Commun. Math. Sci.
Vol. 16, No. 5, pp. 1347–1360

REGULARIZED SEMI-SUPERVISED LEAST SQUARES REGRESSION WITH DEPENDENT SAMPLES[*]

HONGZHI TONG[†] AND MICHAEL NG[‡]

Abstract. In this paper, we study regularized semi-supervised least squares regression with dependent samples. We analyze the regularized algorithm based on reproducing kernel Hilbert spaces, and show, with the use of unlabelled data that the regularized least squares algorithm can achieve the nearly minimax optimal learning rate with a logarithmic term for dependent samples. Our new results are better than existing results in the literature.

Keywords. semi-supervised learning; regularization; least squares regression; non-iid sampling.

AMS subject classifications. 68T05; 62J02.

1. Introduction

In this paper, we consider the regularized semi-supervised least squares regression with non-iid sampling. Let X be a compact subset of $\mathbb{R}^n$, $Y\subset[-M,M]$ for some $M>0$. The relation between the input $x\in X$ and output $y\in Y$ is described by a (unknown) probability distribution ρ on $X\times Y$. Let ρ_X be the marginal distribution of ρ on X and $\rho(\cdot|x)$ be the conditional probability distribution at $x\in X$. Then the regression function $f_\rho: X\mapsto Y$ is defined by

$$f_\rho(x):=E[y|x]=\int_Y y d\rho(y|x),\quad x\in X.$$

With a set of labelled samples $D=\{z_i\}_{i=1}^m=\{(x_i,y_i)\}_{i=1}^m$ drawn according to ρ, the regularized least squares regression scheme is stated as follows:

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

where $\lambda>0$ is a regularization parameter, and $(\mathcal{H}_K,\|\cdot\|_K)$ is a reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel K.

The analysis of regularized least square algorithm has received attention in statistics and learning theory. The main task is to estimate $\|f_{D,\lambda}-f_\rho\|_{L^2_{\rho_X}}$. Note the fact

$$\|f_{D,\lambda}-f_\rho\|_{L^2_{\rho_X}}\le\|f_{D,\lambda}-f_\lambda\|_{L^2_{\rho_X}}+\|f_\lambda-f_\rho\|_{L^2_{\rho_X}}\tag{1.2}$$

where f_λ is a regularization function defined by

$$f_\lambda:=\arg\min_{f\in\mathcal{H}_K}\left\{\int_{X\times Y}(y-f(x))^2d\rho+\lambda\|f\|_K^2\right\}.\tag{1.3}$$

The first term on the right-hand side of (1.2) is sample error and the second term is approximation error. Let L_K be the integral operator on $\mathcal{H}_K$ (or $L^2_{\rho_X}$) defined by

$$L_Kg:=\int_X K_xg(x)d\rho_X,$$

[*] Received: December 25, 2017; Accepted (in revised form): April 28, 2018. Communicated by Lexing Ying.

[†] School of Statistics, University of International Business and Economics, Beijing 100029, P.R. China (tonghz@uibe.edu.cn).

[‡] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (mng@math.hkbu.edu.hk).

where K_x is the function $K(\cdot,x)$ in $\mathcal{H}_K$. To estimate the approximation error, a regularity condition on the regression function f_ρ is required:

$$f_\rho = L_K^r(g_\rho) \text{ for some } g_\rho \in L^2_{\rho_X} \text{ with } r > 0. \tag{1.4}$$

For independent samples with respect to ρ, it is proved in [3] that, if (1.4) holds with some $r \geq \frac{1}{2}$ (which implies $f_\rho \in \mathcal{H}_K$) and the eigenvalues of L_K satisfy $\lambda_i \sim i^{-2\alpha}$ for some $\alpha > \frac{1}{2}$, then the minimax optimal learning rate of regularized least squares algorithm is $\mathcal{O}(m^{\frac{-2\alpha}{4\alpha r+1}})$.

In practice, dependent samples often occur. In [13], 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 contained is equivalent to that contained by only $\hat{m} < m$ independent samples. This implies that for an algorithm, if we have the rate of $\mathcal{O}(m^{-\gamma})$ for independent samples with $\gamma > 0$, then the rate for the mixing sequences is $\mathcal{O}(\hat{m}^{-\gamma})$. For example, when an exponentially strongly mixing sequence is considered, it is shown that the effective number of observations is $\hat{m} = \mathcal{O}(m^{\frac{t}{t+1}})$, where t is given in (2.1), see Dehling and Philipp [6]. Based on this effective number of observations, some capacity-dependent learning rates were derived in the literature, see [19]. On the other hand, capacity-independent learning rates were studied [16, 17]. In [16], if the approximation error

$$\|f_\lambda - f_\rho\|_{L^2_{\rho_X}} \leq \lambda^r \|g_\rho\|_{L^2_{\rho_X}} \tag{1.5}$$

holds for $r \in (0, \frac{1}{2})$, then the learning rate of the regularized least squares algorithm is $\mathcal{O}\left(m^{\frac{-2r}{3+2r}}(\log m)^{\frac{1}{2t}}\right)$. It is further improved to $\mathcal{O}\left(m^{\frac{-3r}{4(1+r)}}(\log m)^{\frac{1}{2t}}\right)$ in [17]. There is still a gap between the above learning rates and the optimal rate under independent samples. The main contribution of this paper is to make use of unlabelled data to study regularized semi-supervised least squares regression with dependent samples. The main advantage of using unlabelled data is that the learning rate can be enhanced. It should be pointed out that using unlabelled data to improve the learning rate of least squares regression with the i.i.d sampling has been studied in the literature, see, e.g. [1,4,12]. In practice, the cost of labelling can be expensive, but the cost of collecting data samples is quite cheap [9,10]. By using the analysis on reproducing kernel Hilbert spaces, we show that the regularized least squares algorithm can achieve the nearly minimax optimal learning rate $\mathcal{O}\left(m^{-\frac{r}{2r+\beta}}(\log m)^{\frac{3}{2t}}\right)$ where $\beta \leq 1$ and $r \in (0, \frac{1}{2})$. It is obvious that our result are better than those in [16, 17].

The outline of this paper is given as follows. In Section 2, we present preliminaries about dependent samples and probabilistic results. In Section 3, we present the framework of the regularized semi-supervised least squares algorithm, and establish the new results of learning rate when unlabelled data is used in the framework. Some concluding remarks are given in Section 4.

