Noname manuscript No.
(will be inserted by the editor)

Distributed kernel-based gradient descent algorithms

Shao-Bo Lin · **Ding-Xuan Zhou**

Received: date / Accepted: date

Abstract We study the generalization ability of distributed learning equipped with a divide-and-conquer approach and gradient descent algorithm in a reproducing kernel Hilbert space (RKHS). Using special spectral features of the gradient descent algorithms and a novel integral operator approach, we provide optimal learning rates of *distributed gradient descent algorithms* in probability and partly conquer the saturation phenomenon in the literature in the sense that the maximum number of local machines to guarantee the optimal learning rates does not vary if the regularity of the regression function goes beyond a certain quantity. We also find that additional un-labeled data can help relaxing the restriction on the number of local machines in distributed learning.

Keywords Learning theory · Distributed learning · Gradient descent algorithm · Integral operator

Mathematics Subject Classification (2000) MSC 68T05 · MSC 94A20 · 41A35

1 Introduction

Distributed learning based on a divide-and-conquer approach has triggered enormous recent research activities in various areas such as optimization [27], data mining [26], and machine learning [13]. This learning strategy breaks up a big problem into manageable pieces, operates learning algorithms on each piece on individual machines or

The work described in this paper is partially supported by the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006] and by the National Natural Science Foundation of China [Grant No. 61502342].

S. B. Lin
College of Mathematics and Information Science, Wenzhou University, Wenzhou, China
E-mail: sblin1983@gmail.com

D. X. Zhou
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China
E-mail: mazhou@cityu.edu.hk

processors, and then puts the individual solutions together to get a final global output. In this way, distributed learning is feasible to conquer big data challenges [30], promote the privacy protection [2], and reduce communication risks [21]. A number of high-adaptive and fault-tolerant distributed data management systems have been practically developed based on distributed learning. Typical examples include the *Hadoop* [9] and *Spark* [1] systems.

Theoretical foundations of distributed learning form a hot topic in machine learning and have been attempted recently in the framework of learning theory [19, 28, 17, 14, 5]. For example, a variance estimate for distributed conditional maximum entropy models was provided in [19]. Optimal learning rates in expectation for distributed regularized least squares were established in [28] under some eigenfunction assumptions, which were improved in [17] by removing the eigenfunction assumptions with a novel integral operator method. In [14], as well as in an independent work [5], optimal learning rates in expectation for distributed spectral algorithms were presented.

This paper aims at refined analysis of distributed learning with kernel-based gradient descent algorithms. Given a Mercer kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ on a compact metric space $\mathscr{X}$ (input space), and a data set $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathscr{X} \times \mathscr{Y}$ with $\mathscr{Y} \subseteq \mathbb{R}$ being the output space, the kernel-based gradient descent algorithm can be stated iteratively with $f_{0,D} = 0$ as

$$f_{t+1,D} = f_{t,D} - \frac{\beta}{|D|} \sum_{(x,y)\in D} (f_{t,D}(x) - y) K_x, \tag{1}$$

where $\beta > 0$ is a step size, $K_x = K(\cdot, x)$ and $|D|$ denotes the cardinality of the set D. The *distributed kernel-based gradient descent algorithm* considered in this paper starts with a partition of the data set D into m disjoint subsets $\{D_j\}_{j=1}^m$. Then it assigns each data subset D_j to a local machine to produce a local estimator f_{t,D_j} by using (1). Finally, these local estimators are communicated to a central processor to derive a global estimator $\overline{f}_{t,D}$ by taking a weighted average

$$\overline{f}_{t,D} = \sum_{j=1}^m \frac{|D_j|}{|D|} f_{t,D_j}. \tag{2}$$

The gradient descent algorithm (1) can be regarded as a special spectral algorithm [18], so optimal learning rates for the distributed algorithm (2) may be obtained from general results for distributed spectral algorithms in [14, 5]. However, the generality of the results in [14, 5] for general spectral algorithms imposes a saturation phenomenon with respect to the number of local machines in the sense that the maximal m to guarantee optimal learning rates no longer improves when the regression function goes beyond a certain level of regularity (see Section 3 for a detailed description). The first purpose of this paper is to conquer this saturation phenomenon by means of special features of the gradient descent algorithm. Using two representations of the difference between f_{t,D_j} and its data-free limit f_t (to be given in Section 4), we shall provide a new error decomposition for distributed kernel-based gradient descent algorithms. With this, the recently developed integral operator approach for distributed learning [14, 17] will be used to obtain optimal learning rates without

saturation. Different from the previous results in [28,17,14,5] established in expectation, our learning rates are in probability. As a consequence, we deduce almost sure convergence of distributed kernel-based gradient descent algorithms by using the Borel-Canttelli Lemma. The second purpose of this paper is to propose the use of additional un-labeled data to enhance the performance of the distributed algorithm (2). We prove that by inputting some additional un-labeled data, the maximal number m of local machines to guarantee the optimal learning rate of $\overline{f}_{t,D}$ can be enlarged (See Section 3 for detailed comparisons).

2 Main Results

Our analysis is carried out in a standard least squares regression framework. Let the sample $D = \{(x_i, y_i)\}_{i=1}^N$ be independently drawn according to ρ, a Borel probability measure on $\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$. Our primary objective is the regression function defined by

$$f_\rho(x) = \int_{\mathscr{Y}} y d\rho(y|x), \qquad x \in \mathscr{X},$$

where $\rho(y|x)$ denotes the conditional distribution at x induced by ρ. Throughout this paper, we assume $\int_{\mathscr{Y}} y^2 d\rho < \infty$ and

$$\int_{\mathscr{Y}} \left(e^{\frac{|y-f_\rho(x)|}{M}} - \frac{|y-f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \le \frac{\gamma^2}{2M^2}, \qquad \forall x \in \mathscr{X}, \tag{3}$$

where M and γ are positive constants. Condition (3) was adopted in [6] to derive confidence-based error estimates for regularized least squares and in [3] for spectral algorithms. It can be found in [23, page 103] or [3] that (3) is equivalent to the following momentum condition (up to a change of constants)

$$\int_{\mathscr{Y}} |y - f_\rho(x)|^\ell d\rho(y|x) \le \frac{1}{2} \ell! \gamma^2 M^{\ell-2}, \qquad \forall \ell \ge 2, x \in \mathscr{X}.$$

Hence (3) is a broad model for the noise of the output y and it is satisfied if the noise is uniformly bounded, Gaussian or sub-Gaussian [20].

Let $L^2_{\rho_X}$ be the Hilbert space of ρ_X square integrable functions on $\mathscr{X}$, with norm denoted by $\|\cdot\|_\rho$, and $\mathscr{H}_K$ be the reproducing kernel Hilbert space associated with the Mercer kernel K. Since $\mathscr{X}$ is compact and K is a Mercer kernel, $\kappa = \sqrt{\sup_{x \in \mathscr{X}} K(x,x)} < \infty$. Furthermore, $K : \mathscr{X} \times \mathscr{X} \to \mathscr{R}$ defines an integral operator L_K on $\mathscr{H}_K$ (or $L^2_{\rho_X}$) by

$$L_K(f) = \int_{\mathscr{X}} K_x f(x) d\rho_X, \qquad f \in \mathscr{H}_K \quad (\text{or } f \in L^2_{\rho_X}).$$

Our error analysis for the *distributed gradient descent algorithm* is stated in terms of the following *regularity condition*

$$f_\rho = L_K^r(h_\rho), \ \text{ for some } r > 0 \text{ and } h_\rho \in L^2_{\rho_X}, \tag{4}$$

where L_K^r denotes the r-th power of $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ as a compact and positive operator. We use the *effective dimension* $\mathscr{N}(\lambda)$ to measure the complexity of $\mathscr{H}_K$ with respect to ρ_X which is defined to be the trace of the operator $(L_K + \lambda I)^{-1} L_K$, that is,

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

2.1 Optimal learning rates

The following error estimate for the *distributed gradient descent algorithm* (2) is the first result of this paper and will be proved in Section 5.

Theorem 1 *Let $0 < \delta < 1$, $0 < \beta \le \kappa^{-2}$. Assume (3) and (4) with $r > 1/2$, then for $t \in \mathbb{N}$ and $\lambda = t^{-1}$, with confidence at least $1 - \delta$, there holds*

$$\|\overline{f}_{t,D} - f_\rho\|_\rho \le C \left\{ t^{-r} + \log(t+1) \widetilde{\mathscr{A}_{D,\lambda}} \log^4 \frac{12m}{\delta} + \mathscr{A}_{D,\lambda} \log \frac{8}{\delta} \right\}, \tag{5}$$

where C is a constant depending only on $M, \gamma, \beta, \kappa, \|h_\rho\|_\rho$ and r, and

$$\mathscr{A}_{D,\lambda} = \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{\mathscr{N}(\lambda)}}{\sqrt{|D|}}, \qquad \widetilde{\mathscr{A}_{D,\lambda}} = \max_{1 \le j \le m} \left[\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}} \right)^2 + 1 \right] \frac{\mathscr{A}^2_{D_j,\lambda}}{\sqrt{\lambda}}. \tag{6}$$

For optimal learning rates of *distributed gradient descent algorithms*, we also need to quantify the effective dimension $\mathscr{N}(\lambda)$ with a parameter $0 < s \le 1$ and a constant $C_0 \ge 1$ as

$$\mathscr{N}(\lambda) \le C_0 \lambda^{-s}, \qquad \forall \lambda > 0. \tag{7}$$

When $s = 1$, condition (7) always holds with the constant $C_0 \ge \mathrm{Tr}(L_K)$. For $0 < s < 1$, the above condition is slightly more general than an eigenvalue decaying assumption in the literature [6]. Indeed, let $\{(\sigma_\ell, \phi_\ell)\}_\ell$ be a set of normalized eigenpairs of the operator L_K on $\mathscr{H}_K$ with $\{\phi_\ell\}_{\ell=1}^\infty$ forming an orthonormal basis of $\mathscr{H}_K$. If $\sigma_n \le C_0 n^{-\alpha}$ for some $\alpha > 1$ and $C_0 \ge 1$, then the eigenvalues of the operator $(\lambda I + L_K)^{-1} L_K$ are $\{\frac{\sigma_\ell}{\lambda + \sigma_\ell}\}_\ell$ and we have

$$\begin{aligned} \mathscr{N}(\lambda) &= \sum_{\ell=1}^\infty \frac{\sigma_\ell}{\lambda + \sigma_\ell} \le \sum_{\ell=1}^\infty \frac{C_0 \ell^{-\alpha}}{\lambda + C_0 \ell^{-\alpha}} = \sum_{\ell=1}^\infty \frac{C_0}{C_0 + \lambda \ell^\alpha} \\ &\le \int_0^\infty \frac{C_0}{C_0 + \lambda t^\alpha} dt = \mathscr{O}(\lambda^{-1/\alpha}). \end{aligned}$$

Therefore, (7) follows from the eigenvalue decaying assumption $\sigma_n = \mathscr{O}(n^{-1/s})$ with $0 < s < 1$.

