Journal of Statistical Planning and Inference 139 (2009) 3579 – 3587

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Least-square regularized regression with non-iid sampling☆

Zhi-Wei Pan*, Quan-Wu Xiao

Joint Advanced Research Center, University of Science and Technology of China and City University of Hong Kong, Suzhou, Jiangshu 215123, China

ARTICLE INFO

Article history:
Received 26 June 2008
Received in revised form
2 December 2008
Accepted 8 April 2009
Available online 18 April 2009

MSC:
68T05
62J02

Keywords:
Least-square regularized regression
Sampling with non-identical distributions
Strong mixing condition
Reproducing kernel Hilbert space

ABSTRACT

We study the least-square regression learning algorithm generated by regularization schemes in reproducing kernel Hilbert spaces. A non-iid setting is considered: the sequence of probability measures for sampling is not identical and the sampling may be dependent. When the sequence of marginal distributions for sampling converges exponentially fast in the dual of a Hölder space and the sampling process satisfies a polynomial strong mixing condition, we derive learning rates for the learning algorithm.

1. Introduction and main results

In this paper we consider the least-square regularized learning algorithm for regression with non-iid sampling.

Let (X,d) be a compact metric space (input space). Each $x\in X$ is assigned a probability measure $\rho_x=\rho(\cdot|x)$ on $Y:=\mathbb{R}$. We define our target function for learning by

$$f_\rho(x)=\int_Y y\,d\rho(y|x),\quad x\in X. \tag{1.1}$$

In the setting of regression in learning theory, $\{\rho_x\}$ are conditional distributions of a probability measure on $Z:=X\times Y$ and f_ρ is the regression function.

Our learning algorithm is a kernel method. We say that $K:X\times X\to\mathbb{R}$ is a *Mercer kernel* if it is continuous, symmetric and positive semidefinite in the sense that the matrix $(K(x_i,x_j))_{i,j=1}^l$ is positive semidefinite for any $\{x_1,\dots,x_l\}\subset X$. The *reproducing kernel Hilbert space* (RKHS) $\mathscr{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_x:=K(x,\cdot):x\in X\}$ with the inner product $\langle\cdot,\cdot\rangle_K$ given by $\langle K_x,K_y\rangle_K=K(x,y)$.

Let $\kappa=\sup_{x\in X}\sqrt{K(x,x)}$. Then the reproducing property means that

$$\langle K_x,f\rangle_K=f(x),\quad \forall x\in X,\ f\in\mathscr{H}_K. \tag{1.2}$$

☆This work was supported partially by the Research Grants Council of Hong Kong [Project no. CityU 104007], National Science Fund for Distinguished Young Scholars of China [Project no. 10529101], and National Basic Research Program of China [Project no. 973-2006CB303102].

* Corresponding author.
E-mail addresses: 50009186@student.cityu.edu.hk (Z.-W. Pan), qwxiao@mail.ustc.edu.cn (Q.-W. Xiao).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.04.007

It follows that

$$|f(x)| \leqslant \|f\|_{C(X)} \leqslant \kappa \|f\|_K, \quad f \in \mathscr{H}_K, \ x \in X. \tag{1.3}$$

The least-square regularized regression algorithm associated with the Mercer kernel K and a sample $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ is defined as

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathscr{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \tag{1.4}$$

where $\lambda \geqslant 0$ is a constant called the *regularization parameter*. It is usually chosen as $\lambda = \lambda(m)$ to depend on m and $\lim_{m\to\infty} \lambda(m) = 0$.

Throughout the paper, we assume that for some $M \geqslant 0$, ρ_X is supported on $[-M, M]$. That means $|y| \leqslant M$ almost surely and hence $|f_\rho(x)| \leqslant M$.

The aim of this paper is to study the learning performance of (1.4) with non-iid sampling. The model we take throughout the paper is from Smale and Zhou (2009) based on a sequence of probability measures $\{\rho^{(i)}\}$ on Z, such that the conditional distribution of $\rho^{(i)}$ at x equals ρ_x for every $x \in X$. The probability distribution of each pair (x_i, y_i) is $\rho^{(i)}$.

In the special case of iid sampling, the sequence $\{\rho^{(i)}\}$ is identical. There have been in the literature satisfactory learning rates such as Caponetto and De Vito (2007) and Wu et al. (2006) for capacity dependent learning rates, and Bousquet and Elisseeff (2002), De Vito et al. (2005), Smale and Zhou (2007) and Zhang (2003) for capacity independent learning rates. For a setting (Smale and Zhou, 2009) of non-identical distributions, the sampling points $\{x_i\}$ are drawn from different marginal distributions, and error analysis was done in Smale and Zhou (2009) under the assumption of independence. Shannon sampling (Smale and Zhou, 2004) and randomized sampling (Zhou and Zhou, to appear) are also examples of this setting. For dependent sampling such as weakly dependent sampling, there is an increasing literature (Modha and Masry, 1996; Steinwart et al., 2008; Sun and Wu, to appear; Xu and Chen, 2008).

The main purpose of the paper is to study learning ability of the least-square regularized regression algorithm (1.4) with non-iid sampling. Our setting does not require independence or identity, since either of them is a rather restrictive assumption in some real data analysis. The learning ability of the algorithm will be measured by learning rates.

Relaxing the independence condition, we assume the sampling sequence to be a stationary process satisfying the following mixing condition.

