Journal of Machine Learning Research 18 (2017) 1-22 Submitted 11/16; Revised 4/17; Published 4/17

Distributed Semi-supervised Learning with Kernel Ridge Regression

Xiangyu Chang XIANGYUCHANG@GMAIL.COM
Center of Data Science and Information Quality
School of Management
Xi'an Jiaotong University, Xi'an, China
Shao-Bo Lin[*] SBLIN1983@GMAIL.COM
Department of Statistics
Wenzhou University, Wenzhou, China
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: Andreas Christmann

Abstract

This paper provides error analysis for distributed semi-supervised learning with kernel ridge regression (DSKRR) based on a divide-and-conquer strategy. DSKRR applies kernel ridge regression (KRR) to data subsets that are distributively stored on multiple servers to produce individual output functions, and then takes a weighted average of the individual output functions as a final estimator. Using a novel error decomposition which divides the generalization error of DSKRR into the approximation error, sample error and distributed error, we find that the sample error and distributed error reflect the power and limitation of DSKRR, compared with KRR processing the whole data. Thus a small distributed error provides a large range of the number of data subsets to guarantee a small generalization error. Our results show that unlabeled data play important roles in reducing the distributed error and enlarging the number of data subsets in DSKRR. Our analysis also applies to the case when the regression function is out of the reproducing kernel Hilbert space. Numerical experiments including toy simulations and a music-prediction task are employed to demonstrate our theoretical statements and show the power of unlabeled data in distributed learning.

Keywords: learning theory, distributed learning, kernel ridge regression, semi-supervised learning, unlabeled data, error decomposition

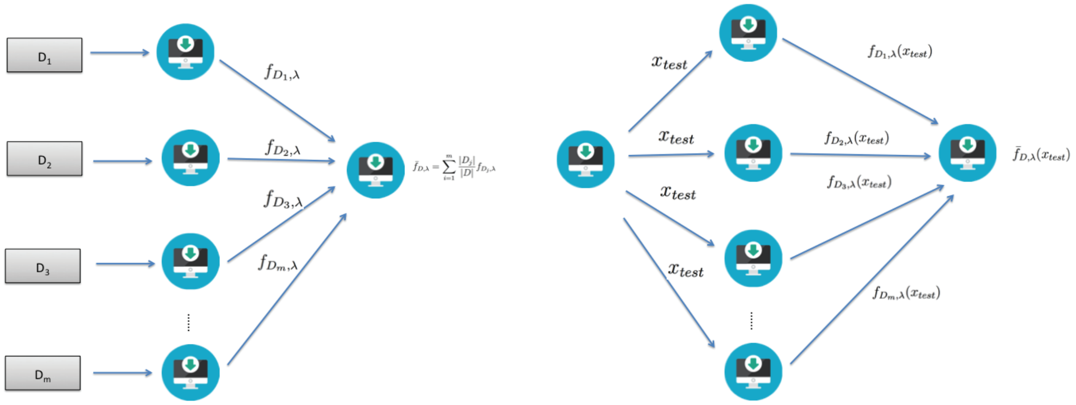
1. Introduction

Data from practical applications in medicine, finance, business and other fields are often stored distributively across multiple servers (called local processors) and may not be shared for reasons of preserving privacy. This requires privacy-preserving machine learning algorithms to discover population-wide patterns of the data without revealing any individual's sensitive information. Distributed learning (Balcan et al., 2012; Yan et al., 2013; Li et al.,

[*]. Corresponding author

2015; Xie et al., 2016) provides a promising way to tackle privacy-preserving learning problems.

Let m be the number of local processors and $D_j = \{(x_{i,j}, y_{i,j})\}_{i=1}^{|D_j|}$ be the data subset stored in the j-th local processor with $1 \leq j \leq m$ and $D = \bigcup_{j=1}^{m} D_j$ be the disjoint union of $\{D_j\}_{j=1}^{m}$, where $|D_j|$ denotes the cardinality of D_j. Distributed learning firstly gets an estimator $f_{D_j,\lambda}$ on each local processor based on D_j and some (regularization) parameter λ, and then launches the individual estimators to a central processor to get a final estimator $\overline{f}_{D,\lambda}$. To get an estimation for a new query point x_{test}, due to the purpose of privacy-preserving, x_{test} is firstly transmitted to each local processor to get an estimation $f_{D_j,\lambda}(x_{test})$, then all $f_{D_j,\lambda}(x_{test})$ with $1 \leq j \leq m$ are communicated to the central processor to synthesize the final estimation $\overline{f}_{D,\lambda}(x_{test})$. Flows of training and testing for distributed learning are exhibited in Figure 1. Distributed learning can thus be a privacy-preserving strategy in the sense that except for a real number $f_{D_j,\lambda}(x_{test})$, the individual's data information in each local processor is unknown to each other. The aim of statistics and learning theory is to verify that $\overline{f}_{D,\lambda}(x_{test})$ is a good approximation of the unknown but definite output of x_{text}.


D₁
D₂
D₃
Dm
$f_{D_1,\lambda}$
$f_{D_2,\lambda}$
$f_{D_3,\lambda}$
$f_{D_m,\lambda}$
$\bar{f}_{D,\lambda} = \sum_{i=1}^{m} \frac{|D_j|}{|D|} f_{D_j,\lambda}$
x_{test}
x_{test}
x_{test}
x_{test}
$f_{D_1,\lambda}(x_{test})$
$f_{D_2,\lambda}(x_{test})$
$f_{D_3,\lambda}(x_{test})$
$f_{D_m,\lambda}(x_{test})$
$\bar{f}_{D,\lambda}(x_{test})$

(a) Flow of training

(b) Flow of testing

Figure 1: Training and testing of distributed learning for privacy-preserving

In this paper, we focus on the distributed kernel ridge regression (DKRR). Let $(\mathcal{H}_K, \|\cdot\|_K)$ be the reproduced kernel Hilbert space (RKHS) induced by a Mercer kernel K on a metric (input) space $\mathcal{X}$. In a standard setting (Mann et al., 2009; Zhang et al., 2013; Lin et al., 2016), DKRR is defined with a regularization parameter $\lambda > 0$ by

$$\overline{f}_{D,\lambda} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} f_{D_j,\lambda}, \tag{1}$$

where

$$f_{D_j,\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D_j|} \sum_{(x,y) \in D_j} (f(x) - y)^2 + \lambda \|f\|_K^2 \right\}, \qquad j = 1, \ldots, m. \tag{2}$$

The optimal learning rate for DKRR was presented in (Zhang et al., 2015) under some eigenfunction assumptions which were removed in (Lin et al., 2016) by using a novel integral operator approach. In these existing results on rigorous analysis for DKRR, there are still two critical problems which greatly hinder applications of DKRR in practice. One is that the optimal learning rate is built upon a strict condition bounding the number of local processors, making DKRR infeasible for large m. The other is that the target function called regression function f_ρ is assumed to be in $\mathcal{H}_K$ to achieve the optimal learning rate, which is difficult to verify in practice.

The aim of the present paper is to consider distributed semi-supervised learning with kernel ridge regression (DSKRR) and demonstrate that using additional unlabeled data in a semi-supervised setting can overcome the aforementioned hurdles of DKRR. Let $D_j \cup \tilde{D}_j(x)$ be the subset of the data for semi-supervised learning which is stored on the j-th local processor, where $\tilde{D}_j(x) = \{x_1^j, \dots, x^j_{|\tilde{D}_j|}\}$. Based on $D_j \cup \tilde{D}_j(x), j = 1, \dots, m$, we construct a training set associated with $D_j \cup \tilde{D}_j(x)$ as

$$D_j^* = \{(x_i^*, y_i^*)\}_{i=1}^{|D_j^*|}$$

with

$$x_i^* = \begin{cases} x_i, & \text{if } (x_i, y_i) \in D_j, \\ \tilde{x}_i, & \text{if } \tilde{x}_i \in \tilde{D}_j(x), \end{cases} \qquad \text{and} \qquad y_i^* = \begin{cases} \frac{|D_j^*|}{|D_j|} y_i, & \text{if } (x_i, y_i) \in D_j, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

and denote $D^* = \bigcup_{j=1}^m D_j^*$. Define DSKRR by

$$\overline{f}_{D^*,\lambda} = \sum_{j=1}^m \frac{|D_j^*|}{|D^*|} f_{D_j^*,\lambda}. \tag{4}$$

We revise the standard integral operator framework (Caponnetto and De Vito, 2007; Smale and Zhou, 2007; Lin et al., 2016) to analyze the learning performance of algorithm (4). The main tool is a novel error decomposition that decomposes the generalization error of algorithm (4) into the approximation error, sample error and a new term called distributed error. The approximation error, which reflects the difference between a data-free limit of KRR and f_ρ, is standard. The sample error reflects the advantage of weighted averaging in (4) in the sense of scaling the generalization error of individual $f_{D_j^*,\lambda}$, $j = 1, 2, \dots, m$, with an additional factor $\frac{|D_j^*|}{|D^*|}$. The distributed error describes the difference between the distributed algorithm (4) and KRR processing the whole data in one single processor via presenting a restriction to the number of local processors. We find that additional unlabeled data play crucial roles in deducing a small distributed error and thus relaxing heavily the restriction on m to achieve the optimal learning rate for DKRR. We also prove that DSKRR leads to satisfactory estimates for the sample error when $f_\rho \notin \mathcal{H}_K$, which is beyond the standard setting with $f_\rho \in \mathcal{H}_K$ in (Zhang et al., 2015; Lin et al., 2016). Experimental studies are carried out to verify our theoretical analysis.

