COMMUN. MATH. SCI.
Vol. 19, No. 6, pp. 1533–1548

REGULARIZED LEAST SQUARE KERNEL REGRESSION FOR STREAMING DATA*

XIAOQING ZHENG†, HONGWEI SUN‡, AND QIANG WU§

Abstract. We study the use of kernel ridge regression (KRR) in the block-wise streaming data. The algorithm works in an online manner: when a new data block comes in, the algorithm computes a local estimator based on the incoming data block and updates the predictive model by weighted average of all local estimators. Assuming the block data sizes increase at a mild rate and the regularization parameters are selected adaptively according to the sample size of all available data at the time of updating the model, we prove the convergence of the average KRR estimator. The rate is optimal when the regression function can be well approximated by the reproducing kernel Hilbert space in the L^2 sense.

Keywords. Learning theory; kernel ridge regression; streaming data; online learning; adaptive underregularization.

AMS subject classifications. 68T05; 68Q32; 68W27.

1. Introduction

Data collection is usually a dynamic process and data come in as a stream. This process may last for a long period or even look endless in the foreseeable future. In practical applications such as business operations, technology development, or scientific research, people may need to analyze only part of the data before the whole data are available and gradually update the knowledge and models as more data come in.

The streaming data may be received block-wise or instance-wise. The former usually occurs when the data are collected by different entities (e.g. different research groups who work on the same scientific problem) or different branches of the same entity. Instance-wise streaming data seem more common in practice and online learning by gradient descent method has been widely studied in the machine learning literature. However, in some scenarios it may be preferred to analyze the instance-wise streaming data in a block-wise manner. For instance, in the dynamic pricing problems the price is usually not updated each time when an instance of sales information becomes available because customers may not like the price changing too frequently [1, 16]. So, there arises a natural requirement on developing appropriate approaches to analyze block-wise streaming data (either real or pseudo).

In this paper we study the use of kernel ridge regression (KRR) for block-wise streaming data. To describe our approach, let us first recall the settings of regression analysis and the kernel ridge regression for a single data set. Let X be a compact metric space, $Y \subset \mathbb{R}$, and $Z = X \times Y$ be a probability space equipped with a Borel probability distribution ρ which can be decomposed into the conditional probability distribution $\rho(\cdot|x)$ on Y and the marginal probability distribution ρ_X on X, i.e., $\rho(x,y) = \rho(\cdot|x) \times \rho_X(x)$. The mean regression function

*Received: July 14, 2020; Accepted (in revised form): January 27, 2021. Communicated by Lexing Ying.

†School of Mathematical Science, University of Jinan, Jinan 250022, Shangdong, P.R. China (1321222973@qq.com).

‡School of Mathematical Science, University of Jinan, Jinan 250022, Shangdong, P.R. China (ss_sunhw@ujn.edu.cn).

§Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA (qwu@mtsu.edu).

$$f_\rho(x) = \int_Y y\, d\rho(y|x),$$

reveals functional relation between input data x and output data y and can be used for statistical inference and predictive analytics. Let $D = \{(x_i, y_i)\}_{i=1}^N$ be a sample of i.i.d observations collected according to ρ. The purpose of regression analysis is to learn a good estimator for f_ρ. Let $K : X \times X \to \mathbb{R}$ be a continuous, symmetric and positive semi-definite function, called a Mercer kernel. The reproducing kernel Hilbert space H_K associated with the kernel K is the completion of the linear span of the set of functions $\{K_x := K(x,\cdot) : x \in X\}$ with the inner product $\langle \cdot,\cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x,y)$. Given a data set D and an RKHS H_K, the KRR estimator for f_ρ is defined by

$$f_{D,\lambda} = \arg\min_{f \in H_K} \left\{ \frac{1}{|D|} \sum_{(x,y) \in D} (f(x) - y)^2 + \lambda \|f\|_K^2 \right\}, \tag{1.1}$$

where $\lambda > 0$ is a regularization parameter. It has been well studied by a vast literature in learning theory; see e.g. [4,13] and many references therein.

Our approach for regression analysis of block-wise streaming data analysis is as follows. Let $D_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$ be a data block received at time t and composed of n_t observations. Let $f_t = f_{D_t,\lambda_t}$ be the KRR estimator obtained with the data D_t and regularization parameter λ_t. The estimator that will be used for statistical inference and forecasting at time t is defined incrementally by

$$\begin{cases} F_1 = f_1; \\ F_t = \dfrac{N_{t-1}}{N_t} F_{t-1} + \dfrac{n_t}{N_t} f_t, \text{ for } t \geq 2, \end{cases} \tag{1.2}$$

where $N_t = \sum_{s=1}^t n_s$ is the number of all observations available at time t. Note that

$$F_t = \sum_{s=1}^t \frac{n_s}{N_t} f_s, \text{ for all } t \in \mathbb{N}.$$

We call F_t a block-wise streaming data based average kernel ridge regression (BSD-AKRR) estimator. This approach does not need to retrieve the historical data to update the estimator. It is also computationally efficient because there is no need of communications between the incoming data block and historical ones.

Our approach is motivated by recent research on distributed kernel ridge regression (DKRR) [9,18]. Consequently, on the one hand, both approaches share great similarities. On the other hand, they also have some essential differences because of the incremental feature of BSD-AKRR.

Distributed learning was extensively studied in recent years for its high efficiency to handle big data [6–11,18]. The divide and conquer method is one of the distributed learning strategies that does not require mutual communication between the local machines and thus is computationally efficient and privacy protecting. In the context of regression analysis, the DKRR based on the divide and conquer strategy will first partition a big data set D of N observations, which is supposed not processable by a single machine, into m subsets, $D = \bigcup_{t=1}^m D_t$, to different local machines, or these subsets may have already been naturally distributed on different local machines. Then a KRR estimator f_t is learned from D_t and the DKRR estimator is defined by the weighted average $F_D = \sum_{t=1}^m \frac{n_t}{N} f_t$. We see both BSD-AKRR and DKRR are based on the weighted average of local KRR estimators from subsets and do not require mutual communication

between data blocks. Their similarity is clear and seems needless to say. At the first glance, one may even have the illusive intuition that these two approaches are identical.

To see the difference between BSD-AKRR and DKRR, recall that underregularization has been shown essential for DKRR to achieve minimax optimal rates. By the theory in [9, 18], the DKRR estimator F_T is minimax optimal if the regularization parameters λ_t for all subsets D_t are chosen the same as $\lambda_t = N^{-\theta}$ for some $\theta > 0$ depending on the regularity of the regression function and the complexity of the kernel space, that is, the choice of λ_t depends on the size N of the whole data set, not on the size n_t of the subset itself. Comparing with learning with a single data set where the optimality requires $\lambda_t = n_t^{-\theta}$, distributed approach requires a much smaller regularization parameter on the subsets to learn local estimators f_s, which may result in overfitting the data. This parameter selection strategy is called underregularization. It requires knowing the size of the whole data to start the training process. This, however, is impractical for streaming data. At each given time t we have no access to the future data that will be collected after time t and hence do not know the whole data size. Even for the available data blocks, since BSD-AKRR does not retrieve the historical data, we can only have f_s be underregularized according to the size N_s of all available data. All previous learned estimators f_s, $s < t$ have regularization parameters selected according to $N_s < N_t$ and are not fully underregularized. Such an adaptive parameter selection, constrained by the incremental feature of streaming data, makes BSD-AKRR essentially different from DKRR.

