Optimal Convergence for Distributed Learning with Stochastic Gradient Methods and Spectral Algorithms

Junhong Lin junhong.lin@epfl.ch

Volkan Cevher volkan.cevher@epfl.ch

Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne
CH1015-Lausanne, Switzerland

Editor:

Abstract

We study generalization properties of distributed algorithms in the setting of nonparametric regression over a reproducing kernel Hilbert space (RKHS). We first investigate distributed stochastic gradient methods (SGM), with mini-batches and multi-passes over the data. We show that optimal generalization error bounds (up to logarithmic factor) can be retained for distributed SGM provided that the partition level is not too large. We then extend our results to spectral algorithms (SA), including kernel ridge regression (KRR), kernel principal component analysis, and gradient methods. Our results are superior to the state-of-the-art theory. Particularly, our results show that distributed SGM has a smaller theoretical computational complexity, compared with distributed KRR and classic SGM. Moreover, even for non-distributed SA, they provide the first optimal, capacity-dependent convergence rates, for the case that the regression function may not be in the RKHS.

Keywords: Kernel Methods, Stochastic Gradient Methods, Spectral Regularization, Distributed Learning

1. Introduction

In statistical learning theory, a set of N input-output pairs from an unknown distribution is observed. The aim is to learn a function which can predict future outputs given the corresponding inputs. The quality of a predictor is often measured in terms of the mean-squared error. In this case, the conditional mean, which is called as the regression function, is optimal among all the measurable functions (Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

In nonparametric regression problems, the properties of the regression function are not known a priori. Nonparametric approaches, which can adapt their complexity to the problem, are key to good results. Kernel methods is one of the most common nonparametric approaches to learning (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). It is based on choosing a RKHS as the hypothesis space in the design of learning algorithms. With an appropriate reproducing kernel, RKHS can be used to approximate any smooth function.

The classical algorithms to perform learning task are regularized algorithms, such as KRR (also called as Tikhonov regularization in inverse problems), kernel principal component regression (KPCR, also known as spectral cut-off regularization in inverse problems), and more generally, SA. From the point of view of inverse problems, such approaches amount to solving an empirical, linear operator equation with the empirical covariance operator replaced by a regularized one (Engl et al., 1996; Bauer et al., 2007; Gerfo et al., 2008). Here, the regularization term controls the complexity of the solution to against over-fitting and to ensure best generalization ability. Statistical results on generalization error had been developed in (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) for KRR and in (Caponnetto, 2006; Bauer et al., 2007) for SA.

Another type of algorithms to perform learning tasks is based on iterative procedure (Engl et al., 1996). In this kind of algorithms, an empirical objective function is optimized in an iterative way with no explicit constraint or penalization, and the regularization against overfitting is realized by early-stopping the empirical procedure. Statistical results on generalization error and the regularization roles of the number of iterations/passes have been investigated in (Zhang and Yu, 2005; Yao et al., 2007) for gradient methods (GM, also known as Landweber algorithm in inverse problems), in (Caponnetto, 2006; Bauer et al., 2007) for accelerated gradient methods (AGM, known as ν-methods in inverse problems) in (Blanchard and Krämer, 2010) for conjugate gradient methods (CGM), and in (Lin and Rosasco, 2017b) for (multi-pass) SGM. Interestingly, GM and AGM can be viewed as special instances of SA (Bauer et al., 2007), but CGM and SGM can not (Blanchard and Krämer, 2010; Lin and Rosasco, 2017b).

The above mentioned algorithms suffer from computational burdens at least of order $O(N^2)$ due to the nonlinearity of kernel methods. Indeed, a standard execution of KRR requires $O(N^2)$ in space and $O(N^3)$ in time, while SGM after T-iterations requires $O(N)$ in space and $O(NT)$ (or T^2) in space. Such approaches would be prohibitive when dealing with large-scale learning problems. These thus motivate one to study distributed learning algorithms (Mcdonald et al., 2009; Zhang et al., 2012). The basic idea of distributed learning is very simple: randomly divide a dataset of size N into m subsets of equal size, compute an independent estimator using a fixed algorithm on each subset, and then average the local solutions into a global predictor. Interestingly, distributed learning technique has been successfully combined with KRR (Zhang et al., 2015; Lin et al., 2017) and more generally, SA (Guo et al., 2017; Blanchard and Mucke, 2016b), and it has been shown that statistical results on generalization error can be retained provided that the number of partitioned subsets is not too large. Moreover, it was highlighted (Zhang et al., 2015) that distributed KRR not only allows one to handle large datasets that restored on multiple machines, but also leads to a substantial reduction in computational complexity versus the standard approach of performing KRR on all N samples.

In this paper, we study distributed SGM, with multi-passes over the data and mini-batches. The algorithm is a combination of distributed learning technique and (multi-pass) SGM (Lin and Rosasco, 2017b): it randomly partitions a dataset of size N into m subsets of equal size, computes an independent estimator by SGM for each subset, and then averages the local solutions into a global predictor. We show that with appropriate choices of algorithmic parameters, optimal generalization error bounds up to a logarithmic factor can be achieved for distributed SGM provided that the partition level m is not too large.

The proposed configuration has certain advantages on computational complexity. For example, without considering any benign properties of the problem such as the regularity of the regression function (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) and a capacity assumption on the RKHS (Zhang, 2005; Caponnetto and De Vito, 2007), even implementing on a single machine, distributed SGM has a convergence rate of order $O(N^{-1/2}\log N)$, with a computational complexity $O(N)$ in space and $O(N^{3/2})$ in time, compared with $O(N)$ in space and $O(N^2)$ in time of classic SGM performing on all N samples, or $O(N^{3/2})$ in space and $O(N^2)$ in time of distributed KRR. Moreover, the approach dovetails naturally with parallel and distributed computation: we are guaranteed a superlinear speedup with m parallel processors (though we must still communicate the function estimates from each processor).

The proof of the main results is based on a similar (but a bit different) error decomposition from (Lin and Rosasco, 2017b), which decomposes the excess risk into three terms: bias, sample and computational variances. The error decomposition allows one to study distributed GM and distributed SGM simultaneously. Different to those in (Lin and Rosasco, 2017b) which rely heavily on the intrinsic relationship of GM with the square loss, in this paper, an integral operator approach (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) is used, combining with some novel and refined analysis, see Subsection 6.2 for further details.

We then extend our analysis to distributed SA and derive similar optimal results on generalization error for distributed SA, based on the fact that GM is a special instance of SA.

This paper is an extended version of the conference version (Lin and Cevher, 2018) where results for distributed SGM are given only. In this version, we additionally provide statistical results for distributed SA, including their proofs, as well as some further discussions.

We highlight that our contributions are as follows.

- We provide the first results with optimal convergence rates (up to a logarithmic factor) for distributed SGM, showing that distributed SGM has a smaller theoretical computational complexity, compared with distributed KRR and non-distributed SGM. As a byproduct, we derive optimal convergence rates (up to a logarithmic factor) for non-distributed SGM, which improve the results in (Lin and Rosasco, 2017b).
- Our results for distributed SA improves previous results from (Zhang et al., 2015) for distributed KRR, and from (Guo et al., 2017) for distributed SA, with a less strict condition on the partition number m. Moreover, they provide the first optimal rates for distributed SA in the non-attainable cases (i.e., the regression function may not be in the RKHS).
- As a byproduct, we provide the first results with optimal, capacity-dependent rates for the non-distributed SA in the non-attainable cases, filling a theoretical gap since (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) for KRR using the integral-operator approach.

The remainder of the paper is organized as follows. Section 2 introduces the supervised learning setting. Section 3 describes distributed SGM, and then presents theoretical results on generalization error for distributed SGM, following with simple comments. Section 4 introduces distributed SA, and then gives statistical results on generalization error. Section 5 discusses and compares our results with related work. Section 6 provides the proofs for distributed SGM. Finally, proofs for auxiliary lemmas and results for distributed SA are provided in the appendix.

2. Supervised Learning Problems

We consider a supervised learning problem. Let ρ be a probability measure on a measurable space $Z = X \times Y$, where X is a compact-metric input space and $Y \subseteq \mathbb{R}$ is the output space. ρ is fixed but unknown. Its information can be only known through a set of samples $\bar{\mathbf{z}} = \{z_i = (x_i, y_i)\}_{i=1}^N$ of $N \in \mathbb{N}$ points, which we assume to be i.i.d.. We denote $\rho_X(\cdot)$ the induced marginal measure on H of ρ and $\rho(|x)$ the conditional probability measure on $\mathbb{R}$ with respect to $x \in H$ and ρ. We assume that ρ_X has full support in X throughout.

The quality of a predictor $f : X \to Y$ can be measured in terms of the expected risk with a square loss defined as

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho(z). \tag{1}$$

In this case, the function minimizing the expected risk over all measurable functions is the regression function given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X. \tag{2}$$

The performance of an estimator $f \in L^2_{\rho_X}$ can be measured in terms of generalization error (excess risk), i.e., $\mathcal{E}(f) - \mathcal{E}(f_\rho)$. It is easy to prove that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \tag{3}$$

Here, $L^2_{\rho_X}$ is the Hilbert space of square integral functions with respect to ρ_X, with its induced norm given by $\|f\|_\rho = \|f\|_{L^2_{\rho_X}} = \left(\int_X |f(x)|^2 d\rho_X\right)^{1/2}$.

Kernel methods are based on choosing the hypothesis space as a RKHS. Recall that a reproducing kernel K is a symmetric function $K : X \times X \to \mathbb{R}$ such that $(K(u_i, u_j))_{i,j=1}^{\ell}$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^{\ell}$ in X. The reproducing kernel K defines a RKHS $(H, \|\cdot\|_H)$ as the completion of the linear span of the set $\{K_x(\cdot) := K(x, \cdot) : x \in X\}$ with respect to the inner product $\langle K_x, K_u \rangle_H := K(x, u)$.

Given only the samples $\bar{\mathbf{z}}$, the goal is to learn the regression function through efficient algorithms.

3. Distributed Learning with Stochastic Gradient Methods

In this section, we first state the distributed SGM. We then present theoretical results for distributed SGM and non-distributed SGM, following with simple discussions.

3.1 Distributed SGM

Throughout this paper, as that in (Zhang et al., 2015), we assume that[1] the sample size $N = mn$ for some positive integers n, m, and we randomly decompose $\bar{\mathbf{z}}$ as $\mathbf{z}_1 \cup \mathbf{z}_2 \cup \cdots \cup \mathbf{z}_m$ with $|\mathbf{z}_1| = |\mathbf{z}_2| = \cdots = |\mathbf{z}_m| = n$. For any $s \in [m]$, we write $\mathbf{z}_s = \{(x_{s,i}, y_{s,i})\}_{i=1}^n$. We study distributed SGM, with mini-batches and multi-pass over the data, as detailed in Algorithm 1. For any $t \in \mathbb{N}^+$, the set of the first t positive integers is denoted by $[t]$.

Algorithm 1 Distributed learning with stochastic gradient methods

Input: Number of partitions m, mini-batch size $b \leq N/m$, total number of iterations T, step-size sequence $\{\eta_t > 0\}_{t=1}^T$, and kernel function $K(\cdot, \cdot)$

1: Divide $\bar{\mathbf{z}}$ evenly and uniformly at random into the m disjoint subsets, $\mathbf{z}_1, \cdots, \mathbf{z}_m$.

2: For every $s \in [m]$, compute a local estimate via b-minibatch SGM over the sample $\mathbf{z}_s$: $f_{s,1} = 0$ and

$$f_{s,t+1} = f_{s,t} - \eta_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} (f_{s,t}(x_{s,j_{s,i}}) - y_{s,j_{s,i}}) K_{x_{s,j_{s,i}}}, \qquad t \in [T]. \tag{4}$$

Here, $j_{s,1}, j_{s,2}, \cdots, j_{s,bT}$ are i.i.d. random variables from the uniform distribution on $[n]$.[2]

3: Take the averaging over these local estimators: $\bar{f}_T = \frac{1}{m} \sum_{s=1}^m f_{s,T}$.

Output: the function $\bar{f}_T$

In the algorithm, at each iteration t, for each $s \in [m]$, the local estimator updates its current solution by subtracting a scaled gradient estimate. It is easy to see that the gradient estimate at each iteration for the s-th local estimator is an unbiased estimate of the full gradient of the

1. For the general case, one can consider the weighted averaging scheme, as that in (Lin et al., 2017), and our analysis still applies with a simple modification.

empirical risk over $\mathbf{z}_s$. The global predictor is the average over these local solutions. In the special case $m = 1$, the algorithm reduces to the classic multi-pass SGM.

There are several free parameters, the step-size η_t, the mini-batch size b, the total number of iterations/passes, and the number of partition/subsets m. All these parameters will affect the algorithm's generalization properties and computational complexity. In the coming subsection, we will show how these parameters can be chosen so that the algorithm can generalize optimally, as long as the number of subsets m is not too large. Different choices on η_t, b, and T correspond to different regularization strategies. In this paper, we are particularly interested in the cases that both η_t and b are fixed as some universal constants that may depend on the local sample size n, while T is tuned.

The total number of iterations T can be bigger than the local sample size n, which means that the algorithm can use the data more than once, or in another words, we can run the algorithm with multiple passes over the data. Here and in what follows, the number of (effective) 'passes' over the data is referred to $\frac{bt}{n}$ after t iterations of the algorithm.

The numerical realization of the algorithm and its performance on a synthesis data can be found in (Lin and Cevher, 2018). The space and time complexities for each local estimator are

$$O(n) \quad \text{and} \quad O(bnT), \tag{5}$$

respectively. The total space and time complexities of the algorithm are

$$O(N) \quad \text{and} \quad O(bNT). \tag{6}$$

3.2 Generalization Properties for Distributed Stochastic Gradient Methods

In this section, we state our results for distributed SGM, following with simple discussions. Throughout this paper, we make the following assumptions.

Assumption 1 *H is separable and K is continuous. Furthermore, for some $\kappa \in [1, \infty[$,*

$$K(x, x) \leq \kappa^2, \quad \forall x \in X, \tag{7}$$

and for some $M, \sigma \geq 0$,

$$\int_Y y^2 d\rho(y|x) \leq M^2,$$

$$\int_Y (f_\rho(x) - y)^2 d\rho(y|x) \leq \sigma^2, \quad \rho_X\text{-almost surely.} \tag{8}$$

The above assumptions are quite common in statistical learning theory, see e.g., (Steinwart and Christmann, 2008; Cucker and Zhou, 2007). The constant σ from Equation (8) measures the noise level of the studied problem. The condition $\int_Y y^2 d\rho(y|x) \leq M^2$ implies that the regression function is bounded almost surely,

$$|f_\rho(x)| \leq M. \tag{9}$$

It is trivially satisfied when Y is bounded, for example, $Y = \{-1, 1\}$ in the classification problem. To state our first result, we define an inclusion operator $\mathcal{S}_\rho : H \to L^2_{\rho_X}$, which is continuous under Assumption (7).

Corollary 1 *Assume that $f_\rho \in H$ and*

$$m \leq N^{\beta}, \quad 0 \leq \beta < \frac{1}{2}.$$

Consider Algorithm 1 with any of the following choices on η_t, b and T.
1) $\eta_t = \eta \simeq m/\sqrt{N}$ for all $t \in [T_*]$, $b = 1$, and $T_* = N/m$.
2) $\eta_t = \eta = \simeq \frac{1}{\log N}$ for all $t \in [T_*]$, $b \simeq \sqrt{N}/m$, and $T_* \simeq \sqrt{N} \log N$.
Then,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{T_*+1} - f_\rho\|_\rho^2 \lesssim N^{-1/2} \log N.$$

Here and throughout this section, we use the notations $a_1 \lesssim a_2$ to mean $a_1 \leq C a_2$ for some positive constant C depending only on $\kappa, M, \sigma, \|\mathcal{S}_\rho\|, \|f_\rho\|_H$, and $a_1 \simeq a_2$ to mean $a_2 \lesssim a_1 \lesssim a_2$.

The above result provides generalization error bounds for distributed SGM with two different choices on step-size η_t, mini-batch size b and total number of iterations/passes. The convergence rate is optimal up to a logarithmic factor, in the sense that it nearly matches the minimax rate $N^{-1/2}$ in (Caponnetto and De Vito, 2007) and the convergence rate $N^{-1/2}$ for KRR (Smale and Zhou, 2007; Caponnetto and De Vito, 2007). The number of passes to achieve optimal error bounds in both cases is roughly one. The above result asserts that distributed SGM generalizes optimally after one pass over the data for two different choices on step-size and mini-batch size, provided that the partition level m is not too large. In the case that $m \simeq \sqrt{N}$, according to (6), the computational complexities are $O(N)$ in space and $O(N^{1.5})$ in time, comparing with $O(N)$ in space and $O(N^2)$ in time of classic SGM.

Corollary 1 provides statistical results for distributed SGM without considering any further benign assumptions about the learning problem, such as the regularity of the regression function and the capacity of the RKHS. In what follows, we will show how the results can be further improved, if we make these two benign assumptions.

The first benign assumption relates to the regularity of the regression function. We introduce the integer operator $\mathcal{L} : L^2_{\rho_X} \to L^2_{\rho_X}$, defined by $\mathcal{L}f = \int_X f(x) K(x, \cdot) d\rho_X$. Under Condition (7), $\mathcal{L}$ is positive trace class operators (Cucker and Zhou, 2007), and hence $\mathcal{L}^\zeta$ is well defined using the spectral theory.

Assumption 2 *There exist $\zeta > 0$ and $R > 0$, such that $\|\mathcal{L}^{-\zeta} f_\rho\|_\rho \leq R$.*

This assumption characterizes how large the subspace that the regression function lies in. The bigger the ζ is, the smaller the subspace is, the stronger the assumption is, and the easier the learning problem is, as $\mathcal{L}^{\zeta_1}(L^2_{\rho_X}) \subseteq \mathcal{L}^{\zeta_2}(L^2_{\rho_X})$ if $\zeta_1 \geq \zeta_2$. Moreover, if $\zeta = 0$, we are making no assumption, and if $\zeta = \frac{1}{2}$, we are requiring that there exists some $f_H \in H$ such that $f_H = f_\rho$ almost surely (Steinwart and Christmann, 2008, Section 4.5).

The next assumption relates to the capacity of the hypothesis space.