Definition 1. A stationary process $\{z_i\}$ is said to be α-mixing or strongly mixing if

$$\alpha(j) = \sup_{A \in \mathfrak{R}_1^k, B \in \mathfrak{R}_{k+j}^\infty, k \geqslant 1} |P(A \cap B) - P(A)P(B)| \to 0 \tag{1.5}$$

as $j \to \infty$, where $\mathfrak{R}_1^k$ and $\mathfrak{R}_{k+j}^\infty$ denote the σ-algebra of events generated by the random variables $\{z_i : 1 \leqslant i \leqslant k\}$ and $\{z_i : i \geqslant k+j\}$, respectively. It is said to satisfy an exponential strongly mixing condition, if for some positive constants a, b and c, we have

$$\alpha(i) \leqslant a \exp(-ci^b), \quad \forall i \geqslant 1. \tag{1.6}$$

It satisfies a polynomial strongly mixing condition, if for some positive constants a and b, we have

$$\alpha(i) \leqslant a i^{-b}, \quad \forall i \geqslant 1. \tag{1.7}$$

While the setting of dependent sampling (with identical distribution) has been intensively studied in the literature (e.g. Steinwart et al., 2008), the non-identical setting is less understood. The main difficulty lies in finding rules for non-identical distributions under suitable conditions. In this paper we keep the standard mixing condition for dependent sampling (Modha and Masry, 1996), and use ideas from Smale and Zhou (2009) to improve the understanding of non-identical setting.

Let $\rho_X^{(i)}$ be the marginal distribution of $\rho^{(i)}$ on X. To replace the identity of the sampling sequence $\{\rho^{(i)}\}$, we assume as in Smale and Zhou (2009) that the marginal distribution $\{\rho_X^{(i)}\}$ converges to a probability measure ρ_X in the dual $(C^s(X))^*$ of a Hölder space. Recall the Hölder space $C^s(X)$ with $0 \leqslant s \leqslant 1$, which consists of all continuous functions on X with the following norm finite:

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

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

$$\| \rho_X^{(i)} - \rho_X \|_{(C^s(X))^*} \leqslant C\alpha^i, \quad \forall i \in \mathbb{N}. \tag{1.8}$$

Remark 1. Condition (1.8) is equivalent to

$$\left|\int_X f(x)\,d\rho_X^{(i)} - \int_X f(x)\,d\rho_X\right| \leqslant C\alpha^i(\|f\|_\infty + |f|_{C^s(X)}), \quad \forall f \in C^s(X),\ i \in \mathbb{N}. \tag{1.9}$$

The exponential convergence of $\{\rho_X^{(i)}\}$ in $(C(X))^*$, the space of signed finite measures with $s = 0$, implies (1.8) for any $0 < s \leqslant 1$. So the setting with condition (1.8) is more general than the exponential convergence of $\rho_X^{(i)}$ as measures.

Definition 3. For a probability measure μ, we define an integral operator $L_{K,\mu} : L^2_\mu \to L^2_\mu$ as

$$L_{K,\mu}f = \int_X K_v f(v)\,d\mu(v).$$

It is a compact operator and its power $L^r_{K,\mu}$ is well-defined. The function f_ρ is said to satisfy the *regularity condition* (of order r) if

$$f_\rho = L^r_{K,\rho_X}(g_\rho) \quad \text{for some } g_\rho \in L^2_{\rho_X}(X). \tag{1.10}$$

Let $0 \leqslant s \leqslant 1$ be a fixed Hölder exponent used in the exponential convergence of measures.

Definition 4. We say that the Mercer kernel K satisfies the *kernel condition* (of order s) if for some constant $\kappa_{2s} > 0$, $K \in C^s(X \times X)$ and for all $u_1, u_2, v_1, v_2 \in X$

$$|K(u_1, v_1) - K(u_2, v_1) - K(u_1, v_2) + K(u_2, v_2)| \leqslant \kappa_{2s}(d(u_1, u_2))^s (d(v_1, v_2))^s. \tag{1.11}$$

Let us state our main results on the error analysis which will be proved in Section 4.

Theorem 1. *Assume that* $\{z_i\}$ *satisfies the* α*-mixing condition* (1.7) *with* $b > 0$ *and* $\{\rho_X^{(i)}\}$ *converges exponentially in* $(C^s(X))^*$ *with* $0 < s \leqslant 1$ *satisfying* (1.8). *Suppose* K *satisfies* (1.11) *and* f_ρ *has the regularity property* (1.10) *with* $\frac{1}{2} < r \leqslant \frac{3}{2}$. *If* $b \geqslant 1$, *by taking* $\lambda = m^{-b/(2br+1)}$, *we have*

$$\mathbb{E}_{z_1,\ldots,z_m}(\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leqslant \widetilde{C}\sqrt{\log m}\, m^{-b(r-1/2)/(2br+1)}, \tag{1.12}$$

where $\widetilde{C}$ *is a constant independent of* m. *If* $0 < b < 1$, *by taking* $\lambda = m^{-b/(2r+1)}$, *we have*

$$\mathbb{E}_{z_1,\ldots,z_m}(\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leqslant \widetilde{C} m^{-b(r-1/2)/(2r+1)}. \tag{1.13}$$

Learning rates (1.12) and (1.13) measure the error in the $\|\cdot\|_K$-metric which was first studied in Smale and Zhou (2007). The learning rates given there in the i.i.d. setting are of order $\mathcal{O}(m^{-(r-1/2)/(2r+1)})$ under assumption (1.10). When the index b in mixing condition (1.7) tends to infinity, the power for our learning rates in (1.12) approaches $(r - 1/2)/2r$ which is even better than the power $(r - 1/2)/(2r + 1)$ in Smale and Zhou (2007). In fact, the power for the learning rates in Smale and Zhou (2007) can be improved to $(r - 1/2)/2r$, as explained in the remark after the proof of Theorem 1 in Section 4. With this improvement, the power in (1.12) under mixing condition (1.7) with $b > 0$ is consistent to that in the i.i.d. setting by taking $b \to \infty$.

Learning rates for the regression algorithm (1.4) are usually measured in the $\|\cdot\|_{L^2_{\rho_X}}$-metric (Caponetto and De Vito, 2007; Wu et al., 2006; Zhang, 2003). In our non-iid setting, the learning rates in the $\|\cdot\|_{L^2_{\rho_X}}$-metric can be stated as follows.

