Optimal Rates of Distributed Regression with Imperfect Kernels

Hongwei Sun   SS_SUNHW@UJN.EDU.CN
School of Mathematical Sciences, University of Jinan
Jinan, Shandong, P. R. China

Qiang Wu   QWU@MTSU.EDU
Department of Mathematical Sciences, Middle Tennessee State University
Murfreesboro, TN 37132, USA

Editor: John Shawe-Taylor

Abstract

Distributed machine learning systems have been receiving increasing attentions for their efficiency to process large scale data. Many distributed frameworks have been proposed for different machine learning tasks. In this paper, we study the distributed kernel regression via the divide and conquer approach. The learning process consists of three stages. Firstly, the data is partitioned into multiple subsets. Then a base kernel regression algorithm is applied to each subset to learn a local regression model. Finally the local models are averaged to generate the final regression model for the purpose of predictive analytics or statistical inference. This approach has been proved asymptotically minimax optimal if the kernel is perfectly selected so that the true regression function lies in the associated reproducing kernel Hilbert space. However, this is usually, if not always, impractical because kernels that can only be selected via prior knowledge or a tuning process are hardly perfect. Instead it is more common that the kernel is good enough but imperfect in the sense that the true regression can be well approximated by but does not lie exactly in the kernel space. We show distributed kernel regression can still achieve capacity independent optimal rate in this case. To this end, we first establish a general framework that allows to analyze distributed regression with response weighted base algorithms by bounding the error of such algorithms on a single data set, provided that the error bounds have factored the impact of unexplained variance of the response variable. Then we perform a leave one out analysis of the kernel ridge regression and bias corrected kernel ridge regression, which in combination with the aforementioned framework allows us to derive sharp error bounds and capacity independent optimal rates for the associated distributed kernel regression algorithms. As a byproduct of the thorough analysis, we also prove the kernel ridge regression can achieve rates faster than $O(N^{-1})$ (where N is the sample size) in the noise free setting which, to our best knowledge, are first observed and novel in regression learning.

Keywords: Learning Theory, Distributed regression learning, Optimal rates, Imperfect kernels, Leave one out analysis

1. Introduction

Distributed machine learning systems have been receiving increasing attentions for their efficiency to process large scale data. Many distributed frameworks have been proposed for different machine learning tasks; see for instance Dean and Ghemawat (2004); Dobriban and

Sheng (2020); Kraska et al. (2013); Boyd et al. (2011); Xu et al. (2018); Zhang et al. (2015). Among others, the divide and conquer approach has been proved easy to implement but efficient for statistical estimation and predictive analytics. This approach is used when the data is too big to be analyzed by one computer node and usually consists of three stages. First, the data is randomly partitioned into multiple subsets. In some applications the data may be naturally stored in different locations as a result of data collection process and there is no need for further partitioning. Second, a base algorithm is selected according to the learning task and applied to each subset to learn a local model. Finally, all local models are averaged to generate the final model. This approach is computationally efficient because the second stage can be easily parallelized. Also, because the local model training does not require mutual communication between the computing nodes, it can largely preserve privacy and confidentiality.

In the context of nonlinear regression analysis distributed kernel methods implemented via the divide and conquer approach have been widely studied and showed asymptotically minimax optimal in many situations. In particular, if the kernel is perfectly selected so that the true regression function lies in the associated reproducing kernel space, the minimax optimality was verified for kernel ridge regression (Zhang et al., 2015; Szabó et al., 2016; Lin et al., 2017), kernel spectral algorithm (Guo et al., 2017a), kernel based gradient descent (Lin and Zhou, 2018), bias corrected regularization kernel network (Guo et al., 2017b), and minimum error entropy (Hu et al., 2020). However, it is usually, if not always, impractical to select perfect kernels for real world problems. More commonly one has to select an imperfect kernel by some prior knowledge and/or a tuning process. Such a kernel is empirically optimal within a family of candidate kernels, usually good enough for applications, but there is no guarantee of perfectness. A typical example is the widely used Gaussian radial basis kernel. It is effective in most nonlinear data analysis problems because of its universality. But it is well known that its associated reproducing kernel Hilbert space consists of only infinitely differentiable functions. Functions that are not infinitely differentiable can be well approximated by the kernel but cannot lie exactly in the associated kernel space. In this situation the learning rates obtained in the literature are suboptimal.

The primary goal of this study is to verify the capacity independent optimality of distributed kernel regression algorithms when the kernel is imperfect. We focus on the use of kernel ridge regression (KRR) and bias corrected kernel ridge regression (BCKRR) in the divide and conquer approach. For this purpose, we propose a framework to analyze a broad class of distributed regression methods and conduct rigorous leave one out analyses. More specifically, we make the following four contributions.

- First, we introduce the concept of response weighted regression algorithm, which covers a broad class of regression algorithms and has both KRR and BCKRR as examples. We propose a general framework to analyze the learning performance of response weighted distributed regression algorithms and show that, for such algorithms, it suffices to study the learning performance of the response weighted algorithm on a single data set provided that it characterizes the impact of the unexplained variance of the response variable on the learning performance. This makes the analysis of such distributed learning algorithms much easier. Because both KRR and BCKRR fall into this framework, we will utilize it to analyze the distributed KRR and distributed BCKRR in this study.

- Second, we conduct a leave one out analysis of the KRR algorithm and prove capacity independent error bounds, which are sharp in the sense that they lead to optimal capacity independent rates regardless the kernel is perfect or imperfect. While the idea of leave one out analysis was originally developed in Bousquet and Elisseeff (2002); Zhang (2003), our analysis is more rigorous so that the error bounds factor in the impact of the unexplained variance of the response variable and therefore we are able to utilize them in combination with the aforementioned framework to derive sharp error bounds and optimal learning rates for distributed KRR. Furthermore, our analysis also greatly relaxes the restriction on the number of local machines used in the distributed regression. In particular, our results indicate that fast rates can still be achieved with the number of local machines increasing at the order of $O(\sqrt{N})$ (with N being the sample size) if the true regression does lie in the kernel space. This is a significant relaxation compared with those in the literature; see detailed comparison in Section 2.3.

- Third, we conduct a rigorous leave one out analysis of the BCKRR and utilize the results to derive error bounds and learning rates for distributed BCKRR algorithm. Again, the results are optimal from a capacity independent viewpoint when the kernel is imperfect. As the BCKRR was proposed by the idea of bias correction, its original formula involves a two-step procedure and admits an operator representation which, if not impossible, is unsuitable for leave one analysis in a natural way. To overcome this difficulty we prove two alternative formulae for the algorithm, among which the recentering regularization formula defines the target function by a Tikhonov regularization scheme and allows us to conduct leave one analysis naturally. Moreover, these two perspectives also shed light on the design of other bias corrected algorithms whose solutions do not have an explicit representation like KRR. See Section 6.1 and the discussions in Section 7 for details.

- Last, as a byproduct of our leave one out analysis, we derive super fast learning rates for both KRR and BCKRR when the unexplained variance of the response variable becomes zero, that is, the response value is determined and noise free for any fixed input. If the kernel is perfect, the rate can be faster than $O(N^{-1})$ and even as fast as $O(N^{-2})$ in the best situation. To our best knowledge, such super fast rates for kernel regression are first observed and novel in learning theory research.

The rest of the paper is organized as follows. In Section 2 we describe the problem setting, algorithms, and the main results. Discussions and detailed comparisons between our results and those in the literature will be given. Empirical studies will be used to illustrate the effectiveness of imperfect kernels in distributed regression. In Section 3 we propose a general framework for the analysis of response weighted distributed regression algorithms. Some preliminary lemmas were proved in Section 4. Then in Section 5 and Section 6 we conduct leave one out analyses of KRR and BCKRR, respectively. The results are then used to prove our main results regarding the error bounds and learning rates of distributed KRR and distributed BCKRR. We close with conclusions and discussions in Section 7.

2. Problem setting and main results

Let $\mathcal{X}$ be the sample space of input variable x and $\mathcal{Y}$ the sample space of the response variable y. They are linked by a joint probability measure ρ on the product space $\mathcal{X} \times \mathcal{Y}$. The goal is to learn the mean regression function that minimizes the mean squared prediction error, i.e.,

$$f^* = \arg\min \mathcal{E}(f) \qquad \text{where } \mathcal{E}(f) = \mathbf{E}[(y - f(x))^2].$$

This is usually implemented by minimizing the empirical mean squared error or its regularized version when we have in hand a sampled data set $D = \{(x_i, y_i), i = 1, \ldots, N\}$. If f^* is linear, multiple linear regression or the regularized methods such as ridge regression or LASSO performs well. When f^* is nonlinear, kernel ridge regression can be used to search a good approximating function in a suitable reproducing kernel Hilbert space.

Let K be a Mercer kernel, namely, a continuous, symmetric, and positive-semidefinite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The inner product defined by $\langle K(x, \cdot), K(t, \cdot) \rangle_K = K(x, t)$ induces a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated to the kernel K. The space is the closure of the function class spanned by $\{K_x = K(x, \cdot) : x \in \mathcal{X}\}$. The reproducing property $f(x) = \langle f, K_x \rangle_K$ leads to $|f(x)| \leq \sqrt{K(x,x)}\|f\|_K$. Thus if $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)} < \infty$, then $\mathcal{H}_K$ can be embedded into $C(\mathcal{X})$ and $\|f\|_\infty \leq \kappa\|f\|_K$. We refer to Aronszajn (1950) for more other properties of RKHS. The kernel ridge regression (KRR) estimates the true regression function f^* by the function $f_{D,\lambda} \in \mathcal{H}_K$ minimizing the regularized sample mean squared error,

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

where $\lambda > 0$ is a regularization parameter. It is a popular kernel method for nonlinear regression analysis. Its predictive consistency has been extensively studied in the literature; see e.g. Evgeniou et al. (2000); Bousquet and Elisseeff (2002); Zhang (2003); De Vito et al. (2005); Wu et al. (2006); Bauer et al. (2007); Caponnetto and De Vito (2007); Smale and Zhou (2007); Sun and Wu (2009b); Steinwart et al. (2009) and many references therein. Its applications were also extensively explored and shown successful in many problem domains.

By the famous representer theorem, $f_{D,\lambda}$ admits a representation

$$f_{D,\lambda}(x) = \sum_{i=1}^N c_i K(x_i, x)$$

with $\mathbf{c} = (c_1, \ldots, c_N)^\top \in \mathbb{R}^N$ solved from the linear system $(\lambda N \mathbb{I} + \mathbb{K})\mathbf{c} = \mathbf{y}$ where $\mathbb{I}$ is the identity matrix, $\mathbb{K} = [K(x_i, x_j)]_{i,j=1}^N$ is the kernel matrix on the input data and $\mathbf{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$. Let $S : \mathcal{H}_K \to \mathbb{R}^N$ be be the sampling operator defined by $Sf = (f(x_1), \ldots, f(x_N))^\top$ for $f \in \mathcal{H}_K$. Its dual operator S^* is given by $S^*\mathbf{c} = \sum_{i=1}^N c_i K(x_i, \cdot) \in \mathcal{H}_K$ for $\mathbf{c} \in \mathbb{R}^N$. In Smale and Zhou (2007) it is proved that $f_{D,\lambda}$ has an operator representation

$$f_{D,\lambda} = \frac{1}{N} \left(\lambda I + \frac{1}{N} S^* S \right)^{-1} S^* \mathbf{y}. \tag{2}$$

One objective of this paper is to study the performance of the distributed version of this method.

4

We need several assumptions that are used throughout the paper. Recall the true regression function is $f^*(x) = \mathbf{E}[y|x]$ and define $\sigma^2 = \mathbf{E}[\text{var}(y|x)] = \mathbf{E}[(y - f^*(x))^2]$. Notice that

$$\text{var}(y) = \mathbf{E}\Big[\text{var}(y|x)\Big] + \text{var}\Big(\mathbf{E}[y|x]\Big) = \sigma^2 + \text{var}\Big(f^*(x)\Big).$$

The second term on the right is the part of variance of y that is explained by the regression function while the first term is unexplained. Our first assumption is on the finiteness of unexplained variance.

Assumption 1. The unexplained variance of the response variable is finite, i.e., $\sigma^2 < \infty$.

If in particular $\sigma^2 = 0$, then $y = f^*(x)$ is determined for each given $x \in \mathcal{X}$ and we call the regression problem is noise free. Note also for any function f independent of (x, y) there holds

$$\mathbf{E}\left[(y - f(x))^2\right] = \mathbf{E}\left[(f^*(x) - f(x))^2\right] + \sigma^2. \tag{3}$$

This identity will be repeatedly used in the proof of our main results.

Next we need the so called source condition. To state it, let $\rho_{\mathcal{X}}$ denote the marginal distribution on $\mathcal{X}$. Define

$$L_K f(x) = \int_{\mathcal{X}} K(x, t) f(t) \mathrm{d}\rho_{\mathcal{X}}(t).$$

Then L_K defines a compact operator both on $L^2_{\rho_{\mathcal{X}}}$ (the space of square integrable functions with respect to the probability measure $\rho_{\mathcal{X}}$) and $\mathcal{H}_K$. Let τ_i and ϕ_i be the eigenvalues and eigenfunctions of L_K as an operator on $L^2_{\rho_{\mathcal{X}}}$. Then $\{\phi_i\}_{i=1}^{\infty}$ form an orthogonal basis of $L^2_{\rho_{\mathcal{X}}}$ and

$$L_K f = \sum_{i=1}^{\infty} \tau_i \langle f, \phi_i \rangle_{L^2_{\rho_{\mathcal{X}}}} \phi_i, \qquad \forall\, f \in L^2_{\rho_{\mathcal{X}}}.$$

Also, $\{\psi_i = \sqrt{\tau_i}\phi_i : \tau_i \neq 0\}$ form an orthonormal basis of $\mathcal{H}_K$ and, as an operator on $\mathcal{H}_K$,

$$L_K f = \sum_{i:\tau_i \neq 0} \tau_i \langle f, \psi_i \rangle_K \psi_i, \qquad \forall\, f \in \mathcal{H}_K.$$

It is easy to verify that L_K is a population version of the operator $\frac{1}{N} S^* S$ (as operators on $\mathcal{H}_K$).