The following corollary, to be proved in Section 5, exhibits the concrete learning rates of the distributed kernel-based gradient descent algorithm (2). Denote $\lceil a \rceil$ as the smallest integer not less than $a > 0$.

Corollary 1 *Let $0 < \delta < 1$ and $0 < \beta \le \kappa^{-2}$. Assume (3), (7) with $0 < s \le 1$, (4) with $r > 1/2$, and $|D_1| = |D_2| = \cdots = |D_m|$. If $t = \left\lceil |D|^{\frac{1}{2r+s}} \right\rceil$ and*

$$m \le \frac{|D|^{\frac{r-1/2}{2r+s}}}{\log^5 |D| + 1}, \tag{8}$$

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

$$\|\overline{f}_{t,D} - f_\rho\|_\rho \le C' |D|^{-\frac{r}{2r+s}} \log^4 \frac{12}{\delta},$$

where C' is a constant depending only on $M, \gamma, \beta, \kappa, \|h_\rho\|_\rho$, C_0 and r.

Applying the probability to expectation formula for nonnegative random variables

$$E[\xi] = \int_0^\infty \text{Prob}\,[\xi > t]\, dt \tag{9}$$

to $\|\overline{f}_{t,D} - f_\rho\|_\rho^2$, we can easily deduce the following optimal learning rate in expectation.

Corollary 2 *Let $0 < \beta \le \kappa^{-2}$. Assume (3), (7) with $0 < s \le 1$, (4) with $r > 1/2$, and $|D_1| = |D_2| = \cdots = |D_m|$. If $t = \left\lceil |D|^{\frac{1}{2r+s}} \right\rceil$ and (8) holds, then*

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

Based on the confidence-based error estimate in Corollary 1, we can derive almost sure convergence of the *distributed gradient descent algorithm* (2).

Corollary 3 *Let $0 < \beta \le \kappa^{-2}$. Assume (3), (7) with $0 < s \le 1$, (4) with $r > 1/2$, and $|D_1| = |D_2| = \cdots = |D_m|$. If $t = \left\lceil |D|^{\frac{1}{2r+s}} \right\rceil$ and (8) holds, then for arbitrary $\varepsilon > 0$, there holds*

$$\lim_{|D|\to\infty} |D|^{\frac{r}{2r+s}(1-\varepsilon)} \|\overline{f}_{t,D} - f_\rho\|_\rho = 0.$$

2.2 Allowing more local machines by using additional un-labeled data

Although optimal learning rates of the algorithm (2) were stated in the previous subsection, the restriction (8) on the number of local machines seems a bit strict. In this subsection, we show that this restriction can be relaxed by using additional un-labeled data. Utilizing un-labeled data was studied in [7] for a different purpose of improving learning rates for spectral algorithms when $f_\rho \notin \mathscr{H}_K$. It was also adopted in [4] for this purpose for kernel-based conjugate gradient algorithms. The idea of applying un-labeled data to relaxing the restrictions on the number local processors is motivated by our earlier empirical experiments done for distributed regularized least squares. These experiments and theoretical analysis carried out afterwards can be found in [8].

Let $\tilde{D}_j(x)=\{x_1^j,\ldots,x_{|\tilde{D}_j|}^j\}$ be drawn independently according to ρ_X. We then introduce the training set associated with labeled and un-labeled data in each local machine as

$$D_j^*=D_j\cup\tilde{D}_j=\{x_i^*,y_i^*\}_{i=1}^{|D_j^*|}$$

with

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

where $D_j(x)=\{x:(x,y)\in D \text{ for some } y\in\mathscr{Y}\}$. Let $D^*=\cup_{j=1}^m D_j^*$. We can obtain the following enhanced results.

Theorem 2 *Let $0<\delta<1$, $0<\beta\le\kappa^{-2}$. Assume (3) and (4) with $r>1/2$, then for $t\in\mathbb{N}$ and $\lambda=t^{-1}$ with confidence at least $1-\delta$, there holds*

$$\|\overline{f}_{t,D^*}-f_\rho\|_\rho\le C\left\{t^{-r}+\log(t+1)\widetilde{\mathscr{A}_{D,D^*,\lambda}}\log^4\frac{12m}{\delta}+\mathscr{A}_{D,\lambda}\log\frac{8}{\delta}\right\},\qquad(10)$$

where

$$\widetilde{\mathscr{A}_{D,D^*,\lambda}}:=\max_{1\le j\le m}\left[\left(\frac{\mathscr{A}_{D_j^*,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\frac{\mathscr{A}_{D_j^*,\lambda}\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}.\qquad(11)$$

Based on Theorem 2, we can relax the restriction on m as follows.

Corollary 4 *Let $0<\delta<1$ and $0<\beta\le\kappa^{-2}$. Assume (3), (7) with $0<s\le1$, (4) with $r>1/2$, $|D_1|=|D_2|=\ldots=|D_m|$ and $|D_1^*|=|D_2^*|=\ldots=|D_m^*|$. If $t=\left\lceil|D|^{\frac{1}{2r+s}}\right\rceil$ and*

$$m\le\frac{\min\left\{|D^*|^{1/2}|D|^{-\frac{s+1}{4r+2s}},|D^*|^{1/3}|D|^{\frac{2r+s-2}{6r+3s}}\right\}}{\log^5|D|+1},\qquad(12)$$

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

$$\|\overline{f}_{t,D^*}-f_\rho\|_\rho\le C'|D|^{-\frac{r}{2r+s}}\log^4\frac{12}{\delta}.\qquad(13)$$

3 Related Work and Discussions

The kernel-based kernel gradient descent algorithm algorithms (1) can be viewed as a special case of spectral algorithms, which is well known in the context of inverse problems [12].

To describe this in detail, we define an empirical integral operator $L_{K,D(x)}$ by

$$L_{K,D(x)}(f)=\frac{1}{|D|}\sum_{x\in D(x)}f(x)K_x,\qquad f\in\mathscr{H}_K.$$

The gradient descent algorithm (1) can be rewritten as $f_{0,D}=0$ and

$$f_{t+1,D}=f_{t,D}-\beta(L_{K,D(x)}f_{t,D}-\hat{f}_{K,D})=(I-\beta L_{K,D(x)})f_{t,D}+\beta\hat{f}_{K,D}, \tag{14}$$

where $\hat{f}_{K,D}=\frac{1}{|D|}\sum_{(x_i,y_i)\in D}y_iK_{x_i}$. It follows directly that

$$f_{t,D}=\sum_{k=0}^{t-1}\beta\pi'_{k+1}(L_{K,D(x)})\hat{f}_{K,D}, \tag{15}$$

where π'_{k+1} denotes the polynomial ($\pi'_t\equiv 1$),

$$\pi'_{k+1}(u)=\Pi_{\ell=k+1}^{t-1}(1-\beta u)=(1-\beta u)^{t-k-1}$$

and $\pi'_{k+1}(L_{K,D(x)})$ is defined by spectral calculus [18,14]. Therefore, the gradient descent algorithm (1) is a member of the family of spectral algorithms [18] corresponding to the filter function

$$g_\lambda(u)=\sum_{k=0}^{t-1}\beta\pi'_{k+1}(u)=\frac{1-(1-\beta u)^t}{u},\qquad u>0 \tag{16}$$

with $\lambda=1/t$.

As a typical example of spectral algorithms, the gradient descent algorithm (1) has the advantage of overcoming the saturation phenomenon of the regularized least squares [18]. Furthermore, the computational complexity of algorithm (1) is $\mathcal{O}(|D|^2)$, which is much smaller than that of the regularized least squares [25]. Learning rates of gradient descent algorithms have been studied in [25,3,7,20,10,14]. To be more specific, an integral operator approach developed in [22] was used in [25] to derive learning rates for algorithm (1) in the special case of $s=1$ in (7), which were improved to be almost optimal in [3] by noting that algorithm (1) is a special spectral algorithm. For the general case of $0<s<1$ in (7), almost optimal learning rates of spectral algorithms including the gradient descent algorithms (1) were established in [7], but additional un-labeled data were required. In [20], optimal learning rates of gradient descent algorithms were established for $r=1/2$ in (4) without un-labeled data. Optimal learning rates of spectral algorithms including (1) were derived in our recent paper [14] for $r\geq 1/2$ by using a novel integral operator approach.

Remark 1 After the submission in January 2016 of our previous paper [14] on distributed spectral algorithms, we found two independent nice papers in arxiv: [10] in May 2016 and [5] in October 2016. For the classical spectral algorithms, optimal learning rates were established in [10] under assumptions (4), (3) and some eigenvalue decaying conditions. For the distributed spectral algorithms, optimal learning rates were obtained in [5] under the effective dimension assumption (7).

