# Learning theory of distributed spectral algorithms

2017 Inverse Problems 33 074009

(http://iopscience.iop.org/0266-5611/33/7/074009)

Inverse Problems **33** (2017) 074009 (29pp) https://doi.org/10.1088/1361-6420/aa72b2

# Learning theory of distributed spectral algorithms

**Zheng-Chu Guo**[1], **Shao-Bo Lin**[2] **and Ding-Xuan Zhou**[2]

[1] School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, People's Republic of China
[2] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China

E-mail: guozhengchu@zju.edu.cn, sblin1983@gmail.com and mazhou@cityu.edu.hk

Received 1 January 2016, revised 25 April 2017
Accepted for publication 12 May 2017
Published 21 June 2017

**Abstract**

Spectral algorithms have been widely used and studied in learning theory and inverse problems. This paper is concerned with distributed spectral algorithms, for handling big data, based on a divide-and-conquer approach. We present a learning theory for these distributed kernel-based learning algorithms in a regression framework including nice error bounds and optimal minimax learning rates achieved by means of a novel integral operator approach and a second order decomposition of inverse operators. Our quantitative estimates are given in terms of regularity of the regression function, effective dimension of the reproducing kernel Hilbert space, and qualification of the filter function of the spectral algorithm. They do not need any eigenfunction or noise conditions and are better than the existing results even for the classical family of spectral algorithms.

Keywords: distributed learning, spectral algorithm, integral operator, learning rate

## 1. Introduction

In the big data era, data of high volume may necessarily be stored distributively across multiple servers rather than on one machine. This makes many traditional learning algorithms requiring access to the entire data set infeasible. Distributed learning, based on a divide-and-conquer approach, provides a promising way to tackle this problem and therefore has recently triggered enormous research activities [13, 14, 21]. This strategy applies a specific learning algorithm to one data subset on each server, to produce an individual output (function), and then synthesizes a global output by utilizing some average of the individual outputs. The learning performance of distributed learning has been observed in many practical applications to be as good as that of a big machine which could process the whole data. We are interested in error analysis of such learning algorithms.

Distributed learning with kernel-based regularized least squares was studied in [22], and optimal (minimax) learning rates were derived with a matrix analysis approach by making full use of the linearity of the algorithm under some assumptions on eigenfunctions of the integral operator associated with the kernel. These assumptions were successfully removed in the recent paper [11] with an integral operator approach and also by the linearity of the least squares algorithm.

In this paper, we consider a family of more general learning algorithms, spectral algorithms, and present a learning theory for distributed spectral algorithms. In particular, optimal learning rates will be provided by means of a novel integral operator approach. As a by-product, for the classical spectral algorithms for regression, we shall improve the existing learning rates in the literature by removing a logarithmic factor and give optimal learning rates.

Spectral algorithms were proposed to solve ill-posed linear inverse problems (see e.g. [8]) and employed [2, 12] for regression by noticing connections between learning theory and inverse problems [7]. For learning functions on a compact metric space $\mathcal{X}$ (input space), a spectral algorithm is defined in terms of a Mercer (continuous, symmetric and positive semidefinite) kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ with $\kappa = \sqrt{\sup_{x\in\mathcal{X}} K(x,x)}$ and a filter function $g_\lambda : [0, \kappa^2] \to \mathbf{R}$ with a parameter $\lambda > 0$ acting on spectra of empirical integral operators. For a sample $D = \{(x_i, y_i)\}_{i=1}^N$ with $y_i \in \mathcal{Y} \subseteq \mathbf{R}$ (output space), the empirical integral operator $L_{K,D}$ associated with the kernel $K$ and the input data $D(\mathbf{x}) = \{x_1, \cdots, x_N\}$ is defined on the reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \langle , \rangle_K)$ associated with $K$ by

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

where $K_x = K(\cdot, x)$ and $|D| = N$ denotes the cardinality of $D$. Given the kernel $K$, the filter function $g_\lambda$, and the sample $D$, the spectral algorithm is defined by

$$f_{D,\lambda} = g_\lambda(L_{K,D}) \frac{1}{|D|} \sum_{(x,y)\in D} y K_x. \tag{1}$$

Here $g_\lambda(L_{K,D})$ is an operator on $\mathcal{H}_K$ defined by spectral calculus: if $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})_i\}$ is a set of normalized eigenpairs of $L_{K,D}$ with the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ forming an orthonormal basis of $\mathcal{H}_K$, then $g_\lambda(L_{K,D}) = \sum_i g_\lambda(\sigma_i^{\mathbf{x}}) \phi_i^{\mathbf{x}} \otimes \phi_i^{\mathbf{x}} = \sum_i g_\lambda(\sigma_i^{\mathbf{x}}) \langle \cdot, \phi_i^{\mathbf{x}} \rangle_K \phi_i^{\mathbf{x}}$. This operator acts on the function $f_\rho^D := \frac{1}{|D|} \sum_{(x,y)\in D} y K_x = \frac{1}{N} \sum_{i=1}^N y_i K_{x_i} \in \mathcal{H}_K$ to produce the output function $f_{D,\lambda} \in \mathcal{H}_K$ of spectral algorithm (1).

We take a regression framework in learning theory [6] modelled with a probability measure $\rho$ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and assume throughout the paper that a random sample $D = \{(x_i, y_i)\}_{i=1}^N$ is drawn independently according to $\rho$. The learning problem for regression aims at estimating the regression function $f_\rho : \mathcal{X} \to \mathbf{R}$ defined by conditional means as

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

where $\rho(y|x)$ is the conditional distribution of $\rho$ at $x$. The output function $f_{D,\lambda}$ produced by spectral algorithm (1) is a good estimator of the regression function $f_\rho$ when the filter function $g_\lambda$ is chosen properly and the size $N$ of the sample $D$ is large enough.

As a kernel method, spectral algorithm (1) implements learning tasks in the hypothesis space $\mathcal{H}_K$. What is special about this hypothesis space is its reproducing property asserting that

$f(x)=\langle f,K_x\rangle_K$ for any $f\in\mathcal{H}_K$ and $x\in\mathcal{X}$. It tells us that the empirical integral operator satisfies $L_{K,D}(f)=\frac{1}{|D|}\sum_{x\in D(\mathbf{x})}\langle f,K_x\rangle_K K_x$. So $L_{K,D}$ is a finite-rank positive operator on $\mathcal{H}_K$, and its spectrum is contained in the interval $[0,\kappa^2]$ since $\|K_x\|_K=\sqrt{K(x,x)}\leqslant\kappa$. Expressing $L_{K,D}$ in terms of its normalized eigenpairs $\{(\sigma_i^{\mathbf{x}},\phi_i^{\mathbf{x}})_i\}$ as $L_{K,D}=\sum_i\sigma_i^{\mathbf{x}}\langle\cdot,\phi_i^{\mathbf{x}}\rangle_K\phi_i^{\mathbf{x}}$, we see that the filter function $g_\lambda$ acting on $L_{K,D}$ provides an approximate inverse $g_\lambda(L_{K,D})=\sum_i g_\lambda(\sigma_i^{\mathbf{x}})\langle\cdot,\phi_i^{\mathbf{x}}\rangle_K\phi_i^{\mathbf{x}}$ of $L_{K,D}$ by regularizing the eigenvalue reciprocal $\frac{1}{\sigma_i^{\mathbf{x}}}$ to $g_\lambda(\sigma_i^{\mathbf{x}})$.

Let $\rho_X$ be the marginal distribution of $\rho$ on $\mathcal{X}$ and $(L^2_{\rho_X},\|\cdot\|_\rho)$ be the Hilbert space of $\rho_X$ square integrable functions on $\mathcal{X}$. Define the integral operator $L_K$ on $\mathcal{H}_K$ or $L^2_{\rho_X}$ associated with the Mercer kernel $K$ by

$$L_K(f)=\int_X f(x)K_x\mathrm{d}\rho_X.$$

When the sample size $N$ is large, the function $f_\rho^D=\frac{1}{N}\sum_{i=1}^N y_iK_{x_i}\in\mathcal{H}_K$ is a good approximation of its mean $\int_{\mathcal{Z}}yK_x\mathrm{d}\rho=\int_{\mathcal{X}}f_\rho(x)K_x\mathrm{d}\rho_X=L_K(f_\rho)$ and the empirical integral operator $L_{K,D}$ approximates $L_K$ well. Hence spectral algorithm (1) produces a good estimator $f_{D,\lambda}=g_\lambda(L_{K,D})f_\rho^D$ of the regression function $f_\rho$ when $f_\rho\in\mathcal{H}_K$ and $g_\lambda(L_{K,D})$ is an approximate inverse of $L_{K,D}$ or $L_K$.

Let us illustrate the role in inverting $L_{K,D}$ approximately of the filter function $g_\lambda$ as an approximation of the reciprocal function $\frac{1}{\sigma}$ by describing two typical spectral algorithms, Tikhonov regularization or kernel-based regularized least squares algorithm and spectral cut-off [2, 12, 17].

**Example 1 (Tikhonov regularization).** This spectral algorithm has a filter function $g_\lambda:[0,\kappa^2]\to\mathbf{R}$ with a parameter $\lambda>0$ given by $g_\lambda(\sigma)=\frac{1}{\sigma+\lambda}$. Its output function $f_{D,\lambda}$ equals the solution to the following regularized least squares minimization problem

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

To see this, we use the sampling operator $S_{\mathbf{x}}:\mathcal{H}_K\to\mathbf{R}^N$ defined [15] by $S_{\mathbf{x}}(f)=(f(x_i))_{i=1}^N=(\langle f,K_{x_i}\rangle_K)_{i=1}^N$, where the inner product on $\mathbf{R}^N$ is the normalized one given by $\langle c,\tilde{c}\rangle_{\mathbf{R}^N}=\frac{1}{N}\sum_{i=1}^N c_i\tilde{c}_i$ for $c=(c_i)_{i=1}^N,\tilde{c}=(\tilde{c}_i)_{i=1}^N\in\mathbf{R}^N$. If we denote $\mathbf{y}=(y_i)_{i=1}^N\in\mathbf{R}^N$, then the minimization problem (2) can be expressed as

$$f_{D,\lambda}=\arg\min_{f\in\mathcal{H}_K}\left\{\|S_{\mathbf{x}}(f)-\mathbf{y}\|_{\mathbf{R}^N}^2+\lambda\|f\|_K^2\right\},$$

and its unique solution [8] is $f_{D,\lambda}=\left(S_{\mathbf{x}}^TS_{\mathbf{x}}+\lambda I\right)^{-1}S_{\mathbf{x}}^T\mathbf{y}$. Note that the adjoint operator $S_{\mathbf{x}}^T:\mathbf{R}^N\to\mathcal{H}_K$ is given by $S_{\mathbf{x}}^T(c)=\frac{1}{N}\sum_{i=1}^N c_iK_{x_i}$ for $c\in\mathbf{R}^N$. Hence $S_{\mathbf{x}}^TS_{\mathbf{x}}(f)=\frac{1}{N}\sum_{i=1}^N f(x_i)K_{x_i}=L_{K,D}(f)$ for $f\in\mathcal{H}_K$, and $S_{\mathbf{x}}^T\mathbf{y}=\frac{1}{N}\sum_{i=1}^N y_iK_{x_i}=f_\rho^D$. These identities tell us that $f_{D,\lambda}$ defined by (2) equals $(L_{K,D}+\lambda I)^{-1}f_\rho^D=g_\lambda(L_{K,D})f_\rho^D$, the expression in (1) when $g_\lambda(\sigma)=\frac{1}{\lambda+\sigma}$.

It is well-known that when $\lambda=0$, the minimization problem (2) is ill-posed and its solution is not unique in general. This can also be seen from the expression (1) since the operator $L_{K,D}=\sum_i\sigma_i^{\mathbf{x}}\langle\cdot,\phi_i^{\mathbf{x}}\rangle_K\phi_i^{\mathbf{x}}$ is not invertible and the operator $g_0(L_{K,D})=L_{K,D}^{-1}$ is not well-defined. When $\lambda>0$, the minimization problem (2) becomes well-posed and its unique solution is given by means of the operator $g_\lambda(L_{K,D})=(L_{K,D}+\lambda I)^{-1}=\sum_i\frac{1}{\sigma_i^{\mathbf{x}}+\lambda}\langle\cdot,\phi_i^{\mathbf{x}}\rangle_K\phi_i^{\mathbf{x}}$.

Here the filter function $g_\lambda$ maps the eigenvalue $\sigma_i^{\mathbf{x}} \in [0, \kappa^2]$ in the spectrum of $L_{K,D}$ to a regularized reciprocal $\frac{1}{\sigma_i^{\mathbf{x}}+\lambda}$ instead of the reciprocal itself.

