Distributed Regression Learning with Dependent Samples

To cite this article: Xiaoqing Zheng et al 2019 J. Phys.: Conf. Ser. 1213 022002

View the article online for updates and enhancements.

You may also like

- Online measurement of distributed capacitance in distribution network
  Xiuhua Zhang

- Applicability evaluation of blockchain technology in distributed power transaction
  Han Zhang, Song Xue, Li Ma et al.

- Guest Editor's introduction: Selected papers from the 4th USENIX Conference on Object-Oriented Technologies and Systems
  Joe Sventek
Distributed Regression Learning with Dependent Samples

Xiaoqing Zheng\textsuperscript{1a} Hongwei Sun\textsuperscript{2b} Mengjuan Pang\textsuperscript{3c}

\textsuperscript{1}School of Mathematical Science  
University of Jinan  
Jinan 250022, P.R. China  
\textsuperscript{2}School of Mathematical Science  
University of Jinan  
Jinan 250022, P.R. China  
\textsuperscript{3}School of Mathematical Science  
University of Jinan  
Jinan 250022, P.R. China  

\textsuperscript{a}18254138790@163.com \textsuperscript{b}ss_sunhw@ujn.edu.cn \textsuperscript{c}pangmengjuan2015@163.com

ABSTRACT: Distributed learning is an effective way to analyze big data. In distributed regression, a typical approach is to partition the sample set \( \{ z_i \}_{i=1}^N \) into \( m \) disjoint data subsets of equal size, and then applies the kernel ridge regression algorithm to each sample subset to derive a local estimator, then averages them to get the global estimator. This paper mainly considers distributed regression learning with dependent samples of regularized least squares with mixing inputs that is involved in pre-existing literature [15]. Error bound in the \( K \)–metric has been derived and a novel error division method has been used to prove the asymptotic convergence for this distributed regularization learning. Learning rate of this algorithm will be obtained under a standard regularity condition on the regression function and the polynomial decay strongly mixing condition. It is proved that distributed learning is applicable to not only the i. i. d. samples but also dependent samples.

1. INTRODUCTION

Kernel based regularization networks (KRN) have been widely applied in data analysis. This paper aims at KRN algorithms for regression learning, which can be described as follows. Let \((X,\mathcal{A})\) be a compact metric space, \(Y = \mathbb{R} \). Suppose that \(\rho\) is a fixed but unknown probability distribution on \(Z = X \times Y\). The regression function is defined as

\[
 f_{\rho} = \int_Y y \rho(y|x)\,dx, x \in X.
\]

where \(\rho(y|x)\) is the conditional probability measure at \(x\) induced by \(\rho\). By the sample set \(D = \{(x_i, y_i)\}_{i=1}^N\) drawn identically according to the distribution \(\rho\), the kernel based regularized least squares algorithm learns an empirical regressor by

\[
 f_{D,\lambda} = \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{(x_i, y_i) \in D} \left( f(x) - y \right)^2 + \lambda f^\top f \right\}
\]
Here $H_{\kappa}$ is the reproducing kernel Hilbert space (RKHS) associated to the kernel $K$ (more properties see [2]), which is the completion of the linear span of the set of function $\{K_x = K(x, \cdot): x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_k$ given by $\langle K_x, K_{x'} \rangle_k = K(x, x')$, this paper takes bounded kernels, a.e.

$$\kappa = \sup_{x \neq y} \sqrt{K(x, x)} < \infty.$$  

The performance of the algorithm (2) has been studied thoroughly, seeing [16, 21, 4, 7] and reference therein.

The samples are usually assumed to be independent, but independence is a restrictive assumption and may be difficult to guarantee in many real data analysis. Thus, the learning with dependent samples has attracted a lot of attentions [12, 13, 19, 20, 22]. In this paper, the sample sequence $\{(x_i, y_i)\}_{i=1}^b$ is assumed to be the strongly mixing sequence satisfying the $\alpha$–mixing condition which has been shown very common in sampling process [1, 5, 13].