As a special class of distributed spectral algorithms, optimal learning rates of the distributed gradient descent algorithm (2) have been provided in [14,5]. That is, under the conditions of Corollary 1, if

$$m\leq|D|^{\min\{\frac{2}{2r+s},\frac{2r-1}{2r+s}\}}, \tag{17}$$

then

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

We see from (17) that the restriction on the number of local machines suffers from a saturation phenomenon in the sense that when $r>3/2$, the maximal m to guarantee the optimal learning rate does not improve as r increases and is the same as that of $r=3/2$. This is quite different from the case when $1/2\le r\le 3/2$. In the present paper, we use special features of the distributed kernel-based gradient descent algorithms and provide optimal learning rates in confidence under the assumption (8). Comparing (8) with (17), we find that the saturation is partly overcome in the sense that the maximal m to guarantee the optimal learning rate is strictly increasing with respect to r and

$$\frac{|D|^{\frac{r-1/2}{2r+s}}}{\log^4|D|+1}\ge\tilde{C}_r|D|^{\frac{2}{2r+s}},\qquad \text{if } r>\frac{5}{2},$$

where $\tilde{C}_r$ is a constant depending only on r. It should be mentioned that when $r\le 5/2$, our result is a little worse than that in [14], because

$$\frac{|D|^{\frac{r-1/2}{2r+s}}}{\log^4|D|+1}\le\tilde{C}_r|D|^{\frac{2}{2r+s}}.$$

We think the reason is that we devote to the confidence-based error estimate for distributed kernel-based gradient descent algorithms requiring a deterministic error decomposition, which is totally different from the previous methods [28,17,14,5] focusing on deriving error decompositions for distributed learning in expectation. Based on the confidence-based error estimate, we can derive the almost sure convergence of algorithm (2). We believe that using some delicate techniques in integral operators, our restriction on m can be relaxed to

$$m\le|D|^{\frac{2r-1}{2r+s}} \tag{18}$$

for arbitrary $r>1/2$.

Adopting un-labeled data to improve learning rates of spectral algorithms was proposed in [7]. Corollary 4 in our paper shows that unlabeled data can also be used to enlarge the range of the number of local machines. In fact, if $|D^*|=|D|$ and $r>1/2$, we have

$$|D^*|^{1/2}|D|^{-\frac{s+1}{4r+2s}}=|D|^{\frac{2r-1}{4r+2s}},$$

and

$$|D^*|^{1/3}|D|^{\frac{2r+s-2}{6r+3s}}>|D|^{\frac{2r-1}{4r+2s}}.$$

Then, (12) coincides with (8). However, if $|D^*|>|D|$, we obtain

$$|D|^{\frac{2r-1}{4r+2s}}<\min\left\{|D^*|^{1/2}|D|^{-\frac{s+1}{4r+2s}},|D^*|^{1/3}|D|^{\frac{2r+s-2}{6r+3s}}\right\},$$

which shows an essential advantage of using un-labeled data in distributed learning. In particular, when $|D^*|=|D|^2$, it is derived from Corollary 4 that if

$$m\le\frac{|D|^{\frac{2r-2/3}{2r+s}}}{\log^5|D|+1}, \tag{19}$$

then (13) holds with confidence at least $1-\delta$. It should be noticed that the restriction (19) is even weaker than the restriction (18).

By combining our approach with results in [7], we conjecture that optimal learning rates of distributed kernel-based gradient descent algorithms can be derived when the regression function is outside $\mathscr{H}_K$ by adding un-labeled data in the learning process, as done for distributed regularized least squares in [8]. This paper is focused on distributed learning with the gradient descent algorithm. It would be nice to extend our analysis to other algorithms [24,15,16] by using un-labeled data.

4 Error Decomposition Based on Integral Operators

Our error decomposition is motivated by some special features of the gradient descent algorithm and a recent developed integral operator approach [14,17]. Our main novelty is to use two special representations of $f_{t,D}-f_t$ (with $\{f_t\}$ to be defined by (20) below) to derive an error decomposition in a deterministic sense, different from the decomposition in [28,17,14,5] involving the expectation of the generalization error.

4.1 Special representations for gradient descent algorithms

To demonstrate our ideas, we need data-free limits of the sequence $\{f_{t,D}\}$ defined as a sequence $\{f_t\}_t$ by $f_0=0$ and

$$f_{t+1}=f_t-\beta L_K(f_t-f_\rho). \tag{20}$$

The *first novelty* of our error decomposition is to decompose the iteration relation $f_{t+1}=(I-\beta L_K)f_t+\beta L_K f_\rho$ from (20) in terms of the empirical integral operator $L_{K,D(x)}$ as

$$f_{t+1}=\left(I-\beta L_{K,D(x)}\right)f_t+\beta\left(L_{K,D(x)}-L_K\right)f_t+\beta L_K f_\rho.$$

It follows by induction that

$$f_t=\sum_{k=0}^{t-1}\beta\pi'_{k+1}(L_{K,D(x)})\left[\left(L_{K,D(x)}-L_K\right)f_k+L_K f_\rho\right]. \tag{21}$$

This together with (15) yields the first representation for $f_{t,D}-f_t$ as

$$f_{t,D}-f_t=\sum_{k=0}^{t-1}\beta\pi'_{k+1}(L_{K,D(x)})\chi_{k,D}, \tag{22}$$

where

$$\chi_{k,D}=\hat{f}_{K,D}-L_K f_\rho+(L_K-L_{K,D(x)})f_k.$$

Furthermore, from [25, Proposition 4.3], we can get the second representation for $f_{t,D}-f_t$ as

$$f_{t,D}-f_t=\sum_{k=0}^{t-1}\beta\pi'_{k+1}(L_K)\chi^*_{k,D} \tag{23}$$

with

$$\chi^*_{k,D} = \hat{f}_{K,D} - L_K f_\rho + (L_K - L_{K,D(x)}) f_{k,D}.$$

The above two representations of $f_{t,D} - f_t$ will play essential roles in our analysis.

4.2 Special features of the gradient descent algorithm

To present the error decomposition, we unify (23) and (22) to be

$$F_1 = \sum_{k=0}^{t-1} \beta \pi'_{k+1}(L_K) G_k, \qquad \text{and} \qquad F_2 = \sum_{k=0}^{t-1} \beta \pi'_{k+1}(L_{K,D(x)}) G_k$$

with $G_k \in \mathscr{H}_K$ and bound the norm as

$$\max\left\{\|F_1\|_\rho, \sqrt{\lambda}\|F_1\|_K\right\} = \max\left\{\left\|L_K^{1/2} F_1\right\|_K, \sqrt{\lambda}\|F_1\|_K\right\} \le \left\|(L_K+\lambda I)^{\frac{1}{2}} F_1\right\|_K$$
$$= \left\|\sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K)(L_K+\lambda I)^{-\frac{1}{2}} G_k\right\|_K, \tag{24}$$

and

$$\max\left\{\|F_2\|_\rho, \sqrt{\lambda}\|F_2\|_K\right\} = \max\left\{\left\|L_K^{1/2} F_2\right\|_K, \sqrt{\lambda}\|F_2\|_K\right\} \le \left\|(L_K+\lambda I)^{\frac{1}{2}} F_2\right\|_K$$
$$\le \left\|(L_K+\lambda I)^{\frac{1}{2}}(L_{K,D(x)}+\lambda I)^{-1/2}\right\| \left\|(L_{K,D(x)}+\lambda I)^{\frac{1}{2}} F_2\right\|_K$$
$$= \mathscr{Q}_{D,\lambda} \left\|\sum_{k=0}^{t-1} \beta (L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})(L_{K,D(x)}+\lambda I)^{-\frac{1}{2}} G_k\right\|_K, \tag{25}$$

where $\lambda > 0$ can be arbitrarily chosen and $\mathscr{Q}_{D,\lambda}$ is an operator norm defined by

$$\mathscr{Q}_{D,\lambda} = \left\|(L_K+\lambda I)^{\frac{1}{2}}(L_{K,D(x)}+\lambda I)^{-\frac{1}{2}}\right\|. \tag{26}$$

The *second novelty* of our error decomposition is to bound the norm (26) tightly using our work in [17, 14] and to use special features of the gradient descent algorithm for estimating the norms concerning the operator $\beta(L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})$ and $\beta(L_K+\lambda I)\pi'_{k+1}(L_K)$ as follows.

Lemma 1 *For $\lambda > 0$, $0 < \beta \le \kappa^{-2}$, $t \in \mathbb{N}$ and $k = 0, 1, \ldots, t-1$, we have*

$$\max\left\{\left\|\beta(L_K+\lambda I)\pi'_{k+1}(L_K)\right\|, \left\|\beta(L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})\right\|\right\} \le \frac{1}{t-k} + \beta\lambda \tag{27}$$

and

$$\max\left\{\left\|\sum_{k=0}^{t-1} \beta(L_K+\lambda I)\pi'_{k+1}(L_K)\right\|, \left\|\sum_{k=0}^{t-1} \beta(L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})\right\|\right\} \le 1 + \beta\lambda t. \tag{28}$$

Proof. We only prove (27) and (28) for the operator norms concerning $L_{K,D(x)}$. The inequalities concerning the operator L_K can be derived by using the same method. Let $\{\sigma_i^{\mathbf{x}}\}_i$ be the set of all eigenvalues of the operator $L_{K,D(x)}$ on $\mathscr{H}_K$. Then $0 \le \sigma_i^{\mathbf{x}} \le \|L_{K,D(x)}\| \le \kappa^2$ and the symmetric operator $\beta(L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})$ has eigenvalues

$$\beta(\sigma_i^{\mathbf{x}}+\lambda)\pi'_{k+1}(\sigma_i^{\mathbf{x}}) = (\beta\sigma_i^{\mathbf{x}}+\beta\lambda)\,(1-\beta\sigma_i^{\mathbf{x}})^{t-k-1}.$$

Since $0<\beta\le\kappa^{-2}$, these eigenvalues are nonnegative and bounded by

$$\beta\sigma_i^{\mathbf{x}}\,(1-\beta\sigma_i^{\mathbf{x}})^{t-k-1}+\beta\lambda \le \frac{1}{t-k}+\beta\lambda.$$

Here we have used the fact that the univariate function $u(1-u)^{t-k-1}$ defined on the interval $[0,1]$ takes its maximum values at $u=\frac{1}{t-k}$ and satisfies

$$0 \le u(1-u)^{t-k-1} \le \frac{1}{t-k}, \qquad \forall 0 \le u \le 1.$$

Then the first desired norm estimate (27) follows.

