Journal of Machine Learning Research 18 (2017) 1-31 Submitted 11/15; Revised 6/17; Published 9/17

Distributed Learning with Regularized Least Squares

Shao-Bo Lin SBLIN1983@GMAIL.COM
Department of Mathematics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
Xin Guo X.GUO@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Editor: Ingo Steinwart

Abstract

We study distributed learning with the least squares regularization scheme in a reproducing kernel Hilbert space (RKHS). By a divide-and-conquer approach, the algorithm partitions a data set into disjoint data subsets, applies the least squares regularization scheme to each data subset to produce an output function, and then takes an average of the individual output functions as a final global estimator or predictor. We show with error bounds and learning rates in expectation in both the L^2-metric and RKHS-metric that the global output function of this distributed learning is a good approximation to the algorithm processing the whole data in one single machine. Our derived learning rates in expectation are optimal and stated in a general setting without any eigenfunction assumption. The analysis is achieved by a novel second order decomposition of operator differences in our integral operator approach. Even for the classical least squares regularization scheme in the RKHS associated with a general kernel, we give the best learning rate in expectation in the literature.

Keywords: Distributed learning, divide-and-conquer, error analysis, integral operator, second order decomposition.

1. Introduction and Distributed Learning Algorithms

In the era of big data, the rapid expansion of computing capacities in automatic data generation and acquisition brings data of unprecedented size and complexity, and raises a series of scientific challenges such as storage bottleneck and algorithmic scalability (Zhou et al., 2014). To overcome the difficulty, some approaches for generating scalable approximate algorithms have been introduced in the literature such as low-rank approximations of kernel matrices for kernel principal component analysis (Schölkopf et al., 1998; Bach, 2013), incomplete Cholesky decomposition (Fine, 2002), early-stopping of iterative optimization algorithms for gradient descent methods (Yao et al., 2007; Raskutti et al., 2014; Lin et al., 2016), and greedy-type algorithms. Another method proposed recently is distributed learning based on a divide-and-conquer approach and a particular learning algorithm implemented in individual machines (Zhang et al., 2015; Shamir and Srebro, 2014). This method

produces distributed learning algorithms consisting of three steps: partitioning the data into disjoint subsets, applying a particular learning algorithm implemented in an individual machine to each data subset to produce an individual output (function), and synthesizing a global output by utilizing some average of the individual outputs. This method can significantly reduce computing time and lower single-machine memory requirements. For practical applications in medicine, finance, business and some other areas, the data are often stored naturally across multiple servers in a distributive way and are not combined for reasons of protecting privacy and avoiding high costs. In such situations, the first step of data partitioning is not needed. It has been observed in many practical applications that when the number of data subsets is not too large, the divide-and-conquer approach does not cause noticeable loss of performance, compared with the learning scheme which processes the whole data on a single big machine. Theoretical attempts have been recently made in (Zhang et al., 2013, 2015) to derive learning rates for distributed learning with least squares regularization schemes under certain assumptions.

This paper aims at error analysis of the *distributed learning* with regularized least squares and its approximation to the algorithm processing the whole data in one single machine. Recall (Cristianini and Shawe-Taylor, 2000; Evgeniou et al., 2000) that in a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ induced by a Mercer kernel K on an input metric space $\mathcal{X}$, with a sample $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = \mathbb{R}$ is the output space, the least squares regularization scheme can be stated as

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

Here $\lambda > 0$ is a regularization parameter and $|D| =: N$ is the cardinality of D. This learning algorithm is also called kernel ridge regression in statistics and has been well studied in learning theory. See e.g. (De Vito et al., 2005; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Bauer et al., 2007; Smale and Zhou, 2007; Steinwart and Christmann, 2008). The regularization scheme (1) can be explicitly solved by using a standard matrix inversion technique, which requires costs of $\mathcal{O}(N^3)$ in time and $\mathcal{O}(N^2)$ in memory. However, this matrix inversion technique may not conquer challenges on storages or computations arising from big data.

The distributed learning algorithm studied in this paper starts with partitioning the data set D into m disjoint subsets $\{D_j\}_{j=1}^m$. Then it assigns each data subset D_j to one machine or processor to produce a local estimator $f_{D_j,\lambda}$ by the least squares regularization scheme (1). Finally, these local estimators are communicated to a central processor, and a global estimator $\overline{f}_{D,\lambda}$ is synthesized by taking a weighted average

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

of the local estimators $\{f_{D_j,\lambda}\}_{j=1}^m$. This algorithm has been studied with a matrix analysis approach in (Zhang et al., 2015) where some error analysis has been conducted under some eigenfunction assumptions for the integral operator associated with the kernel, presenting error bounds in expectation.

In this paper we shall use a novel integral operator approach to prove that $\overline{f}_{D,\lambda}$ is a good approximation of $f_{D,\lambda}$. We present a representation of the difference $\overline{f}_{D,\lambda} - f_{D,\lambda}$ in terms of empirical integral operators, and analyze the error $\overline{f}_{D,\lambda} - f_{D,\lambda}$ in expectation without any eigenfunction assumptions. As a by-product, we present the best learning rates in expectation for the least squares regularization scheme (1) in a general setting, which surprisingly has not been done for a general kernel in the literature (see detailed comparisons below).

2. Main Results

Our analysis is carried out in the standard least squares regression framework with a Borel probability measure ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where the input space $\mathcal{X}$ is a compact metric space. The sample D is independently drawn according to ρ. The Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines an integral operator L_K on $\mathcal{H}_K$ as

$$L_K(f) = \int_{\mathcal{X}} K_x f(x) d\rho_X, \qquad f \in \mathcal{H}_K, \tag{3}$$

where K_x is the function $K(\cdot, x)$ in $\mathcal{H}_K$ and ρ_X is the marginal distribution of ρ on $\mathcal{X}$.

2.1 Error Bounds for the Distributed Learning Algorithm

Our error bounds in expectation for the distributed learning algorithm (2) require the uniform boundedness condition for the output y, that is, for some constant $M > 0$, there holds $|y| \le M$ almost surely. Our bounds are stated in terms of the approximation error

$$\|f_\lambda - f_\rho\|_\rho, \tag{4}$$

where f_λ is the data-free limit of (1) defined by

$$f_\lambda = \arg\min_{f \in \mathcal{H}_K} \left\{ \int_{\mathcal{Z}} (f(x) - y)^2 d\rho + \lambda \|f\|_K^2 \right\}, \tag{5}$$

$\|\cdot\|_\rho$ denotes the norm of $L^2_{\rho_X}$, the Hilbert space of square integrable functions with respect to ρ_X, and f_ρ is the regression function (conditional mean) of ρ defined by

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

with $\rho(\cdot|x)$ being the conditional distribution of ρ induced at $x \in \mathcal{X}$.

Since K is continuous, symmetric and positive semidefinite, L_K is a compact positive operator of trace class and $L_K + \lambda I$ is invertible. Define a quantity measuring the complexity of $\mathcal{H}_K$ with respect to ρ_X, the *effective dimension* (Zhang, 2005), to be the trace of the operator $(L_K + \lambda I)^{-1} L_K$ as

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left((L_K + \lambda I)^{-1} L_K\right), \qquad \lambda > 0. \tag{6}$$

In Section 6 we shall prove the following first main result of this paper concerning error bounds in expectation of $\overline{f}_{D,\lambda} - f_{D,\lambda}$ in $\mathcal{H}_K$ and in $L^2_{\rho_X}$. Denote $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$.

Theorem 1 *Assume $|y| \le M$ almost surely. If $|D_j| = \frac{N}{m}$ for $j = 1, \ldots, m$, then we have*

$$E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho \le C_\kappa \left\{ M\sqrt{\lambda m}\mathcal{C}_m \left\{1 + \sqrt{m\mathcal{C}_m}\right\} + \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} m \left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \sqrt{m}\mathcal{C}_m\right)\right\}$$

and

$$E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_K \le C_\kappa \left\{ M\sqrt{m}\mathcal{C}_m \left\{1 + \sqrt{m\mathcal{C}_m}\right\} + \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} m \left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \sqrt{m}\mathcal{C}_m\right)\right\},$$

where C_κ is a constant depending only on κ, and $\mathcal{C}_m$ is the quantity given in terms of $m, N, \lambda, \mathcal{N}(\lambda)$ by

$$\mathcal{C}_m := \frac{m}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}.$$

To derive explicit learning rates from the general error bounds in Theorem 1, one can quantify the increment of $\mathcal{N}(\lambda)$ by the following *capacity assumption*, a characteristic of the complexity of the hypothesis space (Caponnetto and De Vito, 2007; Blanchard and Krämer, 2010),

$$\mathcal{N}(\lambda) \le c\lambda^{-\beta}, \qquad \forall \lambda > 0 \tag{7}$$

for some constants $\beta > 0$ and $c > 0$. Let $\{(\lambda_\ell, \varphi_\ell)\}_\ell$ be a set of normalized eigenpairs of L_K on $\mathcal{H}_K$ with $\{\varphi_\ell\}_\ell$ being an orthonormal basis of $\mathcal{H}_K$ and $\{\lambda_\ell\}_\ell$ arranged in a non-increasing order. A sufficient condition for the capacity assumption (7) with $0 < \beta < 1$ is $\lambda_\ell = O(\ell^{-1/\beta})$, which can be found in, e.g. Caponnetto and De Vito (2007).

Remark 2 *The sufficient condition $\lambda_\ell = O(\ell^{-1/\beta})$ with the index $\beta = \frac{d}{2\tau} < 1$ is satisfied by the Sobolev space $W^\tau(B(\mathbb{R}^d))$ with the smoothness index $\tau > d/2$ on a ball $B(\mathbb{R}^d)$ of the Euclidean space $\mathbb{R}^d$ when the marginal distribution ρ_X is the uniform distribution on $B(\mathbb{R}^d)$, see (Steinwart et al., 2009; Edmunds and Triebel, 1996).*

Condition (7) with $\beta = 1$ always holds true with the choice of the constant $c = \kappa^2$. In fact, the eigenvalues of the operator $(L_K + \lambda I)^{-1} L_K$ are $\{\frac{\lambda_\ell}{\lambda_\ell + \lambda}\}_\ell$. So its trace is given by $\mathcal{N}(\lambda) = \sum_\ell \frac{\lambda_\ell}{\lambda_\ell + \lambda} \le \sum_\ell \frac{\lambda_\ell}{\lambda} = \frac{\mathrm{Tr}(L_K)}{\lambda} \le \kappa^2 \lambda^{-1}$.

In the existing literature on learning rates for the classical least squares algorithm (1), the regularization parameter λ is often taken to satisfy the restriction $\frac{\mathcal{N}(\lambda)}{N\lambda} = O(1)$ as in (Caponnetto and De Vito, 2007) up to a logarithmic factor or in (Steinwart et al., 2009) under some assumptions on (K, ρ_X) (see (14) below). Here, to derive learning rates for $E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho$ with $m \ge 1$ corresponding to the distributed learning algorithm (2), we consider λ to satisfy the following restriction with some constant $C_0 > 0$,

$$0 < \lambda \le C_0 \quad \text{and} \quad \frac{m\mathcal{N}(\lambda)}{N\lambda} \le C_0. \tag{8}$$

Corollary 3 *Assume $|y| \le M$ almost surely. If $|D_j| = \frac{N}{m}$ for $j = 1, \ldots, m$, and λ satisfies (8), then we have*

$$E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho \le \widetilde{C}_\kappa \sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} \left\{ M\sqrt{\lambda} + \frac{m\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} \right\}$$

and

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_K \le \widetilde{C}_\kappa\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}}\left\{M+\frac{m\|f_\rho-f_\lambda\|_\rho}{\sqrt{N\lambda}}\right\},$$

where $\widetilde{C}_\kappa$ *is a constant depending only on* κ, C_0, *and the largest eigenvalue of* L_K.

In the special case that $f_\rho \in \mathcal{H}_K$, the approximation error can be bounded as $\|f_\lambda - f_\rho\|_\rho \le \|f_\rho\|_K\sqrt{\lambda}$. A more general *regularity condition* can be imposed for the regression function as

$$f_\rho = L_K^r(g_\rho) \quad \text{for some } g_\rho \in L^2_{\rho_X},\ r>0, \tag{9}$$

where the integral operator L_K is regarded as a compact positive operator on $L^2_{\rho_X}$ and its rth power L_K^r is well defined for any $r>0$. The condition (9) means f_ρ lies in the range of L_K^r, and the special case $f_\rho \in \mathcal{H}_K$ corresponds to the choice $r=1/2$. It implies $\|f_\lambda - f_\rho\|_\rho \le \lambda^r\|g_\rho\|_\rho$ by Lemma 21 below. Thus, under condition (9), we obtain from Corollary 3, by choosing λ to minimize the order of the error bound subject to the restriction (8), the following nice convergence rates for the error $\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|$ of the distributed learning algorithm (2).

