Journal of Machine Learning Research 18 (2017) 1-25 Submitted 7/17; Revised 10/17; Published 10/17

Learning Theory of Distributed Regression with Bias Corrected Regularization Kernel Network

Zheng-Chu Guo GUOZHENGCHU@ZJU.EDU.CN
School of Mathematical Sciences, Zhejiang University,
Hangzhou 310027, P. R. China
Lei Shi LEISHI@FUDAN.EDU.CN
Shanghai Key Laboratory for Contemporary Applied Mathematics,
School of Mathematical Sciences, Fudan University,
Shanghai 200433, 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 learning is an effective way to analyze big data. In distributed regression, a typical approach is to divide the big data into multiple blocks, apply a base regression algorithm on each of them, and then simply average the output functions learnt from these blocks. Since the average process will decrease the variance, not the bias, bias correction is expected to improve the learning performance if the base regression algorithm is a biased one. Regularization kernel network is an effective and widely used method for nonlinear regression analysis. In this paper we will investigate a bias corrected version of regularization kernel network. We derive the error bounds when it is applied to a single data set and when it is applied as a base algorithm in distributed regression. We show that, under certain appropriate conditions, the optimal learning rates can be reached in both situations.

Keywords: Distributed learning, kernel method, regularization, bias correction, error bound

1. Introduction

Data acquisition become much faster and easier as the development of technology. In this big data era, distributed learning has received considerable attention and is shown to be an effective way to analyze data that is so big and cannot be handled by a single machine. Among various distributed learning paradigms, a simple one is to divide the whole data set into multiple blocks, apply a base learning algorithm to each block, and then average the results from different blocks (Rosenblatt and Nadler, 2016; Zhang et al., 2015). This process, though simple, has some advantages. First, it is computationally efficient because the second stage can be easily parallelized. Second, because no mutual communication is required, the data security or confidentiality can be well protected. Last, recent research shows this method is consistent and sometimes reaches optimal learning rate (Zhang et al., 2015; Lin et al., 2017). Thus its asymptotic effectiveness is theoretically guaranteed.

In distributed learning the performance highly depends on the selection of the base algorithm in the second stage. Assume a big data set D of N observations is randomly divided into m blocks, $D_1, D_2, \dots, D_m$, which are assumed to be of the same size at the moment so that D_i are independent and identically distributed if the entire sample set D is independently drawn from some unknown distribution ρ. Let $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m$ be the estimators obtained by applying a base algorithm on these data blocks. Assume each estimator $\hat{f}_i$ has bias b and variance v . Then the mean squared error of $\hat{f}_i$ is $\mathrm{mse}(\hat{f}_i) = b^2 + v$ while the average estimator

$$\bar{f} = \frac{1}{m}\sum_{i=1}^{m} \hat{f}_i$$

has $\mathrm{mse}(\bar{f}) = b^2 + \frac{v}{m}$. On a single data block the algorithm usually trades off the bias and variance well to achieve the optimal performance. In distributed learning, however, the variance shrinks fast when m is large but the bias keeps unchanging during the average process. In this case, the bias may dominate the learning performance. An algorithm (or a model selection strategy) that is optimal for a single block is not necessarily still optimal for distributed learning. Instead, distributed learning prefers algorithms of small bias as the base learning algorithm on each block. Therefore, when a base learning algorithm is biased, bias correction is expected to play a role to improve the performance. The purpose of this paper is to investigate the application of biased corrected regularization kernel network for distributed regression analysis.

In regression analysis, the data $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_{|D|}, y_{|D|})\}$ is a set of observations collected for input variable X of predictors and a scalar response variable Y, where $|D|$ is the sample size of the data set D. Assume they are linked by

$$y_i = f^*(x_i) + \epsilon_i, \qquad i = 1, 2, \dots, |D|,$$

where x_i comes from a compact metric space (e.g., a bounded subset in $\mathbb{R}^p$), $y_i \in \mathbb{R}$, and ϵ_i is a zero-mean noise. The target is to recover the unknown true model f^* as accurate as possible to understand the impact of predictors and predict the response on unobserved data. Numerous regression methods have been developed in the literature, e.g. ridge regression, LASSO, and regularization kernel network (RKN). Among them, the regularization kernel network is a popular kernel method for nonlinear regression analysis. Its predictive consistency has been extensively studied in a vast 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 (2009); Steinwart et al. (2009); Guo et al. (2017b) and the references therein. Its applications were also extensively explored and shown successful in many problem domains. More recently, a bias corrected version for RKN, or BCRKN for short, was proposed in Wu (2017) to improve the performance of block wise data processing. In Wu (2017), the asymptotic bias and variance of BCRKN on a single data set was characterized, which indicates BCRKN has smaller bias than RKN and thus implies its efficiency in learning with block wise data intuitively. Empirical study also confirmed this. However, without rigorous analysis of the error bounds, there is lack of theoretical guarantee. In this paper, we will derive the error bounds and learning rates of BCKRN both for a single data set and for distributed regression. This will provide a theoretical guarantee for the use of BCRKN from a learning theory perspective.

The rest of this paper will be arranged as follows. In Section 2 we will describe the BCRKN algorithm and state our main results. In particular, we show that BCRKN can achieve the minimax optimal rates in both single data learning and distributed learning. Moreover, BCRKN relaxes the saturation effect of RKN. In Section 3 we discuss the relations of our results with existing work and conduct some comparisons. The proofs of our results are given in Sections 4-8.

2. Main results

Let $\mathcal{X}$ denote the input space which is assumed to be a compact metric space. A Mercer kernel on $\mathcal{X}$ is a continuous, symmetric, and positive-semidefinite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The function class spanned by $\{K_x = K(x, \cdot) : x \in \mathcal{X}\}$ and equipped with the inner product satisfying $\langle K_x, K_t \rangle_K = K(x,t)$ forms a pre-Hilbert space. Its completion is called a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated to the kernel K, with the name coming after the reproducing property $f(x) = \langle f, K(x, \cdot)\rangle_K$, $\forall f \in \mathcal{H}_K$. Note that $|f(x)| \leq \sqrt{K(x,x)}\|f\|_K$ for all $f \in \mathcal{H}_K$. Consequently, with $\kappa = \sup_{x\in\mathcal{X}} \sqrt{K(x,x)} < \infty$, $\mathcal{H}_K$ can be embedded into $C(\mathcal{X})$ and $\|f\|_\infty \leq \kappa \|f\|_K$. More other properties of RKHS that will not be used in this paper can be found in Aronszajn (1950).

Given the data D and the RKHS $\mathcal{H}_K$, RKN estimates the true model by

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

where $\lambda > 0$ is a regularization parameter that trades off the fitting error and model complexity. The well known representer theorem (Wahba, 1990) tells that

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

with the coefficients $\mathbf{c} = (c_1, \ldots, c_{|D|})^\top$ satisfying $(\lambda|D|I + \mathbf{K})\mathbf{c} = \mathbf{y}_D$ where I represents an identity matrix (or operator), $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^{|D|}$ is the kernel matrix on the input data $\mathbf{x}_D = \{x_1, \ldots, x_{|D|}\}$ and $\mathbf{y}_D = (y_1, \cdots, y_{|D|})^\top$ is the vector of the response data. Let $S_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ be the sampling operator defined by

$$S_D f = (f(x_1), \ldots, f(x_{|D|}))^\top, \qquad \forall\, f \in \mathcal{H}_K.$$

Its dual operator S_D^* is given by

$$S_D^* \mathbf{c} = \sum_{i=1}^{|D|} c_i K_{x_i} \in \mathcal{H}_K, \qquad \forall\, \mathbf{c} \in \mathbb{R}^{|D|}.$$

Then $f_{D,\lambda}$ has the following operator representation (Smale and Zhou, 2007)

$$f_{D,\lambda} = \tfrac{1}{|D|} \left(\lambda I + \tfrac{1}{|D} S_D^* S_D\right)^{-1} S_D^* \mathbf{y}_D. \tag{2}$$

Note that the operator $\frac{1}{|D|}S_D^*S_D$ is a sample version of the integral operator

$$L_K f(x) = \mathbf{E}_t\left[K(x,t)f(t)\right] = \int_{\mathcal{X}} K(x,t)f(t)\mathrm{d}\rho_{\mathcal{X}}(t)$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ on $\mathcal{X}$. Recall that L_K defines a compact, symmetric, and positive operator on $\mathcal{H}_K$. In the sequel we also use the notation $L_{K,D} = \frac{1}{|D|}S_D^*S_D$ and write

$$f_{D,\lambda} = (\lambda I + L_{K,D})^{-1}\left(\tfrac{1}{|D|}S_D^*\mathbf{y}_D\right).$$

By the aid of operator representation (2), the asymptotic bias of RKN can be characterized as $-\lambda(\lambda I + L_K)^{-1}f^*$. The bias corrected regularization kernel network (BCRKN) is defined by subtracting a plug-in estimator of the bias (Wu, 2017)

$$f_{D,\lambda}^{\sharp} = f_{D,\lambda} + \lambda\left(\lambda I + L_{K,D}\right)^{-1} f_{D,\lambda}. \tag{3}$$

It is also verified in Wu (2017) that

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

with $\mathbf{c}^{\sharp} = \mathbf{c} + \lambda\left(\lambda I + \frac{1}{n}\mathbf{K}\right)^{-1}\mathbf{c}$. The effectiveness of BCRKN has been tested empirically by a variety of simulations and real applications in Wu (2017). The main purpose of this paper is to verify its effectiveness in distributed regression from a learning theory perspective.

To perform rigorous error analysis and present our main results, we need some notations and assumptions that are used throughout the paper. Firstly, we assume $|y| \leq M$ almost surely for some constant $M > 0$. This implies the true regression function f^* satisfies $\|f^*\|_\infty \leq M$.