The above proof also shows that $\|\pi'_{k+1}(L_{K,D(x)})\| \le 1$ for $k \in \{0,\dots,t-1\}$. To verify the second estimate, we note that the symmetric operator $\beta L_{K,D(x)}$ has eigenvalues $0 \le \beta\sigma_i^{\mathbf{x}} \le 1$. It follows that the operator $I-\beta L_{K,D(x)}$ is positive and (16) with $u = L_{K,D(x)}$ yields

$$\left\| \sum_{k=0}^{t-1} \beta L_{K,D(x)} \pi'_{k+1}(L_{K,D(x)}) \right\| \le 1.$$

Then the second desired norm estimate (28) follows. □

4.3 A novel error decomposition for gradient descent algorithm

To derive the error decomposition for *distributed gradient descent algorithms*, we shall use the representation (22) and Lemma 1 and derive a novel error decomposition for the gradient descent algorithm in the following proposition.

Proposition 1 *Let $\lambda > 0$ and $0 < \beta \le \kappa^{-2}$. If (4) holds with $r > 1/2$, then*

$$\begin{aligned} &\max\left\{\|f_{t,D}-f_t\|_\rho, \sqrt{\lambda}\|f_{t,D}-f_t\|_K\right\} \\ &\le (1+\lambda t\beta)\,\mathscr{D}_{D,\lambda}^2\left(\mathscr{P}_{D,\lambda}+\mathscr{R}_{D,\lambda}\|f_\rho\|_K\right) + \sum_{\ell=0}^{t-1}\left(\frac{1}{t-\ell}+\lambda\beta\right)\|f_\ell-f_\rho\|_K \mathscr{D}_{D,\lambda}^2\mathscr{R}_{D,\lambda}, \end{aligned}$$

where

$$\mathscr{P}_{D,\lambda} := \left\|(L_K+\lambda I)^{-1/2}(L_K f_\rho - \hat{f}_{K,D})\right\|_K,$$

and

$$\mathscr{R}_{D,\lambda} := \left\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,D(x)})\right\|.$$

Proof. For arbitrary $t \geq 0$, it follows from (22) and (25) that

$$\begin{aligned}
&\max\left\{\|f_{t,D}-f_t\|_\rho, \sqrt{\lambda}\|f_{t,D}-f_t\|_K\right\} \\
&\leq \mathscr{Q}_{D,\lambda}\left\|\sum_{k=0}^{t-1}\beta(L_{K,D(x)}+\lambda I)\pi'_{k+1}(L_{K,D(x)})(L_{K,D(x)}+\lambda I)^{-1/2}\chi_{k,D}\right\|_K \\
&\leq \mathscr{Q}_{D,\lambda}\left\|\sum_{\ell=0}^{t-1}\beta(L_{K,D(x)}+\lambda I)\pi'_{\ell+1}(L_{K,D(x)})(L_{K,D(x)}+\lambda I)^{-1/2}(\hat{f}_{K,D}-L_K f_\rho)\right\|_K \\
&+ \mathscr{Q}_{D,\lambda}\left\|\sum_{\ell=0}^{t-1}\beta(L_{K,D(x)}+\lambda I)\pi'_{\ell+1}(L_{K,D(x)})(L_{K,D(x)}+\lambda I)^{-1/2}(L_K-L_{K,D(x)})(f_\ell-f_\rho)\right\|_K \\
&+ \mathscr{Q}_{D,\lambda}\left\|\sum_{\ell=0}^{t-1}\beta(L_{K,D(x)}+\lambda I)\pi'_{\ell+1}(L_{K,D(x)})(L_{K,D(x)}+\lambda I)^{-1/2}(L_K-L_{K,D(x)})f_\rho\right\|_K \\
&=: \mathscr{Q}_{D,\lambda}(A_{1,t,\lambda,D}+A_{2,t,\lambda,D}+A_{3,t,\lambda,D}). \qquad (29)
\end{aligned}$$

Concerning $A_{1,t,\lambda,D}$, (28) and the definitions of $\mathscr{P}_{D,\lambda}$ and $\mathscr{Q}_{D,\lambda}$ yield

$$\begin{aligned}
A_{1,t,\lambda,D} &\leq \left\|\sum_{\ell=0}^{t-1}\beta(L_{K,D(x)}+\lambda I)\pi'_{\ell+1}(L_{K,D(x)})\right\|\|(L_{K,D(x)}+\lambda I)^{-1/2}(\hat{f}_{K,D}-L_K f_\rho)\|_K \\
&\leq (1+\beta\lambda t)\|(L_{K,D(x)}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\|\|(L_K+\lambda I)^{-1/2}(\hat{f}_{K,D}-L_K f_\rho)\|_K \\
&= (1+\beta\lambda t)\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda},
\end{aligned}$$

where we have used $\|AB\| = \|BA\|$ for positive operators A, B in the last equality. Since (4) holds for $r > 1/2$, we have $f_\rho \in \mathscr{H}_K$. Then (28) together with the definition of $\mathscr{R}_{D,\lambda}$ and

$$\|Af\|_K \leq \|A\|\|f\|_K \qquad (30)$$

for positive operator A and $f \in \mathscr{H}_K$ yields

$$A_{3,t,\lambda,D} \leq (1+\beta\lambda t)\mathscr{Q}_{D,\lambda}\mathscr{R}_{D,\lambda}\|f_\rho\|_K.$$

Furthermore, (27) and (30) imply

$$\begin{aligned}
A_{2,t,\lambda,D} &\leq \sum_{\ell=0}^{t-1}\left\|\beta(L_{K,D(x)}+\lambda I)\pi'_{\ell+1}(L_{K,D(x)})\right\|\mathscr{Q}_{D,\lambda}\mathscr{R}_{D,\lambda}\|f_\ell-f_\rho\|_K \\
&\leq \sum_{\ell=0}^{t-1}\left(\frac{1}{t-\ell}+\lambda\beta\right)\|f_\ell-f_\rho\|_K\mathscr{Q}_{D,\lambda}\mathscr{R}_{D,\lambda}.
\end{aligned}$$

Inserting bounds of $A_{1,t,\lambda,D}$, $A_{2,t,\lambda,D}$ and $A_{3,t,\lambda,D}$ into (29), we have

$$\begin{aligned}
\max\left\{\|f_{t,D}-f_t\|_\rho, \sqrt{\lambda}\|f_{t,D}-f_t\|_K\right\} &\leq (1+\lambda t\beta)\mathscr{Q}_{D,\lambda}^2\left(\mathscr{P}_{D,\lambda}+\mathscr{R}_{D,\lambda}\|f_\rho\|_K\right) \\
&+ \sum_{\ell=0}^{t-1}\left(\frac{1}{t-\ell}+\lambda\beta\right)\|f_\ell-f_\rho\|_K\mathscr{Q}_{D,\lambda}^2\mathscr{R}_{D,\lambda}.
\end{aligned}$$

This completes the proof of Proposition 1. □

4.4 Error decomposition for distributed gradient descent algorithm

By the aid of Proposition 1, we can use the representation formula (23) to derive the error decomposition of *distributed gradient descent algorithms* in Proposition 2. The main novelty is that our error decomposition is exhibited deterministically rather than in expectation, which makes our analysis totally different from [28, 17, 14, 5].

Proposition 2 *Let $\lambda > 0$ and $0 < \beta \le \kappa^{-2}$. If (4) holds with $r > 1/2$, then*

$$\|\overline{f}_{t,D} - f_\rho\|_\rho \le \|f_t - f_\rho\|_\rho + \mathscr{L}_{D,t,\lambda} + \mathscr{G}_{D,t,\lambda}, \tag{31}$$

where

$$\mathscr{G}_{D,t,\lambda} := \mathscr{R}_{D,\lambda} \sum_{k=0}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) \|f_k - f_\rho\|_K + (1+\lambda\beta t)(\mathscr{P}_{D,\lambda} + \mathscr{R}_{D,\lambda}\|f_\rho\|_K), \tag{32}$$

and

$$\mathscr{L}_{D,t,\lambda} := \max_{1\le j\le m} \frac{\mathscr{Q}^2_{D_j,\lambda}\mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}} \sum_{k=1}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) \left[(1+\lambda t\beta)\left(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda}\|f_\rho\|_K\right) + \sum_{\ell=0}^{k-1}\left(\frac{1}{k-\ell} + \lambda\beta\right)\|f_\ell - f_\rho\|_K \mathscr{R}_{D_j,\lambda}\right]. \tag{33}$$

Proof. Applying (23) to D_j for each fixed $j \in \{1,\dots,m\}$, we have

$$\overline{f}_{t,D} - f_t = \sum_{k=0}^{t-1} \beta\pi'_{k+1}(L_K) \sum_{j=1}^m \frac{|D_j|}{|D|} \chi^*_{k,D_j}.$$