2. Preliminaries

Let $Z = \{z_i\}_{i \geq 1}$ be a stationary random sequence with unknown distribution ρ. It implies that z_i $(i \geq 1)$ have the same distribution ρ. The σ-algebras generated by $(z_a, z_{a+1}, \cdots, z_b)$ is denoted by $\mathcal{A}_a^b$. In order to employ dependent samples, we need to figure out their correlation. To estimate the correlation between the σ-algebras $\mathcal{A}_1^i$ and $\mathcal{A}_{i+n}^\infty$, various mixing coefficients have been proposed and used in the literature [8,14,18]:

$$\alpha(Z,n) := \sup_{A \in \mathcal{A}_1^i, B \in \mathcal{A}_{i+n}^\infty} |\rho(A \cap B) - \rho(A)\rho(B)|,$$

$$\beta(Z,n) := E \sup_{B\in\mathcal{A}_{i+n}^{\infty}} |\rho(B) - \rho(B|\mathcal{A}_1^i)|,$$

$$\phi(Z,n) := \sup_{A\in\mathcal{A}_1^i, B\in\mathcal{A}_{i+n}^{\infty}} |\rho(B) - \rho(B|A)|.$$

It is well-known that these coefficients satisfy

$$2\alpha(Z,n) \le \beta(Z,n) \le \phi(Z,n),$$

see [2]. We remark that small coefficients refer to weak dependence among samples. Therefore, the α-mixing coefficients are the weakest notion for describing the dependence structure. In this paper, we focus only on α-mixing processes.

Definition 2.1. *A stochastic process $Z = \{z_i\}_{i\ge 1}$ is called α-mixing if there holds*

$$\lim_{n\to\infty} \alpha(Z,n) = 0.$$

Moreover, a stochastic process Z is called geometrically α-mixing, if

$$\alpha(Z,n) \le ce^{-bn^t}, \ n \ge 1. \tag{2.1}$$

for some constants $b > 0$, $c \ge 0$ and $t > 0$.

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 $\|\xi\|_u = (E\|\xi\|_{\mathcal{H}}^u)^{1/u}$ if $1 \le u < +\infty$ and $\|\xi\|_\infty = \sup\|\xi\|_{\mathcal{H}}$. The following lemma was proved by Dehling and Philipp [6].

Lemma 2.1. *Let ξ and χ be random variables with values in a separable Hilbert space $\mathcal{H}$ measurable σ-algebras $\mathcal{J}$ and $\mathcal{K}$ 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*

$$|E < \xi, \chi > - < E\xi, E\chi > | \le 15\alpha^{1/t}(\mathcal{J},\mathcal{K})\|\xi\|_u\|\chi\|_v,$$

where $\alpha(\mathcal{J},\mathcal{K})$ is the α-coefficient of $\mathcal{J}$ and $\mathcal{K}$ defined as

$$\alpha(\mathcal{J},\mathcal{K}) := \sup_{A\in\mathcal{J}, B\in\mathcal{K}} |\rho(A\cap B) - \rho(A)\rho(B)|.$$

By using this lemma, we derive the following results.

Lemma 2.2. *Let $Z = \{z_i\}_{i\ge1}$ be a stationary α-mixing sequence that satisfies (2.1). For a random variable ξ on (Z,ρ) with values in a Hilbert space $\mathcal{H}$, we have, for any $0 < \delta \le \infty$*

$$E\left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)\right\|^2 \le \frac{1}{m}\|\xi\|_2^2 + \frac{30C_1}{m}\left(\frac{2+\delta}{\delta}\right)^{1/t}\|\xi\|_{2+\delta}^2,$$

where C_1 is a constant given by $c\int_0^\infty b^{-1/t}e^{-y^t}dy$.

Proof. We first note that

$$\left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)\right\|^2 = \left\|\frac{1}{m}\sum_{i=1}^m \xi(z_i)\right\|^2 - \frac{2}{m}\sum_{i=1}^m \langle \xi(z_i), E(\xi)\rangle + \|E(\xi)\|^2.$$

By taking expectations on both sides of the above equality, we have

$$\begin{aligned}E\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-E(\xi)\right\|^2&=E\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)\right\|^2-\frac{2}{m}\sum_{i=1}^{m}\langle E\xi(z_i),E(\xi)\rangle+\|E(\xi)\|^2\\&=E\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)\right\|^2-\|E(\xi)\|^2.\end{aligned}\tag{2.2}$$

By using Lemma 2.1 with $u=v=2+\delta, t=\frac{2+\delta}{\delta}$ with $\delta>0$, we have for $j<i$

$$\begin{aligned}E\langle\xi(z_i),\xi(z_j)\rangle&\leq\langle E\xi(z_i),E\xi(z_j)\rangle+15\alpha^{\frac{\delta}{2+\delta}}(Z,i-j)\|\xi(z_i)\|_{2+\delta}\|\xi(z_j)\|_{2+\delta}\\&=\|E\xi\|^2+15\alpha^{\frac{\delta}{2+\delta}}(Z,i-j)\|\xi\|_{2+\delta}^2.\end{aligned}\tag{2.3}$$

Here we compute the term in (2.2) as follows:

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)\right\|^2&=\frac{1}{m}\sum_{i=1}^{m}\left\langle\xi(z_i),\frac{1}{m}\xi(z_i)+\frac{1}{m}\sum_{j\neq i}\xi(j)\right\rangle\\&=\frac{1}{m^2}\sum_{i=1}^{m}\|\xi(z_i)\|^2+\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\langle\xi(z_i),\xi(z_j)\rangle.\end{aligned}$$

By taking the expectation of above equality and using (2.3), we obtain

$$\begin{aligned}E\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)\right\|^2&=\frac{1}{m^2}\sum_{i=1}^{m}E\|\xi(z_i)\|^2+\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}E\langle\xi(z_i),\xi(z_j)\rangle\\&\leq\frac{1}{m}E\|\xi\|^2+\frac{m-1}{m}\|E\xi\|^2+\frac{30}{m}\|\xi\|_{2+\delta}^2\sum_{l=1}^{m-1}\alpha^{\frac{\delta}{2+\delta}}(Z,l)\\&=\frac{1}{m}\|\xi\|_2^2+\frac{m-1}{m}\|E\xi\|^2+\frac{30}{m}\|\xi\|_{2+\delta}^2\sum_{l=1}^{m-1}\alpha^{\frac{\delta}{2+\delta}}(Z,l).\end{aligned}\tag{2.4}$$