Because L_K is compact and admits an eigen-decomposition form, L_K^r is well defined for all $r > 0$. In particular, let $\overline{\mathcal{H}_K}$ be the closure of $\mathcal{H}_K$ in $L^2_{\rho_{\mathcal{X}}}$. Then for each $f \in \overline{\mathcal{H}_K}$ we have $L_K^{\frac{1}{2}} f \in \mathcal{H}_K$ and

$$\|f\|_{L^2_{\rho_{\mathcal{X}}}} = \|L_K^{\frac{1}{2}} f\|_K. \tag{4}$$

If the kernel K is universal in the sense that $\mathcal{H}_K$ is dense in $C(\mathcal{X})$, then $\overline{\mathcal{H}_K} = L^2_{\rho_{\mathcal{X}}}$ and (4) holds for all $f \in L^2_{\rho_{\mathcal{X}}}$. The source condition is stated as follows.

Assumption 2. There exist some $u^* \in L^2_{\rho_{\mathcal{X}}}$ and $r > 0$ such that $f^* = L_K^r u^*$.

If the source condition holds with $r \geq \frac{1}{2}$ then $f^* \in \mathcal{H}_K$ and the kernel is perfect for learning the regression function. If $r < \frac{1}{2}$, then f^* does not lie in $\mathcal{H}_K$ and the kernel is imperfect. We will focus our study on the imperfect case in this paper.

Before moving on to the distributed KRR, let us first recall that if the source condition holds with $0 < r \leq 1$, KRR can reach the optimal capacity independent rate $O(N^{-\frac{2r}{1+2r}})$ (Zhang, 2003). If $\frac{1}{2} \leq r \leq 1$ and in addition the capacity of the reproducing kernel Hilbert space as measured by the effective dimension satisfies

$$\mathcal{N}(\lambda) = \mathrm{Tr}((\lambda I + L_K)^{-1}L_K) \leq C_0 \lambda^{-\beta} \tag{5}$$

for some $C_0 > 0$ and $0 < \beta < 1$, KRR reaches the minimax optimal capacity dependent rate $O(N^{-\frac{2r}{\beta+2r}})$ (Caponnetto and De Vito, 2007). However, to our best knowledge, if $r < \frac{1}{2}$ meaning that the kernel is imperfect, such minimax optimality has never been verified in the literature. Note all our results in this paper will not assume any capacity conditions.

2.1 Distributed kernel ridge regression

In the context of distributed kernel regression we divide the whole data D into m disjoint subset $D = \bigcup_{\ell=1}^{m} D_\ell$. Without loss of generality we assume all data sets are of equal size $n = N/m$ and denote $D_\ell = \{(x_{\ell,1}, y_{\ell,1}), \ldots, (x_{\ell,n}, y_{\ell,n})\}$. Let $f_{D_\ell,\lambda}$ be the local estimator learned from D_ℓ by using KRR method (1). The distributed KRR defines the final global estimator by

$$\overline{f_{D,\lambda}} = \sum_{\ell=1}^{m} \frac{n}{N} f_{D_\ell,\lambda} = \frac{1}{m} \sum_{\ell=1}^{m} f_{D_\ell,\lambda}. \tag{6}$$

This approach has been studied in Zhang et al. (2015); Lin et al. (2017) and the minimax optimality was verified for $\frac{1}{2} \leq r \leq 1$. In this paper we prove the following capacity independent bounds for all $r > 0$.

Theorem 1 *Assume $\sigma^2 < \infty$ and $f^* = L_K^r u^*$ for some $u^* \in L_{\rho_\mathcal{X}}^2$ and $0 < r \leq 1$.*

(i) *If $0 < r \leq \frac{1}{2}$, then there exists a constant $C_1 > 0$ independent of N, n, m, or λ such that*

$$\mathbf{E}\left[\left\|\overline{f_{D,\lambda}} - f^*\right\|_{L_{\rho_\mathcal{X}}^2}^2\right] \leq C_1 \left\{\frac{m\sigma^2}{N^2\lambda^2} + \frac{\sigma^2}{N\lambda} + \lambda^{2r}\left(1 + \frac{m^2}{N^2\lambda^2} + \frac{m}{N\lambda}\right)\right\}.$$

Consequently, if $m \leq N^{\frac{2r}{1+2r}}$, then with the choice $\lambda = N^{-\frac{1}{1+2r}}$, we have

$$\mathbf{E}\left[\left\|\overline{f_{D,\lambda}} - f^*\right\|_{L_{\rho_\mathcal{X}}^2}^2\right] = O\left(N^{-\frac{2r}{1+2r}}\right).$$

(ii) *If $\frac{1}{2} < r \leq 1$ and $\lambda n \geq 1$, then there exists a constant $C_2 > 0$ such that*

$$\mathbf{E}\left[\left\|\overline{f_{D,\lambda}} - f^*\right\|_{L_{\rho_\mathcal{X}}^2}^2\right] \leq C_2 \left\{\frac{m\sigma^2}{N^2\lambda^2} + \frac{\sigma^2}{N\lambda} + \lambda^{2r} + \frac{\lambda^r m^{\frac{r}{2}+\frac{1}{4}}}{N^{\frac{r}{2}+\frac{1}{4}}}\right\}.$$

Consequently, if $m \leq N^{\frac{4r^2+1}{(1+2r)^2}}$, then with the choice $\lambda = N^{-\frac{1}{1+2r}}$, we have

$$\mathbf{E}\left[\left\|\overline{f_{D,\lambda}} - f^*\right\|_{L_{\rho_\mathcal{X}}^2}^2\right] = O\left(N^{-\frac{2r}{1+2r}}\right).$$

Note if $r > 1$ the error bounds and convergence rates are the same as $r = 1$ due to the saturation effect of KRR and have been omitted. The error bounds in Theorem 1 are sharp and the rates are capacity independent optimal. A detailed comparison with the results in the literature is given in Section 2.3.

2.2 Bias correction

The bias corrected kernel ridge regression (BCKRR) was proposed in Wu (2017) to efficiently handle block wise data, for which the distributed learning is an example. Notice the operator representation (2) implies $f_{D,\lambda}$ has asymptotic bias $-\lambda(\lambda I + L_K)^{-1} f^*$. The BCKRR is formulated by subtracting an empirical estimate of the asymptotic bias:

$$f_{D,\lambda}^\sharp = f_{D,\lambda} + \lambda \left(\lambda I + \frac{1}{N} S^* S \right)^{-1} f_{D,\lambda}. \tag{7}$$

Similar to distributed KRR, a distributed BCKRR can be designed by applying BCKRR on each subset and averaging the local estimators to obtain the global estimator as

$$\overline{f_{D,\lambda}^\sharp} = \frac{1}{m} \sum_{\ell=1}^m f_{D_\ell,\lambda}^\sharp.$$

The bias and variance of BCKRR had been characterized in Wu (2017) and the distributed BCKRR was studied in Guo et al. (2017b). Those studies have shown that the BCKRR benefits the block wise data analysis both theoretically and empirically. Similar to the distributed KRR case, the distributed BCKRR has been shown asymptotically minimax optimal when the kernel is perfect. But when the kernel is imperfect, Guo et al. (2017b) derived the rate of $O(N^{-\frac{r}{\beta+2r}})$ under the capacity condition (5). It implies a capacity independent rate of $O(N^{-\frac{r}{1+2r}})$, which is far from optimal. Even with the capacity condition, if it is weak, say $\beta > \frac{1}{2} - r$, the rate in Guo et al. (2017b) is still worse than the capacity independent rate $O(N^{-\frac{2r}{1+2r}})$. One of our main contributions is to address this question.

Theorem 2 *Assume $\sigma^2 < \infty$ and $f^* = L_K^r u^*$ for some $u^* \in L_{\rho_{\mathcal{X}}}^2$ with some $0 < r \leq 2$. If $\lambda n \geq 1$, then there exists a constant $\widetilde{C} > 0$ independent of N, n, m, or λ such that*

$$\mathbf{E}\left[\left\| \overline{f_{D,\lambda}^\sharp} - f^* \right\|_{L_{\rho_{\mathcal{X}}}^2}^2 \right] \leq \widetilde{C} \begin{cases} \dfrac{\sigma^2}{N\lambda} + \lambda^{2r}, & \text{if } 0 < r \leq \frac{1}{2}; \\[3mm] \dfrac{\sigma^2}{N\lambda} + \lambda^{2r} + \dfrac{\lambda^r m^{\frac{r}{2}+\frac{1}{4}}}{N^{\frac{r}{2}+\frac{1}{4}}}, & \text{if } \frac{1}{2} < r \leq 1; \\[3mm] \lambda^{2r} + \dfrac{\lambda^r m^{\frac{r}{2}+\frac{1}{4}}}{N^{\frac{r}{2}+\frac{1}{4}}} + \dfrac{\lambda^{r+\frac{1}{2}} m^{\frac{1}{2}}}{N^{\frac{1}{2}}} + \dfrac{\sigma^2}{\lambda N}, & \text{if } 1 < r \leq \frac{3}{2}; \\[3mm] \lambda^{2r} + \dfrac{\lambda m}{N} + \dfrac{\lambda^{r+\frac{1}{2}} m^{\frac{1}{2}}}{N^{\frac{1}{2}}} + \dfrac{\sigma^2}{\lambda N}, & \text{if } \frac{3}{2} < r \leq 2. \end{cases}$$

Consequently, if choosing $\lambda = N^{-\frac{1}{2r+1}}$ and the number of local machines satisfies $m \leq N^{\theta}$ with

$$\theta = \begin{cases} \dfrac{2r}{1+2r}, & \text{if } 0 < r \leq \frac{1}{2}; \\[3mm] \dfrac{4r^2+1}{(1+2r)^2}, & \text{if } \frac{1}{2} < r \leq \frac{1+\sqrt{2}}{2}; \\[3mm] \dfrac{2}{1+2r}, & \text{if } \frac{1+\sqrt{2}}{2} < r \leq 2, \end{cases}$$

then we have

$$\mathbf{E}\left[\left\|\overline{f_{D,\lambda}^{\sharp}} - f^*\right\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] = O\left(N^{-\frac{2r}{1+2r}}\right).$$

2.3 Comparison with literature

The minimax rate analysis of regularized least square algorithms has been extensively studied in statistics and learning theory literature. If the eigenvalues of the operator L_K decays as $\tau_i \leq i^{-\frac{1}{\beta}}$, which implies the capacity condition (5), then the minimax learning rate of kernel ridge regression is $O(N^{-\frac{2r}{2r+\beta}})$ for $\frac{1}{2} \leq r \leq 1$; see e.g. Caponnetto and De Vito (2007); Lin et al. (2017); Blanchard and Mücke (2018). To our best knowledge, this rate has never been proved for $r < \frac{1}{2}$ without imposing additional conditions.

Note that the capacity condition (5) roughly measures the smoothness of kernel K. The smoother the kernel is, the smaller the β. For example, it is proved in Mendelson and Neeman (2010) that, if $K \in C^{\alpha}(\mathcal{X}^2 \times \mathcal{X}^2)$ with some integer $\alpha \geq 1$ and $\mathcal{X}^2$ is locally the graph of a Lipschitz function, then (5) is satisfied with $\beta = \left(\frac{\alpha}{2\dim(\mathcal{X})} + \frac{1}{2}\right)^{-1}$. Moreover, the Mercer theorem guarantees all kernels satisfies (5) with $\beta = 1$. Therefore, $\beta = 1$ corresponds to the worst situation and imposing a capacity condition (5) with $\beta = 1$ is equivalent to no capacity condition. We see the rate $O(N^{-\frac{2r}{2r+1}})$ is minimax optimal in a capacity independent sense. It is proved in Zhang (2003) that kernel ridge regression can achieve this capacity independent optimal rate when $r \leq \frac{1}{2}$.

Concerning the distributed kernel ridge regression via the divide and conquer technique, under the assumptions that $\mathbf{E}[|\phi_i(x)|^{2k}] \leq A^{2k}$ for some $k > 2$ and constant $A < \infty$, $\lambda_i \leq ai^{-\frac{1}{\beta}}$, and $f^* \in \mathcal{H}_K$ (i.e. $r = \frac{1}{2}$), it is proved in Zhang et al. (2015) that the optimal learning rate of $O(N^{-\frac{1}{1+\beta}})$ can be achieved by restricting the number of local processors

$$m \leq c_{\alpha} \left(\frac{N^{\frac{k-4-k\beta}{1+\beta}}}{A^{4k} \log^k N}\right)^{\frac{1}{k-2}}.$$

This unfortunately does not apply to the case $\beta = 1$. Later in Lin et al. (2017) the minimax optimal rate was verified for all $r \in [\frac{1}{2}, 1]$ if $m \leq N^{\min(\frac{3(2r-1)+\beta}{5(2r+\beta)}, \frac{2r-1}{2r+\beta})}$ and in Guo et al. (2017a) the restriction was relaxed to $m \leq N^{\min(\frac{2}{2r+\beta}, \frac{2r-1}{2r+\beta})}$. Although their results apply to $\beta = 1$, the restriction on m is strict. If $r = \frac{1}{2}$, the data is essentially not able to be distributed according to their results. The error bounds in Lin et al. (2017); Guo et al.

