality is trivial. To prove the second one, consider

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

as an $HS(\mathcal{H}_K)$-valued random variable. Notice that $\mathbb{E} K_x \otimes K_x = L_K$. By Lemma 3.3 with $u = v = \infty$ and $t = 1$ we have for $i \neq j$

$$\begin{aligned}\mathbb{E}\big\langle K_{x_i} \otimes K_{x_i}, K_{x_j} \otimes K_{x_j}\big\rangle_{HS} &\leq \big\langle \mathbb{E} K_{x_i} \otimes K_{x_i}, \mathbb{E} K_{x_j} \otimes K_{x_j}\big\rangle_{HS} \\ &\quad + 15 \alpha_{|i-j|} \|K_{x_i} \otimes K_{x_i}\|_\infty \|K_{x_j} \otimes K_{x_j}\|_\infty \\ &\leq \|L_K\|_{HS}^2 + 15 \kappa^4 \alpha_{|i-j|}\end{aligned}$$

where we have used the fact

$$\|K_x \otimes K_x\|_\infty = \sup_{x \in X} \|K_x \otimes K_x\|_{HS} = \sup_{x \in X} \|K_x\|_K^2 = \sup_{x \in X} K(x, x) = \kappa^2.$$

Together with the fact

$$\mathbb{E}\langle K_{x_i} \otimes K_{x_i}, K_{x_i} \otimes K_{x_i}\rangle_{HS} = \mathbb{E}\|K_{x_i} \otimes K_{x_i}\|_{HS}^2 \leq \kappa^4$$

we get

$$\mathbb{E}\left\|\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \otimes K_{x_i}\right\|_{HS}^2 \leq \frac{\kappa^4}{m} + \frac{m^2 - m}{m^2} \|L_K\|_{HS}^2 + \frac{30 \kappa^4}{m} \sum_{\ell=1}^{m-1} \alpha_\ell.$$

Then simple computation gives the desired estimate. □

Lemma 5.2 *Under assumptions of Theorem* 2.2 *we have*

$$\mathbb{E}\| f_{\mathbf{z},\gamma} - f_\gamma \|_{\rho_X} \leq \left(\gamma^{-1/2} + \gamma^{-3/2} \sqrt{\mathbb{E}\left\|L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2}\right) \sqrt{\mathbb{E}\|\Delta\|_K^2}$$

where

$$\Delta = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - f_\gamma(x_i)\right) K_{x_i} - L_K\left(f_\rho - f_\gamma\right)$$

Proof By Lemma 3.1 (2) we can write

$$\| f_{\mathbf{z},\gamma} - f_\gamma \|_{\rho_X} = \left\| L_K^{\frac{1}{2}}\left(f_{\mathbf{z},\gamma} - f_\gamma\right)\right\|_K = \left\| L_K^{\frac{1}{2}}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1}\Delta\right\|_K .$$

Using the fact $\|L_K^{\frac{1}{2}}(L_K+\gamma I)^{-1}\| \le \gamma^{-1/2}$ we have

$$\begin{aligned}
&\left\| L_K^{\frac{1}{2}}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1}\right\| \\
&\quad\le \left\| L_K^{\frac{1}{2}}(L_K+\gamma I)^{-1}\right\| + \left\| L_K^{\frac{1}{2}}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1} - L_K^{\frac{1}{2}}(L_K+\gamma I)^{-1}\right\| \\
&\quad\le \gamma^{-1/2} + \left\| L_K^{\frac{1}{2}}(L_K+\gamma I)^{-1}\left(L_K - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right)\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \gamma I\right)^{-1}\right\| \\
&\quad\le \gamma^{-1/2} + \gamma^{-3/2}\left\| L_K - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\| .
\end{aligned}$$

Therefore,

$$\| f_{\mathbf{z},\gamma} - f_\gamma \|_{\rho_X} \le \left(\gamma^{-1/2} + \gamma^{-3/2}\left\| L_K - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|\right)\|\Delta\|_K$$

and our conclusion follows by using the Schwartz inequality. □

Now we can prove our main results.

Proof of Theorem 2.2 The conclusion follows by combining Lemmas 5.2, 5.1, 4.2, 4.3, the Markov inequality, and Lemma 3.2. □

The proof of Theorem 2.3 is similar. We omit the details.

6 Learning rates

We use the refined bounds to deduce the learning rates and prove Corollaries in Section 2.