Assumption 3 *For some $\gamma \in [0, 1]$ and $c_\gamma > 0$, $\mathcal{L}$ satisfies*

$$\operatorname{tr}(\mathcal{L}(\mathcal{L} + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \textit{for all } \lambda > 0. \tag{10}$$

The left hand-side of (10) is called effective dimension (Zhang, 2005) or degrees of freedom (Caponnetto and De Vito, 2007). It is related to covering/entropy number conditions, see (Steinwart and Christmann, 2008). The condition (10) is naturally satisfied with $\gamma = 1$, since $\mathcal{L}$ is a trace class operator which implies that its eigenvalues $\{\sigma_i\}_i$ satisfy $\sigma_i \lesssim i^{-1}$. Moreover, if the eigenvalues of $\mathcal{L}$ satisfy a polynomial decaying condition $\sigma_i \sim i^{-c}$ for some $c > 1$, or if $\mathcal{L}$

is of finite rank, then the condition (10) holds with $\gamma = 1/c$, or with $\gamma = 0$. The case $\gamma = 1$ is refereed as the capacity independent case. A smaller γ allows deriving faster convergence rates for the studied algorithms, as will be shown in the following results.

Making these two assumptions, we have the following general results for distributed SGM.

Theorem 1 *Under Assumptions 2 and 3, let $\eta_t = \eta$ for all $t \in [T]$ with η satisfying*

$$0 < \eta \leq \frac{1}{4\kappa^2 \log T}. \tag{11}$$

Then for any $t \in [T]$ and $\tilde{\lambda} = n^{\theta-1}$ with $\theta \in [0,1]$, the following results hold.
1) For $\zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{t+1} - f_\rho\|_\rho^2 \leq ((\tilde{\lambda}\eta t)^2 \vee Q_{\gamma,\theta,n}^{2\zeta\vee 1} \vee \log t)[C_5 \frac{(R + \mathbf{1}_{\{2\zeta<1\}}\|f_\rho\|_\infty)^2}{(\eta t)^{2\zeta}} + C_8 \frac{\sigma^2}{N\tilde{\lambda}^\gamma} + C_{10}\frac{M^2\eta}{mb}]. \tag{12}$$

2) For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{t+1} - f_\rho\|_\rho^2 \leq ((\tilde{\lambda}\eta t)^{2\zeta} \vee Q_{\gamma,\theta,n} \vee (\frac{(\eta t)^{2\zeta-1}}{n^{(\zeta-1/2)\wedge 1}}) \vee \log t)[C_5 \frac{R^2}{(\eta t)^{2\zeta}} + C_8 \frac{\sigma^2}{N\tilde{\lambda}^\gamma} + C_{10}\frac{M^2\eta}{mb}]. \tag{13}$$

Here,

$$Q_{\gamma,\theta,n} = 1 \vee [\gamma(\theta^{-1} \wedge \log n)] \tag{14}$$

and C_5, C_6, C_8, C_{10} are positive constants depending only on $\kappa^2, \zeta, c_\gamma, \|\mathcal{L}\|$ which will be given explicitly in the proof, see (62), (63), (67) and (74).

In the above result, we only consider the setting of a fixed step-size. Results with a decaying step-size can be directly derived following our proofs in the coming sections, combining with some basic estimates from (Lin and Rosasco, 2017b). The error bound from (12) depends on the number of iteration t, the step-size η, the mini-batch size, the number of sample points N and the partition level m. It holds for any pseudo regularization parameter $\tilde{\lambda}$ where $\tilde{\lambda} \in [n^{-1}, 1]$. When $t \leq n/\eta$, for $\zeta \leq 1$, we can choose $\tilde{\lambda} = (\eta t)^{-1}$, and ignoring the logarithmic factor and constants, (12) reads as

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{t+1} - f_\rho\|_\rho^2 \lesssim \frac{1}{(\eta t)^{2\zeta}} + \frac{(\eta t)^\gamma}{N} + \frac{\eta}{mb}. \tag{15}$$

The right-hand side of the above inequality is composed of three terms. The first term is related to the regularity parameter ζ of the regression function f_ρ, and it results from estimating bias. The second term depends on the sample size N, and it results from estimating sample variance. The last term results from estimating computational variance due to random choices of the sample points. In comparing with the error bounds derived for classic SGM performed on a local machine, one can see that averaging over the local solutions can reduce sample and computational variances, but keeps bias unchanged. As the number of iteration t increases, the bias term decreases, and the sample variance term increases. This is a so-called trade-off problem in statistical learning theory. Solving this trade-off problem leads to the best choice on number of iterations. Notice that the computational variance term is independent of the number of iterations t and it depends on the step-size, the mini-batch size, and the partition level. To derive optimal rates, it is necessary to choose a small step-size, and/or a large mini-batch size, and a suitable partition level. In what follows, we provide different choices of these algorithmic parameters, corresponding to different regularization strategies, while leading to the same optimal convergence rates up to a logarithmic factor.

Corollary 2 *Under Assumptions 2 and 3, let $\zeta \leq 1$, $2\zeta + \gamma > 1$ and*

$$m \leq N^{\beta}, \quad \text{with } 0 \leq \beta < \frac{2\zeta + \gamma - 1}{2\zeta + \gamma}. \tag{16}$$

Consider Algorithm 1 with any of the following choices on η_t, b and T_*.
1) $\eta_t \simeq n^{-1}$ for all $t \in [T_*]$, $b = 1$, and $T_* \simeq N^{\frac{1}{2\zeta+\gamma}} n$.
2) $\eta_t \simeq n^{-1/2}$ for all $t \in [T_*]$, $b \simeq \sqrt{n}$, and $T_* \simeq N^{\frac{1}{2\zeta+\gamma}} \sqrt{n}$.
3) $\eta_t \simeq N^{-\frac{2\zeta}{2\zeta+\gamma}} m$ for all $t \in [T_*]$, $b = 1$, and $T_* \simeq N^{\frac{2\zeta+1}{2\zeta+\gamma}}/m$.
4) $\eta_t \simeq \frac{1}{\log N}$ for all $t \in [T_*]$, $b \simeq N^{\frac{2\zeta}{2\zeta+\gamma}}/m$, and $T_* \simeq N^{\frac{1}{2\zeta+\gamma}} \log N$.
Then,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{T_*+1} - f_\rho\|_\rho^2 \lesssim N^{-\frac{2\zeta}{2\zeta+\gamma}} \log N.$$

We add some comments on the above theorem. First, the convergence rate is optimal up to a logarithmic factor, as it is almost the same as that for KRR from (Caponnetto and De Vito, 2007; Smale and Zhou, 2007) and also it nearly matches the minimax lower rate $O(N^{-\frac{2\zeta}{2\zeta+\gamma}})$ in (Caponnetto and De Vito, 2007). In fact, let $\mathcal{P}(\gamma,\zeta)$ ($\gamma \in (0,1)$ and $\zeta \in [1/2,1]$) be the set of probability measure ρ on Z, such that Assumptions 1-3 are satisfied. Then the following minimax lower rate is a direct consequence of (Caponnetto and De Vito, 2007, Theorem 2):

$$\liminf_{N\to\infty} \inf_{f^N} \sup_{\rho \in \mathcal{P}(\gamma,\zeta)} \Pr\left(\bar{\mathbf{z}} \in Z^N : \mathbb{E}\|\mathcal{S}_\rho f^N - f_\rho\|_\rho^2 > C N^{\frac{-2\zeta}{2\zeta+\gamma}}\right) = 1,$$

for some constant $C > 0$ independent on N, where the infimum in the middle is taken over all algorithms as a map $Z^N \ni \bar{\mathbf{z}} \mapsto f^N \in H$. Alternative minimax lower rates (perhaps considering other quantities, R and σ^2) could be found in (Caponnetto and De Vito, 2007, Theorem 3) and (Blanchard and Mucke, 2016a, Theorem 3.5). Second, distributed SGM saturates when $\zeta > 1$. The reason for this is that averaging over local solutions can only reduce sample and computational variances, not bias. Similar saturation phenomenon is also observed when analyzing distributed KRR in (Zhang et al., 2015; Lin et al., 2017). Third, the condition $2\zeta + \gamma > 1$ is equivalent to assuming that the learning problem can not be too difficult. We believe that such a condition is necessary for applying distributed learning technique to reduce computational costs, as there are no means to reduce computational costs if the learning problem itself is not easy. Fourth, as the learning problem becomes easier (corresponds to a bigger ζ), the faster the convergence rate is, and moreover the larger the number of partition m can be. Finally, different parameter choices leads to different regularization strategies. In the first two regimes, the step-size and the mini-batch size are fixed as some prior constants (which only depends on n), while the number of iterations depends on some unknown distribution parameters. In this case, the regularization parameter is the number of iterations, which in practice can be tuned by using cross-validation methods. Besides, the step-size and the number of iterations in the third regime, or the mini-batch size and the number of iterations in the last regime, depend on the unknown distribution parameters, and they have some regularization effects. The above theorem asserts that distributed SGM with differently suitable choices of parameters can generalize optimally, provided the partition level m is not too large.

3.3 Optimal Rate for Multi-pass SGM on a Single Dataset

As a direct corollary of Theorem 1, we derive the following results for classic multi-pass SGM.

Corollary 3 *Under Assumptions 2 and 3, consider Algorithm 1 with $m = 1$ and any of the following choices on η_t, b and T_*.*
1) $\eta_t \simeq N^{-1}$ for all $t \in [T_*]$, $b = 1$, and $T_* \simeq N^{\alpha+1}$.
2) $\eta_t \simeq N^{-1/2}$ for all $t \in [T_*]$, $b \simeq \sqrt{N}$, and $T_* \simeq N^{\alpha+1/2}$.
3) $\eta_t \simeq N^{-2\zeta\alpha}$ for all $t \in [T_*]$, $b = 1$, and $T_* \simeq N^{\alpha(2\zeta+1)}$.
4) $\eta_t \simeq \frac{1}{\log N}$ for all $t \in [T_*]$, $b \simeq N^{2\zeta\alpha}$, and $T_* \simeq N^{\alpha} \log N$.
Here,

$$\alpha = \frac{1}{(2\zeta + \gamma) \vee 1}.$$

Then,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_{T_*+1} - f_\rho\|_\rho^2 \lesssim \begin{cases} N^{-\frac{2\zeta}{2\zeta+\gamma}} \log N, & \text{if } 2\zeta + \gamma > 1; \\ N^{-2\zeta} \log N, & \text{otherwise.} \end{cases} \tag{17}$$

The above results provide generalization error bounds for multi-pass SGM trained on a single dataset. The derived convergence rate is optimal in the minimax sense (Caponnetto and De Vito, 2007; Blanchard and Mucke, 2016a) up to a logarithmic factor. Note that SGM does not have a saturation effect, and optimal convergence rates can be derived for any $\zeta \in]0, \infty]$. Corollary 3 improves the result in (Lin and Rosasco, 2017b) in two aspects. First, the convergence rates are better than those (i.e., $O(N^{-\frac{2\zeta}{2\zeta+\gamma}} \log N)$ if $2\zeta + \gamma \geq 1$ or $O(N^{-2\zeta} \log^4 N)$ otherwise) from (Lin and Rosasco, 2017b). Second, the above theorem does not require the extra condition $m \geq m_\delta$ made in (Lin and Rosasco, 2017b).

4. Distributed Learning with Spectral Algorithms

In this section, we first state distributed SA. We then present theoretical results for distributed SA, following with simple discussions. Finally, we give convergence results for classic SA.

4.1 Distributed Spectral Algorithms

In this subsection, we present distributed SA. We first recall that a filter function is defined as follows.

Definition 4 (Filter functions) *Let Λ be a subset of $\mathbb{R}_+$. A class of functions $\{\widetilde{G}_\lambda : [0, \kappa^2] \to [0, \infty[, \lambda \in \Lambda\}$ is said to be filter functions with qualification τ ($\tau \geq 0$) if there exist some positive constants $E, F_\tau < \infty$ such that*

$$\sup_{\alpha \in [0,1]} \sup_{\lambda \in \Lambda} \sup_{u \in]0, \kappa^2]} |u^\alpha \widetilde{G}_\lambda(u)| \lambda^{1-\alpha} \leq E, \tag{18}$$

and

$$\sup_{\alpha \in [0,\tau]} \sup_{\lambda \in \Lambda} \sup_{u \in]0, \kappa^2]} |(1 - \widetilde{G}_\lambda(u) u)| u^\alpha \lambda^{-\alpha} \leq F_\tau. \tag{19}$$

In the algorithm, λ is a regularization parameter which should be appropriately chosen in order to achieve best performance. In practice, it can be tuned by using the cross-validation methods. SA is associated with some given filter functions. Different filter functions correspond

Algorithm 2 Distributed learning with spectral algorithms

Input: Number of partitions m, filter function $\widetilde{G}_\lambda$, and kernel function $K(\cdot,\cdot)$

1: Divide $\bar{\mathbf{z}}$ evenly and uniformly at random into m disjoint subsets, $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_m$

2: For every $s \in [m]$, compute a local estimate via SA over the samples $\mathbf{z}_s$: [3]

$$g_\lambda^{\mathbf{z}_s} = \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}_s})\frac{1}{n}\sum_{i=1}^n y_{s,i}K_{s,i}, \quad \mathcal{T}_{\mathbf{x}_s} = \frac{1}{n}\sum_{i=1}^n \langle \cdot, K_{x_{s,i}}\rangle K_{x_{s,i}}$$

3: Take the averaging over these local estimators: $\bar{g}_\lambda^{\bar{\mathbf{z}}} = \frac{1}{m}\sum_{s=1}^m g_\lambda^{\mathbf{z}_s}$

Output: the function $\bar{g}_\lambda^{\bar{\mathbf{z}}}$

to different regularization algorithms. The following examples provide several common filter functions, which leads to different types of regularization methods, see e.g. (Gerfo et al., 2008; Bauer et al., 2007).

Example 1 (KRR) *The choice $\widetilde{G}_\lambda(u) = (u+\lambda)^{-1}$ corresponds to Tikhonov regularization or the regularized least squares algorithm. It is easy to see that $\{G_t(u) : \lambda \in \mathbb{R}_+\}$ is a class of filter functions with qualification $\tau = 1$, and $E = F = 1$.*

Example 2 (GM) *Let $\{\eta_k > 0\}_k$ be such that $\eta_k\kappa^2 \le 1$ for all $k \in \mathbb{N}$. Then as will be shown in Section 6,*

$$\widetilde{G}_\lambda(u) = \sum_{k=1}^t \eta_k \prod_{i=k+1}^t (1-\eta_i u)$$

where we identify $\lambda = (\sum_{k=1}^t \eta_k)^{-1}$, corresponds to gradient methods or Landweber iteration algorithm. The qualification τ could be any positive number, $E = 1$, and $F_\tau = (\tau/\mathrm{e})^\tau$.

Example 3 (Spectral cut-off) *Consider the spectral cut-off or truncated singular value decomposition (TSVD) defined by*

$$\widetilde{G}_\lambda(u) = \begin{cases} u^{-1}, & \text{if } u \ge \lambda, \\ 0, & \text{if } u < \lambda. \end{cases}$$

Then the qualification τ could be any positive number and $E = F_\tau = 1$.

Example 4 (KRR with bias correction) *The function $\widetilde{G}_\lambda(u) = \lambda(\lambda+x)^{-2} + (\lambda+x)^{-1}$ corresponds to KRR with bias correction. It is easy to show that the qualification $\tau = 2$, $E = 2$ and $F_\tau = 1$.*

The implementation of the algorithms is very standard using the representation theorem, for which we thus skip the details.

3. Let L be a self-adjoint, compact operator over a separable Hilbert space. $\widetilde{G}_\lambda(L)$ is an operator on L defined by spectral calculus: suppose that $\{(\sigma_i,\psi_i)\}_i$ is a set of normalized eigenpairs of L with the eigenfunctions $\{\psi_i\}_i$ forming an orthonormal basis of H, then $\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}_s}) = \sum_i \widetilde{G}_\lambda(\sigma_i)\psi_i \otimes \psi_i$.

4.2 Optimal Convergence for Distributed Spectral Algorithms

We have the following general results for distributed SA.

Theorem 2 *Under Assumptions 2 and 3, let $\widetilde{G}_\lambda$ be a filter function with qualification $\tau \geq (\zeta \vee 1)$, and $\bar{g}_\lambda^{\bar{\mathbf{z}}}$ be given by Algorithm 2. Then for any $\tilde{\lambda} = n^{\theta-1}$ with $\theta \in [0,1]$, the following results hold.*
1) For $\zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 \leq (Q_{\gamma,\theta,n}^{2\zeta\vee 1} \vee \frac{\tilde{\lambda}^2}{\lambda^2})[C_5'(R + \mathbf{1}_{\{2\zeta<1\}}\|f_\rho\|_\infty)^2 \lambda^{2\zeta} + C_8' \frac{\sigma^2}{N\tilde{\lambda}^\gamma}]. \tag{20}$$

2) For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 \leq (\frac{\lambda^{1-2\zeta}}{n^{(\zeta-1/2)\wedge 1}} \vee Q_{\gamma,\theta,n} \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda^{2\zeta}})[C_6' R^2 \lambda^{2\zeta} + C_8' \frac{\sigma^2}{N\tilde{\lambda}^\gamma}]. \tag{21}$$

Here, $Q_{\gamma,\theta,n}$ is given by (14)*, and C_5', C_6' and C_8' are positive constants depending only κ, ζ, E, F_τ, c_γ and $\|\mathcal{L}\|$.*

The above results provide generalization error bounds for distributed SA. The upper bound depends on the number of partition m, the regularization parameter λ and total sample size N. When the regularization parameter $\lambda > 1/n$, by setting $\tilde{\lambda} = \lambda$, the derived error bounds for $\zeta \leq 1$ can be simplified as

$$\mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 \lesssim \lambda^{2\zeta} + \frac{1}{N\lambda^\gamma}.$$

There are two terms in the upper bound. They are raised from estimating bias and sample variance. Note that there is a trade-off between the bias term and the sample variance term. Solving this trade-off leads to the best choice on regularization parameter. Note also that similar to that for distributed SGM, distributed SA also saturates when $\zeta > 1$.