(2017a) also apply to $r \leq \frac{1}{2}$ but only lead to suboptimal rate $O(N^{-\frac{2r}{1+\beta}})$ with the restriction $m = O(1)$. If $\beta = 1$, the capacity independent rate is $O(N^{-r})$.

In this paper our primary interest is to understand the performance of distributed kernel regression when the kernel is imperfect, i.e., when $r < \frac{1}{2}$ and thus the true regression function f^* is not in $\mathcal{H}_K$. For this case, Theorem 1 (i) tells that, with the restriction $m \leq N^{\frac{2r}{1+2r}}$, the capacity independent optimal rate $O(N^{-\frac{2r}{1+2r}})$ for $r \leq \frac{1}{2}$ can be achieved. Note this rate is even faster than the existing capacity dependent rate $O(N^{-\frac{2r}{1+\beta}})$ when $r < \frac{\beta}{2}$, let alone the capacity independent rate $O(N^{-r})$. In other words, we proved faster rates under weaker restrictions on m. More importantly, unlike previous results that requires $m = O(1)$ and thus disclaims the effectiveness of distributed kernel regression algorithms in this situation, our results instead verified their feasibility.

If f^* lies in $\mathcal{H}_K$, i.e. $r \geq \frac{1}{2}$, it is as expected that the rate $O(N^{-\frac{2r}{1+2r}})$ in Theorem 1 is worse than the minimax capacity dependent rate in the literature. But comparing the restrictions on the number of local machines, we see our restriction is much relaxed. In particular, when r is close to $\frac{1}{2}$, the restrictions in the literature approaches to $O(1)$. But our restriction still allows about $O(\sqrt{N})$ local machines while preserving fast rates. It is a sacrifice of convergence rates for more local machines and may be useful for applications where analysis of super big data is necessary.

Concerning the distributed BCKRR, similar conclusions can be made. When $r < \frac{1}{2}$, the rate in Theorem 2 is faster and the restriction is more relaxed than the results in Guo et al. (2017b) while when $r \geq \frac{1}{2}$ the rate is worse than those in Guo et al. (2017b) but the restriction on m is greatly relaxed. As noted in Wu (2017); Guo et al. (2017b) a main theoretical advantage of BCKRR is to relax the saturation effect of KRR. Comparing Theorem 2 with Theorem 1, we see the the rate of distributed BCKRR continues improving beyond $r > 1$ and ceases to improve until $r = 2$. The restriction on m also continues to get relaxed to $N^{\frac{\sqrt{2}}{1+\sqrt{2}}}$ until $r = \frac{1+\sqrt{2}}{2}$.

It is worth mentioning that Shi (2019) studied coefficient regularization approach for distributed kernel regression and obtained capacity dependent optimal rate $O(N^{-\frac{2r}{2r+\beta}})$ for $\frac{1}{2} - \frac{\beta}{2} \leq r < \frac{1}{2}$, which implies the capacity independent optimal rate $O(N^{-\frac{2r}{2r+1}})$ for all $0 < r < \frac{1}{2}$. But the requirement $m < N^{-\frac{2r}{1+2r}} / \log N$ is slightly more restrictive. Therefore, on the one hand, the techniques there may be adapted to analyze DKRR and DBCKRR for capacity dependent optimal rates under the same range of r. On the other hand, they are not supposed to supersede leave one out analysis and provide sharpest capacity independent error bounds.

2.4 Empirical effectiveness of imperfect kernels

In this subsection we illustrate the empirical effectiveness of using imperfect kernels in distributed regression. To this end, we adopt the example used in Zhang et al. (2015); Guo et al. (2017b). The true regression function is given by $f^*(x) = \min(x, 1 - x)$ with $x \sim \text{Uniform}[0, 1]$ and the observations are generated by the additive noise model $y_i = f^*(x_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma^2 = \frac{1}{5}$. We consider two kernels: the Sobolev space kernel $K_S(x, t) = 1 + \min(x, t)$ and the Guassian kernel $K_G(x, t) = \exp(-(x - t)^2/0.3)$.

Recall that f^* belongs to $\mathcal{H}_{K_S}$ with $\|f^*\|_{K_S} = 1$. So K_S is a perfect kernel for this problem. Notice that the reproducing kernel Hilbert space associated to a Gaussian kernel consists only infinitely differentiable functions and even polynomials may not lie in the space (Minh, 2010). We conclude f^* does not lie in $\mathcal{H}_{K_G}$ and K_G is an imperfect kernel. We generate $N = 4098$ sample points and use number of partitions $m \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. Mean squared errors between the estimated function and the true regression function is used to measure the performance.

According to Zhang et al. (2015), if the Soboleve space kernel K_S is used, the theoretically optimal choice of the regularization parameter is $\lambda = N^{-\frac{2}{3}}$. When Guassian kernel is used, by Theorem 1 and Theorem 2, the optimal choice of λ should depend on the index r in the source condition, which, unfortunately, is unknown. By $0 < r < \frac{1}{2}$ we know the optimal choice should be $N^{-\alpha}$ with $\frac{1}{2} < \alpha < 1$ and $\alpha = \frac{2}{3}$ seems an acceptable choice. So we will also use $\lambda = N^{-\frac{2}{3}}$ for the Guassian kernel to make the first comparison between the four distributed kernel regression algorithms, namely, distributed KRR with Sobolev space kernel (DKRR-S), distributed BCKRR with Sobolev space kernel (DBCKRR-S), distributed KRR with Guassian kernel (DKRR-G), and distributed BCKRR with Gaussian kernel (DBCKRR-G). To do this, for each aforementioned m value and each algorithm, we repeat the experiment 50 times and report the mean squared errors in Figure 1(a). The results indicate that, even if K_G is imperfect for the problem, it performs comparable with the perfect kernel K_S when m is small and may even outperforms K_S when m becomes large.

To our best knowledge, all rate analysis literature of distributed kernel regression, including Zhang et al. (2015); Lin et al. (2017); Lin and Zhou (2018); Guo et al. (2017b,a) and this study, suggest the optimal regularization parameter be selected as $\lambda = N^{-\alpha}$ with α an index depending on the regularity of the true regression function f^*. While this is very helpful for researchers to understand the optimality of the algorithms, it is less informative for their practical use because such an index is unknown for real problems. Furthermore, distributed kernel regression becomes necessary only if the data is too big to be processed by a single machine. In this situation, it is imaginable that globally tuning the optimal parameter is either impossible or too time consuming. At the same time, note that distributed kernel regression requires underregularization, meaning that the regularization parameter must be chosen according to the total sample size N, not on the local sample size n. To resolve all these problems, Guo et al. (2017b) proposed a practical strategy to tune the parameter for distributed kernel regression. It first cross-validates the regularization parameter locally to get optimal choice $\lambda_{\ell,n} = n^{-\alpha_\ell}$ for each subset for D_ℓ and then use an underregularized parameter

$$\lambda_\ell = \lambda_{\ell,n}^{\frac{\log N}{\log n}} = N^{-\alpha_\ell} \tag{8}$$

to train the local model on D_ℓ. To mimic a real problem, now let us assume we have no information about the true regression function and the regularization parameters have to be selected by using the above tuning strategy. We run the four algorithms again for each m value and report the mean squared errors in Figure 1(b). We see that, when a theoretically optimal regularization parameter is not available and λ has to be tuned from data, the performance of all four methods deteriorates faster as m increases. But Guassian

kernel, though imperfect, seems less sensitive to the number of local machines and so the performance deteriorates slower than Sobolev space kernel.

Finally, the results in both plots indicate that, regardless the choices of kernels and regularization parameters, bias correction always helps to improve the learning performance and relax the restriction on the number of local machines.

Figure 1: Mean squared error of distributed kernel regression with Sobolev space kernel and Guassian kernel when (a) the regularization parameters is fixed as $\lambda = N^{-\frac{2}{3}}$ and (b) the regularization parameter is locally tuned and underregularized according to equation (8).
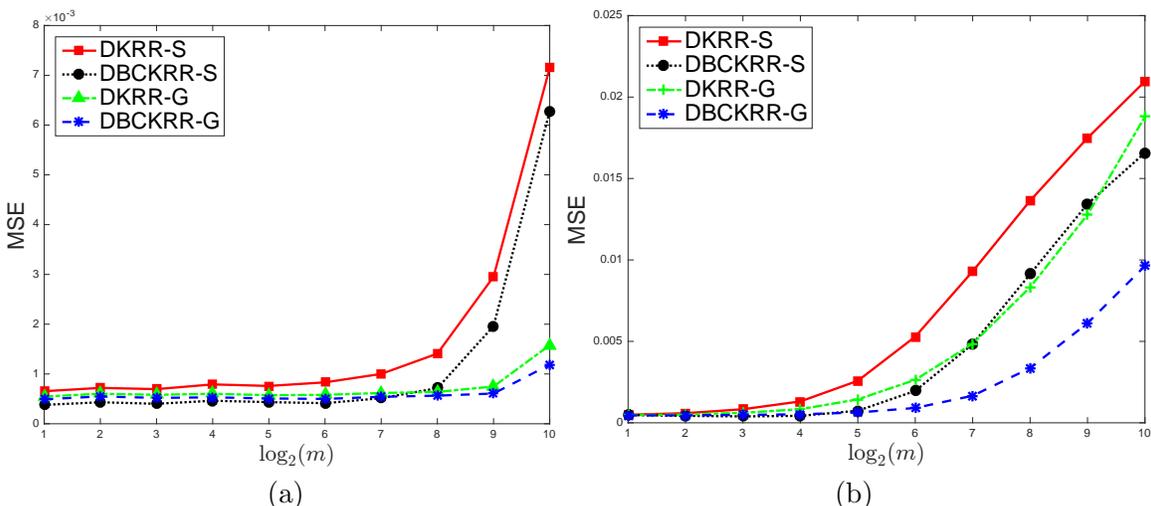
3. A framework for fast rate analysis of distributed regression

In this section we first establish a general procedure to prove fast rate for a class of regression algorithms that possess certain special features and apply to the two algorithms we study in this paper.

Definition 3 *Let Λ be a parameter space and $\mathcal{H}$ a set of hypothesis functions. A regression algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^N \times \Lambda \to \mathcal{H}$ that tunes parameters in Λ is* response weighted *if for any data $D \in (\mathcal{X} \times \mathcal{Y})^N$ and any $\lambda \in \Lambda$ there exists a vector of functions $g_1(D_x, \lambda), \dots, g_N(D_x, \lambda)$ which depend only on the input data $D_x = \{x_1, \dots, x_N\}$ and the parameter λ but not on the output data $D_y = \{y_i\}_{i=1}^N$ such that*

$$\mathcal{A}(D, \lambda) = \sum_{i=1}^{N} y_i g_i(D_x, \lambda).$$

There are many regression algorithms belonging to the response weighted family. By the representation of $f_{D,\lambda}$ and $f_{D,\lambda}^{\sharp}$, it is easy to verify that both KRR and BCKRR belong to this family. There are more examples in the literature. For instance, traditional multiple

11

linear regression, the kernel smooth estimators, the stochastic gradient descent algorithms associated with the least square loss are all response weighted regression algorithms.

Denote by $\overline{\mathcal{A}}$ the distributed algorithm that applies the divide and conquer approach and uses $\mathcal{A}$ as the base algorithm on local subsets. It is easy to derive the following lemma.

Lemma 4 *If $\mathcal{A}$ is a response weighted regression algorithm, then $\overline{\mathcal{A}}$ is also a response weighted regression algorithm.*

This lemma simply says that the response weighted feature can be inherited by the distributed algorithm. This property allows to derive fast rates for such distributed algorithms by studying error bounds of the base algorithm, as shown in the following theorem.

Theorem 5 *Let $\mathcal{A}$ be a response weighted regression algorithm and, when applying to a data D of N observations with parameter λ, have the error bound as*

$$\mathbf{E}\left[\|\mathcal{A}(D, \lambda) - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \varepsilon(N, \lambda, \sigma^2).$$

Then the corresponding distributed algorithm $\overline{\mathcal{A}}$ will have the error bound

$$\mathbf{E}\left[\|\overline{\mathcal{A}}(D, \lambda) - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \frac{\varepsilon(n, \lambda, \sigma^2)}{m} + \varepsilon(n, \lambda, 0)$$

assuming the whole data D is equally split into m subsets with each subset containing n observations.