The main purpose of this paper is to analyze the performance of BSD-AKRR from a learning theory perspective and derive the optimal strategy for its regularization parameters. The main contributions include three aspects:

(i) We show that BSD-AKRR does not converge if the data blocks are of equal size. This can be illustrated both from a theoretical analysis perspective and by a concrete counterexample.

(ii) Mild growth condition on the data block sizes can guarantee the convergence of BSD-AKRR. The regularization parameters can be selected either locally according to the sample size of the block to be processed or by the underregularization strategy, that is, according to the total sample size of all data blocks available at the time of processing an incoming one, while the latter may give faster convergence rate in case the block size grows fast and the target regression function is well approximated by H_K.

(iii) It is preferred the data block grows at a steady pace for the optimal rates. If the block growth is fluctuating and not controlled, only suboptimal convergence can be obtained.

The rest of this paper is arranged as follows. In Section 2 we describe the assumptions for our analysis and state our convergence results for growing data blocks. Discussions and comparisons with the literature will also be presented. The proofs will be given in Section 3 and Section 4. We close with discussions on the divergence phenomena for equal-sized data blocks in Section 5.

2. Assumptions and main results

In this section, we describe the assumptions and main results for our error analysis of BSD-AKRR. We will adopt the well known source condition and the integral operator technique and perform our error analysis [3, 12].

In most real applications both the input and output values are bounded. In this paper we focus on this situation and make the following assumption.

ASSUMPTION 2.1. *Assume that* $|y| \le M$ *for some constant* $M > 0$ *almost surely.*

This assumption implies $|f_\rho(x)| \le M$ almost surely and finite mean conditional variance, i.e.,

$$\sigma^2 = \mathbf{E}\left[(y - f_\rho(x))^2\right] < M^2.$$

Let $L^2_{\rho_X}$ represent the Hilbert space of square integrable functions with respect to the marginal distribution ρ_X and $L_K \colon L^2_{\rho_X} \to L^2_{\rho_X}$ be the integral operator associated with the Mercer kernel K, defined by

$$L_K f = \int_X K(\cdot, t) f(t)\, d\rho_X(t).$$

Then L_K is a positive compact operator not only from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, but also from H_K to H_K [12, 14]. Moreover, if ρ_X is non-degenerate in the sense that any open set of X has positive measure, then the square root operator $L_K^{1/2}$ is an isomorphism from $(\overline{H_K}\ , \|\cdot\|_{L^2_{\rho_X}})$ to H_K. Therefore, for any $f \in \overline{H_K}$ there holds $L_K^{1/2} f \in H_K$ and $\|L_K^{1/2} f\|_K = \|f\|_{L^2_{\rho_X}}$; for more details see e.g. [14].

The second assumption is the regularity of regression function, as measured by the so-called source condition.

ASSUMPTION 2.2. *For some* $0 < \beta \le 1$, *there holds*

$$f_\rho = L_K^\beta(g_\rho) \quad \textit{with} \quad g_\rho \in L^2_{\rho_X}, \tag{2.1}$$

where L_K^β *denotes the* β*-th power of* L_K *on* $L^2_{\rho_X}$.

Since L_K is a compact and positive operator, let $\{\lambda_s\}_{s=1}^\infty$ be the set of positive eigenvalues of L_K and $\{\phi_s\}_{s \ge 1}$ be the corresponding unit eigenvectors, then L_K^β is defined by

$$L_K^\beta = \sum_{s=1}^\infty \lambda_s^\beta \phi_s \otimes \phi_s.$$

Assumption 2.2 has been widely adopted in the literature of learning theory, especially for the analysis of KRR. Note that $\beta = \frac{1}{2}$ is equivalent to $f_\rho \in H_K$.

For $a, b \in \mathbb{R}$, denote $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, $a_+ = a \vee 0$, define $\frac{a}{0} = +\infty$ if $a > 0$ and the real function

$$\vartheta(t) = \begin{cases} 1 & \text{if } t = -1; \\ 0 & \text{otherwise.} \end{cases}$$

Let $N_t = \sum_{s=1}^t n_s$ be the size of all data available at time t. Our main results are stated in the following theorems.

THEOREM 2.1. *Under Assumption 2.1 and Assumption 2.2, if the sample sizes of the data blocks satisfy* $n_s \ge a_0 s^p$ *for some absolute constants* $a_0 > 0$ *and* $p > 0$, *then by taking* $\lambda_s = n_s^{-\theta}$ *with some* $0 < \theta \le \frac{3}{4}$, *there exists an absolute constant* C *independent of* n_s, λ_s, *or* N_t *such that*

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \le C N_t^{-(1-\theta) \wedge \frac{2\beta\theta p}{1+p}} \left(\log(N_t)\right)^{\vartheta(p(1-2\beta\theta))}. \tag{2.2}$$

Theorem 2.1 guarantees the convergence of the algorithm BSD-AKRR provided that the data blocks are growing and the regularization parameters are suitably selected. Also, note that the condition on the growth of block sizes is very mild. It can be easily fulfilled with $p=1$ even if every data block has one more data point than its precedent block.

In Theorem 2.1 the convergence is stated with regularization parameter λ_s chosen locally according to the sample size n_s of the data block D_s, not the total sample size N_s available at s. In other words, no underregularization has been implemented. In Theorem 2.2 below we consider the convergence of BSD-AKRR with underregularizaiton.

THEOREM 2.2. *Under Assumption 2.1 and Assumption 2.2, if the sample sizes of the data blocks satisfy $n_s \geq a_0 s^p$ for some absolute constants $a_0 > 0$ and $p > 0$, then by taking $\lambda_s = N_s^{-\theta}$ with some $0 < \theta \leq \frac{p}{p+1} \wedge \frac{1}{2\beta}$, there exists a constant C independent of the total sample size N_t at time t such that*

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \leq C N_t^{-(1-\theta)\wedge\zeta_1\wedge\zeta_2} \left(\log(N_t)\right)^{\vartheta(1-p)}, \tag{2.3}$$

where $\zeta_1 = 1 - 2(1-\beta)\theta - \left(\frac{2-p}{2(p+1)} \vee 0\right)$, and $\zeta_2 = 2\beta\theta(1 - \frac{2\beta\theta}{p+1})$.

If $p \geq 2$ and $\beta \geq \frac{1}{2}$, we see $\zeta_1 \geq 1-\theta$ and $\zeta_2 \geq \frac{2\beta\theta p}{p+1}$. The convergence rate in Theorem 2.2 becomes $O(N_t^{-(1-\theta)\wedge\zeta_2})$ and is faster than that given by Theorem 2.1, indicating that underregularization helps improving the learning performance when the sample sizes of the incoming data blocks increase fast and H_K contains f_ρ. At the same time, we notice that if $p < 2$ and $\beta \leq \frac{1}{2}$, since $\zeta_1 < 1-\theta$, the rate in Theorem 2.2 may be slower, indicating underregularization may not always help.

Recall that when $\beta > \frac{1}{2}$ and the effective dimension of H_K satisfies

$$\mathcal{N}(\lambda) = \mathrm{Trace}\left((\lambda I + L_K)^{-1} L_K\right) = O(\lambda^{-\alpha}), \tag{2.4}$$

then KRR can reach minimax optimal rate of $O(|D|^{-\frac{2\beta}{2\beta+\alpha}})$ [2, 3]. Since all kernels satisfy (2.4) with $\alpha = 1$, the rate $O(|D|^{-\frac{2\beta}{2\beta+1}})$ corresponds to the capacity-independent optimal rate. When $\beta < \frac{1}{2}$, the capacity-independent optimal rate was proved by the leave-one-out analysis [17]. Notice that, however, neither Theorem 2.1 nor Theorem 2.2 is able to give the optimal rate $O(N_t^{-\frac{2\beta}{2\beta+1}})$. A plausible explanation is that, when the data blocks do not increase at a steady pace, the magnitude of future data blocks may blow up and result in insufficient underregularization of early data blocks. To avoid this, we place an upper bound on n_s to control the growing speed of the data blocks so that their size increases at a steady pace. Theorem 2.3 below shows that optimal rates become obtainable.