Corollary 5 *Under the assumptions of Theorem 2, let $2\zeta + \gamma > 1$, $\lambda = N^{-\frac{1}{2\zeta+\gamma}}$ and the number of partitions satisfies (16). Then*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 \lesssim N^{-\frac{2\zeta}{2\zeta+\gamma}}. \tag{22}$$

The convergence rate from the above corollary is optimal as it matches exactly the minimax rate in (Caponnetto and De Vito, 2007), and it is better than the rate for distributed SGM from Theorem 1, where the latter has an extra logarithmic factor. According to Corollary 5, distributed SA with an appropriate choice of regularization parameter λ can generalize optimally, if the number of partitions is not too large. To the best of our knowledge, the above corollary is the first optimal statistical result for distributed SA considering the non-attainable case (i.e. ζ can be less than 1/2). Moreover, the requirement on the number of partitions $m < N^{\frac{2\zeta+\gamma-1}{2\zeta+\gamma}}$ to achieve optimal generalization error bounds is much weaker than that ($m \leq N^{\frac{2\zeta-1}{2\zeta+\gamma}}$) in (Guo et al., 2017; Blanchard and Mucke, 2016b).

4.3 Optimal Rates for Spectral Algorithms on a Single Dataset

The following results provide generalization error bounds for classic SA.

Corollary 6 *Under Assumptions 2 and 3, let $\widetilde{G}_\lambda$ be a filter function with qualification $\tau \geq (\zeta \vee 1)$, and $g_\lambda^{\mathbf{z}_1}$ be given by Algorithm 2 with $\lambda = N^{-\frac{1}{1\vee(2\zeta+\gamma)}}$ and $m = 1$. Then*

$$\mathbb{E}\|\mathcal{S}_\rho g_\lambda^{\mathbf{z}_1} - f_\rho\|_\rho^2 \lesssim \begin{cases} N^{-\frac{2\zeta}{2\zeta+\gamma}}, & \text{if } 2\zeta+\gamma > 1; \\ N^{-2\zeta}(1 \vee \log N^\gamma), & \text{otherwise.} \end{cases} \tag{23}$$

Here, $a_1 \lesssim a_2$ means $a_1 \leq Ca_2$ for some positive constant C which is depending only on $\kappa, c_\gamma, \zeta, M, \sigma, \|\mathcal{L}\|, E,$ and F_τ.

The above results assert that SA generalizes optimally if the regularization parameter is well chosen. To the best of our knowledge, the derived result is the first one with optimally capacity-dependent rates in the non-attainable case for a general SA. Note that unlike distributed SA, classic SA does not have a saturation effect.

5. Discussion

In this section, we briefly review some of the related results in order to facilitate comparisons. For ease of comparisons, we summarize some of the results and their computational costs in Table 1.

We first briefly review convergence results on generalization error for KRR, and more generally, SA. Statistical results for KRR with different convergence rates have been shown in, e.g., (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Wu et al., 2006; Steinwart and Christmann, 2008; Steinwart et al., 2009). Particularly, Smale and Zhou (2007) proved convergence rates of order $O(N^{-\frac{2\zeta}{1+(2\zeta\vee 1)}})$ with $0 < \zeta \leq 1$, without considering the capacity assumption. Caponnetto and De Vito (2007) gave optimally capacity-dependent convergence rate of order $O(N^{-\frac{2\zeta}{2\zeta+\gamma}})$ but only for the case that $1/2 \leq \zeta \leq 1$. The above two are based on integral operator approaches. Using an alternative argument related to covering-number or entropy-numbers, Wu et al. (2006) provided convergence rate $O(n^{-\frac{2\zeta}{1+\gamma}})$, and (Steinwart and Christmann, 2008, Theorem 7.23) provides convergence rate $O(n^{-\frac{2\zeta}{(2\zeta+\gamma)\vee 1}})$, assuming that $0 < \zeta \leq 1/2$, $\gamma \in (0,1)$ and $|y| \lesssim 1$ almost surely. For GM, Yao et al. (2007) derived convergence rate of order $O(N^{-\frac{2\zeta}{2\zeta+2}})$ (for $\zeta \in]0,\infty[$), without considering the capacity assumption. Involving the capacity assumption, Lin and Rosasco (2017b) derived convergence rate of order $O(N^{\frac{-2\zeta}{2\zeta+\gamma}} \log^2 N)$ if $2\zeta+\gamma > 1$, or $O(N^{-2\zeta} \log^4 N)$ if $2\zeta+\gamma \leq 1$. Note that both proofs from (Yao et al., 2007; Lin and Rosasco, 2017b) rely on the special separable properties of GM with the square loss. For SA, statistical results on generalization error with different convergence rates have been shown in, e.g., (Caponnetto, 2006; Bauer et al., 2007; Blanchard and Mucke, 2016a; Dicker et al., 2017; Lin et al., 2017). The best convergence rate shown so far (without making any extra unlabeled data as that in (Caponnetto, 2006)) is $O(N^{-\frac{2\zeta}{2\zeta+\gamma}})$ (Blanchard and Mucke, 2016a; Dicker et al., 2017; Lin et al., 2017) but only for the attainable case, i.e., $\zeta \geq 1/2$. These results also apply to GM, as GM can be viewed as a special instance of SA. Note that some of these results also require the extra assumption that the sample size N is large enough. In comparisons, Corollary 6 provides the best convergence rates for SA, considering both the non-attainable and attainable cases and without making any extra assumption. Note that our derived error bounds are in expectation,

Algorithm	Ass.	# Processors m	Rate	Local Memory & Time	Memory & Time
KRR (Smale and Zhou, 2007)	$\zeta \in]0,1]$, $\gamma = 1$	1	$N^{-\frac{2\zeta}{(2\zeta \vee 1)+1}}$	$\times$	N^2 & N^3
KRR (Caponnetto and De Vito, 2007)	$\zeta \in [\frac{1}{2},1]$, $\gamma \in]0,1]$, $N \geq N_\delta$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$\times$	-
KRR (Steinwart and Christmann, 2008)[4]	$\zeta \in [0,\frac{1}{2}]$, $\gamma \in]0,1[$, $\|y\| \lesssim 1$	1	$N^{-\frac{(2\zeta}{2\zeta+\gamma)\vee 1}}$	$\times$	-
KRR [Corollary 6]	$\boldsymbol{\zeta \in]0,1], 2\zeta+\gamma > 1}$	1	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$\times$	-
KRR [Corollary 6]	$\boldsymbol{2\zeta+\gamma \leq 1}$	1	$\boldsymbol{N^{-2\zeta} \log N^{\gamma}}$	$\times$	-
GM (Yao et al., 2007)	$\gamma = 1$	1	$N^{-\frac{2\zeta}{2\zeta+2}}$	$\times$	N & $N^2 N^{\frac{1}{2\zeta+2}}$
GM (Dicker et al., 2017)	$\zeta \in [\frac{1}{2},\infty[$, $\gamma \in]0,1]$, $N \geq N_0$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$\times$	N & $N^2 N^{\frac{1}{2\zeta+\gamma}}$
GM (Lin and Rosasco, 2017b)	$2\zeta+\gamma > 1$, $N \geq N_\delta$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}} \log^2 N$	$\times$	N & $N^2 N^{\frac{1}{2\zeta+\gamma}}$
GM (Lin and Rosasco, 2017b)	$2\zeta+\gamma \leq 1$, $N \geq N_\delta$	1	$N^{-2\zeta} \log^4 N$	$\times$	N & N^3
GM [Corollary 6]	$\boldsymbol{2\zeta+\gamma > 1}$	1	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$\times$	N & $N^2 N^{\frac{1}{2\zeta+\gamma}}$
GM [Corollary 6]	$\boldsymbol{2\zeta+\gamma \leq 1}$	1	$\boldsymbol{N^{-2\zeta} \log N^{\gamma}}$	$\times$	N & N^3
SA (Guo et al., 2017)	$\zeta \in [\frac{1}{2},\tau], \gamma \in]0,1]$	1	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$\times$	$-$
SA [Corollary 6]	$\zeta \leq \tau$, $\boldsymbol{2\zeta+\gamma > 1}$	1	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$\times$	$-$
SA [Corollary 6]	$\zeta \leq \tau, \boldsymbol{2\zeta+\gamma \leq 1}$	1	$\boldsymbol{N^{-2\zeta} \log N^{\gamma}}$	$\times$	$-$
OL (Ying and Pontil, 2008)	$\gamma = 1$	1	$N^{-\frac{2\zeta}{2\zeta+1}} \log N$	$\times$	N & N^2
AveOL (Dieuleveut and Bach, 2016)	$\zeta \in]0,1]$, $2\zeta+\gamma > 1$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$\times$	N & N^2
AveOL (Dieuleveut and Bach, 2016)	$\boldsymbol{2\zeta+\gamma \leq 1}$	1	$\boldsymbol{N^{-2\zeta}}$	$\times$	N & N^2
SGM (Lin and Rosasco, 2017b)	$2\zeta+\gamma > 1, N \geq N_\delta$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}} \log^2 N$	$\times$	N & $N^2 N^{\frac{1-\gamma}{2\zeta+\gamma}}$
SGM (Lin and Rosasco, 2017b)	$2\zeta+\gamma \leq 1, N \geq N_\delta$	1	$N^{-2\zeta} \log^4 N$	$\times$	N & $N^{3-\gamma}$
SGM [Corollary 3]	$\boldsymbol{2\zeta+\gamma > 1}$	1	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$\times$	N & $N^2 N^{\frac{1-\gamma}{2\zeta+\gamma}}$
SGM [Corollary 3]	$\boldsymbol{2\zeta+\gamma \leq 1}$	1	$\boldsymbol{N^{-2\zeta} \log N^{\gamma}}$	$\times$	N & $N^{3-\gamma}$
NyKRR (Rudi et al., 2015)	$\zeta \in [\frac{1}{2},1]$, $\gamma \in]0,1]$, $N \geq N_\delta$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$\times$	$N^{\frac{2\zeta+\gamma+1}{2\zeta+\gamma}}$ & $N^{\frac{2\zeta+2+\gamma}{2\zeta+\gamma}}$
NySGM (Lin and Rosasco, 2017a)	$\zeta \in [\frac{1}{2},1]$, $\gamma \in]0,1]$, $N \geq N_\delta$	1	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$\times$	$N^{\frac{2}{2\zeta+\gamma}\vee 1}$ & $N^{\frac{2\zeta+2}{2\zeta+\gamma}}$
DKRR & DSA (Guo et al., 2017)	$\zeta \in [\frac{1}{2},1]$, $\gamma \in]0,1]$	$N^{\frac{2\zeta-1}{2\zeta+\gamma}}$	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$	$N^{\frac{2(1+\gamma)}{2\zeta+\gamma}}$ & $N^{\frac{3(1+\gamma)}{2\zeta+\gamma}}$	$N^{\frac{2\zeta+2\gamma+1}{2\zeta+\gamma}}$ & $N^{\frac{2\zeta+2+3\gamma}{2\zeta+\gamma}}$
DKRR & DSA [Corollary 5]	$\boldsymbol{\zeta \in]0,1], 2\zeta+\gamma > 1}$	$N^{\frac{2\zeta+\gamma-1}{2\zeta+\gamma}}$	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$N^{\frac{2}{2\zeta+\gamma}}$ & $N^{\frac{3}{2\zeta+\gamma}}$	$N^{\frac{2\zeta+\gamma+1}{2\zeta+\gamma}}$ & $N^{\frac{2\zeta+2+\gamma}{2\zeta+\gamma}}$
DSGM [Corollary 2.(3)]	$\boldsymbol{\zeta \in]0,1], 2\zeta+\gamma > 1}$	$N^{\frac{2\zeta+\gamma-1}{2\zeta+\gamma}}$	$\boldsymbol{N^{-\frac{2\zeta}{2\zeta+\gamma}}}$	$N^{\frac{1}{2\zeta+\gamma}}$ & $N^{\frac{2}{2\zeta+\gamma}}$	$\boldsymbol{N}$ & $\boldsymbol{N^{\frac{2\zeta+\gamma+1}{2\zeta+\gamma}}}$

Table 1

Summary of assumptions and results for distributed SGM (DSGM) and related approaches including KRR, GM, SA, one-pass SGM (OL), one-pass SGM with averaging (AveOL), SGM, Nyström KRR (NyKRR), Nyström SGM (NySGM), distributed KRR (DKRR) and distributed SA (DSA).

but it is not difficult to derive error bounds in high probability using our approach, and we will report this result in a future work.

We next briefly review convergence results for SGM. SGM (Robbins and Monro, 1951) has been widely used in convex optimization and machine learning, see e.g. (Cesa-Bianchi et al., 2004; Nemirovski et al., 2009; Bottou et al., 2016) and references therein . In what follows, we will briefly recall some recent works on generalization error for nonparametric regression on a RKHS considering the square loss. We will use the term "online learning algorithm" (OL) to mean one-pass SGM, i.e, SGM that each sample can be used only once. Different variants of OL, either with or without regularization, have been studied. Most of them take the form

$$f_{t+1} = (1-\lambda_t) f_t - \eta_t (f_t(x_t) - y_t) K_{x_t}, t = 1 \cdots, N.$$

Here, the regularization parameter λ_t could be zero (Zhang, 2004; Ying and Pontil, 2008), or a positive (Smale and Yao, 2006; Ying and Pontil, 2008) and possibly time-varying constant (Tarres and Yao, 2014). Particularly, Tarres and Yao (2014) studied OL with time-varying

4. The results from (Steinwart and Christmann, 2008) are based on entropy-numbers arguments while the other results summarized for KRR in the table are based on integral-operator arguments.

regularization parameters and convergence rate of order $O(N^{\frac{-2\zeta}{2\zeta+1}})$ ($\zeta \in [\frac{1}{2}, 1]$) in high probability was proved. Ying and Pontil (2008) studied OL without regularization and convergence rate of order $O(N^{-\frac{2\zeta}{2\zeta+1}})$ in expectation was shown. Both convergence rates from (Ying and Pontil, 2008; Tarres and Yao, 2014) are capacity-independently optimal and they do not take the capacity assumption into account. Considering an averaging step (Polyak and Juditsky, 1992) and a proof technique motivated by (Bach and Moulines, 2013), Dieuleveut and Bach (2016) proved capacity-dependently optimal rate $O(N^{-\frac{2\zeta}{(2\zeta+\gamma)\vee 1}})$ for OL in the case that $\zeta \leq 1$. Recently, Lin and Rosasco (2017b) studied (multi-pass) SGM, i.e, Algorithm 1 with $m = 1$. They showed that SGM with suitable parameter choices, achieves convergence rate of order $O(N^{-\frac{2\alpha}{(2\alpha+\gamma)\vee 1}} \log^{\beta} N)$ with $\beta = 2$ when $2\alpha+\gamma > 1$ or $\beta = 4$ otherwise, after some number of iterations. In comparisons, the derived results for SGM in Corollary 3 are better than those from (Lin and Rosasco, 2017b), and the convergence rates are the same as those from (Dieuleveut and Bach, 2016) for averaging OL when $\zeta \leq 1$ and $2\zeta+\gamma > 1$. For the case $2\zeta+\gamma \leq 1$, the convergence rate $O(N^{-2\zeta}(1 \vee \log N^{\gamma}))$ for SGM in Corollary 3 is worser than $O(N^{-2\zeta})$ in (Dieuleveut and Bach, 2016) for averaging OL. However, averaging OL saturates for $\zeta > 1$, while SGM does not.

To meet the challenge of large-scale learning, a line of research focus on designing learning algorithms with Nyström subsampling, or more generally sketching. Interestingly, the latter has also been applied to compressed sensing, low rank matrix recovery and kernel methods, see e.g. (Candès et al., 2006; Yurtsever et al., 2017; Yang et al., 2012) and references therein. The basic idea of Nyström subsampling is to replace a standard large matrix with a smaller matrix obtained by subsampling (Smola and Schölkopf, 2000; Williams and Seeger, 2000). For kernel methods, Nyström subsampling has been successfully combined with KRR (Alaoui and Mahoney, 2015; Rudi et al., 2015; Yang et al., 2017) and SGM (Lu et al., 2016; Lin and Rosasco, 2017a). Generalization error bounds of order $O(N^{\frac{-2\zeta}{2\zeta+\gamma}})$ (Rudi et al., 2015; Lin and Rosasco, 2017a) were derived, provided that the subsampling level is suitably chosen, considering the case $\zeta \in [\frac{1}{2}, 1]$. Computational advantages of these algorithms were highlighted. Here, we summarize their convergence rates and computational costs in Table 1, from which we see that distributed SGM has advantages on both memory and time.

Another line of research for large-scale learning focus on distributed (parallelizing) learning. Distributed learning, based on a divide-and-conquer approach, has been used for, e.g., perceptron-based algorithms (Mcdonald et al., 2009), parametric smooth convex optimization problems (Zhang et al., 2012), and sparse regression (Lee et al., 2017). Recently, this approach has been successfully applied to learning algorithms with kernel methods, such as KRR (Zhang et al., 2015), and SA (Guo et al., 2017; Blanchard and Mucke, 2016a). Zhang et al. (2015) first studied distributed KRR and showed that distributed KRR retains optimal rates $O(N^{-\frac{2\zeta}{2\zeta+\gamma}})$ (for $\zeta \in [\frac{1}{2}, 1]$) provided the partition level is not too large. The number of partition to retain optimal rate shown in (Zhang et al., 2015) for distributed KRR depends on some conditions which may be less well understood and thus potentially leads to a suboptimal partition number. Lin et al. (2017) provided an alternative and refined analysis for distributed KRR, leading to a less strict condition on the partition number. Guo et al. (2017) extended the analysis to distributed SA, an proved optimal convergence rate for the case $\zeta \geq 1/2$, if the number of partitions $m \leq N^{\frac{2\zeta-1}{2\zeta+\gamma}}$. In comparison, the condition on partition number from Theorem 2 for distributed SA is less strict. Moreover, Theorem 2 shows that distributed SA can retain optimal rate even in the non-attainable case. According to Corollary 2, distributed SGM with appropriate choices of parameters can achieve optimal rate if the partition number is not too large. In comparison of the derived results for distributed SA with those for distributed SGM, we see from Table 1 that the latter has advantages on both memory and time. The most related to our works are (Zinkevich et al., 2010; Jain et al., 2016). Zinkevich et al. (2010) studied distributed OL for optimization problems over a finite-dimensional domain, and proved convergence results assuming that the objective function is strongly convex. Jain et al. (2016) considered distributed OL with averaging for least square regression problems over a finite-dimension space and proved certain convergence results that may depend on the smallest eigenvalue of the covariance matrix. These results do not apply to our cases, as we consider distributed multi-pass SGM for nonparametric regression over a RKHS and our objective function is not strongly convex. We finally remark that using a partition approach (Thomann et al., 2016; Tandon et al., 2016), one can also scale up the kernel methods, with a computational advantage similar as those of using distributed learning technique.

