Indefinite Kernel Network with Dependent Sampling

Hongwei Sun* and Qiang Wu†

* School of Mathematical Science, University of Jinan,

Shandong Provincial Key Laboratory of Network based Intelligent Computing

Jinan 250022, People's Republic of China

E-mail: ss_sunhw@ujn.edu.cn

† Department of Mathematical Sciences, Middle Tennessee State University

Murfreesboro, TN 37132, USA

E-mail: wuqiangmath@gmail.com

Abstract

We study the asymptotical properties of indefinite kernel network with coefficient regularization and dependent sampling. The framework under investigation is different from classical kernel learning. Positive definiteness is not required by the kernel function and the samples are allowed to be weakly dependent with the dependence measured by a strong mixing condition. By a new kernel decomposition technique introduced in [25], two reproducing kernel Hilbert spaces and their associated kernel integral operators are used to characterize the properties and learnability of the hypothesis function class. Capacity independent error bounds and learning rates are deduced.

Keywords: Kernel network; indefinite kernel; regression learning; regularization; α-mixing condition; consistency.

1 Introduction

The aim of this paper is to establish the mathematical foundation of indefinite kernel network for regression learning with dependent sampling. For this purpose we study its asymptotical properties, prove the consistency, and estimate the learning rates.

In the context of statistical learning theory [22], the framework of regression learning is usually described as follows: Let X be a domain of $\mathbb{R}^n$ and $Y = \mathbb{R}$, ρ be a non-degenerate Borel probability distribution on $Z = X \times Y$. The regression function $f_\rho : X \to Y$ is given by

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

where $\rho(y|x)$ is the conditional distribution of y for given x. The target of regression learning is to find an good approximation of f_ρ from a set of observations $\mathbf{z} = \big\{(x_i,\, y_i)\big\}_{i=1}^m \in Z^m$ drawn from the unknown probability measure ρ.

In kernel network for regression learning, a kernel function $K : X \times X \to \mathbb{R}$ plays the role of basis function and the regression function is learned from superpositions of the kernel functions. To be precise, let

$$\mathscr{H}_{K,\mathbf{x}} = \left\{ f_{\mathbf{c}}(x) = \sum_{i=1}^m c_i K(x, x_i) : \mathbf{c} = (c_1, \ldots, c_m) \in \mathbb{R}^m, m \in \mathbb{N} \right\}.$$

Then the approximation is searched within $\mathscr{H}_{K,\mathbf{x}}$ by minimizing the least square error. Using a penalty term to stabilize this ill posed problem, we get the regularized kernel network

$$f_{\mathbf{z}} = f_{\mathbf{c}_\mathbf{z}} \quad \text{where} \quad \mathbf{c}_\mathbf{z} = \arg\min_{\mathbf{c}\in\mathbb{R}^m} \frac{1}{m} \sum_{i=1}^m (y_i - f_{\mathbf{c}}(x_i))^2 + \lambda m \sum_{i=1}^m c_i^2. \tag{1.1}$$

Unlike the traditional kernel methods where the kernel function is always positive definite, in this paper we relax the requirement of the kernel function. We assume it is a quite general bivariate function satisfying only the continuity and uniform boundedness on X which we call indefinite kernel. Note that if the kernel is positive definite, the kernel matrix $K_\mathbf{x} = (K(x_i, x_j)_{i,j=1}^m$ is positive definite. Then $\mathbf{c}K_\mathbf{x}\mathbf{c}^\top$ is a good choice of the regularizer and the algorithm reduces to the traditional kernel regression which has been extensively studied in the literature; see [4, 8–10, 18, 19, 29] and references therein.

When K is an indefinite kernel, since the kernel matrix is not necessarily positive definite, the above quadratic form cannot play the role of the stabilizer. Instead ℓ^2 norm of the coefficient vector $\mathbf{c}$ is a good choice in this setting. Indefinite kernels find their applications in many areas in recent years, see [11, 14, 17]. Several machine learning algorithms associated to indefinite kernels or matrices were developed. For instance, learning with Reproducing Kernel Krein Spaces was proposed in [7] and support vector machines with indefinite kernels was introduce in [12, 28]. These works motivated the study of the mathematical foundations for the learning with indefinite kernels [21, 24–26].

Although some studies have been done for learning with indefinite kernels, there are questions left unanswered. For (1.1) we first proved a capacity independent result in [21] where the indefinite kernel is only assumed continuous and uniformly bounded. The consistency under so weak conditions is encouragingly supportive for indefinite kernel learning. However, the method used there is tricky and complicated. It has two shortcomings: firstly it leads to slow convergence rates and do not support the competition of indefinite kernels versus positive kernels. Secondly, no evidence shows that it can be easily adapted to study the dependent sampling which is very common in practice. More recently the second author proposed a new technique in [25] that used two reproducing kernel Hilbert spaces and their associated kernel integral operators to characterize the properties of hypothesis function class $\mathscr{H}_{K,\mathbf{x}}$ and was able to improve the learning rate. This technique also shed light to the study of this algorithm with dependent sampling.