THEOREM 2.3. *Under Assumption 2.1 and Assumption 2.2, if the sample sizes of the data blocks satisfy $a_1 s^p \leq n_s \leq a_2 s^p$ for some absolute constants $0 < a_1 < a_2$ and $p > 0$, then by taking $\lambda_s = N_s^{-\theta}$ with some $0 < \theta \leq \frac{3p}{4(p+1)}$, there exists a constant C independent of the total sample size N_t at time t such that*

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \leq C N_t^{-(1-\theta)\wedge(2\beta\theta)}. \tag{2.5}$$

If in addition $\beta > \frac{1}{6}$ and $p \geq \frac{4}{6\beta-1}$ then by taking $\theta = \frac{1}{2\beta+1}$ we have

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] = O\left(N_t^{-\frac{2\beta}{2\beta+1}}\right). \tag{2.6}$$

We see from (2.5) that, by imposing an upper bound on n_s, the convergence rate of BSD-AKRR is further improved. The estimation (2.6) verifies that the capacity-independent optimal rate is achieved when $\beta > \frac{1}{6}$, with suitable adaptive underregularization choice of the parameters λ_s according to all available sample size N_s up to time s. The facts $\frac{n_t}{N_{t-1}} = O(\frac{1}{t}) \to 0$ and $\frac{n_t}{n_{t-1}} \to 1$ as $t \to \infty$ indicate that, to obtain sharp convergence rates, it is preferred that the data blocks are asymptotically of equal size and none of them dominates.

In all three theorems, the choice for the parameter θ implies that the algorithm fails to converge if the block data sizes are equal ($p = 0$) or diminishes ($p < 0$). This delivers a message to practitioners: if an incoming data block is not sufficiently large, one should wait for more data to form a large data block and update the model because otherwise the learning performance may not improve. It should be emphasized that data block growth requirement for convergence guarantee is not only due to technical difficulty, but is an inherent feature of blockwise data processing. If all blocks are of equal size, we can show the divergence of BSD-AKRR both from a theoretical perspective and by an illustrative counter example. See Section 5 for details.

2.1. Connections and comparisons with distributed learning

Although our result for the block wise streaming data is essentially different from the distributed learning, it is still interesting to make some comparisons between them. In the literature of distributed kernel regression [5,7,9,10,18], assuming the whole data are known and randomly split into multiple subsets and the regularization parameter is selected appropriately according to the sample size of the whole data, the minimax optimal rate of $O(|D|^{-\frac{2\beta}{2\beta+\alpha}})$ was proved for $\beta \geq \frac{1}{2}$ while a suboptimal rate $O(|D|^{-\frac{2\beta}{1+\alpha}})$ was obtained for $\beta < \frac{1}{2}$. Moreover, for the best rates, the number of subsets is restricted to be $O(|D|^{\frac{2\beta-1}{2\beta+\alpha}})$ if $\beta > \frac{1}{2}$ and $O(1)$ if $\beta \leq \frac{1}{2}$. This is equivalent to requesting that the sample size of each subset must be larger than $O(|D|^{\frac{\alpha+1}{2\beta+\alpha}})$ for $\beta > \frac{1}{2}$ and $O(|D|)$ for $\beta \leq \frac{1}{2}$. In other words, if more data becomes available and the total sample size increases, one needs to redistribute the subsets so that each subset contains sufficiently many samples.

In our results for the BSD-AKRR, since the future is not known at each point of time $s < t$, the local estimators f_s are not sufficiently underregularized particularly for small s. This is probably the reason of suboptimal rate for $\beta < \frac{1}{6}$. When $\beta > \frac{1}{6}$, since we consider capacity-independent rate which corresponds to the worse capacity condition $\alpha = 1$, we generally should not expect our rates to be better than the capacity-dependent ones. But we see that if $\alpha > 2\beta$ and $\frac{1}{6} < \beta < \frac{1}{2}$, then $\frac{2\beta}{2\beta+1} > \frac{2\beta}{1+\alpha}$ and hence our rate in Theorem 2.3 is sharper than that obtained in the literature of distributed kernel regression. Furthermore, at each time t we see the number of data blocks is $t = O(N_t^{1/(p+1)})$, which greatly relaxes the constraint on the number of subsets allowed in distributed kernel regression. In BSD-AKRR, when more data become available, the existing data blocks will not change in order to minimize historical data retrieval. Instead, incoming data blocks are required to grow bigger and bigger, making BSD-AKRR essentially different from distributed kernel regression. What is in common for them is that convergence cannot be guaranteed if all data blocks are of certain *fixed* equal size.

As both BSD-AKRR and distributed kernel regression implement block wise data processing, such similarity allowed us to adapt the techniques developed here to distributed kernel regression after this paper was completed. We refined the learning theory analysis of distributed kernel regression and obtained capacity-independent optimal rates under relaxed restrictions. The results are reported in a recent manuscript [15],

which has been posted on arxiv.org.

3. Error bound for local estimators

To analyze the learning performance of BSD-AKRR algorithm (1.2), we recall $F_t = \sum_{s=1}^{t} \frac{n_s}{N_t} f_s$ and write

$$\begin{aligned}
\mathbf{E}\left[\|F_t - f_\rho\|_{L^2_{\rho_X}}^2\right] &= \mathbf{E}\left[\left\|\sum_{s=1}^{t} \frac{n_s}{N_t}(f_s - f_\rho)\right\|_{L^2_{\rho_X}}^2\right] \\
&= \sum_{s=1}^{t} \frac{n_s^2}{N_t^2} \mathbf{E}\left[\|f_s - f_\rho\|_{L^2_{\rho_X}}^2\right] + \sum_{i\neq j} \frac{n_i n_j}{N_t^2} \langle \mathbf{E}f_i - f_\rho, \mathbf{E}f_j - f_\rho\rangle_{L^2_{\rho_X}} \\
&= \sum_{s=1}^{t} \frac{n_s^2}{N_t^2} \left\{\mathbf{E}\left[\|f_s - f_\rho\|_{L^2_{\rho_X}}^2\right] - \|\mathbf{E}f_s - f_\rho\|_{L^2_{\rho_X}}^2\right\} + \|\mathbf{E}F_t - f_\rho\|_{L^2_{\rho_X}}^2 \\
&= \sum_{s=1}^{t} \frac{n_s^2}{N_t^2} \mathbf{E}\left[\|f_s - \mathbf{E}f_s\|_{L^2_{\rho_X}}^2\right] + \left\|\sum_{s=1}^{t} \frac{n_s}{N_t}(\mathbf{E}f_s - f_\rho)\right\|_{L^2_{\rho_X}}^2 . \qquad (3.1)
\end{aligned}$$