The $\alpha$–coefficient of two $\sigma$–fields $J$ and $D$ is defined as

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

For a sequence of samples $\{z_i\}_{i=1}^m$, it is denoted by $M_{\alpha}$ the $\sigma$–field generated by random variables $z_{i-1}, z_{i-2}, \ldots, z_0$. The random sequence $z_i, i \geq 1$ is said to satisfy a $\alpha$–mixing condition if

$$\alpha_j = \sup_{i=1} M_{\alpha_1} \rightarrow 0, \quad \text{as} \quad i \rightarrow \infty.$$  

For any continuous map $f: Z \rightarrow Z$, where $Z$ is a topological space and a measurable space with the Borel $\sigma$–algebra. It is certainly that the $\sigma$–field $M_{\alpha}$ generated by random variables $f(z_0), f(z_1), \ldots, f(z_i)$ satisfies $M_{\alpha} \subset M_{\alpha_i}$. Denote the strongly mixing coefficient of the sample sequence $f(z_i), i \geq 1$ by $\alpha_j$, then there holds

$$\alpha_j \leq \alpha_i, \forall i \geq 1.$$  

which shows that if $z_i, i \geq 1$ satisfies the $\alpha$–mixing condition, then the sample sequence $f(z_i), i \geq 1$ also satisfies the $\alpha$–mixing condition.

In order to process with big data, distributed learning has been explored for many algorithms, such as [3] and [10] for distributed spectral kernel algorithm, ref [6] for distributed semi-supervised kernel ridge regression learning, ref [11] for distributed bias correction regularization network and ref [14] for distributed gradient descent algorithm. This paper mainly explores distributed KRN with dependent samples. The algorithm is described as follows, firstly divides the data set $D$ into $m$ disjoint subsets $\{D_j\}_{j=1}^m$ equally, so the number of each subset is $n = \frac{N}{m}$, then applies the regularized least squares regression to every data subset $D_j$ to derive the local estimator $f_{D_j}$, a.e.

$$f_{D_j} = \arg \min_{f \in M_{\alpha}} \left\{ \frac{1}{n} \sum_{x \in D_j} \left( f(x) - y \right)^2 + \lambda \|f\|_{\alpha}^2 \right\}$$  

(3)

Finally, the global estimator $f_\alpha$ is synthesized by a weighted average,

$$f_\alpha = \frac{1}{m} \sum_{j=1}^m f_{D_j}\lambda$$  

(4)

The main purpose of this paper is to estimate the error bound and learning rates of the distributed KRN (3) and (4) with the weakly dependent samples, and to reveal the relation between the dependence of sampling and the consistency of the algorithm. The rest of this paper is arranged as follows. In section 2 the assumptions and the main result of this paper have been proposed. The lemmas used in the proof of the main results are given in section 3. In section 4, error bound of the distributed learning scheme with weakly dependent samples will be deduced. Moreover, the learning rate will be derived in section 5.
2. ASSUMPTIONS AND MAIN RESULTS

This section presents the error analysis of the distributed learning with dependent samples. The error estimation is stated in terms of the regularity of the regression function, the weakly dependence of sample sequences and the integral operator technique.

In the sequel, this paper assumes that \( y \leq M \) for some constant \( M > 0 \) almost surely, and takes the regularization parameter \( 0 < \lambda \leq \kappa \). By introducing the regularization function \( f_\lambda = (L_\lambda + \lambda I)^{-1}L_\lambda f_\rho \), the error \( \| f_\lambda - f_\rho \|_\kappa \) in expectation can be decomposed into approximation error and sample error. Here \( L_\lambda^2 \) is the Hilbert space of square integrable functions with respect to \( \rho \), and \( L_\lambda^2 : L_\rho^2 \to L_\rho^2 \) is the integral operator associated with the Mercer kernel \( K \), defined by

\[
L_\lambda f = \int K(\cdot,t) f(t) \rho(t)
\]

\( L_\lambda \) is a positive compact operator from \( L_\rho^2 \) to \( L_\rho^2 \), and the norm of operator \( L_\lambda \) is less than \( \kappa \). For more properties of it see [17, 18].