2. Theoretical Assessments

The generalization ability of DSKRR will be analyzed in a standard learning theory framework, in which the sample in D is assumed to be independently drawn from ρ, a Borel probability measure on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \mathbb{R}$ and $\tilde{D}(x) = \cup_{j=1}^m \tilde{D}_j(x)$ from ρ_X, the marginal distribution of ρ. Our purpose is to estimate how the estimator $\overline{f}_{D^*,\lambda}$ based on D^* approximates the regression function $f_\rho(x) := \int_{\mathcal{Y}} y d\rho(y|x)$ with $\rho(\cdot|x)$ being the conditional distribution of ρ induced at $x \in \mathcal{X}$. Throughout this paper, we assume $|y| \leq M$ almost surely for some constant $M > 0$ and $\mathcal{X}$ is compact, which implies $\|f_\rho\|_\infty \leq M$ almost surely and $\kappa := \sqrt{\sup_{x\in\mathcal{X}} K(x,x)} < \infty$.

As convergence may be as slow as one wants without imposing any restriction on the distribution ρ (Györfy et al., 2002), we introduce the following regularity assumption on f_ρ. Let L_K be the integral operator on $\mathcal{H}_K$ (or $L^2_{\rho_X}$ with norm $\|\cdot\|_\rho\|$) defined by

$$L_K f = \int_{\mathcal{X}} K_x f(x) d\rho_X,$$

where K_x is the function $K(\cdot, x)$ in $\mathcal{H}_K$. The regularity condition in this paper is

$$f_\rho = L_K^r h_\rho, \quad \text{for some } r > 0 \text{ and } h_\rho \in L^2_{\rho_X}, \tag{5}$$

where L_K^r denotes the r-th power of $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$, a compact and positive operator. This regularity condition is standard in learning theory and has been used in a large literature (Bauer et al., 2007; Caponnetto and De Vito, 2007; Smale and Zhou, 2007; Caponnetto and Yao, 2010; Shi et al., 2011; Guo et al., 2016; Hu et al., 2015; Lin and Zhou, 2016).

To derive fast learning rates, we should also present some restrictions on the capacity of $\mathcal{H}_K$. In this paper, we use the effective dimension $\mathcal{N}(\lambda)$ to measure the complexity of $\mathcal{H}_K$ with respect to ρ_X, which is defined to be the trace of the operator $(L_K + \lambda I)^{-1} L_K$, that is

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

We are in a position to present our main results, to be proved in Section 5. We first consider the traditional case of $f_\rho \in \mathcal{H}_K$ by imposing condition (5) with $r \geq 1/2$.

Theorem 1 *Assume $|y| \leq M$ almost surely and that (5) holds with $1/2 \leq r \leq 1$. We have*

$$\begin{aligned} &\max\left\{E\left[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2\right], \lambda E\left[\|\overline{f}_{D^*,\lambda} - f_\rho\|_K^2\right]\right\} \\ \leq\ & C\left[\lambda^{2r} + \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{\mathcal{A}^2_{D_j^*,\lambda}}{\lambda} + 1\right)^2 \left(\lambda^{2r-1}\mathcal{A}^2_{D_j^*,\lambda} + \frac{|D_j^*|}{|D^*|}\mathcal{A}^2_{D_j,\lambda}\right)\right], \end{aligned} \tag{6}$$

where C is a constant independent of m, $|D_j|$, $|D_j^*|$, or λ and

$$\mathcal{A}_{D,\lambda} = \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|}}. \tag{7}$$

To obtain explicit learning rates, we quantify the increment of $\mathcal{N}(\lambda)$ with a parameter $0 < s \leq 1$ and a constant $C_0 > 0$ as

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-s}, \qquad \forall \lambda > 0. \tag{8}$$

Condition (8) with $s = 1$ is always satisfied by taking $C_0 = \mathrm{Tr}(L_K) \leq \kappa^2$. When $0 < s < 1$, condition (8) is more general than the eigenvalue decaying assumption in the literature (Caponnetto and De Vito, 2007; Steinwart et al., 2009). Based on Theorem 1 and condition (8), we derive the following optimal learning rate of algorithm (4).

Corollary 2 *Assume $|y| \leq M$ almost surely, (8) with $0 < s \leq 1$ and (5) with $1/2 \leq r \leq 1$. If $\lambda = |D|^{-\frac{1}{2r+s}}$, $|D_1| = |D_2| = \cdots = |D_m|$, $|D_1^*| = |D_2^*| = \cdots = |D_m^*|$, and*

$$m \leq \min\left\{|D|^{\frac{2r+2s-1}{2r+s}}, |D^*||D|^{\frac{-s-1}{2r+s}}\right\}, \tag{9}$$

then

$$E[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2] = \mathcal{O}\left(|D|^{-\frac{2r}{2r+s}}\right).$$

When $|D^*| = |D|$, the training set D^* defined by (3) for DSKRR is the same as D for DKRR. In such a situation (without using any unlabeled data), the output function $\overline{f}_{D^*,\lambda}$ produced by DSKRR algorithm (4) coincides with $\overline{f}_{D,\lambda}$ generated by DKRR algorithm (1), and the optimal learning rates stated in Corollary 2 are the same as those in (Lin et al., 2016; Guo et al., 2016) achieved under the restriction $m \leq |D|^{\frac{2r-1}{2r+s}}$ given by (9). In particular, for the special case of $r = 1/2$ (that is, $f_\rho \in \mathcal{H}_K$), the number of local processors $m = \mathcal{O}(1)$ does not increase as the sample size $|D|$ becomes large, which is very restrictive and limits the use of distributed learning.

Corollary 2 tells us that additional unlabeled data can be used to relax the above restriction on m. For the special case with $r = 1/2$ and $s \geq 1/2$, when $|D^*| = |D|^{1+\frac{1}{1+s}}$ with additional unlabel data of size $|D|^{1+\frac{1}{1+s}} - |D|$, Corollary 2 asserts that the output function $\overline{f}_{D^*,\lambda}$ produced by DSKRR algorithm (4) achieves the optimal learning rates under the restriction $m \leq |D|^{\frac{1}{1+s}}$. This allows the number of local processors to increase as the sample size $|D|$ does. Thus our analysis demonstrates the usage of additional unlabeled data in distributed learning, which is the first purpose of this paper.

The second purpose of this paper is to extend the range of r in (5) from $r \geq 1/2$ for the standard setting with $f_\rho \in \mathcal{H}_K$ to $0 < r < 1/2$ for considering the situation $f_\rho \notin \mathcal{H}_K$.

Theorem 3 *Assume $|y| \leq M$ almost surely and that (5) holds with $0 < r < 1/2$. Let $0 < \lambda \leq 1$. We have*

$$E\left[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2\right] \leq \overline{C}\left[\lambda^{2r} + \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{\mathcal{A}_{D_j^*,\lambda}^2}{\lambda} + 1\right)^2\left(\lambda^{2r-1}\mathcal{A}_{D_j^*,\lambda}^2 + \frac{|D_j^*|}{|D^*|}\mathcal{A}_{D_j,\lambda}^2\right)\right], \tag{10}$$

where $\overline{C}$ is a constant independent of m, $|D_j|$, $|D_j^*|$ or λ.

Theorem 3 yields the following optimal learning rate for algorithm (4) when $f_\rho \notin \mathcal{H}_K$, which has not been given in the literature of distributed learning (Zhang et al., 2015; Lin et al., 2016; Guo et al., 2016).

Corollary 4 *Assume $|y| \leq M$ almost surely, (8) with $0 < s \leq 1$ and (5) with $0 < r < 1/2$. If $r + s \geq 1/2$, $|D_1| = |D_2| = \cdots = |D_m|$, $|D_1^*| = |D_2^*| = \cdots = |D_m^*|$, $\lambda = |D|^{-\frac{1}{2r+s}}$, $|D^*| \geq |D|^{\frac{s+1}{2r+s}}$ and m satisfies (9), then*

$$E[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2] = \mathcal{O}\left(|D|^{-\frac{2r}{2r+s}}\right).$$

3. Related Work and Discussion

As a state-of-the-art strategy to reduce the computational burden for some specified algorithms, distributed and parallel computation has triggered enormous research activities in the statistical and machine learning communities (Gillick et al., 2006). Distributed learning with ridge regression (Zhang et al., 2013), gradient descent algorithms (Shamir and Srebro, 2014), online learning (Zinkevich et al., 2010) and spectral algorithms (Guo et al., 2016) were proposed and their learning performances have been observed in many practical applications to be as good as that of a big processor which could handle the whole data, provided the number of servers m is not very large. Some theoretical bounds for m were presented in (Zhang et al., 2015; Lin et al., 2016; Guo et al., 2016; Blanchard and Mücke, 2016), asserting that the range of m depends on the regularity of f_ρ, which is difficult to verify in practice. This key problem makes users select only a small m or take m as a parameter in the learning process.

Compared with these results, there are two novelties of our results in this paper, though adding unlabeled data in the learning process causes additional computations. On one hand, if one takes m as a parameter in the learning process, the data should be re-divided again and again and thus it requires a large amount of communications. Our result avoids these re-division and communications in the sense that for a large range of m, if enough unlabeled data are given, optimal learning rates can be achieved. On the other hand, for some applications, data of small size (e.g. data in hospitals) are stored distributively across a great number of processors and cannot be shared each other for preserving privacy. The existing results (Zhang et al., 2015; Lin et al., 2016; Guo et al., 2016; Blanchard and Mücke, 2016) cannot tackle these problems in the sense that m is out of the range for the quantity to guarantee the optimal learning rate. The result in this paper presents a possibility to conquer this problem, provided there are additional unlabeled data.