Proof Let $y^*_{\ell,i} = f^*(x_{\ell,i})$ for $1 \leq \ell \leq m$, $1 \leq i \leq n$ and

$$D^*_\ell = \{(x_{\ell,1}, y^*_{\ell,1}), \ldots, (x_{\ell,n}, y^*_{\ell,n})\}$$

be the noise free observations associated to the subset D_ℓ. We have

$$(\sigma^*)^2 = \mathbf{E}\left[\text{var}\left(y^*_{\ell,i}|x_{\ell,i}\right)\right] = \mathbf{E}\left[\left(y^*_{\ell,i} - f^*(x_{\ell,i})\right)^2\right] = 0$$

and therefore

$$\mathbf{E}\left[\left\|\mathcal{A}(D^*_\ell, \lambda) - f^*\right\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \varepsilon(n, \lambda, 0). \tag{9}$$

Because of the response weighted feature of $\mathcal{A}$ we have

$$\mathbf{E}\left[\mathcal{A}(D_\ell, \lambda)|D_{\ell,x}\right] = \sum_{i=1}^n \mathbf{E}[y_{\ell,i}|x_{\ell,i}]g_i(D_{\ell,x}, \lambda) = \sum_{i=1}^n f^*(x_{\ell,i})g_i(D_{\ell,x}, \lambda) = \mathcal{A}(D^*_\ell, \lambda).$$

Therefore $\mathbf{E}\left[\mathcal{A}(D_\ell, \lambda)\right] = \mathbf{E}\left[\mathcal{A}(D^*_\ell, \lambda)\right]$ and we have

$$\mathbf{E}\left[\|\overline{\mathcal{A}}(D, \lambda) - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right]$$

$$= \mathbf{E}\left[\left\|\frac{1}{m}\sum_{\ell=1}^{m}\left(\mathcal{A}(D_\ell,\lambda) - \mathbf{E}\left[\mathcal{A}(D_\ell,\lambda)\right]\right)\right\|_{L^2_{\rho_\mathcal{X}}}^2\right] + \left\|\frac{1}{m}\sum_{\ell=1}^{m}\left(\mathbf{E}\left[\mathcal{A}(D_\ell,\lambda)\right] - f^*\right)\right\|_{L^2_{\rho_\mathcal{X}}}^2$$

$$\leq \frac{1}{m^2}\sum_{\ell=1}^{m}\mathbf{E}\left[\left\|\mathcal{A}(D_\ell,\lambda) - \mathbf{E}\left[\mathcal{A}(D_\ell,\lambda)\right]\right\|_{L^2_{\rho_\mathcal{X}}}^2\right] + \frac{1}{m}\sum_{\ell=1}^{m}\left\|\mathbf{E}\left[\mathcal{A}(D_\ell,\lambda)\right] - f^*\right\|_{L^2_{\rho_\mathcal{X}}}^2$$

$$= \frac{1}{m^2}\sum_{\ell=1}^{m}\mathbf{E}\left[\left\|\mathcal{A}(D_\ell,\lambda) - f^*\right\|_{L^2_{\rho_\mathcal{X}}}^2\right] + \left(\frac{1}{m} - \frac{1}{m^2}\right)\sum_{\ell=1}^{m}\left\|\mathbf{E}\left[\mathcal{A}(D_\ell,\lambda)\right] - f^*\right\|_{L^2_{\rho_\mathcal{X}}}^2$$

$$\leq \frac{1}{m^2}\sum_{\ell=1}^{m}\varepsilon(n,\lambda,\sigma^2) + \frac{1}{m}\sum_{\ell=1}^{m}\mathbf{E}\left[\left\|\mathcal{A}(D_\ell^*,\lambda) - f^*\right\|_{L^2_{\rho_\mathcal{X}}}^2\right]$$

$$\leq \frac{1}{m}\varepsilon(n,\lambda,\sigma^2) + \varepsilon(n,\lambda,0),$$

where we used (9) in the last step. ■

Theorem 5 states that, if the base algorithm $\mathcal{A}$ admits sharper error bounds for the noise free data (i.e. $\sigma^2 = 0$), then the distributed algorithm will be able to converges fast in the sense that the distributed computing is playing a rule and produces an estimator better than the algorithm on a single subset.

In order to use this general framework to derive error bounds and fast rates for distributed algorithm, however, we must study the performance of the base algorithm on a single data with full exploration of impact of the unexplained variance σ^2. Although both KRR and BCKRR have already been studied in the literature, the error bounds that clearly involve σ^2 and allow us to derive error bounds of their distributed version with imperfect kernels are not available. They will be our main objectives in Sections 5 and 6.

4. Preliminaries

In this section we provide some notations and preliminary lemmas that will be used in the proofs. For any regularization parameter $\lambda > 0$, define

$$f_\lambda = \arg\min_{f \in \mathcal{H}_K}\left\{\|f - f^*\|_{L^2_{\rho_X}}^2 + \lambda\|f\|_K^2\right\}. \tag{10}$$

It is a sample limit version of the KRR estimation $f_{D,\lambda}$ and admits an operator representation

$$f_\lambda = (\lambda I + L_K)^{-1}L_K f^*.$$

It plays an essential role to characterize the approximation error of the two algorithms under study. The following lemma is well known.

Lemma 6 *Under the source condition $f^* = L_K^r u^*$ for some $r > 0$ and $u^* \in L_{\rho_\mathcal{X}}^2$, there hold*

$$\|f_\lambda - f^*\|_{L_{\rho_\mathcal{X}}^2}^2 \leq \lambda^{\min(2r,2)} \|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 ,$$

and

$$\lambda\|f_\lambda\|_K^2 \leq \lambda^{\min(2r,1)} \|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 .$$

In the analysis of BCKRR, we will need the sample limit of $f_{D,\lambda}^\sharp$ defined by

$$f_\lambda^\sharp = f_\lambda + \lambda(\lambda I + L_K)^{-1} f_\lambda = (\lambda I + L_K)^{-2}(2\lambda L_K + L_K^2)f^*. \tag{11}$$

For this function, we have the following approximation property (Guo et al., 2017b).

Lemma 7 *Under the source condition $f^* = L_K^r u^*$ for some $r > 0$ and $u^* \in L_{\rho_\mathcal{X}}^2$, there holds*

$$\|f_\lambda^\sharp - f^*\|_{L_{\rho_\mathcal{X}}^2}^2 \leq \lambda^{\min(2r,4)} \|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 .$$

The following lemma can be derived from operator monotone property; see e.g. Sun and Wu (2009b).

Lemma 8 *If T_1 and T_2 are two bounded self-adjoint positive operators, then for any $\alpha \in [0,1]$ there holds*

$$\|T_1^\alpha - T_2^\alpha\| \leq \|T_1 - T_2\|^\alpha.$$

The Lemma 9 below follows from simple calculation.

Lemma 9 *Let ξ be a random variable with values in a Hilbert space $\mathcal{H}$ and $\{\xi_1, \ldots, \xi_N\}$ be a set of i.i.d. observations for ξ. Then*

$$\mathbf{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \xi_i - \mathbf{E}\xi\right\|^2\right] \leq \frac{\mathbf{E}[\|\xi\|^2]}{N}.$$

Let $\text{HS}(\mathcal{H}_K)$ be the class of Hilbert-Schmidt operators on $\mathcal{H}_K$. It is known $\text{HS}(\mathcal{H}_K)$ forms a Hilbert space with the Hilbert-Schmidt norm which, give an operator A on $\mathcal{H}_K$, is defined by $\|A\|_{\text{HS}}^2 = \text{Trace}(A^*A)$. Then L_K as an operator on $\mathcal{H}_K$ belongs to $\text{HS}(\mathcal{H}_K)$ and $\|L_K\|_{\text{HS}} \leq \kappa^2$; see e.g. Sun and Wu (2009a). Define the rank one operator $K_x \otimes K_x$ by

$$K_x \otimes K_x f = \langle f, K_x \rangle_K K_x = f(x)K_x.$$

Then $\|K_x \otimes K_x\|_{\text{HS}} \leq \kappa^2$. Note for any Hilbert-Schmidt operator A there holds $\|A\| \leq \|A\|_{\text{HS}}$. Applying Lemma 9 to $\xi = K_x \otimes K_x$ we have the following lemma.

Lemma 10 *We have*

$$\mathbf{E}\left[\left\|\frac{1}{N}S^*S - L_K\right\|^2\right] \leq \frac{\kappa^4}{N}$$

Define
$$Q(\lambda, N) = \min_{f \in \mathcal{H}_K} \left\{ \|f - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 + \lambda \|f\|_K^2 \right\} - \lambda \mathbf{E}\left[\|f_{D,\lambda}\|_K^2 \right].$$

Note that $Q(\lambda, N)$ depends on the regularization parameter λ and sample size N, not on the data D. We will need a sharp bound for $Q(\lambda, N)$.

Lemma 11 *Assume $f^* = L_K^r u^*$ for some $r > 0$, $u^* \in L^2_{\rho_{\mathcal{X}}}$.*

(i) *If $0 < r \leq \frac{1}{2}$, then*
$$Q(\lambda, N) \leq 2\lambda^{2r} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2.$$

(ii) *If $\frac{1}{2} < r \leq 1$, then*
$$Q(\lambda, N) \leq \lambda^{2r} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{2\kappa^{2r}}{\lambda^r N^{\frac{r}{2} + \frac{1}{4}}} + \frac{2\kappa}{\sqrt{\lambda N}} \right).$$

(iii) *If $1 < r \leq \frac{3}{2}$, then*
$$Q(\lambda, N) \leq \lambda^2 \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{2\kappa^{2r}}{\lambda N^{\frac{r}{2} + \frac{1}{4}}} + \frac{2\kappa \lambda^{r-1}}{\sqrt{\lambda N}} \right).$$

(iv) *If $r > \frac{3}{2}$, then*
$$Q(\lambda, N) \leq \lambda^2 \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{2\kappa^3}{\lambda N} + \frac{2\kappa}{\sqrt{N}} \right).$$

Proof By the definition (10) we see
$$Q(\lambda, N) = \|f_\lambda - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 + \lambda \|f_\lambda\|_K^2 - \lambda \mathbf{E}\left[\|f_{D,\lambda}\|_K^2 \right]. \tag{12}$$

We see conclusion (i) follows easily from Lemma 6.

To prove conclusions (ii)-(iv), the key is to bound $\|f_\lambda\|_K^2 - \mathbf{E}\left[\|f_{D,\lambda}\|_K^2 \right]$. To this end, first note that

$$
\begin{aligned}
\|f_\lambda\|_K^2 - \mathbf{E}\left[\|f_{D,\lambda}\|_K^2 \right] &= \mathbf{E}\left[\langle f_\lambda - f_{D,\lambda}, \ f_\lambda + f_{D,\lambda} \rangle_K \right] \\
&= 2\mathbf{E}\left[\langle f_\lambda - f_{D,\lambda}, \ f_\lambda \rangle_K - \|f_\lambda - f_{D,\lambda}\|_K^2 \right] \\
&\leq 2\mathbf{E}\left[\langle f_\lambda - f_{D,\lambda}, \ f_\lambda \rangle_K \right] \\
&= 2 \langle f_\lambda - \mathbf{E}[f_{D,\lambda}], \ f_\lambda \rangle_K.
\end{aligned}
$$

By the operator representation (2) of $f_{D,\lambda}$ we have
$$\mathbf{E}[f_{D,\lambda}] = \mathbf{E}\left[\frac{1}{N} \left(\lambda I + \frac{1}{N} S^* S \right)^{-1} S^* S f^* \right].$$

It is easy to verify that $f_\lambda - \mathbf{E}[f_{D,\lambda}] = \mathbf{E}\left[\left(\lambda I + \frac{1}{N} S^* S \right)^{-1} U \right]$ with $U = \left(\frac{1}{N} S^* S - L_K \right)(f_\lambda - f^*)$. Applying Lemma 9 to the $\mathcal{H}_K$-valued random variable $\xi = (f_\lambda(x) - f^*(x)) K_x$ and by Lemma 6 we have
$$\mathbf{E}\left[\|U\|_K^2 \right] \leq \kappa^2 \|f_\lambda - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 N^{-1} \leq \kappa^2 \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \lambda^{\min(2r,2)} N^{-1}.$$

Then, by the assumption $f^* = L_K^r u^*$ for $\frac{1}{2} < r \leq \frac{3}{2}$ and using Lemma 8, Hölder's inequality, and Lemma 10, we obtain

$$
\begin{aligned}
\langle f_\lambda - \mathbf{E}[f_{D,\lambda}], \, f_\lambda \rangle_K &= \mathbf{E}\left[\left\langle \left(\lambda I + \frac{1}{N}S^*S\right)^{-1} U, \, \left(\lambda I + L_K\right)^{-1} L_K^{1+r} u^* \right\rangle_K\right]\\
&= \mathbf{E}\left[\left\langle L_K^{r+\frac{1}{2}}(\lambda I + L_K)^{-1}\left(\lambda I + \frac{1}{N}S^*S\right)^{-1} U, \, L_K^{\frac{1}{2}} u^* \right\rangle_K\right]\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}} \mathbf{E}\left[\left\| L_K^{r-\frac{1}{2}}\left(\lambda I + \frac{1}{N}S^*S\right)^{-1} U \right\|_K\right] \quad\quad (13)\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}} \mathbf{E}\left[\left(\lambda^{-1}\left\| L_K^{r-\frac{1}{2}} - \left(\frac{1}{N}S^*S\right)^{r-\frac{1}{2}}\right\| + \lambda^{r-\frac{3}{2}}\right)\|U\|_K\right]\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}} \mathbf{E}\left[\left(\lambda^{-1}\left\| L_K - \frac{1}{N}S^*S\right\|^{r-\frac{1}{2}} + \lambda^{r-\frac{3}{2}}\right)\|U\|_K\right]\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}} \left\{ \lambda^{-1}\left(\mathbf{E}\left[\left\| L_K - \frac{1}{N}S^*S\right\|^2\right]\right)^{\frac{r}{2}-\frac{1}{4}} + \lambda^{r-\frac{3}{2}}\right\} \left(\mathbf{E}\left[\|U\|_K^2\right]\right)^{\frac{1}{2}}\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}}^2 \left(\frac{\kappa^{2r}\lambda^{\min(r-1,0)}}{N^{\frac{r}{2}+\frac{1}{4}}} + \frac{\kappa\lambda^{\min(2r-\frac{3}{2},r-\frac{1}{2})}}{N^{\frac{1}{2}}}\right).
\end{aligned}
$$

Plugging this estimate into (12) and applying Lemma 6 again we obtain the desired bounds in conclusion (ii) and conclusion (iii).