On the other hand, it is easy to see from (2.1) that

$$\begin{aligned}\sum_{l=1}^{m-1}\alpha^{\frac{\delta}{2+\delta}}(Z,l)&\leq c\sum_{l=1}^{m-1}\exp\left(\frac{-b\delta l^t}{2+\delta}\right)\\&\leq c\int_0^{m-1}\exp\left(\frac{-b\delta x^t}{2+\delta}\right)dx\\&\leq c\left(\frac{2+\delta}{\delta}\right)^{1/t}\int_0^{\infty}b^{-1/t}\exp(-y^t)dy=C_1\left(\frac{2+\delta}{\delta}\right)^{1/t}.\end{aligned}\tag{2.5}$$

By putting (2.5) and (2.4) into (2.2), we get

$$\begin{aligned}E\left\|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-E(\xi)\right\|^2&\leq\frac{1}{m}\|\xi\|_2^2-\frac{1}{m}\|E\xi\|^2+\frac{30C_1\left(\frac{2+\delta}{\delta}\right)^{1/t}}{m}\|\xi\|_{2+\delta}^2\\&\leq\frac{1}{m}\|\xi\|_2^2+\frac{30C_1}{m}\left(\frac{2+\delta}{\delta}\right)^{1/t}\|\xi\|_{2+\delta}^2.\end{aligned}$$

The result follows. ∎

In the next section, we consider Lemma 2.2 for α-mixing sequence of dependent samples and analyze the learning rate with both labelled and unlabelled data.

3. Regularized semi-supervised least squares regression

In addition to a set of labelled samples $D=(z_i)_{i=1}^m=\{(x_i,y_i)\}_{i=1}^m$ drawn according to ρ, we have a sequence of unlabelled data

$$D_{\tilde{\mathbf{x}}}=\{x_{m+1},\cdots,x_{m+l}\}.$$

Here we combine both labelled and unlabelled data together by constructing a combined training set as follows:

$$D^*=\{(x_i,y_i')\}_{i=1}^{m+l},$$

where

$$y_i'=\begin{cases}\frac{m+l}{m}y_i, & if\ 1\le i\le m,\\ 0, & otherwise.\end{cases}$$

Denote by

$$D_{\mathbf{x}}^*=(x_1,\cdots,x_{m+l})\in\mathbb{R}^{m+l},\quad D_{\mathbf{x}}=(x_1,\cdots,x_m)\in\mathbb{R}^m,$$

and

$$D_{\mathbf{y}}^*=(y_1',\cdots,y_m',0,\cdots,0)\in\mathbb{R}^{m+l},\quad D_{\mathbf{y}}=(y_1,\cdots,y_m)\in\mathbb{R}^m.$$

With the training data D^*, the semi-supervised regularized least squares regression scheme is given by

$$f_{D^*,\lambda}:=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m+l}\sum_{i=1}^{m+l}(y_i'-f(x_i))^2+\lambda\|f\|_K^2\right\}. \tag{3.1}$$

We see that the problem in (3.1) is similar to that in (1.1) except that unlabelled data is used. The aim is still to estimate $\|f_{D^*,\lambda}-f_\rho\|_{L^2_{\rho_X}}$ or more precisely $\|f_{D^*,\lambda}-f_\lambda\|_{L^2_{\rho_X}}+\|f_\lambda-f_\rho\|_{L^2_{\rho_X}}$, i.e., the sum of the sample error and the approximation error. Here f_λ is defined in (1.3).

Define the sampling operator $S_{D_{\mathbf{x}}^*}:\mathcal{H}_K\to\mathbb{R}^{m+l}$ as $S_{D_{\mathbf{x}}^*}(f):=(f(x_1),\cdots,f(x_{m+l}))\in\mathbb{R}^{m+l}$. Then its adjoint is

$$S_{D_{\mathbf{x}}^*}^T\mathbf{c}:=\frac{1}{m+l}\sum_{i=1}^{m+l}c_iK_{x_i}$$

for $\mathbf{c}=(c_1,\cdots,c_{m+l})\in\mathbb{R}^{m+l}$. Let $L_{K,D_{\mathbf{x}}^*}$ be the data-dependent approximation of L_K defined by

$$L_{K,D_{\mathbf{x}}^*}f:=S_{D_{\mathbf{x}}^*}^TS_{D_{\mathbf{x}}^*}f=\frac{1}{m+l}\sum_{i=1}^{m+l}f(x_i)K_{x_i}.$$

It is well-known (see [3,15]) that

$$f_\lambda=(L_K+\lambda I)^{-1}L_Kf_\rho \tag{3.2}$$

and

$$f_{D^*,\lambda}=(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}S^T_{D^*_{\mathbf{x}}}D^*_{\mathbf{y}}.$$

It is easy to check

$$S^T_{D^*_{\mathbf{x}}}D^*_{\mathbf{y}}=\frac{1}{m}\sum_{i=1}^{m}y_iK_{x_i}=S^T_{D_{\mathbf{x}}}D_{\mathbf{y}}.$$

Therefore,

$$f_{D^*,\lambda}=(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}S^T_{D_{\mathbf{x}}}D_{\mathbf{y}}. \tag{3.3}$$

We see from (3.3) that $f_{D^*,\lambda}$ involves the labelled data $D_{\mathbf{y}}$ and data-dependent samples $D^*_{\mathbf{x}}$.

3.1. Error analysis.

Firstly, the following lemma can be found in [15], which gives an estimate of the approximation error $\|f_\lambda-f_\rho\|_{L^2_{\rho_X}}$.

Lemma 3.1. *Assume (1.4) with $0<r\leq 1$. There holds*

$$\|f_\lambda-f_\rho\|_{L^2_{\rho_X}}\leq\lambda^r\|g_\rho\|_{L^2_{\rho_X}}. \tag{3.4}$$

Now we focus on estimating the sample error $\|f_{D^*,\lambda}-f_\lambda\|_{L^2_{\rho_X}}$.