We then compare our results with those in two closely related papers (Zhang et al., 2015; Lin et al., 2016). The seminal work (Zhang et al., 2015) considered the learning performance of algorithm (1) when $r = \frac{1}{2}$, i.e., ($f_\rho \in \mathcal{H}_K$). Using a matrix decomposition approach, (Zhang et al., 2015) derived an optimal learning rate of order $\mathcal{O}(|D|^{-1/(s+1)})$ under the assumption that for some constants $k > 2$ and $A < \infty$, the normalized eigenfunctions $\{\phi_\ell\}_\ell$ of L_K in $L^2_{\rho_X}$ satisfy

$$\|\varphi_\ell\|^{2k}_{L^{2k}_{\rho_X}} = E\left[|\phi_\ell(x)|^{2k}\right] \leq A^{2k}, \qquad \forall \ell \in \mathbb{N}. \tag{11}$$

Condition (11) was removed in (Lin et al., 2016), by using a novel integral operator approach based on a second order decomposition of difference of operator inverses. However, the analysis in (Lin et al., 2016) works only for $r > 1/2$. In our Corollary 2, we show that the optimal learning rate for DSKRR can be achieved for all $\frac{1}{2} \leq r \leq 1$ without assuming condition (11), provided additional unlabeled data are used. At the first glance, our approach in algorithm (4) incurs additional computation due to the unlabeled data. However, it is important for privacy-preserving learning when the data are stored in m (fixed) servers with $m > N^{\frac{2r-1}{2r+s}}$ and cannot be shared. Our result in Corollary 4 is new since optimal learning rates for DKRR when $f_\rho \notin \mathcal{H}_K$ have not been provided in the existing literate of distributed learning (Zhang et al., 2015; Lin et al., 2016; Guo et al., 2016).

Unlabeled data exist widely due to the expensive cost of label evaluation. Originally, unlabeled data are considered to be non-informative and often given up. However, with deeper understanding of semi-supervised learning, researchers recognize that unlabeled data can be useful in some special applications (Zhu and Goldberg, 2009) such as manifold learning (Belkin et al., 2006). In learning theory, it was found in (Caponnetto and Yao, 2010) that using unlabeled data can overcome the limitation that the optimal learning rate for KRR is only achievable for $f_\rho \in \mathcal{H}_K$. It was shown there that optimal learning rates for spectral algorithms might be achieved even when $f_\rho \notin \mathcal{H}_K$, provided enough unlabeled data were added in the training process. Similar results have been deduced for kernel-based conjugate gradient descent in (Blanchard and Krämer, 2010; Blanchard and Mücke, 2016). Results in the present paper show that unlabeled data also benefit distributed learning algorithms by allowing more local processors while achieving optimal learning rates. Furthermore, using some ideas from (Caponnetto and Yao, 2010), we succeed in conquering the bottleneck that optimal learning rates for distributed learning algorithms are achievable only when $f_\rho \in \mathcal{H}_K$. The numerical experiments to be reported in the last section motivated the last two authors (Lin and Zhou, 2016) to study the use of unlabeled data in distributed kernel-based gradient descent algorithms. Our results in the present paper also motivated a recent work (Guo et al., 2017) on error analysis for distributed manifold regularization algorithms.

4. Error Decomposition

The main tool in our analysis is a novel error decomposition for DSKRR. For this purpose, we introduce data-free limit and semi-supervised learning version of $f_{D_j,\lambda}$ as

$$f_\lambda = \arg\min_{f\in\mathcal{H}_K}\left\{\int_{\mathcal{X}}(f(x)-f_\rho(x))^2 d\rho_X + \lambda\|f\|_K^2\right\}$$

and

$$f^\diamond_{D_j,\lambda} = E^*[f_{D_j,\lambda}] := E[f_{D_j,\lambda}|D_j(x)].$$

The following proposition gives the error decomposition, whose proof is given at the end of this section.

Proposition 5 *Let $\overline{f}_{D,\lambda}$ be defined by (1). We have*

$$\frac{1}{2}E[\|\overline{f}_{D,\lambda}-f_\rho\|_\rho^2] \leq \|f_\lambda - f_\rho\|_\rho^2 + \sum_{j=1}^m \frac{|D_j|^2}{|D|^2}E\left[\|f_{D_j,\lambda}-f_\lambda\|_\rho^2\right] + \sum_{j=1}^m \frac{|D_j|}{|D|}E\left[\left\|f^\diamond_{D_j,\lambda}-f_\lambda\right\|_\rho^2\right], \tag{12}$$

and if $f_\rho \in \mathcal{H}_K$,

$$\frac{1}{2}E[\|\overline{f}_{D,\lambda}-f_\rho\|_K^2] \leq \|f_\lambda-f_\rho\|_K^2+\sum_{j=1}^m \frac{|D_j|^2}{|D|^2}E\left[\|f_{D_j,\lambda}-f_\lambda\|_K^2\right]+\sum_{j=1}^m \frac{|D_j|}{|D|}E\left[\left\|f_{D_j,\lambda}^\diamond-f_\lambda\right\|_K^2\right]. \tag{13}$$

The three terms on the right-hand side of (12) (or (13)) are the approximation error, sample error and distributed error. The approximation error, independent of the sample, describes the approximation capability of f_λ. The sample error connects the synthesized estimator (1) with the estimator (2). Compared with the sample error of the estimator (2), $E\left[\|f_{D_j,\lambda}-f_\lambda\|_\rho^2\right]$, there is an additional $\frac{|D_j|}{|D|}$ in our error decomposition, which reflects the power of the weighted averaging in (1) and shows the reason why the distributed algorithm (1) possesses similar learning performances as KRR processing the whole data D. Since $E^*[y_i]=f_\rho(x_i)$, it is easy to check that $f_{D_j,\lambda}^\diamond$ is the estimator derived from KRR with the noise-free data $\{(x_i,f_\rho(x_i))\}_{(x_i,y_i)\in D_j}$. This implies

$$f_{D_j,\lambda}^\diamond = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{|D_j|}\sum_{(x,y)\in D_j}(f(x)-f_\rho(x))^2+\lambda\|f\|_K^2\right\}. \tag{14}$$

The distributed error presented in Proposition 5 measures the limitation of the distributed learning algorithm (1). Compared with the sample error $E\left[\|f_{D_j,\lambda}-f_\lambda\|_\rho^2\right]$ of estimator (2), the distributed error $E\left[\left\|f_{D_j,\lambda}^\diamond-f_\lambda\right\|_\rho^2\right]$ focuses on the noise-free data and therefore is smaller than $E\left[\|f_{D_j,\lambda}-f_\lambda\|_\rho^2\right]$. This makes algorithm (1) possess similar learning rates as that of KRR with the whole data D. However, since there are only $|D_j|$ samples, it is impossible to get upper bounds asymptotically as $|D|^{-2r/(2r+s)}$ for $E\left[\left\|f_{D_j,\lambda}^\diamond-f_\lambda\right\|_\rho^2\right]$ when m is large. Thus, a restriction on m to guarantee the optimal learning rate is necessary.

To deduce a wide range of m, we need a tight bound for the distributed error. Noting that $f_{D_j,\lambda}^\diamond$ is independent of the sample outputs, it motivates us to employ the unlabeled part $\tilde{D}_j(x), j=1,\ldots,m$ of the data set D_j^* in designing the algorithm (4). In the following, we combine the traditional integral operator approach (Smale and Zhou, 2004, 2005, 2007) with a recently developed second order decomposition of operator inverses (Lin et al., 2016; Guo et al., 2016) to derive a tight estimate for the distributed error of algorithm (4). Denote the sampling operator $S_D:\mathcal{H}_K\to\mathbb{R}^{|D|}$ (or $L_{\rho_X}^2\to\mathbb{R}^{|D|}$) by

$$S_D f := \{f(x_i)\}_{(x_i,y_i)\in D}.$$

Its adjoint $S_D^T:\mathbb{R}^{|D|}\to\mathcal{H}_K$ (or $\mathbb{R}^{|D|}\to L_{\rho_X}^2$) is given by

$$S_D^T\mathbf{c} := \frac{1}{|D|}\sum_{(x_i,y_i)\in D} c_i K_{x_i}, \qquad \mathbf{c}\in\mathbb{R}^{|D|}.$$

Let $L_{K,D}$ be the data-dependent approximation of L_K defined by

$$L_{K,D}f = S_D^T S_D f = \frac{1}{|D|}\sum_{(x,y)\in D} f(x)K_x.$$

Then it is easy to check (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) that

$$f_{D,\lambda} = (L_{K,D} + \lambda I)^{-1} S_D^T y_D, \qquad f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho \tag{15}$$

and

$$f^\diamond_{D,\lambda} = (L_{K,D} + \lambda I)^{-1} L_{K,D} f_\rho, \tag{16}$$

where I is the identity operator and $y_D := (y_1, \dots, y_{|D|})^T$. The following proposition presents error estimates for the sample and distributed errors.