Corollary 4 *Assume regularity condition (9) for some* $0<r\le 1$, *capacity assumption (7) with* $\beta=\frac{1}{2\alpha}$ *for some* $\alpha>0$, *and* $|y|\le M$ *almost surely. If* $|D_j|=\frac{N}{m}$ *for* $j=1,\dots,m$ *with*

$$m \le N^{\frac{\alpha\max\{2r,1\}+\frac{1}{2}+2\alpha(r-1)}{2\alpha\max\{2r,1\}+1+2\alpha(r-1)}}, \tag{10}$$

and $\lambda=\left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha\max\{2r,1\}+1}}$, *then we have*

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho \le \widetilde{C}_{\kappa,c,r}\frac{1}{\sqrt{m}}\left(\frac{m}{N}\right)^{\frac{\alpha\max\{2r,1\}}{2\alpha\max\{2r,1\}+1}}$$

and

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_K \le \widetilde{C}_{\kappa,c,r}\frac{1}{\sqrt{m}}\left(\frac{m}{N}\right)^{\frac{\alpha\max\{2r-1,0\}}{2\alpha\max\{2r,1\}+1}},$$

where $\widetilde{C}_{\kappa,c,r}$ *is a constant independent of* N *or* m.

In particular, when $f_\rho\in\mathcal{H}_K$ *and* $m\le N^{\frac{1}{2+2\alpha}}$, *taking* $\lambda=\left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$ *yields the rates* $E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho \le \widetilde{C}_{\kappa,c,r}N^{-\frac{\alpha}{2\alpha+1}}m^{-\frac{1}{4\alpha+2}}$ *and* $E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_K \le \widetilde{C}_{\kappa,c,r}\frac{1}{\sqrt{m}}$.

Remark 5 *In Corollary 4, we present learning rates in both* $\mathcal{H}_K$ *and* $L^2_{\rho_X}$ *norms. The rates with respect to the* $L^2_{\rho_X}$ *norm provide estimates for the generalization ability of the algorithm for regression. The rates with respect to the* $\mathcal{H}_K$ *norm give error estimates with respect to the uniform metric due to the relation* $\|f\|_\infty\le\kappa\|f\|_K$, *and might be used to solve some mismatch problems in learning theory where the generalization ability of learning algorithms is measured with respect to a probability measure* μ *different from* ρ_X.

Remark 6 *The learning rates in Corollary 4 are stated for the norms of the difference* $\overline{f}_{D,\lambda}-f_{D,\lambda}$ *which reflects the variance of the distributed learning algorithm (2). These rates decrease as* m *increases (subject to the restriction (10)) and thereby the regularization parameter* λ *increases, which is different from the learning rates presented for* $E\left\|\overline{f}_{D,\lambda}-f_\rho\right\|_\rho$ *in (Zhang et al., 2015). To derive learning rates for* $E\left\|\overline{f}_{D,\lambda}-f_\rho\right\|_\rho$ *by our analysis, the regularization parameter* λ *should be independent of* m, *as shown in Corollary 11 below.*

2.2 Optimal Learning Rates for Least Squares Regularization Scheme

The second main result of this paper is optimal learning rates in expectation for the least squares regularization scheme (1). We can even relax the uniform boundedness to a *moment condition* that for some constant $p \geq 1$,

$$\sigma_\rho^2 \in L^p_{\rho_X}, \tag{11}$$

where σ_ρ^2 is the conditional variance defined by $\sigma_\rho^2(x) = \int_{\mathcal{Y}} (y - f_\rho(x))^2 \, d\rho(y|x)$.

The following learning rates in expectation for the least squares regularization scheme (1) will be proved in Section 5. The existence of f_λ is ensured by $E[y^2] < \infty$.

Theorem 7 *Assume $E[y^2] < \infty$ and moment condition (11) for some $1 \leq p \leq \infty$. Then we have*

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] \leq \left(2 + 56\kappa^4 + 57\kappa^2\right)(1+\kappa)\left(1 + \frac{1}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right)$$
$$\left\{\left(1 + \frac{1}{\sqrt{N\lambda}}\right)\|f_\lambda - f_\rho\|_\rho + \sqrt{\left\|\sigma_\rho^2\right\|_p}\left(\frac{\mathcal{N}(\lambda)}{N}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{N\lambda}\right)^{\frac{1}{2p}}\right\}. \tag{12}$$

If the parameters satisfy $\frac{\mathcal{N}(\lambda)}{N\lambda} = O(1)$, we have the following explicit bound.

Corollary 8 *Assume $E[y^2] < \infty$ and moment condition (11) for some $1 \leq p \leq \infty$. If λ satisfies (8) with $m = 1$, then we have*

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] = O\left(\|f_\lambda - f_\rho\|_\rho + \left(\frac{\mathcal{N}(\lambda)}{N}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{N\lambda}\right)^{\frac{1}{2p}}\right).$$

In particular, if $p = \infty$, that is, the conditional variances are uniformly bounded, then

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] = O\left(\|f_\lambda - f_\rho\|_\rho + \sqrt{\frac{\mathcal{N}(\lambda)}{N}}\right).$$

In particular, when regularity condition (9) is satisfied, we have the following learning rates in expectation.

Corollary 9 *Assume $E[y^2] < \infty$, moment condition (11) for some $1 \leq p \leq \infty$, and regularity condition (9) for some $0 < r \leq 1$. If the capacity assumption $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$ holds with some $\alpha > 0$, then by taking $\lambda = N^{-\frac{2\alpha}{2\alpha\max\{2r,1\}+1}}$ we have*

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] = O\left(N^{-\frac{2r\alpha}{2\alpha\max\{2r,1\}+1} + \frac{1}{2p}\frac{2\alpha-1}{2\alpha\max\{2r,1\}+1}}\right).$$

In particular, when $p = \infty$ (the conditional variances are uniformly bounded), we have

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] = O\left(N^{-\frac{2r\alpha}{2\alpha\max\{2r,1\}+1}}\right).$$

Remark 10 *It was shown in (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Bauer et al., 2007) that under the condition of Corollary 9 with $p = \infty$ and $r \in [\frac{1}{2}, 1]$, the best learning rate, called minimax lower rate of convergence, for learning algorithms with output functions in $\mathcal{H}_K$ is $O\left(N^{-\frac{2r\alpha}{4\alpha r+1}}\right)$. So the convergence rate we obtain in Corollary 9 is optimal.*

Combining bounds for $\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho$ and $\|f_{D,\lambda} - f_\rho\|_\rho$, we shall prove in Section 6 the following learning rates in expectation for the distributed learning algorithm (2) for regression.

Corollary 11 *Assume $|y| \le M$ almost surely and regularity condition (9) for some $\frac{1}{2} < r \le 1$. If the capacity assumption $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$ holds with some $\alpha > 0$, $|D_j| = \frac{N}{m}$ for $j = 1, \ldots, m$, and m satisfies the restriction*

$$m \le N^{\frac{2\alpha(2r-1)}{4\alpha r+1}}, \tag{13}$$

then by taking $\lambda = N^{-\frac{2\alpha}{4\alpha r+1}}$, we have

$$E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_\rho\right] = O\left(N^{-\frac{2\alpha r}{4\alpha r+1}}\right).$$

Remark 12 *Corollary 11 shows that distributed learning with the least squares regularization scheme in a RKHS can achieve the optimal learning rates in expectation, provided that m satisfies the restriction (13). It should be pointed out that our error analysis is carried out under regularity condition (9) with $1/2 < r \le 1$ while the work in (Zhang et al., 2015) focused on the case with $r = 1/2$. When r approaches $1/2$, the number m of local processors under the restriction (13) reduces to 1, which corresponds to the non-distributed case. In our follow-up work, we will consider to relax the restriction (13) in a semi-supervised learning framework by using additional unlabelled data, as done in (Caponnetto and Yao, 2010). The main contribution of our analysis for distributed learning in this paper is to remove an eigenfunction assumption in (Zhang et al., 2015) by using a novel second order decomposition for a difference of operator inverses.*

Remark 13 *In Corollary 11 and Corollary 9, the choice of the regularization parameter λ is independent of the number m of local processors. This is consistent with the results in (Zhang et al., 2015). There have been several approaches for selecting the regularization parameter λ in regularization schemes in the literature including cross-validation (Györfy et al., 2002; Blanchard and Krämer, 2016) and the balancing principle (De Vito et al., 2010). For practical applications of distributed learning algorithms, how to choose λ and m (except the situations when the data are stored naturally in a distributive way) is crucial. Though we only consider the theoretical topic of error analysis in this paper, it would be interesting to develop parameter selection methods for distributed learning.*

3. Comparisons and Discussion

The least squares regularization scheme (1) is a classical algorithm for regression and has been extensively investigated in statistics and learning theory. There is a vast literature on

this topic. Here for a general kernel beyond those for the Sobolev spaces, we compare our results with the best learning rates in the existing literature.

Under the assumption that the orthogonal projection $f_{\mathcal{H}}$ of f_ρ in $L^2_{\rho_X}$ onto the closure of $\mathcal{H}_K$ satisfies regularity condition (9) for some $\frac{1}{2} \le r \le 1$, and that the eigenvalues $\{\lambda_i\}_i$ of L_K satisfy $\lambda_i \approx i^{-2\alpha}$ with some $\alpha > 1/2$, it was proved in (Caponnetto and De Vito, 2007) that

$$\lim_{\tau\to\infty} \limsup_{N\to\infty} \sup_{\rho\in\mathcal{P}(\alpha)} \text{Prob}\left[\|f_{D,\lambda_N} - f_{\mathcal{H}}\|_\rho^2 > \tau\lambda_N^{2r}\right] = 0,$$

where

$$\lambda_N = \begin{cases} N^{-\frac{2\alpha}{4\alpha r+1}}, & \text{if } \frac{1}{2} < r \le 1, \\ \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{2\alpha+1}}, & \text{if } r = \frac{1}{2}, \end{cases}$$

and $\mathcal{P}(\alpha)$ denotes a set of probability measures ρ satisfying some moment decay condition (which is satisfied when $|y| \le M$). This learning rate in probability is stated with a limit as $\tau \to \infty$ and it has a logarithmic factor in the case $r = \frac{1}{2}$. In particular, to guarantee $\|f_{D,\lambda_N} - f_{\mathcal{H}}\|_\rho^2 \le \tau_\eta \lambda_N^{2r}$ with confidence $1-\eta$ by this result, one needs to restrict $N \ge N_\eta$ to be large enough and has the constant τ_η depending on η to be large enough. Using existing tools for error analysis including those in (Caponnetto and De Vito, 2007), we develop a novel second order decomposition technique for a difference of operator inverses in this paper, and succeed in deriving the optimal learning rate in expectation in Corollary 9 by removing the logarithmic factor in the case $r = \frac{1}{2}$.

Under the assumption that $|y| \le M$ almost surely, the eigenvalues $\{\lambda_i\}_i$ satisfying $\lambda_i \le a i^{-2\alpha}$ with some $\alpha > 1/2$ and $a > 0$, and for some constant $C > 0$, the pair (K, ρ_X) satisfying

$$\|f\|_\infty \le C \|f\|_K^{\frac{1}{2\alpha}} \|f\|_\rho^{1-\frac{1}{2\alpha}} \tag{14}$$

for every $f \in \mathcal{H}_K$, it was proved in (Steinwart et al., 2009) that for some constant $c_{\alpha,C}$ depending only on α and C, with confidence $1-\eta$, for any $0 < \lambda \le 1$,

$$\|\pi_M(f_{D,\lambda}) - f_\rho\|_\rho^2 \le 9\mathcal{A}_2(\lambda) + c_{\alpha,C} \frac{a^{1/(2\alpha)} M^2 \log(3/\eta)}{\lambda^{1/(2\alpha)} N}.$$

Here π_M is the projection onto the interval $[-M, M]$ defined (Chen et al., 2004; Wu et al., 2006) by

$$\pi_M(f)(x) = \begin{cases} M, & \text{if } f(x) > M, \\ f(x), & \text{if } |f(x)| \le M, \\ -M, & \text{if } f(x) < -M, \end{cases}$$

and $\mathcal{A}_2(\lambda)$ is the approximation error defined by

$$\mathcal{A}_2(\lambda) = \inf_{f\in\mathcal{H}_K} \left\{\|f - f_\rho\|_\rho^2 + \lambda\|f\|_K^2\right\}.$$

When $f_\rho \in \mathcal{H}_K$, $\mathcal{A}_2(\lambda) = O(\lambda)$ and the choice $\lambda_N = N^{\frac{2\alpha}{2\alpha+1}}$ gives a learning rate of order $\|f_{D,\lambda_N} - f_\rho\|_\rho = O\left(N^{-\frac{\alpha}{2\alpha+1}}\right)$. Here one needs to impose the condition (14) for the functions spaces $L^2_{\rho_X}$ and $\mathcal{H}_K$, and to take the projection onto $[-M, M]$. Our learning rates

in Corollary 9 do not require such a condition for the function spaces, nor do we take the projection. Let us mention the fact from (Steinwart et al., 2009; Mendelson and Neeman, 2010) that condition (14) is satisfied when $\mathcal{H}_K$ is a Sobolev space on a domain of a Euclidean space and ρ_X is the uniform distribution, or when the eigenfunctions $\{\varphi_\ell/\sqrt{\lambda_\ell} : \lambda_\ell > 0\}$ of L_K normalized in $L^2_{\rho_X}$ are uniformly bounded. Recall that $\|\varphi_\ell\|_{L^2_{\rho_X}} = \sqrt{\lambda_\ell}$ since $\lambda_\ell = \lambda_\ell \|\varphi_\ell\|_K^2 = \langle \lambda_\ell \varphi_\ell, \varphi_\ell\rangle_K = \langle L_K\varphi_\ell, \varphi_\ell\rangle_K$ equals

$$\langle \int_{\mathcal{X}} K_x \varphi_\ell(x) d\rho_X(x), \varphi_\ell\rangle_K = \int_{\mathcal{X}} \varphi_\ell(x)\varphi_\ell(x) d\rho_X(x) = \|\varphi_\ell\|^2_{L^2_{\rho_X}}.$$

Learning rates for the least squares regularization scheme (1) in the $\mathcal{H}_K$-metric have been investigated in (Smale and Zhou, 2007).