Proposition 3.1. *Let $f_{D^*,\lambda}$ and f_λ be given by (3.3) and (3.2) respectively. Then we have*

$$\|f_{D^*,\lambda}-f_\lambda\|_{L^2_{\rho_X}}\leq I_1(I_2+I_3\|f_\lambda\|_K),$$

where

$$I_1:=\|(L_K+\lambda I)(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}\|,$$

$$I_2:=\|(L_K+\lambda I)^{-1/2}(L_Kf_\rho-S^T_{D_{\mathbf{x}}}D_{\mathbf{y}})\|_K$$

and

$$I_3:=\|(L_K+\lambda I)^{-1/2}(L_K-L_{K,D^*_{\mathbf{x}}})\|.$$

Proof. We first find that

$$\begin{aligned}f_{D^*,\lambda}-f_\lambda&=(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}S^T_{D_{\mathbf{x}}}D_{\mathbf{y}}-(L_K+\lambda I)^{-1}L_Kf_\rho\\&=(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}(S^T_{D_{\mathbf{x}}}D_{\mathbf{y}}-L_Kf_\rho)+[(L_{K,D^*_{\mathbf{x}}}+\lambda I)^{-1}-(L_K+\lambda I)^{-1}]L_Kf_\rho.\end{aligned}$$

By [5, Corollary 4.13], every $f\in\mathcal{H}_K$ can be written as $f=L_K^{1/2}g$ for some $g\in L^2_{\rho_X}$, and

$$\|g\|_{L^2_{\rho_X}}=\|f\|_K\leq\|(L_K+\lambda I)^{1/2}g\|_K.$$

It follows that

$$\|f_{D^*,\lambda}-f_\lambda\|_{L^2_{\rho_X}}$$

$$
\begin{aligned}
&\leq \|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1}(S_{D_{\mathbf{x}}}^T D_{\mathbf{y}} - L_K f_\rho)\|_K \\
&\quad + \|(L_K+\lambda I)^{1/2}[(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1} - (L_K+\lambda I)^{-1}]L_K f_\rho\|_K \\
&\leq \|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1/2}\|^2 \|(L_K+\lambda I)^{-1/2}(L_K f_\rho - S_{D_{\mathbf{x}}}^T D_{\mathbf{y}})\|_K \\
&\quad + \|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1}(L_K - L_{K,D_{\mathbf{x}}^*})(L_K+\lambda I)^{-1}L_K f_\rho\|_K \\
&\leq \|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1/2}\|^2 \|(L_K+\lambda I)^{-1/2}(L_K f_\rho - S_{D_{\mathbf{x}}}^T D_{\mathbf{y}})\|_K \\
&+ \|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1/2}\|^2 \|(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_{\mathbf{x}}^*})\| \|(L_K+\lambda I)^{-1}L_K f_\rho\|_K.
\end{aligned}
$$

Here, we have used in the second inequality, the fact

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$$

for positive operators A and B. Moveover, we have used in the second and last inequality, the fact

$$\|AB\| = \|A^T B^T\| = \|(BA)^T\| = \|BA\|$$

for self-adjoint operators A and B. By Lemma 1 in [7], we know that

$$\|(L_K+\lambda I)^{1/2}(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1/2}\|^2 \leq \|(L_K+\lambda I)(L_{K,D_{\mathbf{x}}^*}+\lambda I)^{-1}\|.$$

We thus prove the proposition by using (3.2). □

Next, we define a quantity measuring the complexity of $\mathcal{H}_K$ with respect to ρ_X, the effective dimension (see [20]) is the trace of the operator $(L_K+\lambda I)^{-1}L_K$ as

$$\mathcal{N}(\lambda) := Tr((L_K+\lambda I)^{-1}L_K), \ \lambda > 0.$$

Also, we denote $\kappa := \sup_{x\in X} \sqrt{K(x,x)}$.

We are setting out to bound I_1, I_2, I_3 by some lemmas. Some techniques used in the proofs are adopted from [11, 16].

Lemma 3.2. *Let the sampling sequence $D_{\mathbf{x}}^*$ be a stationary α-mixing sequence that satisfies (2.1). Then, for any* $0 < \delta \leq +\infty$,

$$E\left[\|I_3\|^2\right] \leq \frac{\kappa^2\mathcal{N}(\lambda)}{m+l} + \frac{30C_1\kappa^{\frac{4(1+\delta)}{2+\delta}}}{m+l}\left(\frac{2+\delta}{\delta}\right)^{1/t}\left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}}\right)^{2/(2+\delta)}. \tag{3.5}$$

Proof. We would like to apply Lemma 2.2 to the random variable ξ_1 defined by

$$\xi_1(x) := (L_K+\lambda I)^{-1/2} < \cdot, K_x >_K K_x, \ x \in X.$$

It takes values in $HS(\mathcal{H}_K)$, the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}_K$, with inner product $<A, B>_{HS} = Tr(B^T A)$. The norm is given by $\|A\|_{HS}^2 = \sum_i \|Ae_i\|_K^2$, where $\{e_i\}$ is an orthonormal basis of $\mathcal{H}_K$. The space $HS(\mathcal{H}_K)$ is a subspace of the space of bounded linear operators on $\mathcal{H}_K$, denoted as $(L(\mathcal{H}_K), \|\cdot\|)$, with the norm relations

$$\|A\| \leq \|A\|_{HS} \quad \text{and} \quad \|AB\|_{HS} \leq \|A\|_{HS}\|B\|. \tag{3.6}$$

It is easy to check that

$$E(\xi_1) = (L_K+\lambda I)^{-1/2} L_K$$

and

$$\frac{1}{m+l}\sum_{i=1}^{m+l}\xi_1(x_i)=(L_K+\lambda I)^{-1/2}L_{K,D^*_{\mathbf{x}}}.$$

