On the Feasibility of Distributed Kernel Regression for Big Data

Chen Xu[1] cux10@psu.edu
Yongquan Zhang[2] zyqmath@163.com
Runze Li[1] rli@stat.psu.edu

[1] The Methodology Center
The Pennsylvania State University
State College, PA, USA, 16801

[2] Department of Mathematics
China Jiliang University
Hangzhou, Zhejiang, China, 310018

May 6, 2015

Abstract

In modern scientific research, massive datasets with huge numbers of observations are frequently encountered. To facilitate the computational process, a divide-and-conquer scheme is often used for the analysis of big data. In such a strategy, a full dataset is first split into several manageable segments; the final output is then averaged from the individual outputs of the segments. Despite its popularity in practice, it remains largely unknown that whether such a distributive strategy provides valid theoretical inferences to the original data. In this paper, we address this fundamental issue for the distributed kernel regression (DKR), where the algorithmic feasibility is measured by the generalization performance of the resulting estimator. To justify DKR, a uniform convergence rate is needed for bounding the generalization error over the individual outputs, which brings new and challenging issues in the big data setup. Under mild conditions, we show that, with a proper number of segments, DKR leads to an estimator that is generalization consistent to the unknown regression function. The obtained results justify the method of DKR and shed light on the feasibility of using other distributed algorithms for processing big data. The promising preference of the method is supported by both simulation and real data examples.

Keywords: Distributed Algorithm, Kernel Regression, Big Data, Learning Theory, Generalization Bounds.

1 Introduction

The rapid development in data generation and acquisition has made a profound impact on knowledge discovery. Collecting data with unprecedented sizes and complexities is now feasible in many scientific fields. For example, a satellite takes thousands of high resolution images per day; a Walmart store has millions of transactions per week; and Facebook generates billions of posts per month. Such examples also occur in agriculture, geology, finance, marketing, bioinformatics, and Internet studies among others. The appearance of big data brings great opportunities for extracting new information and discovering subtle patterns. Meanwhile, their huge volume also poses many challenging issues

to the traditional data analysis, where a dataset is typically processed on a single machine. In particular, some severe challenges are from the computational aspect, where the storage bottleneck and algorithmic feasibility need to be faced. Designing effective and efficient analytic tools for big data has been a recent focus in the statistics and machine learning communities [24].

In the literature, several strategies have been proposed for processing big data. To overcome the storage bottleneck, *Hadoop* system was developed to conduct distributive storage and parallel processing. The idea of *Hadoop* follows from a natural divide-and-conquer framework, where a large problem is divided into several manageable subproblems and the final output is obtained by combining the corresponding sub-outputs. With the aid of *Hadoop*, many machine learning methods can be re-built to their distributed versions for the big data analysis. For examples, McDonald et al. [14] considered a distributed training approach for structured perception, while Kleiner et al. [10] introduced a distributed bootstrap method. Recently, similar ideas have also been applied to statistical point estimation [11], kernel ridge regression [28], matrix factorization [13], and principal component analysis [26].

To better understand the divide-and-conquer strategy, let us consider an illustrative example as follows. Suppose that a dataset consists of $N = 1,000,000$ random samples $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ with dimension $d = 100$. We assume that the data follow from a linear model $y_i = x_i^T\beta + \varepsilon$ with a random noise ε. The goal of learning is to estimate the regression coefficient β. Let $Y = (y_1, \ldots, y_N)^T$ be the N-dimensional response vector and $X = (x_1, \ldots, x_N)^T$ be the $N \times d$ covariate matrix. Apparently, the huge sample size of this problem makes the single-machine-based least squares estimate $\widehat{\beta} = (X^\top X)^{-1}X^\top Y$ computationally costly. Instead, one may first evenly distribute the N samples into m local machines and obtain m sub-estimates $\widehat{\beta}_j$ based on m independent running. The final estimate of β can then be obtained by averaging the m sub-estimates $\bar{\beta} = \sum_{j=1}^m \widehat{\beta}_j/m$. Compared with the traditional method, such a distributive learning framework utilizes the computing power of multiple machines, which avoids the direct storage and operation on the original full dataset. We further illustrate this framework in Figure 1 and refer to it as a distributed algorithm.

The distributed algorithm provides a computationally viable route for learning with big data. However, it remains largely unknown that whether such a divide-and-conquer scheme indeed provides valid theoretical inferences to the original data. For point estimation, Li et al. [11] showed that the distributed moment estimation is consistent, if an unbiased estimate is obtained for each of the sub-problems. For kernel ridge regression, Zhang et al. [28] showed that, with appropriate tuning parameters, the distributed algorithm does lead to a valid estimation. To provide some insights on the feasibility issue, we numerically compare the estimation accuracy of $\bar{\beta}$ with that of $\widehat{\beta}$ in the aforementioned example. Specifically, we generate x_i independently from $N(0, I_{d\times d})$ and set β based on d independent observations from $U[0, 1]$. The value of y_i is generated from the presumed linear model with $\varepsilon \sim N(0, 1)$. We then randomly distribute the full data to $m \in [2^0, 2^{15}]$ local machines and output $\bar{\beta}$ based on m local ridge estimates $\widehat{\beta}_j$ for $j = 1, \ldots, m$. In Figure 2, we plot the estimation errors

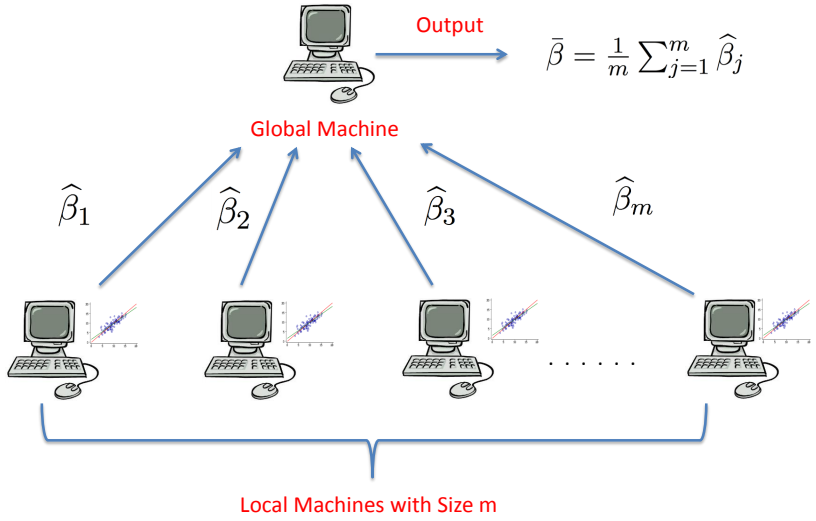

Figure 1: A divide-and-conquer learning framework.

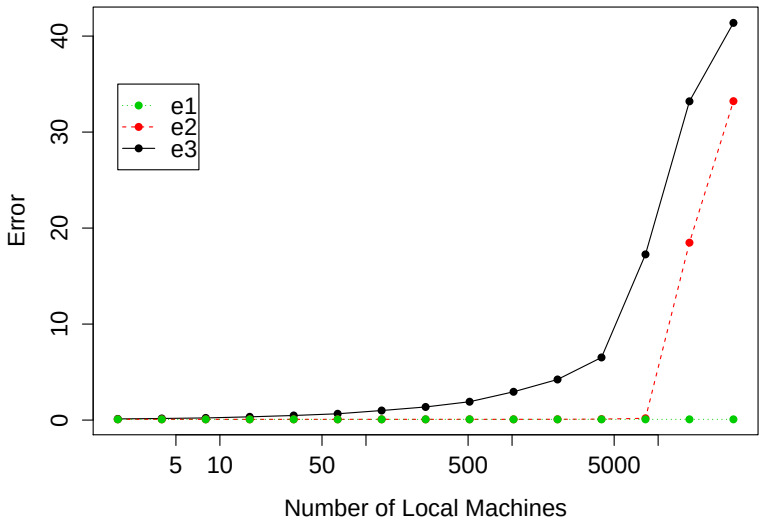

Figure 2: Estimation errors for the distributed regression.

versus the number of local machines m based on three types of estimators: $e_1 = \|\beta - \widehat{\beta}\|_2^2$, $e_2 = \|\beta - \bar{\beta}\|_2^2$, and $e_3 = \min_j \|\beta - \widehat{\beta}_j\|_2^2$. For a wide range of m, it seems that the distributed estimator $\bar{\beta}$ leads to a similar accuracy as the traditional $\widehat{\beta}$ does. However, this argument tends to be false when m is overly large. This observation brings an interesting but fundamental question for using the distributed algorithm in regression: under what conditions the distributed estimator provides an effective estimation of the target function? In this paper, we aim to find an answer to this question and provide more general theoretical support for the distributed regression.