For the distributed learning algorithm (2) with subsets $\{D_j\}_{j=1}^m$ of equal size, under the assumption that for some constants $2 < k \le \infty$ and $A < \infty$, the eigenfunctions $\{\varphi_i\}_i$ satisfy

$$\begin{cases} \sup_{\lambda_i>0} E\left[\left|\frac{\varphi_i(x)}{\sqrt{\lambda_i}}\right|^{2k}\right] \le A^{2k}, & \text{when } k < \infty, \\ \sup_{\lambda_i>0} \left\|\frac{\varphi_i(x)}{\sqrt{\lambda_i}}\right\|_\infty \le A, & \text{when } k = \infty, \end{cases} \tag{15}$$

and that $f_\rho \in \mathcal{H}_K$ and $\lambda_i \le a i^{-2\alpha}$ for some $\alpha > 1/2$ and $a > 0$, it was proved in (Zhang et al., 2015) that for $\lambda = N^{-\frac{2\alpha}{2\alpha+1}}$ and m satisfying the restriction

$$m \le c_\alpha \begin{cases} \left(\frac{N^{\frac{2(k-4)\alpha-k}{2\alpha+1}}}{A^{4k}\log^k N}\right)^{\frac{1}{k-2}}, & \text{when } k < \infty \\ \frac{N^{\frac{2\alpha-1}{2\alpha+1}}}{A^4 \log N}, & \text{when } k = \infty \end{cases} \tag{16}$$

with a constant c_α depending only on α, there holds $E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_\rho^2\right] = O\left(N^{-\frac{2\alpha}{2\alpha+1}}\right)$. This interesting result was achieved by a matrix analysis approach for which the eigenfunction assumption (15) played an essential role.

It might be challenging to check the eigenfunction assumption (15) involving the integral operator $L_K = L_{K,\rho_X}$ for the pair (K, ρ_X). To our best knowledge, besides the case of finite dimensional RKHSs, there exist in the literature only three classes of pairs (K, ρ_X) proved satisfying this eigenfunction assumption: the first class (Steinwart et al., 2009) is the Sobolev reproducing kernels on Euclidean domains together with the unform measures ρ_X, the second (Williamson et al., 2001) is periodical kernels, and the third class can be constructed by a Mercer type expansion

$$K(x, y) = \sum_i \lambda_i \frac{\varphi_i(x)}{\sqrt{\lambda_i}} \frac{\varphi_i(y)}{\sqrt{\lambda_i}}, \tag{17}$$

where $\{\frac{\varphi_i(x)}{\sqrt{\lambda_i}}\}_i$ is an orthonormal system in $L^2_{\rho_X}$. An example of a C^∞ Mercer kernel was presented in (Zhou, 2002, 2003) to show that only the smoothness of the Mercer kernel does not guarantee the uniform boundedness of the eigenfunctions $\{\frac{\varphi_i(x)}{\sqrt{\lambda_i}}\}_i$. Furthermore, it was shown in (Gittens and Mahoney, 2016) that these normalized eigenfunctions associated with radial basis kernels tend to be localized when the radial basis parameters are made smaller, which raises a practical concern for the uniform boundedness assumption on the eigenfunctions.

Remark 14 *To see how the eigenfunctions $\{\varphi_i(x)\}_i$ change with the marginal distribution, we consider a different measure μ induced by a nonnegative function $P \in L^2_{\rho_X}$ as $d\mu = P(x)d\rho_X$. An eigenpair (λ, φ) of the integral operator $L_{K,\mu}$ associated with the pair (K, μ) with $\lambda > 0$ satisfies*

$$L_{K,\mu}\varphi = \int_{\mathcal{X}} K(\cdot, x)\varphi(x)P(x)d\rho_X = \sum_{\lambda_i>0} \lambda_i \frac{\varphi_i}{\sqrt{\lambda_i}} \int_{\mathcal{X}} \frac{\varphi_i(x)}{\sqrt{\lambda_i}}\varphi(x)P(x)d\rho_X = \lambda\varphi.$$

We introduce an index set $I := \{i : \lambda_i > 0\}$, a possibly infinite matrix

$$\mathbb{K}^P = \left(\lambda_i \int_{\mathcal{X}} \frac{\varphi_i(x)}{\sqrt{\lambda_i}} P(x) \frac{\varphi_j(x)}{\sqrt{\lambda_j}} d\rho_X\right)_{i,j\in I}, \tag{18}$$

and express φ in terms of the orthonormal basis $\{\varphi_i\}_{i\in I}$ as $\varphi = \sum_{i\in I} \frac{c_i}{\sqrt{\lambda_i}}\varphi_i$ where the sequence $c = (c_i)_{i\in I}$ is given by $c_i = \langle\varphi, \frac{\varphi_i}{\sqrt{\lambda_i}}\rangle_{L^2_{\rho_X}} = \sqrt{\lambda_i}\langle\varphi, \varphi_i\rangle_K$. Then the eigenpair (λ, φ) satisfies

$$L_{K,\mu}\varphi = \lambda\varphi \quad \Longleftrightarrow \quad \sum_{i\in I} \frac{\varphi_i}{\sqrt{\lambda_i}} \sum_{j\in I} \left(\mathbb{K}^P\right)_{i,j} c_j = \lambda \sum_{i\in I} \frac{c_i}{\sqrt{\lambda_i}}\varphi_i \quad \Longleftrightarrow \quad \mathbb{K}^P c = \lambda c.$$

Thus the eigenpairs of the integral operator $L_{K,\mu}$ associated with (K, μ) can be characterized by those of the possibly infinite matrix $\mathbb{K}^P$ defined by (18). Finding the eigenpairs of $\mathbb{K}^P$ is an interesting question involving matrix analysis in linear algebra and multiplier operators in harmonic analysis. Note that when $P \equiv 1$ corresponding to $\mu = \rho_X$, $\mathbb{K}^P$ is diagonal with diagonal entries $\{\lambda_i\}_i$ and its eigenvectors yield eigenfunctions $\{\varphi_i\}_i$.

From the above observation, we can see that the marginal distribution ρ_X plays an essential role in the eigenfunction assumption (15) which might be difficult to check for a general marginal distribution ρ_X. For example, it is even unknown whether any of the Gaussian kernels on $\mathcal{X} = [0, 1]^d$ satisfies the eigenfunction assumption (15) when ρ_X is a general Borel measure.

Our learning rates stated in Corollary 4 or Corollary 11 do not require the eigenfunction assumption (15). Moreover, our restriction (10) for the number m of local processors in Corollary 4 is more general than (16) when α is close to 1/2: with $r = 1/2$ corresponding to the condition $f_\rho \in \mathcal{H}_K$, our restriction (10) in Corollary 4 is $m \le N^{\frac{1}{4+6\alpha}}$ with the power index tending to $\frac{1}{7}$ while the restriction in (16) with $k = \infty$ takes the form $m \le c_\alpha \frac{N^{\frac{2\alpha-1}{2\alpha+1}}}{A^4 \log N}$ with the power index tending to 0 as $\alpha \to \frac{1}{2}$. Note that the learning rates stated in Corollary 4 are for the difference $\overline{f}_{D,\lambda} - f_{D,\lambda}$ between the output function of the distributed learning algorithm (2) and that of the algorithm (1) using the whole data. In the special case of $r = \frac{1}{2}$, we can see that $E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho \le \widetilde{C}_{\kappa,c,r} N^{-\frac{\alpha}{2\alpha+1}} m^{-\frac{1}{4\alpha+2}}$, achieved by choosing $\lambda = \left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$, is smaller as m becomes larger. Here the dependence of λ on m is crucial for achieving this convergence rate of the sample error: if we fix λ and N, the error bounds for $E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|$ stated in Theorem 1 and Corollary 3 become larger as m increases. On the other hand, as one expects, increasing the number m of local processors would increase

the approximation error for the regression problem, which can also be seen from the bound with $\lambda = \left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$ stated in Theorem 7. The result in Corollary 11 with $r > \frac{1}{2}$ compensates and gives the optimal learning rate $E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_\rho\right] = O\left(N^{-\frac{2r\alpha}{4\alpha r+1}}\right)$ by restricting m as in (13).

Besides the divide-and-conquer technique, there are many other approaches in the literature to reduce the computational complexity in handling big data. These include the localized learning (Meister and Steinwart, 2016), Nyström regularization (Bach, 2013; Rudi et al., 2015), and online learning (Dekel et al., 2012). The divide-and-conquer technique has the advantage of reducing the single-machine complexity of both space and time without a significant lost of prediction power.

It is an important and interesting problem how the big data are decomposed for distributed learning. Here we only study the approach of random decomposition, and the data subsets are regarded as being drawn from the same distribution. There should be better decomposition strategies for some practical applications. For example, intuitively data points should be divided according to their spatial distribution so that the learning process would yield locally optimal predictors, which could then be synthesized to the output function. See (Thomann et al., 2016).

In this paper, we consider the regularized least squares with Mercer kernels. Our results might be extended to more general kernels. A natural class consists of bounded and positive semidefinite kernels as studied in (Steinwart and Scovel, 2012). By means of the Mercer type expansion (17), one needs some assumptions on the system $\{\frac{\varphi_i(x)}{\sqrt{\lambda_i}}\}_i$, the domain $\mathcal{X}$, and the measure ρ_X to relax the continuity of the kernel while keeping compactness of the integral operator. How to minimize the assumptions and to maximize the scope of applications of the framework such as the situation of an input space $\mathcal{X}$ without a metric (Shen et al., 2014; De Vito et al., 2013) is a valuable question to investigate.

Here we only consider distributed learning with the regularized least squares. It would be of great interest and value to develop the theory for distributed learning with other algorithms such as spectral algorithms (Bauer et al., 2007), empirical feature-based learning (Guo and Zhou, 2012; Guo et al., 2017; Shi et al., 2011), the minimum error entropy principle (Hu et al., 2015; Fan et al., 2016), and randomized Kaczmarz algorithms (Lin and Zhou, 2015).

Remark 15 *After the submission of this paper, in our follow-up paper by Z. C. Guo, S. B. Lin, and D. X. Zhou entitled "Learning theory of distributed spectral algorithms" published in Inverse Problems, error analysis and optimal learning rates for distributed learning with spectral algorithms were derived. In late 2016, we learned that similar analysis was carried out for classical (non-distributed) spectral algorithms implemented in one machine by G. Blanchard and N. Mücke in a paper entitled "Optimal rates for regularization of statistical inverse learning problems" (arXiv:1604.04054, April 2016), and by L. H. Dicker, D. P. Foster, and D. Hsu in a paper entitled "Kernel ridge vs. principal component regression: minimax bounds and adaptibility of regularization operators" (arXiv:1605.08839, May 2016). We are indebted to one of the referees for pointing this out to us.*

4. Second Order Decomposition of Operator Differences and Norms

Our error estimates are achieved by a novel second order decomposition of operator differences in our integral operator approach. We approximate the integral operator L_K by the empirical integral operator $L_{K,D(x)}$ on $\mathcal{H}_K$ defined with the input data set $D(x) = \{x_i\}_{i=1}^N = \{x : (x, y) \in D \text{ for some } y \in \mathcal{Y}\}$ as

$$L_{K,D(x)}(f) = \frac{1}{|D|} \sum_{x \in D(x)} f(x) K_x = \frac{1}{|D|} \sum_{x \in D(x)} \langle f, K_x \rangle_K K_x, \qquad f \in \mathcal{H}_K, \tag{19}$$

where the reproducing property $f(x) = \langle f, K_x \rangle_K$ for $f \in \mathcal{H}_K$ is used. Since K is a Mercer kernel, $L_{K,D_j(x)}$ is a finite-rank positive operator and $L_{K,D_j(x)} + \lambda I$ is invertible.