Proposition 6 *Let $f_{D_j,\lambda}$, f_λ and $f^\diamond_{D_j,\lambda}$ be defined by (15) and (16). We have*

$$\max\{\|f_{D_j,\lambda} - f_\lambda\|_\rho, \sqrt{\lambda}\|f_{D_j,\lambda} - f_\lambda\|_K\} \leq \mathcal{Q}^2_{D_j,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j,\lambda}\|f_\lambda\|_K), \tag{17}$$

and

$$\max\{\|f^\diamond_{D_j,\lambda} - f_\lambda\|_\rho, \sqrt{\lambda}\|f^\diamond_{D_j,\lambda} - f_\lambda\|_K\} \leq \mathcal{Q}^2_{D_j,\lambda}\mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}, \tag{18}$$

where

$$\begin{aligned} \mathcal{P}_{D,\lambda} &:= \left\|(L_K + \lambda I)^{-1/2}(L_K f_\rho - S_D^T y_D)\right\|_K, \\ \mathcal{Q}_{D,\lambda} &:= \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1/2}\|, \\ \mathcal{S}_{D,\lambda} &:= \left\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D})\right\|, \end{aligned}$$

and with a bounded measurable function g on $\mathcal{X}$,

$$\mathcal{R}_{D,\lambda,g} := \left\|(L_K + \lambda I)^{-1/2}\left(\int_{\mathcal{X}} g(x) K_x d\rho_X - \frac{1}{|D|}\sum_{(x,y)\in D} g(x) K_x\right)\right\|_K.$$

Proof Since

$$\begin{aligned} & f_{D_j,\lambda} - f_\lambda = (L_{K,D_j} + \lambda I)^{-1} S_{D_j}^T y_{D_j} - (L_K + \lambda I)^{-1} L_K f_\rho \\ = \; & (L_{K,D_j} + \lambda I)^{-1}(S_{D_j}^T y_{D_j} - L_K f_\rho) + [(L_{K,D_j} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}] L_K f_\rho, \end{aligned}$$

and

$$\|f\|_\rho = \|L_K^{1/2} f\|_K = \|L_K^{1/2}(L_K + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2} f\|_K \leq \|(L_K + \lambda I)^{1/2} f\|_K$$

for any $f \in L^2_{\rho_X}$, it follows from $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for positive operators A, B that

$$\begin{aligned} & \max\left\{\|f_{D_j,\lambda} - f_\lambda\|_\rho, \sqrt{\lambda}\|f_{D_j,\lambda} - f_\lambda\|_K\right\} \\ \leq \; & \|(L_K + \lambda I)^{1/2}(L_{K,D_j} + \lambda I)^{-1}(S_{D_j}^T y_{D_j} - L_K f_\rho)\|_K \\ + \; & \|(L_K + \lambda I)^{1/2}[L_{K,D_j} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}] L_K f_\rho\|_K \\ \leq \; & \mathcal{Q}_{D_j,\lambda}\|(L_{K,D_j} + \lambda I)^{-1/2}(S_{D_j}^T y_{D_j} - L_K f_\rho)\|_K \\ + \; & \|(L_K + \lambda I)^{1/2}(L_{K,D_j} + \lambda I)^{-1}(L_K - L_{K,D_j})(L_K + \lambda I)^{-1} L_K f_\rho\|_K \\ \leq \; & \mathcal{Q}^2_{D_j,\lambda}\|(L_K + \lambda I)^{-1/2}(S_{D_j}^T y_{D_j} - L_K f_\rho)\|_K \\ + \; & \mathcal{Q}^2_{D_j,\lambda}\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D_j})\|\|(L_K + \lambda I)^{-1} L_K f_\rho\|_K \\ \leq \; & \mathcal{Q}^2_{D_j,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j,\lambda}\|f_\lambda\|_K). \end{aligned}$$

This completes the proof of (17). Now we turn to prove (18). Due to (15) and (16), we have

$$\begin{aligned}
& f^{\diamond}_{D_j,\lambda} - f_\lambda = (L_{K,D_j}+\lambda I)^{-1}L_{K,D_j}f_\rho - (L_K+\lambda I)^{-1}L_K f_\rho \\
= \ & (L_{K,D_j}+\lambda I)^{-1}(L_{K,D_j}-L_K)f_\rho + [L_{K,D_j}+\lambda I)^{-1} - (L_K+\lambda I)^{-1}]L_K f_\rho \\
= \ & (L_{K,D_j}+\lambda I)^{-1}(L_{K,D_j}-L_K)f_\rho + (L_{K,D_j}+\lambda I)^{-1}[L_K - L_{K,D_j}](L_K+\lambda I)^{-1}L_K f_\rho \\
= \ & -(L_{K,D_j}+\lambda I)^{-1}[L_K - L_{K,D_j}](f_\rho - f_\lambda).
\end{aligned}$$

Then

$$\begin{aligned}
& \max\{\|f^{\diamond}_{D_j,\lambda} - f_\lambda\|_\rho, \sqrt{\lambda}\|f^{\diamond}_{D_j,\lambda} - f_\lambda\|_K\} \\
\leq \ & \|(L_K+\lambda I)^{1/2}(L_{K,D_j}+\lambda I)^{-1}[L_K - L_{K,D_j}](f_\lambda - f_\rho)\|_K \\
\leq \ & \mathcal{Q}_{D_j,\lambda}\|(L_{K,D_j}+\lambda I)^{-1/2}(L_{K,D_j}-L_K)(f_\lambda - f_\rho)\|_K \leq \mathcal{Q}^2_{D_j,\lambda}\mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}.
\end{aligned}$$

This completes the proof of Proposition 6. ■

If $f_\rho \in \mathcal{H}_K$, we have

$$\mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho} = \|(L_K+\lambda I)^{-1/2}(L_{K,D_j}-L_K)(f_\lambda - f_\rho)\|_K \leq \mathcal{S}_{D_j,\lambda}\|f_\lambda - f_\rho\|_K.$$

This implies that the distributed error can be bounded by $\mathcal{Q}^2_{D_j,\lambda}\mathcal{S}_{D_j,\lambda}\|f_\lambda - f_\rho\|_K$. Comparing with the sample error estimate (17) for an individual local processor, there is an additional term $\|f_\lambda - f_\rho\|_K$ in the distributed error estimate, since $\mathcal{P}_{D_j,\lambda}$, and $\mathcal{S}_{D_j,\lambda}$ are asymptotically equal due to Lemma 9 in Section 5. This together with Lemma 8 below shows that the distributed error estimate is essentially smaller than the sample error estimate for local processors under (5) with $r > 1/2$, and presents the reason why DSKRR performs similarly as KRR on the whole data set D, provided m is not so large.

Recall the definitions of $\mathcal{Q}_{D_j,\lambda}$, $\mathcal{R}_{D_j,\lambda,f_\lambda - f_\rho}$ and $\mathcal{S}_{D_j,\lambda}$. These three quantities are independent of the outputs. Thus, the distributed error estimate decreases when additional unlabeled data are given. This explains why unlabeled data can enlarge the range of m to guarantee the optimal learning rates for DSKRR. On the other hand, it follows from the definition of $\mathcal{P}_{D,\lambda}$ that the sample error estimate depends heavily on the labels corresponding to unlabeled data. But (3) implies that for each fixed $1 \leq j \leq m$,

$$S^T_{D^*_j}y_{D^*_j} = \frac{1}{|D^*_j|}\sum_{(x^*,y^*)\in D^*_j} y^* K_{x^*} = \frac{1}{|D^*_j|}\frac{|D^*_j|}{|D_j|}\sum_{(x,y)\in D_j} yK_x = S^T_{D_j}y_{D_j}.$$

Then $\mathcal{P}_{D^*_j,\lambda} = \mathcal{P}_{D_j,\lambda}$ for each $1 \leq j \leq m$, which implies that the sample error estimate does not increase when the unlabeled data are added and shows the necessity of re-weighting of y in our definition in (3).

Thus, the unlabeled data in algorithm (4) can reduce the distributed error estimate without increasing the sample error estimate. The following proposition which can be deduced directly from Propositions 5 and 6, shows the detailed error decomposition for algorithm (4).

Proposition 7 *Let $\overline{f}_{D^*,\lambda}$ be defined by algorithm (4). If condition (5) holds with $0 < r \leq 1$, then for $1/2 \leq r \leq 1$,*

$$\begin{aligned}
&\frac{1}{2}\max\left\{E[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2], \lambda E[\|\overline{f}_{D^*,\lambda} - f_\rho\|_K^2]\right\} \\
\leq \quad & \max\{\|f_\lambda - f_\rho\|_\rho^2, \lambda\|f_\lambda - f_\rho\|_K^2\} + \sum_{j=1}^m \frac{|D_j^*|^2}{|D^*|^2} E\left[\mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2\right] \\
& + \sum_{j=1}^m \frac{|D_j^*|}{|D^*|} E\left[\mathcal{Q}^4_{D_j^*,\lambda}\mathcal{S}^2_{D_j^*,\lambda}\|f_\lambda - f_\rho\|_K^2\right] \qquad (19)
\end{aligned}$$

while for $0 < r < 1/2$,

$$\begin{aligned}
E[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2] \quad \leq \quad & \|f_\lambda - f_\rho\|_\rho^2 + \sum_{j=1}^m \frac{|D_j^*|^2}{|D^*|^2} E\left[\mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2\right] \\
+ \quad & \sum_{j=1}^m \frac{|D_j^*|}{|D^*|} E\left[\mathcal{Q}^4_{D_j^*,\lambda}\mathcal{R}^2_{D_j^*,\lambda,f_\lambda - f_\rho}\right]. \qquad (20)
\end{aligned}$$

We end this section by proving Proposition 5.