Since the learning algorithm in this paper is to find a good approximation for \( f_\rho \) in the function space \( H_\lambda \), but the target function \( f_\rho \) is often not in this hypothesis space. So an appropriate approximation condition is needed to derive the error bound. By taking the following frequently used approximation condition [16, 21],

\[
f_\rho = L_\beta^\alpha (u_\beta), \quad \text{for some } \frac{1}{2} \leq \beta \leq 1 \quad \text{and} \quad u_\beta \in L_\rho^2,
\]

where \( L_\beta^\alpha \) denoted the \( \beta \)-th power of the positive operator \( L_\lambda \) on \( L_\rho^2 \).

**Theorem 1.** Assume \( y \leq M \) almost surely and the sample sequences are assumed to satisfy \( \alpha \)-mixing condition and \( f_\rho = L_\beta^\alpha u_\beta \) with \( \frac{1}{2} \leq \beta \leq 1 \) and \( u_\beta \in L_\rho^2 \), then for \( \delta = \infty \), \( E \| f_\lambda - f_\rho \|_\kappa \) is bounded by \( a_0 \) times

\[
\lambda^2 m^{-1} n^{-1} \left(1 + \sum_{i=1}^{n} \alpha_i\right) + \lambda^{2\beta-1} n^{-1} \left(1 + \sum_{i=1}^{n} \alpha_i\right)^2 + \lambda_2 m^{-1} \sum_{i=1}^{n} \alpha_{21-i} + \lambda^{2\beta-1}
\]

(6)

Here \( \alpha_i \) is a constant only depending on \( \kappa, M \), \( \| f_\rho \|_\kappa \) and \( \| u_\beta \|_\rho^2 \).

**Corollary 1.** Under the assumption of Theorem 1, if the \( \alpha \)-mixing coefficients satisfy a polynomial decay, i.e., \( \alpha_i \leq a i^{-s} \) for some \( a > 0 \) and \( t > 0 \), when \( m = N^t \) with \( 0 < s < 1 \) by taking \( \lambda = N^{-h} \), it can be deduced that

When \( t \geq 1 \),

\[
E \| f_\lambda - f_\rho \|_\kappa \leq \begin{cases} O \left( N^{\frac{2\beta-1}{2\beta+1}} \log N \right), & 0 < s \leq \frac{2\beta-1}{2\beta+1} \\ O \left( N^{\frac{2\beta-1}{2\beta+1}} \log N \right), & \frac{2\beta-1}{2\beta+1} < s < 1. \end{cases}
\]

When \( 0 < t < 1 \),

\[
E \| f_\lambda - f_\rho \|_\kappa \leq \begin{cases} O \left( N^{\frac{2\beta-1}{2\beta+1}} \right), & 0 < s \leq \frac{2\beta-1}{2\beta+1} \\ O \left( N^{\frac{2\beta-1}{2\beta+1}} \right), & \frac{2\beta-1}{2\beta+1} < s < 1. \end{cases}
\]
3. PRELIMINARY LEMMAS

Denote the j-th data set \( D_j = \{ x'_i, y'_i \} : i = 1, \ldots, n \) \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \), the associated sampling operator \( S_{D_j} : H_k \rightarrow R^* \) is defined by

\[
S_{D_j} f = \left( f(x'_i) \right)_{i=1}^n, \quad \text{for any } f \in H_k.
\]

Its adjoint operator is

\[
S_{D_j}^* \alpha = \sum_{i=1}^n \alpha_i K(x'_i, \cdot).
\]

The integral operator \( L_k \) can be approximated by the sampling operator

\[
\frac{1}{n} \sum_{j=1}^n S_{D_j} = \frac{1}{n} \sum_{j=1}^n K(z'_j \otimes K_{x'_j} \otimes EK_{x'_j} \otimes K_{x'_j}) = L_k.
\]

If \( \xi \) be a random variable with \( u \) th moment in a Hilbert space and \( 1 \leq u \leq + \infty \), denote \( \| \xi \|_u = \sup \| \xi \|_{L^u} \). The following lemma gives a way to deal with the dependence of samples, it has been proved in [8, 9].