The operator difference in our study is that of the inverses of two operators defined by

$$Q_{D(x)} = \left(L_{K,D(x)} + \lambda I\right)^{-1} - \left(L_K + \lambda I\right)^{-1}. \tag{20}$$

If we denote $A = L_{K,D(x)} + \lambda I$ and $B = L_K + \lambda I$, then our second order decomposition for the operator difference $Q_{D(x)}$ is a special case of the following second order decomposition for the difference $A^{-1} - B^{-1}$.

Lemma 16 *Let A and B be invertible operators on a Banach space. Then we have*

$$A^{-1} - B^{-1} = B^{-1} \{B - A\} B^{-1} + B^{-1} \{B - A\} A^{-1} \{B - A\} B^{-1}. \tag{21}$$

Proof We can decompose the operator $A^{-1} - B^{-1}$ as

$$A^{-1} - B^{-1} = B^{-1} \{B - A\} A^{-1}. \tag{22}$$

This is the first order decomposition.

Write the last term A^{-1} as $B^{-1} + (A^{-1} - B^{-1})$ and apply another first order decomposition similar to (22) as

$$A^{-1} - B^{-1} = A^{-1} \{B - A\} B^{-1}.$$

It follows from (22) that

$$A^{-1} - B^{-1} = B^{-1} \{B - A\} \left\{B^{-1} + A^{-1} \{B - A\} B^{-1}\right\}.$$

Then the desired identity (21) is verified. The lemma is proved. ■

To demonstrate how the second order decomposition leads to tight error bounds for the classical least squares regularization scheme (1) with the output function $f_{D,\lambda}$, we recall the following formula (see e.g. (Caponnetto and De Vito, 2007; Smale and Zhou, 2007))

$$f_{D,\lambda} - f_\lambda = \left(L_{K,D(x)} + \lambda I\right)^{-1} \Delta_D, \quad \Delta_D := \frac{1}{|D|} \sum_{z \in D} \xi_\lambda(z) - E[\xi_\lambda], \tag{23}$$

where Δ_D is induced by the random variables ξ_λ with values in the Hilbert space $\mathcal{H}_K$ defined as

$$\xi_\lambda(z) = (y - f_\lambda(x)) K_x, \qquad z = (x, y) \in \mathcal{Z}. \tag{24}$$

Then we can express $f_{D,\lambda} - f_\lambda$ by means of the notation $Q_{D(x)} = \left(L_{K,D(x)} + \lambda I\right)^{-1} - (L_K + \lambda I)^{-1}$ as

$$f_{D,\lambda} - f_\lambda = \left[Q_{D(x)}\right] \Delta_D + (L_K + \lambda I)^{-1} \Delta_D.$$

Due to the identity between the $L^2_{\rho_X}$ norm and the $\mathcal{H}_K$ metric

$$\|g\|_\rho = \|L_K^{\frac{1}{2}} g\|_K, \qquad \forall g \in L^2_{\rho_X}, \tag{25}$$

we can estimate the error $\|f_{D,\lambda} - f_\lambda\|_\rho = \|L_K^{\frac{1}{2}} (f_{D,\lambda} - f_\lambda)\|_K$ by bounding the $\mathcal{H}_K$ norm of the following expression obtained from the second order decomposition (21) with $B - A = L_K - L_{K,D(x)}$

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

Combining this expression with the operator norm bound $\left\| L_K^{\frac{1}{2}} (L_K + \lambda I)^{-\frac{1}{2}} \right\| \leq 1$ and the notation

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

for convenience, we find

$$\begin{aligned} \left\| L_K^{\frac{1}{2}} \left[Q_{D(x)}\right] \Delta_D \right\|_K \leq & \Xi_D \left\| (L_K + \lambda I)^{-1} \Delta_D \right\|_K + \Xi_D \left\| \left(L_{K,D(x)} + \lambda I \right)^{-1} \right\| \\ & \Xi_D \left\| (L_K + \lambda I)^{-\frac{1}{2}} \Delta_D \right\|_K. \end{aligned}$$

By decomposing $(L_K + \lambda I)^{-1} \Delta_D$ as $(L_K + \lambda I)^{-\frac{1}{2}} (L_K + \lambda I)^{-\frac{1}{2}} \Delta_D$ and using the bounds $\left\| \left(L_{K,D(x)} + \lambda I \right)^{-1} \right\| \leq \frac{1}{\lambda}$, $\left\| (L_K + \lambda I)^{-\frac{1}{2}} \right\| \leq 1/\sqrt{\lambda}$, we know that

$$\left\| L_K^{\frac{1}{2}} \left[Q_{D(x)}\right] \Delta_D \right\|_K \leq \left(\frac{\Xi_D}{\sqrt{\lambda}} + \frac{\Xi_D^2}{\lambda} \right) \left\| (L_K + \lambda I)^{-\frac{1}{2}} \Delta_D \right\|_K \tag{27}$$

and

$$\|f_{D,\lambda} - f_\lambda\|_\rho \leq \left(\frac{\Xi_D}{\sqrt{\lambda}} + \frac{\Xi_D^2}{\lambda} + 1 \right) \left\| (L_K + \lambda I)^{-\frac{1}{2}} \Delta_D \right\|_K. \tag{28}$$

Thus the classical least squares regularization scheme (1) can be analyzed after estimating the operator norm $\Xi_D = \left\| (L_K + \lambda I)^{-\frac{1}{2}} \left\{ L_K - L_{K,D(x)} \right\} \right\|$ and the function norm $\left\| (L_K + \lambda I)^{-\frac{1}{2}} \Delta_D \right\|_K$. In the following two lemmas, to be proved in the appendix, these norms are estimated in terms of effective dimensions by methods in the existing literature (Caponnetto and De Vito, 2007; Bauer et al., 2007; Blanchard and Krämer, 2010).

Lemma 17 *Let D be a sample drawn independently according to ρ. Then the following estimates for the operator norm $\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left\{L_K-L_{K,D(x)}\right\}\right\|$ hold.*

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

(b) For any $0<\delta<1$, with confidence at least $1-\delta$, there holds

$$\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left\{L_K-L_{K,D(x)}\right\}\right\|\leq\mathcal{B}_{|D|,\lambda}\log(2/\delta),\tag{29}$$

where we denote the quantity

$$\mathcal{B}_{|D|,\lambda}=\frac{2\kappa}{\sqrt{|D|}}\left\{\frac{\kappa}{\sqrt{|D|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\right\}.\tag{30}$$

(c) For any $d>1$, there holds

$$\left\{E\left[\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left\{L_K-L_{K,D(x)}\right\}\right\|^d\right]\right\}^{\frac{1}{d}}\leq(2d\Gamma(d)+1)^{\frac{1}{d}}\mathcal{B}_{|D|,\lambda},$$

where Γ is the Gamma function defined for $d>0$ by $\Gamma(d)=\int_0^\infty u^{d-1}\exp\{-u\}\,du$.

Lemma 18 *Let D be a sample drawn independently according to ρ and g be a measurable bounded function on $\mathcal{Z}$ and ξ_g be the random variable with values in $\mathcal{H}_K$ defined by $\xi_g(z)=g(z)K_x$ for $z=(x,y)\in\mathcal{Z}$. Then the following statements hold.*

(a) $E\left[\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K^2\right]=\mathcal{N}(\lambda)$.

(b) For almost every $x\in\mathcal{X}$, there holds $\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K\leq\frac{\kappa}{\sqrt{\lambda}}$.

(c) For any $0<\delta<1$, with confidence at least $1-\delta$, there holds

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

Remark 19 *In the existing literature, the first order decomposition (22) was used. To bound the norm of $\left(L_{K,D(x)}+\lambda I\right)^{-1}\Delta_D$ by this approach, one needs to either use the brute force estimate $\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|\leq\frac{1}{\lambda}$ leading to suboptimal learning rates, or applying the approximation of L_K by $L_{K,D(x)}$ which is valid only with confidence and leads to confidence-based estimates with λ depending on the confidence level. In our second order decomposition, we decompose the inverse operator $\left(L_{K,D(x)}+\lambda I\right)^{-1}$ further after the first order decomposition (22). This leads to finer estimates with an additional factor $\|(B-A)B^{-\frac{1}{2}}\|$ in the second term of the bound (21) and gives the refined error bound (28).*

5. Deriving Error Bounds for Least Squares Regularization Scheme

In this section we prove our main result on error bounds for the least squares regularization scheme (1), which illustrates how to apply the second order decomposition (21) for operator differences in our integral operator approach.

Proposition 20 *Assume $E[y^2] < \infty$ and moment condition (11) for some $1 \le p \le \infty$. Then*

$$\begin{aligned} E\left[\|f_{D,\lambda} - f_\lambda\|_\rho\right] &\le \left(2 + 56\kappa^4 + 57\kappa^2\right)\left(1 + \frac{1}{(|D|\lambda)^2} + \frac{\mathcal{N}(\lambda)}{|D|\lambda}\right) \\ &\quad \left\{\kappa^{\frac{1}{p}}\sqrt{\left\|\sigma_\rho^2\right\|_p}\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{|D|\lambda}\right)^{\frac{1}{2p}} + \kappa\frac{\|f_\lambda - f_\rho\|_\rho}{\sqrt{|D|\lambda}}\right\}. \end{aligned}$$

Proof We apply the error bound (28) and the Schwarz inequality to get

$$E\left[\|f_{D,\lambda} - f_\lambda\|_\rho\right] \le \left\{E\left[\left(1 + \frac{\Xi_D}{\sqrt{\lambda}} + \frac{\Xi_D^2}{\lambda}\right)^2\right]\right\}^{1/2}\left\{E\left[\left\|(L_K + \lambda I)^{-1/2}\Delta_D\right\|_K^2\right]\right\}^{1/2}. \tag{31}$$

The first factor in the above bound involves $\Xi_D = \left\|(L_K + \lambda I)^{-\frac{1}{2}}\left\{L_K - L_{K,D(x)}\right\}\right\|$ and it can be estimated by applying Lemma 17 as

$$\begin{aligned} \left\{E\left[\left(1 + \frac{\Xi_D}{\sqrt{\lambda}} + \frac{\Xi_D^2}{\lambda}\right)^2\right]\right\}^{1/2} &\le 1 + \left\{E\left[\frac{\Xi_D^2}{\lambda}\right]\right\}^{1/2} + \left\{E\left[\frac{\Xi_D^4}{\lambda^2}\right]\right\}^{1/2} \\ &\le 1 + \left\{\frac{\kappa^2\mathcal{N}(\lambda)}{|D|\lambda}\right\}^{1/2} + \left\{\frac{49\mathcal{B}_{|D|,\lambda}^4}{\lambda^2}\right\}^{1/2} \\ &\le 2 + \frac{56\kappa^4}{(|D|\lambda)^2} + \frac{57\kappa^2\mathcal{N}(\lambda)}{|D|\lambda}. \end{aligned} \tag{32}$$

To deal with the factor in the bound (31), we separate $\Delta_D = \frac{1}{|D|}\sum_{z\in D}\xi_\lambda(z) - E[\xi_\lambda]$ as

$$\Delta_D = \Delta_D' + \Delta_D'',$$

where

$$\Delta_D' := \frac{1}{|D|}\sum_{z\in D}\xi_0(z), \quad \Delta_D'' := \frac{1}{|D|}\sum_{z\in D}\left(\xi_\lambda - \xi_0\right)(z) - E[\xi_\lambda]$$

are induced by another random variable ξ_0 with values in the Hilbert space $\mathcal{H}_K$ defined by

$$\xi_0(z) = (y - f_\rho(x))K_x, \qquad z = (x, y) \in \mathcal{Z}. \tag{33}$$

Then the second factor in (31) can be separated as

$$\begin{aligned} &\left\{E\left[\left\|(L_K + \lambda I)^{-1/2}\Delta_D\right\|_K^2\right]\right\}^{1/2} \\ &\le \left\{E\left[\left\|(L_K + \lambda I)^{-1/2}\Delta_D'\right\|_K^2\right]\right\}^{1/2} + \left\{E\left[\left\|(L_K + \lambda I)^{-1/2}\Delta_D''\right\|_K^2\right]\right\}^{1/2}. \end{aligned} \tag{34}$$

Let us bound the first term of (34). Observe that

$$(L_K+\lambda I)^{-1/2}\Delta'_D=\sum_{z\in D}\frac{1}{|D|}(y-f_\rho(x))\,(L_K+\lambda I)^{-1/2}(K_x).$$