Note that

$$\left\|\sum_{s=1}^{t} \frac{n_s}{N_t}(\mathbf{E}f_s - f_\rho)\right\|_{L^2_{\rho_X}}^2 \leq \left\{\sum_{s=1}^{t} \frac{n_s}{N_t}\|\mathbf{E}f_s - f_\rho\|_{L^2_{\rho_X}}\right\}^2 \leq \sum_{s=1}^{t} \frac{n_s}{N_t}\|\mathbf{E}f_s - f_\rho\|_{L^2_{\rho_X}}^2 .$$

Therefore we have

$$\mathbf{E}\left[\|F_t - f_\rho\|_{L^2_{\rho_X}}^2\right] \leq \sum_{s=1}^{t} \frac{n_s^2}{N_t^2} \mathbf{E}\left[\|f_s - \mathbf{E}f_s\|_{L^2_{\rho_X}}^2\right] + \sum_{s=1}^{t} \frac{n_s}{N_t}\|\mathbf{E}f_s - f_\rho\|_{L^2_{\rho_X}}^2 . \qquad (3.2)$$

This tells us that to bound the error of F_t, the key is to bound the variance and bias of all local estimators. In the sequel of this section we will focus on the estimation of $\mathbf{E}\left[\|f_s - \mathbf{E}f_s\|_{L^2_{\rho_X}}^2\right]$ and $\|\mathbf{E}f_s - f_\rho\|_{L^2_{\rho_X}}^2$.

For the sample subset $D_s = \{z_{s,i} = (x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$, the sampling operator $S_{D_s}: H_K \to \mathbb{R}^{n_s}$ is defined by

$$S_{D_s} f := (f(x_{s,i}))_{i=1}^{n_s} \quad \text{for } f \in H_K.$$

Its adjoint operator $S_{D_s}^*: \mathbb{R}^{n_s} \to H_K$ is

$$S_{D_s}^* \mathbf{c} := \frac{1}{n_s} \sum_{i=1}^{n_s} c_i K_{x_{s,i}} \quad \text{for } \mathbf{c} = (c_1, \cdots, c_{n_s}) \in \mathbb{R}^{n_s}.$$

It is proved in [12, 14] that

$$f_s = (\lambda_s I + S_{D_s}^* S_{D_s})^{-1} S_{D_s}^* \mathbf{y}_s, \qquad (3.3)$$

where $\mathbf{y}_s = (y_{s,1} \ldots, y_{s,n_s}) \in \mathbb{R}^{n_s}$ is the vector of response values on the subset D_s. The sample limit version of f_s associated with the regularization parameter λ_s, defined by $f_{\lambda_s} = (\lambda_s I + L_K)^{-1} L_K f_\rho$, plays an essential role for the error analysis. In the sequel, without loss of generality we assume $0 < \lambda_s \leq 1$, and denote that

$$\kappa \doteq \sup_{x\in X} \sqrt{K(x,x)} < \infty.$$

Under the assumption (2.1) we have

$$\|f_\rho - f_{\lambda_s}\|_{L^2_{\rho_X}} = \|\lambda_s L_K^\beta (\lambda_s I + L_K)^{-1} g_\rho\|_{L^2_{\rho_X}} \le \lambda_s^\beta \|g_\rho\|_{L^2_{\rho_X}}. \tag{3.4}$$

Let $\eta_s(z) = (f_\rho(x) - f_{\lambda_s}(x))K_x$ and

$$\Delta_s = \frac{1}{n_s} \sum_{z \in D_s} (f_\rho(x) - f_{\lambda_s}(x))K_x - L_K(f_\rho - f_{\lambda_s})$$

be the deviation of the sample mean of η_s on the subset D_s from its expectation. Then

$$\mathbf{E} f_s - f_{\lambda_s} = \mathbf{E}\left[\left(S_{D_s}^* S_{D_s} + \lambda_s I\right)^{-1} \Delta_s\right]. \tag{3.5}$$

To bound the variance and bias of f_s, we need the following two lemmas.

LEMMA 3.1. *Assume $|y| \le M$ almost surely, $f_\rho = L_K^\beta g_\rho$ for some $0 < \beta \le 1$ and $g_\rho \in L^2_{\rho_X}$. We have*

$$\mathbf{E}\left[\|\Delta_s\|_K^2\right] = \mathbf{E}\left[\left\|\frac{1}{n_s}\sum_{z\in D_s}\eta_s(z) - \mathbf{E}\eta_s\right\|_K^2\right] \le \kappa^2 \|g_\rho\|^2_{L^2_{\rho_X}} \lambda_s^{2\beta} n_s^{-1}.$$

LEMMA 3.2. *We have* $\mathbf{E}\left[\|L_K - S_{D_s}^* S_{D_s}\|^2\right] \le \frac{\kappa^4}{n_s}$.

The proofs of Lemma 3.1 and Lemma 3.2 follow from standard calculations. We omit the details.

In Proposition 3.2 and (3.12) below we state our bound for the variance of f_s. We remark that a similar bound had been proved for the case $0 < \beta \le \frac{1}{2}$ by the leave-one-out analysis in [17]. But in our proof, the upper bound of variance holds for all $\beta > 0$, and it seems that the variance has no relation with the source condition (2.2).

Consider the leave-one-out estimate. Plugging one more i.i.d. observation $z_{s,n_s+1} = (x_{s,n_s+1}, y_{s,n_s+1})$ into the sample set D_s, for any $1 \le i \le n_s + 1$, let

$$f_{s^{\setminus i}} = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{n_s}\sum_{j\ne i, j=1}^{n_s+1}\left(f(x_{s,j}) - y_{s,j}\right)^2 + \lambda_s\|f\|_K^2\right\},$$

$$g_s = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{n_s}\sum_{j=1}^{n_s+1}\left(f(x_{s,j}) - y_{s,j}\right)^2 + \lambda_s\|f\|_K^2\right\}.$$

The following lemma can been proved by the similar method proposed in [17].

LEMMA 3.3. *For all $1 \le i \le n_s + 1$,*

$$\|g_s - f_{s^{\setminus i}}\|_K \le \frac{\kappa}{n_s\lambda_s}|g_s(x_{s,i}) - y_{s,i}|.$$

PROPOSITION 3.1. *Assume $|y| \le M$ almost surely. There holds*

$$\mathbf{E}\left[\|f_s - f_\rho\|^2_{L^2_{\rho_X}}\right] \le \|f_{\lambda_s} - f_\rho\|^2_{L^2_{\rho_X}} + \frac{n_s\lambda_s}{n_s+1}\left(\|f_{\lambda_s}\|_K^2 - \mathbf{E}\|g_s\|_K^2\right) + M^2\left(\frac{\kappa^4}{n_s^2\lambda_s^2} + \frac{2\kappa^2}{n_s\lambda_s}\right).$$

Proof. Recall $\sigma^2 = \mathbf{E}(f_\rho(x)-y)^2$ and $\mathbf{E}(f(x)-y)^2 = \mathbf{E}(f_\rho(x)-f(x))^2+\sigma^2$.

$$\begin{aligned}
\mathbf{E}\left[\|f_s - f_\rho\|^2_{L^2_{\rho_X}}\right] &= \mathbf{E}\left[\left(f_{s\setminus n_s+1}(x_{s,n_s+1})-y_{s,n_s+1}\right)^2\right]-\sigma^2\\
&= \mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(f_{s\setminus i}(x_{s,i})-y_{s,i}\right)^2\right]-\sigma^2\\
&= \mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(f_{s\setminus i}(x_{s,i})-y_{s,i}\right)^2-\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]\\
&\quad+\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]-\sigma^2. \qquad (3.6)
\end{aligned}$$