Since $\sum_{j=1}^m \frac{|D_j|}{|D|} = 1$ and

$$\sum_{j=1}^m \frac{|D_j|}{|D|} \hat{f}_{K,D_j} = \sum_{j=1}^m \frac{|D_j|}{|D|} \frac{1}{|D_j|} \sum_{(x,y)\in D_j} yK_x = \frac{1}{|D|} \sum_{(x,y)\in D} yK_x = \hat{f}_{K,D},$$

we have

$$\sum_{j=1}^m \frac{|D_j|}{|D|} \chi^*_{k,D_j} = \sum_{j=1}^m \frac{|D_j|}{|D|} (L_K - L_{K,D_j(x)}) f_{k,D_j} + \hat{f}_{K,D} - L_K f_\rho.$$

Then

$$\begin{aligned} \|\overline{f}_{t,D} - f_t\|_\rho &\le \left\| \sum_{k=0}^{t-1} \beta\pi'_{k+1}(L_K) \sum_{j=1}^m \frac{|D_j|}{|D|} (L_K - L_{K,D_j(x)}) f_{k,D_j} \right\|_\rho \\ &\quad + \left\| \sum_{k=0}^{t-1} \beta\pi'_{k+1}(L_K)(L_K f_\rho - \hat{f}_{K,D}) \right\|_\rho \\ &=: I_1 + I_2. \end{aligned}$$

Bounding I_2 is easy. In fact, we know from (24) that

$$I_2 \le \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K)(L_K+\lambda I)^{-\frac{1}{2}}(L_K f_\rho - \hat{f}_{K,D}) \right\|_K .$$

This together with (28) and the definition of $\mathscr{P}_{D,\lambda}$ yields

$$I_2 \le \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \right\| \mathscr{P}_{D,\lambda} \le (1+\lambda\beta t)\mathscr{P}_{D,\lambda}. \tag{34}$$

Bounding I_1 is more technical. Using (24) again and the triangle inequality, we have

$$\begin{aligned} I_1 \le & \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \sum_{j=1}^{m} \frac{|D_j|}{|D|}(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)})(f_{k,D_j} - f_k) \right\|_K \\ & + \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \sum_{j=1}^{m} \frac{|D_j|}{|D|}(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)})(f_k - f_\rho) \right\|_K \\ & + \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \sum_{j=1}^{m} \frac{|D_j|}{|D|}(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)}) f_\rho \right\|_K \\ =: & I_{1,1}+I_{1,2}+I_{1,3}. \end{aligned}$$

For $f \in \mathscr{H}_K$, we have

$$\sum_{j=1}^{m} \frac{|D_j|}{|D|} L_{K,D_j(x)} f = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \frac{1}{|D_j|} \sum_{x\in D_j(x)} K_x f(x) = \frac{1}{|D|} \sum_{x\in D(x)} K_x f(x) = L_{K,D(x)} f. \tag{35}$$

Then, it is easy to see

$$\begin{aligned} I_{1,2} &\le \sum_{k=0}^{t-1} \left\| \beta (L_K+\lambda I)\pi'_{k+1}(L_K)(L_K+\lambda I)^{-1/2}(L_K - L_{K,D(x)})(f_k - f_\rho) \right\|_K \\ &\le \sum_{k=0}^{t-1} \left\| \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \right\| \left\| (L_K+\lambda I)^{-1/2}(L_K - L_{K,D(x)})(f_k - f_\rho) \right\|_K . \end{aligned}$$

Combining this with (27), $f_\rho \in \mathscr{H}_K$ and the definition of $\mathscr{R}_{D,\lambda}$ yields

$$I_{1,2} \le \mathscr{R}_{D,\lambda} \sum_{k=0}^{t-1} \left(\beta\lambda + \frac{1}{t-k} \right) \| f_k - f_\rho \|_K . \tag{36}$$

Concerning $I_{1,3}$, we use (28) and (35) to get

$$I_{1,3} \le \left\| \sum_{k=0}^{t-1} \beta (L_K+\lambda I)\pi'_{k+1}(L_K) \right\| \mathscr{R}_{D,\lambda} \| f_\rho \|_K \le \mathscr{R}_{D,\lambda}(1+\beta\lambda t) \| f_\rho \|_K . \tag{37}$$

To bound $I_{1,1}$, we use (27), $\sum_{j=1}^m \frac{|D_j|}{|D|} = 1$ $f_0 = f_{0,D_j} = 0$ and Jensen's inequality to obtain

$$\begin{aligned} I_{1,1} &\le \sum_{k=1}^{t-1} \left\|\beta(L_K+\lambda I)\pi'_{k+1}(L_K)\right\| \left\| \sum_{j=1}^m \frac{|D_j|}{|D|}(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)})(f_{k,D_j} - f_k)\right\|_K \\ &\le \sum_{k=1}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right) \sum_{j=1}^m \frac{|D_j|}{|D|} \left\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)})(f_{k,D_j} - f_k)\right\|_K \\ &= \sum_{j=1}^m \frac{|D_j|}{|D|} \sum_{k=1}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right) \left\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,D_j(x)})(f_{k,D_j} - f_k)\right\|_K \\ &\le \max_{1\le j\le m} \sum_{k=1}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right) \|f_{k,D_j} - f_k\|_K \mathscr{R}_{D_j,\lambda}. \end{aligned}$$

But Proposition 1 with D and t being replaced by D_j and k yields that

$$\begin{aligned} &\|f_{k,D_j} - f_k\|_K \\ &\le \frac{\mathscr{Q}^2_{D_j,\lambda}}{\sqrt{\lambda}} \left[(1+\lambda k\beta)\left(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda}\|f_\rho\|_K\right) + \sum_{\ell=0}^{k-1}\left(\frac{1}{k-\ell} + \lambda\beta\right)\|f_\ell - f_\rho\|_K \mathscr{R}_{D_j,\lambda}\right]. \end{aligned}$$

It follows that

$$\begin{aligned} I_{1,1} \le \max_{1\le j\le m} &\frac{\mathscr{Q}^2_{D_j,\lambda}\mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}} \sum_{k=1}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right) \\ &\times \left[(1+\lambda k\beta)\left(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda}\|f_\rho\|_K\right) + \sum_{\ell=0}^{k-1}\left(\frac{1}{k-\ell} + \lambda\beta\right)\|f_\ell - f_\rho\|_K \mathscr{R}_{D_j,\lambda}\right]. \end{aligned} \tag{38}$$

This together with (34), (36), (37) and (38) gives

$$\begin{aligned} \|\overline{f}_{t,D} - f_t\|_\rho \le \max_{1\le j\le m} &\frac{\mathscr{Q}^2_{D_j,\lambda}\mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}} \sum_{k=1}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right) \\ \times &\left[(1+\lambda k\beta)\left(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda}\|f_\rho\|_K\right) + \sum_{\ell=0}^{k-1}\left(\frac{1}{k-\ell} + \lambda\beta\right)\|f_\ell - f_\rho\|_K \mathscr{R}_{D_j,\lambda}\right] \\ + \mathscr{R}_{D,\lambda} &\sum_{k=0}^{t-1}\left(\beta\lambda + \frac{1}{t-k}\right)\|f_k - f_\rho\|_K + (1+\lambda\beta t)(\mathscr{P}_{D,\lambda} + \mathscr{R}_{D,\lambda}\|f_\rho\|_K). \end{aligned}$$

Then (31) follows from the triangle inequality

$$\|\overline{f}_{t,D} - f_\rho\|_\rho \le \|f_t - f_\rho\|_\rho + \|\overline{f}_{t,D} - f_t\|_\rho.$$

This completes the proof of Proposition 2. □

5 Proofs

To prove our main results, we need to bound the quantities $\mathscr{Q}_{D,\lambda}$, $\mathscr{R}_{D,\lambda}$ and $\mathscr{P}_{D,\lambda}$ by the following probability estimates.

Lemma 2 *Let D be a sample drawn independently according to ρ and $0<\delta<1$. If (3) holds, then each of the following estimates holds with confidence at least $1-\delta$,*

$$\mathscr{Q}^2_{D,\lambda} \le 2\left(\frac{2(\kappa^2+\kappa)\mathscr{A}_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+2, \tag{39}$$

$$\mathscr{R}_{D,\lambda} \le 2(\kappa^2+\kappa)\mathscr{A}_{D,\lambda}\log\big(2/\delta\big), \tag{40}$$

$$\mathscr{P}_{D,\lambda} \le 2(\kappa M+\gamma)\mathscr{A}_{D,\lambda}\log\big(2/\delta\big). \tag{41}$$

These inequalities are well studied in the literature. The first two can be found in [17,14] while the last one can be found in [6].

We are in a position to prove the main results of this paper.

Proof of Theorem 1. We follow our error decomposition (31) described in Proposition 2. We need the following bounds for f_t-f_ρ for $t\ge 1$ under the regularity assumption (4) with $r>1/2$, stated as Theorem 2.10 in [25],

$$\|f_t-f_\rho\|_\rho \le \|h_\rho\|_\rho(2r\kappa^2/e)^r t^{-r}, \tag{42}$$

$$\|f_t-f_\rho\|_K \le \|h_\rho\|_\rho[(2r-1)\kappa^2/e]^{r-1/2}t^{-r+\frac{1}{2}}. \tag{43}$$

Then, we use (43) and Lemma 2 to bound $\mathscr{L}_{D,t,\lambda}$ and $\mathscr{G}_{D,t,\lambda}$, respectively.