If $r > \frac{3}{2}$, then by (13), we have

$$
\begin{aligned}
\langle f_\lambda - \mathbf{E}[f_{D,\lambda}], \, f_\lambda \rangle_K &\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}} \mathbf{E}\left[\left(\lambda^{-1}\left\| L_K - \frac{1}{N}S^*S\right\| + 1\right)\|U\|_K\right]\\
&\leq \|u^*\|_{L^2_{\rho_\mathcal{X}}}^2 \left(\frac{\kappa^3}{N} + \frac{\kappa\lambda}{N^{\frac{1}{2}}}\right).
\end{aligned}
$$

The conclusion (iv) follows by plugging this estimate into (12). ∎

Next we define

$$
Q^\sharp(\lambda, N) = \|f_\lambda^\sharp - f^*\|_{L^2_{\rho_\mathcal{X}}}^2 + \lambda\left(\mathbf{E}\left[\|f_\lambda^\sharp - f_{D,\lambda}\|_K^2\right] - \mathbf{E}\left[\|f_{D,\lambda}^\sharp - f_{D,\lambda}\|_K^2\right]\right).
$$

It will be used to derive the error bound for BCKRR when $r \geq 1$.

Lemma 12 *Assume $f^* = L_K^r u^*$ for some $r \geq \frac{1}{2}$ and $u^* \in L^2_{\rho_\mathcal{X}}$. If $\lambda \leq 1$, then*

$$
Q^\sharp(\lambda, N) \leq 20(\kappa^{2r} + \kappa^2 + \kappa)\|u^*\|_{L^2_{\rho_\mathcal{X}}}^2 \left(\lambda^{\min(2r,4)} + \frac{\lambda^{\min(2r-1,1)}}{N} + \frac{\lambda^{\min(r,1)}}{N^{\min(\frac{r}{2}+\frac{1}{4},1)}}\right.
$$

$$+\frac{\lambda^{\min(2r-\frac{1}{2},r+\frac{1}{2},2)}}{N^{\frac{1}{2}}}\bigg)+\frac{4\kappa^2\sigma^2}{\lambda N}.$$

Proof First note that

$$\mathbf{E}\left[\|f_\lambda^\sharp-f_{D,\lambda}\|_K^2\right]-\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f_{D,\lambda}\|_K^2\right]\le 2\mathbf{E}\left[\left\langle f_\lambda^\sharp-f_{D,\lambda}^\sharp,\ f_\lambda^\sharp-f_{D,\lambda}\right\rangle_K\right].$$

Let

$$f_{\mathbf{x},\lambda}^\sharp=\mathbf{E}\left[f_{D,\lambda}^\sharp|D_x\right]=\left(\lambda I+\frac{1}{N}S^*S\right)^{-2}\left(2\lambda I+\frac{1}{N}S^*S\right)\frac{1}{N}S^*Sf^*$$

and

$$f_{\mathbf{x},\lambda}=\mathbf{E}\left[f_{D,\lambda}|D_x\right]=\left(\lambda I+\frac{1}{N}S^*S\right)^{-1}\frac{1}{N}S^*Sf^*.$$

Then we have that

$$\mathbf{E}\left[\left\langle f_{D,\lambda}^\sharp,\ f_{D,\lambda}\right\rangle_K\right]-\mathbf{E}\left[\left\langle f_{\mathbf{x},\lambda}^\sharp,\ f_{\mathbf{x},\lambda}\right\rangle_K\right]$$

$$=\mathbf{E}\left[\left\langle\left(\lambda I+\frac{1}{N}S^*S\right)^{-2}\left(2\lambda I+\frac{1}{N}S^*S\right)\frac{1}{N}S^*(\mathbf{y}-Sf^*),\left(\lambda I+\frac{1}{N}S^*S\right)^{-1}\frac{1}{N}S^*(\mathbf{y}-Sf^*)\right\rangle_K\right]$$

$$=\mathbf{E}\left[\left\|\left(\lambda I+\frac{1}{N}S^*S\right)^{-1}\frac{1}{N}S^*(\mathbf{y}-Sf^*)\right\|_K^2\right]+\mathbf{E}\left[\left\|\lambda^{\frac{1}{2}}\left(\lambda I+\frac{1}{N}S^*S\right)^{-\frac{3}{2}}\frac{1}{N}S^*(\mathbf{y}-Sf^*)\right\|_K^2\right]$$

$$\le 2\lambda^{-2}\mathbf{E}\left[\left\|\frac{1}{N}S^*(\mathbf{y}-Sf^*)\right\|_K^2\right]\le\frac{2\kappa^2\sigma^2}{\lambda^2 N}.$$

Therefore,

$$\mathbf{E}\left[\|f_\lambda^\sharp-f_{D,\lambda}\|_K^2\right]-\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f_{D,\lambda}\|_K^2\right]\le 2\mathbf{E}\left[\left\langle f_\lambda^\sharp-f_{\mathbf{x},\lambda}^\sharp,\ f_\lambda^\sharp-f_{\mathbf{x},\lambda}\right\rangle_K\right]+\frac{4\kappa^2\sigma^2}{\lambda^2 N}.$$

By the fact $\lambda^2 f_\lambda^\sharp=(2\lambda L_K+L_K^2)(f^*-f_\lambda^\sharp)$ we have

$$f_\lambda^\sharp-f_{\mathbf{x},\lambda}^\sharp=\left(\lambda I+\frac{1}{N}S^*S\right)^{-2}\left\{\left(2\lambda I+\frac{1}{N}S^*S\right)\frac{1}{N}S^*S-(2\lambda I+L_K)L_K\right\}(f_\lambda^\sharp-f^*)$$

$$=\left(\lambda I+\frac{1}{N}S^*S\right)^{-2}\left(2\lambda I+\frac{1}{N}S^*S\right)\left(\frac{1}{N}S^*S-L_K\right)(f_\lambda^\sharp-f^*)$$

$$+\left(\lambda I+\frac{1}{N}S^*S\right)^{-2}\left(\frac{1}{N}S^*S-L_K\right)L_K(f_\lambda^\sharp-f^*).$$

Applying Lemma 9 to the random variable $\xi=(f^*(x)-f_\lambda^\sharp(x))K_x$ and by Lemma 7 we obtain

$$\mathbf{E}\left[\left\|\left(\frac{1}{N}S^*S-L_K\right)(f_\lambda^\sharp-f^*)\right\|_K^2\right]\le\frac{\kappa^2\|f_\lambda^\sharp-f^*\|_{L_{\rho_\mathcal{X}}^2}^2}{N}\le\frac{\kappa^2\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2\lambda^{\min(2r,4)}}{N}.$$

Similarly, we have

$$\mathbf{E}\left[\left\|\left(\frac{1}{N}S^*S-L_K\right)L_K(f_\lambda^\sharp-f^*)\right\|_K^2\right]\le\frac{\kappa^2\|L_K(f_\lambda^\sharp-f^*)\|_{L_{\rho_\mathcal{X}}^2}^2}{N}\le\frac{\kappa^2\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 p^2\lambda^{\min(2r+2,4)}}{N}.$$

17

Hence we have

$$\mathbf{E}\left[\left\|f_\lambda^\sharp - f_{\mathbf{x},\lambda}^\sharp\right\|_K^2\right] \le \frac{10\kappa^2\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 \lambda^{\min(2r-2,0)}}{N}.$$

By Lemma 8 and Lemma 10 we have

$$\mathbf{E}\left[\left\|f_{\mathbf{x},\lambda}^\sharp - f_{\mathbf{x},\lambda}\right\|_K^2\right]$$

$$= \mathbf{E}\left[\left\|\lambda\Big(\lambda I + \frac{1}{N}S^*S\Big)^{-2}\frac{1}{N}S^*SL_K^r u^*\right\|_K^2\right]$$

$$\le 2\lambda^2\mathbf{E}\left[\left\|\Big(\lambda I + \frac{1}{N}S^*S\Big)^{-2}\frac{1}{N}S^*S\Big(L_K^{\min(r-\frac{1}{2},1)} - \big(\frac{1}{N}S^*S\big)^{\min(r-\frac{1}{2},1)}\Big)L_K^{\max(r-1,\frac{1}{2})}u^*\right\|_K^2\right]$$

$$+ 2\lambda^2\mathbf{E}\left[\left\|\Big(\lambda I + \frac{1}{N}S^*S\Big)^{-2}\big(\frac{1}{N}S^*S\big)^{\min(r+\frac{1}{2},2)}L_K^{\max(r-1,\frac{1}{2})}u^*\right\|_K^2\right]$$

$$\le 2\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 \kappa^{\max(4r-6,0)}\left(\mathbf{E}\left[\left\|L_K^{\min(r-\frac{1}{2},1)} - \big(\frac{1}{N}S^*S\big)^{\min(r-\frac{1}{2},1)}\right\|^2\right] + \lambda^{\min(2r-1,2)}\right)$$

$$\le 2\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2 \kappa^{\max(4r-6,0)}\left(\frac{\kappa^{\min(4r-2,4)}}{N^{\min(r-\frac{1}{2},1)}} + \lambda^{\min(2r-1,2)}\right).$$

So we have

$$\mathbf{E}\left[\left\langle f_\lambda^\sharp - f_{\mathbf{x},\lambda}^\sharp,\ f_\lambda^\sharp - f_{\mathbf{x},\lambda}\right\rangle_K\right] = \mathbf{E}\left[\left\|f_\lambda^\sharp - f_{\mathbf{x},\lambda}^\sharp\right\|_K^2\right] + \mathbf{E}\left[\left\langle f_\lambda^\sharp - f_{\mathbf{x},\lambda}^\sharp,\ f_{\mathbf{x},\lambda}^\sharp - f_{\mathbf{x},\lambda}\right\rangle_K\right]$$

$$\le 10(\kappa^{2r} + \kappa^2 + \kappa)\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2\left(\frac{\lambda^{\min(2r-2,0)}}{N} + \frac{\lambda^{\min(r-1,0)}}{N^{\min(\frac{r}{2}+\frac{1}{4},1)}} + \frac{\lambda^{\min(2r-\frac{3}{2},r-\frac{1}{2},1)}}{N^{\frac{1}{2}}}\right)$$

and therefore

$$\lambda\left(\mathbf{E}\left[\|f_\lambda^\sharp - f_{D,\lambda}\|_K^2\right] - \mathbf{E}\left[\|f_{D,\lambda}^\sharp - f_{D,\lambda}\|_K^2\right]\right)$$

$$\le 20(\kappa^{2r} + \kappa^2 + \kappa)\|u^*\|_{L_{\rho_\mathcal{X}}^2}^2\left(\frac{\lambda^{\min(2r-1,1)}}{N} + \frac{\lambda^{\min(r,1)}}{N^{\min(\frac{r}{2}+\frac{1}{4},1)}} + \frac{\lambda^{\min(2r-\frac{1}{2},r+\frac{1}{2},2)}}{N^{\frac{1}{2}}}\right) + \frac{4\kappa^2\sigma^2}{\lambda N}.$$

This together with Lemma 7 gives the desired conclusion. ∎

5. Error bounds for KRR and distributed KRR

The main result of this section is the following error bound for KRR on a single data set.

Theorem 13 *Assume $f^* = L_K^r u^*$ for some $0 < r \le 2$ and $u^* \in L_{\rho_\mathcal{X}}^2$.*

(i) *If $0 < r \leq \frac{1}{2}$, then*

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq 2\lambda^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{\kappa^4}{N^2\lambda^2} + \frac{2\kappa^2}{N\lambda}\right) + \left(\frac{\kappa^4}{N^2\lambda^2} + \frac{2\kappa^2}{N\lambda}\right)\sigma^2.$$

(ii) *If $r \geq \frac{1}{2}$ and λ is chosen so that $\lambda N \geq 1$, then there is a constant $c_1 > 0$ such that*

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq c_1 \left(\lambda^{\min(2r,2)} + \frac{\lambda^{\min(r,1)}}{N^{\min(\frac{r}{2}+\frac{1}{4},1)}} + \frac{\sigma^2}{\lambda N}\right).$$

The proof of Theorem 13 will be proved in Section 5.1 below. With this theorem, the following corollary on the capacity independent rate of KRR follows immediately.

Corollary 14 *Under the assumption of Theorem 13, by choosing $\lambda = N^{-\max\left(\frac{1}{1+2r}, \frac{1}{3}\right)}$ we have*

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] = O\left(N^{-\min\left(\frac{2r}{1+2r}, \frac{2}{3}\right)}\right).$$

If, in addition, $\sigma^2 = 0$, then we can choose $\lambda = N^{-1}$ to obtain

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] = \begin{cases} O\left(N^{-2r}\right), & \text{if } 0 < r \leq \frac{1}{2}; \\ O\left(N^{-\left(\frac{3r}{2}+\frac{1}{4}\right)}\right), & \text{if } \frac{1}{2} < r \leq 1; \\ O\left(N^{-\left(\frac{r}{2}+\frac{5}{4}\right)}\right), & \text{if } 1 < r \leq \frac{3}{2}; \\ O\left(N^{-2}\right), & \text{if } r \geq \frac{3}{2}. \end{cases}$$

While the capacity independent rate of $O(N^{-\frac{2r}{1+2r}})$ is well known in the literature and it is expected the noise free learning should intuitively be better, it is still surprising to see the rate for noise free learning in Corollary 14 can be faster than $O(N^{-1})$. Moreover, KRR was known to suffer from a saturation effect that says the rate ceases to increase for $r \geq 1$. We see from Corollary 14 that for noise free learning the rate can continue to increase beyond $r \geq 1$ and the saturation effect occurs when $r \geq \frac{3}{2}$. To our best knowledge, both the super fast rate and the relaxed saturation effect for noise free regression learning are novel observations.

We can now prove Theorem 1 by combining Theorem 5 and Theorem 13.