To study asymptotical properties of indefinite kernel network with dependent sampling and establish its mathematical foundation will be the main contribution of this paper. For the kernel function, except for the continuity and uniform boundedness we will further assume a regular condition which will be stated in Section 2. For dependent sampling we consider strongly mixing sequence which has been shown very common in sampling process (see [1, 2, 6, 13] and the references therein) and adopted in statistical learning theory, see [15, 20, 27, 30]. We assume the dependence between samples is

measured by α-mixing condition: for two σ-fields $\mathcal{J}$ and $\mathcal{D}$, define the α-coefficient as

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

For 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 random sequence $z_i,\ i \geq 1$, is said to satisfy a α-mixing condition if

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

Except for indefinite kernel and dependent sampling assumption in this framework, we will also relax a widely used boundedness restriction on the output variables y. Instead we use a weaker condition: for some constants $c,\ M \geq 1$,

$$\int_Z |y|^l d\rho \leq c\, l!\, M^l, \quad \forall l \in \mathbb{N}. \tag{1.2}$$

Note the boundedness assumption excludes the usual Gaussian noise while assumption (1.2) covers it. This assumption is well known in probability theory and was introduced in learning theory in [5, 23].

This paper will be arranged as follows. In Section 2 we discuss the technique proposed in [25]. In Section 3 we use it to study the convergence of empirical indefinite kernel integral operator with dependent sampling. In Section 4 we prove a capacity independent error bound for the indefinite kernel network algorithm. The consistency and convergence rates are stated in Section 5 as corollaries.

2 Structures of indefinite kernels

In this section we study the properties of indefinite kernels and the associated integral operators.

For a continuous kernel function $K(x,y)$ we denote the associated integral operator by L_K, defined as

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

If K is bounded, then L_K is a compact linear operator on $L^2_{\rho_X}$.

Define two kernels

$$\tilde{K}(x,t) = \int_X K(x,u)K(t,u)d\rho_X(u),$$
$$\hat{K}(x,t) = \int_X K(v,x)K(v,t)d\rho_X(v).$$

It is obvious that both $\tilde{K}$ and $\hat{K}$ are Mercer kernels, i.e., symmetric and positive definite kernels. Hence their associated integral operators are symmetric and positive. Let $\varphi_l, l \in \mathbb{N}$ be the orthonormal eigenfunction sequence of integral operator $L_{\tilde{K}}$, associated with its positive eigenvalue $\sigma_l^2, l \in \mathbb{N}$. Suppose these eigenvalues are in a non-increasing order. Mercer Theorem states that

$$\tilde{K}(x,t) = \sum_{i=1}^{\infty} \sigma_l^2 \varphi_(x)\varphi_l(t),$$

where the series converges absolutely and uniformly on $X \times X$. This gives that

$$L_{\tilde{K}} = \sum_{l=1}^{\infty} \sigma_l^2 \varphi_l \otimes \varphi_l.$$

It is obvious to notice that

$$L_{\hat{K}} = L_K^* L_K, \text{ and } L_{\tilde{K}} = L_K L_K^*.$$

Moreover, by the polar decomposition of L_K (see [16]), there is a partial isometry operator U from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, such that

$$L_K = (L_K L_K^*)^{1/2} U^* = L_{\tilde{K}}^{1/2} U^*.$$

Denote $\psi_l = U\varphi_l, l \in \mathbb{N}$, which also forms an orthonormal system of $L^2_{\rho_X}(X)$. We have

$$L_K = \sum_{l=1}^{\infty} \sigma_l \varphi_l \otimes \psi_l$$

and

$$L_{\hat{K}} = L_K^* L_K = U L_{\tilde{K}}^{1/2} L_{\tilde{K}}^{1/2} U^* = \sum_{l=1}^{\infty} \sigma_l^2 \psi_l \otimes \psi_l.$$

The latter tells that σ_l^2 and ψ_l are the eigenvalues and eigenfunctions of $L_{\hat{K}}$. By Mercer Theorem we have

$$\hat{K}(x,t) = \sum_{l=1}^{\infty} \sigma_l^2 \psi_l(x)\psi_l(t),$$

where the series converges absolutely and uniformly on $X \times X$.

Next we introduce a kernel condition which was used in [25] and enables the analytic expression of K.

Kernel Condition. $\kappa_0^2 = \sup_{x\in X} \sum_{l=1}^{\infty} \sigma_l \varphi_l^2(x) < \infty, \quad \kappa_1^2 = \sup_{t\in X} \sum_{l=1}^{\infty} \sigma_l \psi_l^2(t) < \infty.$

Denote $\kappa = \max\{\kappa_0, \kappa_1\}$. Kernel Condition ensures that $\sum_{i=1}^{\infty} \sigma_i \varphi_i(x)\psi_i(t)$ converges to $K(x,t)$ absolutely and uniformly on $X \times X$. It encourages us to consider the following two Mercer kernels,

$$K_0(x,t) = \sum_{l=1}^{\infty} \sigma_l \varphi_l(x)\varphi_l(t),$$
$$K_1(x,t) = \sum_{l=1}^{\infty} \sigma_l \psi_l(x)\psi_l(t).$$

The corresponding kernel integral operators are

$$L_{K_0} = \sum_{l=1}^{\infty} \sigma_l \varphi_l \otimes \varphi_l = L_{\tilde{K}}^{\frac{1}{2}},$$
$$L_{K_1} = \sum_{l=1}^{\infty} \sigma_l \psi_l \otimes \psi_l = L_{\hat{K}}^{\frac{1}{2}}.$$

The associated RKHS $\mathscr{H}_{K_0}$ and $\mathscr{H}_{K_1}$ will be simply denoted as $\mathscr{H}_0$ and $\mathscr{H}_1$ in the following. They can be characterized by

$$\mathscr{H}_0 = \left\{ f = \sum_{l=1}^{\infty} f_l \phi_l : \ \sum_{i=1}^{\infty} \frac{f_l^2}{\sigma_l} < \infty \right\},$$
$$\mathscr{H}_1 = \left\{ f = \sum_{l=1}^{\infty} g_l \psi_l : \ \sum_{i=1}^{\infty} \frac{g_l^2}{\sigma_l} < \infty \right\}.$$

Lemma 2.1 below summarizes the properties that will be used in later sections.