Step 1. Estimating $\mathscr{G}_{D,t,\lambda}$. Since (4) holds with $r>1/2$, we have

$$\|f_\rho\|_K = \|L_K^r h_\rho\|_K \le \|L_K^{r-1/2}\|\|L_K^{1/2}h_\rho\|_K \le \kappa^{2r-1}\|h_\rho\|_\rho. \tag{44}$$

The above inequality together with (43), (32), $\lambda=1/t$ and $f_0=0$ yields

$$\begin{aligned}
\mathscr{G}_{D,t,\lambda} &= \mathscr{R}_{D,\lambda}(\beta\lambda+t^{-1})\|f_\rho\|_K \\
&\quad+\sum_{k=1}^{t-1}(\beta\lambda+(t-k)^{-1})\|f_k-f_\rho\|_K\mathscr{R}_{D,\lambda}+(1+\lambda\beta t)(\mathscr{P}_{D,\lambda}+\mathscr{R}_{D,\lambda}\|f_\rho\|_K) \\
&\le \|h_\rho\|_\rho[(2r-1)\kappa^2/e]^{r-1/2}\mathscr{R}_{D,\lambda}\sum_{k=1}^{t-1}[\beta\lambda+(t-k)^{-1}]k^{-r+1/2} \\
&\quad+(1+\beta)(1+2\kappa^{2r-1}\|h_\rho\|_\rho)(\mathscr{P}_{D,\lambda}+\mathscr{R}_{D,\lambda}).
\end{aligned} \tag{45}$$

Notice that

$$\begin{aligned}
\sum_{k=1}^{t-1}\frac{k^{-r+\frac{1}{2}}}{t-k} &\le \sum_{1\le k\le t/2}\frac{2}{t}k^{-r+\frac{1}{2}}+\sum_{t/2<k\le t-1}2^{r-\frac{1}{2}}t^{-r+\frac{1}{2}}\frac{1}{t-k} \\
&\le C_r'\begin{cases} t^{-r+\frac{1}{2}}\log(t+1), & \text{when } \frac{1}{2}<r\le\frac{3}{2}, \\ t^{-1}, & \text{when } r>\frac{3}{2}, \end{cases}
\end{aligned} \tag{46}$$

where C_r' is a constant given by

$$C_r' = \begin{cases} \frac{3}{\frac{3}{2}-r} + 2^{r-\frac{1}{2}}, & \text{when } \frac{1}{2} < r < \frac{3}{2}, \\ 8, & \text{when } r = \frac{3}{2}, \\ \left(\frac{2(2r-1)}{2r-3} + 2^{r-\frac{1}{2}}\right) \min_{\ell \in \mathbb{N}} \left\{1 + \ell^{-r+\frac{3}{2}} \log(\ell+1)\right\}, & \text{when } r > \frac{3}{2}, \end{cases}$$

and

$$\sum_{k=1}^{t-1} k^{-r+\frac{1}{2}} \le C_r' \begin{cases} t^{-r+\frac{3}{2}}, & \text{when } \frac{1}{2} < r < \frac{3}{2}, \\ \log(t+1), & \text{when } r = \frac{3}{2}, \\ 1, & \text{when } r > \frac{3}{2}. \end{cases} \tag{47}$$

We obtain

$$\sum_{k=1}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) k^{-r+1/2} \le 2C_r'(1+\beta) B_{t,\lambda,r},$$

where

$$B_{t,\lambda,r} := \left[t^{-1} + \lambda + t^{-r+1/2} + \lambda t^{-r+3/2}\right] \log(t+1).$$

Since $\lambda = 1/t$, we have

$$B_{t,\lambda,r} = 2\left[t^{-1} + t^{-r+1/2}\right] \log(t+1).$$

Due to $r > 1/2$, there exists some constant $C_r \ge 1$ depending only on r such that

$$\max\left\{t^{-1}\log(t+1), t^{-r+1/2}\log(t+1)\right\} \le C_r, \qquad \forall t \ge 1. \tag{48}$$

So

$$B_{t,\lambda,r} \le 4C_r.$$

Then, we have

$$\sum_{k=1}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) k^{-r+1/2} \le 8C_r C_r'(1+\beta). \tag{49}$$

Plugging (49) into (45), we obtain

$$\mathscr{G}_{D,t,\lambda} \le C_1 \left(\mathscr{R}_{D,\lambda} + \mathscr{P}_{D,\lambda}\right), \tag{50}$$

where

$$C_1 = (1+\beta)\max\left\{1 + 2\kappa^{2r-1}\|h_\rho\|_\rho, 8C_r C_r'\|h_\rho\|_\rho [(2r-1)\kappa^2/e]^{r-1/2}\right\}.$$

It follows from Lemma 2 that there exist two subsets $\mathscr{Z}_{1,\delta}^{|D|}$ and $\mathscr{Z}_{2,\delta}^{|D|}$ of $\mathscr{Z}^{|D|}$ with measures at least $1-\delta$ such that for arbitrary $D \subset \mathscr{Z}_{1,\delta}^{|D|} \cap \mathscr{Z}_{2,\delta}^{|D|}$ there holds

$$\mathscr{R}_{D,\lambda} \le 2(\kappa^2+\kappa)\mathscr{A}_{D,\lambda} \log\big(2/\delta\big), \qquad \text{and} \qquad \mathscr{P}_{D,\lambda} \le 2(\kappa M + \gamma)\mathscr{A}_{D,\lambda} \log\big(2/\delta\big).$$

The above estimates together with (50) yield that for arbitrary $D \subset \mathscr{Z}_{1,\delta}^{|D|} \cap \mathscr{Z}_{1,\delta}^{|D|}$, there holds

$$\mathscr{G}_{D,t,\lambda} \leq C_2 \mathscr{A}_{D,\lambda} \log\big(2/\delta\big),$$

where

$$C_2 = 2C_1(\kappa^2 + \kappa + \kappa M + \gamma).$$

Then, with confidence at least $1-\delta/2$, there holds

$$\mathscr{G}_{D,t,\lambda} \leq C_2 \mathscr{A}_{D,\lambda} \log\big(8/\delta\big). \tag{51}$$

Step 2. Estimating $\mathscr{L}_{D,t,\lambda}$. Due to (43) and (44), we have from (33) that

$$\mathscr{L}_{D,t,\lambda} \leq \max_{1\leq j\leq m} \frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}} \sum_{k=1}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) \big\{(1+\lambda t\beta)(\mathscr{P}_{D_j,\lambda} + \kappa^{2r-1}\|h_\rho\|_\rho \mathscr{R}_{D_j,\lambda})$$
$$+ (k^{-1}+\lambda\beta)\kappa^{2r-1}\|h_\rho\|_\rho \mathscr{R}_{D_j,\lambda} + \|h_\rho\|_\rho [(2r-1)\kappa^2/e]^{r-1/2} \sum_{\ell=1}^{k-1} \left(\frac{1}{k-\ell} + \lambda\beta\right) \ell^{-r+1/2} \mathscr{R}_{D_j,\lambda}\big\},$$

where we denote $\sum_{\ell=1}^{0} a_\ell = 0$. Then, it follows from $\lambda = 1/t$ and (49) that

$$\begin{aligned}
\mathscr{L}_{D,t,\lambda} \leq & \max_{1\leq j\leq m} \frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}} \sum_{k=1}^{t-1} \left(\beta\lambda + \frac{1}{t-k}\right) \big\{(1+\beta)(1+\kappa^{2r-1}\|h_\rho\|_\rho)(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda}) \\
+ & \big[(k^{-1}+\lambda\beta) + 8C_r C_r'(1+\beta)[(2r-1)/e]^{r-1/2}\big] \kappa^{2r-1}\|h_\rho\|_\rho \mathscr{R}_{D_j,\lambda}\big\} \\
\leq & 2(1+\beta)^2(1+\kappa^{2r-1}\|h_\rho\|_\rho) \max_{1\leq j\leq m} \frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda})}{\sqrt{\lambda}} \log(t+1) \\
+ & 2(1+\beta)^2\big[1 + 8C_r C_r'[(2r-1)/e]^{r-1/2}\big] \kappa^{2r-1}\|h_\rho\|_\rho \max_{1\leq j\leq m} \frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}^2}{\sqrt{\lambda}} \log(t+1) \\
\leq & C_3 \log(t+1) \max_{1\leq j\leq m} \frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j,\lambda})}{\sqrt{\lambda}},
\end{aligned} \tag{52}$$

where

$$C_3 := 2(1+\beta)^2 \max\{1 + \kappa^{2r-1}\|h_\rho\|_\rho, \big[1 + 8C_r C_r'[(2r-1)/e]^{r-1/2}\big] \kappa^{2r-1}\|h_\rho\|_\rho\}.$$

Furthermore, Lemma 2 implies that for each fixed j, there exist three subsets $\mathscr{Z}_{1,\delta}^{|D_j|}$, $\mathscr{Z}_{2,\delta}^{|D_j|}$ and $\mathscr{Z}_{3,\delta}^{|D_j|}$ of $\mathscr{Z}^{|D_j|}$ with measures at least $1-\delta$ such that for $D_j \subset \mathscr{Z}_{1,\delta}^{|D_j|} \cap \mathscr{Z}_{2,\delta}^{|D_j|} \cap \mathscr{Z}_{3,\delta}^{|D_j|}$ there holds

$$\mathscr{R}_{D_j,\lambda} \leq 2(\kappa^2+\kappa)\mathscr{A}_{D_j,\lambda} \log\big(2/\delta\big), \qquad \mathscr{P}_{D_j,\lambda} \leq 2(\kappa M + \gamma)\mathscr{A}_{D_j,\lambda} \log\big(2/\delta\big),$$

and

$$\mathscr{Q}_{D_j,\lambda}^2 \leq 2\left(\frac{2(\kappa^2+\kappa)\mathscr{A}_{D_j,\lambda} \log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 2.$$