Theorem 2. *Under the assumption of* Theorem 1, *if* $b > 1$, by *taking* $\lambda = m^{-1/2}$, *we have*

$$\mathbb{E}_{z_1,\ldots,z_m}\left(\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}\right) \leqslant \widetilde{C}(\log m)^{3/4} m^{-1/4}. \tag{1.14}$$

Learning rate (1.14) does not depend on r and b: regularity property (1.10) with any $r > \frac{1}{2}$ for f_ρ ensures enough decay of the approximation error in the $\|\cdot\|_{L^2_{\rho_X}}$-metric, see (4.2). This is essentially different from the analysis in the $\|\cdot\|_K$-metric. Except the logarithmic term, the power $\frac{1}{4}$ for our learning rate in (1.14) is the same as that in the i.i.d. setting given in Smale and Zhou (2007) under assumption (1.10) with $r = \frac{1}{2}$.

Remark 2. Learning rates (1.12)–(1.14) are given in expectation because our estimates depend on an inequality in expectation (Lemma 1 below, cited from Dehling and Philipp, 1982) to deal with dependency. To our best knowledge, due to the dependency, there is no exponential probability inequality in the literature which can be applied to our non-iid setting with mixing conditions.

It would be interesting to get confidence-based learning rates (of type $m^{-\theta}\log 1/\delta$ with confidence $1-\delta$), using exponential probability inequalities for dependent random variables.

2. Bounding the drift error and approximation error

We estimate the error between $f_{\mathbf{z},\lambda}$ and f_ρ by the technique of integral operators (Smale and Zhou, 2007). Learning rates (1.12) and (1.13) will be derived by taking the choice of λ as stated in Theorem 1 from more general error bound (4.1) in Theorem 4.

Write $\mu=(1/m)\sum_{i=1}^m \rho_X^{(i)}$. A noise-free limit of $f_{\mathbf{z},\lambda}$ is given by

$$f_{\lambda,\mu}=\arg\min_{f\in\mathscr{H}_K}\left\{\int_X (f(x)-f_\rho(x))^2 d\mu(x)+\lambda\|f\|_K^2\right\}. \tag{2.1}$$

We shall estimate the error $f_{\mathbf{z},\lambda}-f_\rho$ by decomposing it into three parts:

$$f_{\mathbf{z},\lambda}-f_\rho=\{f_{\mathbf{z},\lambda}-f_{\lambda,\mu}\}+\{f_{\lambda,\mu}-f_{\lambda,\rho_X}\}+\{f_{\lambda,\rho_X}-f_\rho\}. \tag{2.2}$$

2.1. Approximation error

The last term of (2.2) is incurred by the regularization parameter and is called the *approximation error* (Zhou, 2002). It does not depend on the sample. By Proposition 3 in Smale and Zhou (2009), we have the following proposition.

Proposition 1. *If f_ρ satisfies condition* (1.10) *with $\frac{1}{2}<r\leqslant\frac{3}{2}$, then for any $\lambda>0$ we have*

$$\|f_{\lambda,\rho_X}-f_\rho\|_K\leqslant\|g_\rho\|_{L^2_{\rho_X}}\lambda^{r-1/2}. \tag{2.3}$$

2.2. Drift error

The middle term of (2.2) is caused by the difference of the marginal distribution $\{\rho_X^{(i)}\}$ from the limit ρ_X. It is called the *drift error* and can be stated as follows.

Proposition 2. *Let the marginal distribution sequence $\{\rho_X^{(i)}\}$ satisfy exponential convergence condition* (1.8) *with $0\leqslant s\leqslant 1$. If f_ρ satisfies regularity condition* (1.10) *with $\frac{1}{2}<r\leqslant\frac{3}{2}$, and K satisfies kernel condition* (1.11), *then*

$$\|f_{\lambda,\mu}-f_{\lambda,\rho_X}\|_K\leqslant\frac{M_1}{m}\lambda^{r-3/2}\|g_\rho\|_{L^2_{\rho_X}}, \tag{2.4}$$

where $M_1=C(\alpha/(1-\alpha))(\kappa+\kappa_{2s})\sqrt{\kappa^2+2|K|_{C^s(X\times X)}+\kappa_{2s}}$.

Proof. When condition (1.11) is valid, it was proved in Zhou (2003) that $\mathscr{H}_K$ is included in $C^s(X)$ with the inclusion bounded as

$$\|f\|_{C^s(X)}\leqslant(\kappa+\kappa_{2s})\|f\|_K,\quad \forall f\in\mathscr{H}_K. \tag{2.5}$$

Proposition 1 in Smale and Zhou (2009) tells us that

$$\|f_{\lambda,\mu}-f_{\lambda,\rho_X}\|_K\leqslant\frac{C_K}{\lambda}\|\mu-\rho_X\|_{(C^s(X))^*}\|f_{\lambda,\rho_X}-f_\rho\|_{C^s(X)}, \tag{2.6}$$

where $C_K:=\sqrt{\kappa^2+2|K|_{C^s(X\times X)}+\kappa_{2s}}$ is a constant depending only on K. Then

$$\begin{aligned}\|f_{\lambda,\mu}-f_{\lambda,\rho_X}\|_K&\leqslant\frac{C_K}{\lambda}\left\|\frac{1}{m}\sum_{i=1}^m\rho_X^{(i)}-\rho_X\right\|_{(C^s(X))^*}\|f_{\lambda,\rho_X}-f_\rho\|_{C^s(X)}\\&\leqslant\frac{C_K}{\lambda}\frac{1}{m}\sum_{i=1}^m\|\rho_X^{(i)}-\rho_X\|_{(C^s(X))^*}\|f_{\lambda,\rho_X}-f_\rho\|_{C^s(X)}\\&\leqslant\frac{C_K}{\lambda}\frac{1}{m}\sum_{i=1}^m C\alpha^i(\kappa+\kappa_{2s})\|f_{\lambda,\rho_X}-f_\rho\|_K\\&\leqslant\frac{C_K}{\lambda}\frac{C}{m}\frac{\alpha}{1-\alpha}(\kappa+\kappa_{2s})\|f_{\lambda,\rho_X}-f_\rho\|_K,\end{aligned}$$

where $\|f_{\lambda,\rho_X} - f_\rho\|_K$ is the approximation error. Due to Proposition 1,

$$\|f_{\lambda,\mu} - f_{\lambda,\rho_X}\|_K \leqslant \frac{C_K}{\lambda}\frac{C}{m}\frac{\alpha}{1-\alpha}(\kappa+\kappa_{2s})\lambda^{r-1/2}\|g_\rho\|_{L^2_{\rho_X}}$$
$$= \frac{C_K C\frac{\alpha}{1-\alpha}(\kappa+\kappa_{2s})}{m}\lambda^{r-3/2}\|g_\rho\|_{L^2_{\rho_X}}.$$

This proves the proposition. □

3. Estimating the sample error

To deal with the dependence, we need the following probability inequality proved by Dehling and Philipp (1982) (the inequality for real-valued random variables is due to Davydov and Yu, 1970).