Each term in this expression is unbiased because $\int_{\mathcal{Y}}(y-f_\rho(x))\,d\rho(y|x)=0$. This unbiasedness and the independence tell us that

$$\begin{aligned}\left\{E\left[\left\|(L_K+\lambda I)^{-1/2}\Delta'_D\right\|_K^2\right]\right\}^{1/2} &= \left\{\frac{1}{|D|}E\left[\left\|(y-f_\rho(x))\left[(L_K+\lambda I)^{-1/2}\right](K_x)\right\|_K^2\right]\right\}^{1/2}\\ &= \left\{\frac{1}{|D|}E\left[\sigma_\rho^2(x)\left\|\left[(L_K+\lambda I)^{-1/2}\right](K_x)\right\|_K^2\right]\right\}^{1/2}. \quad (35)\end{aligned}$$

If $\sigma_\rho^2\in L^\infty$, then $\sigma_\rho^2(x)\le\left\|\sigma_\rho^2\right\|_\infty$ and by Lemma 18 we have

$$\left\{E\left[\left\|(L_K+\lambda I)^{-1/2}\Delta'_D\right\|_K^2\right]\right\}^{1/2}\le\sqrt{\left\|\sigma_\rho^2\right\|_\infty}\sqrt{\mathcal{N}(\lambda)/|D|}.$$

If $\sigma_\rho^2\in L^p_{\rho_X}$ with $1\le p<\infty$, we take $q=\frac{p}{p-1}$ ($q=\infty$ for $p=1$) satisfying $\frac{1}{p}+\frac{1}{q}=1$ and apply the Hölder inequality $E[|\xi\eta|]\le(E[|\xi|^p])^{1/p}(E[|\eta|^q])^{1/q}$ to $\xi=\sigma_\rho^2$, $\eta=\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K^2$ to find

$$E\left[\sigma_\rho^2(x)\left\|\left[(L_K+\lambda I)^{-1/2}\right](K_x)\right\|_K^2\right]\le\left\|\sigma_\rho^2\right\|_p\left\{E\left[\left\|\left[(L_K+\lambda I)^{-1/2}\right](K_x)\right\|_K^{2q}\right]\right\}^{1/q}.$$

But

$$\left\|\left[(L_K+\lambda I)^{-1/2}\right](K_x)\right\|_K^{2q-2}\le\left(\kappa/\sqrt{\lambda}\right)^{2q-2}$$

and $E\left[\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K^2\right]=\mathcal{N}(\lambda)$ by Lemma 18. So we have

$$\begin{aligned}\left\{E\left[\left\|(L_K+\lambda I)^{-1/2}\Delta'_D\right\|_K^2\right]\right\}^{1/2} &\le \left\{\frac{1}{|D|}\left\|\sigma_\rho^2\right\|_p\left\{\frac{\kappa^{2q-2}}{\lambda^{q-1}}\mathcal{N}(\lambda)\right\}^{1/q}\right\}^{1/2}\\ &= \sqrt{\left\|\sigma_\rho^2\right\|_p}\kappa^{\frac{1}{p}}\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{|D|\lambda}\right)^{\frac{1}{2p}}.\end{aligned}$$

Combining the above two cases, we know that for either $p=\infty$ or $p<\infty$, the first term of (34) can be bounded as

$$\left\{E\left[\left\|(L_K+\lambda I)^{-1/2}\Delta'_D\right\|_K^2\right]\right\}^{1/2}\le\sqrt{\left\|\sigma_\rho^2\right\|_p}\kappa^{\frac{1}{p}}\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{|D|\lambda}\right)^{\frac{1}{2p}}.$$

The second term of (34) can be bounded easily as

$$\begin{aligned}\left\{E\left[\left\|(L_K+\lambda I)^{-1/2}\Delta_D''\right\|_K^2\right]\right\}^{1/2} &\leq \frac{1}{\sqrt{|D|}}\left\{E\left[(f_\rho(x)-f_\lambda(x))^2\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K^2\right]\right\}^{1/2}\\ &\leq \frac{1}{\sqrt{|D|}}\left\{E\left[(f_\rho(x)-f_\lambda(x))^2\frac{\kappa^2}{\lambda}\right]\right\}^{1/2}=\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D|\lambda}}.\end{aligned}$$

Putting the above estimates for the two terms of (34) into (31) and applying the bound (32) involving Ξ_D, we know that $E\left[\|f_{D,\lambda}-f_\lambda\|_\rho\right]$ is bounded by

$$\left(2+\frac{56\kappa^4}{(|D|\lambda)^2}+\frac{57\kappa^2\mathcal{N}(\lambda)}{|D|\lambda}\right)\left(\sqrt{\left\|\sigma_\rho^2\right\|_p}\kappa^{\frac{1}{p}}\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{|D|\lambda}\right)^{\frac{1}{2p}}+\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D|\lambda}}\right).$$

Then our desired error bound follows. The proof of the proposition is complete. ■

Proof of Theorem 7 Combining Proposition 20 with the triangle inequality $\|f_{D,\lambda}-f_\rho\|_\rho \leq \|f_{D,\lambda}-f_\lambda\|_\rho+\|f_\lambda-f_\rho\|_\rho$, we know that

$$\begin{aligned}E\left[\|f_{D,\lambda}-f_\rho\|_\rho\right] &\leq \|f_\lambda-f_\rho\|_\rho+\left(2+56\kappa^4+57\kappa^2\right)\left(1+\frac{1}{(|D|\lambda)^2}+\frac{\mathcal{N}(\lambda)}{|D|\lambda}\right)\\ &\quad\left\{\kappa^{\frac{1}{p}}\sqrt{\left\|\sigma_\rho^2\right\|_p}\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{|D|\lambda}\right)^{\frac{1}{2p}}+\frac{\kappa}{\sqrt{|D|\lambda}}\|f_\lambda-f_\rho\|_\rho\right\}.\end{aligned}$$

Then the desired error bound holds true, and the proof of Theorem 7 is complete. ■

Proof of Corollary 8 By the definition of effective dimension,

$$\mathcal{N}(\lambda)=\sum_\ell\frac{\lambda_\ell}{\lambda_\ell+\lambda}\geq\frac{\lambda_1}{\lambda_1+\lambda}.$$

Combining this with the restriction (8) with $m=1$, we find $\mathcal{N}(\lambda)\geq\frac{\lambda_1}{\lambda_1+C_0}$ and $N\lambda\geq\frac{\lambda_1}{(\lambda_1+C_0)C_0}$. Putting these and the restriction (8) with $m=1$ into the error bound (12), we know that

$$\begin{aligned}&E\left[\|f_{D,\lambda}-f_\rho\|_\rho\right]\leq\left(2+56\kappa^4+57\kappa^2\right)(1+\kappa)\left(1+\frac{(\lambda_1+C_0)^2C_0^2}{\lambda_1^2}+C_0\right)\\ &\left\{\left(1+\sqrt{(\lambda_1+C_0)C_0/\lambda_1}\right)\|f_\lambda-f_\rho\|_\rho+\sqrt{\left\|\sigma_\rho^2\right\|_p}\left(\frac{\mathcal{N}(\lambda)}{N}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{N\lambda}\right)^{\frac{1}{2p}}\right\}.\end{aligned}$$

Then the desired bound follows. The proof of Corollary 8 is complete. ■

To prove Corollary 9, we need the following bounds (Smale and Zhou, 2007) for $\|f_\lambda-f_\rho\|_\rho$ and $\|f_\lambda-f_\rho\|_K$.

Lemma 21 *Assume regularity condition (9) with $0 < r \le 1$. There holds*

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|g_\rho\|_\rho. \tag{36}$$

Furthermore, if $1/2 \le r \le 1$, then we have

$$\|f_\lambda - f_\rho\|_K \le \lambda^{r-1/2} \|g_\rho\|_\rho. \tag{37}$$

Proof of Corollary 9 By Lemma 21, regularity condition (9) with $0 < r \le 1$ implies

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|g_\rho\|_\rho.$$

If

$$\mathcal{N}(\lambda) \le C_0 \lambda^{-\frac{1}{2\alpha}}, \qquad \forall \lambda > 0$$

for some constant $C_0 \ge 1$, then the choice $\lambda = N^{-\frac{2\alpha}{2\alpha \max\{2r,1\}+1}}$ yields

$$\frac{\mathcal{N}(\lambda)}{N\lambda} \le \frac{C_0 \lambda^{-\frac{1}{2\alpha}-1}}{N} = C_0 N^{\frac{2\alpha+1}{2\alpha \max\{2r,1\}+1}-1} \le C_0.$$

So (8) with $m = 1$ is satisfied. With this choice we also have

$$\begin{aligned}\left(\frac{\mathcal{N}(\lambda)}{N}\right)^{\frac{1}{2}(1-\frac{1}{p})} \left(\frac{1}{N\lambda}\right)^{\frac{1}{2p}} &\le C_0^{\frac{1}{2}(1-\frac{1}{p})} N^{-\frac{2\alpha \max\{2r,1\}}{2\alpha \max\{2r,1\}+1}\frac{1}{2}(1-\frac{1}{p})} N^{-\frac{2\alpha \max\{2r,1\}+1-2\alpha}{2\alpha \max\{2r,1\}+1}\frac{1}{2p}} \\ &= C_0^{\frac{1}{2}(1-\frac{1}{p})} N^{-\frac{\alpha \max\{2r,1\}}{2\alpha \max\{2r,1\}+1}+\frac{1}{2p}\frac{2\alpha-1}{2\alpha \max\{2r,1\}+1}}.\end{aligned}$$

Putting these estimates into Corollary 8, we know that

$$\begin{aligned}E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] &= O\left(N^{-\frac{2\alpha r}{2\alpha \max\{2r,1\}+1}} + N^{-\frac{\alpha \max\{2r,1\}}{2\alpha \max\{2r,1\}+1}+\frac{1}{2p}\frac{2\alpha-1}{2\alpha \max\{2r,1\}+1}}\right) \\ &= O\left(N^{-\frac{\alpha \min\{2r,\max\{2r,1\}\}}{2\alpha \max\{2r,1\}+1}+\frac{1}{2p}\frac{2\alpha-1}{2\alpha \max\{2r,1\}+1}}\right).\end{aligned}$$

But we find

$$\min\{2r, \max\{2r,1\}\} = 2r$$

by discussing the two different cases $0 < r < \frac{1}{2}$ and $\frac{1}{2} \le r \le 1$. Then our conclusion follows immediately. The proof of Corollary 9 is complete. ■

6. Proof of Error Bounds for the Distributed Learning Algorithm

To analyze the error $\overline{f}_{D,\lambda} - f_{D,\lambda}$ for distributed learning, we recall the notation $Q_{D(x)} = \left(L_{K,D(x)} + \lambda I\right)^{-1} - (L_K + \lambda I)^{-1}$ for the difference of inverse operators and use the notation $Q_{D_j(x)}$ involving the data subset D_j. The empirical integral operator $L_{K,D_j(x)}$ is defined with D replaced by the data subset D_j. For our error analysis for the distributed learning algorithm (2), we need the following error decomposition for $\overline{f}_{D,\lambda} - f_{D,\lambda}$.

Lemma 22 *Assume $E[y^2] < \infty$. For $\lambda > 0$, we have*

$$\begin{aligned}\overline{f}_{D,\lambda} - f_{D,\lambda} &= \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left[\left(L_{K,D_j(x)} + \lambda I \right)^{-1} - \left(L_{K,D(x)} + \lambda I \right)^{-1} \right] \Delta_j \\ &= \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left[Q_{D_j(x)} \right] \Delta_j' + \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left[Q_{D_j(x)} \right] \Delta_j'' - \left[Q_{D(x)} \right] \Delta_D, \quad (38)\end{aligned}$$

where

$$\Delta_j := \frac{1}{|D_j|} \sum_{z \in D_j} \xi_\lambda(z) - E[\xi_\lambda], \quad \Delta_D := \frac{1}{|D|} \sum_{z \in D} \xi_\lambda(z) - E[\xi_\lambda],$$

and

$$\Delta_j' := \frac{1}{|D_j|} \sum_{z \in D_j} \xi_0(z), \quad \Delta_j'' := \frac{1}{|D_j|} \sum_{z \in D_j} \left(\xi_\lambda - \xi_0 \right)(z) - E[\xi_\lambda].$$

Proof Recall the expression (23) for $f_{D,\lambda} - f_\lambda$. When the data subset D_j is used, we have

$$f_{D_j,\lambda} - f_\lambda = \left(L_{K,D_j(x)} + \lambda I \right)^{-1} \Delta_j.$$

So we know that

$$\overline{f}_{D,\lambda} - f_\lambda = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left\{ f_{D_j,\lambda} - f_\lambda \right\} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left(L_{K,D_j(x)} + \lambda I \right)^{-1} \Delta_j.$$

We can decompose Δ_D as

$$\Delta_D = \frac{1}{|D|} \sum_{z \in D} \xi_\lambda(z) - E[\xi_\lambda] = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left\{ \frac{1}{|D_j|} \sum_{z \in D_j} \xi_\lambda(z) - E[\xi_\lambda] \right\} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \Delta_j$$

in the expression (23) for $f_{D,\lambda} - f_\lambda$ and find

$$f_{D,\lambda} - f_\lambda = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left(L_{K,D(x)} + \lambda I \right)^{-1} \Delta_j.$$

Then the first desired expression for $\overline{f}_{D,\lambda} - f_{D,\lambda}$ follows.