The second term of (3.6) is bounded by

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{n_s+1}\sum_{j=1}^{n_s+1}\left(g_s(x_{s,j})-y_{s,j}\right)^2\right]\\
&=\frac{n_s}{n_s+1}\mathbf{E}\left[\frac{1}{n_s}\sum_{j=1}^{n_s+1}\left(g_s(x_{s,j})-y_{s,j}\right)^2+\lambda_s\|g_s\|^2_K\right]-\frac{n_s\lambda_s}{n_s+1}\mathbf{E}\|g_s\|^2_K\\
&\le\frac{n_s}{n_s+1}\mathbf{E}\left[\frac{1}{n_s}\sum_{j=1}^{n_s+1}\left(f_{\lambda_s}(x_{s,j})-y_{s,j}\right)^2+\lambda_s\|f_{\lambda_s}\|^2_K\right]-\frac{n_s\lambda_s}{n_s+1}\mathbf{E}\|g_s\|^2_K\\
&=\|f_{\lambda_s}-f_\rho\|^2_{L^2_{\rho_X}}+\frac{n_s\lambda_s}{n_s+1}\left(\|f_{\lambda_s}\|^2_K-\mathbf{E}\|g_s\|^2_K\right)+\sigma^2. \qquad (3.7)
\end{aligned}$$

By the boundedness of the output data, we have the following simpler bound,

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{n_s+1}\sum_{j=1}^{n_s+1}\left(g_s(x_{s,j})-y_{s,j}\right)^2\right]\\
&\le\frac{n_s}{n_s+1}\mathbf{E}\left[\frac{1}{n_s}\sum_{j=1}^{n_s+1}\left(g_s(x_{s,j})-y_{s,j}\right)^2+\lambda_s\|g_s\|^2_K\right]\le M^2. \qquad (3.8)
\end{aligned}$$

By Lemma 3.3, the first part of (3.6) can be estimated as

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(f_{s\setminus i}(x_{s,i})-g_s(x_{s,i})\right)\left(f_{s\setminus i}(x_{s,i})+g_s(x_{s,i})-2y_{s,i}\right)\right]\\
&=\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(f_{s\setminus i}(x_{s,i})-g_s(x_{s,i})\right)^2\right]\\
&\quad+\mathbf{E}\left[\frac{2}{n_s+1}\sum_{i=1}^{n_s+1}\left(f_{s\setminus i}(x_{s,i})-g_s(x_{s,i})\right)\left(g_s(x_{s,i})-y_{s,i}\right)\right]\\
&\le\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\frac{\kappa^4}{n_s^2\lambda_s^2}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]+\mathbf{E}\left[\frac{2}{n_s+1}\sum_{i=1}^{n_s+1}\frac{\kappa^2}{n_s\lambda_s}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]\\
&=\left(\frac{\kappa^4}{n_s^2\lambda_s^2}+\frac{2\kappa^2}{n_s\lambda_s}\right)\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]
\end{aligned}$$

$$\leq M^2\left(\frac{\kappa^4}{n_s^2\lambda_s^2}+\frac{2\kappa^2}{n_s\lambda_s}\right). \tag{3.9}$$

Plugging estimates in (3.7) and (3.9) into (3.6), Proposition 3.1 is proved. ∎

PROPOSITION 3.2. *Under the condition that $|y|\leq M$ almost surely and $0<\lambda_s\leq 1$, there holds*

$$\mathbf{E}\left[\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\right]+\lambda_s\mathbf{E}\left[\|f_s-f_{\lambda_s}\|_K^2\right]\leq M^2(\kappa^2+2)^2\Big(\frac{1}{n_s\lambda_s}+\frac{1}{n_s^2\lambda_s^2}\Big).$$

Proof. By Proposition 3.1, there holds

$$\begin{aligned}
\mathbf{E}\left[\|f_s-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\right]&=\mathbf{E}\left[\|f_s-f_\rho\|_{L^2_{\rho_X}}^2\right]+2\mathbf{E}\left[\langle f_s-f_{\lambda_s},f_\rho-f_{\lambda_s}\rangle_{L^2_{\rho_X}}\right]-\|f_\rho-f_{\lambda_s}\|_{L^2_{\rho_X}}^2\\
&\leq M^2\left(\frac{\kappa^4}{n_s^2\lambda_s^2}+\frac{2\kappa^2}{n_s\lambda_s}\right)+\frac{n_s\lambda_s}{n_s+1}\left(\|f_{\lambda_s}\|_K^2-\mathbf{E}\left[\|g_s\|_K^2\right]\right)\\
&\quad+2\mathbf{E}[\langle f_s-f_{\lambda_s},L_K(f_\rho-f_{\lambda_s})\rangle_K]\\
&\leq M^2\left(\frac{\kappa^4}{n_s^2\lambda_s^2}+\frac{2\kappa^2}{n_s\lambda_s}\right)+\lambda_s\left(\|f_{\lambda_s}\|_K^2-\mathbf{E}\left[\|f_s\|_K^2\right]\right)+\lambda_s\mathbf{E}\left[\|f_s\|_K^2\right]\\
&\quad+2\lambda_s\mathbf{E}[\langle f_s-f_{\lambda_s},f_{\lambda_s}\rangle_K]-\frac{n_s\lambda_s}{n_s+1}\mathbf{E}\left[\|g_s\|_K^2\right]-\frac{\lambda_s}{n_s+1}\|f_{\lambda_s}\|_K^2.\\
&\leq M^2\left(\frac{\kappa^4}{n_s^2\lambda_s^2}+\frac{2\kappa^2}{n_s\lambda_s}\right)-\lambda_s\mathbf{E}\left[\|f_s-f_{\lambda_s}\|_K^2\right]+\Lambda,
\end{aligned}\tag{3.10}$$

where we used the fact

$$\lambda_s\left(\|f_{\lambda_s}\|_K^2-\mathbf{E}\left[\|f_s\|_K^2\right]\right)+2\lambda_s\mathbf{E}[\langle f_s-f_{\lambda_s},f_{\lambda_s}\rangle_K]=-\lambda_s\mathbf{E}\left[\|f_s-f_{\lambda_s}\|_K^2\right]$$

and

$$\Lambda=\lambda_s\mathbf{E}\left[\|f_s\|_K^2\right]-\frac{n_s\lambda_s}{n_s+1}\mathbf{E}\left[\|g_s\|_K^2\right].$$

We know from the definition of g_s that $\lambda_s\|g_s\|_K^2\leq M^2$ and hence

$$\begin{aligned}
\Lambda&=\lambda_s\mathbf{E}\left\{\|f_s-g_s\|_K^2+2\langle f_s-g_s,g_s\rangle_K+\frac{1}{n_s+1}\|g_s\|_K^2\right\}\\
&\leq\lambda_s\mathbf{E}\left[\|f_s-g_s\|_K^2\right]+2M\left[\lambda_s\mathbf{E}\|f_s-g_s\|_K^2\right]^{\frac{1}{2}}+\frac{M^2}{n_s+1}.
\end{aligned}\tag{3.11}$$

By Lemma 3.3 and (3.8) we have

$$\begin{aligned}
\mathbf{E}\left[\|f_s-g_s\|_K^2\right]&=\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\|f_{s\setminus i}-g_s\|_K^2\right]\\
&\leq\frac{\kappa^2}{n_s^2\lambda_s^2}\mathbf{E}\left[\frac{1}{n_s+1}\sum_{i=1}^{n_s+1}\left(g_s(x_{s,i})-y_{s,i}\right)^2\right]\leq\frac{M^2\kappa^2}{n_s^2\lambda_s^2}.
\end{aligned}$$

Plugging this bound into (3.11) and by (3.10), we prove the desired estimation. ∎

Note that our bound in Proposition 3.2 is independent of the source condition in Assumption 2.2. By the fact that $\mathbf{E}\left[\|\zeta-\mathbf{E}\zeta\|^2\right]\leq\mathbf{E}\left[\|\zeta-h\|^2\right]$ for all vector-valued random variables ζ and any vector h, we have

$$\mathbf{E}\left[\|f_s-\mathbf{E}f_s\|_{L^2_{\rho_X}}^2\right]\leq M^2(\kappa^2+2)^2\left\{\lambda_s^{-2}n_s^{-2}+\lambda_s^{-1}n_s^{-1}\right\}. \tag{3.12}$$

Now we turn to estimating the bias $\|\mathbf{E}f_s - f_\rho\|^2_{L^2_{\rho_X}}$.