Lemma 1. *Let ξ and η be random variables with values in a separable Hilbert space $\mathscr{H}$ measurable σ-field $\mathscr{F}$ and $\mathscr{G}$, respectively. If $u, v, t \geqslant 1$ (possibly $+\infty$) with $u^{-1}+v^{-1}+t^{-1}=1$, then*

$$|\mathbb{E}(\xi,\eta) - (\mathbb{E}\xi, \mathbb{E}\eta)| \leqslant 15\alpha^{1/t}(\mathscr{F},\mathscr{G})\|\xi\|_u\|\eta\|_v, \tag{3.1}$$

where $\alpha(\mathscr{F},\mathscr{G}) = \sup_{A\in\mathscr{F},B\in\mathscr{G}}|P(A\bigcap B) - P(A)P(B)|$ and $\|\xi\|_u = (\mathbb{E}\|\xi\|^u_{\mathscr{H}})^{1/u}$.

Proposition 3. *Let the random sequence $\{z_i\}$ satisfy the α-mixing condition, and the marginal distribution sequence $\{\rho_X^{(i)}\}$ satisfy the exponential convergence condition* (1.8). *Let $\delta > 0$. If $f_\rho \in \mathscr{H}_K$, then*

$$\mathbb{E}_{z_1,\ldots,z_m}\{\|f_{\mathbf{z},\lambda} - f_{\lambda,\mu}\|_K\} \leqslant \frac{\kappa\|f_\rho\|_K}{\sqrt{m\lambda}} + 6\kappa(M+\kappa\|f_\rho\|_K)^{\delta/(4+2\delta)}\|f_\rho\|_K^{2/(2+\delta)}\frac{\sqrt{\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}}}{\sqrt{m\lambda}\lambda^{\delta/(4+2\delta)}}. \tag{3.2}$$

Proof. Denote $\mathbf{x} = \{x_i\}_{i=1}^m$. Recall the operator $S_{\mathbf{x}}^T S_{\mathbf{x}} : \mathscr{H}_K \to \mathscr{H}_K$ given by $S_{\mathbf{x}}^T S_{\mathbf{x}} f = \sum_{i=1}^m f(x_i)K_{x_i}$. By an expression obtained in Theorem 1 of Smale and Zhou (2005, 2007),

$$f_{\mathbf{z},\lambda} - f_{\lambda,\mu} = \left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f_{\lambda,\mu}(x_i))K_{x_i} - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\}. \tag{3.3}$$

Denote the random variable ξ with values in $\mathscr{H}_K$ given by $\xi(z) = (y - f_{\lambda,\mu}(x))K_x$. Then

$$\|f_{\mathbf{z},\lambda} - f_{\lambda,\mu}\|_K \leqslant \frac{\Delta}{\lambda},$$

where

$$\Delta := \left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\|_K. \tag{3.4}$$

Taking inner products in $\mathscr{H}_K$, we have

$$\mathbb{E}_{z_1,\ldots,z_m}(\Delta^2) \leqslant I_1 + I_2 - 2I_3 + \|L_{K,\mu}(f_\rho - f_{\lambda,\mu})\|_K^2$$
$$:= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_{z_i}(y_i - f_{\lambda,\mu}(x_i))^2 K(x_i,x_i) + \frac{1}{m^2}\sum_{i\neq j}\mathbb{E}_{z_i,z_j}\langle\xi(z_i),\xi(z_j)\rangle_K$$
$$- \frac{2}{m}\sum_{i=1}^m \mathbb{E}_{z_i}(y_i - f_{\lambda,\mu}(x_i))L_{K,\mu}(f_\rho - f_{\lambda,\mu})(x_i) + \|L_{K,\mu}(f_\rho - f_{\lambda,\mu})\|_K^2.$$

Since $\|L_{K,\mu}^{1/2} f\|_K = \|f\|_{L^2_\mu}$ for $f \in L^2_\mu$, we know that

$$I_3 = \int_X (f_\rho(x) - f_{\lambda,\mu}(x))L_{K,\mu}(f_\rho - f_{\lambda,\mu})(x)\,d\mu = \|L_{K,\mu}^{1/2}(f_\rho - f_{\lambda,\mu})\|^2_{L^2_\mu} = \|L_{K,\mu}(f_\rho - f_{\lambda,\mu})\|_K^2.$$

Next we estimate the crucial part I_2 involving the weak dependence. When $i \neq j$, we apply Lemma 1 to $\xi = \xi(z_i)$ and $\eta = \xi(z_j)$ with $u = v = 2+\delta, t = (2+\delta)/\delta$ and see

$$\mathbb{E}_{z_i,z_j}\langle\xi(z_i),\xi(z_j)\rangle_K - \langle\mathbb{E}_{z_i}\xi(z_i), \mathbb{E}_{z_j}\xi(z_j)\rangle_K \leqslant 15(\alpha(|i-j|))^{\delta/(2+\delta)}\|\xi(z_i)\|_{2+\delta}\|\xi(z_j)\|_{2+\delta}. \tag{3.5}$$