Under the kernel-based regression setup, we propose to take the generalization consistency as a criterion for measuring the feasibility of the distributed algorithms. That is, we regard an algorithm is theoretically feasible if its generalization error tends to zero as the number of observations N goes to infinity. To justify the distributed regression, a uniform convergence rate is needed for bounding the generalization error over the m sub-estimators. This brings new and challenging issues in analysis for the big data setup. Under mild conditions, we show that the distributed kernel regression (DKR) is feasible when the number of its distributed sub-problems is moderate. Our result is applicable to many commonly used regression models, which incorporate a variety of loss, kernel, and penalty functions. Moreover, the feasibility of DKR does not rely on any parametric assumption on the true model. It therefore provides a basic and general understanding for the distributed regression analysis. We demonstrate the promising performance of DKR via both simulation and real data examples.

The rest of the paper is organized as follows. In Section 2, we introduce model setup and formulate the DKR algorithm. In Section 3, we establish the generalization consistency and justify the feasibility of DKR. In Section 4, we show numerical examples to support the good performance of DKR. Finally, we conclude the paper in Section 5 with some useful remarks.

2 Distributed Kernel Regression

2.1 Notations

Let $Y \in [-M, M] \subset \mathbb{R}$ be a response variable bounded by some $M > 0$ and $X \in \mathcal{X} \subset \mathbb{R}^d$ be its d-dimensional covariate drawn from a compact set $\mathcal{X}$. Suppose that $Z = X \times Y$ follows from a fixed but unknown distribution ρ with its support fully filled on $\mathcal{Z} = [-M, M] \times \mathcal{X}$. Let $S = \{z_i = (y_i, x_i), i = 1, \ldots, N\}$ be N independent observations collected from Z. The goal of study is to estimate the potential relationship $f^* : \mathcal{X} \to R$ between X and Y through analyzing S.

Let $\ell(.)$ be a nonnegative loss function and f be an arbitrary mapping from $\mathcal{X}$ to $\mathbb{R}$. We use

$$\mathcal{E}(f) = \mathbb{E}_z[\ell(f, z)] = \int_{\mathcal{Z}} \ell(f, z) d\rho$$

to denote the expected risk of f. The minimizer $f_\rho = \arg\min \mathcal{E}(f)$ is called

the regression function, which is an oracle estimate under ℓ and thus serves as a benchmark for other estimators. Since ρ is unknown, f_ρ is only conceptual. Practically, it is common to estimate f^* through minimizing a regularized empirical risk

$$\min_{f\in\mathcal{F}} \Big\{\mathcal{E}_S(f) + \lambda\|f\|\Big\}, \tag{1}$$

where $\mathcal{F}$ is a user-specified hypothesis space, $\mathcal{E}_S(f) = \sum_{i=1}^N \ell(f, z_i)/N$ is the empirical risk, $\|.\|$ is a norm in $\mathcal{F}$, and $\lambda \geq 0$ is a regularization parameter.

Framework (1) covers a broad range of regression methods. In the machine learning community, it is popular to set $\mathcal{F}$ by a reproducing kernel Hilbert space (RKHS). Specifically, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, symmetric, and semi-positive definite kernel function. The RKHS $\mathcal{H}_K = \overline{\text{span}}\{K(x, .), x \in \mathcal{X}\}$ is a Hilbert space of L^2-integrable functions induced by K. For any $f = \sum_i \alpha_i K(u_i, .)$ and $g = \sum_i \beta_i K(v_i, .)$, their inner product is defined by

$$< f, g >_K = \sum_{i,j} \alpha_i\beta_j K(u_i, v_j)$$

and the kernel L_2 norm is given by $\|f\|_K^2 =< f, f >_K$. It is easy to verify that

$$f(x) =< f, K(x, \cdot) >_{\mathcal{H}_K} \tag{2}$$

for any $f \in \mathcal{H}_K$. Therefore, K is a reproducing kernel of $\mathcal{H}_K$. Readers may refer to [1] [21] for more detailed discussions about RKHS.

Let $\mathcal{C}(\mathcal{X})$ denote the space of continuous functions on $\mathcal{X}$. It is known that $\mathcal{H}_K$ is dense in $\mathcal{C}(\mathcal{X})$ with appropriate choices of K [15]. This property makes $\mathcal{H}_K$ a highly flexible space to estimate an arbitrary $f^* \in \mathcal{C}(\mathcal{X})$. In this paper, we follow framework (1) with $\mathcal{F} = \mathcal{H}_K$ and $\|.\| = \|.\|_K^p$ for some $p > 0$.

2.2 The DKR Algorithm

We now consider (1) in the big data setup. In particular, we assume that sample S is too big to be processed in a single machine and thus we need to use its distributed version. Suppose S is evenly and randomly assigned to m local machines, with each machine processing $n = N/m$ samples. We denote by S_j, $j = 1, 2, \ldots, m$ the sample segment assigned to the jth machine. The global estimator is then constructed through taking average of the m local estimators. Specifically, by setting $\mathcal{F} = \mathcal{H}_K$ in (1), this strategy leads to the distributed kernel regression (DKR), which is described as Algorithm 1.

By representer theorem [17], f_j in step 2 of DKR can be constructed from $\text{span}\{K(x_i, .), x_i \in S_j\}$. This allows DKR to be practically carried out within finite n-dimensional subspaces. The distributive framework of DKR enables parallel processing and thus is appealing to the analysis of big data. With $m = 1$, DKR reduces to the regular kernel-based learning, which has received a great deal of attention in the literature [18] [23] [27]. With quadratic ℓ and $p = 2$, Zhang et. al. [28] conducted a feasibility analysis for DKR with $m > 1$. Unfortunately, their results are built upon the close-form solution of f_j and thus are not applicable to other DKR cases. In this work, we attempt to provide a more general feasibility result for using DKR in dig data.

Algorithm 1 The DKR Algorithm

Input: S, K, λ, m

Output: $\bar{f}$

1: Randomly split S into m sub-samples $S_1, \ldots, S_m$ and store them separately on m local machines.

2: Let $T_M[.]$ be a truncation operator with a cutoff threshold M. For $j = 1, 2, \ldots, m$, find a local estimator based on S_j by

$$\widehat{f}_j = T_M\left[f_j\right],$$

where

$$f_j = \arg\min_{f \in \mathcal{H}_K} \Big\{ \frac{1}{n} \sum_{z_i \in S_j} \ell(f, z_i) + \lambda \|f\|_K^p \Big\}.$$

3: Combine $\widehat{f}_j$s to get a global estimator

$$\bar{f} = \frac{1}{m} \sum_{j=1}^{m} \widehat{f}_j.$$

3 Consistency of DKR

3.1 Preliminaries and Assumptions

In regression analysis, a good estimator of f^* is expected not only to fit training set S but also to predict the future samples from Z. In the machine learning community, such an ability is often referred to as the generalization capability. Recall that f_ρ is a conceptual oracle estimator, which enjoys the lowest generalization risk in a given loss. The goodness of $\bar{f}$ can be typically measured by

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) = \mathbb{E}_z[\ell(\bar{f}, z) - \ell(f_\rho, z)]. \tag{3}$$

A feasible (consistent) $\bar{f}$ is then required to have generalization error (3) converge to zero as $N \to \infty$. When the quadratic loss is used, the convergence of (3) also leads to the convergence of $\|\bar{f} - f_\rho\|_2$, which responds to the traditional notion of consistency in statistics.