Recall the set of normalized (in $\mathcal{H}_K$) eigenfunctions $\{\varphi_i\}_i$ of L_K. It is an orthonormal basis of $\mathcal{H}_K$. If we regard L_K as an operator on $L^2_{\rho_X}$, the normalized eigenfunctions in $L^2_{\rho_X}$ are $\{\frac{\varphi_i}{\sqrt{\lambda_i}}\}_i$ ($\lambda_i>0$ is the eigenvalue corresponding to φ_i); they form an orthonormal basis of the orthogonal complement of the eigenspace associated with the zero eigenvalue. By the Mercer Theorem,

$$K(x,\bar{x})=\sum_i\varphi_i(x)\varphi_i(\bar{x}). \tag{3.7}$$

By the definition of the HS norm and (3.7), we have

$$\begin{aligned}\|\xi_1(x)\|^2_{HS}&=\sum_i\|(L_K+\lambda I)^{-1/2}<\cdot,K_x>_K K_x\varphi_i\|^2_K\\&=\sum_i\|(L_K+\lambda I)^{-1/2}\varphi_i(x)K_x\|^2_K\\&=\sum_i\varphi_i^2(x)\|(L_K+\lambda I)^{-1/2}\sum_l\varphi_l(x)\varphi_l\|^2_K\\&=\sum_i\varphi_i^2(x)\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\\&=K(x,x)\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\|\xi_1\|_2^2=E\|\xi_1(x)\|^2_{HS}\ \le\ \kappa^2E\left[\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right]\ =\ \kappa^2\sum_l\frac{\|\varphi_l\|^2_{L^2_{\rho_X}}}{\lambda+\lambda_l}&=\kappa^2\sum_l\frac{\lambda_l}{\lambda+\lambda_l}\\&=\kappa^2\mathcal{N}(\lambda).\end{aligned}$$

For $\delta>0$, we obtain

$$\begin{aligned}\|\xi_1\|_{2+\delta}=(E\|\xi_1(x)\|^{2+\delta}_{HS})^{1/(2+\delta)}&\le\left[E\left(K(x,x)\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right)^{\frac{2+\delta}{2}}\right]^{1/(2+\delta)}\\&\le\kappa\left[E\left(\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right)\left(\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right)^{\frac{\delta}{2}}\right]^{1/(2+\delta)}\\&\le\kappa\left(\frac{\kappa}{\sqrt{\lambda}}\right)^{\frac{\delta}{2+\delta}}\left[E\left(\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right)\right]^{1/2+\delta}\\&\le\kappa^{\frac{2+2\delta}{2+\delta}}\left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}}\right)^{1/(2+\delta)},\end{aligned}$$

and for $\delta=\infty$,

$$\|\xi_1\|_\infty=\sup\|\xi_1(x)\|_{HS}=\sup\left(K(x,x)\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right)^{1/2}\le\frac{\kappa^2}{\sqrt{\lambda}}.$$

By applying Lemma 2.2 to $\xi_1(x)$, we have

$$\begin{aligned}E[\|I_3\|_{HS}^2]&=E\left[\|(L_K+\lambda I)^{-1/2}(L_K-L_{K,D_{\mathbf{x}}^*})\|_{HS}^2\right]\\&=E\left\|\frac{1}{m+l}\sum_{i=1}^{m+l}\xi_1(x_i)-E(\xi_1)\right\|_{HS}^2\\&\le\frac{\kappa^2\mathcal{N}(\lambda)}{m+l}+\frac{30C_1\kappa^{\frac{4(1+\delta)}{2+\delta}}}{m+l}\left(\frac{2+\delta}{\delta}\right)^{1/t}\left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}}\right)^{2/(2+\delta)}.\end{aligned}$$

Then (3.5) follows from the first inequality of (3.6). □

Lemma 3.3. *Let the sampling sequence $D_{\mathbf{x}}^*$ be a stationary α-mixing sequence that satisfies (2.1), g be a measurable bounded function on $X\times Y$ and ξ_g be the random variable with values on $\mathcal{H}_K$ given by $\xi_g(z)=g(z)K_x$ for $z=(x,y)\in X\times Y$. Then, for any $0<\delta\le\infty$*

$$\begin{aligned}&E\left[\left\|(L_K+\lambda I)^{-1/2}\left(\frac{1}{m}\sum_{1=1}^{m}\xi_g(z_i)-E(\xi_g)\right)\right\|_K^2\right]\\&\le\|g\|_\infty^2\left(\frac{\mathcal{N}(\lambda)}{m}+\frac{30C_1\kappa^{\frac{2\delta}{2+\delta}}}{m}\left(\frac{2+\delta}{\delta}\right)^{1/t}\left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}}\right)^{2/(2+\delta)}\right).\end{aligned}\tag{3.8}$$

Proof. Consider the random variable ξ_2 defined by

$$\xi_2(z)=(L_K+\lambda I)^{-1/2}(\xi_g(z)),\ z\in X\times Y.$$

It takes values in $\mathcal{H}_K$. By (3.7), we obtain

$$\begin{aligned}\|\xi_2\|_2^2=E\|\xi_2(z)\|_K^2=E\left[\|(L_K+\lambda I)^{-1/2}(g(z)K_x)\|_K^2\right]&\le\|g\|_\infty^2E\left[\|(L_K+\lambda I)^{-1/2}\sum_l\varphi_l(x)\varphi_l\|_K^2\right]\\&=\|g\|_\infty^2E\left[\sum_l\frac{\varphi_l^2(x)}{\lambda+\lambda_l}\right]\\&=\|g\|_\infty^2\sum_l\frac{\|\varphi_l\|_{L^2_{\rho_X}}^2}{\lambda+\lambda_l}\\&=\|g\|_\infty^2\sum_l\frac{\lambda_l}{\lambda+\lambda_l}=\|g\|_\infty^2\mathcal{N}(\lambda).\end{aligned}$$

For $\delta>0$, we get

$$\|\xi_2\|_{2+\delta}=(E\|\xi_2(z)\|_K^{2+\delta})^{1/(2+\delta)}\le\|g\|_\infty\left(E\left[\|(L_K+\lambda I)^{-1/2}K_x)\|_K^{2+\delta}\right]\right)^{1/(2+\delta)}$$

$$
\begin{aligned}
&= \|g\|_\infty \left(E\left[\sum_l \frac{\varphi_l^2(x)}{\lambda+\lambda_l} \right] \left[\sum_l \frac{\varphi_l^2(x)}{\lambda+\lambda_l} \right]^{\frac{\delta}{2}} \right)^{\frac{1}{2+\delta}} \\
&\le \|g\|_\infty \kappa^{\frac{\delta}{2+\delta}} \left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}} \right)^{\frac{1}{2+\delta}}.
\end{aligned}
$$

and for $\delta = \infty$, we have

$$
\begin{aligned}
\|\xi_2\|_\infty = \sup \|\xi_2(z)\|_K \le \|g\|_\infty \sup \left\| (L_K+\lambda I)^{-1/2} K_x) \right\|_K
&= \|g\|_\infty \sup \left[\sum_l \frac{\varphi_l^2(x)}{\lambda+\lambda_l} \right]^{1/2} \\
&\le \|g\|_\infty \sup \left(\frac{K(x,x)}{\lambda} \right)^{1/2} \\
&\le \frac{\kappa \|g\|_\infty}{\sqrt{\lambda}}.
\end{aligned}
$$