**Example 2 (Spectral cut-off).** The spectral cut-off or truncated singular value decomposition has a filter function $g_\lambda$ with a parameter $\lambda > 0$ given by

$$g_\lambda(\sigma) = \begin{cases} \frac{1}{\sigma}, & \text{if } \sigma \geqslant \lambda, \\ 0, & \text{if } \sigma < \lambda. \end{cases}$$

Here the filter function $g_\lambda$ eliminates the eigenvalues of $L_{K,D}$ in the spectrum interval $[0, \lambda)$ and maps those eigenvalues $\sigma_i^{\mathbf{x}}$ on $[\lambda, \kappa^2]$ to their reciprocals $\frac{1}{\sigma_i^{\mathbf{x}}}$. So the operator $g_\lambda(L_{K,D})$ is given by spectral calculus as $g_\lambda(L_{K,D}) = \sum_{i:\sigma_i^{\mathbf{x}} \geqslant \lambda} \frac{1}{\sigma_i^{\mathbf{x}}} \langle \cdot, \phi_i^{\mathbf{x}} \rangle_K \phi_i^{\mathbf{x}}$, and it provides an approximate inverse of $L_{K,D}$ when the truncation parameter $\lambda$ is small.

We may regard the least squares minimization problem (2) in example 1 as a Tikhonov regularization solution [8] to an ill-posed linear inverse problem with noisy data $D = \{(x_i, y_i)\}_{i=1}^N$. Both of the output and input have noise. The data-free limit of (2) takes the form

$$f_\lambda = \arg \min_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}.$$

If we use the inclusion operator $S_\rho : \mathcal{H}_K \to L^2_{\rho_X}$ defined by $S_\rho(f)(x) = f(x) = \langle f, K_x \rangle_K$, then we know that $f_\lambda = \left(S_\rho^T S_\rho + \lambda I\right)^{-1} S_\rho^T f_\rho$. But $S_\rho^T(h) = \int_{\mathcal{X}} h(x) K_x \mathrm{d}\rho_X$. So $S_\rho^T S_\rho(f) = L_K(f)$. Since $L_K$ is a compact positive operator on $\mathcal{H}_K$, $L_K + \lambda I$ is invertible for any $\lambda > 0$. When $f_\rho \in \mathcal{H}_K$, we have $f_\lambda = (L_K + \lambda I)^{-1} L_K(f_\rho)$, the data-free limit of (1) given by $g_\lambda(L_K) L_K(f_\rho)$ with $g_\lambda(\sigma) = \frac{1}{\lambda + \sigma}$. This expression for the data-free limit hints the so-called saturation phenomenon [12, 20] for the regularized least squares algorithm: its learning rate $\mathcal{O}(N^{-\frac{2r}{2r+\beta}})$, stated with a complexity parameter $\beta > 0$ and a regularity parameter $r > 0$ given in (8) and (6) below, is saturated at $r = 1$ and ceases to improve when the regularity of the regression function goes beyond as $r > 1$. Some other members in the family of spectral algorithms may have improved learning rates together with some other advantages in solving various learning tasks. For example, the Landweber iteration [12] (or gradient descent [20]), with the filter function $g_\lambda(\sigma) = \sum_{i=1}^{t-1} (1 - \sigma)^i$ parameterized by $\lambda = \frac{1}{t}$ for an integer $t \in \mathbf{N}$, can have improved learning rates, from $\mathcal{O}(N^{-\frac{2}{2+\beta}})$ to $\mathcal{O}(N^{-\frac{2r}{2r+\beta}})$ with arbitrary $r > 1$. A corresponding parameter for a general spectral algorithm, threshold for improvements of learning rates, is called the qualification $\nu_g > 0$ (to be defined below) which can even be infinity (for spectral cut-off). Deriving learning rates of spectral algorithm (1) requires rigorous analysis for the approximation of $(L_K + \lambda I)^{-1}$ by $(L_{K,D} + \lambda I)^{-1}$, which is key in our study.

The spectral algorithms (1) are implemented by spectral calculus of the empirical integral operator $L_{K,D}$ or singular value decompositions of the Gramian matrix $(K(x_i, x_j))_{i,j=1}^N$. When the data set $D$ comes distributively or has a very large sample size $N$, it is natural for us to consider distributed learning with spectral algorithms. Let $D = \cup_{j=1}^m D_j$ be a disjoint union of data subsets $D_j$. The distributed learning algorithms studied in this paper take the form of a weighted average of the output functions $\{f_{D_j,\lambda}\}$ produced by the spectral algorithms (1) implemented on individual data subsets $\{D_j\}$ as

$$\overline{f}_{D,\lambda} = \sum_{j=1}^m \frac{|D_j|}{|D|} f_{D_j,\lambda} = \sum_{j=1}^m \frac{|D_j|}{|D|} g_\lambda(L_{K,D_j}) \frac{1}{|D_j|} \sum_{(x,y) \in D_j} y K_x. \tag{3}$$

In the special case with $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$, (3) coincides with the algorithm of distributed learning with regularized least squares considered in [11, 22].

The main purpose of this paper is to derive optimal learning rates for the distributed spectral algorithms (3) by a novel integral operator method. We shall show that the distributed spectral algorithms (3) can achieve the same learning rates as the spectral algorithms (1) (acting on the whole data set), provided $|D_j|$ is not too small for each $D_j$.

## 2. Main results

In this section, we present optimal learning rates of the distributed spectral algorithms (3). These learning rates are stated in terms of the regularity of the regression function, the complexity of the RKHS, and the qualification of the filter function, a characteristic of a spectral algorithm making our analysis essentially different from the previous work for the regularized least squares in [11, 22].

### *2.1. Filter function and examples of spectral algorithms*

The learning performance of a spectral algorithm depends on its filter function $g_\lambda$ with qualification $\nu_g \geqslant \frac{1}{2}$ defined as follows.

**Definition 1.** We say that $g_\lambda : [0, \kappa^2] \to \mathbf{R}$, with $0 < \lambda \leqslant \kappa^2$, is a filter function with qualification $\nu_g \geqslant \frac{1}{2}$ if there exists a positive constant $b$ independent of $\lambda$ such that

$$\sup_{0<\sigma\leqslant\kappa^2} |g_\lambda(\sigma)| \leqslant \frac{b}{\lambda}, \qquad \sup_{0<\sigma\leqslant\kappa^2} |g_\lambda(\sigma)\sigma| \leqslant b, \tag{4}$$

and

$$\sup_{0<\sigma\leqslant\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leqslant \gamma_\nu \lambda^\nu, \qquad \forall\ 0 < \nu \leqslant \nu_g, \tag{5}$$

where $\gamma_\nu > 0$ is a constant depending only on $\nu \in (0, \nu_g]$.

In examples 1 and 2, we may take the constants as $b = 1, \gamma_\nu = 1$, while the qualification is $\nu_g = 1$ for Tikhonov regularization and $\nu_g = \infty$ for spectral cut-off.

Below are some other examples of spectral algorithms with different filter functions.

**Example 3 (Landweber iteration).** If $g_\lambda(\sigma) = \sum_{i=0}^{t-1}(1-\sigma)^i$ with $\lambda = \frac{1}{t}$ for some $t \in \mathbf{N}$, then algorithm (1) corresponds to the Landweber iteration or gradient descent algorithm. For this filter function, we have $b = 1$ and the qualification $\nu_g$ is infinite. The constant $\gamma_\nu$ equals 1 if $\nu \in (0, 1]$ and $\nu^\nu$ if $\nu > 1$.

**Example 4 (Accelerated Landweber iteration).** Accelerated Landweber iteration or semi-iterative regularization can be regarded as a generalization of the Landweber iteration where the filter function is defined as $g_\lambda(\sigma) = p_t(\sigma)$ with $\lambda = t^{-2}, t \in \mathbb{N}$, and $p_t$ a polynomial of degree $t-1$. Here $b = 2$ and the qualification is usually finite [8]. A special case of the accelerated Landweber iteration is the $\nu$ method. We refer the readers to [2, 8, 12] and references therein for more details about the $\nu$ method and accelerated Landweber iteration.

### *2.2. Optimal minimax rates of distributed spectral algorithms*

Our error bounds for the distributed spectral algorithms are based on the following regularity condition

$$f_\rho = L_K^r(u_\rho) \text{ for some } r > 0 \text{ and } u_\rho \in L^2_{\rho X}, \tag{6}$$

where $L_K^r$ denotes the $r$th power of $L_K$ on $L^2_{\rho_X}$ since $L_K : L^2_{\rho X} \to L^2_{\rho X}$ is a compact and positive operator.

We shall use the *effective dimension* $\mathcal{N}(\lambda)$ to measure the complexity of $\mathcal{H}_K$ with respect to $\rho_X$, which is defined to be the trace of the operator $(\lambda I + L_K)^{-1}L_K$, that is

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

Throughout the paper we assume that for some constant $M > 0, |y| \leqslant M$ almost surely. Our analysis can be extended to more general situations by assuming some exponential decay or moment conditions [4].

In section 6, we shall prove the following error bounds in expectation for the distributed spectral algorithms (3). Recall that $D = \cup_{j=1}^m D_j$ is a disjoint union of data subsets $D_j$.

**Theorem 1.** *If the regularity condition* (6) *holds with* $1/2 \leqslant r \leqslant \nu_g$*, then there exists a constant* $\tilde{C}$ *independent of* $m$ *or* $|D_j|$ *such that for* $1/2 \leqslant r \leqslant \min\{3/2, \nu_g\}$,

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] \leqslant \tilde{C} \sum_{j=1}^m \frac{|D_j|}{|D|} \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^{\max\{2,2r\}} \left(\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} + \lambda^{2r}\right),$$

*and for* $3/2 < r \leqslant \nu_g$,

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2]$$
$$\leqslant \tilde{C} \sum_{j=1}^m \frac{|D_j|}{|D|} \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^3 \left(\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} + \frac{\lambda}{|D|} + \lambda^{2r} + \lambda^2\mathcal{B}^2_{|D_j|,\lambda} + \frac{\lambda^3}{|D_j|}\right).$$

*Here* $\mathcal{B}_{|D_j|,\lambda}$ *is a quantity defined by*

$$\mathcal{B}_{|D_j|,\lambda} = \frac{2\kappa}{\sqrt{|D_j|}}\left\{\frac{\kappa}{\sqrt{|D_j|\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}. \tag{7}$$

To obtain explicit learning rates, we need to quantify the effective dimension $\mathcal{N}(\lambda)$ with a parameter $0 < \beta \leqslant 1$ and a constant $C_0 > 0$ as

$$\mathcal{N}(\lambda) \leqslant C_0\lambda^{-\beta}, \qquad \forall \lambda > 0. \tag{8}$$

The condition (8) with $\beta = 1$ is always satisfied by taking the constant $C_0 = \mathrm{Tr}(L_K) \leqslant \kappa^2$.

For $0 < \beta < 1$, the condition (8) is slightly more general than the eigenvalue decaying assumption in the literature (e.g. [4]). Indeed, let $\{(\lambda_\ell, \phi_\ell)\}_\ell$ be a set of normalized eigenpairs of $L_K$ on $\mathcal{H}_K$ with $\{\phi_\ell\}_{\ell=1}^\infty$ forming an orthonormal basis of $\mathcal{H}_K$, and let

$$L_K = \sum_{\ell=1}^\infty \lambda_\ell \langle \cdot, \phi_\ell \rangle_K \phi_\ell$$

be the spectral decomposition. If $\lambda_n \leqslant c_0 n^{-1/\beta}$ for some $0 < \beta < 1$ and $c_0 \geqslant 1$, then

$$\mathcal{N}(\lambda) = \sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{\lambda + \lambda_\ell} \leqslant \sum_{\ell=1}^{\infty} \frac{c_0 \ell^{-1/\beta}}{\lambda + c_0 \ell^{-1/\beta}} = \sum_{\ell=1}^{\infty} \frac{c_0}{c_0 + \lambda \ell^{1/\beta}}$$
$$\leqslant \int_0^{\infty} \frac{c_0}{c_0 + \lambda t^{1/\beta}} \mathrm{d}t = \mathcal{O}(\lambda^{-\beta}).$$

If the condition (8) for the effective dimension is imposed, then we can derive the following learning rates of the distributed spectral algorithms (3) to be proved in section 6.