We conclude this section with some further questions. First, in this paper, we assume that all parameter choices are given priorly. In practice, these parameters can be possibly tuned by cross-validation method. Second, the derived rate for SGM and SA when $2\zeta + \gamma \leq 1$ is $O(N^{-2\zeta}(1 \vee \log N^{\gamma}))$, which is worser than $O(N^{-2\zeta})$ of averaging OL (Dieuleveut and Bach, 2016). It would be interesting to improve the rate, or to derive a minimax rate for the case $2\zeta + \gamma \leq 1$. Third, all results stated in this paper are in expectation, and it would be interesting to derive high-probability results (possibly by a proof technique from (London, 2017)).

6. Proofs for Distributed SGM

In this section, we provide the proofs of our main theorems for distributed SGM. We begin with some basic notations. For ease of readability, we also make a list of notations in the appendix.

6.1 Notations

$\mathbb{E}[\xi]$ denotes the expectation of a random variable ξ. $\|\cdot\|_\infty$ denotes the supreme norm with respect to ρ_X. For a given bounded operator $L : H' \to H''$, $\|L\|$ denotes the operator norm of L, i.e., $\|L\| = \sup_{f\in H', \|f\|_{H'}=1} \|Lf\|_{H''}$. Here H' and H'' are two separable Hilbert spaces (which could be the same).

We introduce the inclusion operator $\mathcal{S}_\rho : H \to L^2_{\rho_X}$, which is continuous under Assumption 1. Furthermore, we consider the adjoint operator $\mathcal{S}_\rho^* : L^2_{\rho_X} \to H$, the covariance operator $\mathcal{T} : H \to H$ given by $\mathcal{T} = \mathcal{S}_\rho^*\mathcal{S}_\rho$, and the operator $\mathcal{L} : L^2_{\rho_X} \to L^2_{\rho_X}$ given by $\mathcal{S}_\rho\mathcal{S}_\rho^*$. It can be easily proved that $\mathcal{S}_\rho^* f = \int_X K_x f(x) d\rho_X(x)$ and $\mathcal{T} = \int_X \langle \cdot, K_x\rangle_H K_x d\rho_X(x)$. The operators $\mathcal{T}$ and $\mathcal{L}$ can be proved to be positive trace class operators (and hence compact). In fact, by (7),

$$\|\mathcal{L}\| = \|\mathcal{T}\| \leq \mathrm{tr}(\mathcal{T}) = \int_X \mathrm{tr}(K_x \otimes K_x) d\rho_X(x) = \int_X \|K_x\|_H^2 d\rho_X(x) \leq \kappa^2. \tag{24}$$

For any function $f \in H$, the H-norm can be related to the $L^2_{\rho_X}$-norm by $\sqrt{\mathcal{T}}$: (Bauer et al., 2007)

$$\|\mathcal{S}_\rho f\|_\rho = \left\|\sqrt{\mathcal{T}} f\right\|_H, \tag{25}$$

and furthermore according to the singular value decomposition of $\mathcal{S}_\rho$,

$$\|\mathcal{L}^{-\frac{1}{2}}\mathcal{S}_\rho f\|_\rho \leq \|f\|_H. \tag{26}$$

We define the sampling operator (with respect to any given set $\mathbf{x} \subseteq X$ of cardinality n) $\mathcal{S}_{\mathbf{x}} : H \to \mathbb{R}^n$ by $(\mathcal{S}_{\mathbf{x}} f)_i = f(x_i) = \langle f, K_{x_i}\rangle_H$, $i \in [n]$, where the norm $\|\cdot\|_{\mathbb{R}^n}$ is the standard Euclidean norm times $1/\sqrt{n}$. Its adjoint operator $\mathcal{S}_{\mathbf{x}}^*: \mathbb{R}^n \to H$, defined by $\langle \mathcal{S}_{\mathbf{x}}^* \mathbf{y}, f\rangle_H = \langle \mathbf{y}, \mathcal{S}_{\mathbf{x}} f\rangle_{\mathbb{R}^n}$ for $\mathbf{y} \in \mathbb{R}^n$ is thus given by

$$\mathcal{S}_{\mathbf{x}}^* \mathbf{y} = \frac{1}{n}\sum_{i=1}^n y_i K_{x_i}. \tag{27}$$

Moreover, we can define the empirical covariance operator (with respect to $\mathbf{x}$) $\mathcal{T}_{\mathbf{x}}: H \to H$ such that $\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^* \mathcal{S}_{\mathbf{x}}$. Obviously,

$$\mathcal{T}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \langle \cdot, K_{x_i}\rangle_H K_{x_i}.$$

By (7), similar to (24), we have

$$\|\mathcal{T}_{\mathbf{x}}\| \le \operatorname{tr}(\mathcal{T}_{\mathbf{x}}) \le \kappa^2. \tag{28}$$

For any $\tilde{\lambda} > 0$, for notational simplicity, we let $\mathcal{T}_{\tilde{\lambda}} = \mathcal{T} + \tilde{\lambda}$, $\mathcal{T}_{\mathbf{x}\tilde{\lambda}} = \mathcal{T}_{\mathbf{x}} + \tilde{\lambda}$, and

$$\mathcal{N}(\tilde{\lambda}) = \operatorname{tr}(\mathcal{L}(\mathcal{L} + \tilde{\lambda})^{-1}) = \operatorname{tr}(\mathcal{T}(\mathcal{T} + \tilde{\lambda})^{-1}).$$

For any $f \in H$ and $x \in X$, the following well known reproducing property holds:

$$\langle f, K_x\rangle_H = f(x). \tag{29}$$

and following from the above, Cauchy-Schwarz inequality and (7), one can prove that

$$|f(x)| = |\langle f, K_x\rangle_H| \le \|f\|_H \|K_x\|_H \le \kappa \|f\|_H \tag{30}$$

For any $s \in [m]$, we denote the set of random variables $\{j_{s,i}\}_{b(t-1)+1 \le i \le bt}$ by $\mathbf{J}_{s,t}$, $\{j_{s,1}, j_{s,2}, \cdots, j_{s,bT}\}$ by $\mathbf{J}_s$, and $\{\mathbf{J}_1, \cdots, \mathbf{J}_m\}$ by $\mathbf{J}$. Note that $j_{s,1}, j_{s,2}, \cdots, j_{s,bT}$ are conditionally independent given $\mathbf{z}_s$.

6.2 Error Decomposition

The key to our proof is an error decomposition. To introduce the error decomposition, we need to introduce two auxiliary sequences.

The first auxiliary sequence is generated by distributed GM. For any $s \in [m]$, the GM over the sample set $\mathbf{z}_s$ is defined by $g_{s,1} = 0$ and

$$g_{s,t+1} = g_{s,t} - \eta_t \left(\mathcal{T}_{\mathbf{x}_s} g_{s,t} - \mathcal{S}_{\mathbf{x}_s}^* \mathbf{y}_s\right), \qquad t = 1, \ldots, T, \tag{31}$$

where $\{\eta_t > 0\}$ is a step-size sequence given by Algorithm 1. The average estimator over these local estimators is given by

$$\bar{g}_t = \frac{1}{m}\sum_{s=1}^m g_{s,t}. \tag{32}$$

The second auxiliary sequence is generated by distributed pseudo GM as follows. For any $s \in [m]$, the pseudo GM over the input set $\mathbf{x}_s$ is defined by $h_{s,1} = 0$ and

$$h_{s,t+1} = h_{s,t} - \eta_t \left(\mathcal{T}_{\mathbf{x}_s} h_{s,t} - \mathcal{L}_{\mathbf{x}_s} f_\rho\right), \qquad t = 1, \ldots, T. \tag{33}$$

The average estimator over these local estimators is given by

$$\bar{h}_t = \frac{1}{m}\sum_{s=1}^m h_{s,t}. \tag{34}$$

In the above, for any given inputs set $\mathbf{x} \subseteq X^{|\mathbf{x}|}$, $\mathcal{L}_{\mathbf{x}} : L^2_{\rho_X} \to H$ is defined as that for any $f \in L^2_{\rho_X}$ such that $\|f\|_\infty < \infty$,

$$\mathcal{L}_{\mathbf{x}} f = \frac{1}{|\mathbf{x}|} \sum_{x \in \mathbf{x}} f(x) K_x. \tag{35}$$

Note that (33) can not be implemented in practice, as $f_\rho(x)$ is unknown in general.

We state the error decomposition as follows.

Proposition 1 *We have that for any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{f}_t - f_\rho\|_\rho^2 = \mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2 + \mathbb{E}[\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2] + \mathbb{E}\|\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t)\|_\rho^2. \tag{36}$$

The error decomposition is similar as (but a bit different from) (Lin and Rosasco, 2017b, Proposition 1) for classic multi-pass SGM. There are three terms in the right-hand side of (36). The first term depends on the regularity of the regression function (Assumption 2) and it is called as *bias*. The second term depends on the noise level σ^2 from (8) and it is called as *sample variance*. The last term is caused by the random estimates of the full gradients and it is called as *computational variance*. In the following subsections, we will estimate these three terms separately. Total error bounds can be thus derived by substituting these estimates into the error decomposition.

The proof idea is quite simple. According to Lemmas 7, 21 and 22, in order to proceed the analysis, we only need to estimate bias, sample and computational variance of a local estimator. In order to estimate local bias and local sample variance, as given in Lemma 8, we rewrite $g_{s,t}$ and $h_{s,t}$ as the special forms induced by a filter function G_t of GM. The strategy here for estimating local bias and sample variance is different from that in (Lin and Rosasco, 2017b) which relies on the following error decomposition and iterative relationship motivated by (Lin and Zhou, 2015):

$$\|g_{s,t} - f_\rho\|_\rho \le \|r_t - f_\rho\|_\rho + \|g_{s,t} - r_t\|_\rho \quad and$$

$$g_{s,t+1} - r_{t+1} = \sum_{k=1}^t \eta_k \prod_{j=k+1}^t (I - \eta_j \mathcal{T}_{\mathbf{x}_s})(\mathcal{T} r_k - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}} r_k + \mathcal{S}_{\mathbf{x}_s}^* \mathbf{y}_s),$$

where the population sequence $\{r_t\}_t$ is defined by $r_1 = 0$ and

$$r_{t+1} = (I - \mathcal{T}) r_t + \mathcal{S}_\rho^* f_\rho. \tag{37}$$

Instead, in this paper, we use spectral theory from functional analysis to proceed the estimations. Our main novelties lies in a new error bound for $\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}} \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|$, see Lemma 19 (which allows one to derive optimal rates in the non-attainable cases without requiring the sample size is large enough, and refines the error bounds on $\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}} \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|$ by involving Assumption 3 in the logarithmic factor), and the estimation on local bias, as well as some other refined analysis. For estimating local bias, we introduce a new key error decomposition, in order to cover both the non-attainable case and the unbounded-output case.

All the missing proofs of propositions and lemmas in this section can be found in Appendix B.

6.3 Estimating Bias

In this subsection, we estimate bias, i.e., $\mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2$. We first give the following lemma, which asserts that the bias term can be estimated in terms of the bias of a local estimator.

Lemma 7 *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_t - f_\rho\|_\rho^2 \le \mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2.$$

To estimate the bias of the local estimator, $\mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2$, we next introduce some preliminary notations and lemmas.

$\Pi_{t+1}^T(L) = \prod_{k=t+1}^T (I - \eta_k L)$ for $t \in [T-1]$ and $\Pi_{T+1}^T(L) = I$, for any operator $L : H \to H$, where H is a Hilbert space and I denotes the identity operator on H. Let $k, t \in \mathbb{N}$. We use the following conventional notations: $1/0 = +\infty$, $\prod_k^t = 1$ and $\sum_k^t = 0$ whenever $k > t$. $\Sigma_k^t = \sum_{i=k}^t \eta_i$, $\lambda_{k:t} = (\Sigma_k^t)^{-1}$, and specially $\lambda_{1:t}$ is abbreviated as λ_t. Define the function $G_t : \mathbb{R} \to \mathbb{R}$ by

$$G_t(u) = \sum_{k=1}^t \eta_k \prod_{k=t+1}^T (I - \eta_k u). \tag{38}$$

Throughout this paper, we assume that the step-size sequence satisfies $\eta_t \in]0, \kappa^{-2}]$ for all $t \in \mathbb{N}$. Thus, $G_t(u)$ and $\Pi_k^t(u)$ are non-negative on $]0, \kappa^2]$. For notational simplicity, throughout the rest of this subsection, we will drop the index $s = 1$ for the first local estimator whenever it shows up, i.e, we abbreviate $h_{1,t}$ as h_t, $\mathbf{z}_1$ as $\mathbf{z}$, and $\mathcal{T}_{\mathbf{x}_1}$ as $\mathcal{T}_{\mathbf{x}}$, etc.

The key idea for our estimation on bias is that $\{h_t\}_t$ can be well approximated by the population sequence $\{r_t\}_t$, a deterministic sequence depending on the regression function f_ρ.

We first have the following observations.

Lemma 8 *The sequence $\{r_t\}_t$ defined by (37) can be rewritten as*

$$r_{t+1} = G_t(\mathcal{T})\mathcal{S}_\rho^* f_\rho. \tag{39}$$

Similarly, for any $s \in [m]$, the sequences $\{g_{s,t}\}_t$ and $\{h_{s,t}\}_t$ defined by (31) and (33) can be rewritten as

$$g_{s,t+1} = G_t(\mathcal{T}_{\mathbf{x}_s})\mathcal{S}_{\mathbf{x}_s}^* \mathbf{y}_s,$$

and

$$h_{s,t+1} = G_t(\mathcal{T}_{\mathbf{x}_s})\mathcal{L}_{\mathbf{x}_s} f_\rho.$$

Proof Using the relationship (37) iteratively, introducing with $r_1 = 0$, one can prove the first conclusion. ■

According to the above lemma, we know that GM can be rewritten as a form of SA with filter function $\widetilde{G}_\lambda(\cdot) = G_t(\cdot)$. In the next lemma, we further develop some basic properties for this filter function.

Lemma 9 *For all $u \in [0, \kappa^2]$,*
1) $u^\alpha G_t(u) \le \lambda_t^{\alpha-1}$, $\forall \alpha \in [0,1]$.
2) $(1 - uG_t(u))u^\alpha = \Pi_1^t(u)u^\alpha \le (\alpha/\mathrm{e})^\alpha \lambda_t^\alpha$, $\quad \forall \alpha \in [0, \infty[$.
3) $\Pi_k^t(u)u^\alpha \le (\alpha/\mathrm{e})^\alpha \lambda_{k:t}^\alpha$, $\quad \forall t, k \in \mathbb{N}$.

According to Lemma 9, $G_t(\cdot)$ is a filter function indexed with regularization parameter $\lambda = \lambda_t$, and the qualification τ can be any positive number, and $E = 1$, $F_\tau = (\tau/\mathrm{e})^\tau$. Using Lemma 9 and the spectral theorem, one can get the following results.

Lemma 10 *Let L be a compact, positive operator on a separable Hilbert space H such that $\|L\| \le \kappa^2$. Then for any $\tilde{\lambda} \ge 0$,*
1) $\|(L+\tilde{\lambda})^\alpha G_t(L)\| \le \lambda_t^{\alpha-1}(1 + (\tilde{\lambda}/\lambda_t)^\alpha), \ \forall \alpha \in [0,1]$.
2) $\|(I - LG_t(L))(L+\tilde{\lambda})^\alpha\| = \|\Pi_1^t(L)(L+\tilde{\lambda})^\alpha\| \le 2^{(\alpha-1)_+}((\alpha/\mathrm{e})^\alpha + (\tilde{\lambda}/\lambda_t)^\alpha)\lambda_t^\alpha, \quad \forall \alpha \in [0,\infty[$.
3) $\|\Pi_{k+1}^t(L)L^\alpha\| \le (\alpha/\mathrm{e})^\alpha \lambda_{k:t}^\alpha, \ \forall k,t \in \mathbb{N}$.

To proceed the proof, we introduce the following basic lemmas on operators.

Lemma 11 *(Fujii et al., 1993, Cordes inequality) Let A and B be two positive bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \le \|AB\|^s, \quad \textit{when } 0 \le s \le 1.$$

Lemma 12 *Let H_1, H_2 be two separable Hilbert spaces and $\mathcal{S} : H_1 \to H_2$ a compact operator. Then for any function $f : [0, \|\mathcal{S}\|] \to [0,\infty[$,*

$$f(\mathcal{S}\mathcal{S}^*)\mathcal{S} = \mathcal{S}f(\mathcal{S}^*\mathcal{S}).$$

Proof The result can be proved using singular value decomposition of a compact operator. ■

Lemma 13 *Let A and B be two non-negative bounded linear operators on a separable Hilbert space with $\max(\|A\|, \|B\|) \le \kappa^2$ for some non-negative κ^2. Then for any $\zeta > 0$,*

$$\|A^\zeta - B^\zeta\| \le C_{\zeta,\kappa}\|A - B\|^{\zeta \wedge 1}, \tag{40}$$

where

$$C_{\zeta,\kappa} = \begin{cases} 1 & \textit{when } \zeta \le 1, \\ 2\zeta\kappa^{2\zeta-2} & \textit{when } \zeta > 1. \end{cases} \tag{41}$$

Proof The proof is based on the fact that u^ζ is operator monotone if $0 < \zeta \le 1$. While for $\zeta \ge 1$, the proof can be found in, e.g., (Dicker et al., 2017). ■

Using Lemma 10, one can prove the following results, which give some basic properties for the population sequence $\{r_t\}_t$.