**Lemma 3.1.** Let the random sequence \( z \) satisfies the \( \alpha \) – mixing condition, If \( \xi \) is measurable with respect to \( M^n \), \( \eta \) is measurable with respect to \( M^n \), \( \xi \) and \( \eta \) have \( \infty \) – th moment, if \( \alpha = \nu = \infty \), \( t = 1 \). then

\[
\| E(\xi, \eta) - (E \xi, E \eta) \| \leq 15 \alpha \| \xi \|_\infty \| \eta \|_\infty.
\]

Denoted the random variable valued in \( H_k \) by \( \xi(z) = (y - f_\alpha K_z) \), then there have the following Lemma 3.2.

**Lemma 3.2.** Let \( \Delta = E \left| \frac{1}{n} \sum_{i=1}^n \xi(z_i) - L_k \left( f_{\alpha} - f_j \right) \right| \) for \( \delta = + \infty \), there holds

\[
\Delta \leq \frac{30 \kappa^2}{n} (M + \kappa \| f \|_\infty) \left( 1 + \sum_{i=1}^n \alpha_i \right).
\]

Proof: Note the fact that \( E \xi = L_k \left( f_{\alpha} - f_j \right) \). Using Lemma 3.1 with \( \alpha = \nu = \infty \), for \( j < i \) there have

\[
E(\xi(z), \xi(z)) \leq E(\xi(z), E \xi(z)) + 15 \alpha (M^n, M^n) \| \xi(z) \|_\infty \| \xi(z) \|_\infty \\
\leq \| \xi(z) \|_\infty \| f_{\alpha} - f_j \|_\infty + 15 \alpha \| \xi(z) \|_\infty \| f_{\alpha} - f_j \|_\infty
\]

Let \( \Delta = E \left| \frac{1}{n} \sum_{i=1}^n \xi(z_i) - L_k \left( f_{\alpha} - f_j \right) \right| \), Then direct computation leads to

\[
\Delta \leq \frac{1}{n} \| \xi \|_\infty + \frac{30 \kappa}{n} \sum_{i=1}^n \alpha_i \| f \|_\infty
\]

It suffices to estimate \( \| \xi \|_\infty \) and \( \| f \|_\infty \). By the definition of \( f_\alpha \), \( E(y - f_\alpha(x)) \) can be bounded as follow.

\[
E(y - f_\alpha(x)) \leq \inf_{f_\alpha \in H_k} E(y - f_\alpha(x)) + \Delta \| f \|_\infty \leq E \| y \|_\infty \leq M^2
\]

which implies

\[
\| \xi \|_\infty = E \left\{ (y - f_\alpha(x)) K(x,x) \right\} \leq \kappa^2 E(y - f_\alpha(x)) \leq \kappa^2 M^2
\]

and

\[
\| f \|_\infty = \left\| y - f_\alpha(x) \right\| \leq \kappa (M + \kappa \| f \|_\infty)
\]

Plugging the estimates (9) and (10) into (8), this completes the proof of Lemma 3.2.

**Lemma 3.3.** For any \( \alpha \) – mixing sequence \( \{ x'_j \} \), there holds
\[
E \left| f_k - \frac{1}{n} S^*_n S_n \right| \leq E \left| f_k - \frac{1}{n} S^*_n S_n \right| \leq \frac{K^4}{n} \left( 1 + 30 \sum_{i=1}^{n} \alpha_i \right)
\]

The above two lemmas are Lemma 4.2 and Lemma 5.1 in reference [15].

### 4. ERROR BOUND FOR THE DISTRIBUTED LEARNING ALGORITHM

To estimate \( E \left\| \overline{f}_n - f_i \right\| \), the error can be divided into the approximation error and the sample error by the regularization function \( f_{\lambda} \),

\[
E \left\| \overline{f}_n - f_i \right\| \leq 2E \left\| \overline{f}_n - f_{\lambda} \right\| + 2E \left\| f_{\lambda} - f_i \right\|.
\]

The approximation error has been estimated in the literature [17],

\[
\left\| f_{\lambda} - f_i \right\| \leq \lambda^{-\frac{1}{2}} \left\| \overline{f}_n \right\|.
\]