**Corollary 1.** *Assume the regularity condition* (6) *with* $1/2 \leqslant r \leqslant \nu_g$. *If* (8) *holds with* $0 < \beta \leqslant 1$, $|D_1| = |D_2| = \cdots = |D_m|$, $\lambda = N^{-\frac{1}{2r+\beta}}$, *and*

$$m \leqslant N^{\min\left\{\frac{2}{2r+\beta}, \frac{2r-1}{2r+\beta}\right\}}, \tag{9}$$

*then*

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] = \mathcal{O}\left(N^{-\frac{2r}{2r+\beta}}\right). \tag{10}$$

Our error analysis for the distributed spectral algorithms (3) is carried out by making some improvements of existing methods and results [2, 7, 12] on spectral algorithms (1). As a by-product, we shall prove in section 5 the following error estimates for $\|f_{D,\lambda} - f_\rho\|_\rho$ concerning the classical spectral algorithms (1).

**Theorem 2.** *If the regularity condition* (6) *holds with* $1/2 \leqslant r \leqslant \nu_g$, *then for any* $\delta \in (0, 1)$, *with confidence at least* $1 - \delta$, $\|f_{D,\lambda} - f_\rho\|_\rho$ *is bounded by*

$$\begin{cases} C\left(\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right)^{\max\{1,r\}} \left(\mathcal{B}_{|D|,\lambda} + \lambda^r\right) (\log\frac{4}{\delta})^4, & \text{if } \frac{1}{2} \leqslant r \leqslant \min\left\{\frac{3}{2}, \nu_g\right\}, \\ C\left(\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right)\left(\mathcal{B}_{|D|,\lambda} + \left(\frac{\lambda}{|D|}\right)^{\frac{1}{2}} + \lambda^r\right) (\log\frac{6}{\delta})^3, & \text{if } \frac{3}{2} < r \leqslant \nu_g, \end{cases}$$

*where C is a constant independent of* $|D|$ *or* $\delta$. *If* (8) *holds with* $0 < \beta \leqslant 1$ *and* $\lambda = N^{-\frac{1}{2r+\beta}}$, *then for any* $0 < \delta < 1$, *with confidence at least* $1 - \delta$, *we have*

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant \tilde{C} N^{-\frac{r}{2r+\beta}} \left(\log 6/\delta\right)^4, \tag{11}$$

*where* $\tilde{C}$ *is a constat independent of N or* $\delta$. *Moreover,*

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho^2\right] = \mathcal{O}\left(N^{-\frac{2r}{2r+\beta}}\right). \tag{12}$$

To facilitate the use of the distributed spectral algorithms in practice, we discuss the important issue of parameter selection by the strategy of cross-validation in two situations.

The first situation is when the data are stored naturally across multiple servers in a distributive way. This happens in some practical applications in medicine, finance, business, and some other areas, and the data are not combined for reasons of protecting privacy or avoiding high costs. In such a circumstance, the number of data subsets $m$ is fixed. We propose a cross-validation approach for selecting the regularization parameter $\lambda$ for the distributed spectral algorithms, as done for the regularized least squares in [5]. Without loss of generality, we assume that each data subset $D_j$ has an even sample size and is the disjoint union of two subsets, $D_j^1$ (the training set) and $D_j^2$ (the validation set), of equal cardinality $|D_j^1| = |D_j^2| = |D_j|/2$. Let $\Lambda$

be a set of positive parameters. We use the training data subsets $\{D_j^1\}_{j=1}^m$ and the distributed spectral algorithm to get a set of estimators $\{\overline{f}_{D^1,\lambda}\}_{\lambda\in\Lambda}$ with $D^1 = \cup_{j=1}^m D_j^1$, and then utilize the validation subsets $\{D_j^2\}_{j=1}^m$ to select an optimal parameter $\lambda^*$ by

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \frac{1}{m}\sum_{j=1}^{m}\frac{1}{|D_j^2|}\sum_{z=(x,y)\in D_j^2}\left(\overline{f}_{D^1,\lambda}(x)-y\right)^2. \tag{13}$$

The final estimator is $\overline{f}_{D^1,\lambda^*}$. Some theoretical analysis for the cross-validation strategy can be found in [5, theorem 3].

The other situation we consider for parameter selection in distributed learning is when the divide-and-conquer approach is used to reduce the computational complexity and memory requirements. Here there exist two parameters in the distributed spectral algorithm: the number of data subsets $m$ and the regularization parameter $\lambda$. We assume that the data set $D$ has size $|D| = 2^n$ for some $n\in\mathbf{N}$. We choose $1\leqslant k_{\min}\leqslant k_{\max}\leqslant n$, take $m = 2^k$ with $k\in\mathcal{K}_{\text{index}} := \{k_{\min}, k_{\min}+1, \ldots, k_{\max}\}$, and divide the data set $D$ into $m = 2^k$ disjoint data subsets $\{D_{j,k}\}_{j=1}^{2^k}$ of size $2^{n-k}$. Here the integer $k_{\min}$ is chosen according to the processing ability (for data sets of size at most $2^{n-k_{\min}}$) of the local processors and $k_{\max}$ is set according to the size requirement (at least $2^{n-k_{\max}}$) of data subsets in the local processors. By dividing each data subset $D_{j,k}$ into the disjoint union of two subsets $D_{j,k}^1$ and $D_{j,k}^2$ of equal cardinality, and applying the distributed spectral algorithm to the training data subsets $\{D_{j,k}^1\}_{j=1}^m$ and $\lambda\in\Lambda$, we get a set of estimators $\{\overline{f}_{D^1_{\cdot,k},\lambda}\}_{\lambda\in\Lambda}$ with $D^1_{\cdot,k} = \cup_{j=1}^{2^k} D_{j,k}^1$. Then we can use the validation subsets $\{D_{j,k}^2\}_{j=1}^{2^k}$ to select the optimal parameter pair $(k^*, \lambda^*)$ by

$$(k^*,\lambda^*) = \arg\min\nolimits_{(k,\lambda)\in\mathcal{K}_{\text{index}}\times\Lambda}\frac{1}{2^{n-1}}\sum_{j=1}^{2^k}\sum_{z=(x,y)\in D_{j,k}^2}\left(\overline{f}_{D^1_{\cdot,k},\lambda}(x)-y\right)^2. \tag{14}$$

This yields the final estimator $\overline{f}_{D^1_{\cdot,k^*},\lambda^*}$. The computational cost of the above procedure is in the order of $O\left(\sum_{\lambda\in\Lambda}\sum_{k=k_{\min}}^{k_{\max}}\left(2^{n-k}\right)^3\right) = O\left(|D|^3 2^{-3k_{\min}}|\Lambda|\right)$. However, for each choice of $k$, a procedure of re-partition and re-communication of the data set would be required, which would significantly hinders the use of the proposed distributed spectral algorithms. To the best of our knowledge, a feasible way to select $m$ is to adjust it manually. It is possible to select $m$ manually via a few trails, since (9) yields optimal rates for distributed spectral algorithms for $m\leqslant m_0$ with the largest number of data subsets $m_0$ depending on $r$ and $\beta$ (which is difficult to determine in practice). It would be interesting to develop other efficient data-driven parameter selection strategies for distributed spectral algorithms (even for the well known distributed regularized least squares [22]) which can be rigorously justified in theory.

## 3. Related work and discussion

The study of spectral algorithms (1) has a long history in the literature of inverse problems [8]. The generalization performance of these algorithms have been investigated in the learning theory literature [2, 5, 12]. Let us mention some explicit results related to this paper and demonstrate the slight improvement of our learning rates (12) for the spectral algorithms (1) by removing a logarithmic factor from the existing results. Note that the learning rates (12) coincide with the minimax lower bound proved in [4, theorem 3], and are optimal.

When the regularity condition (6) holds with $1/2 \leqslant r \leqslant \nu_g$, it was proved in [2, corollary 17] that for

$$N > \left(2\sqrt{2}\kappa^2 \log\frac{4}{\delta}\right)^{\frac{4r+2}{2r+3}}, \tag{15}$$

with confidence at least $1-\delta$,

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant CN^{-\frac{r}{2r+1}}\log\frac{4}{\delta}, \tag{16}$$

where $C$ is a constant independent of $\delta$ or $N$. Let $\mathcal{P}_{N,\delta}$ be the event that (15) holds. Then it follows from the definition of expectation that

$$E[\|f_{D,\lambda} - f_\rho\|_\rho^2] \leqslant E[\|f_{D,\lambda} - f_\rho\|_\rho^2|\mathcal{P}_{N,\delta}] + [\|f_{D,\lambda} - f_\rho\|_\rho^2|\mathcal{P}_{N,\delta}^T], \tag{17}$$

where $\mathcal{P}_{N,\delta}^T$ denotes the event that (15) does not hold. Plugging (16) with sufficiently small $\delta$ ($\delta = N^{-2}$ for example) into (17), we obtain

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

This learning rate is almost optimal, but the extra term $\log N$ makes it suboptimal. This extra logarithmic term arises since the confidence-based error bound $\mathcal{O}\left(N^{-\frac{r}{2r+1}}\log\frac{4}{\delta}\right)$ holds only when $N$ is large enough satisfying the restriction (15). Recall that the complexity condition (8) is always satisfied with the universal parameter $\beta = 1$. Our derived learning rates (12) with this universal parameter choice $\beta = 1$ in (8) remove the logarithmic term and make the learning rates optimal.

The learning rates in [2] are derived with the worse but universal parameter $\beta = 1$ in the complexity condition (8), no matter how smooth the kernel $K$ is. This phenomenon was observed in [5] where learning rates of spectral algorithms (1) with a possibly smaller parameter $\beta \in (0, 1]$ in the complexity assumption (8) were provided. Under the conditions (6) and (8), it was proved in [5, theorem 2] that for

$$\lambda = \left(\frac{4C_2\log\frac{6}{\delta}}{N}\right)^{\frac{1}{2r+\beta}}$$

and $2r + \beta \geqslant 2$, with confidence at least $1-\delta$, there holds

$$\|f_{D,\lambda} - f_\rho\|_\rho^2 \leqslant C_3N^{-\frac{2r}{2r+\beta}}\left(\log\frac{6}{\delta}\right)^{\frac{r}{2r+\beta}},$$

where $C_2$ and $C_3$ are constants independent of $\delta$ or $N$. Again, this learning rate is suboptimal since the optimal regularization parameter depends on the confidence level $\delta$. Furthermore, when $2r + \beta < 2$, the result in [5] needs additional semi-supervised samples. Our derived learning rates (12) remove the restriction $2r + \beta \geqslant 2$ and provide the optimal minimax learning rates with $\lambda$ independent of $\delta$.

Distributed learning is a recent development for handling big data [13, 14, 21]. The existing rigorous error analysis of distributed learning algorithms in the framework of learning theory (or nonparametric regression) was originally given in [22] for the kernel-based regularized least squares, the special case of the spectral algorithms (1) with the filter function $g_\lambda(\sigma) = (\sigma + \lambda)^{-1}$. To demonstrate the essential difference between our analysis in this paper and that in [22], we state a key assumption made in [22] on the normalized eigenfunctions $\{\phi_i\}_i$ of the integral operator $L_K$ that for some constants $2 < k \leqslant \infty$ and $A < \infty$, there holds

$$\begin{cases} \sup_{\lambda_i>0} E\left[\left|\frac{\phi_i(x)}{\sqrt{\lambda_i}}\right|^{2k}\right] \leqslant A^{2k}, & \text{when } k < \infty, \\ \sup_{\lambda_i>0} \left\|\frac{\phi_i(x)}{\sqrt{\lambda_i}}\right\|_\infty \leqslant A, & \text{when } k = \infty. \end{cases} \tag{18}$$

Under this key assumption, for the distributed kernel-based regularized least squares with $\{D_j\}_{j=1}^m$ of equal size, the learning rate $E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_\rho^2\right] = \mathcal{O}\left(N^{-\frac{1}{\beta+1}}\right)$ with $0 < \beta < 1$ was derived in [22] when $f_\rho \in \mathcal{H}_K$, $\lambda_i = \mathcal{O}\left(i^{-1/\beta}\right)$, $\lambda = N^{-\frac{1}{\beta+1}}$ and

$$m = \begin{cases} \mathcal{O}\left(\left(\frac{N^{\frac{(k-4)-k\beta}{\beta+1}}}{A^{4k}\log^k N}\right)^{\frac{1}{k-2}}\right), & \text{when } k < \infty, \\ \mathcal{O}\left(\frac{N^{\frac{1-\beta}{\beta+1}}}{A^4 \log N}\right), & \text{when } k = \infty. \end{cases} \tag{19}$$

It shows that, under the regularity condition (6) with $r = \frac{1}{2}$, distributed kernel-based regularized least squares can achieve the optimal learning rate provided that $m$ satisfies (19) and (18) is valid. This result was proved by a matrix analysis approach and the eigenfunction assumption (18) played an important role. There are three differences between our results in this paper and those in [22]. Besides the broader range of our results for the whole family of distributed spectral algorithms and the more general regularity condition (6) with $r \geqslant \frac{1}{2}$, our analysis does not require the essential eigenfunction assumption (18).