Proof [Proof of Theorem 1.] Theorem 13 tells that, if $0 < r \leq \frac{1}{2}$, the error bound for KRR with a single data set of sample size n is

$$\varepsilon(n, \lambda, \sigma^2) \leq 2\lambda^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{\kappa^4}{n^2\lambda^2} + \frac{2\kappa^2}{n\lambda}\right) + \left(\frac{\kappa^4}{n^2\lambda^2} + \frac{2\kappa^2}{n\lambda}\right)\sigma^2.$$

Consequently

$$\varepsilon(n, \lambda, 0) \leq 2\lambda^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{\kappa^4}{n^2\lambda^2} + \frac{2\kappa^2}{n\lambda}\right).$$

19

By Theorem 5 and the fact $N = mn$, we obtain

$$\mathbf{E}\left[\|\overline{f_{D,\lambda}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \frac{\varepsilon(n,\lambda,\sigma^2)}{m} + \varepsilon(n,\lambda,0)$$

$$\leq 4\lambda^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \left(1 + \frac{\kappa^4}{n^2\lambda^2} + \frac{2\kappa^2}{n\lambda}\right) + \left(\frac{\kappa^4}{mn^2\lambda^2} + \frac{2\kappa^2}{mn\lambda}\right)\sigma^2$$

$$\leq C_1\left\{\lambda^{2r}\left(1 + \frac{m^2}{N^2\lambda^2} + \frac{m}{N\lambda}\right) + \frac{\sigma^2 m}{N^2} + \frac{\sigma^2}{N\lambda}\right\}$$

with $C_1 = (1 + \kappa^4)\max(4\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2, 1)$. If $\lambda = N^{-\frac{1}{2r+1}}$ and $m \leq N^{\frac{2r}{1+2r}}$, then $m \leq N\lambda$ and therefore

$$\mathbf{E}\left[\|\overline{f_{D,\lambda}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq 3C_1\left\{\lambda^{2r} + \frac{\sigma^2}{N\lambda}\right\} = 3C_1(1 + \sigma^2)N^{-\frac{2r}{1+2r}}.$$

This proves the claim in (i).

To prove claim (ii) we note that if $\frac{1}{2} \leq r \leq 1$ and $\lambda n \geq 1$, the error bound for KRR is

$$\varepsilon(n,\lambda,\sigma^2) \leq c_1\left(\lambda^{2r} + \frac{\lambda^r}{n^{\frac{r}{2}+\frac{1}{4}}} + \frac{\sigma^2}{\lambda n}\right).$$

By Theorem 5 we have

$$\mathbf{E}\left[\|\overline{f_{D,\lambda}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq 2c_1\left(\lambda^{2r} + \frac{\lambda^r}{n^{\frac{r}{2}+\frac{1}{4}}} + \frac{\sigma^2}{\lambda nm}\right) = 2c_1\left(\lambda^{2r} + \frac{\lambda^r m^{\frac{r}{2}+\frac{1}{4}}}{N^{\frac{r}{2}+\frac{1}{4}}} + \frac{\sigma^2}{\lambda N}\right).$$

With the choice $\lambda = N^{-\frac{1}{2r+1}}$ and $m \leq N^{\frac{4r^2+1}{4r^2+4r+1}}$, we obtain

$$\mathbf{E}\left[\|\overline{f_{D,\lambda}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq 2c_1(2 + \sigma^2)N^{-\frac{2r}{2r+1}}.$$

This finishes the proof. ■

5.1 Leave one out analysis for KRR

In this subsection we bound the error of KRR on a single data set and prove Theorem 13. To this end, we perform a leave one out analysis of KRR. While the idea of leave one out analysis roots in the work of Zhang (2003), we need a more rigorous analysis to quantify the dependence on σ^2.

Let $\widetilde{D}$ be a sample of $N + 1$ observations $\widetilde{D} = \{z_i = (x_i, y_i) : 1 \leq i \leq N + 1\}$, and $\widetilde{D}_{(i)}$ denote the sample of N observations generated by removing the ith observation (x_i, y_i) from $\widetilde{D}$, that is, $\widetilde{D}_{(i)} = \{z_j = (x_j, y_j) : 1 \leq j \leq N + 1, j \neq i\}$. Let

$$\hat{g} = \arg\min_{f \in \mathcal{H}_K}\left\{\frac{1}{N}\sum_{j=1}^{N+1}(f(x_j) - y_j)^2 + \lambda\|f\|_K^2\right\}$$

and define the leave one out estimators associated to $\widetilde{D}_{(i)}$ by

$$\hat{f}_{(i)} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{\substack{j=1 \\ j \neq i}}^{N+1} (f(x_j) - y_j)^2 + \lambda \|f\|_K^2 \right\}.$$

It is easy to verify that $\hat{g} = f_{\widetilde{D}, \frac{N\lambda}{N+1}}$ and $f_{D,\lambda} = \hat{f}_{(N+1)}$. We also note that $\hat{f}_{(i)}$, $1 \leq i \leq N+1$ are identically distributed (but not independent).

The following bound for the leave one out estimator is referred to leave one out stability and has been proved in Zhang (2003). We still give the proof for the purpose of self-containing.

Lemma 15 *We have*

$$\|\hat{f}_{(i)} - \hat{g}\|_K \leq \frac{\kappa}{N\lambda} |\hat{g}(x_i) - y_i|.$$

Proof We denote $\triangle f_i = \hat{f}_{(i)} - \hat{g}$ and

$$\mathcal{E}_i(f) = \sum_{\substack{j=1 \\ j \neq i}}^{N+1} \left(f(x_j) - y_j \right)^2.$$

Recall that, if h is convex, then for any $a, b \in \mathbb{R}$ and $t \in [0, 1]$, there holds

$$h(a + t(b - a)) - h(a) \leq t \big(h(b) - h(a) \big).$$

By the convexity of the square loss, we have

$$\mathcal{E}_i(\hat{g} + t\triangle f_i) - \mathcal{E}_i(\hat{g}) \leq t \left(\mathcal{E}_i(\hat{f}_{(i)}) - \mathcal{E}_i(\hat{g}) \right) \tag{14}$$

and

$$\mathcal{E}_i(\hat{f}_{(i)} - t\triangle f_i) - \mathcal{E}_i(\hat{f}_{(i)}) \leq t \left(\mathcal{E}_i(\hat{g}) - \mathcal{E}_i(\hat{f}_{(i)}) \right). \tag{15}$$

By the definition of $\hat{f}_{(i)}$, we have

$$\frac{1}{N} \mathcal{E}_i(\hat{f}_{(i)}) + \lambda \left\| \hat{f}_{(i)} \right\|_K^2 \leq \frac{1}{N} \mathcal{E}_i(\hat{f}_{(i)} - t\triangle f_i) + \lambda \left\| \hat{f}_{(i)} - t\triangle f_i \right\|_K^2.$$

This together with (15) leads to

$$2t \left\langle \hat{f}_{(i)}, \triangle f_i \right\rangle_K - t^2 \|\triangle f_i\|_K^2 = \left\| \hat{f}_{(i)} \right\|_K^2 - \left\| \hat{f}_{(i)} - t\triangle f_i \right\|_K^2 \leq \frac{t}{N\lambda} \left(\mathcal{E}_i(\hat{g}) - \mathcal{E}_i(\hat{f}_{(i)}) \right). \tag{16}$$

Similarly, by the definition of $\hat{g}$, we have

$$\frac{1}{N} \mathcal{E}_i(\hat{g}) + \lambda \|\hat{g}\|_K^2 + \frac{1}{N} (\hat{g}(x_i) - y_i)^2 \leq \frac{1}{N} \mathcal{E}_i(\hat{g} + t\triangle f_i) + \lambda \|\hat{g} + t\triangle f_i\|_K^2$$

$$+ \frac{1}{N} (\hat{g}(x_i) + t\triangle f_i(x_i) - y_i)^2.$$

By (14), we obtain

$$
\begin{aligned}
-2t \left\langle \hat{g}, \ \triangle f_i \right\rangle_K &- t^2 \|\triangle f_i\|_K^2 = \left\|\hat{g}\right\|_K^2 - \left\|\hat{g} + t\triangle f_i\right\|_K^2 \\
&\leq \frac{t}{N\lambda} \Big(\mathcal{E}_i(\hat{f}_{(i)}) - \mathcal{E}_i(\hat{g}) \Big) + \frac{1}{N\lambda} \Big\{ (\hat{g}(x_i) + t\triangle f_i(x_i) - y_i)^2 - (\hat{g}(x_i) - y_i)^2 \Big\}.
\end{aligned}
\tag{17}
$$

Adding (16) and (17) together, we obtain

$$
\begin{aligned}
2t(1-t) \|\triangle f_i\|_K^2 &\leq \frac{1}{N\lambda} \Big\{ (\hat{g}(x_i) + t\triangle f_i(x_i) - y_i)^2 - (\hat{g}(x_i) - y_i)^2 \Big\} \\
&= \frac{t}{N\lambda} \triangle f_i(x_i) \Big(2\hat{g}(x_i) + t\triangle f_i(x_i) - 2y_i \Big) \\
&\leq \frac{\kappa t}{N\lambda} \|\triangle f_i\|_K \Big| 2\hat{g}(x_i) + t\triangle f_i(x_i) - 2y_i \Big|.
\end{aligned}
$$

Therefore,

$$
2(1-t) \|\triangle f_i\|_K \leq \frac{\kappa}{N\lambda} \Big| 2\hat{g}(x_i) + t\triangle f_i(x_i) - 2y_i \Big|.
$$

Letting $t \to 0$, we obtain

$$
\|\triangle f_i\|_K \leq \frac{\kappa}{N\lambda} \Big| \hat{g}(x_i) - y_i \Big|.
$$

This proves Lemma 15. ∎

We will need the following lemma.

Lemma 16 *We have*

$$
\mathbf{E}\left[\frac{1}{N+1} \sum_{i=1}^{N+1} (\hat{g}(x_i) - y_i)^2 \right] \leq Q\Big(\frac{N\lambda}{N+1}, N+1 \Big) + \sigma^2.
$$

Proof Let $\tilde{\lambda} = \frac{N\lambda}{N+1}$ and $f_{\tilde{\lambda}}$ be defined by (10) associated with the regularization parameter $\tilde{\lambda}$. By the definition of $\hat{g}$, we have

$$
\begin{aligned}
\mathbf{E}&\left[\frac{1}{N+1} \sum_{i=1}^{N+1} (\hat{g}(x_i) - y_i)^2 \right] \\
&= \frac{N}{N+1} \mathbf{E}\left[\frac{1}{N} \sum_{i=1}^{N+1} (\hat{g}(x_i) - y_i)^2 + \lambda \left[\|\hat{g}\|_K^2 \right] \right] - \frac{N\lambda}{N+1} \mathbf{E}\|\hat{g}\|_K^2 \\
&\leq \frac{N}{N+1} \mathbf{E}\left[\frac{1}{N} \sum_{i=1}^{N+1} (f_{\tilde{\lambda}}(x_i) - y_i)^2 + \lambda \|f_{\tilde{\lambda}}\|_K^2 \right] - \frac{N\lambda}{N+1} \mathbf{E}\left[\|\hat{g}\|_K^2 \right] \\
&= \|f_{\tilde{\lambda}} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 + \tilde{\lambda} \|f_{\tilde{\lambda}}\|_K^2 - \tilde{\lambda}\mathbf{E}\|\hat{g}\|_K^2 + \sigma^2 \\
&= Q\Big(\tilde{\lambda}, N+1 \Big) + \sigma^2,
\end{aligned}
$$

where we have used the fact that $f_{\tilde{\lambda}}$ is independent of (x_i, y_i) and the identity (3). ∎

Proposition 17 *There holds that*

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \left(1 + \frac{\kappa^4}{N^2\lambda^2} + \frac{2\kappa^2}{N\lambda}\right) Q\left(\frac{N\lambda}{N+1}, N+1\right) + \left(\frac{\kappa^4}{N^2\lambda^2} + \frac{2\kappa^2}{N\lambda}\right)\sigma^2.$$

Proof By the facts that $f_{D,\lambda} = \hat{f}_{(N+1)}$ and $\hat{f}_{(i)}$, $1 \leq i \leq N+1$ are identically distributed, we have

$$\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] = \mathbf{E}\left[\|\hat{f}_{(N+1)} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] = \frac{1}{N+1}\sum_{i=1}^{N+1}\mathbf{E}\left[\|\hat{f}_{(i)} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right]$$

By the fact that $f_{(i)}$ is independent of (x_i, y_i) and the identity (3) we obtain

$$\begin{aligned}
\mathbf{E}\left[\|f_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] &= \frac{1}{N+1}\sum_{i=1}^{N+1}\mathbf{E}\left[\left(\hat{f}_{(i)}(x_i) - f^*(x_i)\right)^2\right] \\
&= \frac{1}{N+1}\sum_{i=1}^{N+1}\left(\mathbf{E}\left[(\hat{f}_{(i)}(x_i) - y_i)^2\right] - \sigma^2\right) \\
&= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - y_i\right)^2\right] - \sigma^2 \\
&= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - y_i\right)^2 - \frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{g}(x_i) - y_i\right)^2\right] \\
&\quad + \left(\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{g}(x_i) - y_i\right)^2\right] - \sigma^2\right)
\end{aligned}$$

By Lemma 15 and the reproducing property we have

$$|\hat{f}_{(i)}(x_i) - \hat{g}(x_i)| \leq \kappa\|\hat{f}_{(i)} - \hat{g}\|_K \leq \frac{\kappa^2}{N\lambda}|\hat{g}(x_i) - y_i|.$$

Therefore,

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - y_i\right)^2 - \frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{g}(x_i) - y_i\right)^2\right] \\
&= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - \hat{g}(x_i)\right)\left(\hat{f}_{(i)}(x_i) + \hat{g}(x_i) - 2y_i\right)\right] \\
&= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - \hat{g}(x_i)\right)^2 + \frac{2}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}_{(i)}(x_i) - \hat{g}(x_i)\right)\left(\hat{g}(x_i) - y_i\right)\right]
\end{aligned}$$

$$\leq \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\frac{\kappa^4}{N^2\lambda^2}\left(\hat{g}(x_i)-y_i\right)^2 + \frac{2}{N+1}\sum_{i=1}^{N+1}\frac{\kappa^2}{N\lambda}\left(\hat{g}(x_i)-y_i\right)^2\right]$$

$$= \left(\frac{\kappa^4}{N^2\lambda^2}+\frac{2\kappa^2}{N\lambda}\right)\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{g}(x_i)-y_i\right)^2\right].$$

Then

$$\mathbf{E}\left[\|f_{D,\lambda}-f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \left(1+\frac{\kappa^4}{N^2\lambda^2}+\frac{2\kappa^2}{N\lambda}\right)\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{g}(x_i)-y_i\right)^2\right] - \sigma^2$$

and the desired conclusion follows by using Lemma 16. ∎

Note that Theorem 13 follows immediately from Proposition 17 and Lemma 11.