Proof of Proposition 5 For the sake of brevity, we only prove (12), since (13) can be derived by using the same method. Due to the triangle inequality and the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b > 0$, we have

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] \leq 2\|f_\lambda - f_\rho\|_\rho^2 + 2E[\|\overline{f}_{D,\lambda} - f_\lambda\|_\rho^2]. \qquad (21)$$

It follows from $\sum_{j=1}^m \frac{|D_j|}{|D|} = 1$ that

$$\begin{aligned}
\|\overline{f}_{D,\lambda} - f_\lambda\|_\rho^2 &= \left\|\sum_{j=1}^m \frac{|D_j|}{|D|}(f_{D_j,\lambda} - f_\lambda)\right\|_\rho^2 \\
= \quad & \sum_{j=1}^m \frac{|D_j|^2}{|D|^2}\|f_{D_j,\lambda} - f_\lambda\|_\rho^2 + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle f_{D_j,\lambda} - f_\lambda, \sum_{k\neq j}\frac{|D_k|}{|D|}(f_{D_k,\lambda} - f_\lambda)\right\rangle_\rho.
\end{aligned}$$

Taking expectations gives

$$\begin{aligned}
E\left[\|\overline{f}_{D,\lambda} - f_\lambda\|_\rho^2\right] &= \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\lambda\|_\rho^2\right] \\
+ \quad & \sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle E[f_{D_j,\lambda}] - f_\lambda, E[\overline{f}_{D,\lambda}] - f_\lambda - \frac{|D_j|}{|D|}\left(E[f_{D_j,\lambda}] - f_\lambda\right)\right\rangle_\rho.
\end{aligned}$$

But

$$\begin{aligned}
\sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle E_{D_j}[f_{D_j,\lambda}] - f_\lambda, E[\overline{f}_{D,\lambda}] - f_\lambda\right\rangle_\rho &= \left\langle E[\overline{f}_{D,\lambda}] - f_\lambda, E[\overline{f}_{D,\lambda}] - f_\lambda\right\rangle_\rho \\
&= \left\|E[\overline{f}_{D,\lambda}] - f_\lambda\right\|_\rho^2.
\end{aligned}$$

We see that $E\left[\|\overline{f}_{D,\lambda} - f_\lambda\|_\rho^2\right]$ equals

$$\sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\lambda\|_\rho^2\right] - \sum_{j=1}^m \frac{|D_j|^2}{|D|^2}\left\|E[f_{D_j,\lambda}] - f_\lambda\right\|_\rho^2 + \left\|E[\overline{f}_{D,\lambda}] - f_\lambda\right\|_\rho^2.$$

Furthermore, by the Schwarz inequality and $\sum_{j=1}^m \frac{|D_j|}{|D|} = 1$, we have

$$\|E[\overline{f}_{D,\lambda}] - f_\lambda\|_\rho^2 = \left\|\sum_{j=1}^m \frac{|D_j|}{|D|}\left(E[f_{D_j,\lambda}] - f_\lambda\right)\right\|_\rho^2 \le \sum_{j=1}^m \frac{|D_j|}{|D|}\left\|E[f_{D_j,\lambda}] - f_\lambda\right\|_\rho^2.$$

Then

$$E\left[\|\overline{f}_{D,\lambda}] - f_\lambda\|_\rho^2\right] \le \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\lambda\|_\rho^2\right] + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\|E[f_{D_j,\lambda}] - f_\lambda\right\|_\rho^2.$$

Inserting the above inequalities into (21), we find

$$\frac{1}{2}E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] \le \|f_\lambda - f_\rho\|_\rho^2 + \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\lambda\|_\rho^2\right] + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\|E[f_{D_j,\lambda}] - f_\lambda\right\|_\rho^2.$$

According to Jensen's inequality, we obtain

$$\left\|E[f_{D_j,\lambda}] - f_\lambda\right\|_\rho^2 \le E\left[\left\|f^\diamond_{D_j,\lambda} - f_\lambda\right\|_\rho^2\right],$$

which implies

$$\frac{1}{2}E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] \le \|f_\lambda - f_\rho\|_\rho^2 + \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\lambda\|_\rho^2\right] + \sum_{j=1}^m \frac{|D_j|}{|D|} E\left[\left\|f^\diamond_{D_j,\lambda} - f_\lambda\right\|_\rho^2\right].$$

This completes the proof of Proposition 5. ■

5. Proofs

According to Proposition 7, to prove Theorems 1 and 3, we only need to bound $\|f_\lambda - f_\rho\|_\rho$, $\|f_\lambda - f_\rho\|_K$, $\mathcal{Q}_{D_j^*,\lambda}$, $\mathcal{P}_{D_j,\lambda}$, $\mathcal{R}_{D_j^*,\lambda,g}$ and $\mathcal{S}_{D_j^*,\lambda}$. The following two lemmas present bounds for these quantities. The first one can be found in (Smale and Zhou, 2007), which estimates the approximation error of algorithm (4).

Lemma 8 *Assume (5) with* $0 < r \le 1$. *There holds*

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|h_\rho\|_\rho. \tag{22}$$

Furthermore, if $1/2 \le r \le 1$, *then we have*

$$\|f_\lambda - f_\rho\|_K \le \lambda^{r-1/2} \|h_\rho\|_\rho. \tag{23}$$

The second lemma presents bounds for $\mathcal{Q}_{D_j^*,\lambda}$, $\mathcal{P}_{D_j,\lambda}$, $\mathcal{R}_{D_j^*,\lambda,g}$ and $\mathcal{S}_{D_j^*,\lambda}$. In particular, (24) was proved in (Guo et al., 2016), (25) and (27) were given in (Lin et al., 2016) and (26) can be found in (Caponnetto and De Vito, 2007). Recall the quantity $\mathcal{A}_{D,\lambda}$ defined by (7).

Lemma 9 *Let D be a sample drawn independently according to ρ and $0 < \delta < 1$. If $|y| \leq M$ almost surely, then each of the following estimates holds with confidence at least $1-\delta$,*

$$\mathcal{Q}_{D,\lambda}^2 \leq 2\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+2, \tag{24}$$

$$\mathcal{S}_{D,\lambda} \leq 2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\log(2/\delta), \tag{25}$$

$$\mathcal{P}_{D,\lambda} \leq 2M(\kappa+1)\mathcal{A}_{D,\lambda}\log(2/\delta), \tag{26}$$

$$\mathcal{R}_{D,\lambda,g} \leq 2\|g\|_\infty(\kappa+1)\mathcal{A}_{D,\lambda}\log(2/\delta). \tag{27}$$

We can now use Proposition 7 and the above lemmas to prove our main results.

Proof of Theorem 1 As $r \geq 1/2$, Lemma 8 implies that

$$\max\left\{\|f_\lambda - f_\rho\|_\rho, \sqrt{\lambda}\|f_\lambda - f_\rho\|_K\right\} \leq \lambda^r\|h_\rho\|_\rho. \tag{28}$$

For an arbitrary fixed $j \in \{1,\dots,m\}$, it follows from Lemma 9 that there exist three subsets $\mathcal{Z}_{1,\delta}^{|D_j^*|}$, $\mathcal{Z}_{2,\delta}^{|D_j^*|}$ and $\mathcal{Z}_{3,\delta}^{|D_j^*|}$ of $\mathcal{Z}^{|D_j^*|}$ with measures at least $1-\delta/3$ such that for $D_j^* \in \mathcal{Z}_{1,\delta}^{|D_j^*|} \cap \mathcal{Z}_{2,\delta}^{|D_j^*|} \cap \mathcal{Z}_{3,\delta}^{|D_j^*|}$ there holds

$$\begin{aligned}
\mathcal{Q}_{D_j^*,\lambda}^2 &\leq 2\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+2,\\
\mathcal{S}_{D_j^*,\lambda} &\leq 2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}\log\frac{6}{\delta},\\
\mathcal{P}_{D_j,\lambda} &\leq 2M(\kappa+1)\mathcal{A}_{D_j,\lambda}\log\frac{6}{\delta}.
\end{aligned}$$

This together with

$$\|f_\lambda\|_K = \|(L_K+\lambda I)^{-1}L_K f_\rho\|_K \leq \|f_\rho\|_K \leq \|L_K^{r-1/2}\|\|L_K^{1/2}h_\rho\|_K \leq \kappa^{2r-1}\|h_\rho\|_\rho$$

yields that with confidence at least $1-\delta$, there holds

$$\begin{aligned}
&\mathcal{Q}_{D_j^*,\lambda}^2(\mathcal{P}_{D_j,\lambda}+\mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)\\
\leq\ & 16(\kappa+1)\log^3\frac{6}{\delta}\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\left[M\mathcal{A}_{D_j,\lambda}+\kappa^{2r}\|h_\rho\|_\rho\mathcal{A}_{D_j^*,\lambda}\right].
\end{aligned}$$