**Remark.** To illustrate our novelty further, we remark that the eigenfunction assumption (18) for the integral operator $L_K = L_{K,\rho_X}$ associated with the pair $(K, \rho_X)$ could be difficult to verify. Besides the case of finite dimensional RKHSs, in the existing literature, there are only these classes of pairs $(K, \rho_X)$ with this eigenfunction assumption rigorously verified: the Sobolev reproducing kernels on Euclidean domains with normalized Lebesgue measures [18], periodical kernels [19], and kernels artificially constructed by a Mercer type expansion

$$K(x, y) = \sum_i \lambda_i \frac{\phi_i(x)}{\sqrt{\lambda_i}} \frac{\phi_i(y)}{\sqrt{\lambda_i}}, \tag{20}$$

where $\{\frac{\phi_i(x)}{\sqrt{\lambda_i}}\}_i$ is a given orthonormal system in $L^2_{\rho_X}$. To demonstrate the complexity of the eigenfunction assumption (18) as the marginal distribution $\rho_X$ changes, take a different probability measure $\mu$ induced by a nonnegative function $P \in L^2_{\rho_X}$ as $d\mu = P(x)\mathrm{d}\rho_X$. By expanding $\phi = \sum_{i\in I} \frac{c_i}{\sqrt{\lambda_i}}\phi_i$ in terms of the orthonormal basis $\{\phi_i\}_{i\in I}$ of $\mathcal{H}_K$ where $I := \{i : \lambda_i > 0\}$, we can see [11] that for $\lambda_0 > 0$,

$$L_{K,\mu}\phi = \lambda_0\phi \quad \Longleftrightarrow \quad \mathbb{K}^P c = \lambda_0 c,$$

where $c = (c_i)_{i\in I}$ and $\mathbb{K}^P$ is a possibly infinite matrix given by

$$\mathbb{K}^P = \left(\lambda_i \int_{\mathcal{X}} \frac{\phi_i(x)}{\sqrt{\lambda_i}} P(x) \frac{\phi_j(x)}{\sqrt{\lambda_j}} \mathrm{d}\rho_X\right)_{i,j\in I}. \tag{21}$$

Hence the eigenpairs of the integral operator $L_{K,\mu}$ associated with $(K, \mu)$ can be characterized by those of the matrix $\mathbb{K}^P$ defined by (21). Checking an eigenvector condition for the possibly infinite matrix $\mathbb{K}^P$ corresponding to the eigenfunction assumption (18) is a challenging question involving multiplier operators in harmonic analysis. The matrix $\mathbb{K}^P$ is not diagonal in general, except the case when $\mu = \rho_X$ and $P \equiv 1$. Moreover, an example of a $C^\infty$ Mercer kernel was presented in [23] to show that only the smoothness of the Mercer kernel does not guarantee the uniform boundedness of the eigenfunctions $\{\frac{\phi_i(x)}{\sqrt{\lambda_i}}\}_i$. Until now, it is even unknown whether any of the Gaussian kernels on $\mathcal{X} = [0, 1]^d$ satisfies the eigenfunction assumption (18) when $\rho_X$ is a general Borel measure.

Note that the condition (9) on the number $m$ of local processors is more restrictive than (19) imposed in [22] for the special case of regularized least squares and $r = 1/2$. But a main advantage of our results is to remove the eigenfunction assumption (18) which was key in [22] to get less restrictions on $m$. It is a great challenge to derive the optimal learning rates under the conditions imposed in this paper, without the eigenfunction assumption (18), but with less demanding restriction (19) on $m$. Our analysis is achieved by the novelty of using the integral operator approach and second order decomposition of a difference of operator inverses. Similar analysis was carried out in our recent work [11] for the special algorithm of distributed regularized least squares. In this paper we are concerned with distributed learning with the general spectral algorithms (1) rather than the regularized least squares. It has not been considered in the literature, to our best knowledge. One of the advantages of distributing learning with general spectral algorithms such as the spectral cut-off and Landweber iteration is that the saturation of the distributed regularized least squares, stated by the restriction $r \leqslant 1$, is overcome and faster learning rates with $r > 1$ can be achieved. Corollary 1 shows that, if the number $m$ of local machines is not too large, then the learning rates of the distributed spectral algorithm (3) and the spectral algorithm (1) implemented on a 'big' machine with the whole data set are identical.

Comparing corollary 1 with theorem 2, we find the regularization parameters $\lambda$ for achieving the optimal learning rates are identical. Even though each local machine accesses to only $1/m$ of the whole data, it is nonetheless essential to regularize as we process the whole data set, which would lead to overfitting in local machines. However, the $m$-fold weighted average reduces the variance and enables algorithm (3) to achieve optimal learning rates. Though we are interested in using the general spectral algorithms, other than the regularized least squares, mainly in the case $r > 1$ (at least for overcoming the saturation) under the restriction (9), with the power exponent involving $\frac{2r-1}{2r+\beta}$, makes more sense, it would be interesting to relax the restriction (9) by other approaches.

After the submission of this paper in January 2016, we found that similar analysis was independently carried out for spectral algorithms by Dicker, Foster, and Hsu in a paper entitled 'Kernel ridge versus principal component regression: minimax bounds and adaptibility of regularization operators' (arXiv:1605.08839, May 2016) and for spectral algorithms and distributed spectral algorithms by Blanchard and Mücke in a paper entitled 'Parallel spectral algorithms for kernel learning' (arXiv:1610.07487, October 2016).

## 4. Second order decomposition for bounding norms

Our error analysis is based on a combination of an integral operator approach [4, 16, 17], a recently developed technique [11] of second order decomposition of inverse operator differences, and a novel idea for bounding operator products developed below in this paper. The core of the integral operator method is to analyze similarities between the empirical integral operator $L_{K,D}$ and its data-free limit $L_K$, since convergence rates of corresponding schemes associated with $L_K$ are easier to derive. Taking spectral algorithms for example, the data-free limit of algorithms (1) is

$$f_\lambda = g_\lambda(L_K)L_K f_\rho.$$

According to the definition of the filter function and conditions (6), if $r \leqslant \nu_g$, we see from the identity $\|f\|_\rho = \|L_K^{1/2} f\|_K$ for $f \in \mathcal{H}_K$ that

$$\|f_\lambda - f_\rho\|_\rho \leqslant \|L_K^{1/2}(g_\lambda(L_K)L_K - I)L_K^{r-1/2}\|\|L_K^{1/2}u_\rho\|_K \leqslant \gamma_r\lambda^r\|u_\rho\|_\rho. \tag{22}$$

The classical method to analyze similarities between $L_K$ and $L_{K,D}$ is to bound the norm of operator difference $L_K - L_{K,D}$. We refer the readers to [2, 4, 5, 7, 9–11, 17, 20] for details on this method. In particular, it can be found in [4, 20] and [11] that for any $\delta \in (0,1)$, with confidence at least $1-\delta$, there holds

$$\|L_K - L_{K,D}\| \le \|L_K - L_{K,D}\|_{HS} \leqslant \frac{4\kappa^2}{\sqrt{|D|}}\log\frac{2}{\delta}, \tag{23}$$

where $||A||_{HS}$ denotes the Hilbert–Schmidt norm of a Hilbert–Schmidt operator $A$, and

$$\|(\lambda I + L_K)^{-1/2}(L_K - L_{K,D})\| \leqslant \mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}, \tag{24}$$

where $\mathcal{B}_{|D|,\lambda}$ is defined by (7).

In our error decomposition for spectral algorithms, given in proposition 2 below, we use the norms of the product operators $(L_K+\lambda I)(L_{K,D}+\lambda)^{-1}$ and $(L_{K,D}+\lambda I)(L_K+\lambda I)^{-1}$ which also reflect similarities between $L_K$ and $L_{K,D}$, as seen from equation (26) below. In this section, we present tight bounds of the product norms by using the second order decomposition of operator differences. This is the first novelty of our error analysis carried out in this paper.

Let $A$ and $B$ be invertible operators on a Banach space. The first order decomposition of operator differences is

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1} = B^{-1}(B-A)A^{-1}. \tag{25}$$

It follows that

$$A^{-1}B = (A^{-1} - B^{-1})B + I = A^{-1}(B-A) + I. \tag{26}$$

Let $A = L_K + \lambda I$ and $B = L_{K,D} + \lambda I$. We see from (26) that

$$\|(L_K+\lambda I)^{-1}(L_{K,D}+\lambda I)\| \leqslant \|(L_K+\lambda I)^{-1/2}\|\|(L_K+\lambda I)^{-1/2}(L_{K,D}-L_K)\| + 1.$$

Since $\|(\lambda I + L_K)^{-1/2}\| \leqslant 1/\sqrt{\lambda}$, it follows from (24) that with confidence at least $1-\delta$, there holds

$$\|(L_K+\lambda I)^{-1}(L_{K,D}+\lambda I)\| \leqslant \frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}}\log\frac{2}{\delta} + 1. \tag{27}$$

To bound the norm of $(L_K+\lambda I)(L_{K,D}+\lambda I)^{-1}$, the first order decomposition (25) is not sufficient due to the lack of an appropriate bound for $\|(L_{K,D}+\lambda I)^{-1/2}(L_{K,D}-L_K)\|$. We need to decompose $A^{-1} = A^{-1} - B^{-1} + B^{-1}$ in (25) further. This is the second order decomposition of operator difference, which was introduced in [11] to derive optimal learning rates of the regularized least squares. It asserts that if $A$ and $B$ are invertible operators on a Banach space, then (25) yields

$$\begin{aligned} A^{-1} - B^{-1} &= B^{-1}(B-A)A^{-1}(B-A)B^{-1} + B^{-1}(B-A)B^{-1} \\ &= B^{-1}(B-A)B^{-1}(B-A)A^{-1} + B^{-1}(B-A)B^{-1}. \end{aligned} \tag{28}$$

This implies the following decomposition of the operator product

$$BA^{-1} = (B-A)B^{-1}(B-A)A^{-1} + (B-A)B^{-1} + I. \tag{29}$$

Inserting $A = L_{K,D} + \lambda I$ and $B = L_K + \lambda I$ to (29), we obtain the following upper bound for $\|(L_K + \lambda I)(L_{K,D} + \lambda I)^{-1}\|$.

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

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

**Proof.** We apply (29) to the operator $A = L_{K,D} + \lambda I$ and $B = L_K + \lambda I$. Applying the bounds $\|(\lambda I + L_{K,D})^{-1}\| \leqslant 1/\lambda$ and $\|(\lambda I + L_K)^{-1/2}\| \leqslant 1/\sqrt{\lambda}$ gives

$$\begin{aligned} \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\| \leqslant{}& \|(\lambda I + L_K)^{-1/2}(L_K - L_{K,D})\|^2 \lambda^{-1} \\ &+ \|(\lambda I + L_K)^{-1/2}(L_K - L_{K,D})\|\lambda^{-1/2} + 1. \end{aligned}$$

Here we have used the fact that

$$\|L_1 L_2\| = \|(L_1 L_2)^T\| = \|L_2^T L_1^T\| = \|L_2 L_1\|$$

for any self-adjoint operators $L_1, L_2$ on Hilbert spaces. Applying (24) yields

$$\begin{aligned} \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\| &\leqslant \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + \left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}}\right) + 1 \\ &\leqslant 2\left[\left(\frac{\mathcal{B}_{|D|,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right], \end{aligned}$$

which proves our result. □

An upper bound of $\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\|$ was presented in [3], as lemma A.5, asserting that if

$$\|(L_K + \lambda I)^{-1/2}(L_{K,D} - L_K)(L_K + \lambda I)^{-1/2}\| < 1 - \eta \tag{30}$$

for some $0 < \eta < 1$, then

$$\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\| \leqslant \frac{1}{\sqrt{\eta}}.$$

The condition (30) with $0 < \eta < 1$ requires the sample size $N$ to be large enough. As discussed in section 3, this restriction imposes a logarithmic term in error estimates, which makes the obtained learning rates suboptimal. We encourage the readers to compare our proposition 1 with lemma A.5 of [3]. The operator product norm estimates in proposition 1 plays a crucial role in our integral operator approach. In particular, based on the estimate of $\|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\|$, we present a new error decomposition technique for spectral algorithms in the next section.