Lemma 2.1. *Under the Kernel Condition, we have*

(i) $L_K = L_{K_0} U^* = U^* L_{K_1}$;

(ii) $K(\cdot, x) \in \mathscr{H}_0$ *and* $K(x, \cdot) \in \mathscr{H}_1$ *for any* $x \in X$;

(iii) U *is an isometry operator from* $\mathscr{H}_0$ *to* $\mathscr{H}_1$ *and* $UK(\cdot, x) = K_1(\cdot, x)$;
U^* *is an isometry operator from* $\mathscr{H}_1$ *to* $\mathscr{H}_0$ *and* $U^*K(x, \cdot) = K_0(\cdot, x)$;

(iv) *L_K is bounded from $L^2_{\rho_X}$ to $\mathscr{H}_0$ and from $\mathscr{H}_1$ to $\mathscr{H}_0$ with both operator norms bounded by κ^2.*

3 Empirical approximation of integral operators by dependent sampling

In this section we study the empirical approximation of the integral operators by mixing sequences. The following lemma from [3] will be used to measure the effects of the dependance between samples. For a random variable ξ with values in a Hilbert space $\mathcal{H}$ and $1 \le u \le +\infty$, denote the u-th moment as $\|\xi\|_u = (\mathbb{E}\|\xi\|^u_{\mathcal{H}})^{1/u}$ if $1 \le u < \infty$ and $\|\xi\|_\infty = \sup \|\xi\|_{\mathcal{H}}$.

Lemma 3.1. *Let ξ and η be random variables with values in a separable Hilbert space $\mathcal{H}$ measurable $\sigma-$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*

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

For the sampling points $\mathbf{x} = \{x_1, \ldots, x_m\}$, the sampling operator $S : \mathscr{H}_0(\mathscr{H}_1) \to \mathbb{R}^m$ maps function f to a vector $Sf = (f(x_1), \ldots, f(x_m))^\top$. Operators T and T_* are defined as, for any $\mathbf{c} = (c_1, \ldots, c_m) \in \mathbb{R}^m$,

$$\begin{aligned} T\mathbf{c} &= \frac{1}{m}\sum_{i=1}^m c_i K(\cdot, x_i), \\ T_*\mathbf{c} &= \frac{1}{m}\sum_{i=1}^m c_i K(x_i, \cdot). \end{aligned}$$

By Lemma 2.1 (ii), T and T_* are operators from $\mathbb{R}^m$ to $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. It is proved in [25] that

$$\|S\| \le \kappa\sqrt{m}, \text{ and } \|T\| \le \frac{\kappa}{\sqrt{m}}, \quad \|T_*\| \le \frac{\kappa}{\sqrt{m}}.$$

In the sequel we simply denote by $\|\cdot\|_{01}$ and $\|\cdot\|_{10}$ the norms of operators from $\mathscr{H}_0$ to $\mathscr{H}_1$ and from $\mathscr{H}_1$ to $\mathscr{H}_0$ respectively.

Proposition 3.2. *Suppose the set of random sequence $z_i = (x_i, y_i)$, $1 \le i \le m$ satisfies the α-mixing condition. Then*

$$\mathbb{E}\|T_*S - L_K^*\|_{01}^2 \le \frac{\kappa^4}{m}\left(1 + 30\sum_{l=1}^{m-1}\alpha_l\right),$$
$$\mathbb{E}\|TS - L_K\|_{10}^2 \le \frac{\kappa^4}{m}\left(1 + 30\sum_{l=1}^{m-1}\alpha_l\right).$$

Proof. The proof of these two inequalities are similar, we only prove the first one. For any $f \in \mathscr{H}_0$, the reproducing property of $\mathscr{H}_0$, $f(x_i) = \langle K_0(\cdot, x_i), f\rangle_0$, allows us to write

$$T_*Sf = \frac{1}{m}\sum_{i=1}^{m} f(x_i)K(x_i, \cdot) = \left[\frac{1}{m}\sum_{i=1}^{m} K(x_i, \cdot)\otimes K_0(\cdot, x_i)\right] f,$$

implying that

$$T_*S = \frac{1}{m}\sum_{i=1}^{m} K(x_i, \cdot)\otimes K_0(\cdot, x_i).$$

By Lemma 2.1 (iii),

$$\begin{aligned}\mathbb{E}\|T_*S - L_K^*\|_{01}^2 &= \mathbb{E}\|U^*T_*S - U^*L_K^*\|_{00}^2 \\ &= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} K_0(\cdot, x_i)\otimes K_0(\cdot, x_i) - L_{K_0}\right\|_{00}^2 \\ &\le \frac{\kappa^4}{m}\left(1 + 30\sum_{l=1}^{m-1}\alpha_l\right).\end{aligned}$$

The last inequality follows from [20, Lemma 5.1]. □

The invertibility of $\lambda I + TST_*S$ is proved for any $\lambda > 0$ in [25].