Using the probability to expectation formula

$$E[\xi] = \int_0^\infty \operatorname{Prob}\left[\xi > t\right]dt \tag{29}$$

for nonnegative random variables to $\xi_1 = \mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2$ and

$$\mathrm{Prob}\left[\xi_1 > u\right] = \mathrm{Prob}\left[\xi_1^{\frac{1}{2}} > u^{\frac{1}{2}}\right] \leq 6\exp\left\{-[16\mathcal{B}(D_j^*,\lambda)]^{-\frac{1}{3}} u^{\frac{1}{6}}\right\}$$

for $u \geq 256\log^6 6\mathcal{B}(D_j^*,\lambda)^2$, we have

$$\begin{aligned} & E\left[\mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2\right] \leq 256\log^6 6\mathcal{B}(D_j^*,\lambda)^2 \\ + \quad & 6\int_0^\infty \exp\left\{-[16\mathcal{B}(D_j^*,\lambda)]^{-\frac{1}{3}} u^{\frac{1}{6}}\right\} du = 256(6+\log^6 6)\mathcal{B}^2(D_j^*,\lambda)\int_0^\infty u^{6-1}\exp\{-u\}\,du, \end{aligned}$$

where

$$\mathcal{B}(D_j^*,\lambda) := (\kappa+1)\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]\left(M\mathcal{A}_{D_j,\lambda} + \kappa^{2r}\|h_\rho\|_\rho \mathcal{A}_{D_j^*,\lambda}\right). \tag{30}$$

Due to the expression $\int_0^\infty u^{d-1}\exp\{-u\}\,du = \Gamma(d)$ for Gamma functions with $d > 0$, we have

$$E\left[\mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda} + \mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2\right] \leq 256(6+\log^6 6)5!\mathcal{B}^2(D_j^*,\lambda). \tag{31}$$

According to Lemma 9, there exist two subsets $\mathcal{Z}^{|D_j^*|}_{1',\delta}$ and $\mathcal{Z}^{|D_j^*|}_{2',\delta}$ of $\mathcal{Z}^{|D_j^*|}$ with measures at least $1-\delta/2$ such that for $D_j^* \in \mathcal{Z}^{|D_j^*|}_{1',\delta} \cap \mathcal{Z}^{|D_j^*|}_{2',\delta}$ there holds

$$\begin{aligned} & \mathcal{Q}^2_{D_j^*,\lambda}\mathcal{S}_{D_j^*,\lambda}\|f_\lambda - f_\rho\|_K \\ \leq \quad & 4(\kappa^2+\kappa)\log\frac{4}{\delta}\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]\mathcal{A}_{D_j^*,\lambda}\|f_\lambda - f_\rho\|_K. \end{aligned} \tag{32}$$

Since $r \geq 1/2$, plugging (23) into (32), with confidence at least $1-\delta$, there holds

$$\begin{aligned} & \mathcal{Q}^2_{D_j^*,\lambda}\mathcal{S}_{D_j^*,\lambda}\|f_\lambda - f_\rho\|_K \\ \leq \quad & 4\kappa(\kappa+1)\lambda^{r-1/2}\|h_\rho\|_\rho\log\frac{4}{\delta}\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]\mathcal{A}_{D_j^*,\lambda}. \end{aligned}$$

The same method as that in deriving (31) yields

$$\begin{aligned} & E\left[\mathcal{Q}^4_{D_j^*,\lambda}\mathcal{S}^2_{D_j^*,\lambda}\|f_\lambda - f_\rho\|_K^2\right] \\ \leq \quad & 256\kappa^2(\kappa+1)^2(4+\log^6 4)5!\|h_\rho\|_\rho^2\lambda^{2r-1}\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^2\mathcal{A}^2_{D_j^*,\lambda}. \end{aligned} \tag{33}$$

Inserting (28), (31), (30) and (33) into (19), we obtain

$$\begin{aligned}
&\max\left\{E[\|\overline{f}_{D^*,\lambda}-f_\rho\|_\rho^2],\lambda E[\|\overline{f}_{D^*,\lambda}-f_\rho\|_K^2]\right\}\leq 2\lambda^{2r}\|h_\rho\|_\rho^2\\
&+512\kappa^2(\kappa+1)^2(4+\log^6 4)5!\|h_\rho\|_\rho^2\lambda^{2r-1}\sum_{j=1}^m\frac{|D_j^*|}{|D^*|}\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\mathcal{A}^2_{D_j^*,\lambda}\\
&+512(6+\log^6 6)5!(\kappa+1)^2\sum_{j=1}^m\frac{|D_j^*|^2}{|D^*|^2}\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\\
&\times\left[M\mathcal{A}_{D_j,\lambda}+\kappa^{2r}\|h_\rho\|_\rho\mathcal{A}_{D_j^*,\lambda}\right]^2.
\end{aligned}$$

Then (6) follows from $\mathcal{A}_{D_j^*,\lambda}\leq\mathcal{A}_{D_j,\lambda}$ and the above estimate with

$$C:=2\|h_\rho\|_\rho^2+512(6+\log^6 6)5!(\kappa+1)^2[(\kappa+1)^2+1]^2\max\left\{\kappa^2\|h_\rho\|_\rho^2,(M+\|h_\rho\|_\rho\kappa^r)^2\right\}.$$

This completes the proof of Theorem 1. ■

Proof of Corollary 2 Let $\lambda=|D|^{-\frac{1}{(2r+s)}}$. Since $r+s\geq r\geq 1/2$, we obtain from (7), (8) and $|D_1|=\cdots=|D_m|$ that

$$\mathcal{A}_{D_j,\lambda}\leq m|D|^{-\frac{2r+s-1/2}{2r+s}}+\sqrt{C_0m}|D|^{-\frac{r}{2r+s}}\qquad\forall j=1,\dots,m. \tag{34}$$

Since $|D_1^*|=\cdots=|D_m^*|$, we have

$$\mathcal{A}_{D_j^*,\lambda}\leq m|D^*|^{-1}|D|^{\frac{1}{4r+2s}}+\sqrt{C_0m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}},\qquad\forall\ j=1,\dots,m. \tag{35}$$

This implies

$$\lambda^{-1/2}\mathcal{A}_{D_j^*,\lambda}\leq m|D^*|^{-1}|D|^{\frac{2}{4r+2s}}+\sqrt{C_0m}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}}.$$

Due to (9), we have

$$\lambda^{-1/2}\mathcal{A}_{D_j^*,\lambda}\leq\sqrt{C_0}+1. \tag{36}$$

Plugging (34), (35), (36) and (9) into (6) and noticing $\frac{|D_j^*|}{|D^*|}=\frac{1}{m}$ and $m\leq|D|^{\frac{2r+2s-1}{2r+s}}$, we obtain

$$E[\|\overline{f}_{D^*,\lambda}-f_\rho\|_\rho^2]\leq C|D|^{-2r/(2r+s)}+8C(\sqrt{C_0}+2)^6|D|^{-2r/(2r+s)}.$$

This completes the proof of Corollary 2. ■

Proof of Theorem 3 The proof is almost the same as that of Theorem 1. The only difference is that when $0<r<1/2$, we have

$$\|f_\lambda\|_K=\|(L_K+\lambda I)^{-1}L_Kf_\rho\|_K\leq\|(L_K+\lambda I)^{-1}L_K^{1/2+r}\|\|L_K^{1/2}h_\rho\|_K\leq\lambda^{r-1/2}\|h_\rho\|_\rho \tag{37}$$

and

$$\|f_\lambda-f_\rho\|_\infty\leq\|f_\lambda\|_\infty+\|f_\rho\|_\infty\leq\kappa\|f_\lambda\|_K+M\leq M+\kappa\lambda^{r-1/2}\|h_\rho\|_\rho. \tag{38}$$

Since $0<r<1/2$, we also see from Lemma 8 that

$$\|f_\lambda-f_\rho\|_\rho\leq\lambda^r\|h_\rho\|_\rho. \tag{39}$$

It follows from (37), (27) and the same method as that for deriving (31) that

$$\begin{aligned} E\left[\mathcal{Q}^4_{D_j^*,\lambda}(\mathcal{P}_{D_j,\lambda}+\mathcal{S}_{D_j^*,\lambda}\|f_\lambda\|_K)^2\right] \quad &\leq \quad 512(6+\log^6 6)5!(\kappa+1)^2\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2 \\ &\times \quad \left[M\mathcal{A}_{D_j,\lambda}+\kappa\lambda^{r-1/2}\|h_\rho\|_\rho\mathcal{A}_{D_j^*,\lambda}\right]^2. \qquad (40) \end{aligned}$$

Combining (38) and (27) with the same method as that for deriving (33) yields

$$\begin{aligned} & E\left[\mathcal{Q}^4_{D_j^*,\lambda}\mathcal{R}^2_{D_j^*,\lambda,f_\lambda-f_\rho}\right] \qquad (41) \\ \leq \quad & 512(\kappa+1)^2(4+\log^6 4)5!(M+\kappa\lambda^{r-1/2}\|h_\rho\|_\rho)^2\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\mathcal{A}^2_{D_j^*,\lambda}. \end{aligned}$$

Inserting (39), (40) and (41) into (20), we have

$$\begin{aligned} & E\left[\|\overline{f}_{D^*,\lambda}-f_\rho\|_\rho^2\right] \leq 2\lambda^{2r}\|h_\rho\|_\rho^2+1024(\kappa+1)^2(4+\log^6 4)5!(M+\kappa\lambda^{r-1/2}\|h_\rho\|_\rho)^2 \\ & \times\sum_{j=1}^m\frac{|D_j^*|}{|D^*|}\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\mathcal{A}^2_{D_j^*,\lambda} \\ & +1024(6+\log^6 6)5!(\kappa+1)^2\sum_{j=1}^m\frac{|D_j^*|^2}{|D^*|^2}\left[\left(\frac{(\kappa^2+\kappa)\mathcal{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2 \\ & \times\left[M\mathcal{A}_{D_j,\lambda}+\kappa\lambda^{r-1/2}\|h_\rho\|_\rho\mathcal{A}_{D_j^*,\lambda}\right]^2. \end{aligned}$$