Lemma 14 *Let $a \in \mathbb{R}$. Under Assumption 2, the following results hold.*
1) For any $a \le \zeta$, we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho \le ((\zeta - a)/\mathrm{e})^{\zeta-a} R\lambda_t^{\zeta-a}.$$

2) We have

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H \le R \cdot \begin{cases} \lambda_t^{\zeta+a-1}, & \textit{if } -\zeta \le a \le 1-\zeta, \\ \kappa^{2(\zeta+a-1)}, & \textit{if } a \ge 1-\zeta. \end{cases} \tag{42}$$

Proof 1) Using Lemma 12,

$$\mathcal{S}_\rho G_t(\mathcal{T})\mathcal{S}_\rho^* = \mathcal{S}_\rho G_t(\mathcal{S}_\rho^*\mathcal{S}_\rho)\mathcal{S}_\rho^* = G_t(\mathcal{S}_\rho\mathcal{S}_\rho^*)\mathcal{S}_\rho\mathcal{S}_\rho^* = G_t(\mathcal{L})\mathcal{L},$$

and by (39), we have

$$\mathcal{L}^{-a}(\mathcal{S}_\rho r_{t+1} - f_\rho) = \mathcal{L}^{-a}\left(G_t(\mathcal{L})\mathcal{L} - I\right) f_\rho.$$

Taking the ρ-norm, applying Assumption 2, we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho \le \|\mathcal{L}^{\zeta-a}(G_t(\mathcal{L})\mathcal{L} - I)\|R = \|\mathcal{L}^{\zeta-a}\Pi_1^t(\mathcal{L})\|R.$$

Note that the condition (7) implies (24). Applying Part 2) of Lemma 10, one can prove the first desired result.

2) By (39) and Assumption 2,

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H = \|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^* f_\rho\|_H \le \|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^* \mathcal{L}^\zeta\|R.$$

Noting that

$$\begin{aligned}\|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^* \mathcal{L}^\zeta\| &= \|\mathcal{T}^{a-1/2} G_t(\mathcal{T})\mathcal{S}_\rho^* \mathcal{L}^{2\zeta}\mathcal{S}_\rho G_t(\mathcal{T})\mathcal{T}^{a-1/2}\|^{1/2} \\ &= \|G_t^2(\mathcal{T})\mathcal{T}^{2\zeta+2a}\|^{1/2} = \|G_t(\mathcal{T})\mathcal{T}^{\zeta+a}\|,\end{aligned}$$

we thus have

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H \le \|G_t(\mathcal{T})\mathcal{T}^{\zeta+a}\|R.$$

If $0 \le \zeta + a \le 1$, i.e., $-\zeta \le a \le 1-\zeta$, then by using 1) of Lemma 10, we get

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H \le \lambda_t^{\zeta+a-1} R.$$

Similarly, when $a \ge 1-\zeta$, we have

$$\|\mathcal{T}^{a-1/2} r_{t+1}\|_H \le \|G_t(\mathcal{T})\mathcal{T}\|\|\mathcal{T}\|^{\zeta+a-1}R \le \kappa^{2(\zeta+a-1)}R,$$

where for the last inequality we used 1) of Lemma 10 and (24). This thus proves the second desired result. ■

With the above lemmas, we can prove the the following analytic result, which enables us to estimate the bias term in terms of several random quantities.

Lemma 15 *Under Assumption 2, let* $\tilde{\lambda} > 0$,

$$\Delta_1^{\mathbf{z}} = \|\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|^2 \vee 1, \qquad \Delta_3^{\mathbf{z}} = \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|$$

and

$$\Delta_2^{\mathbf{z}} = \|\mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}} r_{t+1} + \mathcal{T} r_{t+1}\|_H.$$

Then the following results hold.
1) For $0 < \zeta \le 1$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \le \left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{\zeta \vee \frac{1}{2}}\right)(C_1 R(\Delta_1^{\mathbf{z}})^{\zeta\vee\frac{1}{2}}\lambda_t^\zeta + 2\sqrt{\Delta_1^{\mathbf{z}}}\lambda_t^{-\frac{1}{2}}\Delta_2^{\mathbf{z}}). \tag{43}$$

2) For $\zeta > 1$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}}\left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{\zeta}\right)(C_2 R\lambda_t^\zeta + 2\lambda_t^{-\frac{1}{2}}\Delta_2^{\mathbf{z}} + C_3 R\lambda_t^{\frac{1}{2}}(\Delta_3^{\mathbf{z}})^{(\zeta-\frac{1}{2})\wedge 1}). \tag{44}$$

Here, C_1, C_2 *and* C_3 *are positive constants depending only on* ζ *and* κ.

Proof Using Lemma 8 with $s=1$, we can estimate $\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho$ as

$$\begin{aligned}\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})\mathcal{L}_{\mathbf{x}} f_\rho - f_\rho\|_\rho \leq & \| \underbrace{\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})[\mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}} r_{t+1} + \mathcal{T} r_{t+1}]}_{\mathbf{Bias.1}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})[\mathcal{S}_\rho^* f_\rho - \mathcal{T} r_{t+1}]}_{\mathbf{Bias.2}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho [I - G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}] r_{t+1}}_{\mathbf{Bias.3}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho r_{t+1} - f_\rho}_{\mathbf{Bias.4}} \|_\rho. \end{aligned} \tag{45}$$

In the rest of the proof, we will estimate the four terms of the r.h.s separately.
Estimating Bias.4
Using 1) of Lemma 14 with $a = 0$, we get

$$\|\mathbf{Bias.4}\|_\rho \leq (\zeta/\mathrm{e})^\zeta \lambda_t^\zeta R. \tag{46}$$

Estimating Bias.1
By a simple calculation, we know that for any $f \in H$,

$$\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\| \|f\|_H.$$

Note that

$$\|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| = \sqrt{\|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1} \mathcal{S}_\rho^*\|} = \sqrt{\|\mathcal{L}\mathcal{L}_{\tilde{\lambda}}^{-1}\|} \leq 1, \tag{47}$$

and that applying 1) of Lemma 10, with (28), we have

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\| \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t})/\sqrt{\lambda_t}.$$

Thus for any $f \in H$, we have

$$\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \|f\|_H. \tag{48}$$

Therefore,

$$\|\mathbf{Bias.1}\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \Delta_2^{\mathbf{z}}. \tag{49}$$

Estimating Bias.2
By (48), we have

$$\|\mathbf{Bias.2}\|_\rho \leq (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \lambda_t^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T} r_{t+1} - \mathcal{S}_\rho^* f_\rho\|_H.$$

Using (with $\mathcal{T} = \mathcal{S}_\rho^* \mathcal{S}_\rho$ and $\mathcal{L} = \mathcal{S}_\rho \mathcal{S}_\rho^*$)

$$\|\mathcal{T} r_{t+1} - \mathcal{S}_\rho^* f_\rho\|_H = \|\mathcal{S}_\rho^* (\mathcal{S}_\rho r_{t+1} - f_\rho)\|_H = \|\mathcal{L}^{1/2} (\mathcal{S}_\rho r_{t+1} - f_\rho)\|_\rho,$$

and applying 1) of Lemma 14 with $a = -1/2$, we get

$$\|\mathbf{Bias.2}\|_\rho \leq ((\zeta + 1/2)/\mathrm{e})^{\zeta + 1/2} (1 + \sqrt{\tilde{\lambda}/\lambda_t}) \sqrt{\Delta_1^{\mathbf{z}}} \lambda_t^\zeta R. \tag{50}$$

Estimating Bias.3
By 2) of Lemma 9,

$$\mathbf{Bias.3} = \mathcal{S}_\rho \Pi_1^t(\mathcal{T}_{\mathbf{x}}) r_{t+1}.$$

When $\zeta \le 1/2$, by a simple calculation, we have

$$\begin{aligned}\|\mathbf{Bias.3}\|_\rho \le & \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \|r_{t+1}\|_H \\ \le & \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \|r_{t+1}\|_H,\end{aligned}$$

where for the last inequality, we used (47). By 2) of Lemma 10, with (28),

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \le \sqrt{\lambda_t}(1/\sqrt{2\mathrm{e}} + \sqrt{\tilde{\lambda}/\lambda_t}), \tag{51}$$

and by 2) of Lemma 14,

$$\|r_{t+1}\|_H \le R \lambda_t^{\zeta - 1/2}.$$

It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}}(\sqrt{\tilde{\lambda}/\lambda_t} + 1/\sqrt{2\mathrm{e}}) R \lambda_t^\zeta.$$

When $1/2 < \zeta \le 1$, by a simple computation, we have

$$\|\mathbf{Bias.3}\|_\rho \le \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta} \mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2-\zeta} r_{t+1}\|_H.$$

Applying (47) and 2) of Lemma 14, we have

$$\|\mathbf{Bias.3}\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta} \mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\| R.$$

By 2) of Lemma 10,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta-1/2}\| = \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \le ((\zeta/\mathrm{e})^\zeta + (\tilde{\lambda}/\lambda_t)^\zeta) \lambda_t^\zeta.$$

Besides, by $\zeta \le 1$ and Lemma 11,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2-\zeta} \mathcal{T}_{\tilde{\lambda}}^{\zeta-1/2}\| = \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}(2\zeta-1)} \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}(2\zeta-1)}\| \le \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}} \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|^{2\zeta-1} \le (\Delta_1^{\mathbf{z}})^{\zeta-\frac{1}{2}}.$$

It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \le (\Delta_1^{\mathbf{z}})^\zeta ((\tilde{\lambda}/\lambda_t)^\zeta + (\zeta/\mathrm{e})^\zeta) R \lambda_t^\zeta.$$

When $\zeta > 1$, we rewrite **Bias.3** as

$$\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2} \cdot \mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2} \cdot \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})(\mathcal{T}_{\mathbf{x}}^{\zeta-1/2} + \mathcal{T}^{\zeta-1/2} - \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}) \mathcal{T}^{1/2-\zeta} r_{t+1}.$$

By a simple calculation, we can upper bound $\|\mathbf{Bias.3}\|_\rho$ by

$$\le \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| (\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| + \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \|\mathcal{T}^{\zeta-1/2} - \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|) \|\mathcal{T}^{1/2-\zeta} r_{t+1}\|.$$

Introducing with (47) and (51), and applying 2) of Lemma 14,

$$\|\mathbf{Bias.3}\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}}(\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_1^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| + (1/\sqrt{2\mathrm{e}} + \sqrt{\tilde{\lambda}/\lambda_t}) \sqrt{\lambda_t} \|\mathcal{T}^{\zeta-1/2} - \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\|) R.$$

By 2) of Lemma 10,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| \leq \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta}\Pi_1^t(\mathcal{T}_{\mathbf{x}})\| \leq 2^{\zeta-1}((\zeta/\mathrm{e})^{\zeta} + (\tilde{\lambda}/\lambda_t)^{\zeta})\lambda_t^{\zeta}.$$

Moreover, by Lemma 13 and $\max(\|\mathcal{T}\|, \|\mathcal{T}_{\mathbf{x}}\|) \leq \kappa^2$,

$$\|\mathcal{T}^{\zeta-1/2} - \mathcal{T}_{\mathbf{x}}^{\zeta-1/2}\| \leq (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|^{(\zeta-1/2)\wedge 1}.$$

Therefore, when $\zeta > 1$, **Bias.3** can be estimated as

$$\begin{aligned}&\|\mathbf{Bias.3}\|_{\rho} \\ \leq& \sqrt{\Delta_1^{\mathbf{z}}}\left(2^{\zeta-1}((\zeta/\mathrm{e})^{\zeta} + (\tilde{\lambda}/\lambda_t)^{\zeta})\lambda_t^{\zeta} + (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(1/\sqrt{2\mathrm{e}} + \sqrt{\tilde{\lambda}/\lambda_t})\sqrt{\lambda_t}(\Delta_3^{\mathbf{z}})^{(\zeta-1/2)\wedge 1}\right)R.\end{aligned}$$

From the above analysis, we know that $\|\mathbf{Bias.3}\|_{\rho}$ can be upper bounded by

$$\begin{cases} \sqrt{\Delta_1^{\mathbf{z}}}(\sqrt{\tilde{\lambda}/\lambda_t} + 1/\sqrt{2\mathrm{e}})R\lambda_t^{\zeta}, & \text{if } \zeta \in]0, 1/2], \\ (\Delta_1^{\mathbf{z}})^{\zeta}(\left(\tilde{\lambda}/\lambda_t\right)^{\zeta} + (\zeta/\mathrm{e})^{\zeta})R\lambda_t^{\zeta}, & \text{if } \zeta \in]1/2, 1], \\ \sqrt{\Delta_1^{\mathbf{z}}}\left(2^{\zeta-1}(\left(\frac{\zeta}{\mathrm{e}}\right)^{\zeta} + (\frac{\tilde{\lambda}}{\lambda_t})^{\zeta})\lambda_t^{\zeta} + (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(\frac{1}{\sqrt{2\mathrm{e}}} + \sqrt{\frac{\tilde{\lambda}}{\lambda_t}})\sqrt{\lambda_t}(\Delta_3^{\mathbf{z}})^{(\zeta-\frac{1}{2})\wedge 1}\right)R, & \text{if } \zeta \in]1, \infty[. \end{cases} \tag{52}$$

Introducing (46), (49), (50) and (52) into (45), and by a simple calculation, one can prove the desired results with

$$C_1 = (\zeta/\mathrm{e})^{\zeta} + 2((\zeta + \frac{1}{2})/\mathrm{e})^{\zeta+\frac{1}{2}} + ((\zeta \vee \frac{1}{2})/\mathrm{e})^{\zeta\vee\frac{1}{2}} + 1,$$

$$C_2 = (2^{\zeta-1} + 1)(\zeta/\mathrm{e})^{\zeta} + 2((\zeta + \frac{1}{2})/\mathrm{e})^{\zeta+\frac{1}{2}} + 2^{\zeta-1},$$

$$\text{and} \quad C_3 = (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(1/\sqrt{2\mathrm{e}} + 1).$$

■

The upper bounds in (43) and (44) depend on three random quantities, $\Delta_1^{\mathbf{z}}$, $\Delta_3^{\mathbf{z}}$ and $\Delta_2^{\mathbf{z}}$. To derive error bounds for the bias term from Lemma 15, it is necessary to estimate these three random quantities. We thus introduce the following lemmas.

Lemma 16 *Let $f : X \to Y$ be a measurable function such that $\|f\|_{\infty} < \infty$, then with probability at least $1 - \delta$ $(0 < \delta < 1/2)$,*

$$\|\mathcal{L}_{\mathbf{x}}f - \mathcal{L}f\|_H \leq 2\kappa\left(\frac{2\|f\|_{\infty}}{|\mathbf{x}|} + \frac{\|f\|_{\rho}}{\sqrt{|\mathbf{x}|}}\right)\log\frac{2}{\delta}.$$

Lemma 17 *Let $0 < \delta < 1/2$. It holds with probability at least $1 - \delta$:*

$$\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS} \leq \frac{6\kappa^2}{\sqrt{|\mathbf{x}|}}\log\frac{2}{\delta}.$$

Here, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm.

Lemma 18 *Let $0 < \delta < 1$ and $\lambda > 0$. With probability at least $1-\delta$, the following holds:*

$$\left\|(\mathcal{T}+\lambda)^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})(\mathcal{T}+\lambda)^{-1/2}\right\| \leq \frac{4\kappa^2\beta}{3|\mathbf{x}|\lambda} + \sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda}}, \quad \beta = \log\frac{4\kappa^2(\mathcal{N}(\lambda)+1)}{\delta\|\mathcal{T}\|}.$$

The proofs of Lemmas 16 and 17 are based on concentration result for Hilbert space valued random variable from (Pinelis and Sakhanenko, 1986), while the proof of Lemma 18 is based on the concentration inequality for norms of self-adjoint operators on a Hilbert space from (Tropp, 2012; Minsker, 2011). For completeness, we give the proofs in the appendix.

We will use Lemmas 16 and 14 to estimate the quantity $\Delta_2^{\mathbf{z}}$. The quantity $\Delta_3^{\mathbf{z}}$ can be estimated by Lemma 17 directly, as $\|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\|_{HS}$. The quantity $\Delta_1^{\mathbf{z}}$ can be estimated by the following lemma, whose proof is based on Lemma 18.

Lemma 19 *Under Assumption 3, let $c,\delta \in (0,1)$, $\lambda = |\mathbf{x}|^{-\theta}$ for some $\theta \geq 0$, and*

$$a_{|\mathbf{x}|,\delta,\gamma}(c,\theta) = \frac{32\kappa^2}{(\sqrt{9+24c}-3)^2}\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \theta\gamma\min\left(\frac{1}{\mathrm{e}(1-\theta)_+}, \log|\mathbf{x}|\right)\right). \tag{53}$$

Then with probability at least $1-\delta$,

$$\|(\mathcal{T}+\lambda)^{-1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{1/2}\|^2 \leq (1+c)a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)(1\vee|\mathbf{x}|^{\theta-1}), \text{ and}$$

$$\|(\mathcal{T}+\lambda)^{1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}\|^2 \leq (1-c)^{-1}a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)(1\vee|\mathbf{x}|^{\theta-1}).$$

Remark 1 *Typically, we will choose $c = 2/3$. In this case,*

$$a_{|\mathbf{x}|,\delta,\gamma}(2/3,\theta) = 8\kappa^2\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \theta\gamma\min\left(\frac{1}{\mathrm{e}(1-\theta)_+}, \log|\mathbf{x}|\right)\right). \tag{54}$$

We have with probability at least $1-\delta$,

$$\|(\mathcal{T}+\lambda)^{1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}\|^2 \leq 3a_{|\mathbf{x}|,\delta,\gamma}(2/3,\theta)(1\vee|\mathbf{x}|^{\theta-1}).$$

Proof We use Lemma 18 to prove the result. Let $c \in (0,1]$. By a simple calculation, we have that if $0 \leq u \leq \frac{\sqrt{9+24c}-3}{4}$, then $2u^2/3 + u \leq c$. Letting $\sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda'}} = u$, and combining with Lemma 18, we know that if

$$\sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda'}} \leq \frac{\sqrt{9+24c}-3}{4},$$

which is equivalent to

$$|\mathbf{x}| \geq \frac{32\kappa^2\beta}{(\sqrt{9+24c}-3)^2\lambda'}, \quad \beta = \log\frac{4\kappa^2(1+\mathcal{N}(\lambda'))}{\delta\|\mathcal{T}\|}, \tag{55}$$

then with probability at least $1-\delta$,

$$\left\|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}\right\| \leq c. \tag{56}$$

Note that from (56), we can prove

$$\|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|^2 \leq c+1, \quad \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 \leq (1-c)^{-1}. \tag{57}$$

Indeed, by simple calculations,

$$\begin{aligned}
&\|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}\mathcal{T}_{\lambda'}^{-1/2}\| = \|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2} + I\| \\
&\leq \|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}\| + \|I\| \leq c+1,
\end{aligned}$$

and (Caponnetto and De Vito, 2007)

$$\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1}\mathcal{T}_{\lambda'}^{1/2}\| = \|(I - \mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2})^{-1}\| \leq (1-c)^{-1}.$$

From the above analysis, we know that for any fixed $\lambda' > 0$ such that (55), then with probability at least $1-\delta$, (57) hold.