Dependent sampling in this paper is isometric sampling, what we need to do in distributed learning with dependent samples is to reduce the influence of \( \alpha \)– mixing coefficient on sample error estimation. When \( i \neq j \), it can be denoted as

\[
\sum_{i,j \in OS, i \neq j} \alpha_{ij} = \sum_{i \in odd} \alpha_{i(i-1)}
\]

(13)

Next, it is time to estimate \( E \left\| \overline{f}_n - f_i \right\| \). Let \( O_a = \{ i : i \text{ is an odd, } 1 \leq i \leq m \} \), \( E_a = \{ i : i \text{ is an even, } 1 \leq i \leq m \} \). Thus, it can be deduced that

\[
E \left\| \overline{f}_n - f_i \right\| \leq 2E \frac{1}{m} \sum_{j=1}^{m} \left( f_{D_{i,j}} - f_i \right) \right\| + 2E \frac{1}{m} \sum_{j=1}^{m} \left( f_{D_{i,j}} - f_i \right)
\]

\[
= \frac{2}{m^2} \sum_{j=1}^{m} E \left\| f_{D_{i,j}} - f_i \right\| + 2 \sum_{j=1}^{m} E \left\| f_{D_{i,j}} - f_i, f_{D_{i,j}} - f_i \right\|
\]

\[
+ \frac{2}{m^2} \sum_{j=1}^{m} E \left\| f_{D_{i,j}} - f_i, f_{D_{i,j}} - f_i \right\|
\]

(14)

The following decomposition of the error \( f_{D_{i,j}} - f_i \) has been proved in [16],

\[
f_{D_{i,j}} - f_i = \left( \frac{1}{n} S^*_n S_n + \lambda I \right) \left( \frac{1}{n} \sum_{j=1}^{n} \xi(z) - E \xi \right)
\]

Here \( \xi(z) = (y - f_i(x)) K \) is the random variable valued in \( H_{\lambda} \). By the decomposition of \( f_{D_{i,j}} - f_i \), and \( \left\| f_i \right\| \leq \left\| f_n \right\| \), then there have

\[
E \left\| f_{D_{i,j}} - f_i \right\| \leq \lambda^{-\frac{1}{2}} \left\| \sum_{j=1}^{n} \xi(z) - E \xi \right\|
\]

\[
\left\| f_{D_{i,j}} - f_i \right\| \leq \lambda^{-\frac{1}{2}} \left( \sum_{j=1}^{n} \xi(z) - E \xi \right)
\]

Here \( b_i = (\kappa + 1) \left( M + \kappa \left\| \overline{f}_n \right\| \right) \) by that \( 0 < \lambda \leq \kappa \).

By Lemma 1, for any fixed \( \delta = 0 \), the second and third term in the right hand of Equation (14) could be bounded as

\[
E \left\{ f_{D_{i,j}} - f_i, f_{D_{i,j}} - f_i \right\}
\]

\[
\leq \left\{ f_{D_{i,j}} - f_i, E f_{D_{i,j}} - f_i \right\} + 15 \alpha_{ij} \left\| f_{D_{i,j}} - f_i \right\| \leq \frac{1}{2} \left( \left\| f_{D_{i,j}} - f_i \right\| + \left\| E f_{D_{i,j}} - f_i \right\| \right) + 15 \alpha_{ij} \left\| f_{D_{i,j}} - f_i \right\|^{\lambda^{-1}} b_i
\]
Here \( \frac{1}{n} \sum_{i=1}^{n} \xi(z) - E\xi \) is denoted by \( \Delta \). Plugging the above estimate and (13) into (14), it can be deduced that

\[
E\left[ f_\Delta - f_k \right] \leq \frac{60b^2}{m\lambda^2} \left( 1 + \frac{\sum_{j=1}^{n} \alpha_j}{\lambda} \right) + \frac{4}{m} \sum_{i=1}^{n} E[f_{D,i} - f_k] + \frac{60b^2}{m\lambda^2} \sum_{i=1}^{n} \alpha_{23-\beta}.
\]  