Note that we can extend the domain of L_K to $L^2_{\rho_{\mathcal{X}}}$ and obtain a compact, symmetric, and positive operator on $L^2_{\rho_{\mathcal{X}}}$, which will be denoted by L. We can in turn say L_K is the restriction of L on $\mathcal{H}_K$. So $Lf = L_K f$ for $f \in \mathcal{H}_K$ and we do not need to differentiate them when operating on functions in $\mathcal{H}_K$. Our second assumption is a regularity condition on the true model:

$$f^* = L^r(u^*) \quad \text{for some } r > 0 \text{ and } u^* \in L^2_{\rho_{\mathcal{X}}}. \tag{4}$$

This assumption has been widely used in the literature of learning theory to characterize the approximation ability of $\mathcal{H}_K$; see e.g. De Vito et al. (2005); Smale and Zhou (2007); Bauer et al. (2007); Zhang et al. (2015) and many references therein. Recall that $L^{\frac{1}{2}}$ is an isomorphism from $\overline{\mathcal{H}_K}$ onto $\mathcal{H}_K$, i.e.

$$\|f\|_{L^2_{\rho_{\mathcal{X}}}} = \|L^{\frac{1}{2}}f\|_K, \qquad \text{for } f \in \overline{\mathcal{H}_K}, \tag{5}$$

where $\overline{\mathcal{H}_K}$ is the closure of $\mathcal{H}_K$ in $L^2_{\rho_{\mathcal{X}}}$. So if $r \geq \frac{1}{2}$, the condition (4) implies $f^* \in \mathcal{H}_K$.

We shall use the effective dimension $\mathcal{N}(\lambda) = \mathrm{Tr}((L_K + \lambda I)^{-1}L_K)$, that is, the trace of $(L_K + \lambda I)^{-1}L_K$, to measure the complexity of $\mathcal{H}_K$ with respect to $\rho_{\mathcal{X}}$. We assume that there exist a constant $C_0 > 0$ and some $0 < \beta \leq 1$ such that for all $\lambda > 0$

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-\beta}. \tag{6}$$

Again this is a natural and widely used assumption in the literature; see e.g. De Vito et al. (2005); Caponnetto and De Vito (2007); Zhang et al. (2013); Lin et al. (2017).

Assume $\kappa \geq 1$ without loss of generality for otherwise all our statements and proofs are still valid by defining $\kappa = \max\{1, \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}\}$. Denote

$$\mathcal{B}_{|D|,\lambda} = \frac{2\kappa}{\sqrt{|D|}}\left\{\frac{\kappa}{\sqrt{|D|\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}.$$

The consistency of RKN as well as BCRKN generally requires the regularization parameter λ to be chosen according to the sample size and satisfy $\lambda \to 0$ and $\lambda|D| \to \infty$ as $|D| \to \infty$. This implies λ is upper bounded by an absolute constant. So, in the sequel, we will assume $\lambda \leq 1$ without loss of generality to simplify our notations and presentations.

As the performance of distributed learning highly depends on the base algorithm, we will conduct a thorough error analysis of BCRKN for a single data set first and then turn to the distributed regression.

2.1 Error bound for learning with a single data set

We derive the following error bounds and learning rates for BCRKN when it is applied on a single data set.

Theorem 1 *If the regularity condition* (4) *holds with* $0 < r \leq 2$ *and* $0 < \lambda \leq 1$*, then for any* $0 < \delta < 1$*, with confidence at least* $1-\delta$*,*

$$\|f^\sharp_{D,\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq C\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)^3 \left(\mathcal{B}_{|D|,\lambda} + \lambda^r\right)\left(\log\frac{4}{\delta}\right)^4, \tag{7}$$

where C *is a constant independent of* $|D|$ *or* δ*. Consequently, we have*

$$\mathbf{E}\left[\|f^\sharp_{D,\lambda} - f^*\|^2_{L^2_{\rho_\mathcal{X}}}\right] \leq 4\Gamma(9)C^2\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)^6 \left(\mathcal{B}_{|D|,\lambda} + \lambda^r\right)^2. \tag{8}$$

Corollary 2 *Assume the regularity condition* (4) *holds with* $0 < r \leq 2$ *and* (6) *holds with* $0 < \beta \leq 1$.

(i) If $0 < r < \frac{1}{2}$*, choose* $\lambda = |D|^{-\frac{1}{1+\beta}}$*. Then for any* $0 < \delta < 1$*, with confidence at least* $1-\delta$*, we have*

$$\|f^\sharp_{D,\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq C_1|D|^{-\frac{r}{1+\beta}}\left(\log\frac{4}{\delta}\right)^4,$$

where C_1 *is a constant independent of* $|D|$ *or* δ*. Consequently,*

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

(ii) If $\frac{1}{2} \leq r \leq 2$*, choose* $\lambda = |D|^{-\frac{1}{2r+\beta}}$*. Then for any* $0 < \delta < 1$*, with confidence at least* $1-\delta$*, we have*

$$\|f^\sharp_{D,\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq C_2|D|^{-\frac{r}{2r+\beta}}\left(\log\frac{4}{\delta}\right)^4,$$

where C_2 is a constant independent of $|D|$ or δ. Consequently,

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

Recall that the minimax optimal learning rate under the assumptions (4) and (6) is $\mathcal{O}\left(|D|^{-\frac{2r}{2r+\beta}}\right)$. Theorem 1 tells that, when $r\geq\frac{1}{2}$, BCRKN achieves the minimax optimal learning rate on a single data set.

Since $f^*\in\mathcal{H}_K$ when $r\geq\frac{1}{2}$, we can also measure the convergence of $f^{\sharp}_{D,\lambda}$ to f^* in $\mathcal{H}_K$. As pointed out in Smale and Zhou (2007), the convergence in $\mathcal{H}_K$ implies the convergence in $C^s(\mathcal{X})$ if $K\in C^{2s}(\mathcal{X}\times\mathcal{X})$, here $C^s(\mathcal{X})$ is the space of all functions on $\mathcal{X}\subset\mathbb{R}^p$ whose partial derivatives up to order s are continuous with $\|f\|_{C^s(\mathcal{X})}=\sum_{|\alpha|\leq s}\|D^\alpha f\|_\infty$. So the convergence in $\mathcal{H}_K$ is much stronger. It is not only for the target function itself, but also for its derivatives.

Theorem 3 *If the regularity condition* (4) *holds with $\frac{1}{2}<r\leq 2$, then for any $0<\delta<1$ with confidence at least $1-\delta$,*

$$\|f^{\sharp}_{D,\lambda}-f^*\|_K\leq C_K\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}}+1\right)^2(\lambda^{-\frac{1}{2}}\mathcal{B}_{|D|,\lambda}+\lambda^{r-\frac{1}{2}})\left(\log\frac{4}{\delta}\right)^3, \tag{9}$$

where C_K is a constant independent of $|D|$ or δ. If (6) *holds with $0<\beta\leq 1$ and $\lambda=|D|^{-\frac{1}{2r+\beta}}$, then for any $0<\delta<1$, with confidence at least $1-\delta$, we have*

$$\|f^{\sharp}_{D,\lambda}-f^*\|_K\leq\tilde{C}_K|D|^{-\frac{r-\frac{1}{2}}{2r+\beta}}\left(\log\frac{4}{\delta}\right)^3, \tag{10}$$

where $\tilde{C}_K$ is a constant independent of $|D|$ or δ. Moreover,

$$\mathbf{E}\left[\|f^{\sharp}_{D,\lambda}-f^*\|^2_K\right]=\mathcal{O}\left(|D|^{-\frac{2r-1}{2r+\beta}}\right). \tag{11}$$

Under the assumptions (4) with $r>\frac{1}{2}$ and (6) with $0<\beta\leq 1$, the minimax optimality of the bound $\mathcal{O}\left(|D|^{-\frac{2r-1}{2r+\beta}}\right)$ in the $\mathcal{H}_K$-metric has been proved in Guo et al. (2016). Theorem 3 indicates that the stronger convergence of BCRKN is also rate optimal in the minimax sense.

When $0<r<\frac{1}{2}$, we are unfortunately not able to obtain the minimax rate by the integral operator technique under the assumption (6). Note that if $\mathcal{H}_K$ is finite dimensional the range of L^r is exactly $\mathcal{H}_K$ for all $r>0$. The assumption (4) always implies $f^*\in\mathcal{H}_K$. So the situation $0<r<\frac{1}{2}$ makes sense only when $\mathcal{H}_K$ is infinite dimensional. In this case, L_K has infinite positive eigenvalues which converge to 0. This imposes the main difficulty of error analysis via integral technique – although $L_{K,D}$ converges well to L_K at a rate $\mathcal{O}(|D|^{-1/2})$, the difference of $(\lambda I+L_{K,D})^{-1}$ and $(\lambda I+L_K)^{-1}$ cannot be well bounded when $\lambda\to 0$. Actually, even for RKN which has been exhaustedly studied in the literature, it is an open problem to obtain the minimax rate under the assumptions (4) and (6). However, if there is sufficient amount of unlabeled data which helps to improve the estimate of the integral operator, minimax rate can be achieved, as has been verified in

Blanchard and Krämer (2010). For this purpose we propose the following semi-supervised approach. Assume, in addition to the labeled data D, we have a sequence of unlabelled data $x_{|D|+1}, \ldots, x_{|D'|}$. We create a fully labeled data set

$$D' = \{(x_1, y_1'), \cdots, (x_{|D|}, y_{|D|}'), (x_{|D|+1}, 0), \cdots, (x_{|D'|}, 0)\},$$

where $y_i' = \frac{|D'|}{|D|} y_i$ for $1 \leq i \leq |D|$. We can apply RKN and BCRKN on D' to obtain semi-supervised estimators $f_{D',\lambda}$ and $f^\sharp_{D',\lambda}$. Note that $D = D'$ when $|D'| = |D|$. So the semi-supervised method can be regarded as an extension of the supervised method while the supervised method is a special case of the semi-supervised method with no unlabeled data. The next theorem confirms that BCRKN can achieve the minimax rate for $0 < r < \frac{1}{2}$ when there are enough unlabeled data.