## 5. Deriving error bounds for spectral algorithms

To analyze the spectral algorithms (1), we introduce a new error decomposition technique, the second novelty of our error analysis. Since we do not impose any additional Lipschitz conditions to the filter function $g_\lambda$, it is not easy to bound the difference between $f_{D,\lambda}$ and $f_\lambda$. We turn to bounding the difference between $f_{D,\lambda}$ and its semi-population version $E^*[f_{D,\lambda}]$, where

$$E^*[f_{D,\lambda}] = E[f_{D,\lambda}|x_1,\ldots,x_N]$$

denotes the conditional expectation of $f_{D,\lambda}$ given $x_1,\ldots,x_N$. Since $f_\rho = E^*[y]$, we then have from (1) that

$$E^*[f_{D,\lambda}] = g_\lambda(L_{K,D})L_{K,D}f_\rho \tag{31}$$

and the triangle inequality

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant \|f_{D,\lambda} - E^*[f_{D,\lambda}]\|_\rho + \|E^*[f_{D,\lambda}] - f_\rho\|_\rho.$$

For the first part, since $\|L_K^{1/2}(\lambda I + L_K)^{-1/2}\| \leqslant 1$, we have

$$\begin{aligned}\|f_{D,\lambda} - E^*(f_{D,\lambda})\|_\rho &= \left\|L_K^{1/2}(f_{D,\lambda} - E^*(f_{D,\lambda}))\right\|_K\\ &\leqslant \left\|(\lambda I + L_K)^{1/2}(f_{D,\lambda} - E^*(f_{D,\lambda}))\right\|_K.\end{aligned}$$

Denote

$$\Delta_D = \frac{1}{|D|}\sum_{(x,y)\in D}(y - f_\rho(x))K_x. \tag{32}$$

It follows from the definition of $f_{D,\lambda}$ and the property (4) of the filter function that

$$\begin{aligned}&\|f_{D,\lambda} - E^*(f_{D,\lambda})\|_\rho\\ &\leqslant \left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}(\lambda I + L_{K,D})^{1/2}g_\lambda(L_{K,D})\Delta_D\right\|_K\\ &\leqslant \|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\|\|g_\lambda(L_{K,D})(\lambda I + L_{K,D})\|\|(\lambda I + L_{K,D})^{-1/2}\\ &(\lambda I + L_K)^{1/2}\|\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K\\ &\leqslant 2b\left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\|^2\left\|(\lambda I + L_K)^{-1/2}\Delta_D\right\|_K.\end{aligned}$$

For the second part, we have

$$\begin{aligned}&\|E^*(f_{D,\lambda}) - f_\rho\|_\rho = \left\|L_K^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K\\ &\leqslant \left\|(\lambda I + L_K)^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K\\ &\leqslant \left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\|\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K.\end{aligned}$$

All the above discussion yields the following error decomposition for the spectral algorithms.

**Proposition 2.** *Let $\Delta_D$ be defined by (32). Then we have*

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant I_1 + I_2,$$

*where*

$$I_1 = 2b\left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\|^2 \left\|(\lambda I + L_K)^{-1/2}\Delta_D\right\|_K,$$
$$I_2 = \left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\| \left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K.$$

The error decomposition presented in proposition 2 is different from that in [2] or [5]. In fact, it can be found in equation (26) of [2] that the difference $L_K - L_{K,D}$ rather than the operator product $(L_K + \lambda I)(L_{K,D} + \lambda I)^{-1}$ is used to analyze similarities between $L_K$ and $L_{K,D}$. The error analysis in [5] needs an additional truncated function

$$f_\lambda^{tr} = P_\lambda f_\rho,$$

where $P_\lambda$ is the orthogonal projector in $L^2_{\rho_X}$ defined by

$$P_\lambda = \Theta_\lambda(L_K),$$

with $\Theta_\lambda(\sigma) = 1$ for $\sigma \geqslant \lambda$ and $\Theta_\lambda(\sigma) = 0$ for $\sigma < \lambda$. Our error decomposition does not require such a truncated function and uses the operator product to analyze similarities between $L_K$ and $L_{K,D}$.

Based on the new error decomposition technique presented in proposition 2, we should bound three terms: $\left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\|$, $\left\|(\lambda I + L_K)^{-1/2}\Delta_D\right\|_K$, and $\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$. To bound the first term, we need proposition 1 and the following lemma about norm estimates for products of powers of two positive operators. It was proved in [1, theorem IX.2.1–2] for positive definite matrices and then provided in [3, lemma A.7] for positive operators on Hilbert spaces.

**Lemma 1.** *Let $s \in [0, 1]$. For positive operators A and B on a Hilbert space we have*

$$\|A^s B^s\| \leqslant \|AB\|^s. \tag{33}$$

*Applying the above lemma to $A = \lambda I + L_K$ and $B = (\lambda I + L_{K,D})^{-1}$, we get from proposition 1 the following result.*

**Lemma 2.** *Let $\lambda > 0, 0 \leqslant s \leqslant 1$ and $0 < \delta < 1$. Then, with confidence at least $1 - \delta$, there holds*

$$\|(\lambda I + L_K)^s(\lambda I + L_{K,D})^{-s}\| \leqslant 2^s \left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]^s.$$

*The second term $\left\|(\lambda I + L_K)^{-1/2}\Delta_D\right\|_K$ can be bounded by the following lemma found in* [3] *or* [4].

**Lemma 3.** *Let $\delta \in (0, 1)$, and D be a sample drawn independently according to $\rho$. If $|y| \leqslant M$ almost surely, then with confidence at least $1 - \delta$, there holds*

$$\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K \leqslant \frac{2M}{\kappa}\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}.$$

*To bound the third term $\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$, we need the following two lemmas. The first one can be found in* [3, *lemma A.6*], *and the second one is derived from properties* (4) *and* (5) *of the filter function $g_\lambda$.*

**Lemma 4.** *For positive Hilbert–Schmidt operators $A$ and $B$ on a Hilbert space with $\|A\|, \|B\| \leqslant \mathcal{C}$ for some constant $\mathcal{C} > 0$, we have for $t \geqslant 1$,*

$$||A^t - B^t||_{HS} \leq t\mathcal{C}^{t-1}||A - B||_{HS}. \tag{34}$$

**Lemma 5.** *For $0 < t \leqslant \nu_g$, we have*

$$\|(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^t\| \leqslant 2^t(b + 1 + \gamma_t)\lambda^t. \tag{35}$$

**Proof.** Recall from the introduction that $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})_i\}$ is a set of normalized eigenpairs of $L_{K,D}$ with the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ forming an orthonormal basis of $\mathcal{H}_K$. Then for each $h \in \mathcal{H}_K$, we have $h = \sum \langle h, \phi_i^{\mathbf{x}}\rangle \phi_i^{\mathbf{x}}$ and $\|h\|_K^2 = \sum_i \langle h, \phi_i^{\mathbf{x}}\rangle^2$. Hence

$$\begin{aligned}
&\|(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^t h\|_K \\
&= \left\| \sum_{i=1}^{\infty} \langle h, \phi_i^{\mathbf{x}}\rangle (g_\lambda(\sigma_i^{\mathbf{x}})\sigma_i^{\mathbf{x}} - 1)(\lambda + \sigma_i^{\mathbf{x}})^t \phi_i^{\mathbf{x}} \right\|_K \\
&= \left\{ \sum_{i=1}^{\infty} \left[\langle h, \phi_i^{\mathbf{x}}\rangle (g_\lambda(\sigma_i^{\mathbf{x}})\sigma_i^{\mathbf{x}} - 1)(\lambda + \sigma_i^{\mathbf{x}})^t\right]^2 \right\}^{1/2} \\
&\leqslant \left\{ \sum_{i=1}^{\infty} \left[|\langle h, \phi_i^{\mathbf{x}}\rangle||g_\lambda(\sigma_i^{\mathbf{x}})\sigma_i^{\mathbf{x}} - 1|2^t(\lambda^t + (\sigma_i^{\mathbf{x}})^t)\right]^2 \right\}^{1/2} \\
&\leqslant 2^t(b + 1 + \gamma_t)\lambda^t \left\{ \sum_{i=1}^{\infty} (\langle h, \phi_i^{\mathbf{x}}\rangle)^2 \right\}^{1/2} = 2^t(b + 1 + \gamma_t)\lambda^t \|h\|_K.
\end{aligned}$$

The last inequalities hold due to properties (4) and (5) of the filter function $g_\lambda$ and the elementary inequality $(c + d)^t \leqslant 2^t(c^t + d^t)$ for any $c, d, t \geqslant 0$. Then our result follows from the definition of the operator norm on $\mathcal{H}_K$. □

Before presenting the proof of theorem 2, we give the following proposition.

**Proposition 3.** *Assume (6) holds with $1/2 \leqslant r \leqslant \nu_g$. If $\frac{1}{2} \leqslant r \leqslant 3/2$, we have*

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant 2b\Xi_D\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K + 2^r(b + 1 + \gamma_r)\|u_\rho\|_\rho \lambda^r \Xi_D^r,$$

*where*

$$\Xi_D = \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\|. \tag{36}$$

*If $\frac{3}{2} < r \leqslant \nu_g$, we have*

$$\|f_{D,\lambda} - f_\rho\|_\rho \leqslant 2b\Xi_D\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K + C_{b,r,\kappa}\|u_\rho\|_\rho \Xi_D^{1/2}(\sqrt{\lambda}\|L_K - L_{K,D}\|_{HS} + \lambda^r),$$

*where $C_{b,r,\kappa}$ is a constant depending only on $b$, $r$ and $\kappa$.*

**Proof.** We estimate the two parts $I_1$ and $I_2$ in proposition 2. For $I_1$, recall the notation $\Xi_D = \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\|$ and by lemma 1 with $s = 1/2$, we have

$$I_1 \leqslant 2b\Xi_D\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K.$$

Now we estimate

$$I_2 = \left\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\right\| \left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$$
$$\leqslant \Xi_D^{1/2}\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K. \tag{37}$$

The regularity condition (6) with $1/2 \leqslant r \leqslant \nu_g$ means $f_\rho = L_K^r u_\rho$ for some $u_\rho \in L^2_{\rho_X}$. To estimate the term $\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$, we consider two cases according to different regularity levels as follows.

Case 1: If $r \in [1/2, 3/2]$, we have

$$\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$$
$$= \left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)L_K^{r-1/2}L_K^{1/2}u_\rho\right\|_K$$
$$= \big\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^{r-1/2}(\lambda I + L_{K,D})^{-(r-1/2)}$$
$$(\lambda I + L_K)^{r-1/2}(\lambda I + L_K)^{-(r-1/2)}L_K^{r-1/2}L_K^{1/2}u_\rho\big\|_K$$
$$\leqslant \|(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^r\|\, \Xi_D^{r-1/2}\|u_\rho\|_\rho,$$

where the last inequality follows from lemma 1 with $s = r - 1/2 \in [0, 1]$, and the bound $\|(\lambda I + L_K)^{-(r-1/2)}L_K^{r-1/2}\| \leqslant 1$. Thus,

$$I_2 \leqslant \|(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^r\|\, \Xi_D^r\|u_\rho\|_\rho.$$

Combining the above estimates and lemma 5 with $t = r$ yields

$$I_2 \leqslant 2^r(b + 1 + \gamma_r)\|u_\rho\|_\rho\lambda^r\Xi_D^r. \tag{38}$$

Case 2: If $r > 3/2$, we decompose $\left\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\right\|_K$ as

$$\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)f_\rho\|_K$$
$$= \|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(L_K)^{r-1/2}L_K^{1/2}u_\rho\|_K$$
$$\leqslant \|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(L_K)^{r-1/2}\|\|u_\rho\|_\rho. \tag{39}$$

By adding and subtracting the operator $(L_{K,D})^{r-1/2}$, we can continue the estimation as

$$\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(L_K)^{r-1/2}\|$$
$$\leqslant \|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)((L_K)^{r-1/2} - (L_{K,D})^{r-1/2})\|$$
$$+ \|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(L_{K,D})^{r-1/2}\|.$$

Since $r > 3/2$, applying lemma 4 to $A = L_K$ and $B = L_{K,D}$ with $t = r - 1/2 > 1$, we have

$$\|(L_K)^{r-1/2} - (L_{K,D})^{r-1/2}\| \leqslant (r - 1/2)\kappa^{2r-3}||L_K - L_{K,D}||_{HS}.$$

Combining this observation with the bound $\|(\lambda I + L_{K,D})^{r-1/2}(L_{K,D})^{r-1/2}\| \leqslant 1$, and applying lemma 5 with $t = 1/2$ or $t = r$, we know that

$$\begin{aligned}
&\|(\lambda I + L_{K,D})^{1/2}(g_\lambda(L_{K,D})L_{K,D} - I)(L_K)^{r-1/2}\| \\
&\leqslant \|(\lambda I + L_{K,D})^{1/2}\left(g_\lambda(L_{K,D})L_{K,D} - I\right)\|(r-1/2)\kappa^{2r-3}\|L_K - L_{K,D}\|_{HS} \\
&\quad + \|(g_\lambda(L_{K,D})L_{K,D} - I)(\lambda I + L_{K,D})^r\| \\
&\leqslant (b+1+\gamma_{\frac{1}{2}})(r-1/2)\kappa^{2r-3}\sqrt{\lambda}\|L_K - L_{K,D}\|_{HS} + 2^r(b+1+\gamma_r)\lambda^r \\
&\leqslant C_{b,r,\kappa}(\sqrt{\lambda}\|L_K - L_{K,D}\|_{HS} + \lambda^r),
\end{aligned}$$

where $C_{b,r,\kappa} = (b+1+\gamma_{\frac{1}{2}})(r-1/2)\kappa^{2r-3} + 2^r(b+1+\gamma_r)$. Putting the above estimates into inequality (39) yields our desired error bound. □

Now we are in a position to prove theorem 2.