6. Error bound for BCKRR and distributed BCKRR

The main result of this section is the following error bounds for BCKRR on a single data set that will be used in combination with Theorem 5 to derive the error bounds for its distributed version.

Theorem 18 *Assume $\sigma^2 < \infty$ and $f^* = L_K^r u^*$ for some $0 < r \leq 2$ and $u^* \in L^2_{\rho_{\mathcal{X}}}$.*

(i) *If $0 < r \leq \frac{1}{2}$, then*

$$\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq 8\lambda^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\left(1+\frac{\kappa^4}{N^2\lambda^2}+\frac{\kappa^2}{N\lambda}\right) + 4\sigma^2\left(\frac{\kappa^4}{N^2\lambda^2}+\frac{\kappa^2}{N\lambda}\right).$$

(ii) *If $\frac{1}{2} < r \leq 1$ and $\lambda N \geq 1$, then there is a constant $\tilde{c}_1 > 0$ such that*

$$\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \tilde{c}_1\left(\lambda^{2r}+\frac{\lambda^r}{N^{\frac{r}{2}+\frac{1}{4}}}+\frac{\sigma^2}{\lambda N}\right).$$

(iii) *If $1 < r \leq \frac{3}{2}$, $\lambda \leq 1$ and $\lambda N \geq 1$, then there is a constant $\tilde{c}_2 > 0$ such that*

$$\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \tilde{c}_2\left(\lambda^{2r}+\frac{\lambda}{N^{\frac{r}{2}+\frac{1}{4}}}+\frac{\lambda^{r+\frac{1}{2}}}{\sqrt{N}}+\frac{\sigma^2}{\lambda N}\right).$$

(iv) *If $r \geq \frac{3}{2}$, $\lambda \leq 1$ and $\lambda N \geq 1$, then there is a constant $\tilde{c}_3 > 0$ such that*

$$\mathbf{E}\left[\|f_{D,\lambda}^\sharp-f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2\right] \leq \tilde{c}_3\left(\lambda^{\min(2r,4)}+\frac{\lambda}{N}+\frac{\sigma^2}{\lambda N}+\frac{\lambda^2}{\sqrt{N}}\right).$$

The proof of Theorem 18 will be proved in Section 6.2 below. The following corollary on the capacity independent rate of BCKRR follows immediately.

Corollary 19 *Under the assumption of Theorem 18, by choosing $\lambda = N^{-\frac{1}{1+2r}}$ we have*

$$\mathbf{E}\left[\|f^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] = O\left(N^{-\frac{2r}{1+2r}}\right).$$

If, in addition, $\sigma^2 = 0$, then we can choose $\lambda = N^{-1}$ to obtain

$$\mathbf{E}\left[\|f^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] = \begin{cases} O\left(N^{-2r}\right), & \text{if } 0 < r \le \frac{1}{2}; \\ O\left(N^{-(\frac{3r}{2}+\frac{1}{4})}\right), & \text{if } \frac{1}{2} < r \le 1; \\ O\left(N^{-(\frac{r}{2}+\frac{5}{4})}\right), & \text{if } 1 < r \le \frac{3}{2}; \\ O\left(N^{-2}\right), & \text{if } r \ge \frac{3}{2}. \end{cases}$$

BCKRR suffers from a saturation effect when $r \ge 2$, which relaxed the saturation effect of KRR (Guo et al., 2017b). For noise free learning, however, we see the result for BCKRR is the same as that for KRR in Corollary 14 and the saturation occurs when $r = \frac{3}{2}$. A plausible interpretation of this phenomenon is that the rate $O(N^{-2})$ might be the fastest rate a regression learning algorithm can achieve and thus bias correction cannot help further improve the rate.

Theorem 2 can now be proved by combining Theorem 5 and Theorem 18. The proof is similar to that of Theorem 1 in Section 5. We omit the details.

6.1 Two alternative perspectives for bias correction

Recall the BCKRR estimator is defined by a two step procedure. Although it admits an explicit expression and an operator representation, neither is suitable for leave one out analysis, if not impossible. In this section we show that BCKRR can be interpreted by two alternative perspectives, fitting the residual and recentering regularization. The latter represents BCKRR as a Tikhonov regularization scheme and plays an essential role in the leave one out analysis of BCKRR in the next subsection. These two perspectives on bias correction may also be of independent interest by themselves; see the discussions in Section 7.

Suppose KRR estimator $f_{D,\lambda}$ fit the data by trading off the fitting error and model complexity. If we further fit the residuals by a function $\hat{h}$ and add it to $f_{D,\lambda}$, the resulted function will have smaller fitting error. In other words, the bias is reduced. The following proposition tells that BCKRR defined in (7) is equivalent to this process.

Proposition 20 *Let $r_i = y_i - f_{D,\lambda}(x_i)$ be the residual and $\hat{h}$ is the KRR estimator obtained by fitting the residual, i.e.,*

$$\hat{h} = \arg\min_{g \in \mathcal{H}_K} \left\{\frac{1}{N}\sum_{i=1}^{N}(r_i - g(x))^2 + \lambda\|g\|^2_K\right\}.$$

Then we have $f^{\sharp}_{D,\lambda} = f_{D,\lambda} + \hat{h}$.

Proof Denote by $\mathbf{r} = (r_1, r_2, \ldots, r_n)^\top$ the column vector of residuals. By the operator representation of the KRR estimator we have

$$\hat{h} = \frac{1}{N}\left(\lambda I + \frac{1}{N}S^*S\right)^{-1}S^*\mathbf{r}. \tag{18}$$

It can be easily verified that

$$S^*\mathbf{r} = S^*(\mathbf{y} - Sf_{D,\lambda}) = S^*\mathbf{y} - \frac{1}{N}S^*S\left(\lambda I + \frac{1}{N}S^*S\right)^{-1}S^*\mathbf{y}$$

$$= \lambda\left(\lambda I + \frac{1}{N}S^*S\right)^{-1}S^*\mathbf{y} = \lambda N f_{D,\lambda}.$$

Plugging it into (18) we obtain

$$\hat{h} = \lambda\left(\lambda I + \frac{1}{N}S^*S\right)^{-1}f_{D,\lambda} = f_{D,\lambda}^\sharp - f_{D,\lambda}.$$

This proves the conclusion. ∎

Recall that KRR is somewhat equivalent to search a minimizer of the fitting error in a ball of radius $\frac{1}{\sqrt{\lambda}}$ centered at the zero function. The following proposition explains BCKRR as a re-search for a minimizer of the fitting error in a ball centered at $f_{D,\lambda}$. This is somewhat equivalent to increasing the searching region to reduce the fitting error and thus implement bias reduction.

Proposition 21 *We have*

$$f_{D,\lambda}^\sharp = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda\|f - f_{D,\lambda}\|_K^2\right\}. \tag{19}$$

Proof By the fact $y_i = f_{D,\lambda}(x_i) + r_i$ and Proposition 20, for all $f \in \mathcal{H}_K$, we have

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - f_{D,\lambda}^\sharp(x_i))^2 + \lambda\|f_{D,\lambda}^\sharp - f_{D,\lambda}\|_K^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(r_i - \hat{h}(x_i))^2 + \lambda\|\hat{h}\|_K^2$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}(r_i - [f(x_i) - f_{D,\lambda}(x_i)])^2 + \lambda\|f - f_{D,\lambda}\|_K^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda\|f - f_{D,\lambda}\|_K^2.$$

Therefore, $f_{D,\lambda}^{\sharp}$ is a minimizer of the recentered risk functional

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \|f - f_{D,\lambda}\|_K^2.$$

Since the recentered risk functional is convex in f, the minimizer is unique. This proves (19). ∎

Proposition 21 writes the BCKRR as a recentering Tikhonov regularization scheme. It enables the use of convex analysis to derive the leave one out error bound for BCKRR.

6.2 Leave one out analysis of BCKRR

We now perform a leave one analysis of BCKRR by the similar framework as that we have done in Section 5. The result will be used to prove Theorem 18. To this end, recall the definitions of $\widetilde{D}$, $\hat{f}_{(i)}$ and $\hat{g}$ in Section 5.1. We define

$$\hat{f}_{(i)}^{\sharp} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{\substack{j=1 \\ j \neq i}}^{N+1} (f(x_j) - y_j)^2 + \lambda \|f - \hat{f}_{(i)}\|_K^2 \right\} \tag{20}$$

$$\hat{g}^{\sharp} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{j=1}^{N+1} (f(x_j) - y_j)^2 + \lambda \|f - \hat{g}\|_K^2 \right\}. \tag{21}$$

Notice that that $\hat{f}_{(N+1)}^{\sharp} = f_{D,\lambda}^{\sharp}$ and $\hat{g}^{\sharp} = f_{\widetilde{D},\tilde{\lambda}}^{\sharp}$ where again $\tilde{\lambda} = \frac{N\lambda}{N+1}$.

Lemma 22 *For all $1 \leq i \leq n+1$, there is*

$$\|\hat{f}_{(i)}^{\sharp} - \hat{g}^{\sharp}\|_K \leq \frac{\kappa}{N\lambda} \left[|\hat{g}(x_i) - y_i| + |\hat{g}^{\sharp}(x_i) - y_i| \right].$$

Proof Recall

$$\mathcal{E}_i(f) = \sum_{\substack{j=1 \\ j \neq i}}^{N+1} (f(x_j) - y_j)^2.$$

Let $\triangle f_i = \hat{f}_{(i)}^{\sharp} - \hat{g}^{\sharp}$. By the convexity of the square function, for any $t \in (0,1)$, we have

$$\mathcal{E}_i(\hat{g}^{\sharp} + t\triangle f_i) - \mathcal{E}_i(\hat{g}^{\sharp}) \leq t \left[\mathcal{E}_i(\hat{f}_{(i)}^{\sharp}) - \mathcal{E}_i(\hat{g}^{\sharp}) \right], \tag{22}$$

and

$$\mathcal{E}_i(\hat{f}_{(i)}^{\sharp} - t\triangle f_i) - \mathcal{E}_i(\hat{f}_{(i)}^{\sharp}) \leq t \left[\mathcal{E}_i(\hat{g}^{\sharp}) - \mathcal{E}_i(\hat{f}_{(i)}^{\sharp}) \right]. \tag{23}$$

By the definition of $\hat{f}_{(i)}^{\sharp}$ in (20), for any $t \in (0,1)$

$$\frac{1}{N} \mathcal{E}_i(\hat{f}_{(i)}^{\sharp}) + \lambda \|\hat{f}_{(i)}^{\sharp} - \hat{f}_{(i)}\|_K^2 \leq \frac{1}{N} \mathcal{E}_i(\hat{f}_{(i)}^{\sharp} - t\triangle f_i) + \lambda \|\hat{f}_{(i)}^{\sharp} - t\triangle f_i - \hat{f}_{(i)}\|_K^2.$$

27

This yields

$$\lambda \left\{ 2t \left\langle \hat{f}_{(i)}^{\sharp} - \hat{f}_{(i)}, \triangle f_i \right\rangle_K - t^2 \|\triangle f_i\|_K^2 \right\} \le \frac{1}{N} \left\{ \mathcal{E}_i(\hat{f}_{(i)}^{\sharp} - t\triangle f_i) - \mathcal{E}_i(\hat{f}_{(i)}^{\sharp}) \right\}. \tag{24}$$

Similarly, by the definition of $\hat{g}^{\sharp}$ in (21), we have for any $t \in (0,1)$

$$\frac{1}{N}\mathcal{E}_i(\hat{g}^{\sharp}) + \frac{1}{N}(\hat{g}^{\sharp}(x_i) - y_i)^2 + \lambda \|\hat{g}^{\sharp} - \hat{g}\|_K^2$$
$$\le \frac{1}{N}\mathcal{E}_i(\hat{g}^{\sharp} + t\triangle f_i) + \frac{1}{N}\big((1-t)\hat{g}^{\sharp}(x_i) + t\hat{f}_{(i)}^{\sharp}(x_i) - y_i\big)^2 + \lambda \|\hat{g}^{\sharp} - \hat{g} + t\triangle f_i\|_K^2.$$

Thus,

$$\lambda \left\{ -2t \left\langle \hat{g}^{\sharp} - \hat{g}, \triangle f_i \right\rangle_K - t^2 \|\triangle f_i\|_K^2 \right\} \le \frac{1}{N} \left\{ \mathcal{E}_i(\hat{g}^{\sharp} + t\triangle f_i) - \mathcal{E}_i(\hat{g}^{\sharp}) \right\}$$
$$+ \frac{t}{N} \left(\hat{f}_{(i)}^{\sharp}(x_i) - \hat{g}^{\sharp}(x_i) \right) \left((2-t)\hat{g}^{\sharp}(x_i) + t\hat{f}_{(i)}^{\sharp}(x_i) - 2y_i \right). \tag{25}$$