Together with the restrictions $0<\lambda\leq 1$ and $2r<1$, this yields (10) with

$$\overline{C}:=2\|h_\rho\|_\rho^2+1024(6+\log^6 6)5!(\|h_\rho\|_\rho+M)^2(\kappa+1)^4[(\kappa^2+\kappa)^2+1]^2.$$

This completes the proof of Theorem 3. ■

Proof of Corollary 4 Due to $r+s\geq 1/2$ and (9), we get (34), (35) and (36). Inserting (34), (35), (36) and (9) into (10), we obtain from $|D^*|\geq|D|^{\frac{1+s}{2r+s}}$ that

$$\begin{aligned} & E[\|\overline{f}_{D^*,\lambda}-f_\rho\|_\rho^2]\leq\overline{C}|D|^{-2r/(2r+s)} \\ + \quad & \overline{C}((\sqrt{C_0}+1)^2+1)^2\sum_{j=1}^m\frac{|D_j^*|}{|D|}\left[\frac{1}{m}\mathcal{A}^2_{D_j,\lambda}+\lambda^{2r-1}\mathcal{A}^2_{D_j^*,\lambda}\right] \\ \leq \quad & [\overline{C}+2\overline{C}((\sqrt{C_0}+1)^2+1)^2((1+\sqrt{C_0})^2)]|D|^{-\frac{2r}{2r+s}}, \end{aligned}$$

where we have used $|D^*|\geq|D|^{\frac{2r+s+1}{2r+s}}$ in the last inequality. This completes the proof of Corollary 4. ■

6. Experimental Verifications

In this section, we report experimental studies to justify the statements in Section 2. We employ two criteria for the comparison. The first criterion is the *global error* (GE) which is the mean square error of a testing set with $N = |D|$ examples used in the training flow. GE provides a baseline to assess the performance of DSKRR. The second criterion is the *average error* (AE) which is the mean square error of algorithm (4). Regularization parameters in all experiments are selected by the 5-fold cross-validation.

6.1 Toy simulations

In this part, we carry out three simulations to verify our theoretical statements. The first two simulations are devoted to verifying Corollaries 2 and 4, respectively. The last simulation focuses on the relation between the generalization performance of algorithm (4) and the size of unlabeled data to show the power of unlabeled data in distributed learning.

Simulation 1: We generate $N = 5000$ examples for training. The inputs $\{x_i\}_{i=1}^N$ are independently drawn according to the uniform distribution on the (hyper-)cube $[0, 1]^d$ with $d = 1$ or $d = 3$. The corresponding outputs are generated from the regression models $y_i = g_j(x_i)+\varepsilon_i, i = 1, \ldots, N, j = 1, 2$, where ε_i is the independent Gaussian noise $\mathcal{N}(0, 1/5)$,

$$g_1(x) := \left\{ \begin{array}{cc} x, & 0 < x \leq 0.5, \\ 1 - x, & 0.5 < x \leq 1, \end{array} \right. \tag{42}$$

and

$$g_2(x) := h_2(\|x\|_2) := \left\{ \begin{array}{cc} (1 - \|x\|_2)^6(35\|x\|_2^2 + 18\|x\|_2 + 3), & 0 < \|x\|_2 \leq 1, x \in \mathbb{R}^3, \\ 0, & \|x\|_2 > 1. \end{array} \right. \tag{43}$$

We also generate 500 test examples $\{(x'_i, y'_i)\}_{i=1}^{500}$ with $\{x'_i\}$ drawn independently according to the uniform distribution and $y'_i = g_j(x'_i), j = 1, 2$. The number m of local processors varies from 2 to 60. The $SN \in \{0, 2000, 4000, 6000, 8000, 10000\}$ unlabeled examples $\{\tilde{x}_i\}_{i=1}^{SN}$ are independently drawn according to the uniform distribution on the (hyper-)cube $[0, 1]^d$. It can be found in (Wu, 1995; Schaback and Wendland, 2006) that $g_1 \in W_1^1$ and $g_2 \in W_3^4$, where W_d^α denotes the α-order Sobolev space on $[0, 1]^d$. Furthermore, it is easy to see that $g_1 \notin W_1^2$ and $g_2 \notin W_3^5$. If we define $K_1(x, x') = 1+\min(x, x')$ and $K_2(x, x') = h_3(\|x-x'\|_2)$ with

$$g_3(x) := h_3(\|x\|_2) := \left\{ \begin{array}{cc} (1 - \|x\|_2)^4(4\|x\|_2 + 1), & 0 < \|x\|_2 \leq 1, x \in \mathbb{R}^3, \\ 0, & \|x\|_2 > 1, \end{array} \right. \tag{44}$$

then we know (Wu, 1995; Schaback and Wendland, 2006) that K_1 and K_2 are reproducing kernels for W_1^1 and W_3^2, respectively. Obviously, $g_1 \in \mathcal{H}_{K_1}$ and $g_2 \in \mathcal{H}_{K_2}$. The $N + SN$ sample points are evenly distributively stored in m local processors and algorithm (4) is applied to the training set. The testing results of GEs and AEs are recorded and shown in Figure 2.

In Figure 2, AE curves are recorded by different SN. When m is not too large, as shown in Figure 2, AEs are always comparable to GEs. Furthermore, there exists an upper bound of the number of local processors, m_{SN} (e.g., $m_{100,00} \approx 30$), lager than which AE curves

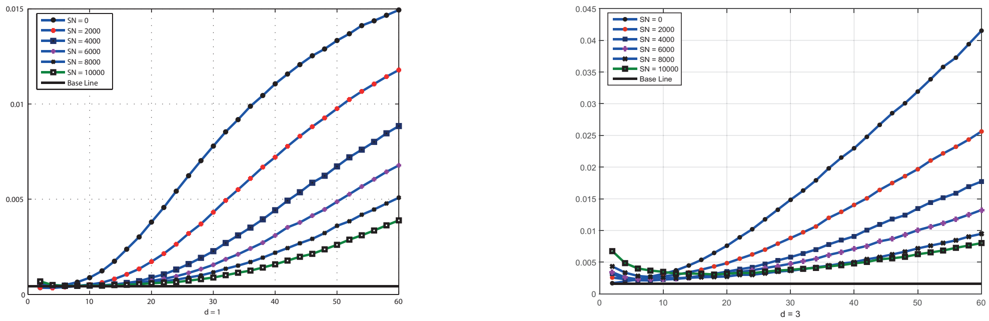


Figure 2: Performances of algorithm (4) with different scale of $|D^*|$ when $f_\rho \in \mathcal{H}_K$

increase dramatically. This result confirms Corollary 2 which indicates that DSKRR can achieve the optimal learning rate as long as m is not too large. Moreover, when we add more unlabeled data into the training set, i.e., increasing SN, the upper bound m_{SN} increases and AEs decrease (e.g., the AE curve for $SN = 100,00$ is almost below all the other curves). This shows the power of adding unlabeled data into training data, as condition (9) shows.

Simulation 2: We generate a training set from the regression model $y_i = g_3(x_i) + \varepsilon_i, i = 1, \ldots, N$ with $N = 5000$, where the inputs $\{x_i\}_{i=1}^N$ are independently drawn according to the uniform distribution on $[0,1]^3$. We also generate 500 test examples $\{(x_i', y_i')\}_{i=1}^{500}$ with x_i' drawn independently according to the uniform distribution and $y_i' = g_3(x_i')$. The number of servers m varies from 2 to 90 and the number of unlabeled examples is $SN \in \{0, 2000, 4000, 6000, 8000, 10000\}$. We utilize $K_3(x,x') = h_2(\|x - x'\|_2)$ as the kernel. It can be found in (Wu, 1995; Schaback and Wendland, 2006) that K_3 is a reproducing kernel for W_3^4. Obviously, $g_3 \notin \mathcal{H}_{K_3}$. GEs and AEs are recorded and shown in Figure 3.

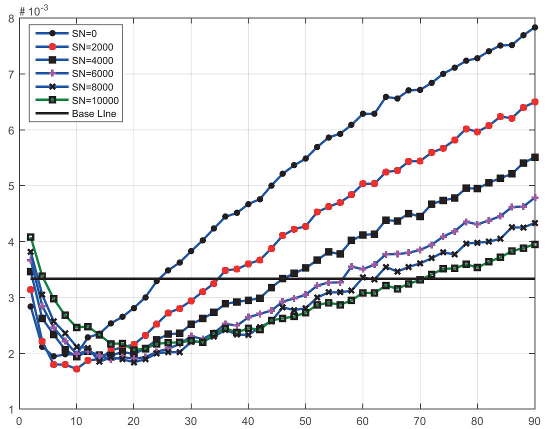


Figure 3: Performances of algorithm (4) with different scale of $|D^*|$ when $f_\rho \notin \mathcal{H}_K$

This simulation shows similar results as Simulation 1. The only difference is that the regression function $g_3 \notin \mathcal{H}_{K_3}$. This demonstrates the statement of Corollary 4, saying that even when the regression function is not in the RHKS, the optimal learning rate of algorithm (4) can be achieved. Based on Simulations 1 and 2, we find that adding unlabeled data into the training set can essentially improve the learning performance of algorithm (1).

Simulation 3: The experimental setting of this simulation is the same as that in Simulation 1. We generate $N = 500$ observations as a training set and 50 observations as a testing set. Here, we fix the size of local processors to be $m = 10$ and vary the number of unlabeled data SN from 50 to 10000. The aim is to describe the relation between the generalization ability of algorithm (4) and the size of unlabeled data. The simulation results are reported in Figure 4.