**Proof of theorem 2.** By proposition 1, for $\delta \in (0,1)$, there exists a subset $\mathcal{Z}^{|D|}_{\delta,1}$ of $\mathcal{Z}^{|D|}$ of measure at least $1-\delta$ such that

$$\Xi_D = \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\| \leqslant 2\left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right], \quad \forall D \in \mathcal{Z}^{|D|}_{\delta,1}.$$

By lemma 3 there exists another subset $\mathcal{Z}^{|D|}_{\delta,2}$ of $\mathcal{Z}^{|D|}$ of measure at least $1-\delta$ such that

$$\|(\lambda I + L_K)^{-1/2}\Delta_D\|_K \leqslant \frac{2M}{\kappa}\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}, \qquad \forall D \in \mathcal{Z}^{|D|}_{\delta,2}.$$

According to (23), there exists a third subset $\mathcal{Z}^{|D|}_{\delta,3}$ of $\mathcal{Z}^{|D|}$ of measure at least $1-\delta$ such that

$$\|L_K - L_{K,D}\|_{HS} \leqslant \frac{4\kappa^2\log\frac{2}{\delta}}{\sqrt{|D|}}, \qquad \forall D \in \mathcal{Z}^{|D|}_{\delta,3}.$$

Put the above observations into proposition 3. When $r \in [\frac{1}{2}, \frac{3}{2}]$, we know that for $D \in \mathcal{Z}^{|D|}_{\delta,1} \bigcap \mathcal{Z}^{|D|}_{\delta,2}$, there holds

$$\begin{aligned}
\|f_{D,\lambda} - f_\rho\|_\rho &\leqslant \tfrac{8Mb}{\kappa}\left[\left(\tfrac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]\mathcal{B}_{|D|,\lambda}\log\tfrac{2}{\delta} \\
&\quad + 4^r(b+1+\gamma_r)\|u_\rho\|_\rho\lambda^r\left[\left(\tfrac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]^r \\
&\leqslant C\left[\left(\tfrac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]^{\max\{1,r\}}\left(\mathcal{B}_{|D|,\lambda}\log\tfrac{2}{\delta} + \lambda^r\right),
\end{aligned}$$

where $C = \frac{8Mb}{\kappa} + 4^r(b+1+\gamma_r)\|u_\rho\|_\rho$. Then our first desired bound is verified by scaling $2\delta$ to $\delta$.

When $r > \frac{3}{2}$, for $D \in \mathcal{Z}^{|D|}_{\delta,1} \bigcap \mathcal{Z}^{|D|}_{\delta,2} \bigcap \mathcal{Z}^{|D|}_{\delta,3}$, there holds

$$
\begin{aligned}
\|f_{D,\lambda}-f_\rho\|_\rho \leqslant{}& \frac{8Mb}{\kappa}\left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta} \\
&+\sqrt{2}C_{b,r,\kappa}\|u_\rho\|_\rho\left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{\frac{1}{2}}\left[4\kappa^2\left(\frac{\lambda}{|D|}\right)^{\frac{1}{2}}\log\frac{2}{\delta}+\lambda^r\right] \\
\leqslant{}& C\left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]\left[\mathcal{B}_{|D|,\lambda}\log\frac{2}{\delta}+\left(\frac{\lambda}{|D|}\right)^{\frac{1}{2}}\log\frac{2}{\delta}+\lambda^r\right],
\end{aligned}
$$

where $C=\frac{8Mb}{\kappa}+\sqrt{2}C_{b,r,\kappa}\|u_\rho\|_\rho(4\kappa^2+1)$. Then we get our first desired bound for $\|f_{D,\lambda}-f_\rho\|_\rho$ by scaling $3\delta$ to $\delta$.

To prove the second desired bound, we notice from the condition $\mathcal{N}(\lambda)\leqslant C_0\lambda^{-\beta}$, and the choices $1/2\leqslant r\leqslant \nu_g$ and $\lambda=N^{-\frac{1}{2r+\beta}}$ that $\mathcal{B}_{|D|,\lambda}$ can be bounded as

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

and

$$
\left(\frac{\mathcal{B}_{|D|,\lambda}}{\sqrt{\lambda}}\right)^2+1\leqslant 4\left(\kappa^2+\kappa\sqrt{C_0}\right)^2+1.
$$

Putting the above observations into our first proved bound, we know that when $r\in[\frac{1}{2},\frac{3}{2}]$, with confidence at least $1-\delta$, $\|f_{D,\lambda}-f_\rho\|_\rho$ is bounded by

$$
\begin{aligned}
&\tilde{C}\left[4\left(\kappa^2+\kappa\sqrt{C_0}\right)^2+1\right]^{\max\{1,r\}}(2\kappa^2+2\kappa\sqrt{C_0}+1)N^{-\frac{r}{2r+\beta}}(\log\frac{4}{\delta})^4 \\
&=\tilde{C}N^{-\frac{r}{2r+\beta}}(\log\frac{4}{\delta})^4,
\end{aligned}
$$

where $\tilde{C}=C\left[4\left(\kappa^2+\kappa\sqrt{C_0}\right)^2+1\right]^{\max\{1,r\}}(2\kappa^2+2\kappa\sqrt{C_0}+1)$.

When $r>\frac{3}{2}$, with confidence at least $1-\delta$, $\|f_{D,\lambda}-f_\rho\|_\rho$ is bounded by

$$
\begin{aligned}
&C\left[4\left(\kappa^2+\kappa\sqrt{C_0}\right)^2+1\right]\left[(2\kappa^2+2\kappa\sqrt{C_0}+1)N^{-\frac{r}{2r+\beta}}+N^{-\frac{2r+\beta+1}{2(2r+\beta)}}\right](\log\frac{6}{\delta})^3 \\
&\leqslant \tilde{C}N^{-\frac{r}{2r+\beta}}(\log\frac{6}{\delta})^3,
\end{aligned}
$$

where $\tilde{C}=2C\left[4\left(\kappa^2+\kappa\sqrt{C_0}\right)^2+1\right]\left(\kappa^2+\kappa\sqrt{C_0}+2\right)$. This proves our second desired bound (11) for $\|f_{D,\lambda}-f_\rho\|_\rho$.

Finally, we apply the formula

$$
E[\xi]=\int_0^\infty \mathrm{Prob}[\xi>t]\mathrm{d}t \tag{40}
$$

for nonnegative random variables to $\xi=\|f_{D,\lambda}-f_\rho\|_\rho^2$ and use the bound

$$\text{Prob}\,[\xi > t] = \text{Prob}\left[\xi^{\frac{1}{2}} > t^{\frac{1}{2}}\right] \leqslant 6\exp\left\{-\tilde{C}^{-\frac{1}{4}} N^{\frac{r}{8r+4\beta}} t^{\frac{1}{8}}\right\}$$

derived from (11) for $t > \tilde{C}\log 6^4 N^{-\frac{r}{2r+\beta}}$. Then

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho^2\right] \leqslant \tilde{C}^2 \log 6^8 N^{-\frac{2r}{2r+\beta}} + 6\int_0^\infty \exp\left\{-\tilde{C}^{-\frac{1}{4}} N^{\frac{r}{8r+4\beta}} t^{\frac{1}{8}}\right\} \mathrm{d}t.$$

The second term equals $48\tilde{C}^2 N^{-\frac{2r}{2r+\beta}} \int_0^\infty u^{8-1}\exp\{-u\}\,\mathrm{d}u$. Due to $\int_0^\infty u^{d-1}\exp\{-u\}\,\mathrm{d}u = \Gamma(d)$ for $d > 0$, we have

$$E[\|f_{D,\lambda} - f_\rho\|_\rho^2] \leqslant (6\Gamma(9) + \log 6^8)\tilde{C}^2 N^{-\frac{2r}{2r+\beta}}.$$

This completes the proof of theorem 2. □

## 6. Proving main results for distributed algorithms

To present the proof, we need an error decomposition for the distributed spectral algorithms (3).

**Proposition 4.** *Let $\overline{f}_{D,\lambda}$ be defined by (3). We have*

$$E[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2] \leqslant \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right] + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\|E[f_{D_j,\lambda}] - f_\rho\right\|_\rho^2. \tag{41}$$

**Proof.** Due to (3) and $\sum_{j=1}^m \frac{|D_j|}{|D|} = 1$, we have

$$\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2 = \left\|\sum_{j=1}^m \frac{|D_j|}{|D|}(f_{D_j,\lambda} - f_\rho)\right\|_\rho^2$$

$$= \sum_{j=1}^m \frac{|D_j|^2}{|D|^2}\|f_{D_j,\lambda} - f_\rho\|_\rho^2 + \sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle f_{D_j,\lambda} - f_\rho, \sum_{k\neq j}\frac{|D_k|}{|D|}(f_{D_k,\lambda} - f_\rho)\right\rangle_\rho.$$

Taking expectations gives

$$E\left[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2\right] = \sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right]$$

$$+ \sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle E_{D_j}[f_{D_j,\lambda}] - f_\rho, E[\overline{f}_{D,\lambda}] - f_\rho - \frac{|D_j|}{|D|}\left(E_{D_j}[f_{D_j,\lambda}] - f_\rho\right)\right\rangle_\rho.$$

But

$$\sum_{j=1}^m \frac{|D_j|}{|D|}\left\langle E_{D_j}[f_{D_j,\lambda}] - f_\rho, E[\overline{f}_{D,\lambda}] - f_\rho\right\rangle_\rho = E\left[\left\langle \overline{f}_{D,\lambda} - f_\rho, E[\overline{f}_{D,\lambda}] - f_\rho\right\rangle_\rho\right]$$

$$= \left\|E[\overline{f}_{D,\lambda}] - f_\rho\right\|_\rho^2.$$

So we know that $E\left[\|\overline{f}_{D,\lambda} - f_\rho\|_\rho^2\right]$ equals

$$\sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right] - \sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \left\|E[f_{D_j,\lambda}] - f_\rho\right\|_\rho^2 + \left\|E[\overline{f}_{D,\lambda}] - f_\rho\right\|_\rho^2.$$

Furthermore, by the Schwarz inequality and $\sum_{j=1}^{m} \frac{|D_j|}{|D|} = 1$, we have

$$\|E[\overline{f}_{D,\lambda}] - f_\rho\|_\rho^2 = \left\|\sum_{j=1}^{m} \frac{|D_j|}{|D|} \left(E[f_{D_j,\lambda}] - f_\rho\right)\right\|_\rho^2 \leqslant \sum_{j=1}^{m} \frac{|D_j|}{|D|} \left\|E[f_{D_j,\lambda}] - f_\rho\right\|_\rho^2. \tag{42}$$

Then the desired bound follows. □

The estimate (42) for the general distributed spectral algorithms is worse than that for the distributed regularized least squares presented in [22], which leads to the restriction (9) on the number $m$ of local processors. It would be interesting to improve this estimate, at least for some families of spectral algorithms by imposing some conditions on the filter function $g_\lambda$.

By using proposition 4, we can now prove theorem 1.