By applying Lemma 2.2 to ξ_2, we derive

$$
\begin{aligned}
&E\left[\left\| (L_K+\lambda I)^{-1/2} \left(\frac{1}{m} \sum_{1=1}^m \xi_g(z_i) - E(\xi_g) \right) \right\|_K^2 \right] \\
&= E\left[\left\| \frac{1}{m} \sum_{i=1}^m \xi_2(z_i) - E(\xi_2) \right\|_K^2 \right] \\
&\le \frac{\|g\|_\infty^2 \mathcal{N}(\lambda)}{m} + \frac{30 C_1 \kappa^{\frac{2\delta}{2+\delta}} \|g\|_\infty^2}{m} \left(\frac{2+\delta}{\delta} \right)^{1/t} \left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}} \right)^{2/(2+\delta)}.
\end{aligned}
$$

The result follows. □

By applying Lemma 3.3, we can get the estimate of $E[I_2^2]$.

COROLLARY 3.1. *Let the sampling sequence $D_{\mathbf{x}}^*$ be a stationary α-mixing sequence that satisfies (2.1). Then for any $0 < \delta \le \infty$*

$$
E\left[I_2^2 \right] \le M^2 \left(\frac{\mathcal{N}(\lambda)}{m} + \frac{30 C_1 \kappa^{\frac{2\delta}{2+\delta}}}{m} \left(\frac{2+\delta}{\delta} \right)^{1/t} \left(\frac{\mathcal{N}(\lambda)}{\lambda^{\delta/2}} \right)^{2/(2+\delta)} \right). \tag{3.9}
$$

Proof. Taking $g(z) = y$ for any $z = (x,y) \in X \times Y$. Then $E(\xi_g) = L_K f_\rho$. Therefore,

$$
I_2 = \|(L_K+\lambda I)^{-1/2} (L_K f_\rho - S_{D_{\mathbf{x}}}^T D_{\mathbf{y}})\|_K = \left\| (L_K+\lambda I)^{1/2} \left(\frac{1}{m} \sum_{1=1}^m \xi_g(z_i) - E(\xi_g) \right) \right\|_K.
$$

Since $|y| \le M$, $\|g\|_\infty \le M$. Then our desirable bound is obtained by using Lemma 3.3. □

The last task is to estimate the bound I_1 by a decomposition of operator product. Let A and B be invertible operators on a Banach space. Then we can get (see [11])

$$
BA^{-1} = (B-A)B^{-1}(B-A)A^{-1} + (B-A)B^{-1} + I. \tag{3.10}
$$

LEMMA 3.4. *Let the sampling sequence $D_{\mathbf{x}}^*$ be a stationary α-mixing sequence that satisfies (2.1). Then. we have*

$$I_1 \le 2(\lambda^{-1} I_3^2 + 1). \tag{3.11}$$

Proof. We apply (3.10) to operators $A = L_{K,D_{\mathbf{x}}^*} + \lambda I$ and $B = L_K + \lambda I$. Since

$$\|(L_{K,D_{\mathbf{x}}^*} + \lambda I)^{-1}\| \le 1/\lambda \quad \text{and} \quad (L_K + \lambda I)^{-1/2} \le 1/\sqrt{\lambda},$$

we find

$$\begin{aligned} I_1 &\le \lambda^{-1} \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D_{\mathbf{x}}^*})\|^2 + \lambda^{-1/2} \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D_{\mathbf{x}}^*})\| + 1 \\ &= \lambda^{-1} I_3^2 + \lambda^{-1/2} I_3 + 1 \ \le \ 2(\lambda^{-1} I_3^2 + 1). \end{aligned}$$

□

3.2. Learning rate. To derive the explicit learning rate of algorithm (3.1), one needs the following assumption as a characteristic of the complexity of the hypothesis space (see [1,3,7,11]):

$$\mathcal{N}(\lambda) \le C_0 \lambda^{-\beta}, \ \forall \lambda > 0. \tag{3.12}$$

for some $0 < \beta \le 1$ and $C_0 > 0$. Now we establish the main results of this paper.

THEOREM 3.1. *Let the sampling sequence $D_{\mathbf{x}}^*$ be a stationary α-mixing sequence that satisfies (2.1). Let $f_{D^*,\lambda}$ be given by (3.3). Assume (3.12) holds. For $0 < r < 1/2$, if $\lambda = m^{\frac{-1}{2r+\beta}}$ and $l \ge m^{\frac{1+\beta}{2r+\beta}} - m$, then, for any $0 < \eta < 1$, with confidence $1 - \eta$,*

$$\|f_{D^*,\lambda} - f_\rho\|_{L^2_{\rho_X}} \le C' \left(\frac{(\log m)^{1/t}}{\eta} \right)^{\frac{3}{2}} m^{-\frac{r}{2r+\beta}}. \tag{3.13}$$

For $1/2 \le r \le 1$, if $\lambda = m^{\frac{-1}{2r+\beta}}$ and $l = 0$, then, for any $0 < \eta < 1$, with confidence $1 - \eta$,

$$\|f_{D,\lambda} - f_\rho\|_{L^2_{\rho_X}} \le \tilde{C} \left(\frac{(\log m)^{1/t}}{\eta} \right)^{\frac{3}{2}} m^{-\frac{r}{2r+\beta}}. \tag{3.14}$$

Here C' and $\tilde{C}$ are constants independent of m, l and η.