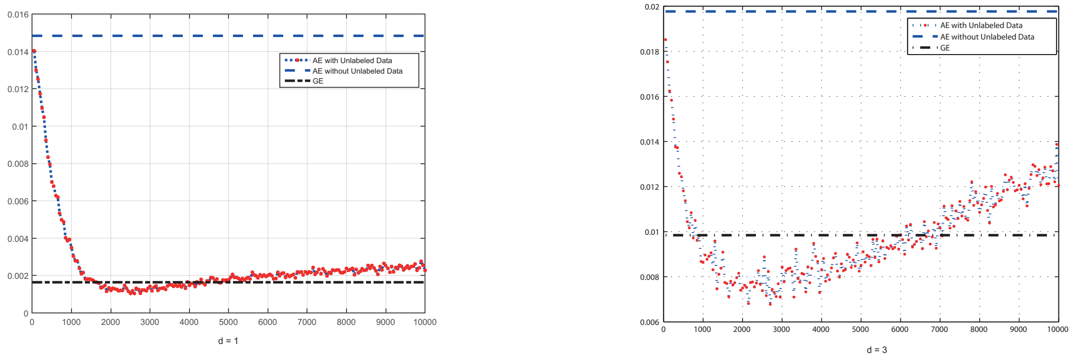


Figure 4: Relation between generalization error and the size of unlabeled data

In Figure 4, we construct two straight lines (black and blue lines) as base lines for comparisons. The black straight line is the GE, while the blue line is the AE value with $SN = 0$. We change the size of unlabeled data and apply algorithm (4) on the training set. It is shown in Figure 4 that the AE's curve decreases dramatically and can achieve the GE value when we add about $10N$ unlabeled data. However, when we add more unlabeled data into the training data, the AE curve begins to increase slowly. This phenomenon indicates two important observations. On one hand, adding unlabeled data into the training set can essentially improve the learning performance of DSKRR. On the other hand, if the size of unlabeled data is too large, the optimal learning rate has reasonable probabilities to be broken out. From our simulations, we suggest that $SN \leq 10N$. Here, we need to mention that for stabilizing the AE's curve, the simulation is repeated 10 times. The drawn curve is the average values of AE.

All these simulations verify our theoretical statements in Section 2 and show the power of unlabeled data in distributed semi-supervised learning.

6.2 Real data experiment

In this part, we focus on the Million Song data (Bertin-Mahieux et al., 2011) that describes a learning task of predicting the year in which a song is released based on audio features associated with the song. The dataset consists of $463,715$ training examples and $51,630$ test examples. Each example is a song released between 1922 and 2011, and the song is represented as a vector of timbre information computed about the song. Each sample point consists of a pair $(x_i, y_i) \in [0,1]^d \times [1922, 2011]$ with $d = 90$.

For the Million Song data, we normalize the feature vectors x so that the timbre signals have mean 0 and standard deviation 1. We also give a feature weight vector $W = (w_1, \ldots, w_d)^\top$ for setting $x_{ij} := w_j x_{ij}$. Here we choose $w_j = 1$ if $j \leq 12$ and $w_j = 0.2$ if $12 < j \leq 90$. Finally, we use the Gaussian kernel $K(x, x') = \exp\left\{-\frac{\|x-x'\|_2^2}{2\beta^2}\right\}$ in our experiments with bandwidth parameter $\beta = 6$ and regularization parameter $\lambda = N^{-1}/2$.

For each feature $X_j = (x_{1j}, \ldots, x_{Nj})^\top$, we denote $x_j^{min} = \min\{x_{1j}, \ldots, x_{Nj}\}$ and $x_j^{max} = \max\{x_{1j}, \ldots, x_{Nj}\}$. Then we generate the unlabeled data $\tilde{x}_{ij}, j = 1, \ldots, SN$, independently from the uniform distribution $U[x_j^{min}, x_j^{max}]$. It should be noted that the distributions for the labeled data and unlabeled data are different, making the learning task a so-called mismatch problem. Algorithm (4) is applied to the training examples with six partitions $m \in \{300, 500, 700, 900, 1100, 1300\}$. Finally, we plot AE curves of DSKRR with 3 different sizes of unlabeled samples in Figure 5. As exhibited in Figure 5, DSKRR has better performance when SN increases if $m \geq 300$. This phenomenon confirms the observation that adding unlabeled data into training examples can improve the order of m, as condition (9) shows.

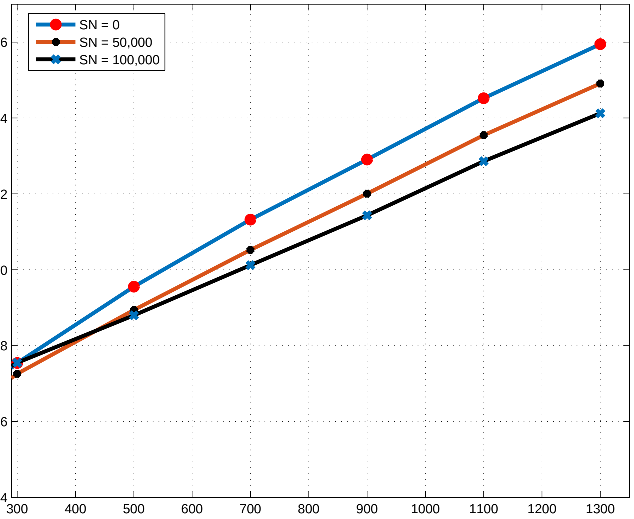


Figure 5: Performance of DSKRR with unlabeled data on Million Song data

Acknowledgments

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 11304114] and the National Natural Science Foundation of China (Grant Numbers 61502342, 11401462, 11461161006, 11471292). This paper was written when the last author visited Shanghai Jiaotong University (SJTU). The hospitality and sponsorships from SJTU and Ministry of Education are greatly appreciated. The authors would like to thank the referees for constructive suggestions and Dr. Yucheng Zhang for discussing with the numerical simulations.

References

M. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity, and privacy. *Conference on Learning Theory (COLT)*, 23: 26.1-26.22, 2011.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23: 52-72, 2007.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learningform labeled and unlabeled examples. *Journal of Machine Learning Research*, 7: 2339-2434, 2006.

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In Proceedings of the 12th International conference on Music Information Retrieval, 18, 2011.

G. Blanchard and N. Krämer. Optimal learning rates for kernel conjugate gradient regression. *Advances in Neural Information Processing Systems*, 226-234, 2010.

G. Blanchard and N. Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14: 763-794, 2016.

G. Blanchard and N. Mücke. Parallelizing spectral algorithms for kernel learning. arXiv preprint arXiv:1610.07487, 2016.

A. Caponnetto and E. DeVito. Optimal rates for the regularized least squares algorithm. *Foundations of Computational Mathematics*, 7: 331-368, 2007.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8: 161-183, 2010.

D. Gillick, A. Faria, and J. DeNero. Mapreduce: Distributed computing for machine learning. Berkley, December 18, 2006.

Z. C. Guo, S. B. Lin, and D. X. Zhou. Distributed learning with spectral algorithms. *Inverse Problems*, Minor revision under review, 2016.

Z. C. Guo, S. B. Lin and L. Shi. Distributed learning with multi-penalty regularization. Submitted, 2017.

L. Györfy, M. Kohler, A. Krzyzak, H. Walk. A Distribution-Free Theory of Nonparametric Regression. Springer-Verlag, Berlin, 2002.

T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13:437–455, 2015.

C. Li, P. Zhou, and T. Jiang. Differential privacy and distributed online learning for wireless big data. *Wireless Communications & Signal Processing (WCSP), 2015 International Conference on. IEEE*, 1-5, 2015.

S. B. Lin, X. Guo, and D. X. Zhou, Distributed learning with regularized least squres. *Journal of Machine Learning Research*, Revision under review, (arXiv 1608.03339), 2016.

S. B. Lin and D. X. Zhou, Distributed kernel gradient descent algorithms. *Constructive Approximation*, To appear.

G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in Neural Information Processing Systems*, 1231-1239, 2009.

R. Schaback and H. Wendland. Kernel techniques: from machine learning to meshless methods. *Acta Numerica*, 15: 543-639, 2006.

O. Shamir and N. Srebro. Distributed stochastic optimization and learning. *In 52nd Annual Allerton Conference on Communication, Control and Computing*, 2014.

L. Shi, Y. L. Feng and D. X. Zhou. Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31: 286-302, 2011.

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41: 279-305, 2004.

S. Smale and D. X. Zhou. Shannon sampling II: Connections to learning theory. *Appllied and Computational Harmonic Analysis*, 19: 285-302, 2005.

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

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *Conference on Learning Theory (COLT Dasgupta and A. Klivans, eds.)*, 22: 79-93, 2009.

Z. M. Wu. Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4: 283-292, 1995.

L. Xie, S. Plis, and A. D. Sarwate. Data-weighted ensemble learning for privacy-preserving distributed learning. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016.

F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transaction on Knowledge and Data Engineering*, 25: 2483-2493, 2013.

Y. C. Zhang, J. Duchi, and M. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14: 3321-3363, 2013.

Y. C. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16: 3299-3340, 2015.

M. Zinkevich, M. Weimer, A. Smola, and L. Li. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 2595-2603, 2010.

X. Zhu and A. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning,* 3: 1-130, 2009.