Now let $\lambda' = a\lambda$ when $\theta \in [0,1)$ and $\lambda' = a|\mathbf{x}|^{-1}$ when $\theta \geq 1$, where for notational simplicity, we denote $a_{|\mathbf{x}|,\delta,\gamma}(c,\theta)$ by a. We will prove that the choice on λ' ensures the condition (55) is satisfied, as thus with probability at least $1-\delta$, (57) holds. Obviously, one can easily prove that $a \geq 1$, using $\kappa^2 \geq 1$ and (24). Therefore, $\lambda' \geq \lambda$, and

$$\|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\sqrt{\lambda'/\lambda},$$

where for the last inequality, we used $\|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|^2 \leq \sup_{u\geq 0}\frac{u+\lambda}{u+\lambda'} \leq 1$ and $\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\|^2 \leq \sup_{u\geq 0}\frac{u+\lambda'}{u+\lambda} \leq \lambda'/\lambda$. Similarly,

$$\|\mathcal{T}_{\lambda}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda}^{1/2}\| \leq \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|\sqrt{\lambda'/\lambda}.$$

Combining with (57), and by a simple calculation, one can prove the desired bounds. What remains is to prove that the condition (55) is satisfied. By Assumption 3 and $a \geq 1$,

$$\beta \leq \log\frac{4\kappa^2(1+c_\gamma a^{-\gamma}|\mathbf{x}|^{(\theta\wedge 1)\gamma})}{\delta\|\mathcal{T}\|} \leq \log\frac{4\kappa^2(1+c_\gamma)|\mathbf{x}|^{\theta\gamma}}{\delta\|\mathcal{T}\|} = \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \theta\gamma\log|\mathbf{x}|.$$

If $\theta \geq 1$, or $\theta\gamma = 0$, or $\log|\mathbf{x}| \leq \frac{1}{(1-\theta)_+ \mathrm{e}}$, then the condition (55) follows trivially. Now consider the case $\theta \in (0,1)$, $\theta\gamma \neq 0$ and $\log|\mathbf{x}| \geq \frac{1}{(1-\theta)_+ \mathrm{e}}$. In this case, we apply (78) to get $\frac{\theta\gamma}{1-\theta}\log|\mathbf{x}|^{1-\theta} \leq \frac{\theta\gamma}{1-\theta}\frac{|\mathbf{x}|^{1-\theta}}{\mathrm{e}}$, and thus

$$\beta \leq \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{1-\theta}\frac{|\mathbf{x}|^{1-\theta}}{\mathrm{e}}.$$

Therefore, a sufficient condition for (55) is

$$\frac{|\mathbf{x}|^{1-\theta}a}{g(c)} \geq \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)}|\mathbf{x}|^{1-\theta}, \quad g(c) = \frac{32\kappa^2}{(\sqrt{9+24c}-3)^2}.$$

From the definition of a in (53),

$$a = g(c)\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)_+}\right),$$

and by a direct calculation, one can prove that the condition (55) is satisfied. The proof is complete. ∎

We also need the following lemma, which enables one to derive convergence results in expectation from convergence results in high probability.

Lemma 20 *Let $F :]0,1] \to \mathbb{R}_+$ be a monotone non-increasing, continuous function, and ξ a nonnegative real random variable such that*

$$\Pr[\xi > F(t)] \le t, \quad \forall t \in (0,1].$$

Then

$$\mathbb{E}[\xi] \le \int_0^1 F(t) dt.$$

The proof of the above lemma can be found in, e.g., (Blanchard and Mucke, 2016a). Now we are ready to state and prove the following result for the local bias.

Proposition 2 *Under Assumptions 2 and 3, we let $\tilde{\lambda} = n^{-1+\theta}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$, the following results hold.*
1) For $0 < \zeta \le 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \le C_5(R + \mathbf{1}_{\{\zeta<1/2\}}\|f_\rho\|_\infty)^2 \left(1 \vee \frac{\tilde{\lambda}^2}{\lambda_t^2} \vee [\gamma(\theta^{-1} \wedge \log n)]^{2\zeta \vee 1} \right) \lambda_t^{2\zeta}.$$

2)For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \le C_6 R^2 \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda_t^{2\zeta}} \vee \lambda_t^{1-2\zeta} \left(\frac{1}{n}\right)^{(\zeta - \frac{1}{2}) \wedge 1} \vee [\gamma(\theta^{-1} \wedge \log n)] \right) \lambda_t^{2\zeta}.$$

Here, C_5 and C_6 are positive constants depending only on κ, ζ and can be given explicitly in the proof.

Remark 2 *In this paper, we did not try to optimize the constants from the error bounds. But one should keep in mind that the constants can be further improved using an alternative proof for some special case, e.g., $\gamma = 0$ (Hsu et al., 2014), or $\zeta \ge 1/2$ (Caponnetto and De Vito, 2007), or $|y| \le M$. Furthermore, by assuming that $|y| \le M$ almost surely (which is not even satisfied with linear measurement model with Gaussian noise), the proof can be further simplified. Note also that, the constants from our error bounds appear to be larger than those from (Hsu et al., 2014; Caponnetto and De Vito, 2007), but our results do not require the extra assumption that the sample size is large enough as those in (Hsu et al., 2014; Caponnetto and De Vito, 2007).*

Proof We will use Lemma 15 to prove the results. To do so, we need to estimate $\Delta_1^{\mathbf{z}}$, $\Delta_2^{\mathbf{z}}$ and $\Delta_3^{\mathbf{z}}$.

By Lemma 19, we have that with probability at least $1 - \delta$,

$$\Delta_1^{\mathbf{z}} \le 3a_{n,\delta,\gamma}(1-\theta) \le (1 \vee \gamma[\theta^{-1} \wedge \log n]) 24\kappa^2 \log \frac{4\kappa^2 \mathrm{e}(c_\gamma + 1)}{\delta\|\mathcal{T}\|}, \tag{58}$$

where $a_{n,\delta,\gamma}(1-\theta) = a_{n,\delta,\gamma}(2/3, 1-\theta)$, given by (54). By Lemma 16, we have that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \leq 2\kappa \left(\frac{2\|r_{t+1} - f_\rho\|_\infty}{n} + \frac{\|\mathcal{S}_\rho r_{t+1} - f_\rho\|_\rho}{\sqrt{n}} \right) \log \frac{2}{\delta}.$$

Applying Part 1) of Lemma 14 with $a = 0$ to estimate $\|\mathcal{S}_\rho r_{t+1} - f_\rho\|_\rho$, we get that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \leq 2\kappa \left(2\|r_{t+1} - f_\rho\|_\infty / n + (\zeta/\mathrm{e})^\zeta R \lambda_t^\zeta / \sqrt{n} \right) \log \frac{2}{\delta}.$$

When $\zeta \geq 1/2$, we know that there exists a $f_H \in H$ such that $\mathcal{S}_\rho f_H = f_\rho$ (Steinwart and Christmann, 2008, Section 4.5). In fact, letting $g = \mathcal{L}^{-\zeta} f_\rho$, for $\zeta \geq 1/2$, f_ρ can be written as

$$f_\rho = \mathcal{L}^\zeta g = (\mathcal{S}_\rho \mathcal{S}_\rho^*)^\zeta g = \mathcal{S}_\rho (\mathcal{S}_\rho^* \mathcal{S}_\rho)^{\zeta - \frac{1}{2}} (\mathcal{S}_\rho^* \mathcal{S}_\rho)^{-\frac{1}{2}} \mathcal{S}_\rho^* g = \mathcal{S}_\rho \mathcal{T}^{\zeta - 1/2} (\mathcal{S}_\rho^* \mathcal{S}_\rho)^{-\frac{1}{2}} \mathcal{S}_\rho^* g.$$

Choosing $f_H = \mathcal{T}^{\zeta - \frac{1}{2}} (\mathcal{S}_\rho^* \mathcal{S}_\rho)^{-\frac{1}{2}} \mathcal{S}_\rho^* g$, as $(\mathcal{S}_\rho^* \mathcal{S}_\rho)^{-\frac{1}{2}} \mathcal{S}_\rho^*$ is partial isometric from $L^2_{\rho_X}$ to H and $\zeta \geq 1/2$, f_H is well defined. Moreover, $\mathcal{S}_\rho f_H = f_\rho$ and

$$\|r_{t+1} - f_H\|_H = \|G_t(\mathcal{T}) \mathcal{S}_\rho^* f_\rho - f_H\|_H = \|G_t(\mathcal{T}) \mathcal{S}_\rho^* \mathcal{S}_\rho f_H - f_H\|_H = \|(G_t(\mathcal{T})\mathcal{T} - I) f_H\|_H,$$

where we used (39) for the first equality. Introducing with $f_H = \mathcal{T}^{\zeta-1} \mathcal{S}_\rho^* g$, with $\|g\|_\rho \leq R$ by Assumption 2,

$$\|r_{t+1} - f_H\|_H \leq \|(G_t(\mathcal{T})\mathcal{T} - I) \mathcal{T}^{\zeta-1} \mathcal{S}_\rho^*\| \|g\|_\rho \leq \|(G_t(\mathcal{T})\mathcal{T} - I)\mathcal{T}^{\zeta - 1/2}\| R.$$

Using Lemma 10 with (24), we get

$$\|r_{t+1} - f_H\|_H \leq ((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} \lambda_t^{\zeta - 1/2} R.$$

Combing with (30),

$$\|r_{t+1} - f_\rho\|_\infty = \|r_{t+1} - f_H\|_\infty \leq \kappa \|r_{t+1} - f_H\|_H \leq \kappa ((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} R \lambda_t^{\zeta - 1/2}.$$

When $\zeta < 1/2$, by Part 2) of Lemma 14, $\|r_{t+1}\|_H \leq R \lambda_t^{\zeta - 1/2}$. Combining with (30), we have

$$\|r_{t+1} - f_\rho\|_\infty \leq \kappa \|r_{t+1}\|_H + \|f_\rho\|_\infty \leq \kappa \lambda_t^{\zeta - 1/2} R + \|f_\rho\|_\infty.$$

From the above analysis, we get that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \leq \log \frac{2}{\delta} \begin{cases} 2\kappa R \big(2\kappa ((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} / (\lambda_t n) + (\zeta/\mathrm{e})^\zeta / \sqrt{\lambda_t n} \big) \lambda_t^{\zeta + 1/2}, & \text{if } \zeta \geq 1/2, \\ 2\kappa \big(2\kappa R / (\lambda_t n) + 2\|f_\rho\|_\infty (n\lambda_t)^{-\zeta - 1/2} + (\zeta/\mathrm{e})^\zeta R / \sqrt{n\lambda_t} \big) \lambda_t^{\zeta + 1/2}, & \text{if } \zeta < 1/2, \end{cases}$$

which can be further relaxed as

$$\Delta_2^{\mathbf{z}} \leq C_4 \tilde{R} (1 \vee (\lambda_t n)^{-1}) \lambda_t^{\zeta + 1/2} \log \frac{2}{\delta}, \quad \widetilde{R} = R + \mathbf{1}_{\{\zeta < 1/2\}} \|f_\rho\|_\infty. \tag{59}$$

where

$$C_4 \leq \begin{cases} 2\kappa \big(2\kappa ((\zeta - 1/2)/\mathrm{e})^{\zeta - 1/2} + (\zeta/\mathrm{e})^\zeta \big), & \text{if } \zeta \geq 1/2, \\ 2\kappa \big(2\kappa + 2 + (\zeta/\mathrm{e})^\zeta \big), & \text{if } \zeta < 1/2. \end{cases}$$

Applying Lemma 17, and combining with the fact that $\|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\|_{HS}$, we have that with probability at least $1-\delta$,

$$\Delta_3^{\mathbf{z}} \leq \frac{6\kappa^2}{\sqrt{n}} \log \frac{2}{\delta}. \tag{60}$$

For $0<\zeta\leq 1$, by Pat 1) of Lemma 15, (58) and (59), we have that with probability at least $1-2\delta$,

$$\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \leq \left(3^{\zeta\vee\frac{1}{2}} C_1 R a_{n,\delta,\gamma}^{\zeta\vee\frac{1}{2}}(1-\theta) + 2\sqrt{3} C_4 \widetilde{R} a_{n,\delta,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{2}{\delta}\right)\left(1\vee\left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{\zeta\vee\frac{1}{2}} \vee \frac{1}{n\lambda_t}\right)\lambda_t^\zeta.$$

Rescaling δ, and then combining with Lemma 20, we get

$$\begin{aligned}
&\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \\
&\leq \int_0^1 \left(3^{\zeta\vee\frac{1}{2}} C_1 a_{n,\delta/2,\gamma}^{\zeta\vee\frac{1}{2}}(1-\theta) + 2\sqrt{3} C_4 a_{n,\delta/2,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{4}{\delta}\right)^2 d\delta \left(1\vee\left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{2\zeta\vee 1} \vee \frac{1}{n^2\lambda_t^2}\right)\lambda_t^{2\zeta}\widetilde{R}^2.
\end{aligned}$$

By a direct computation, noting that since $\tilde{\lambda}\geq n^{-1}$ and $2\zeta\leq 2$,

$$1\vee\left(\frac{\tilde{\lambda}}{\lambda_t}\right)^{2\zeta\vee 1} \vee \frac{1}{n^2\lambda_t^2} \leq 1\vee\left(\frac{\tilde{\lambda}}{\lambda_t}\right)^2,$$

and that for all $b\in\mathbb{R}_+$,

$$\int_0^1 \log^b \frac{1}{t} dt = \Gamma(b+1), \tag{61}$$

one can prove the first desired result with

$$C_5 = 2[C_1^2(48\kappa^2)^{2\zeta\vee 1}(A^{2\zeta\vee 1}+2) + 192\kappa^2 C_4^2(A(\log^2 4 + 2 + 2\log 4) + \log^2 4 + 4\log 4 + 6)],\ A = \log\frac{8\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}. \tag{62}$$

For $\zeta>1$, by Part 2) of Lemma 15, (58), (59) and (60), we know that with probability at least $1-3\delta$,

$$\begin{aligned}
&\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho \\
&\leq \sqrt{3}R(C_2 + 2C_4 + 6\kappa^2 C_3) a_{n,\delta,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{2}{\delta}\left(1\vee\frac{\tilde{\lambda}^\zeta}{\lambda_t^\zeta}\vee\frac{1}{n\lambda_t}\vee\lambda_t^{\frac{1}{2}-\zeta}\left(\frac{1}{n}\right)^{\frac{(\zeta-\frac{1}{2})\wedge 1}{2}}\right)\lambda_t^\zeta.
\end{aligned}$$

Rescaling δ, and applying Lemma 20, we get

$$\begin{aligned}
&\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2 \\
&\leq 3(C_2 + 2C_4 + 6\kappa^2 C_3)^2 R^2 \int_0^1 a_{n,\delta/3,\gamma}(1-\theta)\log^2\frac{6}{\delta} d\delta \left(1\vee\frac{\tilde{\lambda}^{2\zeta}}{\lambda_t^{2\zeta}}\vee\frac{1}{n^2\lambda_t^2}\vee\lambda_t^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1}\right)\lambda_t^{2\zeta}.
\end{aligned}$$

This leads to the second desired result with

$$C_6 = 24\kappa^2(C_2 + 2C_4 + 6\kappa^2 C_3)^2((A+1)\log^2 6 + 2(A+2)\log 6 + 2A + 6),\ A = \log\frac{12\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}. \tag{63}$$

by noting that $n^{-1} \leq \tilde{\lambda}$. The proof is complete. ■

Combining Proposition 2 with Lemma 7, we get the following results for the bias of the fully averaged estimator.

Proposition 3 *Under Assumptions 2 and 3, for any $\tilde{\lambda} = n^{-1+\theta}$ with $\theta \in [0,1]$ and any $t \in [T]$, the following results hold.*
1) For $0 < \zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_{t+1} - f_\rho\|_\rho^2 \leq C_5(R + \mathbf{1}_{\{\zeta<1/2\}}\|f_\rho\|_\infty)^2 \left(1 \vee \frac{\tilde{\lambda}^2}{\lambda_t^2} \vee [\gamma(\theta^{-1} \wedge \log n)]^{2\zeta\vee 1}\right) \lambda_t^{2\zeta}. \tag{64}$$

2) For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho \bar{h}_{t+1} - f_\rho\|_\rho^2 \leq C_6 R^2 \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda_t^{2\zeta}} \vee \lambda_t^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1} \vee [\gamma(\theta^{-1} \wedge \log n)]\right) \lambda_t^{2\zeta}. \tag{65}$$

Here, C_5 and C_6 are given by Proposition 2.

6.4 Estimating Sample Variance

In this section, we estimate sample variance $\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho$. We first introduce the following lemma.

Lemma 21 *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2 = \frac{1}{m}\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2. \tag{66}$$

According to Lemma 21, we know that the sample variance of the averaging over m local estimators can be well controlled in terms of the sample variance of a local estimator. In what follows, we will estimate the local sample variance, $\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2$. Throughout the rest of this subsection, we shall drop the index $s = 1$ for the first local estimator whenever it shows up, i.e., we rewrite $g_{1,t}$ as g_t, $\mathbf{z}_1$ as $\mathbf{z}$, etc.

Proposition 4 *Under Assumption 3, let $\tilde{\lambda} = n^{\theta-1}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 \leq C_8 \frac{\sigma^2}{n\tilde{\lambda}^\gamma}\left(1 \vee \frac{\tilde{\lambda}}{\lambda_t} \vee [\gamma(\theta^{-1} \wedge \log n])\right).$$

Here, C_8 is a positive constant depending only on $\kappa, c_\gamma, \|\mathcal{T}\|$ and will be given explicitly in the proof.