When ℓ is convex, Jensen's inequality implies that

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) \leq \frac{1}{m} \sum_{j=1}^{m} [\mathcal{E}(\widehat{f}_j) - \mathcal{E}(f_\rho)].$$

Therefore, the consistency of $\bar{f}$ is implied by the uniform consistency of the m local estimators $\widehat{f}_j$ for $j = 1, \ldots, m$. Under appropriate conditions, this result may be straightforward in the fixed m setup. However, for analyzing big data, it is particularly desired to have m associated with sample size N. This is because

the number of machines needed in an analysis is usually determined by the scale of that problem. The larger a dataset is, the more machines are needed. This in turn suggests that, in asymptotic analysis, m may diverge to infinity as N increases. This liberal requirement of m poses new and challenging issues to justify $\bar{f}$ under the big data setup.

Clearly, the effectiveness of a learning method relies on the prior assumptions on f_ρ as well as the choice of ℓ. For the convenience of discussion, we assess the performance of DKR under the following conditions.

A1 $f_\rho \in \mathcal{C}(\mathcal{X})$ and $\|f_\rho\|_\infty \leq M$, where $\|.\|_\infty$ denotes the function supremum norm.

A2 The loss function ℓ is convex and nonnegative. For any $f_1, f_2 \in \mathcal{C}(\mathcal{X})$ and $z \in \mathcal{Z}$, there exists a constant L such that

$$|\ell(f_1, z) - \ell(f_2, z)| \leq L\|f_1 - f_2\|_\infty.$$

A3 For any $\omega > 0$ and $g \in \mathcal{C}(\mathcal{X})$, there exists a $f \in \mathcal{H}_K$, such that $\|f - g\|_\infty < \omega$. Moreover, let $\mathcal{B}_R = \{f \in \mathcal{H}_K, \|f\|_\infty \leq R\}$ for some $R > 0$. There exists constants C_0, $s > 0$, such that

$$\log \mathcal{N}_\infty(\mathcal{B}_1, \gamma) \leq C_0 \gamma^{-s},$$

where $\mathcal{N}_\infty(\mathcal{F}, \gamma)$ denotes the covering number of a set $\mathcal{F}$ by balls of radius γ with respect to $\|.\|_\infty$.

Condition A1 is a regularity assumption on f_ρ, which can be trivial in applications. For the quadratic loss, we have $f_\rho(X) = \mathbb{E}(Y|X)$ and thus A1 holds naturally with $Y \in [-M, M]$. Condition A2 requires that $\ell(f, z)$ is Lipschitz continuous in f. It is satisfied by many commonly used loss functions for regression analysis. Condition A3 corresponds to the notion of universal kernel in [15], which implies that $\mathcal{H}_K$ is dense in $\mathcal{C}(\mathcal{X})$. It therefore serves as a prerequisite for estimating an arbitrary $f^* \in \mathcal{C}(\mathcal{X})$ from $\mathcal{H}_K$. A3 also requires that the unit subspace of $\mathcal{H}_K$ has a polynomial complexity. Under our setup, a broad choices of K satisfy this condition, which include the popular Gaussian kernel as a special case [29] [30].

3.2 Generalization Analysis

To justify DKR, we decompose (3) by

$$\begin{aligned} \mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) &= \underbrace{\mathcal{E}_S(f) - \mathcal{E}(f) + \mathcal{E}(\bar{f}) - \mathcal{E}_S(\bar{f})}_{\text{sample error}} && (4) \\ &+ \underbrace{\mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f)}_{\text{hypothesis error}} && (5) \\ &+ \underbrace{\mathcal{E}(f) - \mathcal{E}(f_\rho)}_{\text{approximation error}}, && (6) \end{aligned}$$

where f is an arbitrary element of $\mathcal{H}_\mathcal{K}$. The consistency of $\bar{f}$ is implied if (3) has convergent sub-errors in (4)-(6). Since $f \in \mathcal{H}_\mathcal{K}$ is arbitrary, (6) measures how close the oracle f_ρ can be approximated from the candidate space $\mathcal{H}_\mathcal{K}$. This is a term that purely reflects the prior assumptions on a learning problem. Under Conditions A1-A3, with a f such that $\|f - f_\rho\| \leq N^{-1}$, (6) is naturally bounded by L/N. We therefore carry on our justification by bounding the sample and hypothesis errors.

3.2.1 Sample Error Bound

Let us first work on the sample error (4), which describes the difference between the expected loss and the empirical loss for an estimator. For the convenience of analysis, let us rewrite (4) as

$$\begin{aligned}&\mathcal{E}_S(f) - \mathcal{E}(f) + \mathcal{E}(\bar{f}) - \mathcal{E}_S(\bar{f}) \\ &= \left\{\frac{1}{N}\sum_{i=1}^{N}\xi_1(z_i) - \mathbb{E}_z(\xi_1)\right\} + \left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i)\right\}, \end{aligned} \tag{7}$$

where $\xi_1(z) = \ell(f, z) - \ell(f_\rho, z)$ and $\xi_2(z) = \ell(\bar{f}, z) - \ell(f_\rho, z)$. It should be noted that the randomness of ξ_1 is purely from Z, which makes $\mathbb{E}_z(\xi_1)$ a fixed quantity and $\sum_{i=1}^{N}\xi_1(z_i)/N$ a sample mean of independent observations. For ξ_2, since $\bar{f}$ is an output of S, $\mathbb{E}_z(\xi_2)$ is random in S and $\xi_2(z_i)$s are dependent with each other. We derive a probability bound for the sample error through investigating (7).

To facilitate our proofs, we first state one-side Bernstein inequality as the following lemma.

Lemma 1. *Let $y_1, \ldots, y_N$ be N independently and identically distributed random variables with $\mathbb{E}(y_1) = \mu$ and $var(y_1) = \sigma^2$. If $|y_1 - \mu| \leq T$ for some $T > 0$, then for any $\varepsilon > 0$,*

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} y_i - \mu \geq \varepsilon\right\} \leq \exp\left\{\frac{-N\varepsilon^2}{2(\sigma^2 + \varepsilon T/3)}\right\}.$$

The probability bounds for the two terms of (7) are given respectively in the following propositions.