Proposition 3.3. *Suppose the set of random sequence $z_i = (x_i, y_i)$, $1 \le i \le m$ satisfies a strongly mixing condition. For any $0 < \eta < 1$, with confidence $1 - \eta/2$, the following inequalities hold:*

$$\|T_*S - L_K^*\|_{01} \le \frac{2\kappa^2}{\sqrt{m\eta}}\sqrt{1 + 30\sum_{l=1}^{m-1}\alpha_l} \tag{3.1}$$

$$\|TS - L_K\|_{10} \le \frac{2\kappa^2}{\sqrt{m\eta}}\sqrt{1 + 30\sum_{l=1}^{m-1}\alpha_l}; \tag{3.2}$$

$$\|TST_*S - L_{\widetilde{K}}\|_{00} \leq \frac{4\kappa^4}{\sqrt{m\eta}}\sqrt{1 + 30\sum_{l=1}^{m-1}\alpha_l}; \tag{3.3}$$

$$\|T_*STS - L_{\widehat{K}}\|_{11} \leq \frac{4\kappa^4}{\sqrt{m\eta}}\sqrt{1 + 30\sum_{l=1}^{m-1}\alpha_l}. \tag{3.4}$$

Moreover when λ, m satisfy that

$$8\kappa^4\sqrt{1 + 30\sum_{l=1}^{m-1}\alpha_l} \leq \lambda\sqrt{m\eta}, \tag{3.5}$$

there hold

$$\left\|(\lambda I + TST_*S)^{-1}\right\| \leq \frac{2}{\lambda} \quad \textit{and} \quad \left\|(\lambda I + T_*STS)^{-1}\right\| \leq \frac{2}{\lambda}. \tag{3.6}$$

Proof. By Proposition 3.2 and Markov inequality, inequalities (3.1) and (3.2) hold with confidence $1 - \eta/4$ respectively. Thus with confidence $1 - \eta/2$, these two inequalities hold simultaneously. The inequality (3.3) holds by

$$\|TST_*S - L_{\widetilde{K}}\| \leq \|(TS - L_K)T_*S\| + \|L_K(T_*S - L_K^*)\| \leq \kappa^2\|TS - L_K\| + \kappa^2\|T_*S - L_K^*\|.$$

Similarly we can prove (3.4).

Now we turn to proof of (3.6). Using the invertibility of $\lambda I + TST_*S$, we write

$$\begin{aligned}(\lambda I + TST_*S)^{-1} &= (\lambda I + L_{\widetilde{K}} + TST_*S - L_{\widetilde{K}})^{-1} \\ &= (\lambda I + L_{\widetilde{K}})^{-1}\left\{I + (TST_*S - L_{\widetilde{K}})(\lambda I + L_{\hat{K}})^{-1}\right\}^{-1}.\end{aligned}$$

Inequality (3.3) and condition (3.5) show that, with confidence $1 - \eta/2$,

$$\|(TST_*S - L_{\widetilde{K}})(\lambda I + L_{\hat{K}})^{-1}\| \leq \frac{1}{2}.$$

This ensures the first conclusion in (3.6).

By symmetry, it is straightforward to conclude the invertibility of $\lambda I + T_*STS$ on $\mathscr{H}_1$ and the second inequality in (3.6). □

4 Error analysis

In this section we will prove an error bound for $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}$ which can be used to measure the goodness of the approximation of f_ρ by $f_{\mathbf{z}}$. It is proved in [21, 25] that

$$f_{\mathbf{z}} = T\Big(\lambda I + ST_*ST\Big)^{-1} ST_*\mathbf{y} = \Big(\lambda I + TST_*S\Big)^{-1} TST_*\mathbf{y}.$$

Note that $\mathbb{E}T_*\mathbf{y} = L_K^* f_\rho$, $TS \to L_K$ and $TST_*S \to L_{\widetilde{K}}$ in probability, we introduce the regularizing function

$$f_\lambda = (\lambda I + L_{\widetilde{K}})^{-1} L_{\widetilde{K}} f_\rho$$

as a bridge and decompose the error into sample error and approximation error as follows:

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} \le \|f_{\mathbf{z}} - f_\lambda\|_{L^2_{\rho_X}} + \|f_\lambda - f_\rho\|_{L^2_{\rho_X}}. \tag{4.1}$$

In the next two subsections we will analyze the two types of errors.

4.1 Estimate of approximation error

By no free lunch principle, we have to make some assumptions on the target function. This is the usual way for analysis of learning algorithms. These conditions are usually called prior condition. In this paper we adopt a commonly used prior condition.

Prior Condition. $L_{K_0}^{-\beta} f_\rho \in L^2_{\rho_X}(X)$ for some $\beta > 0$.

This prior condition means that f_ρ belongs to the range of the operator $L_{K_0}^{\beta}$ on $L^2_{\rho_X}$. Note that if $\beta = \frac{1}{2}$ this range is exactly $\mathscr{H}_0$ and if $\beta < \frac{1}{2}$ this range characterizes an interpolation space between $\mathscr{H}_0$ and $L^2_{\rho_X}$ while if $\beta > \frac{1}{2}$ it characterizes a subspace of $\mathscr{H}_0$. By Lemma 2.1 (ii) the empirical approximation $f_{\mathbf{z}}$ is in $\mathscr{H}_0$. It is appropriate to use this space to characterizes the ability of the approximation of f_ρ by $f_{\mathbf{z}}$. The following results are proved in [25, Theorem 5.1 and Lemma 5.3].