**Proof of theorem 1.** We apply the bound (41) in proposition 4.

The proof in the case $\frac{1}{2} \leqslant r \leqslant \frac{3}{2}$ is easy. We apply theorem 2 to the data set $D_j$ with an arbitrarily fixed $j = 1, 2, \ldots, m$, and with confidence at least $1 - \delta$,

$$\|f_{D_j,\lambda} - f_\rho\|_\rho \leqslant C \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^{\max\{1,r\}} \left(\mathcal{B}_{|D_j|,\lambda} + \lambda^r\right) \left(\log\frac{4}{\delta}\right)^4.$$

Using (40), it is easy to derive

$$E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right] \leqslant 4\Gamma(9) C^2 \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^{\max\{2,2r\}} \left(\mathcal{B}_{|D_j|,\lambda} + \lambda^r\right)^2.$$

Then the first term in bound (41) can be estimated as

$$\begin{aligned} &\sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right] \\ &\leqslant 4\Gamma(9) C^2 \sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^{\max\{2,2r\}} \left(\mathcal{B}_{|D_j|,\lambda} + \lambda^r\right)^2. \end{aligned} \tag{43}$$

□

We turn to estimate the second term of (41). For each fixed $j \in \{1, 2, \ldots, m\}$, by Jensen's inequality, we have

$$\|E[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant E[\|E^*[f_{D_j,\lambda}] - f_\rho\|_\rho].$$

The proof of proposition 2 and the bounds (37) and (38) tell us that

$$\|E^*[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant 2^r(b + 1 + \gamma_r)\|u_\rho\|_\rho \lambda^r \Xi_{D_j}^r.$$

It follows that

$$\|E[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant 2^r(b+1+\gamma_r)\|u_\rho\|_\rho \lambda^r E\left[\Xi_{D_j}^r\right]. \tag{44}$$

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

$$\Xi_{D_j} \leqslant 2\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right].$$

Use the expectation formula (40) for the nonnegative random variable $\xi = \Xi_{D_j}^r$ and

$$\text{Prob}[\xi > t] = \text{Prob}[\xi^{\frac{1}{r}} > t^{\frac{1}{r}}] \leqslant 2\exp\left(-\frac{\lambda^{1/2}t^{\frac{1}{2r}}}{\sqrt{2}(\mathcal{B}_{|D_j|,\lambda}^2+\lambda)^{1/2}}\right).$$

We find

$$E\left[\Xi_{D_j}^r\right] \leqslant 2\int_0^\infty \exp\left(-\frac{\lambda^{1/2}t^{\frac{1}{2r}}}{\sqrt{2}(\mathcal{B}_{|D_j|,\lambda}^2+\lambda)^{1/2}}\right)\mathrm{d}t,$$

which equals

$$4r2^r\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^r\int_0^\infty u^{2r-1}\exp\{-u\}\,\mathrm{d}u.$$

Due to $\int_0^\infty u^{2r-1}\exp\{-u\}\,\mathrm{d}u = \Gamma(2r)$ we have

$$E\left[\Xi_{D_j}^r\right] \leqslant r\Gamma(2r)2^{r+2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^r.$$

Inserting the above estimate to (44), we have

$$\|E[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant r\Gamma(2r)2^{2r+2}(b+1+\gamma_r)\|u_\rho\|_\rho\lambda^r\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^r. \tag{45}$$

Combining (41), (43) and (45) we have

$$E[\|\bar{f}_{D,\lambda} - f_\rho\|_\rho^2] \leqslant 4\Gamma(9)C^2\sum_{j=1}^m\frac{|D_j|^2}{|D|^2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{\max\{2,2r\}}\left(\mathcal{B}_{|D_j|,\lambda}+\lambda^r\right)^2$$
$$+ r^2\Gamma^2(2r)(b+1+\gamma_r)^2 2^{4r+4}\|u_\rho\|_\rho^2\lambda^{2r}\sum_{j=1}^m\frac{|D_j|}{|D|}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{2r}.$$

This proves the desired bound in the case $1/2 \leqslant r \leqslant 3/2$.

The proof in the case $r > 3/2$ is more involved. Again we first use theorem 2 to bound $E[\|f_{D_j,\lambda} - f_\rho\|_\rho^2]$ and obtain

$$\sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda} - f_\rho\|_\rho^2\right]$$
$$\leqslant 6\Gamma(7)C^2 \sum_{j=1}^{m} \frac{|D_j|^2}{|D|^2} \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right]^2 \left[\mathcal{B}_{|D_j|,\lambda} + \left(\frac{\lambda}{|D_j|}\right)^{\frac{1}{2}} + \lambda^r\right]^2. \tag{46}$$

Then we estimate

$$\|E[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant E\|E^*[f_{D_j,\lambda}] - f_\rho\|_\rho$$

for each fixed $j$ as in the proof of proposition 3 to have

$$\|E^*[f_{D_j,\lambda}] - f_\rho\|_\rho \leqslant \Xi_{D_j}^{\frac{1}{2}} \left\|(\lambda I + L_{K,D_j})^{\frac{1}{2}}(g_\lambda(L_{K,D_j})L_{K,D_j} - I)L_K^{r-\frac{1}{2}}\right\| \|u_\rho\|_\rho$$
$$\leqslant \Xi_{D_j}^{\frac{1}{2}}\|(g_\lambda(L_{K,D_j})L_{K,D_j} - I)(L_{K,D_j} + \lambda I)^r\|\|u_\rho\|_\rho + \Xi_{D_j}^{\frac{1}{2}}\|u_\rho\|_\rho$$
$$\|(L_{K,D_j} + \lambda I)^{1/2}(g_\lambda(L_{K,D_j})L_{K,D_j} - I)((L_{K,D_j} + \lambda I)^{r-\frac{1}{2}} - (L_K + \lambda I)^{r-\frac{1}{2}})\|$$
$$=: B_{1,D_j} + B_{2,D_j}.$$

The first term $B_{1,D_j}$ above can be estimated easily from proposition 1 and lemma 5: with confidence at least $1 - \delta$, there holds

$$B_{1,D_j} \leqslant 2^{r+1/2} \left[\left(\frac{\mathcal{B}_{|D_j|,\lambda} \log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 1\right]^{1/2} \|u_\rho\|_\rho (b + 1 + \gamma_r)\lambda^r. \tag{47}$$

The most technical part of the proof is to estimate the second term $B_{2,D_j}$ above. We shall do so in two different cases on the following operator decomposition

$$(L_{K,D_j} + \lambda I)^{r-1/2} - (L_K + \lambda I)^{r-1/2}$$
$$= (L_{K,D_j} + \lambda I)\left((L_{K,D_j} + \lambda I)^{r-3/2} - (L_{K,D_j} + \lambda I)^{-1}(L_K + \lambda I)^{r-1/2}\right)$$
$$= (L_{K,D_j} + \lambda I)\Big(((L_{K,D_j} + \lambda I)^{r-3/2} - (L_K + \lambda I)^{r-3/2})$$
$$+ ((L_K + \lambda I)^{-1} - (L_{K,D_j} + \lambda I)^{-1})(L_K + \lambda I)^{r-1/2}\Big).$$

Case 1: $3/2 < r \leqslant 5/2$. In this case $0 < r - 3/2 \leqslant 1$. Then,

$$(L_{K,D_j} + \lambda I)^{r-3/2} - (L_K + \lambda I)^{r-3/2}$$
$$= (L_{K,D_j} + \lambda I)^{r-3/2}\left(I - (L_{K,D_j} + \lambda I)^{-(r-3/2)}(L_K + \lambda I)^{r-3/2}\right)$$

and

$$((L_K + \lambda I)^{-1} - (L_{K,D_j} + \lambda I)^{-1})(L_K + \lambda I)^{r-1/2}$$
$$= (L_K + \lambda I)^{-1}(L_{K,D_j} - L_K)(L_{K,D_j} + \lambda I)^{-1}(L_K + \lambda I)^{r-1/2}.$$

Thus we have

$$\begin{aligned} B_{2,D_j} &\leqslant \Xi_{D_j}^{1/2}\|u_\rho\|_\rho \left\|(L_{K,D_j}+\lambda I)^{1/2}(g_\lambda(L_{K,D_j})L_{K,D_j}-I)(L_{K,D_j}+\lambda I)\right. \\ &\quad \cdot \Big((L_{K,D_j}+\lambda I)^{r-3/2}\left(I-(L_{K,D_j}+\lambda I)^{-(r-3/2)}(L_K+\lambda I)^{r-3/2}\right) \\ &\quad \left. + \ (L_K+\lambda I)^{-1}(L_{K,D_j}-L_K)(L_{K,D_j}+\lambda I)^{-1}(L_K+\lambda I)^{r-1/2}\Big)\right\| \\ &\leqslant \Xi_{D_j}^{1/2}\|u_\rho\|_\rho \left\|(g_\lambda(L_{K,D_j})L_{K,D_j}-I)(L_{K,D_j}+\lambda I)^r\right\| \\ &\quad \cdot \left(1+\left\|(L_{K,D_j}+\lambda I)^{-(r-3/2)}(L_K+\lambda I)^{r-3/2}\right\|\right) \\ &\quad + \Xi_{D_j}^{1/2}\|u_\rho\|_\rho \left\|(g_\lambda(L_{K,D_j})L_{K,D_j}-I)(L_{K,D_j}+\lambda I)^{3/2}\right\| \\ &\quad \cdot \frac{1}{\sqrt{\lambda}}\left\|(L_K+\lambda I)^{-1/2}(L_{K,D_j}-L_K)\right\|\Xi_{D_j}\left\|(L_K+\lambda I)^{r-3/2}\right\|. \end{aligned}$$

Since $0<r-3/2\leqslant 1$, we have from (33) that

$$\left\|(L_{K,D_j}+\lambda I)^{-(r-3/2)}(L_K+\lambda I)^{r-3/2}\right\| \leqslant \Xi_{D_j}^{r-3/2}.$$

It follows from (24), lemma 5, and $\|L_K\|\leqslant\kappa^2$ that there exists a subset $Z_{\delta,1}^{|D_j|}$ of $\mathsf{Z}^{|D_j|}$ with measure $1-\delta$ such that for $D_j\in\mathcal{Z}_{\delta,1}^{|D_j|}$,

$$\begin{aligned} B_{2,D_j} &\leqslant \Xi_{D_j}^{1/2}\|u_\rho\|_\rho 2^r(b+1+\gamma_r)\lambda^r(1+\Xi_{D_j}^{r-3/2}) \\ &+ (2\kappa)^3\Xi_{D_j}^{3/2}\|u_\rho\|_\rho(b+1+\gamma_{3/2})\lambda\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}. \end{aligned}$$

According to proposition 1, there exists a subset $\mathcal{Z}_{\delta,2}^{|D_j|}$ of $\mathcal{Z}_{\delta,1}^{|D_j|}$ with measure $1-2\delta$ such that for $D_j\in\mathcal{Z}_{\delta,2}^{|D_j|}$,

$$\begin{aligned} B_{2,D_j} \leqslant 4^r\|u_\rho\|_\rho &\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{r-1}(b+1+\gamma_r)\lambda^r \\ + (2\kappa)^3\sqrt{2} &\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\|u_\rho\|_\rho(b+1+\gamma_{3/2})\lambda\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta} \\ \leqslant \|u_\rho\|_\rho &\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2} \\ \times &\left[4^r(b+1+\gamma_r)\lambda^r+(2\kappa)^3\sqrt{2}(b+1+\gamma_{3/2})\lambda\mathcal{B}_{|D_j|,\lambda}\log\frac{2}{\delta}\right]. \end{aligned}$$

This together with (47) tells us that with confidence at least $1-\delta$, $\|E^*[f_{D_j,\lambda}]-f_\rho\|_\rho$ is bounded by

$$
\begin{aligned}
&2^{r+1/2}\left[\left(\frac{\mathcal{B}_{|D|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{1/2}\|u_\rho\|_\rho(b+1+\gamma_r)\lambda^r\\
&\quad+\|u_\rho\|_\rho\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{4}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\\
&\quad\times\left[4^r(b+1+\gamma_r)\lambda^r+(2\kappa)^3\sqrt{2}(b+1+\gamma_{3/2})\lambda\mathcal{B}_{|D_j|,\lambda}\log\frac{4}{\delta}\right]\\
&\leqslant\tilde{C}'\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\left[\lambda^r+\lambda\mathcal{B}_{|D_j|,\lambda}\right]\left(\log\frac{4}{\delta}\right)^4,
\end{aligned}
$$

where

$$
\tilde{C}'=\|u_\rho\|_\rho\max\{(2^{r+1/2}+4^r)(b+1+\gamma_r),(2\kappa)^3\sqrt{2}(b+1+\gamma_{3/2})\}. \tag{48}
$$

Then, using the formula (40), we have

$$
E[\|E^*[f_{D_j,\lambda}]-f_\rho\|_\rho]\leqslant 4\tilde{C}'\Gamma(5)\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\left[\lambda^r+\lambda\mathcal{B}_{|D_j|,\lambda}\right].
$$

This means

$$
\|E[f_{D_j,\lambda}]-f_\rho\|_\rho^2\leqslant(4\tilde{C}'\Gamma(5))^2\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{3}\left[\lambda^r+\lambda\mathcal{B}_{|D_j|,\lambda}\right]^2.
$$

The above estimate together with (46) and (41) yields

$$
\begin{aligned}
E[\|\overline{f}_{D,\lambda}-f_\rho\|_\rho^2]\leqslant{}&6\Gamma(7)C^2\sum_{j=1}^m\frac{|D_j|^2}{|D|^2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\left[\mathcal{B}_{|D_j|,\lambda}+\left(\frac{\lambda}{|D_j|}\right)^{\frac{1}{2}}+\lambda^r\right]^2\\
&+(4\tilde{C}'\Gamma(5))^2\sum_{j=1}^m\frac{|D_j|}{|D|}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^3\left[\lambda^r+\lambda\mathcal{B}_{|D_j|,\lambda}\right]^2.
\end{aligned}
$$

This verifies the desired bound in the case $\frac{3}{2}<r\leqslant\frac{5}{2}$.