Proposition 1. *Suppose that Conditions A1-A2 are satisfied. For any $0 < \delta < 1$ and $f \in \mathcal{H}_K$, we have*

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N}\xi_1(z_i) - \mathbb{E}_z(\xi_1) \leq 2L\|f - f_\rho\|_\infty\left(\frac{\log(1/\delta)}{N} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right\} \geq 1 - \delta.$$

Proof. Let f be an arbitrary function in $\mathcal{H}_K$. By Condition A2, we have

$$|\xi_1(z)| = |\ell(f, z) - \ell(f_\rho, z)| \leq L\|f - f_\rho\|_\infty$$

for some constant $L > 0$. This implies that $\text{var}(\xi_1) \leq L^2\|f - f_\rho\|_\infty^2$ and $|\xi_1 - \mathbb{E}_z(\xi_1)| \leq 2L\|f - f_\rho\|_\infty$. By Lemma 1, we have,

$$\mathbb{P}\Big\{\frac{1}{N}\sum_{i=1}^{N}\xi_1(z_i) - \mathbb{E}_z(\xi_1) \geq \varepsilon\Big\} \leq \exp\left\{-\frac{Nt^2}{2(L^2\|f - f_\rho\|_\infty^2 + 2/3L\|f - f_\rho\|_\infty t)}\right\} \tag{8}$$

for any $\varepsilon > 0$. Denoting the right hand side of (8) by δ, we have

$$N\varepsilon^2 + \frac{4}{3}L\|f - f_\rho\|_\infty \log\delta\varepsilon + 2L^2\|f - f_\rho\|_\infty^2 \log\delta = 0. \tag{9}$$

The positive root of (9) is given by

$$\begin{aligned}
\varepsilon^* &= \frac{\frac{4}{3}L\|f - f_\rho\|_\infty \log 1/\delta + L\|f - f_\rho\|_\infty\sqrt{\frac{16}{9}\log^2 1/\delta + 8N\log 1/\delta}}{2N} \\
&\leq L\|f - f_\rho\|_\infty\left(\frac{4\log 1/\delta}{3N} + \sqrt{\frac{2\log 1/\delta}{N}}\right) \\
&\leq 2L\|f - f_\rho\|_\infty\left(\frac{\log(1/\delta)}{N} + \sqrt{\frac{\log(1/\delta)}{N}}\right). \qquad (10)
\end{aligned}$$

The proposition is proved by setting $\varepsilon = \varepsilon^*$ in (8). □

Proposition 2. *Suppose that Conditions A1-A3 are satisfied. For any* $0 < \delta < 1$ *and* $f \in \mathcal{H}_K$*, we have*

$$\mathbb{P}\left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) \leq 12ML\left(\frac{V(N,\delta) + \sqrt{V(N,\delta)N}}{N}\right) + N^{-1/(s+2)}\right\} \geq 1-\delta$$

where $V(N,\delta) = C_0(8LMN^{1/(s+2)})^s - \log\delta$.

Proof. Let $\mathcal{D}_M = \{f \in \mathcal{C}(\mathcal{X}), \|f\|_\infty \leq M\}$. Under Condition A3, $\mathcal{B}_{2M} \subset \mathcal{H}_K$ is dense in $\mathcal{D}_M$. Therefore, for any $\epsilon > 0$, there exists a $g_\epsilon \in \mathcal{B}_{2M}$, such that $\|\bar{f} - g_\epsilon\|_\infty < \epsilon$. By A2, we further have

$$\ell(\bar{f}, z) - \ell(g_\epsilon, z) \leq L\epsilon.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) &= \mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) - [\mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f_\rho)] \\
&\leq \mathcal{E}(g_\epsilon) - \mathcal{E}(f_\rho) - [\mathcal{E}_S(g_\epsilon) - \mathcal{E}_S(f_\rho)] + 2L\epsilon. \qquad (11)
\end{aligned}$$

Let $U_\gamma \subset \mathcal{B}_{2M}$ be a cover of $\mathcal{B}_{2M}$ by balls of radius γ with respect to $\|.\|_\infty$.

With $\epsilon \to 0$, (11) implies that

$$\begin{aligned}
& \mathbb{P}\left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) \geq \varepsilon\right\} \\
\leq\ & \mathbb{P}\left\{\sup_{g\in\mathcal{B}_{2M}} \mathcal{E}(g) - \mathcal{E}(f_\rho) - [\mathcal{E}_S(g) - \mathcal{E}_S(f_\rho)] \geq \varepsilon\right\} \\
\leq\ & \mathbb{P}\left\{\sup_{g\in U_\gamma} \mathcal{E}(g) - \mathcal{E}(f_\rho) - [\mathcal{E}_S(g) - \mathcal{E}_S(f_\rho)] \geq \varepsilon - 2L\gamma\right\} \\
\leq\ & \mathcal{N}_\infty(\mathcal{B}_{2M},\gamma)\max_{g\in U_\gamma}\mathbb{P}\left\{\mathcal{E}(g) - \mathcal{E}(f_\rho) - [\mathcal{E}_S(g) - \mathcal{E}_S(f_\rho)] \geq \varepsilon - 2L\gamma\right\} \\
\leq\ & \mathcal{N}_\infty(\mathcal{B}_{2M},\gamma)\exp\left\{-\frac{N(\varepsilon-2L\gamma)^2}{2[9L^2M^2 + 2(\varepsilon-2L\gamma)LM]}\right\}, \qquad (12)
\end{aligned}$$

where the last inequality follows from Lemma 1. By A3, we have

$$\mathcal{N}_\infty(\mathcal{B}_{2M},\gamma) = \mathcal{N}_\infty(\mathcal{B}_1,\gamma/2M) \leq \exp\{C_0(2M/\gamma)^s\}. \qquad (13)$$

Let $\gamma = \varepsilon/4L$. Inequality (12) together with (13) further implies that

$$\mathbb{P}\left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) \geq \varepsilon\right\} \leq \exp\left\{C_0(\frac{8LM}{\varepsilon})^s - \frac{N(\varepsilon)^2}{72L^2M^2 + 8\varepsilon LM}\right\}. \qquad (14)$$

When $\varepsilon \geq N^{-\tau}$ for some $\tau > 0$, (14) implies that

$$\mathbb{P}\left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) \geq \varepsilon\right\} \leq \exp\left\{C_0(8LMN^\tau)^s - \frac{N(\varepsilon)^2}{72L^2M^2 + 8\varepsilon LM}\right\}. \qquad (15)$$

Denote the right hand side of (15) by δ. Following the similar arguments in (9) - (10), we have

$$\mathbb{P}\left\{\mathbb{E}_z(\xi_2) - \frac{1}{N}\sum_{i=1}^{N}\xi_2(z_i) \geq ML\left(\frac{8V(N,\delta) + 6\sqrt{2V(N,\delta)N}}{N}\right) + N^{-\tau}\right\} \leq \delta, \qquad (16)$$

where $V(N,\delta) = C_0(8LMN^\tau)^s - \log\delta$. The proposition is proved by setting $\tau = 1/(s+2)$, which minimizes the bound order in (16). □

Based on Propositions 1 and 2, decomposition (7) implies directly the following probability bound of the sample error.

Theorem 1. *(Sample Error) Suppose that Conditions A1-A3 are salified. Let* $M' = \max\{2M, \|f - f_\rho\|_\infty\}$. *For any* $f \in \mathcal{H}_K$ *and* $0 < \delta < 1$, *we have, with probability at least* $1-\delta$,

$$\mathcal{E}_S(f) - \mathcal{E}(f) + \mathcal{E}(\bar{f}_\lambda) - \mathcal{E}_S(\bar{f}_\lambda) \leq 6M'L\left\{\frac{T_1(N,\delta)}{N} + \frac{T_2(N,\delta)}{N^{\frac{1}{2}}}\right\} + \frac{1}{N^{\frac{1}{2+s}}}, \qquad (17)$$

where

$$\begin{aligned}
T_1(N,\delta) &= V(N,\delta/2) + \log(2/\delta), \\
T_2(N,\delta) &= \sqrt{V(N,\delta/2)} + \sqrt{\log(2/\delta)}.
\end{aligned}$$

When $\|f - f_\rho\|_\infty$ is bounded, the leading factor in (17) is $\sqrt{V(N, \delta/2)/N}$. In that case, Theorem 1 implies that the sample error (4) has an $O(N^{-1/(2+s)})$ bound in probability. Under our model setup, this result is general for a broad range of continuous estimators that is bounded above.

3.2.2 Hypothesis Error Bound

We now continue our feasibility analysis on the hypothesis error (5), which measures the empirical risk difference between $\bar{f}$ and an arbitrary f. When DKR is conducted with $m = 1$, $\bar{f}$ corresponds to the single-machine-based kernel learning. By setting $\lambda = 0$, the hypothesis error has a natural zero bound by definition. However, this property is no longer valid for a general DKR with $m > 1$.