Theorem 4 *Assume the regularity condition* (4) *with* $0 < r < \frac{1}{2}$. *For any* $0 < \delta < 1$, *we have with confidence at least* $1 - \delta$,

$$\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq \left(\frac{2M}{\kappa} + 4\|u^*\|_{L^2_{\rho_\mathcal{X}}}\right) \left(\frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}} + 1\right)^3 \left(\mathcal{B}_{|D|,\lambda} + \lambda^r\right) \left(\log\frac{4}{\delta}\right)^3. \tag{12}$$

If in addition (6) *holds with* $0 < \beta \leq 1$ *and* $r + \beta \geq \frac{1}{2}$, $\lambda = |D|^{-\frac{1}{2r+\beta}}$, $|D'| \geq |D|^{\frac{1+\beta}{2r+\beta}}$, *then for any* $\delta \in (0,1)$, *with confidence at least* $1 - \delta$, *there holds*

$$\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq C'|D|^{-\frac{r}{2r+\beta}} \left(\log\frac{4}{\delta}\right)^3 \tag{13}$$

where the constant C' *is independent of* δ, $|D|$ *or* $|D'|$ *and will be given explicitly in the proof.*

2.2 Error bound of distributed regression with BCRKN

When BCRKN is used as a base algorithm for distributed regression, a big data set D is split into m blocks $D_1, D_2, \ldots, D_m$. On each block D_j, BCRKN is applied to produce an estimator $f^\sharp_{D_j,\lambda}$, and the weighted average of $f^\sharp_{D_j,\lambda}$,

$$\overline{f}^\sharp_{D,\lambda} = \sum_{j=1}^m \frac{|D_j|}{|D|} f^\sharp_{D_j,\lambda}, \tag{14}$$

is used for the purposes of prediction and inference. For this divide-and-conquer approach, we first give a general error bound for an arbitrary m. Here we do not require each block to have the same sample size.

Theorem 5 *If the regularity condition* (4) *holds with* $\frac{1}{2} \leq r \leq 2$ *and* $\lambda \leq 1$, *then there exists a constant* $\bar{C}$ *independent of* m *or* $|D_j|$ *such that*

$$\mathbf{E}\left[\|\overline{f}^\sharp_{D,\lambda} - f^*\|^2_{L^2_{\rho_\mathcal{X}}}\right] \leq \bar{C} \sum_{j=1}^m \frac{|D_j|}{|D|} \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^6 \left(\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} + \frac{\lambda^2 \mathcal{N}(\lambda)}{|D|} + \lambda^{2r}\right).$$

We next show that the distributed BCRKN (14) can achieve the optimal learning rate provided that m is not too large.

Theorem 6 *Assume the regularity condition* (4) *with* $\frac{1}{2} \leq r \leq 2$. *If* (6) *holds with* $0 < \beta \leq 1$, $|D_1| = |D_2| = \cdots = |D_m|$, $\lambda = |D|^{-\frac{1}{2r+\beta}}$, *and the number of the local machines satisfies*

$$m \leq |D|^{\min\left\{\frac{2}{2r+\beta}, \frac{2r-1}{2r+\beta}\right\}}, \tag{15}$$

then

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

3. Relations to existing work and discussions

The minimax analysis of regularized least square algorithm has received attention in statistics and learning theory literature (DeVore et al., 2004; Györfi et al., 2006; Temlyakov, 2008; Caponnetto and De Vito, 2007; Steinwart et al., 2009). In particular, assume L_K admits an eigendecomposition $L_K = \sum_{i=1}^{\infty} \tau_i \phi_i \otimes \phi_i$, where $\tau_i \geq 0$ and ϕ_i are the eigenvalues and eigenfunctions of L_K, respectively. It is proved in Caponnetto and De Vito (2007) that, if the regularity condition (4) holds with some $r \geq \frac{1}{2}$ and the eigenvalues satisfy $\tau_i \sim i^{-2\alpha}$ for some $\alpha > \frac{1}{2}$, then the minimax optimal learning rate of regularized least square algorithm is $\mathcal{O}(|D|^{-\frac{2\alpha}{4\alpha r+1}})$. It is also proved that RKN can achieve minimax rate if $\frac{1}{2} < r \leq 1$. When $r = \frac{1}{2}$, they obtained a suboptimal rate $\mathcal{O}\left(\left(\frac{\log|D|}{|D|}\right)^{-\frac{2\alpha}{2\alpha+1}}\right)$. In Steinwart et al. (2009), under the additional restriction

$$\|f\|_\infty \leq C\|f\|_K^{\frac{1}{2\alpha}} \|f\|_{L^2_{\rho_{\mathcal{X}}}}^{1-\frac{1}{2\alpha}}, \qquad \forall\ f \in \mathcal{H}_K,$$

it is proved that the projected (or clipped) RKN estimator can achieve the minimax learning rate. More recently, Lin et al. (2017) proved that RKN can achieve minimax learning rate for r in the whole range of $[\frac{1}{2}, 1]$ without any restrictions except for the conditions (4) and $\tau_i \sim i^{-2\alpha}$. It improves the results in Caponnetto and De Vito (2007); Steinwart et al. (2009). When $r \geq 1$, RKN suffers the saturation effect and the learning rate will not improve. Note our condition (6) on the effective dimension is nearly equivalent to $\tau_i \sim i^{-2\alpha}$ with $\beta = \frac{1}{2\alpha}$. The result in Corollary 2 tells that BCRKN can achieve the minimax learning rate for $r \in [\frac{1}{2}, 2]$ and thus relaxes the saturation effect of RKN.

For distributed regression problem, assume all data blocks D_i, $i = 1, \ldots, m$, are of equal size. If RKN is used as the base algorithm, under the assumptions that $\mathbf{E}[|\phi_i(x)|^{2k}] \leq A^{2k}$ for some $k > 2$ and constant $A < \infty$, $\lambda_i \leq a i^{-2\alpha}$, 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 $\mathcal{O}(n^{-\frac{2\alpha}{2\alpha+1}})$ can be achieved by choosing $\lambda = |D|^{-\frac{2\alpha}{2\alpha+1}}$ and restricting the number of local processors

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

Later in Lin et al. (2017) the regularity condition (4) was taken into consideration and it is proved that the distributed regression can achieve the minimax optimal rate for all $r \in [\frac{1}{2}, 1]$ if

$$m \leq |D|^{\min\{\frac{6\alpha(2r-1)+1}{5(4\alpha r+1)}, \frac{2\alpha(2r-1)}{4\alpha r+1}\}}. \tag{16}$$

The method suffers from the saturation effect inherited from RKN. So the learning rate cannot improve with $r > 1$. When BCRKN is applied as the base algorithm for distributed regression, the saturation effect is relaxed and the minimax optimal learning rate can be achieved for the whole range $r \in [\frac{1}{2}, 2]$ as in the single data learning case. Compare (15) with (16) and we see our analysis also relaxes the restriction on the number m of local processors.

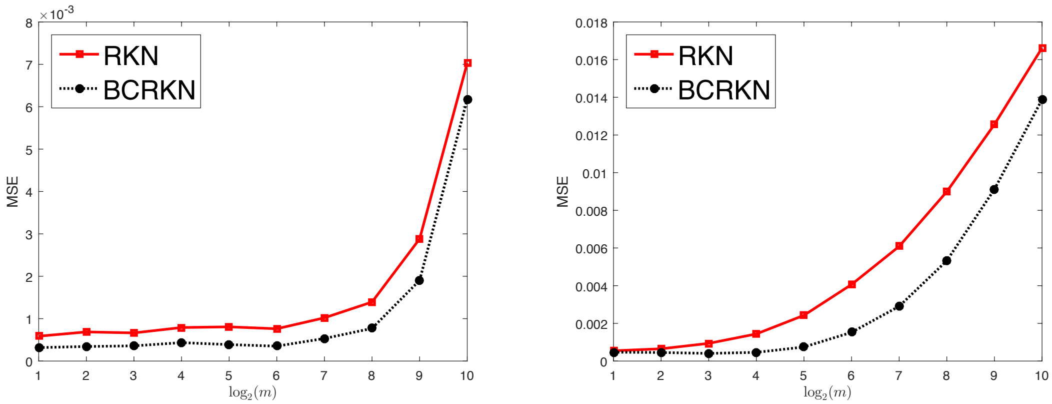
When $r < 1$ we notice that distributed regression with RKN and BCRKN both reach the optimal rates by underregularization, that is, selecting the regularization parameter according to the number of all observations $|D|$, not the number of observations in each block $|D_i|$. But due to the reduced bias the parameter selection of BCRKN is less sensitive and thus could be advantageous in practice. We show this by an illustrative example used in Zhang et al. (2015). Consider the model $f^*(x) = \min\{x, 1-x\}$ with $x \sim \text{Uniform}[0,1]$ and the noise $\epsilon \sim N(0, \sigma^2)$ with $\sigma^2 = \frac{1}{5}$. Let $K(x,t) = 1 + \min\{x,t\}$. Then $f^* \in \mathcal{H}_K$ and $\|f^*\|_K = 1$. We first compare the distributed RKN and the distributed BCRKN when $\lambda = |D|^{-2/3}$, a theoretically optimal choice. We generate $|D| = 4098$ sample points and use number of partitions $m \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. The mean squared errors of two methods are plotted in Figure 1 (a). We see BCRKN slightly outperforms RKN for all m.

Recall that the analyses in Zhang et al. (2015); Lin et al. (2017) and this paper indicate the optimal choice of the regularization parameter is $\lambda = |D|^{-\theta}$ with θ an index depending on the regularity of the true target function f^* and the effective dimension of the integral operator L_K. Clearly both are unknown in practice and thus a theoretical optimal choice of the regularization parameter is actually not available. At the same time, in a big data setting where distributed regression is necessary globally tuning the optimal parameter is either impossible or too time consuming. A reasonable way is to tune the parameter locally to get optimal choice $\lambda_i = |D_i|^{-\theta}$ on D_i and then underregularize it using

$$\lambda = \lambda_i^{\frac{\log|D|}{\log|D_i|}} = |D|^{-\theta}.$$

So we next compare the use of RKN and BCRKN in distributed regression when this parameter selection strategy is used. The results are shown in Figure 1 (b). We see the requirement on the number of local processors becomes more restrictive for both methods, indicating that underregularizing locally optimal parameter does not lead to globally optimal parameter. BCRKN significantly outperforms RKN as m increases, indicating it is less sensitive to the parameter selection when a globally optimal parameter is not available.