By adding and subtracting the operator $(L_K + \lambda I)^{-1}$, writing $\Delta_j = \Delta_j' + \Delta_j''$, and noting $E[\xi_0] = 0$, we know that the first expression implies (38). This proves Lemma 22. ■

Before proving our first main result on the error $\overline{f}_{D,\lambda} - f_{D,\lambda}$ in the $\mathcal{H}_K$ metric and L_ρ^2 metric, we state the following more general result which allows different sizes for data subsets $\{D_j\}$.

Theorem 23 *Assume that for some constant $M > 0$, $|y| \leq M$ almost surely. Then we have*

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho\right] \leq C'_\kappa M\sqrt{\lambda}\left\{\sum_{j=1}^m \left(\frac{|D_j|}{|D|}\right)^2 \left(\mathcal{S}^2_{|D_j|,\lambda} + \mathcal{S}^3_{|D_j|,\lambda}\right)\right\}^{\frac{1}{2}} + C'_\kappa \sum_{j=1}^m \frac{|D_j|}{|D|}$$
$$\frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{|D_j|\lambda}}\left(\sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}} + \mathcal{S}_{|D_j|,\lambda}\right) + C'_\kappa\left(\sqrt{\frac{\mathcal{N}(\lambda)}{|D|\lambda}} + \mathcal{S}_{|D|,\lambda}\right)\left(M\sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} + \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{|D|\lambda}}\right)$$

and

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_K\right] \leq C'_\kappa M\left\{\sum_{j=1}^m \left(\frac{|D_j|}{|D|}\right)^2 \left(\mathcal{S}^2_{|D_j|,\lambda} + \mathcal{S}^3_{|D_j|,\lambda}\right)\right\}^{\frac{1}{2}} + C'_\kappa \sum_{j=1}^m \frac{|D_j|}{|D|}$$
$$\frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{|D_j|\lambda}}\left(\sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}} + \mathcal{S}_{|D_j|,\lambda}\right) + C'_\kappa\left(\sqrt{\frac{\mathcal{N}(\lambda)}{|D|\lambda}} + \mathcal{S}_{|D|,\lambda}\right)\left(M\sqrt{\frac{\mathcal{N}(\lambda)}{|D|\lambda}} + \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{|D|\lambda}}\right),$$

where C'_κ is a constant depending only on κ, and $\mathcal{S}_{|D_j|,\lambda}$ is the quantity given by

$$\mathcal{S}_{|D_j|,\lambda} = \frac{1}{|D_j|^2\lambda^2} + \frac{\mathcal{N}(\lambda)}{|D_j|\lambda}.$$

Proof Recall the expression (38) for $\overline{f}_{D,\lambda} - f_{D,\lambda}$ in Lemma 22. It enables us to express

$$L_K^{1/2}\left\{\overline{f}_{D,\lambda} - f_{D,\lambda}\right\} = J_1 + J_2 + J_3, \tag{39}$$

where the terms J_1, J_2, J_3 are given by

$$J_1 = \sum_{j=1}^m \frac{|D_j|}{|D|}\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta'_j,\ J_2 = \sum_{j=1}^m \frac{|D_j|}{|D|}\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta''_j,\ J_3 = -\left[L_K^{1/2}Q_{D(x)}\right]\Delta_D.$$

These three terms will be dealt with separately in the following.

For the first term J_1 of (39), each summand with $j \in \{1,\ldots,m\}$ can be expressed as $\sum_{z\in D_j}\frac{1}{|D|}(y - f_\rho(x))\left[L_K^{1/2}Q_{D_j(x)}\right](K_x)$, and is unbiased because $\int_{\mathcal{Y}} y - f_\rho(x)d\rho(y|x) = 0$. The unbiasedness and the independence tell us that

$$E\left[\|J_1\|_K\right] \leq \left\{E\left[\|J_1\|_K^2\right]\right\}^{1/2} \leq \left\{\sum_{j=1}^m \left(\frac{|D_j|}{|D|}\right)^2 E\left[\left\|\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta'_j\right\|_K^2\right]\right\}^{1/2}. \tag{40}$$

Let $j \in \{1,\ldots,m\}$. The estimate (27) derived from the second order decomposition (21) yields

$$\left\|\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta'_j\right\|_K^2 \leq \left(\frac{\Xi_{D_j}}{\sqrt{\lambda}} + \frac{\Xi^2_{D_j}}{\lambda}\right)^2 \left\|(L_K + \lambda I)^{-1/2}\Delta'_j\right\|_K^2. \tag{41}$$

Now we apply the formula

$$E[\xi] = \int_0^\infty \text{Prob}\left[\xi > t\right] dt \tag{42}$$

to estimate the expected value of (41). By Part (b) of Lemma 17, for $0 < \delta < 1$, there exists a subset $\mathcal{Z}_{\delta,1}^{|D_j|}$ of $\mathcal{Z}^{|D_j|}$ of measure at least $1-\delta$ such that

$$\Xi_{D_j} \leq \mathcal{B}_{|D_j|,\lambda} \log(2/\delta), \qquad \forall D_j \in \mathcal{Z}_{\delta,1}^{|D_j|}. \tag{43}$$

Applying Part (c) of Lemma 18 to $g(z) = y - f_\rho(x)$ with $\|g\|_\infty \leq 2M$ and the data subset D_j, we know that there exists another subset $\mathcal{Z}_{\delta,2}^{|D_j|}$ of $\mathcal{Z}^{|D_j|}$ of measure at least $1-\delta$ such that

$$\left\| (L_K + \lambda I)^{-1/2} \Delta_j' \right\|_K \leq \frac{2M}{\kappa} \mathcal{B}_{|D_j|,\lambda} \log(2/\delta), \qquad \forall D_j \in \mathcal{Z}_{\delta,2}^{|D_j|}.$$

Combining this with (43) and (41), we know that for $D_j \in \mathcal{Z}_{\delta,1}^{|D_j|} \cap \mathcal{Z}_{\delta,2}^{|D_j|}$,

$$\left\| \left[L_K^{1/2} Q_{D_j(x)} \right] \Delta_j' \right\|_K^2 \leq \left(\frac{\mathcal{B}_{|D_j|,\lambda}^2}{\lambda} + \frac{\mathcal{B}_{|D_j|,\lambda}^4}{\lambda^2} \right) \left(\frac{M}{\kappa} \right)^2 \mathcal{B}_{|D_j|,\lambda}^2 \left(2 \log(2/\delta) \right)^6.$$

Since the measure of the set $\mathcal{Z}_{\delta,1}^{|D_j|} \cap \mathcal{Z}_{\delta,2}^{|D_j|}$ is at least $1-2\delta$, by denoting

$$\mathcal{C}_{|D_j|,\lambda} = 64 \left(\frac{\mathcal{B}_{|D_j|,\lambda}^2}{\lambda} + \frac{\mathcal{B}_{|D_j|,\lambda}^4}{\lambda^2} \right) \left(\frac{M}{\kappa} \right)^2 \mathcal{B}_{|D_j|,\lambda}^2,$$

we see that

$$\text{Prob}\left[\left\| \left[L_K^{1/2} Q_{D_j(x)} \right] \Delta_j' \right\|_K^2 > \mathcal{C}_{|D_j|,\lambda} \left(\log(2/\delta) \right)^6 \right] \leq 2\delta.$$

For $0 < t < \infty$, the equation $\mathcal{C}_{|D_j|,\lambda} \left(\log(2/\delta) \right)^6 = t$ has the solution

$$\delta_t = 2 \exp\left\{ - \left(t / \mathcal{C}_{|D_j|,\lambda} \right)^{1/6} \right\}.$$

When $\delta_t < 1$, we have

$$\text{Prob}\left[\left\| \left[L_K^{1/2} Q_{D_j(x)} \right] \Delta_j' \right\|_K^2 > t \right] \leq 2\delta_t = 4 \exp\left\{ - \left(t / \mathcal{C}_{|D_j|,\lambda} \right)^{1/6} \right\}.$$

This inequality holds trivially when $\delta_t \geq 1$ since the probability is at most 1. Thus we can apply the formula (42) to the nonnegative random variable $\xi = \left\| \left[L_K^{1/2} Q_{D_j(x)} \right] \Delta_j' \right\|_K^2$ and obtain

$$E\left[\left\| \left[L_K^{1/2} Q_{D_j(x)} \right] \Delta_j' \right\|_K^2 \right] = \int_0^\infty \text{Prob}\left[\xi > t\right] dt \leq \int_0^\infty 4 \exp\left\{ - \left(t / \mathcal{C}_{|D_j|,\lambda} \right)^{1/6} \right\} dt$$

which equals $24\Gamma(6)\mathcal{C}_{|D_j|,\lambda}$. Therefore,

$$E\left[\|J_1\|_K\right] \le \left\{\sum_{j=1}^m \left(\frac{|D_j|}{|D|}\right)^2 24\Gamma(6)\mathcal{C}_{|D_j|,\lambda}\right\}^{1/2}$$
$$\le 1536\sqrt{5}\kappa M \left\{\sum_{j=1}^m \left(\frac{|D_j|}{|D|}\right)^2 \lambda\left(\frac{\kappa^2}{|D_j|^2\lambda^2}+\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}\right)^2\left\{1+8\kappa^2\left(\frac{\kappa^2}{|D_j|^2\lambda^2}+\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}\right)\right\}\right\}^{1/2}.$$

For the second term J_2 of (39), we apply (27) again and obtain

$$\left\|\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta_j''\right\|_K \le \left(\frac{\Xi_{D_j}}{\sqrt{\lambda}}+\frac{\Xi_{D_j}^2}{\lambda}\right)\left\|(L_K+\lambda I)^{-1/2}\Delta_j''\right\|_K.$$

Applying the Schwarz inequality and Lemmas 17 and 18, we get

$$\begin{aligned}
E\left[\left\|\left[L_K^{1/2}Q_{D_j(x)}\right]\Delta_j''\right\|_K\right] &\le \left\{E\left[\left(\frac{\Xi_{D_j}}{\sqrt{\lambda}}+\frac{\Xi_{D_j}^2}{\lambda}\right)^2\right]\right\}^{1/2} \\
&\quad \frac{1}{\sqrt{|D_j|}}\left\{E\left[(f_\rho(x)-f_\lambda(x))^2\left\|(L_K+\lambda I)^{-1/2}(K_x)\right\|_K^2\right]\right\}^{1/2} \\
&\le \left(\left\{\frac{\kappa^2\mathcal{N}(\lambda)}{|D_j|\lambda}\right\}^{1/2}+\left\{\frac{49\mathcal{B}_{|D_j|,\lambda}^4}{\lambda^2}\right\}^{1/2}\right)\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D_j|\lambda}} \\
&\le \left(\kappa\sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}}+56\kappa^2\left(\frac{\kappa^2}{(|D_j|\lambda)^2}+\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}\right)\right)\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D_j|\lambda}}.
\end{aligned}$$

It follows that

$$E\left[\|J_2\|_K\right] \le \sum_{j=1}^m \frac{|D_j|}{|D|}\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D_j|\lambda}}\left(\kappa\sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}}+56\kappa^2\left(\frac{\kappa^2}{(|D_j|\lambda)^2}+\frac{\mathcal{N}(\lambda)}{|D_j|\lambda}\right)\right).$$

The last term J_3 of (39) has been handled by (27) and in the proof of Proposition 20 by ignoring the summand 1 in the bound (31), and we find from the trivial bound $\left\|\sigma_\rho^2\right\|_\infty \le 4M^2$ with $p=\infty$ that

$$E\left[\|J_3\|_K\right] \le \left(\kappa\sqrt{\frac{\mathcal{N}(\lambda)}{|D|\lambda}}+56\kappa^2\left(\frac{\kappa^2}{(|D|\lambda)^2}+\frac{\mathcal{N}(\lambda)}{|D|\lambda}\right)\right)\left(2M\left(\frac{\mathcal{N}(\lambda)}{|D|}\right)^{\frac{1}{2}}+\frac{\kappa\|f_\rho-f_\lambda\|_\rho}{\sqrt{|D|\lambda}}\right).$$

Combining the above estimates for the three terms of (39), we see that the desired error bound in the $L^2_{\rho_X}$ metric holds true.