So, for $D_j \subset \mathscr{Z}_{1,\delta}^{|D_j|} \cap \mathscr{Z}_{2,\delta}^{|D_j|} \cap \mathscr{Z}_{3,\delta}^{|D_j|}$, there holds

$$\frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}}(\mathscr{P}_{D_j,\lambda}+\mathscr{R}_{D_j,\lambda}) \le C_4 \left[\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\frac{\mathscr{A}_{D_j,\lambda}^2}{\sqrt{\lambda}}\log^4\frac{2}{\delta},$$

where

$$C_4 := 16(\kappa+1)^4[\kappa^2+\kappa+\kappa M+\gamma].$$

Thus, with confidence at least $1-3\delta$, there holds

$$\frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}}(\mathscr{P}_{D_j,\lambda}+\mathscr{R}_{D_j,\lambda}) \le C_4 \left[\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\frac{\mathscr{A}_{D_j,\lambda}^2}{\sqrt{\lambda}}\log^4\frac{2}{\delta}.$$

This implies that with confidence at least $1-3m\delta$, there holds

$$\max_{1\le j\le m}\frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}}(\mathscr{P}_{D_j,\lambda}+\mathscr{R}_{D_j,\lambda}) \le C_4 \max_{1\le j\le m}\left[\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\frac{\mathscr{A}_{D_j,\lambda}^2}{\sqrt{\lambda}}\log^4\frac{2}{\delta}.$$

Scaling $3m\delta$ to $\frac{\delta}{2}$, we have with confidence at least $1-\delta/2$, there holds

$$\max_{1\le j\le m}\frac{\mathscr{Q}_{D_j,\lambda}^2 \mathscr{R}_{D_j,\lambda}}{\sqrt{\lambda}}(\mathscr{P}_{D_j,\lambda}+\mathscr{R}_{D_j,\lambda}) \le C_4 \max_{1\le j\le m}\left[\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\frac{\mathscr{A}_{D_j,\lambda}^2}{\sqrt{\lambda}}\log^4\frac{12m}{\delta}.$$

All these estimates yield that with confidence at least $1-\delta/2$, there holds

$$\mathscr{L}_{D,t,\lambda} \le C_4C_3\log(t+1)\widetilde{\mathscr{A}_{D,\lambda}}\log^4\frac{12m}{\delta}, \tag{53}$$

where $\widetilde{\mathscr{A}_{D,\lambda}}$ is defined by (6).

Step 3. Deducing learning rate. Plugging (51), (53) and (42) into (31), with confidence $1-\delta$ we have

$$\|\overline{f}_{t,D}-f_\rho\|_\rho \le C\left\{t^{-r}+\log(t+1)\widetilde{\mathscr{A}_{D,\lambda}}\log^4\frac{12m}{\delta}+\mathscr{A}_{D,\lambda}\log\frac{8}{\delta}\right\},$$

where

$$C := \max\left\{\|h_\rho\|_\rho(2r\kappa^2/e)^r, C_4C_3, C_2\right\}.$$

This completes the proof of Theorem 1. □

Proof of Corollary 1. Let $t=\left\lceil |D|^{\frac{1}{(2r+s)}}\right\rceil$. Since $\lambda=t^{-1}$ and $r+s>r>1/2$, we obtain from (6), (7) and $|D_1|=\cdots=|D_m|$ that

$$\mathscr{A}_{D,\lambda} \le |D|^{-1+\frac{1}{4r+2s}}+\sqrt{C_0}|D|^{-\frac{1}{2}+\frac{s}{4r+2s}} \le (\sqrt{C_0}+1)|D|^{-\frac{r}{2r+s}} \tag{54}$$

and

$$\mathscr{A}_{D_j,\lambda} \le m|D|^{-\frac{2r+s-1/2}{2r+s}}+\sqrt{C_0m}|D|^{-\frac{r}{2r+s}} \qquad \forall j=1,\dots,m. \tag{55}$$

But (8) implies

$$m|D|^{-\frac{4r+2s-1}{4r+2s}} \leq \sqrt{m}|D|^{-\frac{r}{2r+s}}.$$

So

$$\mathscr{A}_{D_j,\lambda} \leq (\sqrt{C_0}+1)\sqrt{m}|D|^{-\frac{r}{2r+s}}, \qquad \forall j=1,\ldots,m, \tag{56}$$

and

$$\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}} \leq (\sqrt{C_0}+1)\sqrt{m}|D|^{-\frac{r-1/2}{2r+s}}, \qquad \forall j=1,\ldots,m.$$

Hence $r>1/2$ together with (8) gives

$$\left(\frac{\mathscr{A}_{D_j,\lambda}}{\sqrt{\lambda}}\right)^2+1 \leq (\sqrt{C_0}+1)^2+1, \qquad \forall j=1,\ldots,m.$$

Then,

$$\widetilde{A_{D,\lambda}} \leq [(\sqrt{C_0}+1)^2+1](\sqrt{C_0}+1)^2\sqrt{m}|D|^{-\frac{r-1/2}{2r+s}}\sqrt{m}|D|^{-\frac{r}{2r+s}}. \tag{57}$$

Plugging (8) and (57) into (5) and noting

$$\log^4\frac{12m}{\delta} \leq 16(\log^4\frac{12}{\delta}+\log^4 m) \leq 16\log^4\frac{12}{\delta}(\log^3 m+1) \tag{58}$$

and

$$(\log^4|D|+1)(\log|D|+1) \leq 2(\log^5|D|+1),$$

we have with confidence $1-\delta$,

$$\begin{aligned}\|\overline{f}_{t,D}-f_\rho\|_\rho &\leq C_5\left\{|D|^{-\frac{r}{2r+s}}+|D|^{-\frac{r}{2r+s}}m|D|^{-\frac{r-1/2}{2r+s}}(\log^5|D|+1)+|D|^{-\frac{r}{2r+s}}\right\}\log^4\frac{12}{\delta}\\ &\leq 3C_5|D|^{-\frac{r}{2r+s}}\log^4\frac{12}{\delta},\end{aligned}$$

where

$$C_5 := 32C[(\sqrt{C_0}+1)^2+1](\sqrt{C_0}+1)^2.$$

This completes the proof of Corollary 1. □

Proof of Corollary 2. Applying the formula (9) for nonnegative random variables to $\xi=\|f_{D,\lambda}-f_\rho\|_\rho^2$ and use the bound

$$\text{Prob}\,[\xi>u]=\text{Prob}\left[\xi^{\frac{1}{2}}>u^{\frac{1}{2}}\right] \leq 12\exp\left\{-(C')^{-1/4}N^{\frac{r}{8r+4s}}u^{\frac{1}{8}}\right\}$$

for $u\geq(C'\log^4 12)^2|D|^{-2r/(2r+s)}$ derived from Corollary 1. We find

$$E\left[\|f_{D,\lambda}-f_\rho\|_\rho^2\right] \leq (C'\log^4 12)^2|D|^{-2r/(2r+s)}+12\int_0^\infty \exp\left\{-(C')^{-1/4}N^{\frac{r}{8r+4s}}u^{\frac{1}{8}}\right\}du$$

which equals $(96+\log^8 12)(C')^2|D|^{-\frac{2r}{2r+s}}\int_0^\infty u^{8-1}\exp\{-u\}\,du$. Due to $\int_0^\infty u^{d-1}\exp\{-u\}\,du=\Gamma(d)$ for arbitrary $d>0$, we have

$$E[\|f_{D,\lambda}-f_\rho\|_\rho^2] \leq (96+\log^8 12)(C')^2 7!|D|^{-\frac{2r}{2r+s}}.$$

This completes the proof of Corollary 2. □

To prove Corollary 3, we need the following Borel-Cantelli Lemma [11, page 262]. The Borel-Cantelli Lemma asserts for a sequence $\{\eta_n\}_n$ of events that if the sum of the probabilities is finite $\sum_{n=1}^{\infty} \text{Prob}[\eta_n] < \infty$, then the probability that infinitely many of them occur is 0.

Lemma 3 *Let $\{\eta_n\}$ be a sequence of events in some probability space and $\{\varepsilon_n\}$ be a sequence of positive numbers satisfying $\lim_{n\to\infty} \varepsilon_n = 0$. If*

$$\sum_{n=1}^{\infty} Prob[|\eta_n - \eta| > \varepsilon_n] < \infty,$$

then η_n converges to η almost surely.

Proof of Corollary 3. Let $N := |D|$ and $\delta = \delta_N = N^{-2}$ in Corollary 1. Set $\Psi_N = N^{-\frac{r}{2r+s}}$. By Corollary 1, if $t = \left\lceil |D|^{\frac{1}{2r+s}} \right\rceil$ and (8) holds, then for any N and $\varepsilon > 0$,

$$\text{Prob}\left[\Psi_N^{-1+\varepsilon}\|\overline{f}_{t,D} - f_\rho\|_\rho > C'\Psi_N^{\varepsilon}\left(\log\frac{12}{\delta_N}\right)^4\right] \le \delta_N.$$

Denote $\mu_N = C'\Psi_N^{\varepsilon}\left(\log\frac{12}{\delta_N}\right)^4$. Obviously,

$$\sum_{N=2}^{\infty} \text{Prob}\left[\Psi_N^{-1+\varepsilon}\|\overline{f}_{t,D} - f_\rho\|_\rho > \mu_N\right] \le \sum_{N=2}^{\infty} \delta_N < \infty$$

and $\mu_N \to 0$ when $N \to 0$. Then our conclusion follows from Lemma 3. This completes the proof of Corollary 3. □

Proof of Theorem 2. It follows from (31), (50) and (52) that

$$\begin{aligned}\|\overline{f}_{t,D^*} - f_\rho\|_\rho \le{}& \|f_t - f_\rho\|_\rho + C_1(\mathscr{R}_{D^*,\lambda} + \mathscr{P}_{D^*,\lambda}) \\ &+ C_3\log(t+1)\max_{1\le j\le m}\frac{\mathscr{Q}^2_{D_j^*,\lambda}\mathscr{R}_{D_j^*,\lambda}(\mathscr{P}_{D_j^*,\lambda} + \mathscr{R}_{D_j^*,\lambda})}{\sqrt{\lambda}}.\end{aligned}$$