(15)

Now it is turn to estimate \( \|E_{D,i} - f_k\| \). Applying the expression

\[
f_{D,i} = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} y_i K_{i,i}.
\]

There holds

\[
E[f_{D,i} - f_k] = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \frac{1}{n} S^*_o S_o f_o,
\]

By the facts that

\[
f_i = (L_x + \lambda I)^{-1} L_k f_o, \quad E\left( \frac{1}{n} S^*_o S_o \right) = L_x.
\]

Then

\[
E[f_{D,i} - f_k] = \left[ \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \left( \frac{1}{n} S^*_o S_o f_o - \frac{1}{n} S^*_o S_o f_o \right) \right] = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \left( \left( \frac{1}{n} S^*_o S_o \right) (f_i - f_o) + L_x (f_i - f_o) \right) = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \left( \left( \frac{1}{n} S^*_o S_o \right) (f_i - f_o) \right) = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \left( \left( L_x - \frac{1}{n} S^*_o S_o \right) (f_i - f_o) \right) = \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \left( \left( L_x - \frac{1}{n} S^*_o S_o \right) \left( \frac{1}{n} S^*_o S_o + \lambda I \right)^{-1} \right) \left( L_x - \frac{1}{n} S^*_o S_o \right) (f_i - f_o)
\]

Thus

\[
\|E[f_{D,i} - f_k]\| \leq \lambda^{2\beta-1} \|\|f_i - f_o\|\| \left[ \left( \frac{1}{n} S^*_o S_o \right) \right]^{1/2}.
\]  

(16)

Plugging (16) into (15), and by Lemma 2 and Lemma 3, then (6) holds true. This complete the proof of Theorem 1.

5. LEARNING RATES AND DISCUSSIONS

This section presents the learning rate of the distributed learning with dependent samples. Using the fact that

\[
\sum_{i=t}^{n} l_i = \begin{cases} O(m^{1-t}), & 0 < t < 1; \\ O(\log m), & t = 1; \\ O(1), & t > 1. 
\end{cases}
\]

Since \( f_o \in H_s \), there just need to analyze the \( \beta > \frac{1}{2} \) case. In addition, in order to achieve a fast learning rate, For all the cases, by taking \( \delta = \alpha \), the results follow by direct computation. Moreover, careful computation shows that the \( \log \) term in the corollary 1 may be dropped in case of \( t > 1 \). It can be described the learning rate as the following three aspects.
When \( t > 1 \), under the assumption of Theorem 1 and Corollary 1, by taking 
\[
t_0 = \min \left\{ \frac{1-s}{2}, \frac{t}{2\beta + 1} \right\},
\]
there have
\[
E \left| f_D - f^\star \right| \leq \begin{cases} 
O \left( N^{\frac{\beta-1}{2\beta+1}} \right), & 0 < s \leq \frac{2\beta-1}{2\beta+1}; \\
O \left( N^{\frac{1+\beta(2\beta+1)}{2\beta+1}} \right), & \frac{2\beta-1}{2\beta+1} < s < 1.
\end{cases}
\]
(17)

When \( t = 1 \), it is similar to the proof of the case \( t > 1 \), there have
\[
E \left| f_D - f^\star \right| = \begin{cases} 
O \left( N^{\frac{\beta-1}{2\beta+1}} \log N \right), & 0 < s \leq \frac{2\beta-1}{2\beta+1}; \\
O \left( N^{\frac{1+\beta(2\beta+1)}{2\beta+1}} \log N \right), & \frac{2\beta-1}{2\beta+1} < s < 1.
\end{cases}
\]
(18)

When \( 0 < t < 1 \), by taking 
\[
t_0 = \min \left\{ \frac{t(1-s)}{2}, \frac{t}{2\beta+1} \right\},
\]
Then there have
\[
E \left| f_D - f^\star \right| \leq \begin{cases} 
O \left( N^{\frac{\beta(1-s)}{2\beta+1}} \right), & 0 < s \leq \frac{2\beta-1}{2\beta+1}; \\
O \left( N^{\frac{1+\beta(2\beta+1)}{2\beta+1}} \right), & \frac{2\beta-1}{2\beta+1} < s < 1.
\end{cases}
\]
(19)

Comparing (17) (18) and (19) with the learning rate for \( \frac{1}{2} \leq \beta \leq 1 \) that have been proved in [15], the learning rate is a little worse.