But $\mathbb{E}_{z_i}\xi(z_i)=\int_X(f_\rho(x)-f_{\lambda,\mu}(x))K_x d\rho_X^{(i)}$ and

$$\|\xi(z_i)\|_{2+\delta}^{2+\delta}=\int_Z\|\xi(z_i)\|_K^{2+\delta}d\rho^{(i)}=\int_Z[(y-f_{\lambda,\mu}(x))^2K(x,x)]^{(2+\delta)/2}d\rho^{(i)}$$
$$\leqslant(M+\|f_{\lambda,\mu}\|_\infty)^{\delta/2}\kappa^\delta\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)\,d\rho_X^{(i)}. \tag{3.6}$$

It follows that

$$I_2\leqslant\frac{1}{m^2}\sum_{i\neq j}\int_X\int_X(f_\rho(x)-f_{\lambda,\mu}(x))(f_\rho(w)-f_{\lambda,\mu}(w))K(x,w)\,d\rho_X^{(i)}(x)\,d\rho_X^{(j)}(w)$$
$$+\frac{15}{m^2}\sum_{i\neq j}(\alpha(|i-j|))^{\delta/(2+\delta)}\kappa^{2\delta/(2+\delta)}(M+\|f_{\lambda,\mu}\|_\infty)^{\delta/(2+\delta)}$$
$$\times\left\{\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)d\rho_X^{(i)}\right\}^{1/(2+\delta)}\left\{\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)d\rho_X^{(j)}\right\}^{1/(2+\delta)}.$$

The last term on the right side is bounded by

$$\frac{15}{m^2}\kappa^{2\delta/(2+\delta)}(M+\|f_{\lambda,\mu}\|_\infty)^{\delta/(2+\delta)}\sum_{i=1}^m\left\{\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)d\rho_X^{(i)}\right\}^{2/(2+\delta)}\sum_{j\neq i}(\alpha(|i-j|))^{\delta/(2+\delta)}$$
$$\leqslant\frac{15}{m}\kappa^{2\delta/(2+\delta)}(M+\|f_{\lambda,\mu}\|_\infty)^{\delta/(2+\delta)}\left\{\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)d\mu\right\}^{2/(2+\delta)}2\sum_{l=1}^{m-1}(\alpha(l))^{\delta/(2+\delta)}. \tag{3.7}$$

Taking $f=f_\rho$ in definition (2.1) of $f_{\lambda,\mu}$, we know that

$$\int_X(f_{\lambda,\mu}(x)-f_\rho(x))^2d\mu+\lambda\|f_{\lambda,\mu}\|_K^2\leqslant\lambda\|f_\rho\|_K^2. \tag{3.8}$$

It yields $\|f_{\lambda,\mu}\|_K\leqslant\|f_\rho\|_K$ and

$$\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)\,d\mu\leqslant\kappa^2\lambda\|f_\rho\|_K^2. \tag{3.9}$$

Hence the last term for bounding I_2 is at most

$$\frac{30}{m}\kappa^{2\delta/(2+\delta)}(M+\kappa\|f_\rho\|_K)^{\delta/(2+\delta)}(\kappa^2\lambda\|f_\rho\|_K^2)^{2/(2+\delta)}\sum_{l=1}^{m-1}(\alpha(l))^{\delta/(2+\delta)}$$
$$\leqslant30\kappa^2(M+\kappa\|f_\rho\|_K)^{\delta/(2+\delta)}\|f_\rho\|_K^{4/(2+\delta)}\frac{\lambda^{2/(2+\delta)}}{m}\sum_{l=1}^{m-1}(\alpha(l))^{\delta/(2+\delta)}. \tag{3.10}$$

The first term for bounding I_2 equals

$$\|L_{K,\mu}(f_\rho-f_{\lambda,\mu})\|_K^2-\frac{1}{m^2}\sum_{i=1}^m\int_X\int_X(f_\rho(x)-f_{\lambda,\mu}(x))(f_\rho(w)-f_{\lambda,\mu}(w))K(x,w)\,d\rho_X^{(i)}(x)\,d\rho_X^{(i)}(w),$$

which is bounded by $\|L_{K,\mu}(f_\rho-f_{\lambda,\mu})\|_K^2$ according to the Mercer kernel property. Also, we see that

$$I_1=\frac{1}{m^2}\sum_{i=1}^m\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)\,d\rho_X^{(i)}$$
$$=\frac{1}{m}\int_X(f_\rho(x)-f_{\lambda,\mu}(x))^2K(x,x)\,d\mu\leqslant\frac{\kappa^2\lambda\|f_\rho\|_K^2}{m}.$$

Combining all the estimates for I_1,I_2,I_3, we know that

$$\mathbb{E}_{z_1,\ldots,z_m}(\Delta^2)\leqslant\frac{\kappa^2\lambda\|f_\rho\|_K^2}{m}+30\kappa^2(M+\kappa\|f_\rho\|_K)^{\delta/(2+\delta)}\|f_\rho\|_K^{4/(2+\delta)}\frac{\lambda^{2/(2+\delta)}}{m}\sum_{l=1}^{m-1}(\alpha(l))^{\delta/(2+\delta)}. \tag{3.11}$$

This proves the desired bound. □

When the error is measured in the $\|\cdot\|_{L^2_{\rho_X}}$-metric, it was shown in Caponetto and De Vito (2007) and Smale and Zhou (2007) that bounds for the sample error can be improved by means of the norm relationship $\|L_{K,\mu}^{1/2} f\|_K = \|f\|_{L^2_\mu}$. In our non-iid setting, we have the following estimate for the sample error.