The estimate in the $\mathcal{H}_K$ metric follows from the steps in deriving the error bound in the $L^2_{\rho_X}$ metric except that in the representation (39) the operator $L_K^{1/2}$ in the front disappears. This change gives an additional factor $1/\sqrt{\lambda}$, the bound for the operator $(L_K+\lambda I)^{-1/2}$, and proves the desired error bound in the $\mathcal{H}_K$ metric. ■

Remark 24 *A crucial step in the above error analysis for the distributed learning algorithm is to use the unbiasedness and independence to get (40) where the norm is squared in the expected value. Thus we can obtain the optimal learning rates in expectation for the distributed learning algorithm (2) but have difficulty in getting rates in probability. It would be interesting to derive error bounds in probability by combining our second order decomposition technique with some analysis in the literature (Caponnetto and De Vito, 2007; Blanchard and Krämer, 2010; Steinwart and Christmann, 2008; Wu and Zhou, 2008).*

Proof of Theorem 1 Since $|D_j| = \frac{N}{m}$ for $j = 1, \ldots, m$, the bound in Theorem 23 in the $L^2_{\rho_X}$ metric can be simplified as

$$\begin{aligned}
E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho\right] &\le C'_\kappa M\sqrt{\lambda}\sqrt{m}\left(\frac{m}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right)\left\{1 + \sqrt{m}\left(\frac{m}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right)^{\frac{1}{2}}\right\} \\
&\quad + C'_\kappa \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} m\left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \frac{m\sqrt{m}}{(N\lambda)^2} + \frac{\sqrt{m}\mathcal{N}(\lambda)}{N\lambda}\right) \\
&\quad + C'_\kappa\left(\frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} + M\sqrt{\frac{\mathcal{N}(\lambda)}{N}}\right)\left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \frac{1}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right) \\
&\le C'_\kappa M(\sqrt{m}+1)\sqrt{\lambda}\left(\frac{m}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right)\left\{1 + \sqrt{m}\left(\frac{m}{(N\lambda)^2} + \frac{\mathcal{N}(\lambda)}{N\lambda}\right)^{\frac{1}{2}}\right\} \\
&\quad + C'_\kappa \frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}}(m+1)\left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \frac{m\sqrt{m}}{(N\lambda)^2} + \frac{\sqrt{m}\mathcal{N}(\lambda)}{N\lambda}\right) \\
&\le 2C'_\kappa M\sqrt{\lambda}\mathcal{C}_m\left\{\sqrt{m} + m\sqrt{\mathcal{C}_m}\right\} + 2C'_\kappa\frac{\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}} m\left(\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}} + \sqrt{m}\mathcal{C}_m\right).
\end{aligned}$$

Then the desired error bound in the $L^2_{\rho_X}$ metric with $C_\kappa = 2C'_\kappa$ follows. The proof for the error bound in the $\mathcal{H}_K$ metric is similar. The proof of Theorem 1 is complete. ■

Proof of Corollary 3 As in the proof of Corollary 8, the restriction (8) implies $\mathcal{N}(\lambda) \ge \frac{\lambda_1}{\lambda_1 + C_0}$ and $N\lambda \ge \frac{m\lambda_1}{(\lambda_1 + C_0)C_0}$. It follows that

$$\frac{m}{(N\lambda)^2} \le \frac{(\lambda_1 + C_0)C_0}{\lambda_1}\frac{1}{N\lambda} \le \frac{(\lambda_1 + C_0)^2 C_0}{\lambda_1^2}\frac{\mathcal{N}(\lambda)}{N\lambda}$$

and with $\widetilde{C}'_\kappa := \frac{(\lambda_1 + C_0)^2 C_0}{\lambda_1^2} + 1$,

$$\mathcal{C}_m \le \widetilde{C}'_\kappa \frac{\mathcal{N}(\lambda)}{N\lambda}.$$

Putting these bounds into Theorem 1 and applying the restriction $\frac{m\mathcal{N}(\lambda)}{N\lambda} \le C_0$, we know that

$$\begin{aligned}
E\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho &\le C_\kappa\left(\widetilde{C}'_\kappa\right)^{\frac{3}{2}}\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}}\left(1 + \sqrt{\frac{m\mathcal{N}(\lambda)}{N\lambda}}\right)\left\{M\sqrt{\frac{m\mathcal{N}(\lambda)}{N}} + \frac{m\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}}\right\} \\
&\le C_\kappa\left(\widetilde{C}'_\kappa\right)^{\frac{3}{2}}\left(1 + \sqrt{C_0}\right)\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}}\left\{M\sqrt{C_0}\sqrt{\lambda} + \frac{m\|f_\rho - f_\lambda\|_\rho}{\sqrt{N\lambda}}\right\}
\end{aligned}$$

and

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_K \;\leq\; C_\kappa\left(\widetilde{C}'_\kappa\right)^{\frac{3}{2}}\left(1+\sqrt{C_0}\right)\sqrt{\frac{\mathcal{N}(\lambda)}{N\lambda}}\left\{M\sqrt{C_0}+\frac{m\|f_\rho-f_\lambda\|_\rho}{\sqrt{N}\lambda}\right\}.$$

Then the desired error bounds hold by taking the constant $\widetilde{C}_\kappa = C_\kappa\left(\widetilde{C}'_\kappa\right)^{\frac{3}{2}}\left(1+\sqrt{C_0}\right)^2$. This proves Corollary 3. ■

Proof of Corollary 4 If

$$\mathcal{N}(\lambda)\leq C_0\lambda^{-\frac{1}{2\alpha}}, \qquad \forall\lambda>0$$

for some constant $C_0\geq 1$, then the choice $\lambda=\left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha\max\{2r,1\}+1}}$ satisfies (8). With this choice we also have

$$\frac{m\mathcal{N}(\lambda)}{N\lambda}\leq C_0\left(\frac{m}{N}\right)^{\frac{2\alpha(\max\{2r,1\}-1)}{2\alpha\max\{2r,1\}+1}}.$$

Since regularity condition (9) yields $\|f_\lambda-f_\rho\|_\rho\leq\|g_\rho\|_\rho\lambda^r$ by Lemma 21, we have by Corollary 3,

$$\begin{aligned}E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho \;\leq\; & \widetilde{C}_\kappa\sqrt{C_0}\frac{1}{\sqrt{m}}\left(\frac{m}{N}\right)^{\frac{\alpha(\max\{2r,1\}-1)}{2\alpha\max\{2r,1\}+1}}\left(\|g_\rho\|_\rho\left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha\max\{2r,1\}+1}(r-\frac{1}{2})}\frac{m}{\sqrt{N}}\right.\\ & \left.+M\left(\frac{m}{N}\right)^{\frac{\alpha}{2\alpha\max\{2r,1\}+1}}\right).\end{aligned}$$

The inequality $\left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha\max\{2r,1\}+1}(r-\frac{1}{2})}\frac{m}{\sqrt{N}}\leq\left(\frac{m}{N}\right)^{\frac{\alpha}{2\alpha\max\{2r,1\}+1}}$ is equivalent to

$$m^{1+\frac{2\alpha}{2\alpha\max\{2r,1\}+1}(r-1)}\leq N^{\frac{1}{2}+\frac{2\alpha}{2\alpha\max\{2r,1\}+1}(r-1)}$$

and it can be expressed as (10). Since (10) is valid, we have

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho\leq\widetilde{C}_\kappa\sqrt{C_0}\left(\|g_\rho\|_\rho+M\right)\frac{1}{\sqrt{m}}\left(\frac{m}{N}\right)^{\frac{\alpha\max\{2r,1\}}{2\alpha\max\{2r,1\}+1}}.$$

This proves the first desired convergence rate. The second rate follows easily. This proves Corollary 4. ■

Proof of Corollary 11 By Corollary 9, with the choice $\lambda=N^{-\frac{2\alpha}{4\alpha r+1}}$, we can immediately bound $\|f_{D,\lambda}-f_\rho\|_\rho$ as

$$E\left[\|f_{D,\lambda}-f_\rho\|_\rho\right]=O\left(N^{-\frac{2\alpha r}{4\alpha r+1}}\right).$$

The assumption $\mathcal{N}(\lambda)=O(\lambda^{-\frac{1}{2\alpha}})$ tells us that for some constant $C_0\geq 1$,

$$\mathcal{N}(\lambda)\leq C_0\lambda^{-\frac{1}{2\alpha}}, \qquad \forall\lambda>0.$$

So the choice $\lambda=N^{-\frac{2\alpha}{4\alpha r+1}}$ yields

$$\frac{m\mathcal{N}(\lambda)}{N\lambda}\leq C_0\frac{m\lambda^{-\frac{1+2\alpha}{2\alpha}}}{N}=C_0mN^{\frac{1+2\alpha}{4\alpha r+1}-1}=C_0mN^{\frac{2\alpha(1-2r)}{4\alpha r+1}}. \tag{44}$$

If m satisfies (13), then (8) is valid, and by Corollary 3,

$$\begin{aligned} E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho &\le \widetilde{C}_\kappa\sqrt{C_0}N^{\frac{\alpha(1-2r)}{4\alpha r+1}}\left(\lambda^r\|g_\lambda\|_\rho\frac{m}{\sqrt{N\lambda}}+M\sqrt{\lambda}\right)\\ &\le \widetilde{C}_\kappa\sqrt{C_0}\Big(\|g_\lambda\|_\rho+M\Big)\lambda^r\left(\frac{mN^{\frac{\alpha(1-2r)}{4\alpha r+1}}}{\sqrt{N\lambda}}+N^{\frac{\alpha(1-2r)}{4\alpha r+1}}\lambda^{\frac12-r}\right)\\ &= \widetilde{C}_\kappa\sqrt{C_0}\Big(\|g_\lambda\|_\rho+M\Big)N^{-\frac{2\alpha r}{4\alpha r+1}}\left(mN^{-\frac{2\alpha(2r-1)+\frac12}{4\alpha r+1}}+1\right). \end{aligned}$$

Since (13) is satisfied, we have

$$E\left\|\overline{f}_{D,\lambda}-f_{D,\lambda}\right\|_\rho \le 2\widetilde{C}_\kappa\sqrt{C_0}\Big(\|g_\lambda\|_\rho+M\Big)N^{-\frac{2\alpha r}{4\alpha r+1}},$$

and thereby

$$E\left[\left\|\overline{f}_{D,\lambda}-f_\rho\right\|_\rho\right]=O\left(N^{-\frac{2\alpha r}{4\alpha r+1}}\right).$$

This proves Corollary 11. ■

Acknowledgments

We thank the anonymous referees for their constructive suggestions. The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 11304114]. The corresponding author is Ding-Xuan Zhou.

Appendix

This appendix provides detailed proofs of the norm estimates stated in Lemmas 17 and 18 involving the approximation of L_K by $L_{K,D(x)}$. To this end, we need the following probability inequality for vector-valued random variables in (Pinelis, 1994).