Proof Following from Lemma 8,

$$g_{t+1} - h_{t+1} = G_t(\mathcal{T}_\mathbf{x})(\mathcal{S}_\mathbf{x}^* \mathbf{y} - \mathcal{L}_\mathbf{x} f_\rho).$$

For notational simplicity, we let $\epsilon_i = y_i - f_\rho(x_i)$ for all $i \in [n]$ and $\boldsymbol{\epsilon} = (\epsilon_i)_{1\leq i\leq n}$. Then the above can be written as

$$g_{t+1} - h_{t+1} = G_t(\mathcal{T}_\mathbf{x})\mathcal{S}_\mathbf{x}^* \boldsymbol{\epsilon}.$$

Using the above relationship and the isometric property (25), we have

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1}-h_{t+1})\|_\rho^2 &= \mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho G_t(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_\rho^2 \\ &= \mathbb{E}_{\mathbf{y}}\|\mathcal{T}^{1/2} G_t(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_H^2 \\ &= \frac{1}{n^2}\sum_{l,k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k]\,\mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_l}\otimes K_{x_k}\right).\end{aligned}$$

Here, $\mathbb{E}_{\mathbf{y}}$ denotes the expectation with respect to $\mathbf{y}$ conditional on $\mathbf{x}$. From the definition of f_ρ and the independence of z_l and z_k when $l \neq k$, we know that $\mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k] = 0$ whenever $l \neq k$. Therefore,

$$\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1}-h_{t+1})\|_\rho^2 = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_k^2]\,\mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right).$$

Using the condition (8),

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1}-h_{t+1})\|_\rho^2 \le& \frac{\sigma^2}{n^2}\sum_{k=1}^n \mathrm{tr}\left(G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}G_t(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right) \\ =& \frac{\sigma^2}{n}\,\mathrm{tr}\left(\mathcal{T}(G_t(\mathcal{T}_{\mathbf{x}}))^2\mathcal{T}_{\mathbf{x}}\right) \\ =& \frac{\sigma^2}{n}\,\mathrm{tr}\left(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_{\tilde{\lambda}}^{1/2}(G_t(\mathcal{T}_{\mathbf{x}}))^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\tilde{\lambda}}^{1/2}\right) \\ \le& \frac{\sigma^2}{n}\,\mathrm{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}\mathcal{T}_{\tilde{\lambda}}^{-1/2})\|\mathcal{T}_{\tilde{\lambda}}^{1/2}G_t(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\tilde{\lambda}}^{1/2}\| \\ \le& \frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}\|\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}G_t(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}\|\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\mathcal{T}_{\tilde{\lambda}}^{1/2}\| \\ \le& \frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}\Delta_1^{\mathbf{z}}\|G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}\|\|G_t(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}\| \\ \le& \frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}\Delta_1^{\mathbf{z}}(1+\tilde{\lambda}/\lambda_t),\end{aligned}$$

where $\Delta_1^{\mathbf{z}}$ is given by Lemma 15 and we used 1) of Lemma 10 for the last inequality. Taking the expectation with respect to $\mathbf{x}$, this leads to

$$\mathbb{E}\|\mathcal{S}_\rho(g_{t+1}-h_{t+1})\|_\rho^2 \le \frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}(1+\tilde{\lambda}/\lambda_t)\mathbb{E}[\Delta_1^{\mathbf{z}}].$$

Applying Lemmas 19 and 20, we get

$$\begin{aligned}\mathbb{E}\|\mathcal{S}_\rho(g_{t+1}-h_{t+1})\|_\rho^2 \le& 6\frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}(1\vee(\tilde{\lambda}/\lambda_t))\int_0^1 a_{n,\delta,\gamma}(2/3,1-\theta)d\delta \\ \le& C_7\frac{\sigma^2\mathcal{N}(\tilde{\lambda})}{n}(1\vee(\tilde{\lambda}/\lambda_t)\vee[\gamma(\theta^{-1}\wedge\log n]),\end{aligned}$$

where $C_7 = 48\kappa^2\log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}$. Using Assumption 3, we get the desired result with

$$C_8 = c_\gamma 48\kappa^2\log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}. \tag{67}$$

■

Using the above proposition and Lemma 21, we derive the following results for sample variance.

Proposition 5 *Under Assumption 3, let $\tilde{\lambda} = n^{\theta-1}$ for some $\theta \in [0,1]$. Then for any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_{t+1} - \bar{h}_{t+1})\|_\rho^2 \leq C_8 \frac{\sigma^2}{N\tilde{\lambda}^\gamma}\left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda_t}\right) \vee [\gamma(\theta^{-1} \wedge \log n)]\right). \tag{68}$$

Here, C_8 is the positive constant given by Proposition 4.

6.5 Estimating Computational Variance

In this section, we estimate computational variance, $\mathbb{E}[\|\mathcal{S}_\rho(\bar{f}_t - \bar{h}_t)\|_\rho^2]$. We begin with the following lemma, from which we can see that the global computational variance can be estimated in terms of local computational variances.

Lemma 22 *For any $t \in [T]$, we have*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t)\|_\rho^2 = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2. \tag{69}$$

In what follows, we will estimate the local computational variance, i.e., $\mathbb{E}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2$. As in Subsections 6.3 and 6.4, we will drop the index s for the s-th local estimator whenever it shows up. We first introduce the following two lemmas, see (Lin and Rosasco, 2017b, Lemmas 20 and 24). The empirical risk $\mathcal{E}_{\mathbf{z}}(f)$ of a function f with respect to the samples $\mathbf{z}$ is defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n}\sum_{(x,y)\in\mathbf{z}}(f(x) - y)^2.$$

Lemma 23 *Assume that for all $t \in [T]$ with $t \geq 2$,*

$$\frac{1}{\eta_t}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t-1}\eta_i^2 \leq \frac{1}{4\kappa^2}. \tag{70}$$

Then for all $t \in [T]$,

$$\sup_{k\in[t]} \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(f_k)] \leq \frac{8\mathcal{E}_{\mathbf{z}}(0)\Sigma_1^t}{\eta_t t}. \tag{71}$$

Lemma 24 *For any $t \in [T]$, we have*

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{\kappa^2}{b}\sum_{k=1}^t \eta_k^2 \left\|\mathcal{T}^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(f_k)]. \tag{72}$$

Here, $\mathbb{E}_{\mathbf{J}}$ denotes the expectation with respect to $\mathbf{J}$ conditional on $\mathbf{z}$.

Now, we are ready to state and prove the result for local computational variance as follows.

Proposition 6 *Assume that (70) holds for any $t \in [T]$ with $t \geq 2$. Let $\tilde{\lambda} = n^{-\theta+1}$ for some $\theta \in [0,1]$. For any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq C_9 M^2 (1 \vee [\gamma(\theta^{-1} \wedge \log n)]) b^{-1} \sup_{k\in[t]} \left\{ \frac{\Sigma_1^k}{\eta_k k} \right\} \left(\sum_{k=1}^{t-1} \eta_k^2 (\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2 \right).$$

Here, C_9 is a positive constant depending only on $\kappa, c_\gamma, \|\mathcal{T}\|$ and can be given explicitly in the proof.

Proof Following from Lemmas 24 and 23, we have that,

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{8\kappa^2 \mathcal{E}_{\mathbf{z}}(0)}{b} \sum_{k=1}^{t} \eta_k^2 \left\| \mathcal{T}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \right\|^2 \sup_{k\in[t]} \left\{ \frac{\Sigma_1^k}{\eta_k k} \right\}.$$

Taking the expectation with respect to $\mathbf{y}$ conditional on $\mathbf{x}$, and then with respect to $\mathbf{x}$, noting that $\int_Y y^2 d\rho(y|x) \leq M^2$, we get

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \frac{8\kappa^2 M^2}{b} \sup_{k\in[t]} \left\{ \frac{\Sigma_1^k}{\eta_k k} \right\} \sum_{k=1}^{t} \eta_k^2 \mathbb{E} \left\| \mathcal{T}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \right\|^2.$$

Note that

$$\begin{aligned} \left\| \mathcal{T}^{\frac{1}{2}} \Pi_k^t(\mathcal{T}_{\mathbf{x}}) \right\|^2 &\leq \|\mathcal{T}^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\|^2 \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \Pi_k^t(\mathcal{T}_{\mathbf{x}})\|^2 \leq \Delta_1^{\mathbf{z}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}} (\Pi_k^t(\mathcal{T}_{\mathbf{x}}))^2\| \\ &\leq \Delta_1^{\mathbf{z}} (\|\mathcal{T}_{\mathbf{x}} \Pi_k^t(\mathcal{T}_{\mathbf{x}})\| + \tilde{\lambda}\|\Pi_k^t(\mathcal{T}_{\mathbf{x}})\|) \|\Pi_k^t(\mathcal{T}_{\mathbf{x}})\| \leq \Delta_1^{\mathbf{z}} (\lambda_{k:t}\mathrm{e}^{-1} + \tilde{\lambda}), \end{aligned}$$

where $\Delta_1^{\mathbf{z}}$ is given by Lemma 15 and for the last inequality we used Part 2) of Lemma 10. Therefore,

$$\mathbb{E}\|\mathcal{S}_\rho f_{t+1} - \mathcal{S}_\rho g_{t+1}\|_\rho^2 \leq \mathbb{E}[\Delta_1^{\mathbf{z}}] \frac{8\kappa^2 M^2}{b} \sup_{k\in[t]} \left\{ \frac{\Sigma_1^k}{\eta_k k} \right\} \left(\sum_{k=1}^{t-1} \eta_k^2 (\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2 \right).$$

Using Lemmas 19 and 20, and by a simple calculation, one can upper bound $\mathbb{E}[\Delta_1^{\mathbf{z}}]$ and consequently prove the desired result with C_9 given by

$$C_9 = 192\kappa^4 \log \frac{4\kappa^2 (c_\gamma + 1)\mathrm{e}}{\|\mathcal{T}\|}.$$

The proof is complete. ■

Combining Lemma 22 with Proposition 6, we have the following error bounds for computational variance.

Proposition 7 *Assume that (70) holds for any $t \in [T]$ with $t \geq 2$. Let $\tilde{\lambda} = n^{-\theta+1}$ for some $\theta \in [0,1]$. For any $t \in [T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho (\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \leq C_9 M^2 (1 \vee [\gamma(\theta^{-1} \wedge \log n)]) \frac{1}{mb} \sup_{k\in[t]} \left\{ \frac{\Sigma_1^k}{\eta_k k} \right\} \left(\sum_{k=1}^{t-1} \eta_k^2 (\tilde{\lambda} + \lambda_{k+1:t}\mathrm{e}^{-1}) + \eta_t^2 \right). \tag{73}$$

Here, C_9 is the positive constant from Proposition 6.

6.6 Deriving Total Errors

We are now ready to derive total error bounds for (distributed) SGM and to prove the main theorems for (distributed) SGM of this paper.

Proof of Theorem 1 We will use Propositions 1, 3, 5 and 7 to prove the result.

We first show that the condition (11) implies (70). Indeed, when $\eta_t = \eta$, for any $t \in [T]$

$$\frac{1}{\eta_t}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t-1}\eta_i^2 = \eta\sum_{k=2}^{t}\frac{1}{k} \le \eta\sum_{k=2}^{t}\int_{k-1}^{k}\frac{1}{x}dx = \eta\log t \le \frac{1}{4\kappa^2}$$

where for the last inequality, we used the condition (11). Thus, by Proposition 7, (73) holds. Note also that $\lambda_{k+1:t} = \frac{1}{\eta(t-k)}$ and $\lambda_t = \frac{1}{\eta t}$ as $\eta_t = \eta$. It thus follows from (73) that

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \le C_9 M^2 (1 \vee [\gamma(\theta^{-1} \wedge \log n)]) \frac{\eta}{mb}\left(\tilde{\lambda}\eta(t-1) + \sum_{k=1}^{t-1}\frac{1}{\mathrm{e}(t-k)} + \eta\right).$$

Applying

$$\sum_{k=1}^{t-1}\frac{1}{t-k} = \sum_{k=1}^{t-1}\frac{1}{k} \le 1 + \sum_{k=2}^{t-1}\int_{k-1}^{k}\frac{1}{x}dx \le 1 + \log t,$$

and (11), we get

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{f}_{t+1} - \bar{g}_{t+1}\|_\rho^2 \le C_9 M^2 (1 \vee [\gamma(\theta^{-1} \wedge \log n)] \vee \tilde{\lambda}\eta t \vee \log t) \frac{\eta}{mb}\left(2 + \frac{1}{4\kappa^2}\right).$$

Introducing the above inequality, (64) (or (65)), and (68) into the error decomposition (36), by a direct calculation, one can prove the desired results with

$$C_{10} = C_9\left(2 + \frac{1}{4\kappa^2}\right) = 192\kappa^4 \log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}\left(2 + \frac{1}{4\kappa^2}\right). \tag{74}$$

■

Proof of Corollary 2 In Theorem 1, we let $\tilde{\lambda} = N^{-\frac{1}{2\zeta+\gamma}}$. In this case, with Condition (16), it is easy to show that

$$1 \ge \theta = \frac{\log\tilde{\lambda}}{\log n} + 1 = \frac{\log\tilde{\lambda}}{\log N - \log m} + 1 \ge -\frac{1}{2\zeta+\gamma}\frac{\log N}{\log N - \beta\log N} + 1 > 0.$$

The proof can be done by simply applying Theorem 1 and plugging with the specific choices of η_t, b, and T_*. ■

Proof of Corollary 1 Since $f_\rho \in H$, we know from (26) that Assumption 2 holds with $\zeta = \frac{1}{2}$ and $R \le \|f_\rho\|_H$. As noted in comments after Assumption 3, (10) trivially holds with $\gamma = 1$ and $c_\gamma = \kappa^2$. Applying Corollary 2, one can prove the desired results. ■

Proof of Corollary 3 In Theorem 1, we let $m = 1$ and $n = N$ and $\tilde{\lambda} = N^{\theta-1}$ with $\theta = 1 - \alpha$. Then it is easy to see that

$$\gamma(\theta^{-1} \wedge \log N) \le \begin{cases} \frac{\gamma(2\zeta+\gamma)}{2\zeta+\gamma-1}, & \text{if } 2\zeta+\gamma > 1, \\ \gamma\log N, & \text{if } 2\zeta+\gamma \le 1. \end{cases}$$

Following from (12) or (13), and plugging with the specific choices on η_t, T_*, b, one can prove the desired error bounds. ■

Acknowledgments

The authors thank Yen-Huan Li for his comments. This work was sponsored by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111. It has also received funding from Hasler Foundation Program: Cyber Human Systems (project number 16066), and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n 725594-time-data).

References

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

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

Gilles Blanchard and Nicole Mucke. Optimal rates for regularization of statistical inverse learning problems. *arXiv preprint arXiv:1604.04054*, 2016a.

Gilles Blanchard and Nicole Mucke. Parallelizing spectral algorithms for kernel learning. *arXiv preprint arXiv:1610.07487*, 2016b.

Leon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information theory*, 52(2):489–509, 2006.

Andrea Caponnetto. Optimal learning rates for regularization operators in learning theory. *Technical report*, 2006.

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

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Lee H Dicker, Dean P Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

L Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 2017.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *arXiv preprint arXiv:1610.03774*, 2016.

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

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning*, number CONF, 2018.

Junhong Lin and Lorenzo Rosasco. Optimal rates for learning with Nyström stochastic gradient methods. *arXiv preprint arXiv:1710.07797*, 2017a.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017b.

Junhong Lin and Ding-Xuan Zhou. Learning theory of randomized kaczmarz algorithm. *The Journal of Machine Learning Research*, 16(1):3341–3365, 2015.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

Ben London. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2935–2944, 2017.

Jing Lu, Steven CH Hoi, Jialei Wang, Peilin Zhao, and Zhi-Yong Liu. Large scale online kernel learning. *Journal of Machine Learning Research*, 17(47):1, 2016.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2011.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4): 1574–1609, 2009.

IF Pinelis and AI Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis.* Cambridge university press, 2004.

Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Alex J Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. 2000.

Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference On Learning Theory*, 2009.

Rashish Tandon, Si Si, Pradeep Ravikumar, and Inderjit Dhillon. Kernel ridge regression via partitioning. *arXiv preprint arXiv:1608.01976*, 2016.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9): 5716–5735, 2014.

Philipp Thomann, Ingo Steinwart, Ingrid Blaschzyk, and Mona Meister. Spatial decompositions for large scale SVMs. *arXiv preprint arXiv:1612.00374*, 2016.

Joel A Tropp. User-friendly tools for random matrices: An introduction. Technical report, DTIC Document, 2012.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 661–667. MIT press, 2000.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, pages 476–484, 2012.

Yun Yang, Mert Pilanci, Martin J Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Alp Yurtsever, Madeleine Udell, Joel Aaron Tropp, and Volkan Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine learning*, page 116. ACM, 2004.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

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

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.

Appendix

- In **Appendix A**, we provide a list of notations commonly used in this paper.
- In **Appendix B**, we prove some of the lemmas and propositions from Section 6.
- In **Appendix C**, we prove our main results for distributed SA. We first introduce an error decomposition, which decomposes total errors into bias and sample variance. We then estimate these two terms in the following two subsequent subsections. Plugging the two estimates into the error decomposition, we prove the desired results.