Proof. For $0 < r < \frac{1}{2}$, we set $\delta = \frac{2}{\log m - 1}$, $l \ge m^{\frac{1+\beta}{2r+\beta}} - m$ and $\lambda = m^{\frac{-1}{2r+\beta}}$. Then by Lemma 3.3 and the assumption in (3.12), we have

$$\begin{aligned} E[I_3^2] &\le \frac{\kappa^2 \mathcal{N}(\lambda)}{m+l} + \frac{30 C_1 \kappa^4 (\log m)^{1/t}}{m+l} \left(\frac{(\mathcal{N}(\lambda))^{1-1/\log m}}{\lambda^{1/\log m}} \right) \\ &\le (C_0 \kappa^2 + 30 C_0 C_1 \kappa^4)(\log m)^{1/t} \left[\frac{m^{\beta/(2r+\beta)}}{m+l} + \frac{m^{\beta/(2r+\beta)} e^{\frac{1-\beta}{2r+\beta}}}{m+l} \right] \\ &\le 2(C_0 \kappa^2 + 30 C_0 C_1 \kappa^4) e^{\frac{1-\beta}{2r+\beta}} (\log m)^{1/t} m^{\frac{-1}{2r+\beta}}. \end{aligned}$$

Therefore, for any $0 < \eta < 1$, by using Markov inequality, with confidence $1 - \frac{\eta}{2}$, we obtain

$$I_3 \le \frac{2}{\sqrt{\eta}} ((C_0 \kappa^2 + 30 C_0 C_1 \kappa^4) e^{\frac{1-\beta}{2r+\beta}})^{1/2} (\log m)^{\frac{1}{2t}} m^{\frac{-1}{2(2r+\beta)}}. \tag{3.15}$$

It thus follows from Lemma 3.4 with the same confidence set as in (3.15):

$$I_1 \le \frac{8}{\eta}(\log m)^{\frac{1}{t}}((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1). \tag{3.16}$$

Similarly, by taking $\delta=\frac{2}{\log m-1}$ and $\lambda=m^{\frac{-1}{2r+\beta}}$, and using Corollary 3.1 and the assumption in (3.12), we have

$$\begin{aligned} E[I_2^2] &\le M^2\left(\frac{\kappa^2\mathcal{N}(\lambda)}{m}+\frac{30C_1\kappa^2(\log m)^{1/t}}{m}\left(\frac{(\mathcal{N}(\lambda))^{1-1/\log m}}{\lambda^{1/\log m}}\right)\right)\\ &\le M^2(C_0\kappa^2+30C_0C_1\kappa^2)(\log m)^{1/t}m^{\frac{-2r}{2r+\beta}}\left(1+m^{\frac{1-\beta}{(2r+\beta)\log m}}\right)\\ &\le 2M^2(C_0\kappa^2+30C_0C_1\kappa^2)e^{\frac{1-\beta}{2r+\beta}}(\log m)^{1/t}m^{\frac{-2r}{2r+\beta}}. \end{aligned}$$

Again, by using Markov inequality, it implies that with confidence $1-\frac{\eta}{2}$,

$$I_2 \le \frac{2M}{\sqrt{\eta}}((C_0\kappa^2+30C_0C_1\kappa^2)e^{\frac{1-\beta}{2r+\beta}})^{1/2}(\log m)^{\frac{1}{2t}}m^{\frac{-r}{2r+\beta}}. \tag{3.17}$$

Moreover, by (3.2) and (1.4), we have

$$\begin{aligned} \|f_\lambda\|_K &= \|(L_K+\lambda I)^{-1}L_Kf_\rho\|_K\\ &= \|(L_K+\lambda I)^{-1}L_KL_K^rg_\rho\|_K\\ &\le \|(L_K+\lambda I)^{-1}L_K^{r+1/2}\|\|L_K^{1/2}g_\rho\|_K\\ &\le \lambda^{r-1/2}\|g_\rho\|_{L^2_{\rho_X}}\\ &= m^{\frac{1-2r}{2(2r+\beta)}}\|g_\rho\|_{L^2_{\rho_X}}. \end{aligned} \tag{3.18}$$

Now putting the estimates (3.15)-(3.18) into Proposition 3.1, we get with confidence $1-\eta$

$$\begin{aligned} &\|f_{D^*,\lambda}-f_\lambda\|_{L^2_{\rho_X}}\\ &\le 32\left(\frac{((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1)(\log m)^{1/t}}{\eta}\right)^{\frac{3}{2}}(M+\|g_\rho\|_{L^2_{\rho_X}})m^{\frac{-r}{2r+\beta}}. \end{aligned}$$

By combining the estimate of the approximation error in Lemma 3.1 and the above estimate of the sample error, the results in (3.13) follow by setting

$$C'=32((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1)^{\frac{3}{2}}(M+\|g_\rho\|_{L^2_{\rho_X}})+\|g_\rho\|_{L^2_{\rho_X}}.$$

When $1/2\le r<1$, we take $l=0$ (i.e., without using other unlabelled samples). Then $f_{D^*,\lambda}=f_{D,\lambda}$. By (3.2), we have

$$\|f_\lambda\|_K=\|(L_K+\lambda I)^{-1}L_Kf_\rho\|_K\le\|f_\rho\|_K\le\kappa M. \tag{3.19}$$

By taking $\delta=\frac{2}{\log m-1}$ and $\lambda=m^{\frac{-1}{2r+\beta}}$ in Lemma 3.3, we can see from (3.12) that

$$E[I_3^2]\le\frac{\kappa^2\mathcal{N}(\lambda)}{m}+\frac{30C_1\kappa^4(\log m)^{1/t}}{m}\left(\frac{(\mathcal{N}(\lambda))^{1-1/\log m}}{\lambda^{1/\log m}}\right)$$

$$\leq (C_0\kappa^2+30C_0C_1\kappa^4)(\log m)^{1/t} m^{\frac{-2r}{2r+\beta}}(1+e^{\frac{1-\beta}{2r+\beta}})$$
$$\leq 2(C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}(\log m)^{1/t} m^{\frac{-2r}{2r+\beta}}.$$

It follows from Markov inequality that with confidence $1-\eta/2$

$$I_3 \leq \frac{2}{\sqrt{\eta}}((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}})^{1/2}(\log m)^{\frac{1}{2t}} m^{\frac{-r}{2r+\beta}}. \tag{3.20}$$

Then by Lemma 3.4, with the same confidence set as in (3.20), there holds

$$I_1 \leq \frac{8}{\eta}(\log m)^{\frac{1}{t}}((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1). \tag{3.21}$$

Therefore, putting (3.17) and (3.19)-(3.21) into Proposition 3.1, we get with confidence $1-\eta$,

$$\|f_{D,\lambda}-f_\lambda\|_{L^2_{\rho_X}}$$
$$\leq 32\left(\frac{((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1)(\log m)^{1/t}}{\eta}\right)^{\frac{3}{2}}(\kappa+1)Mm^{\frac{-r}{2r+\beta}}.$$