Proposition 4. *Let the random sequence $\{z_i\}$ satisfy the α-mixing condition, and the marginal distribution sequence $\{\rho_X^{(i)}\}$ satisfy the exponential convergence condition* (1.8). *Let $\delta > 0$. If $f_\rho \in \mathscr{H}_K$, then*

$$\mathbb{E}_{z_1,\dots,z_m}\{\|f_{\mathbf{z},\lambda} - f_{\lambda,\mu}\|_{L^2_{\rho_X}}\} \leqslant \left[1 + \frac{\kappa m^{1/4}}{\sqrt{m\lambda}}\left\{1 + 30\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}\right\}^{1/4}\right] \times \left(\frac{\kappa\|f_\rho\|_K}{\sqrt{m}} + 6\kappa(M + \kappa\|f_\rho\|_K)^{\delta/(4+2\delta)}\|f_\rho\|_K^{2/(2+\delta)}\frac{\sqrt{\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}}}{\sqrt{m}\lambda^{\delta/(4+2\delta)}}\right).$$

Proof. Applying the relation $\|L_{K,\mu}^{1/2} f\|_K = \|f\|_{L^2_\mu}$ to the function $f = f_{\mathbf{z},\lambda} - f_{\lambda,\mu}$ in (3.3) we know that $\mathbb{E}_{z_1,\dots,z_m}\{\|f_{\mathbf{z},\lambda} - f_{\lambda,\mu}\|_{L^2_{\rho_X}}\}$ equals

$$\begin{aligned}
&\mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu}^{1/2}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f_{\lambda,\mu}(x_i))K_{x_i} - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\}\right\|_K \\
&\leqslant \mathbb{E}_{z_1,\dots,z_m}\left\|\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right)^{1/2}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f_{\lambda,\mu}(x_i))K_{x_i} - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\}\right\|_K \\
&\quad + \mathbb{E}_{z_1,\dots,z_m}\left\|\left\{L_{K,\mu}^{1/2} - \left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right)^{1/2}\right\}\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f_{\lambda,\mu}(x_i))K_{x_i} - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\}\right\|_K.
\end{aligned}$$

This in connection with the inequality $\|L_{K,\mu}^{1/2} - ((1/m)S_{\mathbf{x}}^T S_{\mathbf{x}})^{1/2}\| \leqslant \|L_{K,\mu} - (1/m)S_{\mathbf{x}}^T S_{\mathbf{x}}\|^{1/2}$ given as Theorem 2.1 in Sun and Wu (2009) implies

$$\begin{aligned}
\mathbb{E}_{z_1,\dots,z_m}\{\|f_{\mathbf{z},\lambda} - f_{\lambda,\mu}\|_{L^2_{\rho_X}}\} &\leqslant \frac{1}{\sqrt{\lambda}}\mathbb{E}_{z_1,\dots,z_m}\Delta + \mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu} - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^{1/2} \\
&\quad \times \left\|\left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f_{\lambda,\mu}(x_i))K_{x_i} - L_{K,\mu}(f_\rho - f_{\lambda,\mu})\right\}\right\|_K \\
&\leqslant \frac{1}{\sqrt{\lambda}}\mathbb{E}_{z_1,\dots,z_m}\Delta + \sqrt{\mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu} - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|}\cdot\sqrt{\mathbb{E}_{z_1,\dots,z_m}\left(\frac{1}{\lambda}\Delta\right)^2}.
\end{aligned}$$

Besides bound (3.11) for $\mathbb{E}_{z_1,\dots,z_m}\Delta$, we need to estimate the operator norm $\|L_{K,\mu} - (1/m)S_{\mathbf{x}}^T S_{\mathbf{x}}\|$. Denote $\zeta(x) := K_x\langle\cdot, K_x\rangle_K$ which is a rank one operator on $\mathscr{H}_K$ for $x \in X$. Consider ζ to be a random variable with values in $HS(\mathscr{H}_K)$, the Hilbert space of Hilbert–Schmidt operates on $\mathscr{H}_K$, with inner product $\langle A, B\rangle_{HS} = Tr(B^T A)$. Here Tr denotes the trace of a (trace-class) linear operator. The space $HS(\mathscr{H}_K)$ is a subspace of the space of bounded linear operator on $\mathscr{H}_K$, denoted as $(L(\mathscr{H}_K), \|\cdot\|)$, with the norm relations

$$\|A\| \leqslant \|A\|_{HS}, \quad \|AB\|_{HS} \leqslant \|A\|_{HS}\|B\|. \tag{3.12}$$

As in the proof of Proposition 3, we have

$$\begin{aligned}
\mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu} - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|^2 &\leqslant \mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu} - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}\right\|_{HS}^2 \\
&\leqslant \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_{z_i}\langle\zeta(x_i), \zeta(x_i)\rangle_{HS} + \frac{1}{m^2}\sum_{i\neq j}^m \mathbb{E}_{z_i,z_j}\langle\zeta(x_i), \zeta(x_j)\rangle_{HS} - \frac{2}{m}\sum_{i=1}^m \mathbb{E}_{z_i}\langle\zeta(x_i), L_{K,\mu}\rangle_{HS} + \|L_{K,\mu}\|_{HS} \\
&\leqslant \frac{\kappa^4}{m} + \frac{15}{m^2}\sum_{i\neq j}(\alpha(|i-j|))^{\delta/(2+\delta)}\|\zeta(x_i)\|_{2+\delta}\|\zeta(x_j)\|_{2+\delta} \leqslant \frac{\kappa^4}{m}\left(1 + 30\sum_{l=1}^{m-1}(\alpha(l))^{\delta/(2+\delta)}\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}_{z_1,\dots,z_m}\{\|f_{\mathbf{z},\lambda}-f_{\lambda,\mu}\|_{L^2_{\rho_X}}\}\leqslant\frac{1}{\sqrt{\lambda}}\sqrt{\mathbb{E}_{z_1,\dots,z_m}\Delta^2}\left(1+\frac{1}{\sqrt{\lambda}}\sqrt{\mathbb{E}_{z_1,\dots,z_m}\left\|L_{K,\mu}-\frac{1}{m}S_{\mathbf{x}}^TS_{\mathbf{x}}\right\|}\right)$$

can be bounded as stated. This proves Proposition 4. □

4. Deriving learning rates

Combining the bounds in Propositions 1–3, we get the following estimate for the error $f_{\mathbf{z},\lambda}-f_\rho$ in the $\mathscr{H}_K$-metric.