Note that the upper bounds on the number of local processors are constrained by the analysis techniques, not necessarily reflect the true limit on the number of local processors allowed in practice. Also, since underregularization is necessary in distributed regression but globally tuning the optimal parameter is impractical, the impact of parameter selection strategy on the number of local processors is unknown. Further investigation on these issues is of great interest in future research.



(a)

(b)

Figure 1: MSE of distributed RKN and distributed BCRKN. (a) $\lambda = |D|^{-2/3}$ is used. (b) λ is first tuned locally and the underregularized.

Finally, it is worth mentioning that debiasing techniques have also been developed to improve the accuracy of statistical inference of distributed lasso and received considerable attention recently (Bühlmann, 2013; Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Lee et al., 2017; Battey et al., 2015).

4. Preliminary lemmas

The following lemmas can be found in Lin et al. (2017); Guo et al. (2017a).

Lemma 7 *Let D be a sample drawn independently according to ρ and g be a measurable bounded function on $\mathcal{Z}$ and ξ_g be a random variable with values on $\mathcal{H}_K$ given by $\xi_g(z) = g(z)K_x$ for $z = (x, y) \in \mathcal{Z}$. For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds*

$$\left\|(\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D|}\sum_{z\in D} \xi_g(z) - \mathbf{E}[\xi_g]\right)\right\|_K \leq \frac{\|g\|_\infty \log\frac{2}{\delta}}{\kappa}\mathcal{B}_{|D|,\lambda}.$$

Lemma 8 *Let D be a sample drawn independently according to ρ. If $|y| \leq M$ almost surely, then with confidence at least $1 - \delta$, there holds*

$$\left\|(\lambda I + L_K)^{-\frac{1}{2}}(\tfrac{1}{|D|} S_D^* \mathbf{y}_D - L_{K,D} f^*)\right\|_K \leq 2M\mathcal{B}_{|D|,\lambda} \log\frac{2}{\delta}.$$

Lemma 9 *Let D be a sample drawn independently according to ρ. If $|y| \leq M$ almost surely, then for any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds*

$$\Xi_D := \|(\lambda I + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\| \leq \mathcal{B}_{|D|,\lambda} \log\frac{2}{\delta}, \tag{17}$$

If A and B are invertible operators on a Banach space, then by the second order operator decomposition proposed in Lin et al. (2017), we have

$$A^{-1} - B^{-1} = B^{-1}(B - A)B^{-1}(B - A)A^{-1} + B^{-1}(B - A)B^{-1}. \tag{18}$$

This implies the following decomposition of the operator product

$$BA^{-1} = (B - A)B^{-1}(B - A)A^{-1} + (B - A)B^{-1} + I. \tag{19}$$

With $A = L_{K,D} + \lambda I$ and $B = L_K + \lambda I$ in (19), and applying Lemma 9, we have the following bound for $\|(L_K + \lambda I)(L_{K,D} + \lambda I)^{-1}\|$; for the detailed proof see Guo et al. (2017a).

Proposition 10 *For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds*

$$\Omega_D := \|(L_K + \lambda I)(L_{K,D} + \lambda I)^{-1}\| \leq \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}} + 1 \right)^2.$$

Moreover, the confidence set is the same as that in Lemma 9.

Lemma 11 *Let Q be positive random variable. If there are constants $a > 0, b > 0, \tau > 0$ such that for any $0 < \delta \leq 1$, with confident at least $1 - \delta$, there holds $Q \leq a(\log \frac{b}{\delta})^\tau$, then for any $s > 0$ we have $\mathbf{E}[Q^s] \leq a^s b \Gamma(\tau s + 1)$.*

Proof Note the condition implies that for all $t > 0$ there is

$$\Pr\left[Q^{\frac{1}{\tau}} > t\right] \leq b \exp\left(-\frac{t}{a^{1/\tau}}\right).$$

So we have

$$\begin{aligned} \mathbf{E}[Q^s] &= \mathbf{E}\left[\left(Q^{1/\tau}\right)^{\tau s}\right] = \tau s \int_0^\infty t^{\tau s - 1} \Pr\left[Q^{\frac{1}{\tau}} > t\right] dt \\ &\leq \tau s b \int_0^\infty t^{\tau s - 1} \exp\left(-\frac{t}{a^{1/\tau}}\right) dt \\ &= b \tau s \Gamma(\tau s) a^s = a^s b \Gamma(\tau s + 1). \end{aligned}$$

This proves the lemma. ■

5. Error analysis of BCRKN in $L^2_{\rho_{\mathcal{X}}}$ when $r \geq \frac{1}{2}$

We will split the proof of Theorem 1 into three cases: $0 < r < \frac{1}{2}$, $\frac{1}{2} \leq r \leq \frac{3}{2}$, and $\frac{3}{2} < r \leq 2$. In this section we prove it for the second and third cases while leave the first case to Section 7. Denote $\Delta_D = \frac{1}{|D|} S_D^* (\mathbf{y}_D - S_D f^*)$.

Proposition 12 *If $\frac{1}{2} \leq r \leq \frac{3}{2}$, we have*

$$\|f^\sharp_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}} \leq 2\Omega_D \left\|(\lambda I + L_K)^{-\frac{1}{2}} \Delta_D\right\|_K + \lambda^r (\Omega_D)^r \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

Proof By the triangle inequality, we have

$$\left\|f^{\sharp}_{D,\lambda}-f^*\right\|_{L^2_{\rho_{\mathcal{X}}}} \leq \left\|f^{\sharp}_{D,\lambda}-\mathbf{E}^*[f^{\sharp}_{D,\lambda}]\right\|_{L^2_{\rho_{\mathcal{X}}}} + \left\|\mathbf{E}^*[f^{\sharp}_{D,\lambda}]-f^*\right\|_{L^2_{\rho_{\mathcal{X}}}}, \tag{20}$$

where $\mathbf{E}^*[f^{\sharp}_{D,\lambda}] = (2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-2} L_{K,D} f^*$ is the conditional expectation with respect to $\mathbf{y}_D$ given $\mathbf{x}_D$.

For the first term $\left\|f^{\sharp}_{D,\lambda}-\mathbf{E}^*[f^{\sharp}_{D,\lambda}]\right\|_{L^2_{\rho_{\mathcal{X}}}}$, noting that $2\lambda I + L_{K,D}$ and $(\lambda I + L_{K,D})^{-1}$ commute, we have

$$\begin{aligned}
&\left\|f^{\sharp}_{D,\lambda}-\mathbf{E}^*[f^{\sharp}_{D,\lambda}]\right\|_{L^2_{\rho_{\mathcal{X}}}} = \left\|L_K^{\frac{1}{2}}(2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-2}\Delta_D\right\|_K \\
\leq\ & \left\|(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D})^{-\frac{1}{2}}\right\| \left\|(\lambda I + L_{K,D})^{-\frac{1}{2}}(2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-\frac{1}{2}}\right\| \\
&\times \left\|(\lambda I + L_{K,D})^{-\frac{1}{2}}(\lambda I + L_K)^{\frac{1}{2}}\right\| \left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta\right\|_K \\
\leq\ & 2\Omega_D \left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K,
\end{aligned} \tag{21}$$

here we have used the fact (Blanchard and Krämer, 2010) that

$$\|A^s B^s\| \leq \|AB\|^s, \qquad 0 \leq s \leq 1,$$

for positive operators A and B on Hilbert spaces.

For the second term, we have

$$\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^* = [(2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-2} L_{K,D} - I] f^* = \lambda^2 (\lambda I + L_{K,D})^{-2} f^*.$$

By the regularity condition (4),

$$\begin{aligned}
\left\|\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^*\right\|_{L^2_{\rho_{\mathcal{X}}}} &= \lambda^2 \left\|(\lambda I + L_{K,D})^{-2} L^r u^*\right\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\leq \lambda^2 \left\|(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D})^{-2} L_K^{r-\frac{1}{2}} L^{\frac{1}{2}} u^*\right\|_K \\
&\leq \lambda^2 \left\|(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D})^{-\frac{1}{2}}\right\| \left\|(\lambda I + L_{K,D})^{-\frac{3}{2}} L_K^{r-\frac{1}{2}}\right\| \left\|L^{\frac{1}{2}} u^*\right\|_K \\
&\leq \lambda^2 (\Omega_D)^{\frac{1}{2}} \left\|(\lambda I + L_{K,D})^{-\frac{3}{2}} L_K^{r-\frac{1}{2}}\right\| \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.
\end{aligned} \tag{22}$$

Since $\frac{1}{2} \leq r \leq \frac{3}{2}$, we have

$$\begin{aligned}
&\left\|(\lambda I + L_{K,D})^{-\frac{3}{2}} L_K^{r-\frac{1}{2}}\right\| \\
=\ & \left\|(\lambda I + L_{K,D})^{r-2}(\lambda I + L_{K,D})^{-r+\frac{1}{2}}(\lambda I + L_K)^{r-\frac{1}{2}}(\lambda I + L_K)^{-r+\frac{1}{2}} L_K^{r-\frac{1}{2}}\right\|
\end{aligned}$$

$$\begin{aligned}
&\leq \left\|(\lambda I+L_{K,D})^{r-2}\right\| \left\|(\lambda I+L_{K,D})^{-r+\frac{1}{2}}(\lambda I+L_K)^{r-\frac{1}{2}}\right\| \left\|(\lambda I+L_K)^{-r+\frac{1}{2}}L_K^{r-\frac{1}{2}}\right\| \\
&\leq \lambda^{r-2}(\Omega_D)^{r-\frac{1}{2}}.
\end{aligned}$$

Therefore,

$$\left\|\mathbf{E}^*[f^\sharp_{D,\lambda}]-f^*\right\|_{L^2_{\rho_\mathcal{X}}} \leq \lambda^r(\Omega_D)^r\|u^*\|_{L^2_{\rho_\mathcal{X}}}. \tag{23}$$