This together with (3.2) and (3.4) implies that with the confidence $1-\eta$,

$$\|f_{D,\lambda}-f_\rho\|_{L^2_{\rho_X}}$$
$$\leq \left\{32\left(\frac{((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1)(\log m)^{1/t}}{\eta}\right)^{\frac{3}{2}}(\kappa+1)M+\|g_\rho\|_{L^2_{\rho_X}}\right\}m^{\frac{-r}{2r+\beta}}.$$

We thus prove (3.14) with $\tilde{C}=32((C_0\kappa^2+30C_0C_1\kappa^4)e^{\frac{1-\beta}{2r+\beta}}+1)^{\frac{3}{2}}(\kappa+1)M+\|g_\rho\|_{L^2_{\rho_X}}$. □

Remark 3.1. The condition (3.12) with $\beta=1$ is always satisfied with $C_0=\kappa^2$. For $0<\beta<1$, (3.12) is more general than the eigenvalue decaying assumption in the literature (see, e.g. [3]). It is shown in [7] if $\lambda_i \sim i^{-2\alpha}$ for some $\alpha>\frac{1}{2}$, then (3.12) holds with $\beta=\frac{1}{2\alpha}$. So the results in Theorem 3.1 state that the regularized semi-supervised least squares algorithm can achieve the nearly minimax optimal learning rate for dependent samples and $r\in(0,1]$, except an extra logarithmic term compared with the minimax optimal learning rate $\mathcal{O}(m^{\frac{-2\alpha}{4\alpha r+1}})$ of the regularized least squares algorithm for independent samples.

Also it is obvious that for $r\in(0,\frac{1}{2})$, the learning rate of the proposed method in Theorem 3.1 is better than that $\mathcal{O}\left(m^{\frac{-3r}{4(1+r)}}(\log m)^{\frac{1}{2t}}\right)$ stated in [17].

Remark 3.2. In our setting, the parameter t (2.1) is used to measure the correlations of the samples. If t is large, the samples are more “independent”. The learning rate in Theorem 3.1 will get better; this is reflected by the logarithmic term $(\log m)^{\frac{3}{2t}}$ (t only appears in this term) becoming smaller. In particular, when $t\to\infty$, the samples are completely independent and the logarithmic term disappears. The results in Theorem 3.1 accordingly reduce to the same minimax optimal learning rate as that for the independent sampling.

4. Concluding remarks

We have studied and analyzed regularized least squares regression with dependent samples. Because of dependent samples, the learning rate of the regularized least squares algorithm is not optimal. In order to improve the learning rate of the regularized least squares regression with dependent samples, we make use of unlabelled data, i.e., we collect more data samples without label. In practice, the cost of labelling can be expensive, but the cost of collecting data samples is quite cheap. Our idea is to formulate a regularized semi-supervised least squares regression using all data samples (even dependent) and limited labels. The theoretical results show that except for a logarithmic term which is caused by the correlations of the samples, the learning rate of the proposed method is nearly minimax optimal no matter $f_\rho \in \mathcal{H}_K$ or not.

Acknowledgements. The authors thank the referees for their valuable comments and suggestions. The research is supported in part by HKRGC GRF 12302715, 12306616, 12200317, and HKBU RC-ICRS/16-17/03.

REFERENCES

[1] G. Blanchard and K. Krämer, *Optimal rates for kernel conjugate gradient regression*, Advances in Neural Information Processing Systems, 226–234, 2010.

[2] R.C. Bradley, *Introduction to Strong Mixing Conditions*, Kendrick Press, Heber City, UT, 1-3, 2007.

[3] A. Caponnetto and E. De Vito, *Optimal rates for the least squares algorithm*, Found. Comput. Math., 7:331–368, 2007.

[4] X. Chang, S.B. Lin, and D.X. Zhou, *Distributed semi-supervised learning with kernel ridge regression*, J. Mach. Learn. Res., 18:1–22, 2017.

[5] F. Cuker and D.X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University, Cambridge, 2007.

[6] H. Dehling and W. Philipp, *Almost sure invariance principles for weakly dependent vectorvalued random variables*, Ann. Probab., 10:689–701, 1982.

[7] Z.C. Guo, S.B. Lin, and D.X. Zhou, *Learning theory of distributed spectral algorithms*, Inverse Problems, 33(7):074009, 2017.

[8] I.A. Ibragimov, *Some limit theorems for stationary processes*, Theory Probab. Appl., 7:349–382, 1962.

[9] L. Jing and M. Ng, *Sparse label-indicator optimization methods for image classification*, IEEE Transactions on Image Processing, 23:1002–1014, 2014.

[10] X. Kong, M. Ng, and Z. Zhou, *Transductive multi-label learning via label set propagation*, IEEE Transactions on Knowledge and Data Engineering, 25:704–719, 2013.

[11] S.B. Lin, X. Guo, and D.X. Zhou, *Distributed learning with regularized least squares*, J. Mach. Learn. Res., 18:1–31, 2017.

[12] S.B. Lin and D.X. Zhou, *Distributed kernel-based gradient descent algorithms*, Constr. Approx., 47:249–276, 2018.

[13] D.S. Modha and E. Masry, *Minimum complexity regression estimation with weakly dependent observations*, IEEE. Trans. Inform. Theory, 42:2133–2145, 1996.

[14] M. Rosenblatt, *A central limit theorem and a strong mixing condition*, Proc. Nat. Ac. Sc. USA, 42:43–47, 1956.

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

[16] H.W. Sun and Q. Wu, *Regularized least square regression with dependent samples*, Adv. Comput. Math., 32:175–189, 2010.

[17] H.W. Sun and Q. Wu, *A note on application of integral operator in learning theory*, Appl. Comput. Harmon. Anal., 26:416–421, 2009.

[18] V.A. Wolkonski and Y.A. Rozanov, *Some limit theorems for random functions*, Theory Probab. Appl., 4:178–197, 1959.

[19] Y.L. Xu and D.R. Chen, *Learning rates of regularized regression for exponentially strongly mixing sequence*, J. Statist. Plann. Inference, 138:2180–2189, 2008.

[20] T. Zhang, *Learning bounds for kernel regression using effective data dimensionality*, Neural Computation, 17:2077–2098, 2005.