PROPOSITION 3.3. *Assume $|y| \le M$ almost surely, $f_\rho = L_K^\beta g_\rho$ for some $0 < \beta \le 1$ and $g_\rho \in L^2_{\rho_X}$. There holds*

$$\|\mathbf{E}f_s - f_\rho\|^2_{L^2_{\rho_X}} \le (20\kappa^4 + 2)\|g_\rho\|^2_{L^2_{\rho_X}} \left\{ \lambda_s^{2\beta-2} n_s^{-\frac{3}{2}} + \lambda_s^{2\beta} \right\}.$$

Proof. First notice that

$$\|\mathbf{E}f_s - f_\rho\|^2_{L^2_{\rho_X}} \le 2\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}} + 2\|f_{\lambda_s} - f_\rho\|^2_{L^2_{\rho_X}}.$$

To estimate the first term on the right-hand side, by the formula (3.5), we have

$$\begin{aligned}
\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}} &= \left\| \mathbf{E}\left[L_K^{\frac{1}{2}} \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} \Delta_s \right] \right\|^2_K \\
&\le 2 \left\| \mathbf{E}\left[\left(L_K^{\frac{1}{2}} - (S^*_{D_s} S_{D_s})^{\frac{1}{2}} \right) \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} \Delta_s \right] \right\|^2_K \\
&\quad + 2 \left\| \mathbf{E}\left[(S^*_{D_s} S_{D_s})^{\frac{1}{2}} \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} \Delta_s \right] \right\|^2_K \\
&= 2 \left\| \mathbf{E}\left[\left(L_K^{\frac{1}{2}} - (S^*_{D_s} S_{D_s})^{\frac{1}{2}} \right) \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} \Delta_s \right] \right\|^2_K \\
&\quad + 2 \left\| \mathbf{E}\left[\left\{ (S^*_{D_s} S_{D_s})^{\frac{1}{2}} \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} - L_K^{\frac{1}{2}} (L_K + \lambda_s I)^{-1} \right\} \Delta_s \right] \right\|^2_K,
\end{aligned}$$

where we have used the fact that $\mathbf{E}[\Delta_s] = 0$. By the operator monotone inequality $\|A^k - B^k\| \le \|A - B\|^k$ for any $0 < k \le 1$ and two positive operators A, B, and Hölder inequality, we obtain

$$\begin{aligned}
\|\mathbf{E}f_s - f_{\lambda_s}\|^2_{L^2_{\rho_X}} \le{}& 2\lambda_s^{-2} \mathbf{E}\|\Delta_s\|^2_K \times \left(\mathbf{E}\left\| L_K - S^*_{D_s} S_{D_s} \right\|^2 \right)^{\frac{1}{2}} + 2\mathbf{E}\|\Delta_s\|^2_K \\
&\times \mathbf{E}\left\| (S^*_{D_s} S_{D_s})^{\frac{1}{2}} \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} - L_K^{\frac{1}{2}} (L_K + \lambda_s I)^{-1} \right\|^2. \quad (3.13)
\end{aligned}$$

For any two positive operators A and B defined on a Hilbert Space, and $\lambda > 0$, we have

$$\begin{aligned}
&A^{\frac{1}{2}} (A + \lambda I)^{-1} - B^{\frac{1}{2}} (B + \lambda I)^{-1} \\
&= (A + \lambda I)^{-1} \left(A^{\frac{1}{2}} (B + \lambda I) - (A + \lambda I) B^{\frac{1}{2}} \right) (B + \lambda I)^{-1} \\
&= (A + \lambda I)^{-1} \left(\lambda \left(A^{\frac{1}{2}} - B^{\frac{1}{2}} \right) + A^{\frac{1}{2}} \left(B^{\frac{1}{2}} - A^{\frac{1}{2}} \right) B^{\frac{1}{2}} \right) (B + \lambda I)^{-1}.
\end{aligned}$$

Applying it with $A = S^*_{D_s} S_{D_s}$, $B = L_K$, we obtain

$$\mathbf{E}\left\| (S^*_{D_s} S_{D_s})^{\frac{1}{2}} \left(S^*_{D_s} S_{D_s} + \lambda_s I \right)^{-1} - L_K^{\frac{1}{2}} (L_K + \lambda_s I)^{-1} \right\|^2 \le 4\lambda_s^{-2} \left(\mathbf{E}\left\| L_K - S^*_{D_s} S_{D_s} \right\|^2 \right)^{\frac{1}{2}}.$$

Plugging the estimations in Lemma 3.1 and Lemma 3.2 into (3.13) and combining it with (3.4) lead to the desired upper bound for the bias. This proves Proposition 3.3. □

4. Learning rates of the average estimator

In this section we prove the three theorems stated in Section 2. To simplify our notations, we use $a \preceq b$ to denote the relation $a \le Cb$ with a constant C independent of n_s, λ_s, or N_s for all $1 \le s \le t$. The notation $a \sim b$ means $a \preceq b$ and $b \preceq a$ hold simultaneously. We will need the fact that

$$\sum_{s=1}^{t} s^a \sim t^{a+1} \qquad \text{for} \qquad a > -1.$$

Proof. **(Proof of Theorem 2.1.)** Under the condition that $n_s \ge a_0 s^p$ for all $1 \le s \le t$, then

$$N_t = \sum_{s=1}^{t} n_s \ge a_0 \sum_{s=1}^{t} s^p \ge \frac{a_0}{1+p} t^{1+p}. \tag{4.1}$$

By the error decomposition (3.2) and $\lambda_s = n_s^{-\theta}$ with $0 < \theta \le \frac{3}{4}$ we have

$$\begin{aligned} \mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] &\preceq \frac{1}{N_t^2} \sum_{s=1}^{t} \left(n_s^{2\theta} + n_s^{1+\theta}\right) + \frac{1}{N_t} \sum_{s=1}^{t} \left(n_s^{-\frac{1}{2}+2(1-\beta)\theta} + n_s^{1-2\beta\theta}\right) \\ &\preceq \frac{1}{N_t^2} \sum_{s=1}^{t} n_s^{1+\theta} + \frac{1}{N_t} \sum_{s=1}^{t} n_s^{1-2\beta\theta}. \end{aligned} \tag{4.2}$$

By the fact $n_s < N_s \le N_t$ for all $1 \le s \le t$, the first term on the right can be bounded by

$$\frac{1}{N_t^2} \sum_{s=1}^{t} n_s^{1+\theta} \le \frac{1}{N_t^{2-\theta}} \sum_{s=1}^{t} n_s = \frac{1}{N_t^{1-\theta}}. \tag{4.3}$$