Theorem 4.1. *Under the Prior Condition and the Kernel Condition there are constants C_1 and C_2 such that*

$$\|f_\lambda - f_\rho\|_{L^2_{\rho_X}} \le C_1 \lambda^{\min\{\frac{\beta}{2}, 1\}},\ \|L_K^*(f_\lambda - f_\rho)\|_{\mathscr{H}_1} \le C_2 \lambda^{\min\{\frac{1}{4}+\frac{\beta}{2}, 1\}}.$$

4.2 Estimate of sample error

To estimate the sample error, we adopt the following decomposition used [25]:

$$\begin{aligned}
&f_{\mathbf{z}} - f_\lambda = \big(\lambda I + TST_*S\big)^{-1}\big[TS(T_*\mathbf{y} - T_*Sf_\lambda) - L_{\widetilde{K}}(f_\rho - f_\lambda)\big]\\
=&\big(\lambda I + TST_*S\big)^{-1}\Big[TS\Big(T_*\mathbf{y} - T_*Sf_\lambda - L_K^*(f_\rho - f_\lambda)\Big) + (TS - L_K)L_K^*(f_\rho - f_\lambda)\Big]\\
=&\big(\lambda I + TST_*S\big)^{-1}TS\Delta + \big(\lambda I + TST_*S\big)^{-1}(TS - L_K)L_K^*(f_\rho - f_\lambda),
\end{aligned}$$

where

$$\Delta = \frac{1}{m}\sum_{i=1}^m (y_i - f_\lambda(x_i))K(x_i, \cdot) - L_K^*(f_\rho - f_\lambda).$$

Note that both $f_{\mathbf{z}}$ and f_λ are in $\mathscr{H}_0$. By the fact that $L_{K_0}^{\frac{1}{2}}$ is an isometric mapping between $L^2_{\rho_X}$ and $\mathscr{H}_0$, we have

$$\begin{aligned}
\|f_{\mathbf{z}} - f_\lambda\|_{L^2_{\rho_X}} = \|L_{K_0}^{\frac{1}{2}}(f_{\mathbf{z}} - f_\lambda)\|_{\mathscr{H}_0} \le& \|L_{K_0}^{\frac{1}{2}}\big(\lambda I + TST_*S\big)^{-1}TS\|_{10}\|\Delta\|_{\mathscr{H}_1}\\
&+ \|L_{K_0}^{\frac{1}{2}}\big(\lambda I + TST_*S\big)^{-1}\|_{00}\|TS - L_K\|_{10}\|L_K^*(f_\rho - f_\lambda)\|_{\mathscr{H}_1}.
\end{aligned}$$

We next estimate the right hand side term by term.

It is easy to check that $\big(\lambda I + TST_*S\big)^{-1}TS = TS\big(\lambda I + T_*STS\big)^{-1}$. So we can write

$$\begin{aligned}
&L_{K_0}^{\frac{1}{2}}\big(\lambda I + TST_*S\big)^{-1}TS = L_{K_0}^{\frac{1}{2}}TS\big(\lambda I + T_*STS\big)^{-1}\\
=&L_{K_0}^{\frac{1}{2}}(TS - L_K)\big(\lambda I + T_*STS\big)^{-1} + L_{K_0}^{\frac{1}{2}}L_K(\lambda I + T_*STS)\big)^{-1}\\
=&L_{K_0}^{1/2}(TS - L_K)\big(\lambda I + T_*STS\big)^{-1} + L_{K_0}^{\frac{1}{2}}L_K(\lambda I + L_{\hat{K}}\big)^{-1}\\
&+ L_{K_0}^{\frac{1}{2}}L_K(\lambda I + L_{\hat{K}}\big)^{-1}(L_{\hat{K}} - T_*STS)(\lambda I + T_*STS)^{-1}
\end{aligned}$$

For each $g = \sum_{l=1}^\infty g_l\psi_l \in \mathscr{H}_1$,

$$L_{K_0}^{\frac{1}{2}}L_K(\lambda I + L_{\hat{K}}\big)^{-1}g = \sum_{l=1}^\infty \frac{\sigma_l^{3/2}}{\lambda + \sigma_l^2}g_l\phi_l \in \mathscr{H}_0.$$

This gives $\|L_{K_0}^{\frac{1}{2}}L_K(\lambda I + L_{\hat{K}}\big)^{-1}\|_{10} \le \lambda^{-\frac{1}{4}}$. Thus $\|L_{K_0}^{\frac{1}{2}}\big(\lambda I + TST_*S\big)^{-1}TS\|_{10}$ is bounded by

$$\le \lambda^{-\frac{1}{4}} + \Big[\lambda^{-\frac{1}{4}}\|L_{\hat{K}} - T_*STS\|_{11} + \kappa\|L_K - TS\|_{10}\Big]\Big\|\big(\lambda I + TST_*S\big)^{-1}\Big\|_{11}. \qquad (4.2)$$

Similarly we can prove that

$$\|L_{K_0}^{1/2}\big(\lambda I + TST_*S\big)^{-1}\|_{00} \le \lambda^{-3/4} + \lambda^{-3/4}\|L_{\widetilde{K}} - TST_*S\|_{00}\|(\lambda I + TST_*S\big)^{-1}\|_{00}. \quad (4.3)$$

For $\|\Delta\|_{\mathscr{H}_1}$. Denote $\xi(z) = (y - f_\lambda(x))K(x,\cdot)$ to be a $\mathscr{H}_1$ valued random variable. Then $\mathbb{E}\xi = L_K^*(f_\rho - f_\lambda)$ and

$$\Delta = \frac{1}{m}\sum_{i=1}^m \xi(z_i) - L_K(f_\rho - f_\lambda) = \frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi.$$