Lemma 25 *For a random variable ξ on $(\mathcal{Z},\rho)$ with values in a separable Hilbert space $(H,\|\cdot\|)$ satisfying $\|\xi\|\le\widetilde{M}<\infty$ almost surely, and a random sample $\{z_i\}_{i=1}^s$ independent drawn according to ρ, there holds with confidence $1-\widetilde{\delta}$,*

$$\left\|\frac1s\sum_{i=1}^s[\xi(z_i)-E(\xi)]\right\|\le\frac{2\widetilde{M}\log(2/\widetilde{\delta})}{s}+\sqrt{\frac{2E(\|\xi\|^2)\log(2/\widetilde{\delta})}{s}}. \tag{45}$$

Proof of Lemma 17 We apply Lemma 25 to the random variable η_1 defined by

$$\eta_1(x)=(L_K+\lambda I)^{-1/2}\langle\cdot,K_x\rangle_K K_x,\qquad x\in\mathcal{X} \tag{46}$$

It takes values in $HS(\mathcal{H}_K)$, the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}_K$, with inner product $\langle A,B\rangle_{HS}=\mathrm{Tr}(B^TA)$. Here Tr denotes the trace of a (trace-class) linear operator. The norm is given by $\|A\|_{HS}^2=\sum_i\|Ae_i\|_K^2$ where $\{e_i\}$ is an orthonormal basis

of $\mathcal{H}_K$. The space $HS(\mathcal{H}_K)$ is a subspace of the space of bounded linear operators on $\mathcal{H}_K$, denoted as $(L(\mathcal{H}_K), \|\cdot\|)$, with the norm relations

$$\|A\| \leq \|A\|_{HS}, \qquad \|AB\|_{HS} \leq \|A\|_{HS}\|B\|. \tag{47}$$

Now we use effective dimensions to estimate norms involving η_1. The random variable η_1 defined by (46) has mean $E(\eta_1) = (L_K + \lambda I)^{-1/2} L_K$ and sample mean $(L_K + \lambda I)^{-1/2} L_{K,D(x)}$. Recall the set of normalized (in $\mathcal{H}_K$) eigenfunctions $\{\varphi_i\}_i$ of L_K. It is an orthonormal basis of $\mathcal{H}_K$. If we regard L_K as an operator on $L^2_{\rho_X}$, the normalized eigenfunctions in $L^2_{\rho_X}$ are $\{\frac{1}{\sqrt{\lambda_i}}\varphi_i\}_i$ and they form an orthonormal basis of the orthogonal complement of the eigenspace associated with eigenvalue 0. By the Mercer Theorem, we have the following uniform convergent Mercer expansion

$$K(x,y) = \sum_i \lambda_i \frac{1}{\sqrt{\lambda_i}}\varphi_i(x)\frac{1}{\sqrt{\lambda_i}}\varphi_i(y) = \sum_i \varphi_i(x)\varphi_i(y). \tag{48}$$

Take the orthonormal basis $\{\varphi_i\}_i$ of $\mathcal{H}_K$. By the definition of the HS norm, we have

$$\|\eta_1(x)\|^2_{HS} = \sum_i \left\| (L_K + \lambda I)^{-1/2} \langle \cdot, K_x\rangle_K K_x\varphi_i \right\|^2_K.$$

For a fixed i,

$$\langle \cdot, K_x\rangle_K K_x\varphi_i = \varphi_i(x)K_x,$$

and $K_x \in \mathcal{H}_K$ can be expended by the orthonormal basis $\{\varphi_\ell\}_\ell$ as

$$K_x = \sum_\ell \langle \varphi_\ell, K_x\rangle_K \varphi_\ell = \sum_\ell \varphi_\ell(x)\varphi_\ell. \tag{49}$$

Hence

$$\begin{aligned}
\|\eta_1(x)\|^2_{HS} &= \sum_i \left\| \varphi_i(x) \sum_\ell \varphi_\ell(x) \left(L_K + \lambda I\right)^{-1/2} \varphi_\ell \right\|^2_K \\
&= \sum_i \left\| \varphi_i(x) \sum_\ell \varphi_\ell(x) \frac{1}{\sqrt{\lambda_\ell + \lambda}} \varphi_\ell \right\|^2_K = \sum_i \left(\varphi_i(x)\right)^2 \sum_\ell \frac{\left(\varphi_\ell(x)\right)^2}{\lambda_\ell + \lambda}.
\end{aligned}$$

Combining this with (48), we see that

$$\|\eta_1(x)\|^2_{HS} = K(x,x) \sum_\ell \frac{\left(\varphi_\ell(x)\right)^2}{\lambda_\ell + \lambda}, \qquad \forall x \in \mathcal{X} \tag{50}$$

and

$$E\left[\|\eta_1(x)\|^2_{HS}\right] \leq \kappa^2 E\left[\sum_\ell \frac{\left(\varphi_\ell(x)\right)^2}{\lambda_\ell + \lambda}\right] = \kappa^2 \sum_\ell \frac{\int_{\mathcal{X}} \left(\varphi_\ell(x)\right)^2 d\rho_X}{\lambda_\ell + \lambda}.$$

But

$$\int_{\mathcal{X}} \left(\varphi_\ell(x)\right)^2 d\rho_X = \|\varphi_\ell\|^2_{L^2_{\rho_X}} = \left\| \sqrt{\lambda_\ell} \frac{1}{\sqrt{\lambda_\ell}} \varphi_\ell \right\|^2_{L^2_{\rho_X}} = \lambda_\ell. \tag{51}$$

So we have

$$E\left[\|\eta_1\|_{HS}^2\right] \leq \kappa^2 \sum_{\ell} \frac{\lambda_\ell}{\lambda_\ell + \lambda} = \kappa^2 \mathrm{Tr}\left(\left(L_K + \lambda I\right)^{-1} L_K\right) = \kappa^2 \mathcal{N}(\lambda) \tag{52}$$

and

$$E\left\|\frac{1}{|D|}\sum_{x\in D(x)} \eta_1(x) - E[\eta_1]\right\|_{HS}^2 = E\left[\left\|\left(L_K + \lambda I\right)^{-1/2}\left\{L_K - L_{K,D(x)}\right\}\right\|_{HS}^2\right] \leq \frac{\kappa^2 \mathcal{N}(\lambda)}{|D|}.$$

Then our desired inequality in Part (a) follows from the first inequality of (47).

From (49) and (50), we find a bound for η_1 as

$$\|\eta_1(x)\|_{HS} \leq \kappa \frac{1}{\sqrt{\lambda}} \sqrt{\sum_{\ell} \left(\varphi_\ell(x)\right)^2} \leq \frac{\kappa}{\sqrt{\lambda}} \sqrt{K(x,x)} \leq \frac{\kappa^2}{\sqrt{\lambda}}, \qquad \forall x \in \mathcal{X}.$$

Applying Lemma 25 to the random variable η_1 with $\widetilde{M} = \frac{\kappa^2}{\sqrt{\lambda}}$, we know by (47) that with confidence at least $1-\delta$,

$$\begin{aligned} \left\|E[\eta_1] - \frac{1}{|D|}\sum_{x\in D(x)} \eta_1(x)\right\| &\leq \left\|E[\eta_1] - \frac{1}{|D|}\sum_{x\in D(x)} \eta_1(x)\right\|_{HS} \\ &\leq \frac{2\kappa^2 \log(2/\delta)}{|D|\sqrt{\lambda}} + \sqrt{\frac{2\kappa^2 \mathcal{N}(\lambda)\log(2/\delta)}{|D|}}. \end{aligned}$$

Writing the above bound by taking a factor $\frac{2\kappa \log\left(2/\delta\right)}{\sqrt{|D|}}$, we get the desired bound (29).

Recall $\mathcal{B}_{|D|,\lambda}$ defined by (30). Apply the formula (42) for nonnegative random variables to $\xi = \left\|\left(L_K + \lambda I\right)^{-1/2}\left\{L_K - L_{K,D(x)}\right\}\right\|^d$ and use the bound

$$\mathrm{Prob}\left[\xi > t\right] = \mathrm{Prob}\left[\xi^{\frac{1}{d}} > t^{\frac{1}{d}}\right] \leq 2\exp\left\{-\frac{t^{\frac{1}{d}}}{\mathcal{B}_{|D|,\lambda}}\right\}$$

derived from (29) for $t \geq \log^d 2\mathcal{B}_{|D|,\lambda}$. We find

$$E\left[\left\|\left(L_K + \lambda I\right)^{-1/2}\left\{L_K - L_{K,D(x)}\right\}\right\|^d\right] \leq \log^d 2\mathcal{B}_{|D|,\lambda} + \int_0^\infty 2\exp\left\{-\frac{t^{\frac{1}{d}}}{\mathcal{B}_{|D|,\lambda}}\right\} dt.$$

The second term on the right-hand side of above equation equals $2d\mathcal{B}_{|D|,\lambda}^d \int_0^\infty u^{d-1}\exp\{-u\}\,du$. Then the desired bound in Part (c) follows from $\int_0^\infty u^{d-1}\exp\{-u\}\,du = \Gamma(d)$ and the lemma is proved. ■

Proof of Lemma 18 Consider the random variable η_2 defined by

$$\eta_2(z) = \left(L_K + \lambda I\right)^{-1/2}(K_x), \qquad z = (x,y) \in \mathcal{Z}. \tag{53}$$

It takes values in $\mathcal{H}_K$. By (49), it satisfies

$$\|\eta_2(z)\|_K = \left\| (L_K + \lambda I)^{-1/2} \left(\sum_\ell \varphi_\ell(x) \varphi_\ell \right) \right\|_K = \left(\sum_\ell \frac{(\varphi_\ell(x))^2}{\lambda_\ell + \lambda} \right)^{1/2}.$$

So

$$E\left[\left\| (L_K + \lambda I)^{-1/2} (K_x) \right\|_K^2 \right] = E\left[\sum_\ell \frac{(\varphi_\ell(x))^2}{\lambda_\ell + \lambda} \right] = \mathcal{N}(\lambda).$$

This is the statement of Part (a).

We also see from (49) that for $x \in \mathcal{X}$,

$$\|\eta_2(z)\|_K = \left(\sum_\ell \frac{(\varphi_\ell(x))^2}{\lambda_\ell + \lambda} \right)^{1/2} \le \frac{1}{\sqrt{\lambda}} \left(\sum_\ell (\varphi_\ell(x))^2 \right)^{1/2} = \frac{1}{\sqrt{\lambda}} \sqrt{K(x,x)} \le \frac{\kappa}{\sqrt{\lambda}}.$$

This verifies the statement of Part (b).

For Part (c), we consider another random variable η_3 defined by

$$\eta_3(z) = (L_K + \lambda I)^{-1/2} (g(z) K_x), \qquad z = (x, y) \in \mathcal{Z}. \tag{54}$$

It takes values in $\mathcal{H}_K$ and satisfies

$$\|\eta_3(z)\|_K = |g(z)| \left\| (L_K + \lambda I)^{-1/2} (K_x) \right\|_K = |g(z)| \left(\sum_\ell \frac{(\varphi_\ell(x))^2}{\lambda_\ell + \lambda} \right)^{1/2}.$$

So

$$\|\eta_3(z)\|_K \le \frac{\kappa \|g\|_\infty}{\sqrt{\lambda}}, \qquad z \in \mathcal{Z}$$

and

$$E\left[\|\eta_3\|_K^2 \right] \le \|g\|_\infty^2 E\left[\sum_\ell \frac{(\varphi_\ell(x))^2}{\lambda_\ell + \lambda} \right] = \|g\|_\infty^2 \mathcal{N}(\lambda).$$

Applying Lemma 25 verifies the statement in Part (c). The proof of the lemma is complete. ■

References

F. Bach. Sharp analysis of low-rank kernel matrix approximations. *Proceedings of the 26th Annual Conference on Learning Theory*, 185-209, 2013.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23:52-72, 2007.

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

G. Blanchard and N. Krämer. Convergence rates for kernel conjugate gradient for random design regression. *Analysis and Applications*, 14:763-794, 2016.

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

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8:161-183, 2010.

D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143-1175, 2004.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and X. Lin. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165-202, 2012.

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

E. De Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10:455-479, 2010.

E. De Vito, V. Umanità, and S. Villa. An extension of Mercer theorem to vector-valued measurable kernels. *Applied and Computational Harmonic Analysis*, 34:339-351, 2013.

D.E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators.* Cambridge University Press, Cambridge, 1996.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advance in Computional Mathematics*, 13:1-50, 2000.

J. Fan, T. Hu, Q. Wu, and D.X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis*, 41:164-189, 2016.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243-264, 2002.

A. Gittens and M. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17:1-65, 2016.

X. Guo and D. X. Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Applied and Computational Harmonic Analysis*, 32:389-400, 2012.

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

L. Györfy, M. Kohler, A. Krzyzak, H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag, Berlin, 2002.

T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13:437-455, 2015.

J. H. Lin, L. Rosasco, and D. X. Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(77): 1-38, 2016.

J. H. Lin and D. X. Zhou. Learning theory of randomized Kaczmarz algorithm. *Journal of Machine Learning Research*, 16:3341-3365, 2015.

S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526-565, 2010.

M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *Journal of Machine Learning Research*, 17: 1-44, 2016.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22:1679-1706, 1994.

G. Raskutti, M. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335-366, 2014.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 1657-1665, 2015.

B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *IEEE Transactions on Information Theory*, 10:1299-1319, 1998.

O. Shamir and N. Srebro. Distributed stochastic optimization and learning. *In the 52nd Annual Allerton Conference on Communication, Control and Computing*, 2014.

W.J. Shen, H.S. Wong, Q.W. Xiao, X. Guo, and S. Smale. Introduction to the peptide binding problem of computational immunology: new results. *Foundations of Computational Mathematics*, 14:951-984, 2014.

L. Shi, Y.L. Feng, and D.X. Zhou. Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31:286-302, 2011.

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

I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, New York, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *In Proceedings of the 22nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.), pp. 79-93, 2009.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363-417, 2012.

P. Thomann, I. Steinwart, I. Blaschzyk, and M. Meister. Spatial decompositions for large scale SVMs. ArXiv:1612.00374v1, 2016.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47:2516-2532, 2001.

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

Q. Wu and D. X. Zhou. Learning with sample dependent hypothesis spaces. *Computers and Mathematics with Applications*, 56:2896-2907, 2008.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289-315, 2007.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17:2077-2098, 2005.

Y. C. Zhang, J. Duchi, and M. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321-3363, 2013.

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

D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18:739-767, 2002.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743-1752, 2003.

Z. H. Zhou, N. V. Chawla, Y. Jin, and G. J. Williams. Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Computational Intelligence Magazine*, 9:62-74, 2014.