Note that for any $0 < \alpha \le 1$, by Hölder inequality and (4.1), we have

$$\sum_{s=1}^{t} n_s^\alpha \le \left[\sum_{s=1}^{t} \left(n_s^\alpha\right)^{\frac{1}{\alpha}}\right]^\alpha \times t^{1-\alpha} \preceq N_t^{\frac{1+\alpha p}{1+p}}$$

and, for any $\alpha \le 0$, we have

$$\sum_{s=1}^{t} n_s^\alpha \preceq \sum_{s=1}^{t} s^{p\alpha} \preceq t^{(1+\alpha p)\vee 0} (\log t)^{\vartheta(p\alpha)} \preceq N_t^{\frac{1+\alpha p}{1+p}\vee 0} (\log(N_t))^{\vartheta(p\alpha)}.$$

Under the condition $0 < \theta \le \frac{3}{4}$, we have $-\frac{1}{2} \le 1 - 2\beta\theta < 1$. Therefore, we can bound the second term on the right of (4.2) by

$$\frac{1}{N_t} \sum_{s=1}^{t} n_s^{1-2\beta\theta} \preceq N_t^{-\left(\frac{2\beta\theta p}{p+1}\wedge 1\right)} (\log(N_t))^{\vartheta(p(1-2\beta\theta))}. \tag{4.4}$$

Combining the estimations in (4.3) and (4.4) we have

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \preceq N_t^{-(1-\theta)\wedge\frac{2\beta\theta p}{1+p}} (\log(N_t))^{\vartheta(p(1-2\beta\theta))}.$$

This proves Theorem 2.1. ∎

Proof. **(Proof of Theorem 2.2.)** By the error decomposition (3.2) and $\lambda_s = N_s^{-\theta}$ we have

$$\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \preceq \frac{1}{N_t^2}\sum_{s=1}^{t}\left(N_s^{2\theta} + n_s N_s^{\theta}\right) + \frac{1}{N_t}\sum_{s=1}^{t}\left(n_s^{-\frac{1}{2}} N_s^{2(1-\beta)\theta} + n_s N_s^{-2\beta\theta}\right). \quad (4.5)$$

It is easy to verify that, if $0 < \theta \le \frac{p}{p+1}$, then

$$\frac{1}{N_t^2}\sum_{s=1}^{t} N_s^{2\theta} \le \frac{t}{N_t^{2-2\theta}} \preceq \frac{1}{N_t^{2-2\theta-\frac{1}{p+1}}} \preceq \frac{1}{N_t^{1-\theta}} \quad (4.6)$$

and

$$\frac{1}{N_t^2}\sum_{s=1}^{t} n_s N_s^{\theta} \le \frac{1}{N_t^{2-\theta}}\sum_{s=1}^{t} n_s = \frac{1}{N_t^{1-\theta}}. \quad (4.7)$$

By $n_s \ge a_0 s^p$,

$$\begin{aligned}\frac{1}{N_t}\sum_{s=1}^{t} n_s^{-\frac{1}{2}} N_s^{2(1-\beta)\theta} &\preceq \frac{1}{N_t^{1-2(1-\beta)\theta}}\sum_{s=1}^{t} s^{-\frac{p}{2}} \\ &\preceq \begin{cases} N_t^{-1+2(1-\beta)\theta} t^{1-\frac{p}{2}} \preceq N_t^{-1+2(1-\beta)\theta+\frac{2-p}{2(p+1)}}, & \text{if } p<2; \\ N_t^{-1+2(1-\beta)\theta}\log t \preceq N_t^{-1+2(1-\beta)\theta}\log N_t, & \text{if } p=2; \\ N_t^{-1+2(1-\beta)\theta}, & \text{if } p>2. \end{cases} \end{aligned} \quad (4.8)$$

When $0 < \theta \le \frac{1}{2\beta}$. Denote $\nu = 1 - \frac{2\beta\theta}{p+1}$, $s^* = \max\{1 \le s \le t : N_s \le N_t^{\nu}\}$. Then

$$\begin{aligned}\frac{1}{N_t}\sum_{s=1}^{t} n_s N_s^{-2\beta\theta} &\preceq \frac{1}{N_t}\sum_{s=1}^{s^*} n_s^{1-2\beta\theta} + \frac{1}{N_t^{1+2\beta\theta\nu}}\sum_{s=s^*+1}^{t} n_s \\ &\preceq \frac{1}{N_t}\left(\sum_{s=1}^{s^*} n_s\right)^{1-2\beta\theta} t^{2\beta\theta} + N_t^{-2\beta\theta\nu} \\ &\preceq N_t^{-1+(1-2\beta\theta)\nu+\frac{2\beta\theta}{p+1}} + N_t^{-2\beta\theta\nu} \\ &\preceq N_t^{-2\beta\theta\nu}. \end{aligned} \quad (4.9)$$

Combining the estimations in (4.6)-(4.9), we obtain the desired conclusion. ∎

Proof. **(Proof of Theorem 2.3.)** If $n_s \sim s^p$ for all $1 \le s \le t$, then $N_s \sim s^{p+1}$. Recalling the error decomposition (3.2) and the choice of $\lambda_s = N_s^{-\theta} \sim s^{-\theta(p+1)}$ we have

$$\begin{aligned}&\mathbf{E}\left[\|F_t - f_\rho\|^2_{L^2_{\rho_X}}\right] \\ &\preceq \sum_{s=1}^{t}\left(\frac{s^{2\theta(p+1)}}{t^{2(p+1)}} + \frac{s^{p+\theta(p+1)}}{t^{2(p+1)}}\right) + \sum_{s=1}^{t}\left(\frac{s^{2(1-\beta)\theta(p+1)-p/2}}{t^{(p+1)}} + \frac{s^{p-2\beta\theta(p+1)}}{t^{p+1}}\right) \\ &\preceq t^{1-2(1-\theta)(p+1)} + t^{-(1-\theta)(p+1)}\end{aligned}$$

$$+\sum_{s=1}^{t}\left(\frac{s^{\max\{2(1-\beta)\theta(p+1)-p/2,\theta(1+p)-1\}}}{t^{(p+1)}}+\frac{s^{\max\{p-2\beta\theta(p+1),\theta(1+p)-1\}}}{t^{p+1}}\right)$$

$$\preceq t^{1-2(1-\theta)(p+1)}+t^{-(1-\theta)(p+1)}+t^{2(1-\beta)\theta(p+1)-\frac{3p}{2}}+t^{-2\beta\theta(p+1)}.$$

If $\theta\leq\frac{3p}{4(p+1)}$, then

$$1-2(1-\theta)(p+1)\leq-(1-\theta)(p+1),$$
$$2(1-\beta)\theta(p+1)-3p/2\leq-2\beta\theta(p+1).$$

Therefore,

$$\mathbf{E}\left[\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\right]\preceq t^{-(1-\theta)(p+1)}+t^{-2\beta\theta(p+1)}.$$

The conclusion (2.5) follows by noting that $N_t\sim t^{p+1}$. Under the assumption $\beta>\frac{1}{6}$ and $p\geq\frac{4}{6\beta-1}$, the conclusion (2.6) is an easy consequence of the choice $\theta=\frac{1}{2\beta+1}$. ∎