When ℓ is convex, we have (5) bounded by

$$\begin{aligned} \mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f) &= \frac{1}{N}\sum_{i=1}^{N} \ell\left(\frac{1}{m}\sum_{j=1}^{m} \widehat{f}_j, z_i\right) - \frac{1}{N}\sum_{i=1}^{N} \ell\left(f, z_i\right) \\ &\leq \frac{1}{m}\sum_{j=1}^{m} \left\{\mathcal{E}_S(\widehat{f}_j) - \mathcal{E}_S(f)\right\}. \end{aligned} \tag{18}$$

This implies that the hypothesis error of $\bar{f}$ is bounded by a uniform bound of the hypothesis errors over the m sub-estimators. We formulate this idea as the following theorem.

Theorem 2. *(Hypothesis Error) Suppose that Conditions A1-A3 are satisfied. For any $0 < \delta < 1$ and $f \in \mathcal{H}_K$, we have, with probability at least $1 - \delta$,*

$$\mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f) \leq 6LM' \left(\frac{T_1(n, \delta/2)}{n} + \frac{T_2(n, \delta/2)}{n^{\frac{1}{2}}}\right) + \frac{1}{n^{\frac{1}{2+s}}} + 2\lambda \|f\|_K^p,$$

where M', T_1, and T_2 are defined in Theorem 1.

Proof. Without loss of generality, we prove the theorem for $\bar{f}$ with $m > 1$. Recall that DKR spilt S into m segments $S_1, \ldots, S_m$. Let S/S_j be the sample set with S_j removed from S and $\mathcal{E}_Q = \sum_{z_i \in Q} \ell(f, z_i)/q$ be the empirical risk for a sample set Q of size q. Under A2, we have ℓ is convex and thus

$$\begin{aligned} \mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f) &\leq \frac{1}{m}\sum_{j=1}^{m} \left\{\mathcal{E}_S(\widehat{f}_j) - \mathcal{E}_S(f)\right\} \\ &= \frac{1}{m}\sum_{j=1}^{m} \left[\frac{m}{N}(\mathcal{E}_{S_j}(\widehat{f}_j) - \mathcal{E}_{S_j}(f)) + \frac{N-m}{N}(\mathcal{E}_{S/S_j}(\widehat{f}_j) - \mathcal{E}_{S/S_j}(f))\right] \\ &= \frac{1}{m}\sum_{j=1}^{m} \left[\frac{m}{N} B_j + \frac{N-m}{N} U_j\right], \end{aligned} \tag{19}$$

where $B_j = (\mathcal{E}_{S_j}(\widehat{f}_j) - \mathcal{E}_{S_j}(f))$ and $U_j = (\mathcal{E}_{S/S_j}(\widehat{f}_j) - \mathcal{E}_{S/S_j}(f))$.

Let us first work on the first term of (19). By definition of $\widehat{f}_j$, we know that

$$\mathcal{E}_{S_j}(\widehat{f}_j) + \lambda \|\widehat{f}_j\|_K^p \leq \mathcal{E}_{S_j}(f_j) + \lambda \|f_j\|_K^p \leq \mathcal{E}_{S_j}(f) + \lambda \|f\|_K^p$$

Therefore,

$$B_j = \mathcal{E}_{S_j}(\widehat{f}_j) - \mathcal{E}_{S_j}(f)) \leq \lambda \|f\|_K^p - \lambda \|\widehat{f}_j\|_K^p \leq \lambda \|f\|_K^p. \tag{20}$$

This implies that the first term of (19) is bounded by $m\lambda \|f\|_K^p / N$.

We now turn to bound the second term of (19). Specifically, we further decompose U_j by

$$\begin{aligned} U_j &= u_{1j} + u_{2j} + u_{3j} + u_{4j} + B_j \\ &\leq u_{1j} + u_{2j} + u_{3j} + u_{4j} + \lambda \|f\|_K^p, \end{aligned}$$

where

$$\begin{aligned} u_{1j} &= \mathcal{E}_{S/S_j}(\widehat{f}_j) - \mathcal{E}_{S/S_j}(f_\rho) - \mathcal{E}(\widehat{f}_j) + \mathcal{E}(f_\rho) \\ u_{2j} &= \mathcal{E}(f) - \mathcal{E}(f_\rho) - \mathcal{E}_{S/S_j}(f) + \mathcal{E}_{S/S_j}(f_\rho) \\ u_{3j} &= \mathcal{E}_{S_j}(f) - \mathcal{E}_{S_j}(f_\rho) + \mathcal{E}(f_\rho) - \mathcal{E}(f) \\ u_{4j} &= \mathcal{E}(\widehat{f}_j) - \mathcal{E}(f_\rho) - \mathcal{E}_{S_j}(\widehat{f}_j) + \mathcal{E}_{S_j}(f_\rho) \end{aligned}$$

Note that $\widehat{f}_j$ is independent of S/S_j. Proposition 1 readily implies that, with probability at least $1-\delta$,

$$\begin{aligned} u_{1j} &\leq 4LM \left(\frac{\log(1/\delta)}{N-n} + \sqrt{\frac{\log(1/\delta)}{N-n}} \right), \\ u_{2j} &\leq 2L\|f - f_\rho\|_\infty \left(\frac{\log(1/\delta)}{N-n} + \sqrt{\frac{\log(1/\delta)}{N-n}} \right), \\ u_{3j} &\leq 2L\|f - f_\rho\|_\infty \left(\frac{\log(1/\delta)}{n} + \sqrt{\frac{\log(1/\delta)}{n}} \right). \end{aligned}$$

Also, by applying Proposition 2 with $m = 1$, we have, with probability at least $1-\delta$,

$$u_{4j} \leq 12ML \left(\frac{V(n,\delta) + \sqrt{V(n,\delta)n}}{n} \right) + n^{-1/(s+2)},$$

with the same V defined in Proposition 2. Consequently, we have, with probability at least $1-\delta$,

$$U_j \leq 6LM' \left(\frac{V(n,\delta/4) + \log(4/\delta)}{n} + \frac{\sqrt{\log 4/\delta} + \sqrt{V(n,\delta/4)}}{n^{\frac{1}{2}}} \right) + \frac{1}{n^{\frac{1}{2+s}}} + \lambda \|f\|_K^p, \tag{21}$$

where $M' = \max\{2M, \|f - f_\rho\|_\infty\}$.

Inequalities (20) and (21) further imply that, with probability at least $1-\delta$

$$\mathcal{E}_S(\bar{f}) - \mathcal{E}_S(f) \leq 6LM' \left(\frac{T_1(n,\delta/2)}{n} + \frac{T_2(n,\delta/2)}{n^{\frac{1}{2}}} \right) + \frac{1}{n^{\frac{1}{2+s}}} + 2\lambda \|f\|_K^p.$$

The theorem is therefore proved. □

Theorem 2 implies that, with appropriate f and λ, the hypothesis error of DKR has an $O(n^{-1/(2+s)})$ bound in probability. This results is applicable to a general $\bar{f}$ with $m \geq 1$, which incorporates the diverging m situations.

3.3 Generalization Bound of DKR

With the aid of Theorems 1-2, we obtain a probability bound for the generalization error of $\bar{f}$ as the following theorem.

Theorem 3. *(Generalization Error) Suppose that Conditions A1-A3 are satisfied. When N is sufficiently large, for any $0 < \delta < 1$,*

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) \leq 24LM \left(\frac{T_1(n, \delta/4)}{n} + \frac{T_2(n, \delta/4)}{n^{\frac{1}{2}}} \right) + \frac{2+L}{n^{\frac{1}{2+s}}} + 2\lambda \|f_0\|_K^p$$

with probability at least $1 - \delta$, where $f_0 \in \mathcal{H}_\mathcal{K}$ and $\|f_0 - f_\rho\|_\infty \leq N^{-1}$.