Theorem 3. *Assume that $\{z_i\}$ satisfies α-mixing condition (1.7) and $\{\rho_X^{(i)}\}$ satisfies exponential convergence (1.8) with $0<s\leqslant 1$. If K satisfies kernel condition (1.11) and f_ρ has regularity property (1.10) for some $\frac{1}{2}<r\leqslant\frac{3}{2}$, then for any $\delta>0$, we have*

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K)\leqslant\widetilde{C}^*\left\{\lambda^{r-1/2}+\frac{\lambda^{r-1/2}}{m\lambda}+\frac{1}{\sqrt{m\lambda}}+\frac{\sqrt{\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}}}{\sqrt{m\lambda}\lambda^{\delta/(4+2\delta)}}\right\},\tag{4.1}$$

where $\widetilde{C}^*=\|g_\rho\|_{L^2_{\rho_X}}+M_1\|g_\rho\|_{L^2_{\rho_X}}+\kappa\|f_\rho\|_K+6\kappa\|f_\rho\|_K^{2/(2+\delta)}(M+\kappa\|f_\rho\|_K)^{\delta/(4+2\delta)}$.

In the same way, the following estimate for the error $f_{\mathbf{z},\lambda}-f_\rho$ in the $L^2_{\rho_X}$-metric is a consequence of Propositions 2 and 4 and a trivial bound for the approximation error when $f_\rho\in\mathscr{H}_K$:

$$\|f_{\lambda,\rho_X}-f_\rho\|^2_{L^2_{\rho_X}}+\lambda\|f_{\lambda,\rho_X}\|_K^2\leqslant\lambda\|f_\rho\|_K^2\quad\forall\lambda>0.\tag{4.2}$$

Theorem 4. *Assume that $\{z_i\}$ satisfies α-mixing condition (1.7) and $\{\rho_X^{(i)}\}$ satisfies exponential convergence (1.8) with $0<s\leqslant 1$. If K satisfies kernel condition (1.11) and f_ρ has regularity property (1.10) for some $\frac{1}{2}<r\leqslant\frac{3}{2}$, then for any $\delta>0$, we have*

$$\begin{aligned}&\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_{L^2_{\rho_X}})\\&\leqslant\widetilde{C}^{**}\left\{\lambda^{1/2}+\frac{\lambda^{r-1/2}}{m\lambda}+\left(\frac{1}{\sqrt{m}}+\frac{\sqrt{\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}}}{\sqrt{m}\lambda^{\delta/(4+2\delta)}}\right)\left[1+\frac{m^{1/4}}{\sqrt{m\lambda}}+\frac{m^{1/4}}{\sqrt{m\lambda}}\left(\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}\right)^{1/4}\right]\right\},\end{aligned}$$

where $\widetilde{C}^{**}=\|f_\rho\|_K+\kappa M_1\|g_\rho\|_{L^2_{\rho_X}}+\kappa\|f_\rho\|_K+6\kappa\|f_\rho\|_K^{2/(2+\delta)}(M+\kappa\|f_\rho\|_K)^{\delta/(4+2\delta)}(1+3\kappa)$.

We are in a position to prove Theorem 1 on learning rates in the K-metric stated in the Introduction.

Proof of Theorem 1. When $1\leqslant b<\infty$, we take $\delta=2/(b-1)>0$. Then $\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}\leqslant\sum_{i=1}^{m}ai^{-1}\leqslant a\log m$. By Theorem 4, we have

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K)\leqslant\widetilde{C}^*\left\{\lambda^{r-1/2}+\frac{\lambda^{r-1/2}}{m\lambda}+\frac{1}{\sqrt{m\lambda}}+\frac{\sqrt{a}\sqrt{\log m}}{\sqrt{m\lambda}\lambda^{1/2b}}\right\}.\tag{4.3}$$

Thus when $\lambda^{r-1/2}=1/\sqrt{m\lambda}\lambda^{1/2b}$, that is, $\lambda=m^{-b/(2br+1)}$, we have

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K)\leqslant(3+\sqrt{a})\widetilde{C}^*\sqrt{\log m}m^{-b(r-1/2)/(2br+1)}.\tag{4.4}$$

This proves (1.12).

When $0<b<1$, we take $\delta=\infty$. Then $\sum_{i=1}^{m-1}(\alpha(i))^{\delta/(2+\delta)}=\sum_{i=1}^{m-1}\alpha(i)\leqslant(a/(1-b))m^{1-b}$, and we conclude from Theorem 4 that

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K)\leqslant\widetilde{C}^*\left\{\lambda^{r-1/2}+\frac{\lambda^{r-1/2}}{m\lambda}+\frac{1}{\sqrt{m\lambda}}+\frac{\sqrt{\dfrac{a}{1-b}}m^{(1-b)/2}}{\sqrt{m\lambda}\lambda^{1/2}}\right\}.\tag{4.5}$$

When $\lambda=m^{-b/(2r+1)}$, this yields

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K)\leqslant\left(3+\sqrt{\frac{a}{1-b}}\right)\widetilde{C}^*m^{-b(r-1/2)/(2r+1)}.\tag{4.6}$$

This proves Theorem 1. □

Remark 3. In the above proof, we make full use of assumption (1.10) with $\frac{1}{2} < r \leqslant \frac{3}{2}$ and apply bound (2.3) from Proposition 1 for the approximation error. In Smale and Zhou (2007) the estimate for the sample error $\|f_{\mathbf{z},\lambda} - f_{\lambda,\rho_X}\|_K$ (stated as Theorem 1 there) was provided for general measures without special condition (1.10). Hence the learning rates there were not optimal when f_ρ satisfies (1.10) with $\frac{1}{2} < r \leqslant \frac{3}{2}$. If one uses (2.3) in this special case, the learning rate for $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ in Smale and Zhou (2007) can be improved to $\mathcal{O}(m^{-(r-1/2)/2r})$, which would be consistent to our learning rate (1.12).

Another mixing condition is induced by ϕ-mixing coefficients as follows.