By Lemma 3.1 and the fact $\|K(x_i,\cdot)\|_{\mathscr{H}_1} \le \kappa$, for any $\delta > 0$

$$\begin{aligned}
\mathbb{E}\|\Delta\|_{\mathscr{H}_1}^2 &= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\|\xi(z_i)\|_{\mathscr{H}_1}^2 + \frac{2}{m^2}\sum_{i<j}\mathbb{E}\langle\xi(z_i),\xi(z_j)\rangle_{\mathscr{H}_1} - \|\mathbb{E}\xi\|_{\mathscr{H}_1}^2 \\
&\le \frac{1}{m}\mathbb{E}\|\xi\|_{\mathscr{H}_1}^2 + \frac{30}{m^2}\sum_{i<j}\alpha_{j-i}^{\frac{\delta}{2+\delta}}\big(\mathbb{E}\|\xi\|_{\mathscr{H}_1}^{2+\delta}\big)^{\frac{2}{2+\delta}} \\
&\le \frac{\kappa^2}{m}\mathbb{E}(y - f_\lambda(x))^2 + \frac{30\kappa^2}{m}\left(\sum_{l=1}^{m-1}\alpha_l^{\frac{\delta}{2+\delta}}\right)\big(\mathbb{E}|y - f_\lambda(x)|^{2+\delta}\big)^{\frac{2}{2+\delta}}. \qquad (4.4)
\end{aligned}$$

Since $\mathbb{E}y^2 \le 2cM^2$ and $\|f_\lambda\|_{L^2_{\rho_X}}^2 \le \|f_\rho\|_{L^2_{\rho_X}}^2 \le \mathbb{E}y^2$ we have

$$\mathbb{E}(y - f_\lambda(x))^2 \le 2(\mathbb{E}y^2 + \|f_\lambda\|_{L^2_{\rho_X}}^2) \le 8cM^2.$$

Let $\lceil\delta\rceil$ denote the minimal integer that is equal to or larger than δ. By (1.2)

$$\big(\mathbb{E}|y|^{2+\delta}\big)^{\frac{2}{2+\delta}} \le \big(\mathbb{E}|y|^{2+\lceil\delta\rceil}\big)^{\frac{2}{2+\lceil\delta\rceil}} \le \big(c(2+\lceil\delta\rceil)!M^{2+\lceil\delta\rceil}\big)^{\frac{2}{2+\lceil\delta\rceil}} \le c(3+\delta)^2M^2.$$

For f_λ, we have

$$|f_\lambda(x)| \le \kappa\|f_\lambda\|_{\mathscr{H}_0} = \kappa\|(\lambda I + L_{K_0}^2)^{-1}L_{K_0}^{\frac{3}{2}+\beta}L_{K_0}^{-\beta}f_\rho\|_{L^2_{\rho_X}} \le C_3\kappa\lambda^{\min\{\frac{2\beta-1}{4},0\}}$$

where $C_3 = \|L_{K_0}^{-\min\{\beta,\frac{1}{2}\}}f_\rho\|_{L^2_{\rho_X}}$. Thus

$$(\mathbb{E}|f_\lambda(x)|^{2+\delta})^{\frac{2}{2+\delta}} \le \left(\|f_\lambda\|_{L^2_{\rho_X}}^2 C_3^\delta\kappa^\delta\lambda^{\delta\min\{\frac{2\beta-1}{4},0\}}\right)^{\frac{2}{2+\delta}} \le C_4\lambda^{\min\{\frac{\delta(2\beta-1)}{2(2+\delta)},0\}}.$$

with $C_4 = (2cM^2C_3^\delta\kappa^\delta)^{\frac{2}{2+\delta}} \le 2cM^2(C_3^2\kappa^2+1)$. Then we get

$$\begin{aligned}
\big(\mathbb{E}|y - f_\lambda(x)|^{2+\delta}\big)^{\frac{2}{2+\delta}} &\le 2\left(\big(\mathbb{E}|y|^{2+\delta}\big)^{\frac{2}{2+\delta}} + \big(\mathbb{E}|f_\lambda(x)|^{2+\delta}\big)^{\frac{2}{2+\delta}}\right) \\
&\le C_5\left(1 + \delta^2 + \lambda^{\min\{\frac{\delta(2\beta-1}{2(2+\delta)},0\}}\right)
\end{aligned}$$

where $C_5 = \max\{20cM^2, 2C_4\}$. We finally obtain

$$\mathbb{E}\|\Delta\|^2_{\mathscr{H}_1} \leq C_6 m^{-1} \left[1 + \left(1 + \delta^2 + \lambda^{\min\{\frac{\delta(2\beta-1)}{2(2+\delta)}, 0\}}\right) \left(\sum_{l=1}^{m-1} \alpha_l^{\frac{\delta}{2+\delta}}\right)\right] \tag{4.5}$$

Plugging (4.2), (4.3), (4.5) into (4.1) and applying Theorem 4.1, Proposition 3.2, we can obtain the sample error bound. The regularization parameter λ depends on the sample size m, $\lambda = \lambda(m)$, and tends to zero as $m \to +\infty$. Thus, we assume $0 < \lambda \leq 1$ in the sequel.