From the definitions of D^*, we obtain

$$\hat{f}_{K,D^*} = \frac{1}{|D^*|}\sum_{(x_i^*,y_i^*)\in D^*} y_i^* K_{x_i^*} = \frac{1}{|D^*|}\sum_{(x_i,y_i)\in D}\frac{|D^*|}{|D|} y_i K_{x_i} = \hat{f}_{K,D},$$

and

$$\hat{f}_{K,D_j^*} = \frac{1}{|D_j^*|}\sum_{(x_i^*,y_i^*)\in D_j^*} y_i^* K_{x_i^*} = \frac{1}{|D_j^*|}\sum_{(x_i,y_i)\in D_j}\frac{|D_j^*|}{|D_j|} y_i K_{x_i} = \hat{f}_{K,D_j}.$$

Then we obtain

$$\mathscr{P}_{D,\lambda} = \mathscr{P}_{D^*,\lambda}, \qquad \text{and} \qquad \mathscr{P}_{D_j,\lambda} = \mathscr{P}_{D_j^*,\lambda}, \qquad \forall j = 1,\dots,m.$$

Thus,

$$\|\overline{f}_{t,D^*} - f_\rho\|_\rho \le \|f_t - f_\rho\|_\rho + C_1(\mathscr{R}_{D^*,\lambda} + \mathscr{P}_{D,\lambda}) + C_3 \log(t+1) \max_{1\le j\le m} \frac{\mathscr{Q}^2_{D_j^*,\lambda}\mathscr{R}_{D_j^*,\lambda}(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j^*,\lambda})}{\sqrt{\lambda}}. \quad (59)$$

A similar argument as that in the proof of Theorem 1 together with Lemma 2 and $\mathscr{A}_{D^*,\lambda} \le \mathscr{A}_{D,\lambda}$ yields that with confidence at least $1-\delta/2$, there holds

$$C_1(\mathscr{R}_{D^*,\lambda} + \mathscr{P}_{D,\lambda}) \le C_2 \mathscr{A}_{D,\lambda} \log\big(8/\delta\big) \quad (60)$$

and with confidence $1-\delta/2$, there holds

$$\max_{1\le j\le m} \frac{\mathscr{Q}^2_{D_j^*,\lambda}\mathscr{R}_{D_j^*,\lambda}(\mathscr{P}_{D_j,\lambda} + \mathscr{R}_{D_j^*,\lambda})}{\sqrt{\lambda}} \le C_4 \widetilde{\mathscr{A}_{D,D^*,\lambda}} \log^4\big(12m/\delta\big), \quad (61)$$

where $\widetilde{\mathscr{A}_{D,D^*,\lambda}}$ is defined by (11). Plugging (42), (60) and (61) into (59), we obtain with confidence at least $1-\delta$, there holds

$$\|\overline{f}_{t,D^*} - f_\rho\|_\rho \le C\left[t^{-r} + \mathscr{A}_{D,\lambda}\log\big(8/\delta\big) + \widetilde{\mathscr{A}_{D,D^*,\lambda}}\log^4\big(12m/\delta\big)\log(t+1)\right].$$

This completes the proof of Theorem 2. □

Proof of Corollary 4. Since $\lambda = 1/t$, $|D_1^*| = \cdots = |D_m^*|$ and $t = \left\lceil |D|^{\frac{1}{(2r+s)}} \right\rceil$, we have

$$\mathscr{A}_{D_j^*,\lambda} \le m|D^*|^{-1}|D|^{\frac{1}{4r+2s}} + \sqrt{C_0 m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}}, \qquad \forall\ j = 1,\dots,m. \quad (62)$$

This means

$$\lambda^{-1/2}\mathscr{A}_{D_j^*,\lambda} \le m|D^*|^{-1}|D|^{\frac{1}{2r+s}} + \sqrt{C_0 m}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}}.$$

Due to (12), $r > 1/2$ and $|D| \le |D^*|$, we have

$$\left(\lambda^{-1/2}\mathscr{A}_{D_j^*,\lambda}\right)^2 + 1 \le (\sqrt{C_0}+1)^2 + 1, \qquad \forall j = 1,\dots,m. \quad (63)$$

Furthermore, based on (12), we have

$$m|D^*|^{-1}|D|^{\frac{1}{4r+2s}} \le \sqrt{m}|D^*|^{-1/2}|D|^{\frac{s}{4r+2s}}.$$

Therefore

$$\lambda^{-1/2}\mathscr{A}_{D_j^*,\lambda} \le (\sqrt{C_0}+1)\sqrt{m}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}}. \quad (64)$$

Plugging (63), (64) and (55) into (11), we get

$$\widetilde{\mathscr{A}_{D,D^*,\lambda}} \le [(\sqrt{C_0}+1)^2+1](\sqrt{C_0}+1) \times \left[m\sqrt{m}|D|^{-\frac{2r+s-1/2}{2r+s}}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}} + \sqrt{C_0}m|D|^{-\frac{r}{2r+s}}|D^*|^{-1/2}|D|^{\frac{s+1}{4r+2s}}\right]. \quad (65)$$

Inserting (54), (65) and (58) into (10), we obtain from (12) and $t = \lceil |D|^{1/(2r+s)} \rceil$ that with confidence $1-\delta$

$$\|\overline{f}_{t,D^*} - f_\rho\|_\rho \le C'|D|^{-\frac{r}{2r+s}}\log^4\frac{12}{\delta},$$

where C' is the constant in Corollary 1. This completes the proof of Corollary 4. □

References

1. T. Ameet, Spark Meetup: MLbase, Distributed Machine Learning with Spark. slideshare.net. Spark User Meetup, San Francisco, California. 6 August (2013).
2. M. Balcan, A. Blum, S. Fine, Y. Mansour, Distributed learning, communication complexity, and privacy, COLT, 23, (2012).
3. F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, J. Complex., 23, 52-72 (2007).
4. G. Blanchard, N. Krämer, Optimal learning rates for kernel conjugate gradient regression, Advances in Neural Information Processing Systems, 226-234 (2010).
5. G. Blanchard, N. Mücke, Parallelizing spectral algorithms for kernel learning, arXiv preprint arXiv:1610.07487 (2016).
6. A. Caponnetto, E. DeVito, Optimal rates for the regularized least squares algorithm, Found. Comp. Math., 7, 331-368 (2007).
7. A. Caponnetto, Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. Anal. Appl., 8, 161-183 (2010).
8. X. Chang, S. B. Lin, D. X. Zhou, Distributed semi-supervised learning with kernel ridge regression, J. Mach. Learn. Res., minor revision to be made (2016).
9. J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51, 107-113 (2008).
10. L. H. Dicker, D. P. Foster, D. Hsu, Kernel ridge vs. principal component regression: minimax bounds and adaptability of regularization operators, arXiv preprint arXiv:1605.08839 (2016).
11. R. M. Dudley, Real Analysis and Probability, Cambridge Studies in Advanced Mathematics **74**, Cambridge University Press, Cambridge (2002).
12. H. W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Mathematics and its Applications **375**, Kluwer Academic Publishers Group, Dordrecht (1996).
13. D. Gillick, A. Faria, J. DeNero, Mapreduce: Distributed computing for machine learning, Berkley, December 18 (2006).
14. Z. C. Guo, S. B. Lin and D. X. Zhou, Distributed learning with spectral algorithms. Inverse Problems, minor revisision under review (2016).
15. T. Hu, J. Fan, Q. Wu, D. X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl., 13, 437-455 (2015).
16. J. H. Lin, D. X. Zhou. Learning theory of randomized Kaczmarz algorithm. J. Mach. Learn. Res., 16, 3341-3365 (2015).
17. S. B. Lin, X. Guo, D. X. Zhou, Distributed learning with regularized least squares, J. Mach. Learn. Res., revision under review (arXiv 1608.03339) (2016).
18. L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, A. Verri, Spectral algorithms for supervised learning, Neural Comput., 20, 1873-1897 (2008).
19. G. Mann, R. McDonald, M. Mohri, N. Silberman, D. Walker, Efficient large-scale distributed training of conditional maximum entropy models, NIPS, 1231-1239 (2009).
20. G. Raskutti, M. Wainwright, B. Yu, Early stopping and non-parametric regression: an optimal data-dependent stopping rule, J. Mach. Learn. Res., 15, 335-366 (2014).
21. O. Shamir, N. Srebro, Distributed stochastic optimization and learning, In 52nd Annual Allerton Conference on Communication, Control and Computing, (2014).
22. S. Smale, D. X. Zhou, Learning theory estimates via integral operators and their approximations. Constr. Approx., 26, 153-172 (2007).
23. S. van de Vaart, J. Wellner, Weak Convergence and Empirical Process: with Applications to Statistics, Springer Series in Statistics, Springer, New York (1996).
24. Q. Wu, D. X. Zhou, Learning with sample dependent hypothesis space. Comput. Math. Appl., 56, 2896-2907 (2008).
25. Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constr. Approx., 26, 289-315 (2007).
26. X. D. Wu, X. Q. Zhu, G. Q. Wu and W. Ding. Data mining with big data. IEEE Trans. Know. Data Engin., 26, 97-107 (2014).
27. Y. C. Zhang, J. Duchi, M. Wainwright, Communication-efficient algorithms for statistical optimization, J. Mach. Learn. Res., 14, 3321-3363 (2013).
28. Y. C. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, J. Mach. Learn. Res., 16, 3299-3340 (2015).

29. D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory, 49, 1743-1752 (2003).
30. Z. H. Zhou, N. V. Chawla, Y. Jin, G. J. Williams, Big data opportunities and challenges: Discussions from data analytics perspectives, IEEE Computational Intelligence Magazine, 9, 62-74 (2014).