Then the conclusion follows by combining (21) and (23). ∎

Proposition 13 *If* $\frac{3}{2}<r\leq 2$, *we have*

$$\|f^\sharp_{D,\lambda}-f^*\|_{L^2_{\rho_\mathcal{X}}} \leq 2\Omega_D\left\|(\lambda I+L_K)^{-\frac{1}{2}}\Delta_D\right\|_K + \lambda\Xi_D(\Omega_D)^{\frac{3}{2}}\kappa^{2r-3}\|u^*\|_{L^2_{\rho_\mathcal{X}}} + \lambda^r\Omega_D\|u^*\|_{L^2_{\rho_\mathcal{X}}}.$$

Proof The proof is similar to Proposition 12. First, $\|f^\sharp_{D,\lambda}-f^*\|_{L^2_{\rho_\mathcal{X}}}$ can be divided into two terms by (20). The first term has been estimated in Proposition 12 as (21). We now focus on the second term. To this end, by (22), we only need to estimate $\left\|(\lambda I+L_{K,D})^{-\frac{3}{2}}L_K^{r-\frac{1}{2}}\right\|$. When $\frac{3}{2}\leq r<2$, we have

$$\begin{aligned}
&(\lambda I+L_{K,D})^{-\frac{3}{2}}L_K^{r-\frac{1}{2}} \\
=\ &(\lambda I+L_{K,D})^{-\frac{1}{2}}\left[(\lambda I+L_{K,D})^{-1}-(\lambda I+L_K)^{-1}\right]L_K^{r-\frac{1}{2}} \\
&+(\lambda I+L_{K,D})^{-\frac{1}{2}}(\lambda I+L_K)^{-1}L_K^{r-\frac{1}{2}} \\
=\ &(\lambda I+L_{K,D})^{-\frac{1}{2}}(\lambda I+L_K)^{-1}(L_K-L_{K,D})(\lambda I+L_{K,D})^{-1}(\lambda I+L_K)(\lambda I+L_K)^{-1}L_K^{r-\frac{1}{2}} \\
&\quad+(\lambda I+L_{K,D})^{-\frac{1}{2}}(\lambda I+L_K)^{\frac{1}{2}}(\lambda I+L_K)^{-\frac{3}{2}}L_K^{r-\frac{1}{2}}.
\end{aligned}$$

By the bounds $\left\|(\lambda I+L_{K,D})^{-\frac{1}{2}}\right\|\leq\frac{1}{\sqrt{\lambda}}$, $\left\|(\lambda I+L_K)^{-\frac{1}{2}}\right\|\leq\frac{1}{\sqrt{\lambda}}$, and $\|L_K\|\leq\kappa^2$, we have,

$$\begin{aligned}
&\left\|(\lambda I+L_{K,D})^{-\frac{3}{2}}L_K^{r-\frac{1}{2}}\right\| \\
\leq\ &\frac{1}{\lambda}\left\|(\lambda I+L_K)^{-\frac{1}{2}}(L_K-L_{K,D})\right\| \left\|(\lambda I+L_{K,D})^{-1}(\lambda I+L_K)\right\| \left\|(\lambda I+L_K)^{-1}L_K^{r-\frac{1}{2}}\right\| \\
&\quad+\left\|(\lambda I+L_{K,D})^{-\frac{1}{2}}(\lambda I+L_K)^{\frac{1}{2}}\right\| \left\|(\lambda I+L_K)^{-\frac{3}{2}}L_K^{r-\frac{1}{2}}\right\| \\
\leq\ &\lambda^{-1}\Omega_D\left\|(\lambda I+L_K)^{-\frac{1}{2}}(L_K-L_{K,D})\right\|\kappa^{2r-3}+\lambda^{r-2}(\Omega_D)^{\frac{1}{2}}.
\end{aligned}$$

Therefore, putting the above bound back into (22) yields

$$\begin{aligned}
\left\|\mathbf{E}^*[f^\sharp_{D,\lambda}]-f^*\right\|_{L^2_{\rho_\mathcal{X}}} \leq\ &\lambda\left\|(\lambda I+L_K)^{-\frac{1}{2}}(L_K-L_{K,D})\right\|(\Omega_D)^{\frac{3}{2}}\kappa^{2r-3}\|u^*\|_{L^2_{\rho_\mathcal{X}}} \\
&+\lambda^r\Omega_D\|u^*\|_{L^2_{\rho_\mathcal{X}}}.
\end{aligned} \tag{24}$$

Now the conclusion follows by plugging (21) and (24) into (20). ■

Now we are ready to prove Theorem 1 and Corollary 2 for $r \geq \frac{1}{2}$.

Proof of Theorem 1: Case $\frac{1}{2} \leq r \leq 2$. By Lemma 8, we have with confidence at least $1 - \frac{\delta}{2}$,

$$\left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K \leq 2M\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}. \tag{25}$$

By Lemma 9 and Proposition 10, we obtain that, with confidence at least $1 - \frac{\delta}{2}$,

$$\Xi_D = \|(\lambda I + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\| \leq \mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta} \tag{26}$$

and

$$\Omega_D = \|(L_K + \lambda I)(L_{K,D} + \lambda I)^{-1}\| \leq \left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^2 \tag{27}$$

hold simultaneously.

When $\frac{1}{2} \leq r \leq \frac{3}{2}$, we apply (25) and (27) to Proposition 12 and obtain

$$\begin{aligned}
\|f^{\sharp}_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}} &\leq 4M\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^2\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta} + \lambda^r\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^{2r}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\leq (4M + \|u^*\|_{L^2_{\rho_{\mathcal{X}}}})\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)^3(\mathcal{B}_{|D|,\lambda} + \lambda^r)\left(\log\frac{4}{\delta}\right)^4
\end{aligned}$$

When $\frac{3}{2} < r \leq 2$, we apply (25), (26) and (27) to Proposition 13 and obtain

$$\begin{aligned}
\|f^{\sharp}_{D,\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}} &\leq 4M\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^2\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta} \\
&\quad + \lambda\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^3\kappa^{2r-3}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\quad + \lambda^r\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^2\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\leq (4M + 2\kappa^{2r-3}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}})\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)^3(\mathcal{B}_{|D|,\lambda} + \lambda^r)\left(\log\frac{4}{\delta}\right)^4.
\end{aligned}$$

So (7) is proved for all $\frac{1}{2} \leq r \leq 2$. Applying Lemma 11 with $b = 4$, $\tau = 4$, and $s = 2$, we obtain the desired estimation in (8). ■

Proof of Corollary 2 (ii). With $\frac{1}{2} \leq r \leq 2$ and the choice of $\lambda = |D|^{-\frac{1}{2r+\beta}}$, we have

$$B_{|D|,\lambda} \leq \frac{2\kappa}{\sqrt{|D|}}\left\{\frac{\kappa|D|^{\frac{1}{2(2r+\beta)}}}{\sqrt{|D|}} + \sqrt{C_0}|D|^{\frac{\beta}{2(2r+\beta)}}\right\} \leq 2\kappa\left(\kappa + \sqrt{C_0}\right)|D|^{-\frac{r}{2r+\beta}} \tag{28}$$

and

$$\frac{B_{|D|,\lambda}}{\sqrt{\lambda}} + 1 \leq 2\kappa(\kappa + \sqrt{C_0})|D|^{-\frac{r}{2r+\beta}}|D|^{\frac{1}{2r+\beta}} + 1 \leq 2\kappa\left(\kappa + \sqrt{C_0}\right) + 1. \tag{29}$$

Then the conclusions follow from Theorem 1. ■

6. Error analysis in $\mathcal{H}_K$

In this section, we derive the error bound for $\|f^{\sharp}_{D,\lambda} - f^*\|_K$ and prove the convergence of BCRKN in $\mathcal{H}_K$. It is similar to the error analysis in $L^2_{\rho_{\mathcal{X}}}$.

Proposition 14 *If $r \in [\frac{1}{2}, \frac{3}{2}]$, we have*

$$\|f^{\sharp}_{D,\lambda} - f^*\|_K \leq 2\lambda^{-\frac{1}{2}}(\Omega_D)^{\frac{1}{2}}\left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K + \lambda^{r-\frac{1}{2}}(\Omega_D)^{r-\frac{1}{2}}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

If $r \in (\frac{3}{2}, 2]$, we have

$$\begin{aligned}\|f^{\sharp}_{D,\lambda} - f^*\|_K \quad &\leq \quad 2\lambda^{-\frac{1}{2}}(\Omega_D)^{\frac{1}{2}}\left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K + \lambda^{\frac{1}{2}}\Xi_D\Omega_D\kappa^{2r-3}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\ &\quad + \lambda^{r-\frac{1}{2}}\Omega_D\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.\end{aligned}$$

Proof By the triangle inequality in $\mathcal{H}_K$, we have

$$\|f^{\sharp}_{D,\lambda} - f^*\|_K \leq \|f^{\sharp}_{D,\lambda} - \mathbf{E}^*[f^{\sharp}_{D,\lambda}]\|_K + \|\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^*\|_K.$$

To estimate the first term, we see that

$$f^{\sharp}_{D,\lambda} - \mathbf{E}^*[f^{\sharp}_{D,\lambda}] = (2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-2}\Delta_D.$$

Then

$$\begin{aligned}&\left\|f^{\sharp}_{D,\lambda} - \mathbf{E}^*[f^{\sharp}_{D,\lambda}]\right\|_K = \|(2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-2}\Delta_D\|_K \\ \leq\ & \left\|(2\lambda I + L_{K,D})(\lambda I + L_{K,D})^{-\frac{3}{2}}\right\|\left\|(\lambda I + L_{K,D})^{-\frac{1}{2}}(\lambda I + L_K)^{\frac{1}{2}}\right\|\left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K \\ \leq\ & 2\lambda^{-\frac{1}{2}}(\Omega_D)^{\frac{1}{2}}\left\|(\lambda I + L_K)^{-\frac{1}{2}}\Delta_D\right\|_K.\end{aligned}$$