Theorem 4.2. *Suppose that $0 < \lambda \leq 1$. For any $\delta > 0$ and $0 < \eta < 1$, with confidence $1 - \eta$, the sample error $\|f_{\mathbf{z}} - f_\lambda\|_{L^2_{\rho_X}}$ is bounded by*

$$C_7 \left\{ \lambda^{-\frac{1}{4}} (m\eta)^{-\frac{1}{2}} \left[1 + \left(1 + \delta + \lambda^{\min\{\frac{\delta(2\beta-1)}{4(2+\delta)}, 0\}}\right) \sqrt{\sum_{l=1}^{m-1} \alpha_l^{\frac{\delta}{2+\delta}}}\right] + \lambda^{\min\{\frac{\beta+1}{2}, \frac{5}{4}\}} \right\}$$

provided that

$$8\kappa^4 \sqrt{1 + 30 \sum_{l=1}^{m-1} \alpha_l} \leq \lambda \sqrt{m\eta}. \tag{4.6}$$

The constant C_7 is given as $\frac{1}{2\kappa^2} \max\{(4\kappa^2 + \kappa)\sqrt{2C_6}, C_2\}$.

5 Learning rates

Combining the approximation error bound in Theorem 4.1 and the sample error bound in Theorem 4.2, we can deduce the learning rates when the mixing coefficients decays in certain rate. For this purpose, we need the following simple facts, for $m > 3$,

$$\sum_{\ell=1}^{m-1} \ell^{-t} \leq \begin{cases} \frac{1}{1-t} m^{1-t} & \text{if } 0 < t < 1; \\ 2 \log m & \text{if } t = 1; \\ \frac{t}{t-1} & \text{if } t > 1. \end{cases} \tag{5.1}$$

We have the following learning rate estimates.

Theorem 5.1. *Suppose that the α-mixing coefficients satisfy a polynomial decay, i.e., $\alpha_i \leq a i^{-t}$ for some $a > 0$ and $t > 0$. Then by choosing $\lambda = \lambda(m)$ appropriately, for m large enough we have with confidence $1 - \eta$,*

(1) *if* $0 < t \le 1$,

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = \begin{cases} O\left(m^{-\frac{t\beta}{4}}(\log m)^{\frac{\beta}{2}}\right) & \text{when } 0 < \beta < \frac{3}{2} \\ O\left(m^{-\frac{t\beta}{2\beta+1}}(\log m)^2\right) & \text{when } \frac{3}{2} \le \beta \le 2 \\ O\left(m^{-\frac{2t}{5}}(\log m)^2\right) & \text{when } \beta > 2. \end{cases}$$

(2) *if* $t > 1$,

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = \begin{cases} O(m^{-\frac{\beta}{4}}(\log m)^{\frac{\beta}{2}}) & \text{when } 0 < \beta < \frac{3}{2} \\ O(m^{-\frac{\beta}{2\beta+1}}) & \text{when } \frac{3}{2} \le \beta \le 2; \\ O(m^{-\frac{2}{5}}) & \text{when } \beta > 2. \end{cases}$$

Proof.

Case (1): $0 < t \le 1$**.** For $0 < t < 1$, by (5.1) and $\alpha_i \le ai^{-t}$,

$$\sqrt{\sum_{l=1}^{m-1} \alpha_l^{\frac{\delta}{2+\delta}}} \le \frac{1}{\sqrt{1-t}} a^{\frac{\delta}{4+2\delta}} m^{\frac{2+\delta-\delta t}{4+2\delta}},$$

$$\sqrt{1 + 30\sum_{l=1}^{m-1} \alpha_l} \le \sqrt{\frac{1+30a}{1-t}} m^{\frac{1-t}{2}}.$$

By Theorem 4.1 and Theorem 4.2, with confidence $1-\eta$, we have

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = O\left(\lambda^{\min\{\frac{\beta}{2},1\}} + (2+\delta)\eta^{-\frac{1}{2}}\lambda^{\min\{\frac{\delta(2\beta-1)}{4(2+\delta)},0\}-\frac{1}{4}} m^{-\frac{\delta t}{4+2\delta}}\right).$$

When $0 < \beta < \frac{3}{2}$, the desired learning rate can be achieved by choosing $\lambda = m^{-\frac{t}{2}}\log m$ and $\delta > 0$ such that $\min\{\frac{\delta(2\beta-1)}{4(2+\delta)}, 0\} - \frac{1}{4} + \frac{\delta}{2+\delta} = \min\{\frac{\beta}{2}, 1\}$. Under these choices the condition (4.6) is easy to check since with m big enough

$$\frac{8\kappa^4(1+30a)^{\frac{1}{2}}}{\sqrt{\eta(1-t)}} \le \log m.$$

When $\beta \ge \frac{3}{2}$, we choose $\delta = \log m$ which gives $\frac{\delta}{2+\delta} = 1 - \frac{2}{2+\log m}$ and $m^{\frac{\delta}{2+\delta}} \ge \frac{m}{e^2}$. The desired convergence rate is obtained by taking $\lambda = m^{-r}\log m$ with $r = \frac{2t}{2\min\{\beta,2\}+1}$. Since $r \le \frac{t}{2}$, the condition (4.6) also holds when m is big enough.

If $t = 1$,

$$\sqrt{1 + 30\sum_{l=1}^{m-1} \alpha_l} \le \sqrt{2(1+30a)}\sqrt{\log m}.$$

We see the above choices for λ still ensure the condition (4.6) for m large enough. Hence the rates can be verified too.