Adding (24) and (25) together and applying (22), (23), we obtain

$$2(1-t)\|\triangle f_i\|_K^2 + 2 \left\langle \hat{g} - \hat{f}_{(i)}, \triangle f_i \right\rangle_K$$
$$\le \frac{1}{N\lambda} \left(\hat{f}_{(i)}^{\sharp}(x_i) - \hat{g}^{\sharp}(x_i) \right) \left((2-t)\hat{g}^{\sharp}(x_i) + t\hat{f}_{(i)}^{\sharp}(x_i) - 2y_i \right).$$

Letting $t \to 0$, we have

$$\|\triangle f_i\|_K^2 + \left\langle \hat{g} - \hat{f}_{(i)}, \triangle f_i \right\rangle_K \le \frac{1}{N\lambda} \left(\hat{f}_{(i)}^{\sharp}(x_i) - \hat{g}^{\sharp}(x_i) \right) \left(\hat{g}^{\sharp}(x_i) - y_i \right)$$
$$= \frac{1}{N\lambda} \triangle f_i(x_i) \left(\hat{g}^{\sharp}(x_i) - y_i \right).$$

Therefore,

$$\|\triangle f_i\|_K^2 \le \|\hat{g} - \hat{f}_{(i)}\|_K \|\triangle f_i\|_K + \frac{\kappa}{N\lambda} \|\triangle f_i\|_K |\hat{g}^{\sharp}(x_i) - y_i|,$$

which implies

$$\|\triangle f_i\|_K \le \|\hat{g} - \hat{f}_{(i)}\|_K + \frac{\kappa}{N\lambda} |\hat{g}^{\sharp}(x_i) - y_i|,$$

By Lemma 15, we obtain the desired bound. ∎

Proposition 23 *We have*

$$\mathbf{E}\left[\|f_{D,\lambda}^{\sharp} - f^*\|_{L^2_{\rho_{\mathcal{X}}}}^2 \right] \le \left(1 + \frac{4\kappa^2}{N\lambda} + \frac{4\kappa^4}{N^2\lambda^2} \right) Q\left(\frac{N\lambda}{N+1}, N+1 \right) + \left(\frac{4\kappa^2}{N\lambda} + \frac{4\kappa^4}{N^2\lambda^2} \right) \sigma^2.$$

Proof Note that $\hat{f}^{\sharp}_{(i)}$, $1 \leq i \leq N+1$, are identically distributed, and $\hat{f}^{\sharp}_{(i)}$ is independent of the observation (x_i, y_i). By the fact $f^{\sharp}_{D,\lambda} = \hat{f}^{\sharp}_{(N+1)}$, we have

$$
\begin{aligned}
\mathbf{E}\left[\|f^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] &= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left\|\hat{f}^{\sharp}_{(i)} - f^*\right\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] \\
&= \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{f}^{\sharp}_{(i)}(x_i) - y_i)^2\right] - \sigma^2 \\
&= \left\{\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{g}^{\sharp}(x_i) - y_i)^2\right] - \sigma^2\right\} \\
&\quad + \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}^{\sharp}_{(i)}(x_i) - \hat{g}^{\sharp}(x_i)\right)^2\right] \\
&\quad + 2\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\left(\hat{f}^{\sharp}_{(i)}(x_i) - \hat{g}^{\sharp}(x_i)\right)\left(\hat{g}^{\sharp}(x_i) - y_i\right)\right] \\
&:= J_1 + J_2 + J_3.
\end{aligned}
\tag{26}
$$

Recall Lemma 16 tells that

$$
\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{g}(x_i) - y_i)^2\right] \leq Q\left(\frac{N\lambda}{N+1}, \ N+1\right) + \sigma^2.
\tag{27}
$$

By the definition of $\hat{g}^{\sharp}$ we also obtain

$$
\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{g}^{\sharp}(x_i) - y_i)^2\right] \leq \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{g}(x_i) - y_i)^2\right] \leq Q\left(\frac{N\lambda}{N+1}, \ N+1\right) + \sigma^2.
\tag{28}
$$

Therefore, we can bound J_1 by

$$
J_1 \leq Q\left(\frac{N\lambda}{N+1}, \ N+1\right).
\tag{29}
$$

To estimate J_2, we apply Lemma 22 and obtain

$$
\begin{aligned}
\left(\hat{f}^{\sharp}_{(i)}(x_i) - \hat{g}^{\sharp}(x_i)\right)^2 &\leq \kappa^2\|\hat{f}^{\sharp}_{(i)} - \hat{g}^{\sharp}\|^2_K \leq \kappa^2\left(\frac{\kappa}{N\lambda}|\hat{g}(x_i) - y_i| + \frac{\kappa}{N\lambda}|\hat{g}^{\sharp}(x_i) - y_i|\right)^2 \\
&\leq 2\kappa^2\left(\frac{\kappa^2}{N^2\lambda^2}(\hat{g}(x_i) - y_i)^2 + \frac{\kappa^2}{N^2\lambda^2}(\hat{g}^{\sharp}(x_i) - y_i)^2\right).
\end{aligned}
$$

Then by (27) and (28) we obtain

$$
J_2 \leq \frac{4\kappa^4}{N^2\lambda^2}\left(Q\left(\frac{N\lambda}{N+1}, \ N+1\right) + \sigma^2\right).
\tag{30}
$$

For J_3, by Lemma 22 again, we have

$$2\left|\left(\hat{f}_{(i)}^\sharp(x_i) - \hat{g}^\sharp(x_i)\right)\left(\hat{g}^\sharp(x_i) - y_i\right)\right| \leq 2\kappa\|\hat{f}_{(i)}^\sharp - \hat{g}^\sharp\|_K|\hat{g}^\sharp(x_i) - y_i|$$

$$\leq 2\kappa\left(\frac{\kappa}{N\lambda}|\hat{g}(x_i) - y_i| + \frac{\kappa}{N\lambda}|\hat{g}^\sharp(x_i) - y_i|\right)|\hat{g}^\sharp(x_i) - y_i|$$

$$\leq \frac{\kappa^2}{N\lambda}(\hat{g}(x_i) - y_i)^2 + \frac{3\kappa^2}{N\lambda}(\hat{g}^\sharp(x_i) - y_i)^2$$

By (27) and (28) again, we have

$$J_3 \leq \frac{4\kappa^2}{N\lambda}\left(Q\left(\frac{N\lambda}{N+1},\ N+1\right) + \sigma^2\right). \tag{31}$$

The desired error bound follows by combining the estimation for J_1, J_2 and J_3. ■

Theorem 18 part (i) and part (ii) now follow immediately from Proposition 23 and Lemma 11.

If $r \geq 1$, the bound in Proposition 23 is still true but the estimation for $Q(\lambda, N)$ is not sufficient to prove sharp bounds for BCKRR. Instead, we will estimate J_1 in (26) alternatively to obtain the following error bound for BCKRR, which together with by Lemma 12 allows to derive the sharp bounds in Theorem 18 part (iii) and part (iv).

Proposition 24 *If $r \geq 1$, we have*

$$\mathbf{E}\left[\|f_{D,\lambda}^\sharp - f^*\|_{L_{\rho_\mathcal{X}}^2}^2\right] \leq Q^\sharp\left(\frac{N\lambda}{N+1},\ N+1\right) + \left(\frac{4\kappa^2}{N\lambda} + \frac{4\kappa^4}{N^2\lambda^2}\right)\left\{Q\left(\frac{N\lambda}{N+1},\ N+1\right) + \sigma^2\right\}.$$

Proof Let $\tilde{\lambda} = \frac{N\lambda}{N+1}$ and $f_{\tilde{\lambda}}^\sharp$ be defined by (11) associated to the parameter $\tilde{\lambda}$. Since $\hat{g}^\sharp = f_{\widetilde{D},\tilde{\lambda}}^\sharp$, we can regard $f_{\tilde{\lambda}}^\sharp$ as the sample limit of $\hat{g}$. Recall the error decomposition in (26). By (19) and a similar process to the proof of Lemma 16, we have

$$\mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(\hat{g}^\sharp(x_i) - y_i)^2\right]$$

$$\leq \frac{N}{N+1}\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N+1}(\hat{g}^\sharp(x_i) - y_i)^2 + \lambda\|\hat{g}^\sharp - \hat{g}\|_K^2\right] - \frac{N\lambda}{N+1}\mathbf{E}\left[\|\hat{g}^\sharp - \hat{g}\|_K^2\right]$$

$$\leq \mathbf{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}(f_{\tilde{\lambda}}^\sharp(x_i) - y_i)^2\right] + \tilde{\lambda}\mathbf{E}\left[\|f_{\tilde{\lambda}}^\sharp - \hat{g}\|_K^2 - \|\hat{g}^\sharp - \hat{g}\|_K^2\right]$$

$$\leq \|f_{\tilde{\lambda}}^\sharp - f^*\|_{L_{\rho_\mathcal{X}}^2}^2 + \sigma^2 + \tilde{\lambda}\mathbf{E}\left[\|f_{\tilde{\lambda}}^\sharp - \hat{g}\|_K^2 - \|\hat{g}^\sharp - \hat{g}\|_K^2\right].$$

Hence we can bound J_1 as

$$J_1 \leq \|f_{\tilde{\lambda}}^\sharp - f^*\|_{L_{\rho_\mathcal{X}}^2}^2 + \tilde{\lambda}\mathbf{E}\left[\|f_{\tilde{\lambda}}^\sharp - \hat{g}\|_K^2 - \|\hat{g}^\sharp - \hat{g}\|_K^2\right] = Q^\sharp\left(\frac{N\lambda}{N+1}, N+1\right).$$

Combining this with the estimations in (30) for J_2 and (31) for J_3 we obtain the desired bound. ■

7. Conclusions and discussions

In this paper, we first proposed a general framework to analyze the performance of response weighted distributed regression algorithms. Then we conducted leave one analyses of KRR and BCKRR, which lead to sharp error bounds and capacity independent optimal rates for both approaches. The error bounds factored in the impact of unexplained variance of the response variable and hence are able to be used in combination with aforementioned framework to deduce sharp error bounds and optimal learning rates for distributed KRR and distributed BCKRR even when the kernel is imperfect in the sense that the true regression function does not lie in the associated reproducing kernel Hilbert space.

Our analysis involves two interesting byproducts. The first one is the super fast rates for noise free learning. To our best knowledge, rates that are faster than $O(N^{-1})$ have been observed for classification learning in some special situations (Smale and Zhou, 2007) but have never been observed for regression learning in the literature. In this paper we first show that they are also possible for regression learning.

The second one is the two alternative perspectives to reformulate BCKRR. They are not only critical for us to analyze the performance of BCKRR in this study, but also shed light on the design of bias corrected algorithms for other machine learning tasks. Recall that the original design of BCKRR in Wu (2017) heavily depends on the explicit operator representation of KRR. In machine learning, most regression or classification algorithms are solved by an iterative optimization process and no explicit analytic solution exists. It is thus impractical to characterize or estimate the bias. However, the idea of fitting residuals may apply to all regression problems while recentering regularization can apply to all regularization schemes. Regression learning usually adopts a loss function $L(|y - f(x)|)$ and minimizes the regularized empirical loss:

$$\hat{f}_L = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(|y_i - f(x_i)|) + \lambda \|f\|_K^2 \right\}.$$

Define the residuals $r_i = y_i - \hat{f}_L(x_i)$ and fit the residual by a function $\hat{g}_L$ such that

$$\hat{g}_L = \arg \min_{g \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(|r_i - g(x_i)|) + \lambda \|g\|_K^2 \right\}.$$

The bias corrected estimator with respect to the loss L can then be defined as $\hat{f}_L^\sharp = \hat{f}_L + \hat{g}_L$. In binary classification one usually uses a loss function of form $L(yf(x))$ where $y \in \{1, -1\}$, for instance, the hinge loss in support vector machines or the logistic loss in logistic regression. Because y_i are labels and residuals are meaningless, it is not appropriate to implement bias correction by fitting residuals in binary classification. Instead, we can still use recentering regularization. Namely, suppose the regularized binary classification algorithm estimates a classifier $\hat{f}_L$ by

$$\hat{f}_L = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(y_i f(x_i)) + \lambda \|f\|_K^2 \right\}.$$

We define the bias corrected classifier by

$$\hat{f}_L^{\sharp} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} L(y_i f(x_i)) + \lambda \|f - \hat{f}_L\|_K^2 \right\}.$$

It would be interesting to study these bias corrected algorithms in the future and investigate their theoretical properties as well as empirical application domains.

Acknowledgment

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, 2005.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation*, 2004.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 2017a.

Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *Journal of Machine Learning Research*, 18 (118):1–25, 2017b.

Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Applied and Computational Harmonic Analysis*, 49 (1):229–256, 2020.

Tim Kraska, Ameet Talwalkar, John C Duchi, Rean Griffith, Michael J Franklin, and Michael I Jordan. MLbase: A distributed machine-learning system. In *6th Biennial Conference on Innovative Data Systems Research (CIDR'13)*, 2013.

Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel gradient descent algorithms. *Constructive Approximation*, 47:249–276, 2018.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

Lei Shi. Distributed learning with indefinite kernels. *Analysis and Applications*, 17(06): 947–975, 2019.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT*, 2009.

Hongwei Sun and Qiang Wu. Application of integral operator for regularized least-square regression. *Mathematical and Computer Modelling*, 49(1):276–285, 2009a.

Hongwei Sun and Qiang Wu. A note on application of integral operator in learning theory. *Applied and Computational Harmonic Analysis*, 26(3):416–421, 2009b.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1): 5272–5311, 2016.

Qiang Wu. Bias corrected regularization kernel network and its applications. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1072–1079, 2017.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Ganggang Xu, Zuofeng Shang, and Guang Cheng. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Tong Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 15(6):1397–1437, 2003.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.