6. CONCLUSION
Although the learning rate derived in this paper is a little worse compared with the learning rate when distributed kernel ridge regression learning is applied to identically and independently sampling, but independence is a restrictive assumption and our analysis reveals the relation between the dependence of sampling and the consistency of distributed kernel ridge regression learning, this is our main contribution.

ACKNOWLEDGEMENTS
The work described in this paper is supported by National Natural Science Foundation of China (Grants No 11671171). The authors would like to thank the referees for constructive suggestions.

REFERENCES
[1] Athreya, K., and Pantula, S. 1986. Mixing properties of harris chains and autoregressive processes. Journal of Applied Probability, vol. 23, no. 4, 880-892.
[2] Aronszajn, N. 1950. Theory of reproducing kernels. Transactions of the American Mathematical Society, vol. 68, no. 3, 337-404.
[3] Blanchard, G., and Mücke, N. 2016. Parallelizing spectral algorithms for kernel learning, Statistics, vol. 2.
[4] Bauer, F., Pereverzev, S., and Rosasco, L. 2007. On regularization algorithms in learning theory, J. Complex, vol. 23, no. 1, 52-72.
[5] Chanda, K. 1974. Strong mixing properties of linear stochastic processes, Journal of Applied Probability, vol. 11, no. 2, 401-408.

[6] Chang, X. Y., Lin, S. B., and Zhou, D. X. 2017. Distributed semi-supervised learning with kernel ridge regression. Journal of Machine Learning Research, vol. 18, no. 46, 1-22.

[7] Caponnetto, A., and DeVito, E. 2007. Optimal rates for the regularized least squares algorithm. Foundations of Computational Mathematics, vol. 7, no. 3, 331-368.

[8] Davydov, U. A. 1970. The invariance principle for stationary processes. Theory of Probability and its Applications, vol. 15, no. 3, 487-498.

[9] Dehling, H., and Philipp, W. 1982. Almost sure invariance principles for weakly dependent vector-valued random variables. The Annals of Probability, vol. 10, no. 3, 689-701.

[10] Guo, Z. C., Lin, S. B., and Zhou, D. C. 2017. Learning theory of distributed spectral algorithm. Inverse Problems, vol. 33, no. 7, 074009.

[11] Guo, Z. C., Shi, L., and Wu, Q. 2017. Learning theory of distributed regression with bias-corrected regularization kernel network. Journal of Machine Learning Research, vol. 18, no. 99-123, 1-25.

[12] Li, L. Q., and Wan, C. G., 2006. Support vector machines with beta-mixing input sequences. Advances in Neural Networks - ISNN, vol. 3971, 928-935.

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

[14] Shamir, O., and Srebro, N. 2014. Distributed stochastic optimization and learning. 52nd Annual Allerton Conference on Communication, Control and Computing, 850-857.

[15] Sun, H. W., Wu, Q., 2010. Regularized least square regression with dependent samples. Advances in Computational Mathematics, vol. 32, no. 2, 175-189.

[16] Smale, S., and Zhou, D. X. 2007. Learning theory estimates via integral operators and their approximation. Constructive Approximation, vol. 26, no. 2, 153-172.

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

[18] Sun, H. W., and Wu, Q. 2009. A note on application of integral operator in learning theory. Applied and Computational Harmonic Analysis, vol. 26, no. 3, 416-421.

[19] Vidyasagar, M., 2003. Learning and generalization with applications to neural networks. Springer Berlin Heidelberg New York.

[20] Withers, C. S. 2000. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. Neural Networks, vol. 3, no. 5, 535-549.

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

[22] Xu, Y. L., and Chen, D. R. 2008. Learning rates of regularized regression for exponentially strongly mixing sequence. Journal of Statistical Planning and Inference, vol. 138, no. 7, 2180-2189.