For the second term, we have

$$\left\|\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^*\right\|_K = \left\|\lambda^2(\lambda I + L_{K,D})^{-2}L^r u^*\right\|_K \leq \lambda^2\|(\lambda I + L_{K,D})^{-2}L_K^{r-\frac{1}{2}}\|\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

Following the same idea as in the proof of Proposition 12, we obtain for $r \in [\frac{1}{2}, \frac{3}{2}]$,

$$\left\|\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^*\right\|_K \leq \lambda^{r-\frac{1}{2}}(\Omega_D)^{r-\frac{1}{2}}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}$$

and following the idea in the proof of Proposition 13, we obtain for $r \in (\frac{3}{2}, 2]$,

$$\left\|\mathbf{E}^*[f^{\sharp}_{D,\lambda}] - f^*\right\|_K \leq \lambda^{\frac{1}{2}}\Xi_D\Omega_D\kappa^{2r-3}\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} + \lambda^{r-\frac{1}{2}}\Omega_D\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

The desired error bounds now follow by combining the estimates for both terms. ■

Proof of Theorem 3. Note that (25), (26) and (27) hold simultaneously with probability at least $1-\delta$. Therefore, when $\frac{1}{2} \leq r \leq \frac{3}{2}$, we have with confidence at least $1-\delta$

$$\begin{aligned}
\|f^{\sharp}_{D,\lambda} - f^*\|_K &\leq 2\lambda^{-\frac{1}{2}} \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1 \right) 2M \mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta} \\
&\quad + \lambda^{r-\frac{1}{2}} \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1 \right)^{2r-1} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\leq (4M + \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}) \left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1 \right)^2 (\lambda^{-\frac{1}{2}} \mathcal{B}_{|D|,\lambda} + \lambda^{r-\frac{1}{2}}) \left(\log \frac{4}{\delta} \right)^3
\end{aligned}$$

and, when $\frac{3}{2} < r \leq 2$,

$$\begin{aligned}
\|f^{\sharp}_{D,\lambda} - f^*\|_K &\leq 2\lambda^{-\frac{1}{2}} \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1 \right) 2M \mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta} \\
&\quad + \lambda^{\frac{1}{2}} \mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta} \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1 \right)^2 \kappa^{2r-3} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\quad + \lambda^{r-\frac{1}{2}} \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1 \right)^2 \|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \\
&\leq \left(4M + 2\kappa^{2r-3} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}\right) \left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}} + 1 \right)^2 (\lambda^{-\frac{1}{2}} \mathcal{B}_{|D|,\lambda} + \lambda^{r-\frac{1}{2}}) \left(\log \frac{4}{\delta} \right)^3.
\end{aligned}$$

This proves the error bound (9). Then (10) follows from estimates (28) and (29), and (11) follows by applying Lemma 11. ■

7. Improve the error analysis by unlabelled data

The error analysis for the semi-supervised approach is more involved. Before we move on, notice that Theorem 1 with $0 < r < \frac{1}{2}$ is a special case of Theorem 4 with $D' = D$ when there is no unlabeled data. So upon finishing Theorem 4, we also obtain Theorem 1 with $0 < r < \frac{1}{2}$.

We need to introduce an intermediate function. Recall L is a compact operator on $L^2_{\rho_{\mathcal{X}}}$. Let $\{\tau_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ be the eigenvalues and eigenfunctions of L. Then $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L^2_{\rho_{\mathcal{X}}}$. Let P_λ be the projection operator on $L^2_{\rho_{\mathcal{X}}}$ that projects each $f \in L^2_{\rho_{\mathcal{X}}}$ onto the subspace spanned by $\{\psi_i : \tau_i \geq \lambda\}$, i.e.

$$P_\lambda f = \sum_{\{i:\tau_i \geq \lambda\}} \langle \psi_i, f \rangle_{L^2_{\rho_{\mathcal{X}}}} \psi_i, \qquad \forall\ f \in L^2_{\rho_{\mathcal{X}}}.$$

By the isomorphism property (5) of $L^{\frac{1}{2}}$, $\{\phi_i = \sqrt{\tau_i}\psi_i : \sigma_i > 0\}$ form an orthonormal basis of $\mathcal{H}_K$. Since $\{i : \tau_i \geq \lambda\}$ is a finite set, it is obvious $P_\lambda f \in \mathcal{H}_K$ for all $f \in L^2_{\rho_\mathcal{X}}$. Define $f_\lambda^{tr} = P_\lambda f^*$. We can bound $\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}}$ as follows.

Proposition 15 *We have*

$$\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq I_1 + I_2 + I_3,$$

where

$$\begin{aligned}
I_1 &= \left\|(\lambda I + L_K)^{\frac{1}{2}}(2\lambda I + L_{K,D'})(\lambda I + L_{K,D'})^{-2}\left(\tfrac{1}{|D'|}S^*_{D'}\mathbf{y}_{D'} - L_{K,D'}f^{tr}_\lambda\right)\right\|_K \\
I_2 &= \left\|\lambda^2(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D'})^{-2}f^{tr}_\lambda\right\|_K, \\
I_3 &= \|f^{tr}_\lambda - f^*\|_{L^2_{\rho_\mathcal{X}}}.
\end{aligned}$$

Proof Note that

$$\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq \|f^\sharp_{D',\lambda} - f^{tr}_\lambda\|_{L^2_{\rho_\mathcal{X}}} + \|f^{tr}_\lambda - f^*\|_{L^2_{\rho_\mathcal{X}}}. \tag{30}$$

Since $f^\sharp_{D',\lambda} - f^{tr}_\lambda \in \mathcal{H}_K$, by the isometry property (5) of $L^{\frac{1}{2}} = L_K^{\frac{1}{2}}$, we have

$$\|f^\sharp_{D',\lambda} - f^{tr}_\lambda\|_{L^2_{\rho_\mathcal{X}}} = \|L_K^{\frac{1}{2}}(f^\sharp_{D',\lambda} - f^{tr}_\lambda)\|_K \leq \|(\lambda I + L_K)^{\frac{1}{2}}(f^\sharp_{D',\lambda} - f^{tr}_\lambda)\|_K. \tag{31}$$

Recall that

$$f^\sharp_{D',\lambda} = f_{D',\lambda} + \lambda(\lambda I + L_{K,D'})^{-1}f_{D',\lambda} = (2\lambda I + L_{K,D'})(\lambda I + L_{K,D'})^{-2}\frac{1}{|D'|}S^*_{D'}\mathbf{y}_{D'}.$$

It is easy to check that

$$\begin{aligned}
f^\sharp_{D',\lambda} - f^{tr}_\lambda &= (2\lambda I + L_{K,D'})(\lambda I + L_{K,D'})^{-2}\left(\frac{1}{|D'|}S^*_{D'}\mathbf{y}_{D'} - L_{K,D'}f^{tr}_\lambda\right) \\
&\quad -\lambda^2(\lambda I + L_{K,D'})^{-2}f^{tr}_\lambda.
\end{aligned}$$

Putting this in (31) we have $\|f^\sharp_{D',\lambda} - f^{tr}_\lambda\|_{L^2_{\rho_\mathcal{X}}}$ bounded by $I_1 + I_2$. Together with (30), we obtain the desired conclusion. ■

Next we estimate the three terms respectively. The third term I_3 can be easily bounded by the following lemma, which has been proved in Caponnetto (2006).

Lemma 16 *We have* $\|f^{tr}_\lambda - f^*\|_{L^2_{\rho_\mathcal{X}}} \leq \lambda^r\|u^*\|_{L^2_{\rho_\mathcal{X}}}$ *and* $\|f^{tr}_\lambda\|_K \leq \lambda^{-\frac{1}{2}+r}\|u^*\|_{L^2_{\rho_\mathcal{X}}}$.

For the first term I_1, we have the following bound.

Proposition 17 *For any* $\delta \in (0,1)$, *with confidence at least* $1-\delta$, *there holds*

$$I_1 \leq \left(\frac{2M}{\kappa} + 2\|u^*\|_{L^2_{\rho_\mathcal{X}}}\right)\left(\frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}} + 1\right)^3\left(\mathcal{B}_{|D|,\lambda} + \lambda^r\right)\left(\log\frac{4}{\delta}\right)^3.$$

Proof Since $2\lambda I + L_{K,D'}$ and $(\lambda I + L_{K,D'})^{-1}$ commute, we have

$$\begin{aligned} I_1 &= \left\| (\lambda I + L_K)^{\frac{1}{2}} (2\lambda I + L_{K,D'})(\lambda I + L_{K,D'})^{-2} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_{K,D'} f_\lambda^{tr} \right) \right\|_K \\ &\leq \left\| (\lambda I + L_K)^{\frac{1}{2}} (\lambda I + L_{K,D'})^{-\frac{1}{2}} \right\| \left\| (2\lambda I + L_{K,D'})(\lambda I + L_{K,D'})^{-1} \right\| \\ &\quad \times \left\| (\lambda I + L_{K,D'})^{-\frac{1}{2}} (\lambda I + L_K)^{\frac{1}{2}} \right\| \left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_{K,D'} f_\lambda^{tr} \right) \right\|_K \\ &\leq 2\Omega_{D'} \left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_{K,D'} f_\lambda^{tr} \right) \right\|_K . \end{aligned}$$

Proposition 10 ensures that, with confidence at least $1 - \frac{\delta}{2}$,

$$\Omega_{D'} \leq \left(\frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}} + 1 \right)^2 \left(\log \frac{4}{\delta} \right)^2 . \tag{32}$$

Now it suffices to consider the term $\left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_{K,D'} f_\lambda^{tr} \right) \right\|_K$. We further divide it into three parts as follows

$$\begin{aligned} \left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_{K,D'} f_\lambda^{tr} \right) \right\| &\leq \left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_K f^* \right) \right\|_K \\ &\quad + \left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K f^* - L_K f_\lambda^{tr}) \right\|_K \\ &\quad + \left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K - L_{K,D'}) f_\lambda^{tr} \right\|_K . \end{aligned}$$