Case 2: $r>5/2$. In this case, we do not need to apply lemma 2 to control the norm of a product operator. Instead, we can apply lemma 4 directly (but lose the advantage of a tighter index $r$) to get

$$
\begin{aligned}
B_{2,D_j}\leqslant\Xi_{D_j}^{1/2}\|u_\rho\|_\rho\Big\|&(L_{K,D_j}+\lambda I)^{1/2}(g_\lambda(L_{K,D_j})L_{K,D_j}-I)(L_{K,D_j}+\lambda I)\\
&\cdot\Big(((L_{K,D_j}+\lambda I)^{r-3/2}-(L_K+\lambda I)^{r-3/2})\\
&+\ (L_K+\lambda I)^{-1}(L_{K,D_j}-L_K)(L_{K,D_j}+\lambda I)^{-1}(L_K+\lambda I)^{r-1/2}\Big)\Big\|.
\end{aligned}
$$

It follows from lemma 5 that

$$B_{2,D_j} \leqslant 2^{3/2}(b+1+\gamma_{3/2})\lambda^{3/2}\Xi_{D_j}^{1/2}\|u_\rho\|_\rho$$
$$\cdot\left((r-3/2)\kappa^{2r-5}\|L_{K,D}-L_K\|_{HS}+\Xi_{D_j}\frac{(2\kappa)^{2r-3}}{\sqrt{\lambda}}\|(L_K+\lambda I)^{-1/2}(L_{K,D_j}-L_K)\|\right).$$

From proposition 1, (23) and (24), with confidence at least $1-\delta$, there holds

$$B_{2,D_j} \leqslant 4(b+1+\gamma_{3/2})\lambda^{3/2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{1/2}\|u_\rho\|_\rho$$
$$\cdot\left((r-3/2)\kappa^{2r-5}\frac{4\kappa^2\log 6/\delta}{\sqrt{|D_j|}}+2\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]\frac{(2\kappa)^{2r-3}}{\sqrt{\lambda}}\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}\right).$$

Combining the above estimate with (47), we obtain, with confidence $1-\delta$,

$$\|E^*[f_{D_j,\lambda}]-f_\rho\|_\rho \leqslant 2^{r+1/2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{1/2}\|u_\rho\|_\rho(b+1+\gamma_r)\lambda^r$$
$$+4(b+1+\gamma_{3/2})\lambda^{3/2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]^{1/2}\|u_\rho\|_\rho$$
$$\cdot\left((r-3/2)\kappa^{2r-5}\frac{4\kappa^2\log 6/\delta}{\sqrt{|D_j|}}+2\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}}{\sqrt{\lambda}}\right)^2+1\right]\frac{(2\kappa)^{2r-3}}{\sqrt{\lambda}}\mathcal{B}_{|D_j|,\lambda}\log\frac{6}{\delta}\right)$$
$$\leqslant C''\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\left(\lambda^r+\lambda^{3/2}\frac{1}{\sqrt{|D_j|}}+\lambda\mathcal{B}_{|D_j|,\lambda}\right)\left(\log\frac{6}{\delta}\right)^4,$$

where

$$C''=\|u_\rho\|_\rho\max\{2^{r+1/2}(b+1+\gamma_r),4(b+1+\gamma_{3/2})(\gamma-3/2)\kappa^{2r-5},8(2b+\gamma_{3/2})(2\kappa)^{2r-3}\}.$$

By (40), we have

$$E[\|E^*[f_{D_j,\lambda}]-f_\rho\|_\rho]\leqslant C''6\Gamma(5)\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^{3/2}\left(\lambda^r+\lambda^{3/2}\frac{1}{\sqrt{|D_j|}}+\lambda\mathcal{B}_{|D_j|,\lambda}\right).$$

Hence

$$\|E[f_{D_j,\lambda}]-f_\rho\|_\rho^2\leqslant(C''6\Gamma(5))^2\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^3\left(\lambda^r+\lambda^{3/2}\frac{1}{\sqrt{|D_j|}}+\lambda\mathcal{B}_{|D_j|,\lambda}\right)^2.$$

The above estimate, together with (46) and (41), tells us that $E[\|\overline{f}_{D,\lambda}-f_\rho\|_\rho^2]$ is bounded by

$$6\Gamma(7)C^2\sum_{j=1}^{m}\frac{|D_j|^2}{|D|^2}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\left[\mathcal{B}_{|D_j|,\lambda}+\left(\frac{\lambda}{|D_j|}\right)^{\frac{1}{2}}+\lambda^r\right]^2$$
$$+(C''6\Gamma(5))^2\sum_{j=1}^{m}\frac{|D_j|}{|D|}\left[\left(\frac{\mathcal{B}_{|D_j|,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^3\left(\lambda^r+\lambda^{3/2}\frac{1}{\sqrt{|D_j|}}+\lambda\mathcal{B}_{|D_j|,\lambda}\right)^2.$$

This proves the desired bound in the case of $r > 5/2$. The proof of theorem 1 is complete. □

**Proof of corollary 1.** Putting the choice $\lambda = N^{-1/(2r+\beta)}$ into condition (8) yields $\mathcal{N}(\lambda) \leqslant C_0N^{\beta/(2r+\beta)}$. For $1/2 \leqslant r \leqslant 3/2$, since $m \leqslant N^{(2r-1)/(2r+\beta)}$ and $|D_1| = |D_2| = \cdots = |D_m|$, we have

$$\frac{\mathcal{N}(\lambda)}{\lambda|D_j|} \leqslant C_0mN^{\frac{1-2r}{2r+\beta}} \leqslant C_0.$$

We also have, for each $j = 1, \ldots, m$,

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

and

$$\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} \leqslant \frac{4\kappa^2}{|D|}\left\{\frac{\kappa}{\sqrt{|D_j|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\right\}^2 \leqslant 4\kappa^2(\kappa+\sqrt{C_0})^2N^{-2r/(2r+\beta)}.$$

Then by theorem 1,

$$\begin{aligned}E[\|\overline{f}_{D,\lambda}-f_\rho\|_\rho^2] &\leqslant \tilde{C}\frac{1}{m}\sum_{j=1}^{m}(4\kappa^2(\kappa+\sqrt{C_0})^2)^3(4\kappa^2(\kappa+\sqrt{C_0})^2+1)N^{-\frac{2r}{2r+\beta}}\\ &= \hat{C}N^{-\frac{2r}{2r+\beta}},\end{aligned}$$

where

$$\hat{C} = \tilde{C}(4\kappa^2(\kappa+\sqrt{C_0})^2)^3(4\kappa^2(\kappa+\sqrt{C_0})^2+1).$$

For $r > 3/2$, since $m \leqslant N^{2/(2r+\beta)}$, we have

$$\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda} \leqslant 4\kappa^2(\kappa+\sqrt{C_0})^2N^{-2r/(2r+\beta)},$$

and

$$\lambda^2\mathcal{B}^2_{|D_j|,\lambda} \leqslant 4\kappa^2(\kappa+\sqrt{C_0})^2N^{-2r/(2r+\beta)}.$$

So by theorem 1, we have

$$\begin{aligned}E[\|\bar{f}_{D,\lambda}-f_\rho\|_\rho^2] &\leqslant \tilde{C}\frac{1}{m}\sum_{j=1}^{m}(4\kappa^2(\kappa+\sqrt{C_0})^2+1)^3(8\kappa^2(\kappa+\sqrt{C_0})^2+3)N^{-\frac{2r}{2r+\beta}}\\ &=\hat{C}'N^{-\frac{2r}{2r+\beta}},\end{aligned}$$

where

$$\hat{C}'=\tilde{C}(4\kappa^2(\kappa+\sqrt{C_0})^2+1)^3(8\kappa^2(\kappa+\sqrt{C_0})^2+3).$$

This completes the proof of corollary 1. □

## Acknowledgments

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. 11401524, No. 11531013, No. 61502342, and No. 11471292]. The corresponding author is Shao-Bo Lin.

## References

[1] Bathia R 1997 *Matrix Analysis* (*Graduate Texts in Mathematics*) vol 169 (New York: Springer)

[2] Bauer F, Pereverzev S and Rosasco L 2007 On regularziation algorithms in learning theory *J. Complex.* **23** 52–72

[3] Blanchard G and Krämer N 2010 Optimal learning rates for kernel conjugate gradient regression *Neural Inf. Processing Systems Conf.* pp 226–34

[4] Caponnetto A and DeVito E 2007 Optimal rates for the regularized least squares algorithm *Found. Comput. Math.* **7** 331–68

[5] Caponnetto A and Yao Y 2010 Cross-validation based adaptation for regularization operators in learning theory *Anal. Appl.* **8** 161–83

[6] Cucker F and Zhou D X 2007 *Learning Theory: an Approximation Theory Viewpoint* (Cambridge: Cambridge University Press)

[7] De Vito E, Rosasco L, Caponnetto A, De Giovannini U and Odone F 2005 Learning from examples as an inverse problem *J. Mach. Learn. Res.* **6** 883–904

[8] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (*Mathematics and its Applications*) vol 375 (Dordrecht: Kluwer)

[9] Hsu D, Kakade S M and Zhang T 2014 Random design analysis of ridge regression *Found. Comput. Math.* **14** 569–600

[10] Hu T, Fan J, Wu Q and Zhou D X 2015 Regularization schemes for minimum error entropy principle *Anal. Appl.* **13** 437–55

[11] Lin S B, Guo X and Zhou D X 2016 Distributed learning with least square regularization *J. Mach. Learn. Res.* (arXiv:1608.03339v2) (submitted in revised version in 2016, under review)

[12] Lo Gerfo L, Rosasco L, Odone F, De Vito E and Verri A 2008 Spectral algorithms for supervised learning *Neural Comput.* **20** 1873–97

[13] Mann G, McDonald R, Mohri M, Silberman N and Walker D 2009 Efficient large-scale distributed training of conditional maximum entropy models *Neural Inf. Processing Systems Conf.* pp 1231–9

[14] Shamir N O 2014 Distributed stochastic optimization and learning *52nd Annual Allerton Conf. on Communication, Control and Computing*

[15] Smale S and Zhou D X 2004 Shannon sampling and function reconstruction from point values *Bull. Am. Math. Soc.* **41** 279–305

[16] Smale S and Zhou D X 2005 Shannon sampling II: connections to learning theory *Appl. Comput. Harmon. Anal.* **19** 285–302

[17] Smale S and Zhou D X 2007 Learning theory estimates via integral operators and their approximations *Constr. Approx.* **26** 153–72

[18] Steinwart I, Hush D and Scovel C 2009 Optimal rates for regularized least squares regression *Proc. of the 22nd Annual Conf. on Learning Theory* ed S Dasgupta and A Klivans pp 79–93

[19] Williamson R C, Smola A J and Schölkopf B 2001 Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators *IEEE Trans. Inf. Theory* **47** 2516–32

[20] Yao Y, Rosasco L and Caponnetto A 2007 On early stopping in gradient descent learning *Constr. Approx.* **26** 289–315

[21] Zhang Y C, Duchi J and Wainwright M 2013 Communication-efficient algorithms for statistical optimization *J. Mach. Learn. Res.* **14** 3321–63

[22] Zhang Y C, Duchi J and Wainwright M 2015 Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates *J. Mach. Learn. Res.* **16** 3299–340

[23] Zhou D X 2002 The covering number in learning theory *J. Complex.* **18** 739–67