Definition 5. A stationary process $\{z_i\}$ is said to be ϕ-mixing if

$$\phi(j) = \sup_{A\in\mathfrak{R}_1^k, B\in\mathfrak{R}_{k+j}^\infty, k\geqslant 1} |P(A|B) - P(A)| \to 0 \quad (\text{as } j \to \infty). \tag{4.7}$$

In this case, the following probability inequality is due to Billingsley (1968) whose proof is also valid for $\mathscr{H}$-valued random variables.

Lemma 2. *Let ξ and η be random variables with values in a separable Hilbert space $\mathscr{H}$ measurable $\mathscr{F}$ and $\mathscr{G}$, respectively. If $p, q \geqslant 1$ satisfy $p^{-1} + q^{-1} = 1$, then*

$$|\mathbb{E}(\xi,\eta) - (\mathbb{E}\xi, \mathbb{E}\eta)| \leqslant 2\phi^{1/p}(\mathscr{F},\mathscr{G})\|\xi\|_p\|\eta\|_q, \tag{4.8}$$

where $\phi(\mathscr{F},\mathscr{G}) = \sup_{A\in\mathscr{F}, B\in\mathscr{G}}|P(A|B) - P(A)|$.

With this inequality, we can do the error analysis as follows.

Theorem 5. *Assume that $\{z_i\}$ satisfies the ϕ-mixing condition (4.7) and $\{\rho_X^{(i)}\}$ satisfies exponential convergence (1.8) with $0 < s \leqslant 1$. If K satisfies kernel condition (1.11) and f_ρ has regularity property (1.10) for some $\frac{1}{2} < r \leqslant \frac{3}{2}$, then for any $\delta > 0$, we have*

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leqslant \widetilde{C}^* \left\{ \lambda^{r-1/2} + \frac{\lambda^{r-1/2}}{m\lambda} + \frac{1}{\sqrt{m\lambda}} + \frac{\sqrt{\sum_{i=1}^{m-1}(\phi(i))^{\delta/(2+\delta)}}}{\sqrt{m\lambda}\lambda^{\delta/(4+2\delta)}} \right\}, \tag{4.9}$$

where $\widetilde{C}^* = \|g_\rho\|_{L^2_{\rho_X}} + M_1\|g_\rho\|_{L^2_{\rho_X}} + \kappa\|f_\rho\|_K + 2\kappa\|f_\rho\|_K^{2/(2+\delta)}(M + \kappa\|f_\rho\|_K)^{\delta/(4+2\delta)}$.

In particular, if $\{\phi(j)\}$ decays as $\phi(j) = \mathcal{O}(j^{-b})$ for some $b > 0$, then we derive easily the following learning rates by taking suitable choices of λ.

Corollary 1. *Under the condition of Theorem 5, if $\phi(j) \leqslant aj^{-b}$ for some $b \geqslant 1$, then*

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leqslant \widetilde{C}\sqrt{\log m} m^{-b(r-1/2)/(2br+1)} \quad \text{by taking } \lambda = m^{-b/(2br+1)},$$

where $\widetilde{C}$ is a constant independent of m. If $\phi(j) \leqslant aj^{-b}$ for some $0 < b < 1$, then

$$\mathbb{E}_{z_1,\dots,z_m}(\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leqslant \widetilde{C} m^{-b(r-1/2)/(2r+1)} \quad \text{by taking } \lambda = m^{-b/(2r+1)}.$$

References

Billingsley, P., 1968. Convergence of Probability Measures. Wiley, New York.
Bousquet, O., Elisseeff, A., 2002. Stability and generation. J. Machine Learning Res. 2, 499–526.
Caponetto, A., De Vito, E., 2007. Optimal rates of regularized least-square algorithm. Found. Comput. Math. 7, 331–368.
Davydov, Y., Yu, A., 1970. The invariance principle for stationary process. Theory Probab. Appl. 14, 487–498.
Dehling, H., Philipp, W., 1982. Almost sure invariance principles for weakly dependent vector-valued random variables. Anal. Probab. 10, 689–701.
De Vito, E., Caponnetto, A., Rosasco, L., 2005. Model selection for regularized least-square algorithm in learning theory. Found. Comput. Math. 5, 59–85.
Modha, D.S., Masry, E., 1996. Minimum complexity regression estimation with weakly dependent observations. IEEE Trans. Inform. Theory 42, 2133–2145.
Smale, S., Zhou, D.X., 2004. Shannon sampling and function reconstruction from point values. Bull. Amer. Math. Soc. 41, 279–305.
Smale, S., Zhou, D.X., 2005. Shannon sampling II: connection to learning theory. Appl. Comput. Harmonic Anal. 19, 285–302.
Smale, S., Zhou, D.X., 2007. Learning theory estimates via integral operators and their approximations. Constr. Approx. 26, 153–172.
Smale, S., Zhou, D.X., 2009. Online learning with Markov sampling. Anal. Appl. 7, 87–113.
Steinwart, I., Hush, D., Scovel, C., 2008. Learning from dependent observations. J. Multivariate Anal. 100, 175–194.
Sun, H.W., Wu, Q., 2009. A note on application of integral operator in learning theory. Appl. Comput. Harmonic Anal. 26, 416–421.
Sun, H.W., Wu, Q., Regularized least square regression with dependent samples. Adv. Comput. Math., to appear. DOI: 10.1007/s10444-008-9099-y.
Wu, Q., Ying, Y., Zhou, D.X., 2006. Learning rates of least-square regularized regression. Found. Comput. Math. 6, 171–192.
Xu, Y.L., Chen, D.R., 2008. Learning rates of regularized regression for exponentially strongly mixing sequence. J. Statist. Plann. Inference 138, 2180–2189.
Zhang, T., 2003. Leave-one-out bounds for kernel methods. Neural Comput. 15, 1397–1437.
Zhou, D.X., 2003. Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inform. Theory 49, 1743–1752.
Zhou, D.X., 2002. The covering number in learning theory. J. Complexity 18, 739–767.
Zhou, X.J., Zhou, D.X., Higher order parzen windows and randomized sampling. Adv. Comput. Math., to appear. DOI: 10.1007/s10444-008-9073-8.