By the definition of y_i', it is easy to check that $\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} = \frac{1}{|D|} S_D^* \mathbf{y}_D$. Applying Lemma 7 with $\xi_g(z) = yK_x$, we obtain, with confidence at least $1 - \frac{\delta}{2}$,

$$\left\| (\lambda I + L_K)^{-\frac{1}{2}} \left(\frac{1}{|D'|} S_{D'}^* \mathbf{y}_{D'} - L_K f^* \right) \right\|_K \leq \frac{M \log \frac{4}{\delta}}{\kappa} \mathcal{B}_{|D|,\lambda}.$$

By Lemma 16, we have

$$\left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K f^* - L_K f_\lambda^{tr}) \right\|_K \leq \| f^* - f_\lambda^{tr} \|_{L^2_{\rho_{\mathcal{X}}}} \leq \lambda^r \| u^* \|_{L^2_{\rho_{\mathcal{X}}}} .$$

For $\left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K - L_{K,D'}) f_\lambda^{tr} \right\|_K$, observe that

$$\left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K - L_{K,D'}) f_\lambda^{tr} \right\|_K \leq \left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K - L_{K,D'}) \right\| \| f_\lambda^{tr} \|_K .$$

By Lemma 16, we have $\| f_\lambda^{tr} \|_K \leq \lambda^{-\frac{1}{2}+r} \| u^* \|_{L^2_{\rho_{\mathcal{X}}}}$. By Lemma 9, we have with confidence at least $1 - \frac{\delta}{2}$,

$$\left\| (\lambda I + L_K)^{-\frac{1}{2}} (L_K - L_{K,D'}) f_\lambda^{tr} \right\|_K \leq \lambda^r \frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}} \| u^* \|_{L^2_{\rho_{\mathcal{X}}}} \log \frac{4}{\delta},$$

with the confidence set the same as that for (32). Combining the above estimations together yields

$$\begin{aligned} & \left\|(\lambda I + L_K)^{-\frac{1}{2}}\left(\frac{1}{|D'|}S^*_{D'}\mathbf{y}_{D'} - L_{K,D'}f^{tr}_\lambda\right)\right\|_K \\ \leq & \left(\frac{M}{\kappa} + \|u^*\|_{L^2_{\rho_\mathcal{X}}}\right)\left(\frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}}+1\right)\left(\mathcal{B}_{|D|,\lambda}+\lambda^r\right)\log\frac{4}{\delta}. \end{aligned} \tag{33}$$

Then our desired result follows by (32) and (33). ∎

Proposition 18 *For any $\delta \in (0,1)$, we have, with confidence at least $1-\frac{\delta}{2}$,*

$$I_2 \leq \left(\frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}}+1\right)\lambda^r\|u^*\|_{L^2_{\rho_\mathcal{X}}}\log\frac{4}{\delta}$$

with the confidence set the same as that for (32)*.*

Proof We have

$$\begin{aligned} I_2 &= \|\lambda^2(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D'})^{-2}f^{tr}_\lambda\|_K \\ &= \lambda^2\left\|(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D'})^{-\frac{1}{2}}(\lambda I + L_{K,D'})^{-\frac{3}{2}}f^{tr}_\lambda\right\|_K \\ &\leq \lambda^2\|(\lambda I + L_K)^{\frac{1}{2}}(\lambda I + L_{K,D'})^{-\frac{1}{2}}\|\|(\lambda I + L_{K,D'})^{-\frac{3}{2}}\|\|f^{tr}_\lambda\|_K \\ &\leq \lambda^r(\Omega_{D'})^{\frac{1}{2}}\|u^*\|_{L^2_{\rho_\mathcal{X}}}, \end{aligned}$$

where we used the bounds $\|(\lambda I + L_{K,D'})^{-\frac{3}{2}}\| \leq \lambda^{-\frac{3}{2}}$ and $\|f^{tr}_\lambda\|_K \leq \lambda^{r-\frac{1}{2}}\|u^*\|_{L^2_{\rho_\mathcal{X}}}$. By (32), we obtain the desired bound and confidence set. ∎

Now we can prove Theorem 4.

Proof of Theorem 4. Plugging the bounds of I_1, I_2, and I_3 into Proposition 15, we obtain the error bound for $\|f^\sharp_{D',\lambda} - f^*\|_{L^2_{\rho_\mathcal{X}}}$ in (12).

If $\lambda = |D|^{-\frac{1}{2r+\beta}}$ with $0 < r \leq \frac{1}{2}$ and $\mathcal{N}(\lambda) \leq C_0\lambda^{-\beta}$, we have

$$B_{|D|,\lambda} \leq \frac{2\kappa}{\sqrt{|D|}}\left\{\kappa|D|^{-\frac{1}{2}}|D|^{\frac{1}{2(2r+\beta)}} + \sqrt{C_0}|D|^{\frac{\beta}{2(2r+\beta)}}\right\} \leq 2\kappa(\kappa+\sqrt{C_0})|D|^{-\frac{r}{2r+\beta}}.$$

Under the condition $|D'| \geq |D|^{\frac{1+\beta}{2r+\beta}}$, we have

$$\begin{aligned} \frac{\mathcal{B}_{|D'|,\lambda}}{\sqrt{\lambda}} &= \frac{2\kappa}{\sqrt{|D'|\lambda}}\left\{\frac{\kappa}{\sqrt{|D'|\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\} \\ &\leq \frac{2\kappa}{\sqrt{|D'|}}|D|^{\frac{1}{2(2r+\beta)}}\left\{\kappa|D'|^{-\frac{1}{2}}|D|^{\frac{1}{2(2r+\beta)}} + \sqrt{C_0}|D|^{\frac{\beta}{2(2r+\beta)}}\right\} \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2\kappa}{\sqrt{|D'|}}|D|^{\frac{1}{2(2r+\beta)}}(\kappa+\sqrt{C_0})|D|^{\frac{\beta}{2(2r+\beta)}} \\
&\leq 2\kappa(\kappa+\sqrt{C_0}).
\end{aligned}
$$

Applying these two estimates to (12), we have for any $\delta \in (0,1)$, with confidence at least $1-\delta$,

$$
\|f^{\sharp}_{D',\lambda} - f^*\|_{L^2_{\rho_{\mathcal{X}}}} \leq C_1 |D|^{-\frac{r}{2r+\beta}} \left(\log\frac{4}{\delta}\right)^3
$$

where

$$
C_1 = \left(\frac{2M}{\kappa} + 4\|u^*\|_{L^2_{\rho_{\mathcal{X}}}}\right)\left[2\kappa(\kappa+\sqrt{C_0})+1\right]^3\left[2\kappa(\kappa+\sqrt{C_0})+1\right].
$$

This completes the proof of Theorem 4. ■

Note we also proved Theorem 1 with $0 < r < \frac{1}{2}$ because it is a special case of Theorem 4 with $D' = D$. So we are in position to prove Corollary 2 (i).

Proof of Corollary 2 (i). When $0 < r \leq \frac{1}{2}$, take $\lambda = |D|^{-\frac{1}{1+\beta}}$. Then

$$
\begin{aligned}
B_{|D|,\lambda} &\leq \frac{2\kappa}{\sqrt{|D|}}\left\{\frac{\kappa|D|^{\frac{1}{2(1+\beta)}}}{\sqrt{|D|}} + \sqrt{C_0}|D|^{\frac{\beta}{2(1+\beta)}}\right\} \\
&\leq 2\kappa(\kappa+\sqrt{C_0})|D|^{-\frac{1}{2(1+\beta)}} \leq 2\kappa(\kappa+\sqrt{C_0})|D|^{-\frac{r}{1+\beta}}
\end{aligned}
$$

and

$$
\frac{B_{|D|,\lambda}}{\sqrt{\lambda}} \leq 2\kappa(\kappa+\sqrt{C_0})|D|^{-\frac{1}{2(1+\beta)}}|D|^{\frac{1}{2(1+\beta)}} = 2\kappa(\kappa+\sqrt{C_0}).
$$

Plugging them into the estimation (7) we obtain the desired learning rate. ■

8. Error analysis for distributed BCRKN

The following lemma is analogous to Guo et al. (2017a, Proposition 4).

Lemma 19 *Let $\overline{f}^{\sharp}_{D,\lambda}$ be defined by (14). We have*

$$
\mathbf{E}\left[\|\overline{f}^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] \leq \sum_{j=1}^m \frac{|D_j|^2}{|D|^2}\mathbf{E}\left[\|f^{\sharp}_{D_j,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\|\mathbf{E}[f^{\sharp}_{D_j,\lambda}] - f^*\right\|^2_{L^2_{\rho_{\mathcal{X}}}}. \quad (34)
$$

Proof of Theorem 5. By Theorem 1, for each fixed $j \in \{1,2,\ldots,m\}$,

$$
\mathbf{E}\left[\|f^{\sharp}_{D_j,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] \leq 4\Gamma(9)C^2\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6\left(\mathcal{B}_{|D_j|,\lambda}+\lambda^r\right)^2.
$$

Then the first term on the right of (34) can be estimated as

$$\sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \mathbf{E}\left[\|f^{\sharp}_{D_j,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}\right] \leq 4\Gamma(9)C^2 \sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^6 \left(\mathcal{B}_{|D_j|,\lambda} + \lambda^r\right)^2. \quad (35)$$

To estimate the second term on the right of (34), for each fixed $j \in \{1, 2, \ldots, m\}$, by Jensen's inequality, we have

$$\left\|\mathbf{E}[f^{\sharp}_{D_j,\lambda}] - f^*\right\|_{L^2_{\rho_{\mathcal{X}}}} \leq \mathbf{E}\left[\|\mathbf{E}^*[f^{\sharp}_{D_j,\lambda}] - f^*\|_{L^2_{\rho_{\mathcal{X}}}}\right].$$

We will bound the second term in two different ways according to the range of r. First consider the case when $\frac{1}{2} \leq r \leq \frac{3}{2}$. The bound (23) in the proof of Proposition 12 tells us that

$$\left\|\mathbf{E}^*[f^{\sharp}_{D_j,\lambda}] - f^*\right\|_{L^2_{\rho_{\mathcal{X}}}} \leq \lambda^r \Omega^r_{D_j} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