Proof. Under Conditions A1 and A3, for any $N \geq 1$, there exists a $f_0 \in \mathcal{H}_\mathcal{K}$ such that $\|f_0 - f_\rho\| < N^{-1}$. Under A2, this also implies that (6) is bounded by $L/N \leq L/n^{1/(2+s)}$. Clearly, when N is sufficiently large, $M' = \max(2M, \|f_0 - f_\rho\|) = 2M$. The theorem is a direct result by applying Theorems 1-2 to (4) and (5) with $f = f_0$. □

Theorem 3 suggests that, if we set $\lambda = o(\|f_0\|_K^{-p} n^{-1/(2+s)})$, the generalization error of $\bar{f}$ is bounded by an $O(n^{-1/(2+s)})$ term in probability. In other words, as $n \to \infty$, a properly tuned DKR leads to an estimator that achieves the oracle predictive power. This justifies the feasibility of using divide-and conquer strategy for the kernel-based regression analysis. Under the assumption that $f_\rho \in \mathcal{H}_\mathcal{K}$, we have $f_0 = f_\rho$ and thus $\bar{f}$ is feasible with $\lambda = o(n^{-1/(2+s)})$. Moreover, when DKR is conducted with Gaussian kernels, Condition A3 is satisfied with any $s > 0$ and thus $\mathcal{E}(\bar{f})$ enjoys a nearly $O_p(n^{-1/2})$ convergence rate to $\mathcal{E}(f_\rho)$.

Theorem 3 provides theoretical support for the distributed learning framework (Algorithm 1). It also reveals that the convergence rate of $\mathcal{E}(\bar{f})$ is related to the scale of local sample size n. This seems to be reasonable, because $\widehat{f}_j$ is biased from f_ρ under a general setup. The individual bias of $\widehat{f}_j$ may diminish as n increase. It, however, would not be balanced off by taking the average of $\widehat{f}_j$s for $j = 1, \ldots, m$. As a result, the generalization bound of $\bar{f}$ is determined by the largest bias among the m $\widehat{f}_j$s. When $\widehat{f}_j$ is (nearly) unbiased, its generalization performance is mainly affected by its variance. In that case, $\bar{f}$ is likely to achieve a faster convergence rate by averaging over $\widehat{f}_j$s. We use the following corollary to show some insights on this point.

Corollary 1. *Suppose that DKR is conducted with the quadratic loss and $\lambda = 0$. If $\mathbb{E}[\widehat{f}_j(x) - f_\rho(x)] = 0$ for any $x \in \mathcal{X}$, then under Conditions A1-A3, we have*

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) = O_p \left(\frac{1}{mn^{\frac{1}{2+s}}} \right).$$

Proof. Let ρ_X be the marginal distribution of X. When the quadratic loss is used, we have

$$\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho) = \|\bar{f} - f_\rho\|_{\rho_X}^2 = \int_{\mathcal{X}} (\bar{f}(X) - f_\rho(X))^2 d\rho_X \tag{22}$$

Since we assume $\mathbb{E}[\widehat{f}_j(x)] = f_\rho(x)$ for any $x \in \mathcal{X}$, (22) implies that

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho)] &= \int_S \int_{\mathcal{X}} (\bar{f}(X) - f_\rho(X))^2 d\rho_X d\rho \\
&= \int_{\mathcal{X}} (\mathbb{E}[\bar{f}(X) - f_\rho(X)])^2 d\rho_X + \int_{\mathcal{X}} \mathbb{E}[\bar{f}(X) - f_\rho(X)]^2 d\rho_X \\
&= \frac{1}{m} \int_{\mathcal{X}} \mathbb{E}[\widehat{f}_1(X) - f_\rho(X)]^2 d\rho_X \\
&= \frac{1}{m} \mathbb{E}[\mathcal{E}(\widehat{f}_1) - \mathcal{E}(f_\rho)]. \qquad (23)
\end{aligned}$$

Applying Theorem 3 with $m = 1$ and $\lambda = 0$, we have, for some generic constant $C > 0$,

$$\mathbb{P}\left\{\mathcal{E}(\widehat{f}_1) - \mathcal{E}(f_\rho) > C \log(8/\delta) n^{-\frac{1}{2+s}}\right\} \leq \delta \tag{24}$$

Let $t = C \log(8/\delta) n^{-\frac{1}{2+s}}$. Inequality (24) implies that

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(\widehat{f}_1) - \mathcal{E}(f_\rho)] &= \int_0^\infty \mathbb{P}\left\{\mathcal{E}(\widehat{f}_1) - \mathcal{E}(f_\rho) > t\right\} dt \\
&\leq \int_0^\infty 8 \exp\left\{-C^{-1} n^{\frac{1}{2+s}} t\right\} dt \\
&\leq 8C n^{-\frac{1}{2+s}}.
\end{aligned}$$

This together with (23) implies that $\mathbb{E}[\mathcal{E}(\bar{f}) - \mathcal{E}(f_\rho)] = O(m^{-1} n^{-\frac{1}{2+s}})$, which further implies the corollary. □

Corollary 1 is only conceptual, because it is usually difficult to construct an unbiased $\widehat{f}_j$ without strong prior knowledge. Nevertheless, it sheds light on designing more efficient DKR with less biased sub-estimators. In practice, this may be conducted by choosing a small λ or using some debiasing techniques in Algorithm 1. In this paper, we focus on providing a general feasibility support for DKR and leave this issue for the future research.

It should also be noted that, under Theorem 3, DKR is feasible only when $n \to \infty$ or equivalently $m = o(N)$. This means that, to have DKR work well, the sample size in each local machine should be large enough. This seems to be a natural condition, because for a large-m-small-n situation, each local output $\widehat{f}_j$ is unlikely to provide a meaningful estimate. As a consequence, the global estimation $\bar{f}_\lambda$ may not be well constructed neither. In real applications, an appropriate m should be used such that the associated DKR achieves a good balance of algorithmic accuracy and computational efficiency.

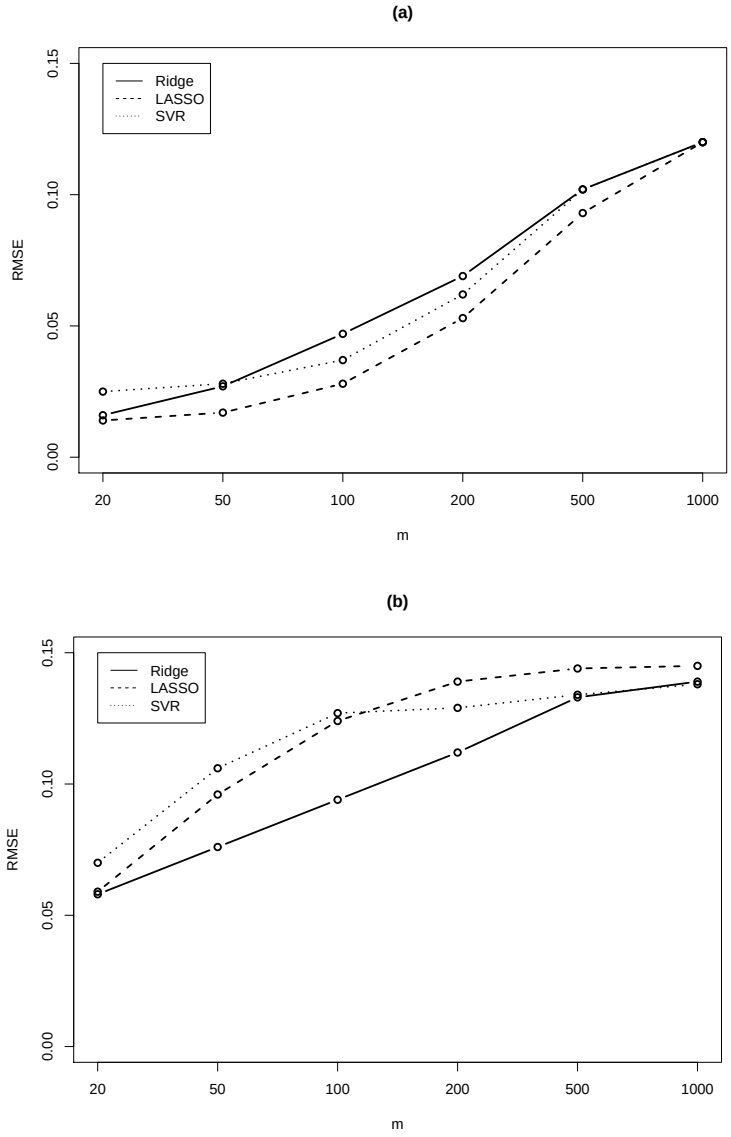

Figure 3: The generalization performance of DKR in Case (i). Plot (a): RMSE for $\bar{f}$; Plot (b): RMSE for $\widehat{f_1}$.