5. BSD-AKRR is divergent for equal-sized data blocks

When all data blocks have the equal sample size, namely, $n_s=n$ is fixed for all $1\leq s\leq t$ and hence $N_s=ns$. Recalling the error decomposition (3.2) and the choice of $\lambda_s=N_s^{-\theta}$ with $0<\theta<\frac{1}{2\beta}$ we have

$$\mathbf{E}\left[\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\right]\preceq\frac{1}{t^2}\sum_{s=1}^{t}\left(\frac{(ns)^{2\theta}}{n^2}+\frac{(ns)^{\theta}}{n}\right)+\frac{1}{t}\sum_{s=1}^{t}\left(\frac{(ns)^{(2-2\beta)\theta}}{n^{3/2}}+(ns)^{-2\beta\theta}\right)$$
$$\preceq\frac{t}{(nt)^{2-2\theta}}+\frac{1}{(nt)^{1-\theta}}+\frac{t^{3/2}}{(nt)^{2\beta\theta-2\theta+3/2}}+(nt)^{-2\beta\theta}. \tag{5.1}$$

When $t\to\infty$ we see the third term on the right-hand side of (5.1) diverges for any $0<\beta<1$. The divergence can be further illustrated as follows. Recalling the error decomposition (3.1) implies

$$\mathbf{E}\left[\|F_t-f_\rho\|^2_{L^2_{\rho_X}}\right]\geq\left\|\sum_{s=1}^{t}\frac{n_s}{N_t}(\mathbf{E}f_s-f_\rho)\right\|^2_{L^2_{\rho_X}}.$$

It is reasonable to assume $\mathbf{E}f_s$ are highly correlated and hence

$$\left\|\sum_{s=1}^{t}\frac{n_s}{N_t}(\mathbf{E}f_s-f_\rho)\right\|^2_{L^2_{\rho_X}}\approx\sum_{s=1}^{t}\frac{n_s}{N_t}\|\mathbf{E}f_s-f_\rho\|^2_{L^2_{\rho_X}}.$$

Denote $g(\lambda)=\lambda^{2\beta-2}n^{-\frac{3}{2}}+\lambda^{2\beta}$. It is easy to verify that the function g reaches minimum value at $\lambda^*=\sqrt{n^{-3/2}(1-\beta)/\beta}$ and therefore

$$g(\lambda)\geq g(\lambda^*)=n^{-\frac{3\beta}{2}}\left\{\left(\tfrac{1-\beta}{\beta}\right)^{\beta-1}+\left(\tfrac{1-\beta}{\beta}\right)^{\beta}\right\}>0 \tag{5.2}$$

regardless of the value of λ. In our analysis we have used Proposition 3.3 to technically bound the bias by

$$\sum_{s=1}^{t}\frac{n_s}{N_t}\|\mathbf{E}f_s-f_\rho\|^2_{L^2_{\rho_X}}\preceq\sum_{s=1}^{t}g(\lambda_s).$$

By (5.2), this bound never vanishes.

One may argue that the above discussion is based on upper bound analysis which may not be sharp. We further illustrate the divergence of BSD-AKRR by a counterexample. Consider Gaussian kernel $K(x,x')=\exp(\|x-x'\|_2^2)$ and the extreme case $n_s=1$. Then for each s, $D_s=\{(x_s,y_s)\}$ contains only one data point and the kernel matrix on the D_s is of dimension 1×1 and has entry value 1. By the representer theorem,

$$f_s(x)=c_sK(x_s,x)=\frac{y_s}{\lambda_s+1}K(x_s,x).$$

and hence

$$\mathbf{E}f_s=\frac{1}{\lambda_s+1}L_Kf_\rho.$$

If λ_s is selected according to $N_s=s$, then $\lambda_s\to 0$ as $s\to\infty$ and therefore

$$\left\|\sum_{s=1}^{t}\frac{n_s}{N_t}\mathbf{E}f_s-f_\rho\right\|_{L^2_{\rho_X}}^2=\left\|\left(\frac{1}{t}\sum_{s=1}^{t}\frac{1}{\lambda_s+1}\right)L_Kf_\rho-f_\rho\right\|_{L^2_{\rho_X}}^2\longrightarrow\|L_Kf_\rho-f_\rho\|_{L^2_{\rho_X}}^2,$$

which clearly does not vanish unless L_K has an eigenvalue 1 and f_ρ happens to be an associated eigenfunction. If λ_s is selected according to $n_s=1$ so that $\lambda_s=\lambda_1$ for all $s\geq 1$, then

$$\left\|\sum_{s=1}^{t}\frac{n_s}{N_t}\mathbf{E}f_s-f_\rho\right\|_{L^2_{\rho_X}}^2=\left\|\frac{1}{\lambda_1+1}L_Kf_\rho-f_\rho\right\|_{L^2_{\rho_X}}^2,$$

which does not vanish either.

Acknowledgments. The work by Hongwei Sun described in this paper is supported by National Natural Science Foundation of China (Grants No. 11671171 and 11871167). The work by Qiang Wu is partially supported by the Simons Foundation Collaboration Grant (Award ID 712916). The three authors made equal contributions to the paper.

REFERENCES

[1] S. Agrawal, Z. Wang, and Y. Ye, *A dynamic near-optimal algorithm for online linear programming*, Oper. Res., 62(4):876–890, 2014. 1

[2] G. Blanchard and N. Mücke, *Optimal rates for regularization of statistical inverse learning problems*, Found. Comput. Math., 18(4):971–1013, 2018. 2

[3] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math., 7(3):331–368, 2007. 2, 2

[4] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007. 1

[5] X. Guo, T. Hu, and Q. Wu, *Distributed minimum error entropy algorithms*, J. Mach. Learn. Res., 21(126):1–31, 2020. 2.1

[6] Z.C. Guo, S.B. Lin, and D.X. Zhou, *Learning theory of distributed spectral algorithms*, Inverse Probl., 33(7):074009, 2017. 1

[7] Z.C. Guo, L. Shi, and Q. Wu, *Learning theory of distributed regression with bias corrected regularization kernel network*, J. Mach. Learn. Res., 18(118):1–25, 2017. 1, 2.1

[8] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou, *Thresholded spectral algorithms for sparse approximations*, Anal. Appl., 15(3):433–455, 2017. 1

[9] S.B. Lin, X. Guo, and D.X. Zhou, *Distributed learning with regularized least squares*, J. Mach. Learn. Res., 18(92):1–31, 2017. 1, 2.1

[10] S.B. Lin and D.X. Zhou, *Distributed kernel gradient descent algorithms*, Constr. Approx., 47:249–276, 2018. 1, 2.1
[11] L. Shi, *Distributed learning with indefinite kernels*, Anal. Appl., 17(06):947–975, 2019. 1
[12] S. Smale and D.-X. Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx., 26:153–172, 2007. 2, 2, 3
[13] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008. 1
[14] H. Sun and Q. Wu, *Application of integral operator for regularized least-square regression*, Math. Comput. Model., 49(1-2):276–285, 2009. 2, 3
[15] H. Sun and Q. Wu, *Optimal rates of distributed regression with imperfect kernels*, J. Mach. Learn. Res., in press, 2021. 2.1
[16] Z. Wang, S. Deng, and Y. Ye, *Close the gaps: A learning-while-doing algorithm for single-product revenue management problems*, Oper. Res., 62(2):318–331, 2014. 1
[17] T. Zhang, *Leave-one-out bounds for kernel methods*, Neural Comput., 15(6):1397–1437, 2003. 2, 3
[18] Y. Zhang, J.C. Duchi, and M.J. Wainwright, *Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates*, J. Mach. Learn. Res., 16:3299–3340, 2015. 1, 2.1