It follows that

$$\|\mathbf{E}[f^{\sharp}_{D_j,\lambda}] - f^*\|_{L^2_{\rho_{\mathcal{X}}}} \leq \lambda^r \|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \mathbf{E}\left[\Omega^r_{D_j}\right]. \quad (36)$$

Applying Proposition 10 to each fixed $j \in \{1, \ldots, m\}$, with confidence at least $1 - \frac{\delta}{2}$, there holds

$$\Omega_{D_j} \leq \left(\frac{\mathcal{B}_{|D_j|,\lambda} \log \frac{4}{\delta}}{\sqrt{\lambda}} + 1\right)^2 \leq \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^2 \left(\log \frac{4}{\delta}\right)^2.$$

By Lemma 11, this implies that for any $s > 0$,

$$\mathbf{E}\left[\Omega^s_{D_j}\right] \leq 2\Gamma(2s+1) \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^{2s}. \quad (37)$$

Applying (37) with $s = r$ to (36) yields

$$\|\mathbf{E}[f^{\sharp}_{D_j,\lambda}] - f^*\|_{L^2_{\rho_{\mathcal{X}}}} \leq 2\Gamma(2r+1)\|u^*\|_{L^2_{\rho_{\mathcal{X}}}} \lambda^r \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^{2r}. \quad (38)$$

Combining (34), (35) and (38), we have

$$\begin{aligned} \mathbf{E}[\|\overline{f}^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}}] \quad &\leq \quad 4\Gamma(9)C^2 \sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^6 \left(\mathcal{B}_{|D_j|,\lambda} + \lambda^r\right)^2 \\ &\quad + 4\Gamma^2(2r+1)\|u^*\|^2_{L^2_{\rho_{\mathcal{X}}}} \lambda^{2r} \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} + 1\right)^{4r}. \end{aligned}$$

This proves the desired bound for $\frac{1}{2} \leq r \leq \frac{3}{2}$.

For $\frac{3}{2} < r \leq 2$, by the bound (24) in the proof of Proposition 13, we have

$$\|\mathbf{E}^*[f^{\sharp}_{D_j,\lambda}] - f^*\|_{L^2_{\rho_{\mathcal{X}}}} \leq \lambda \Xi_{D_j} \Omega^{\frac{3}{2}}_{D_j} \kappa^{2r-3} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}} + \lambda^r \Omega_{D_j} \|u^*\|_{L^2_{\rho_{\mathcal{X}}}}.$$

So,

$$\begin{aligned}\mathbf{E}[\|\mathbf{E}^*[f^\sharp_{D_j,\lambda}]-f^*\|_{L^2_{\rho_\mathcal{X}}}] &\leq \lambda\mathbf{E}\left[\Xi_{D_j}(\Omega_D)^{\frac{3}{2}}\right]\kappa^{2r-3}\|u^*\|_{L^2_{\rho_\mathcal{X}}}+\lambda^r\mathbf{E}[\Omega_{D_j}]\|u^*\|_{L^2_{\rho_\mathcal{X}}} \\ &\leq \lambda\left(\mathbf{E}\left[\Xi^2_{D_j}\right]\right)^{\frac{1}{2}}\left(\mathbf{E}\left[\Omega^3_{D_j}\right]\right)^{\frac{1}{2}}\kappa^{2r-3}\|u^*\|_{L^2_{\rho_\mathcal{X}}}+\lambda^r\mathbf{E}[\Omega_{D_j}]\|u^*\|_{L^2_{\rho_\mathcal{X}}}.\end{aligned}$$

From Lin et al. (2017, Lemma 17), we know

$$\mathbf{E}\left[\left\|(\lambda I+L_K)^{-\frac{1}{2}}(L_K-L_{K,D})\right\|^2\right]\leq\frac{\kappa^2\mathcal{N}(\lambda)}{|D|}$$

and by (37) with $s=3$, we have

$$\mathbf{E}\left[\Omega^3_{D_j}\right]\leq 2\Gamma(7)\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6.$$

Therefore,

$$\begin{aligned}&\|\mathbf{E}[f^\sharp_{D_j,\lambda}]-f^*\|^2_{L^2_{\rho_\mathcal{X}}} \\ \leq\;& 2\lambda^2\mathbf{E}\left[\Xi^2_{D_j}\right]\mathbf{E}\left[\Omega^3_{D_j}\right]\kappa^{4r-6}\|u^*\|^2_{L^2_{\rho_\mathcal{X}}}+2\lambda^{2r}(\mathbf{E}[\Omega_{D_j}])^2\|u^*\|^2_{L^2_{\rho_\mathcal{X}}} \\ \leq\;& 2\lambda^2\frac{\kappa^2\mathcal{N}(\lambda)}{|D_j|}2\Gamma(7)\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6\kappa^{4r-6}\|u^*\|^2_{L^2_{\rho_\mathcal{X}}} \\ &+32\lambda^{2r}\|u^*\|^2_{L^2_{\rho_\mathcal{X}}}\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^4 \\ \leq\;& \left(4\Gamma(7)\kappa^{4r-4}+32\right)\|u^*\|^2_{L^2_{\rho_\mathcal{X}}}\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6\left(\frac{\lambda^2\mathcal{N}(\lambda)}{|D_j|}+\lambda^{2r}\right).\end{aligned}\tag{39}$$

Combining (34), (35) and (39), we have

$$\begin{aligned}&\mathbf{E}[\|\overline{f}^\sharp_{D,\lambda}-f^*\|^2_{L^2_{\rho_\mathcal{X}}}] \\ \leq\;& 4\Gamma(9)C^2\sum_{j=1}^m\frac{|D_j|^2}{|D|^2}\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6\left(\mathcal{B}_{|D_j|,\lambda}+\lambda^r\right)^2 \\ &+\left(4\Gamma(7)\kappa^{4r-4}+32\right)\|u^*\|^2_{L^2_{\rho_\mathcal{X}}}\sum_{j=1}^m\frac{|D_j|}{|D|}\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}+1\right)^6\left(\frac{\lambda^2\mathcal{N}(\lambda)}{|D_j|}+\lambda^{2r}\right).\end{aligned}$$

This proves the conclusion for $\frac{3}{2}<r\leq 2$. ■

Next let us turn to the special case that the local machines are assigned the same number of observations, i.e., $|D_1|=|D_2|=\cdots=|D_m|$.

Proof of Theorem 6. For $\frac{1}{2}\leq r\leq 2$, we choose $\lambda=|D|^{-\frac{1}{2r+\beta}}$. Then by the capacity assumption (6), restriction on the number of local machines $m\leq|D|^{\min\{\frac{2r-1}{2r+\beta},\frac{2}{2r+\beta}\}}$, and the fact $|D_1| = |D_2| = \ldots = |D_m| = \frac{|D|}{m}$, we have

$$\frac{\mathcal{N}(\lambda)}{\lambda|D_j|} \leq C_0 m N^{\frac{1-2r}{2r+\beta}} \leq C_0.$$

It follows that, for each $j = 1, \ldots, m$,

$$\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}} = \frac{2\kappa}{\sqrt{\lambda|D_j|}} \left\{ \frac{\kappa}{\sqrt{|D_j|\lambda}} + \sqrt{\mathcal{N}(\lambda)} \right\} \leq 2\kappa(\kappa + \sqrt{C_0})$$

and

$$\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} \leq 8\kappa^2 \left(\frac{\kappa^2}{|D||D_j|\lambda} + \frac{\mathcal{N}(\lambda)}{|D|} \right) \leq 8\kappa^2(\kappa^2 + C_0)|D|^{-\frac{2r}{2r+\beta}}$$

and

$$\frac{\lambda^2\mathcal{N}(\lambda)}{|D_j|} \leq \frac{C_0|D|^{-\frac{2}{2r+\beta}}|D|^{\frac{\beta}{2r+\beta}}m}{|D|} \leq C_0|D|^{-\frac{2r}{2r+\beta}}.$$

Then by Theorem 5,

$$\mathbf{E}\left[\|\overline{f}^{\sharp}_{D,\lambda} - f^*\|^2_{L^2_{\rho_{\mathcal{X}}}} \right] \leq \bar{C} \left(2\kappa(\kappa + \sqrt{C_0}) + 1 \right)^6 \left(16\kappa^2(\kappa^2 + C_0) + C_0 + 1 \right) |D|^{-\frac{2r}{2r+\beta}}.$$

This completes the proof of Theorem 6. ■

Acknowledgments

The work described in this paper is partially supported by the National Natural Science Foundation of China (Grants No. 11401524, 11531013, 11571078, 11631015, 11671171). Lei Shi is also supported by the Joint Research Fund by National Natural Science Foundation of China and Research Grants Council of Hong Kong (Project No. 11461161006 and Project No. CityU 104012) and Zhuo Xue Program of Fudan University. Qiang Wu is also supported in part by the USDA National Institute of Food and Agriculture (Grant Number 2016-70001-24636). Part of the work was carried out while Zheng-Chu Guo was visiting Shanghai Key Laboratory for Contemporary Applied Mathematics. All authors contributed equally to this paper and are listed alphabetically. The corresponding author is Qiang Wu.

References

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

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed estimation and inference with statistical guarantees. 2015. arXiv preprint arXiv:1509.05457.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

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

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

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MIT, 2006. `http://cbcl.mit.edu/publications/ps/MIT-CSAIL-TR-2006-062.pdf`.

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

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

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. *IMI Preprints*, 22:1–51, 2004.

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

X. Guo, J. Fan, and D.-X. Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *Journal of Machine Learning Research*, 17(1):3058–3091, 2016.

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

Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(3):433–455, 2017b.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with least square regularization. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

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

I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, 2009.

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

V. Temlyakov. Approximation in learning theory. *Constructive Approximation*, 27(1):33–74, 2008.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

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

Q. Wu, Y. Ying, and D.-X. Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

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

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013.

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