4 Numerical Studies

We evaluate the finite sample performance of DKR through both simulation and real data examples. In particular, we assess the distributive strategy for several popular regression methods in terms of both computational efficiency and generalization capability. All numerical studies are implemented by MATLAB 8.2 on a windows workstation with 8-core 3.07GHz CPUs.

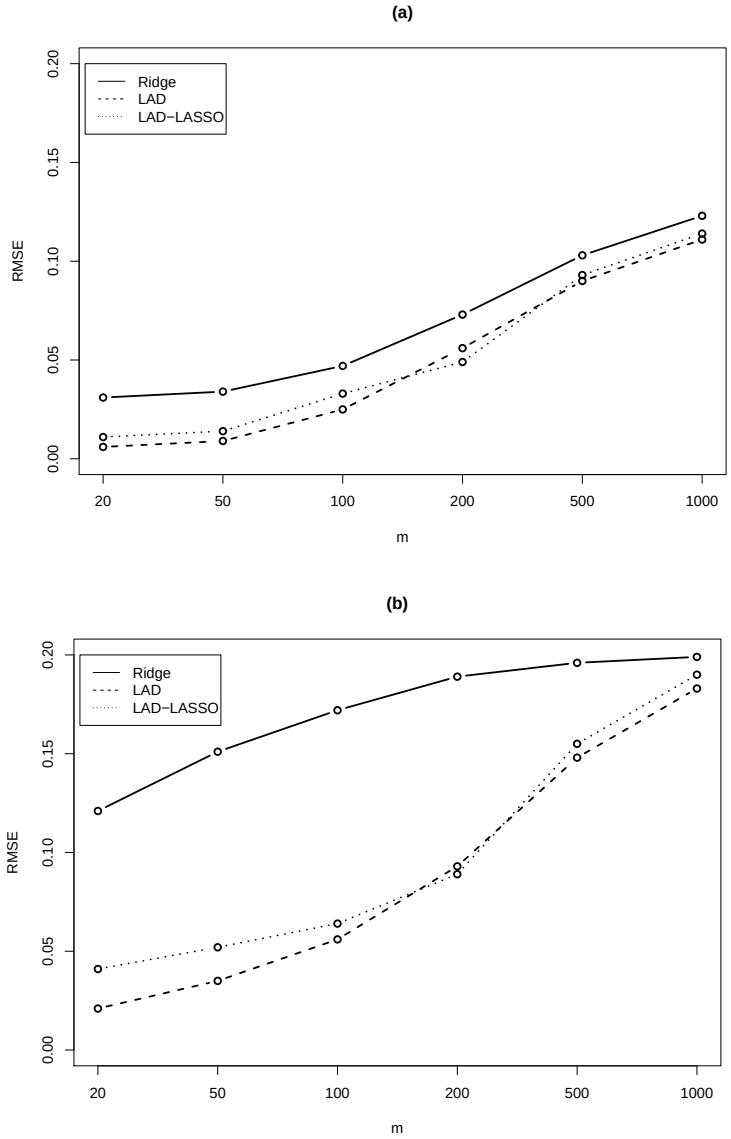

Figure 4: The generalization performance of DKR in Case (ii). Plot (a): RMSE for $\bar{f}$; Plot (b): RMSE for $\widehat{f}_1$.

4.1 Simulation

In the simulation studies, we assess DKR on a hypothetical learning problem with $d = 2$. Specifically, we generate independent observations based on model

$$Y = \text{sinc}(20X_1 - 10) \times \text{sinc}(20X_2 - 10) + \epsilon, \tag{25}$$

where (X_1, X_2) denotes the two attributes of covariate X, ϵ is an observational noise, and

$$\text{sinc}(x) = \begin{cases} \frac{\sin(x)}{x}, & x \neq 0 \\ 1, & x = 0 \end{cases}.$$

The values of (X_1, X_2) are sampled based on a uniform distribution on $[0, 1] \times [0, 1]$.

Table 1: Simulation results: averaged computational time of DKR in second.

	$m =$	20	50	100	200	500	1000
Case (i)	Ridge	27.6	1.91	0.26	0.04	< 0.01	< 0.01
	LASSO	74.8	13.6	4.93	2.54	1.91	1.20
	SVR	0.94	0.25	0.09	0.04	0.02	0.01
Case (ii)	Ridge	28.1	1.94	0.26	0.04	< 0.01	< 0.01
	LAD	112	15.2	2.67	0.76	0.23	0.16
	LAD-LASSO	104	17.4	2.31	0.59	0.17	0.08

We evaluate DKR based on model (25) under two cases: (i) we set $N = 100,000$ and generate data with $\epsilon \sim N(0, 0.2)$; (ii) we generate $N_1 = 80,000$ samples with $\epsilon \sim N(0, 0.1)$ and $N_2 = 20,000$ samples with $\epsilon \sim U[-2, 2]$. The second case is designed such that the data contain about 20% outliers. This setup poses further challenges for DKR in learning the relationship between Y and X.

Regarding the implementation of DKR, we set the number of partition $m = 20, 50, 100, 200, 500$, and 1000, so that the minimum sample size in each local machine is 100. We set the thresholding value $M = 1$ and build the dictionary $\mathcal{H}_K$ by the popular Gaussian kernel

$$K(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2 / \tau^2) \tag{26}$$

with $\tau = 0.05$. In Case (i), we conduct DKR with three popular regression methods under framework (1): ridge regression (L_2-loss plus L_2-regularization), LASSO (L_2-loss plus L_1-regularization), and SVR (ε-intensive-loss plus L_2-regularization); in Case (ii), we conduct DKR based on two robust regression methods: LAD (L_1-loss plus L_2-regularization) and LAD-LASSO (L_1-loss plus L_1-regularization). In our simulations, we choose the tuning parameter λ based on a few pilot runs of DKR with $m = 20$ and use the standard MATLAB packages for computing the corresponding regression estimators.

To assess the generalization capability of DKR, we generate an independent testing set $\{(\tilde{y}_i, \tilde{x}_i),\ i = 1, \ldots, n_t\}$ of size $n_t = 5000$ from model (25) with $\epsilon = 0$ and compute

$$\text{RMSE}(\bar{f}) = \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \left| \bar{f}(\tilde{x}_i) - \tilde{y}_i \right|^2 \right\}^{1/2}.$$

We report the averaged RMSE of DKR for each setup based on 100 repetitions. For comparison, we also report the RMSE of the corresponding standard (non-distributive) regression method based on $1/m$ of the data.