Case (2): $t > 1$. When $0 < \beta \leq \frac{3}{2}$, let $\lambda = m^{-\frac{1}{2}} \log m$. Then, when m is big enough such that

$$\frac{8\kappa^4(1+30a)^{\frac{1}{2}}\sqrt{t}}{\sqrt{\eta(t-1)}} \leq \log m,$$

condition (4.6) holds. Choosing δ satisfying $\frac{\delta}{2+\delta} = \frac{(1+2\beta)}{4t+\min\{2\beta-1,0\}}$ yields $\frac{\delta t}{2+\delta} < 1$ and

$$\sqrt{\sum_{l=1}^{m-1} \alpha_l^{\frac{\delta}{2+\delta}}} \leq \sqrt{\frac{2+\delta}{2+\delta-\delta t}} a^{\frac{\delta}{4+2\delta}} m^{\frac{2+\delta-\delta t}{4+2\delta}}.$$

Thus with confidence $1-\eta$, we have

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = O\left(\lambda^{\frac{\beta}{2}} + \eta^{-\frac{1}{2}} \lambda^{\min\{\frac{\delta(2\beta-1)}{4(2+\delta)},0\}-\frac{1}{4}} m^{-\frac{\delta t}{4+2\delta}}\right) = O\left(m^{-\frac{\beta}{4}} (\log m)^{\frac{\beta}{2}}\right).$$

When $\beta > \frac{3}{2}$, choose $\delta > \frac{2}{t-1}$ and fix it. With confidence $1-\eta$,

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = O\left(\lambda^{\min\{\frac{\beta}{2},1\}} + \eta^{-\frac{1}{2}} \lambda^{-\frac{1}{4}} m^{-\frac{1}{2}}\right).$$

Take $\lambda = m^{-\frac{2}{1+2\min\{\beta,2\}}}$. When

$$\frac{8\kappa^4\sqrt{t(1+30a)}}{\sqrt{\eta(1-t)}} \leq m^{\frac{\min\{\beta,2\}-\frac{3}{2}}{1+2\min\{\beta,2\}}},$$

condition (4.6) holds. The desired learning rate can be easily verified. □

Remark 5.2. *Indefinite kernel network* (1.1) *with independent sampling was studied in [21,25]. The learning rates deudced in [21, Corollary 2.4] is* $m^{-\frac{\beta}{2\beta+6}}$ *for* $0 < \beta \leq 2$, $m^{-\frac{1}{5}}$ *for* $\beta \geq 2$. *This conclusion was improved by [25, Theorem 5.6] to* $m^{-\frac{\beta}{4}}$ *for* $0 < \beta \leq \frac{1}{2}$, $m^{-\frac{\beta}{2\beta+3}}$ *for* $\frac{1}{2} < \beta \leq 2$, *and* $m^{-\frac{2}{7}}$ *for* $\beta \geq 2$. *The rate analysis for independent sampling corresponds to the case* $t = \infty$ *for which the rates in Theorem 5.1 is clearly faster. Therefore, our new rate analysis not only provides rates for dependent sampling, but also improves the existing rate analysis in [21, 25] for independent sampling.*

Acknowledgement

This work is supported by the Nature Science Fund of China (No. 11071276 and No. 11101403).

References

[1] K. Athreya and S. Pantula. Mixing properties of harris chains and autoregressive processes. *Journal of Applied Probability*, 23:880–892, 1986.

[2] K. Athreya and S.G.Pantula. A note on strong mixing of arma processes. *Journal of Applied Probability*, 11:401–408, 1974.

[3] P. Billingsley. *Convergence of Probability Measures*. New York: Wiley, 1968.

[4] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[5] A. Caponnetto and E. D. Vito. Optimal rates for regularized least-squares algorithms. *Found. Comput. Math*, 7:331–368, 2007.

[6] K. Chanda. Strong mixing properties of linear stochastic processes. *Journal of Applied Probability*, 11:401–408, 1974.

[7] C.S.Ong, X.Mary, S.Canu, and A. J. Smola. Learning with non-positive kernels. *Proceedings of the 21st International Coference on Machine Learning*, 2004.

[8] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Soc.*, 39:1–49, 2001.

[9] F. Cucker and D.-X. Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

[10] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13:1–50, 2000.

[11] C. Liu. Gabor-based kernel pca with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:572–581, 2004.

[12] R. Luss and A. d'Aspremont. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1:97–118, 2009.

[13] D. S. Modha. Minimum complexity regression estimation with weakly dependent observations. *IEEE. Transaction Information Theory*, 42:2133–2145, 1996.

[14] E. Pekalska and B. Haasdonk. Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1017–1032, 2009.

[15] L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11:1927–1956, 2010.

[16] W. Rudin. *Functional Analysis*. McGraw-Hill, Inc., 1991.

[17] H. Saigo, J. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20:1682–1689, 2004.

[18] B. Scholkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[19] S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

[20] H. Sun and Q. Wu. Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, 32:175–189, 2010.

[21] H. Sun and Q. Wu. Least square regression with indefinite kernels and coefficient regularization. *Applied and Computational Harmonic Analysis*, 30:96–109, 2011.

[22] V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.

[23] C. Wang and D. X. Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27:55–67, 2011.

[24] Q. Wu. *Classification and Regularization in Learning Theory*. VDM Verlag, 2009.

[25] Q. Wu. Regularization networks with indefinite kernels. 2011. preprint.

[26] Q. W. Xiao and D. X. Zhou. Learning by nonsymmetric kernels with data dependent spaces and ℓ_1-regularizer. 2008. preprint.

[27] Y. L. Xu and D. R. Chen. Learning rates of regularized regression for exponentially strongly mixing sequence. *Journal of Statistical Planning and Inference*, 138:2180–2189, 2008.

[28] Y.Ying, C.Campbelly, and M. Girolami. Analysis of svm with indefinite kernels. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[29] T. Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15:1397–1437, 2003.

[30] B. Zou, L. Li, and Z.B.Xu. The generalization performance of erm algorithm with strongly mixing observations. *Machine Learning*, 75:275–295, 2009.