Proof of Corollary 2.5 We need the following simple facts:

$$\sum_{\ell=1}^{m-1} \ell^{-t} = \begin{cases} O(m^{1-t}) & \text{if } t < 1; \\ O(\log m) & \text{if } t = 1; \\ O(1) & \text{if } t > 1. \end{cases}$$

If $r < 1/2$ and $t \geq 1$, we take $\delta = \frac{2}{t-1}$ (if $t = 1$, $\delta = \infty$) in the error bound in Theorem 2.2. For all the other cases, we take $\delta = \infty$. The results follow by direct computation. Moreover, careful computation shows that the log term in this corollary may be dropped in case of $r > 1/2,\ t < 1$ or $r < 1/2,\ t \neq 1$. □

Proof of Corollary 2.6 We denote

$$B = a \int_0^\infty c^{-1/b} \exp\left(-y^b\right) dy < \infty.$$

Then

$$\sum_{i=1}^{m-1} \alpha^{\frac{\delta}{2+\delta}}(i) \leq a \sum_{i=1}^{m-1} \exp\left(-\frac{c\delta}{2+\delta} i^b\right) \leq a \int_0^{m-1} \exp\left(-\frac{c\delta}{2+\delta} x^b\right) dx$$
$$\leq a \int_0^\infty \left(\frac{2+\delta}{c\delta}\right)^{1/b} \exp\left(-y^b\right) dy \leq B \left(\frac{2+\delta}{\delta}\right)^{1/b}.$$

In the case of $0 < r < \frac{1}{2}$, we take $\delta = \frac{2}{\log m - 1}$ and $\gamma = m^{-2r/(3+2r)}$ in Theorem 2.2 and the result follows from the fact $m^{\delta/(2+\delta)} = m^{1/(\log m)} = \mathrm{e}$.

In case of $r \geq 1/2$ we take $\delta = \infty$ and $\gamma = m^{-\frac{1}{2r+1}}$.

This proves the conclusions. □

7 Discussions

In this paper we studied the learning performance of regularized least square regression with α mixing and ϕ mixing inputs. The capacity independent error bounds and learning rates are derived in terms of integral operator technique. They reveal many interesting phenomena of learning with dependent samples.

Our careful analysis improved the application of the integral technique to error analysis. It gives better error bounds and learning rates for learning with independent samples and sharp error bounds for learning with dependent samples.

In the literature there are some other methods leading to the capacity independent error bounds for independent samples such as stability analysis [5], leave one out analysis [19], and Rademacher average technique [3]. Whether they can be extended to deal with dependent samples is not known and may be interesting subjects for future research.

As we remarked in Section 2, our results even outperform some existing capacity dependent bounds in the literature. Recall this is impossible for independent samples. Therefore, we conjecture these existing capacity dependent results are far from optimal and can be improved.

Acknowledgements This paper was completed during the first author visited Duke University. He would like to thank Professor Sayan Mukherjee for his invitation and valuable discussions. The authors are grateful to Professor Steve Smale, Professor Ding-Xuan Zhou, and two reviewers for their valuable comments.

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
2. Athreya, K.B., Pantula, S.G.: Mixing properties of Harris chains and autoregressive processes. J. Appl. Probab. **23**, 880–892 (1986)
3. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
4. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
5. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. **2**, 499–526 (2002)
6. Cucker, F., Zhou, D.X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007)
7. Davydov, Y.A.: The invariance principle for stationary processes. Theory Probab. Appl. **14**, 487–498 (1970)
8. Dehling, H., Philipp, W.: Almost sure invariance principles for weakly dependent vector-valued random variables. Ann. Probab. **10**, 689–701 (1982)
9. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
10. Li, L.Q., Wan, C.G.: Support vector machines with beta-mixing input sequences. In: Wang, J., et al. (eds.) Lecture Notes on Computer Science, vol. 3971, pp. 928–935. Springer, New York (2006)
11. Modha, D.S.: Minimum complexity regression estimation with weakly dependent observations. IEEE. Trans. Inform. Theory **42**, 2133–2145 (1996)
12. Smale, S., Zhou, D.X.: Shannon sampling and function reconstruction from point values. Bull. Amer. Math. Soc. **41**, 279–305 (2004)
13. Smale, S., Zhou, D.X.: Shannon sampling II: connections to learning theory. Appl. Comput. Harmon. Anal. **19**, 285–302 (2005)
14. Smale, S., Zhou, D.X.: Learning theory estimates via integral operators and their approximations. Constr. Approx. **26**, 153–172 (2007)
15. Vidyasagar, M.: Learning and Generalization with Applications to Neural Networks. Springer, Berlin Heidelberg New York (2003)
16. Withers, C.S.: Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. Neural Netw. **3**, 535–549 (2000)
17. Wu, Q., Ying, Y.M., Zhou, D.X.: Learning rates of least-square regularized regression. Found. Comput. Math. **6**, 171–192 (2006)
18. Xu, Y.L., Chen, D.R.: Learning rates of regularized regression for exponentially strongly mixing sequence. J. Statist. Plann. Inference **138**(7), 2180–2189 (2008)
19. Zhang, T.: Leave-one-out bounds for kernel methods. Neural Comput. **15**, 1397–1437 (2003)