The simulation results are shown in Figures 3-4, where the associated computational cost is given in Table 1. We observe that, when m is moderate, the DKR approach performs quite well in achieving a low RMSE for all tested regression methods. This partially justifies the feasibility result obtained in this work. In our setup, choosing $m \in (50, 100)$ seems to be the most attractive, because the associated DKR estimator enjoys a strong generalization capability

at a low computational cost. Clearly, by using multiple machines, DKR retains almost the same computational cost as the standard non-distributive method using only $1/m$ of the data. Meanwhile, with a moderate m, it significantly improves the resulting estimator over the single machine-based local output. The framework of DKR therefore serves as a viable route for conducting efficient leaning for big data.

It should also be noted that the performance of DKR may deteriorate when m is overly large. In Case (i) with $m = 1000$, DKR does not help much in reducing the RMSE of the single-machine-based estimator. As discussed in Section 3.3, this might be caused by the estimation bias and insufficient sample size for each local machine. In principle, a smaller m helps to improve the effectiveness of DKR, but it also leads to a higher computational cost. In practice, one should conduct DKR with different choices of m and select an appropriate value based on specific situations. It might be a good idea to set m as the smallest value within the affordable computational budget.

DKR also inherits reasonable robustness against outliers from the associated local outputs. This is revealed by the low RMSE of $\bar{f}$ conducted on LAD and LAD-LASSO in Case (ii) with $m \leq 50$.

4.2 Real data example

We apply DKR to analyze a real world dataset, which contains 583,250 instances of Twitter discussions on topics related to new technology in 2013. Each instance is described by $d = 77$ features related to that discussion. It is of interest to predict the number of active discussions (Y) based on these features (X). To facilitate the computing process, we include the instances with $Y \in [20, 200]$ in our analysis, which leads to a training set with size $174,507$ and a testing set with size $19,390$. We standardize each attribute of X such that it has a zero mean and a unit standard deviation. Readers may refer to *Buzz Data* on *http://archive.ics.uci.edu/ml/datasets.html* for more detailed information about this dataset.

Table 2: RMSE for the analysis of Buzz data.

$m =$	40	120	300	500	1000
Ridge	24.8	25.3	25.6	25.9	26.5
LASSO	24.9	25.3	25.6	26.0	26.4
LAD	25.1	25.4	25.9	26.0	26.3

Similar to our simulation studies, we build $\mathcal{H}_K$ based on the Gaussian kernel (26) with $\tau = 10$. We set $m = (40, 120, 300, 500, 1000)$ and apply DKR to the training sample with Ridge, LASSO, and LAD. We summarize the analysis in term of RMSE based on the testing sample, which is shown in Table 2. Like many other social media data, this dataset is known to be noisy and highly skewed. Thus, the results in Table 2 indicate the decent performance of DKR. In this example, we observe that the results are not very sensible to the choice of m. Thus, researchers may prefer a larger m for the computational convenience.

5 Conclusion

In this paper, we studied the distributed kernel regression for learning with big data. DKR follows from a divide-and-conquer framework, which enables distributive storage and parallel computing. In DKR, the performance of the global estimator is determined by a uniform bound over the distributed local estimates. Under mild conditions, we show that DKR provides a consistent estimate that leads to the oracle generalization risk. Our results offer a general theoretical support for DKR, which is applicable to a broad range of regression methods. As the first step, the current work focus only on the feasibility of DKR. It would be important to further investigate its efficiency and develop the corresponding acceleration methods. Also, it is promising to extend the current distributive framework to other learning tasks, such as classification and variable selection. We leave all these interesting topics for the future research.

Acknowledgment

This work is supported in part by NIDA grants P50 DA10075, P50 DA036107, and the Natural Science Foundation of China grant 11301494. The authors are grateful to Dr. Xiangyu Chang at Xi'an Jiaotong University (China) and Dr. Jian Fang at Tulane University for their constructive suggestions to this work.

References

[1] Berlinet, A. and Thomas-Agnan, C. (2004) *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, New York.

[2] Cao, F., Lin, S. and Xu, Z. (2010) Approximation capability of interpolation neural networks. *Neurocomputing*, **74** 457-460.

[3] Chang, F., Dean, J., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A. and Gruber, R. (2008) Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems*, **26**, No. 2, Article 4.

[4] Chu, C., Kim, S., Lin, Y., Yu, Y., Bradski, G., Ng, A. and Olukotun, K. (2006) Map-reduce for machine learning on multicore. *NIPS*, **6** 281-288.

[5] Christmann, A. and Steinwart, I. (2008) Consistency of kernel-based quantile regression. *Applied Stochastic Models in Business and Industry*, **24** 171–183.

[6] Dean, J. and Ghemawat, S. (2008) MapReduce: simplified data processing on large clusters. *Communications of the ACM*, **51** 107-113.

[7] Fan, R., Chang, K., Hsieh, C., Wang, X. and Lin, C. (2008) LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, **9** 1871-1874.

[8] Ghemawat, S., Gobioff, H. and Leung, S. (2003) The Google file system. *ACM SIGOPS Operating Systems Review*, **37** 29-43.

[9] Kimeldorf, G. and Wahba, G. (1971) Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, **33** 82-95.

[10] Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. (2012) The big data bootstrap. *arXiv preprint* arXiv:1206.6415.

[11] Li, R., Lin, D. and Li, B. (2013) Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, **29** 399-409.

[12] Li, Y., Liu, Y. and Zhu, J. (2007) Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, **102** 255-268.

[13] Mackey, L., Talwalkar, A. and Jordan, M. (2011) Divide-and-conquer matrix factorization. *arXiv preprint*, arXiv:1107.0789.

[14] McDonald, R., Hall, K. and Mann, G. (2010) Distributed training strategies for the structured perceptron. *Proceedings of The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 456-464. Los Angeles, CA.

[15] Micchelli, C., Xu, Y. and Zhang, H. (2006) Universal Kernels. *Jounal of Machine Learning Research*, **7** 2651-2667.

[16] Narcowich, F., Ward, J. and Wendland, H. (2006) Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, **24** 175-186.

[17] Schölkopf, B., Herbrich, R. and Smola, A. J. (2001) A Generalized Representer Theorem. *Lecture Notes in Computer Science*, **2111** 416426.

[18] Schölkopf, B. and Smola, A. (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. The MIT Press. Cambridge, MA.

[19] Takeuchi, I., Le, Q., Sears, T. and Smola, A. (2006) Nonparametric quantile estimation. *The Journal of Machine Learning Research*, **7** 1231-1264.

[20] Vapnik, V. (2000) The nature of statistical learning theory. Springer. NewYork, NY.

[21] Wahba, G. (1990) Spline models for observational data. SIAM. Philadelphia, PA.

[22] Wright, J., Ganesh, A., Rao, S., Peng, Y. and Ma, Y. (2009) Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, **22** 2080-2088.

[23] Wu, Q., Ying, Y. and Zhou, D. (2006) Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, **6** 171-192.

[24] Wu, X., Zhu, X., Wu, G. and Ding, W. (2014) Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, **26** 97-107.

[25] Xiang, D., Hu, T. and Zhou, D. (2012) Approximation analysis of learning algorithms for support vector regression and quantile regression. Journal of Applied Mathematics, **2012** pp.17.

[26] Zhao, Q., Meng, D. and Xu, Z. (2012) A recursive divide-and-conquer approach for sparse principal component analysis. *arXiv preprint*, arXiv:1211.7219.

[27] Zhang, T. (2005) Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, **17** 2077-2098.

[28] Zhang, Y., Duchi, J. and Wainwright, M. (2013) Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. *arXiv preprint*, arXiv:1305.5029

[29] Zhou, D. (2002) The covering number in learning theory. *Journal of Complexity*, **18** 739-767.

[30] Zhou, D. (2003) Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, **49** 1743-1752.

[31] Zou, B., Li, L., Xu, Z., Luo, T and Tang, Y (2013) Generalization Performance of Fisher Linear Discriminant Based on Markov Sampling. *IEEE Transactions on Neural Networks and Learning Systems*, **24** 288-300.