Appendix A. List of Notations

Notation	Meaning
H	the hypothesis space, RKHS
X, Y, Z	the input space, the output space and the sample space ($Z = X \times Y$)
ρ, ρ_X	the fixed probability measure on Z, the induced marginal measure of ρ on X
$\rho(\cdot\|x)$	the conditional probability measure on Y w.r.t. $x \in X$ and ρ
N, n, m	the total sample size, the local sample size, the number of partition ($N = nm$)
$\bar{\mathbf{z}}$	the whole samples $\{z_i\}_{i=1}^N$, where each z_i is i.i.d. according to ρ.
$\mathbf{z}_s$	the samples $\{z_{s,i} = (x_{s,i}, y_{s,i})\}_{i=1}^n$ for the s-th local machine, $s \in [m]$
$\mathcal{E}$	the expected risk defined by (1)
κ^2	the constant from the bounded assumption (7) on the hypothesis space H
$\{f_{s,t}\}_t$	the sequence generated by SGM over the local sample $\mathbf{z}_s$, given by (4)
$\{\bar{f}_t\}$	the sequence generated by distributed SGM, i.e., $\bar{f}_t = \frac{1}{m}\sum_{s=1}^m f_{s,t}$
b	the minibatch size of SGM
T	the maximal number of iterations for SGM
$j_{s,i}$ ($j_{s,t}$ etc.)	the random index from the uniform distribution on $[n]$ for SGM performing on the s-th local sample set $\mathbf{z}_s$
$\mathbf{J}_{s,t}$	the set of random indices at t-th iteration of SGM performing on the s-th local sample set $\mathbf{z}_s$
$\mathbf{J}_s$	the set of all random indices for SGM performing on the s-th local sample set $\mathbf{z}_s$ after T iterations
$\mathbf{J}$	the set of all random indices for distributed SGM after T iterations
$\mathbb{E}_{\mathbf{J}_s}$	the expectation with respect to the random variables $\mathbf{J}_s$ (conditional on $\mathbf{z}_s$)
$\mathbb{E}_{\mathbf{J}}$	the expectation with respect to the random variables $\mathbf{J}$ (conditional on $\bar{\mathbf{z}}$)
$\mathbb{E}_{\mathbf{y}}$	the expectation with respect to the random variables $\mathbf{y}$ (conditional on $\mathbf{x}$)
$\{\eta_t\}_t$	the sequence of step-sizes
M, σ	the positive constants from Assumption **??**
$L^2_{\rho_X}$	the Hilbert space of square integral functions from X to $\mathbb{R}$ with respect to ρ_X
f_ρ	the regression function defined (2)
ζ, R	the parameters related to the 'regularity' of f_ρ (see Assumption 2)
γ, c_γ	the parameters related to the effective dimension (see Assumption 3)
$\{g_{s,t}\}_t$	the sequence generated by GM (31) with respect to the s-th local sample set $\mathbf{z}_s$
$\{\bar{g}_t\}_t$	the sequence generated by distributed GM (32)
$\{h_{s,t}\}_t$	the sequence generated by pseudo GM (33) over the s-th local sample set $\mathbf{z}_s$
$\{\bar{h}_t\}_t$	the sequence generated by distributed pseudo GM
$\{r_t\}_t$	the sequence generated by population GM (37)
$\mathcal{S}_\rho$	the inclusion map from $H \to L^2_{\rho_X}$
$\mathcal{S}_\rho^*$	the adjoint operator of $\mathcal{S}_\rho$, $\mathcal{S}_\rho^* f = \int_X f(x) K_x d\rho_X(x)$
$\mathcal{L}$	the operator from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, $\mathcal{L}(f) = \mathcal{S}_\rho \mathcal{S}_\rho^* f = \int_X f(x) K_x \rho_X(x)$
$\mathcal{T}$	the covariance operator from H to H, $\mathcal{T} = \mathcal{S}_\rho^* \mathcal{S}_\rho = \int_X \langle \cdot, K_x \rangle_H K_x d\rho_X(x)$
$\mathcal{S}_{\mathbf{x}}$	the sampling operator from H to $\mathbb{R}^{\|\mathbf{x}\|}$, $(\mathcal{S}_{\mathbf{x}} f)_i = f(x_i), x_i \in \mathbf{x}$
$\mathcal{S}_{\mathbf{x}}^*$	the adjoint operator of $\mathcal{S}_{\mathbf{x}}$, $\mathcal{S}_{\mathbf{x}}^* \mathbf{y} = \frac{1}{\|\mathbf{x}\|}\sum_{i=1}^{\|\mathbf{x}\|} y_i K_{x_i}$
$\mathcal{T}_{\mathbf{x}}$	the empirical covariance operator, $\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^* \mathcal{S}_{\mathbf{x}} = \frac{1}{\|\mathbf{x}\|}\sum_{i=1}^{\|\mathbf{x}\|} \langle \cdot, K_{x_i} \rangle_H K_{x_i}$
$\Pi_{t+1}^T(L)$	$= \Pi_{k=t+1}(I - \eta_k L)$ when $t \in [T-1]$ and $\Pi_{t+1}^T = I$ if $t \geq T$
$\tilde{\lambda}$	a pseudo regularization parameter, $\tilde{\lambda} > 0$
$\mathcal{T}_{\tilde{\lambda}}$,	$\mathcal{T}_{\tilde{\lambda}} = \mathcal{T} + \tilde{\lambda}$

$\mathcal{T}_{\mathbf{x}\tilde{\lambda}}$,	$\mathcal{T}_{\mathbf{x}\tilde{\lambda}} = \mathcal{T}_{\mathbf{x}} + \tilde{\lambda}$
$G_t(\cdot)$	the filter function of GM, (38)
$\widetilde{G}_\lambda(\cdot)$	a general filter function
λ	a regularization parameter $\lambda > 0$
$[t]$	the set $\{1, \cdots, t\}$
$b_1 \lesssim b_2$	$b_1 \leq Cb_2$ for some universal constant $C > 0$
$b_1 \simeq b_2$	$b_2 \lesssim b_1 \lesssim b_2$
$\Delta_1^{\mathbf{z}}, \Delta_2^{\mathbf{z}}, \Delta_3^{\mathbf{z}}$	the random quantities defined in Lemma 15 (or Lemma 30)
$\mathcal{L}_{\mathbf{x}}$	an operator defined by (35)
Σ_k^t	$= \sum_{i=k}^t \eta_i$ ($= 0$ if $k > t$)
λ_t	the regularization parameter of GM ($= (\Sigma_1^t)^{-1}$)
$\lambda_{k:t}$	$= (\Sigma_k^t)^{-1}$ ($= \infty$ if $k > t$)
$a_{\lvert\mathbf{x}\rvert,\delta,\gamma}(c,\theta)$	the quantity defined by (53)
$g_\lambda^{\mathbf{z}_s}$	the estimator defined by SA over the s-th local sample set $\mathbf{z}_s$, see Algorithm 2
$\bar{g}_\lambda^{\bar{\mathbf{z}}}$	the estimator defined by distributed SA, see Algorithm 2
$h_\lambda^{\mathbf{z}_s}$	the estimator defined by pseudo SA over the s-th local sample set $\mathbf{z}_s$, (83)
$\bar{h}_\lambda^{\bar{\mathbf{z}}}$	the estimator defined by distributed pseudo SA, (84)
$\tilde{r}_\lambda$	the function defined by population SA, (86)

Appendix B. Proofs for Section 3

In this section, we provide the missing proofs of lemmas and propositions from Section 3.

B.1 Proof of Proposition 1

For any $s \in [m]$, using an inductive argument, one can prove that (Lin and Rosasco, 2017b)

$$\mathbb{E}_{\mathbf{J}_s|\mathbf{z}_s}[f_{s,t}] = g_{s,t}. \tag{75}$$

Here $\mathbb{E}_{\mathbf{J}_s|\mathbf{z}_s}$ (or abbreviated as $\mathbb{E}_{\mathbf{J}_s}$) denotes the conditional expectation with respect to $\mathbf{J}_s$ given $\mathbf{z}_s$. Indeed, taking the conditional expectation with respect to $\mathbf{J}_{s,t}$ (given $\mathbf{z}_s$) on both sides of (4), and noting that $f_{s,t}$ depends only on $\mathbf{J}_{s,1}, \cdots, \mathbf{J}_{s,t-1}$ (given $\mathbf{z}_s$), one has

$$\mathbb{E}_{\mathbf{J}_{s,t}}[f_{s,t+1}] = f_{s,t} - \eta_t \frac{1}{n} \sum_{i=1}^n (f_{s,t}(x_{s,i}) - y_{s,i}) K_{x_{s,i}},$$

and thus,

$$\mathbb{E}_{\mathbf{J}_s}[f_{s,t+1}] = \mathbb{E}_{\mathbf{J}_s}[f_{s,t}] - \eta_t \frac{1}{n} \sum_{i=1}^n (\mathbb{E}_{\mathbf{J}_s}[f_{s,t}](x_{s,i}) - y_{s,i}) K_{x_{s,i}}, \quad t = 1, \ldots, T,$$

which satisfies the iterative relationship given in (31). Similarly, using the definition of the regression function (2) and an inductive argument, one can also prove that

$$\mathbb{E}_{\mathbf{y}_s}[g_{s,t}] = h_{s,t}. \tag{76}$$

Here, $\mathbb{E}_{\mathbf{y}_s}$ denotes the conditional expectation with respect to $\mathbf{y}_s$ given $\mathbf{x}_s$.

We have

$$\|\mathcal{S}_\rho \bar{f}_t - f_\rho\|_\rho^2 = \|\mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t\|_\rho^2 + \|\mathcal{S}_\rho \bar{g}_t - f_\rho\|_\rho^2 + 2\langle \mathcal{S}_\rho \bar{f}_t - \mathcal{S}_\rho \bar{g}_t, \mathcal{S}_\rho \bar{g}_t - f_\rho \rangle.$$

Taking the conditional expectation with respect to $\mathbf{J}$ (given $\mathbf{z}$) on both sides, using (75) which implies

$$\mathbb{E}_{\mathbf{J}}\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t) = \frac{1}{m}\sum_{s=1}^m \mathcal{S}_\rho \mathbb{E}_{\mathbf{J}_s}[f_{s,t} - g_{s,t}] = 0,$$

we thus have

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\bar{f}_t - f_\rho\|_\rho^2 = \mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\bar{f}_t - \mathcal{S}_\rho\bar{g}_t\|_\rho^2 + \|\mathcal{S}_\rho\bar{g}_t - f_\rho\|_\rho^2.$$

Taking the conditional expectation with respect to $\bar{\mathbf{y}} = \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ (given $\bar{\mathbf{x}} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$), noting that

$$\mathbb{E}_{\bar{\mathbf{y}}}\|\mathcal{S}_\rho\bar{g}_t - f_\rho\|_\rho^2 = \mathbb{E}_{\bar{\mathbf{y}}}[\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2] + \|\mathcal{S}_\rho\bar{h}_t - f_\rho\|_\rho^2 + 2\langle \mathcal{S}_\rho\mathbb{E}_{\bar{\mathbf{y}}}[\bar{g}_t - \bar{h}_t], \mathcal{S}_\rho\bar{h}_t - f_\rho\rangle_\rho$$

and that from (76),

$$\langle \mathcal{S}_\rho\mathbb{E}_{\bar{\mathbf{y}}}[\bar{g}_t - \bar{h}_t], \mathcal{S}_\rho\bar{h}_t - f_\rho\rangle_\rho = \frac{1}{m}\sum_{s=1}^m \langle \mathcal{S}_\rho\mathbb{E}_{\mathbf{y}_s}(g_{s,t} - h_{s,t}), \mathcal{S}_\rho\bar{h}_t - f_\rho\rangle_\rho = 0,$$

we know that

$$\mathbb{E}_{\bar{\mathbf{y}}}\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\bar{f}_t - f_\rho\|_\rho^2 = \mathbb{E}_{\bar{\mathbf{y}}}\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\bar{f}_t - \mathcal{S}_\rho\bar{g}_t\|_\rho^2 + \mathbb{E}_{\bar{\mathbf{y}}}[\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho^2] + \|\mathcal{S}_\rho\bar{h}_t - f_\rho\|_\rho^2,$$

which leads to the desired result. ■

B.2 Proof of Lemma 7

By Jensen's inequality, we can prove the desired result:

$$\mathbb{E}\|\mathcal{S}_\rho\bar{h}_t - f_\rho\|_\rho^2 = \mathbb{E}\left\|\frac{1}{m}\sum_{s=1}^m(\mathcal{S}_\rho h_{s,t} - f_\rho)\right\|_\rho^2 \le \frac{1}{m}\mathbb{E}\sum_{s=1}^m \|\mathcal{S}_\rho h_{s,t} - f_\rho\|_\rho^2 = \mathbb{E}\|\mathcal{S}_\rho h_{1,t} - f_\rho\|_\rho^2.$$

■

B.3 Proof of Lemma 9

1). For $\alpha = 0$ or 1, the proof is straightforward and can be found in (Yao et al., 2007). Indeed, for all $u \in [0, \kappa^2]$, $\Pi_{k+1}^t(u) \le 1$ and thus $G_t(u) \le \sum_{k=1}^t \eta_k = \lambda_t^{-1}$. Moreover, writing $\eta_k u = 1 - (1 - \eta_k u)$, we have

$$uG_t(u) = \sum_{k=1}^t (\eta_k u)\Pi_{k+1}^t(u) = \sum_{k=1}^t (\Pi_{k+1}^t(u) - \Pi_k^t(u)) = 1 - \Pi_1^t(u) \le 1. \tag{77}$$

Now we consider the case $0 < \alpha < 1$. We have

$$u^\alpha G_t(u) = |uG_t(u)|^\alpha |G_t(u)|^{1-\alpha} \le \lambda_t^{\alpha-1},$$

where we used $uG_t(u) \le 1$ and $G_t(u) \le \lambda_t^{-1}$ in the above.

2) By (77), we have $(1 - uG_t(u))u^\alpha = \Pi_1^t(u)u^\alpha$. Then the desired result is a direct consequence of Conclusion 3).

3) The proof can be also found, e.g., in (Lin and Rosasco, 2017b, Page 17). Using the basic inequality

$$1 + x \le \mathrm{e}^x \qquad \text{for all } x \ge -1, \tag{78}$$

with $\eta_l \kappa^2 \le 1$, we get

$$\Pi_{k+1}^t(u) u^\alpha \le \exp\left\{-u\Sigma_{k+1}^t\right\} u^\alpha.$$

The maximum of the function $g(u) = \mathrm{e}^{-cu}u^\alpha$ (with $c > 0$) over $\mathbb{R}_+$ is achieved at $u_{\max} = \alpha/c$, and thus

$$\sup_{u\ge 0} \mathrm{e}^{-cu}u^\alpha = \left(\frac{\alpha}{\mathrm{e}c}\right)^\alpha. \tag{79}$$

Using this inequality with $c = \Sigma_{k+1}^t$, one can prove the desired result. ∎

B.4 Proof of Lemma 10

1) Following from the spectral theorem, one has

$$\|(L+\tilde{\lambda})^\alpha G_t(L)\| \le \sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha G_t(u) \le \sup_{u\in[0,\kappa^2]} (u^\alpha + \tilde{\lambda}^\alpha) G_t(u).$$

Using Part 1) of Lemma 9 to the above, one can prove the first conclusion.
2) Using the spectral theorem,

$$\|\Pi_1^t(L)(L+\tilde{\lambda})^\alpha\| \le \sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u).$$

When $\alpha \le 1$,

$$\sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u) \le \sup_{u\in[0,\kappa^2]} (u^\alpha+\tilde{\lambda}^\alpha) \Pi_1^t(u) \le (\alpha/\mathrm{e})^\alpha \lambda_t^\alpha + \tilde{\lambda}^\alpha,$$

where for the last inequality, we used Part 2) of Lemma 9. Similarly, when $\alpha > 1$, by Hölder's inequality, and Part 2) of Lemma 9,

$$\sup_{u\in[0,\kappa^2]} (u+\tilde{\lambda})^\alpha \Pi_1^t(u) \le 2^{\alpha-1} \sup_{u\in[0,\kappa^2]} (u^\alpha+\tilde{\lambda}^\alpha) \Pi_1^t(u) \le 2^{\alpha-1}((\alpha/\mathrm{e})^\alpha \lambda_t^\alpha + \tilde{\lambda}^\alpha).$$

From the above analysis, one can prove the second conclusion.
3) Simply applying the spectral theorem and 3) of Lemma 9, one can prove the third conclusion. ∎

B.5 Proof of Lemma 16

We first introduce the following concentration result for Hilbert space valued random variable used in (Caponnetto and De Vito, 2007) and based on the results in (Pinelis and Sakhanenko, 1986).

Lemma 25 *Let $w_1, \cdots, w_m$ be i.i.d random variables in a separable Hilbert space with norm $\|\cdot\|$. Suppose that there are two positive constants B and σ^2 such that*

$$\mathbb{E}[\|w_1 - \mathbb{E}[w_1]\|^l] \le \frac{1}{2} l! B^{l-2}\sigma^2, \quad \forall l \ge 2. \tag{80}$$

Then for any $0 < \delta < 1/2$, the following holds with probability at least $1-\delta$,

$$\left\| \frac{1}{m}\sum_{k=1}^m w_m - \mathbb{E}[w_1] \right\| \le 2\left(\frac{B}{m} + \frac{\sigma}{\sqrt{m}}\right)\log\frac{2}{\delta}.$$

In particular, (80) holds if

$$\|w_1\| \le B/2 \ \ a.s., \quad and \quad \mathbb{E}[\|w_1\|^2] \le \sigma^2. \tag{81}$$

Lemmas 16 and 17 can be proved by simply applying the above lemma.

Proof of Lemma 16 Let $\xi_i = f(x_i)K_{x_i}$ for $i = 1, \cdots, |\mathbf{x}|$. Obviously,

$$\mathcal{L}_{\mathbf{x}} f - \mathcal{L} f = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}(\xi_i - \mathbb{E}[\xi_i]),$$

and by Assumption (7), we have

$$\|\xi\|_H \le \|f\|_\infty \|K_x\|_H \le \kappa \|f\|_\infty$$

and

$$\mathbb{E}\|\xi\|_H^2 \le \kappa^2 \|f\|_\rho^2.$$

Applying Lemma 25 with $B' = 2\kappa\|f\|_\infty$ and $\sigma = \kappa\|f\|_\rho$, one can prove the desired result. ■

B.6 Proof of Lemma 17

Let $\xi_i = K_{x_i} \otimes K_{x_i}$, for all $i \in [|\mathbf{x}|]$. Obviously,

$$\mathcal{T} - \mathcal{T}_{\mathbf{x}} = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|}(\mathbb{E}[\xi_i] - \xi_i),$$

and by Assumption (7), $\|\xi_i\|_{HS} = \|K_{x_i}\|_H^2 \le \kappa^2$. Applying Lemma 25 with $B' = 2\kappa^2$ and $\sigma' = \kappa^2$, one can prove the desire result.

■

B.7 Proof of Lemma 18

In order to prove Lemma 18, we introduce the following concentration inequality for norms of self-adjoint operators on a Hilbert space.

Lemma 26 *Let $\mathcal{X}_1, \cdots, \mathcal{X}_m$ be a sequence of independently and identically distributed self-adjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $\mathbb{E}[\mathcal{X}_1] = 0$, and $\|\mathcal{X}_1\| \le B$ almost surely for some $B > 0$. Let $\mathcal{V}$ be a positive trace-class operator such that $\mathbb{E}[\mathcal{X}_1^2] \preccurlyeq \mathcal{V}$. Then with probability at least $1-\delta$, ($\delta \in]0,1[$), there holds*

$$\left\|\frac{1}{m}\sum_{i=1}^m \mathcal{X}_i\right\| \le \frac{2B\beta}{3m} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}}, \qquad \beta = \log\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|\delta}.$$

Proof The proof can be found in, e.g., (Rudi et al., 2015; Dicker et al., 2017). Following from the argument in (Minsker, 2011, Section 4), we can generalize (Tropp, 2012, Theorem 7.3.1) from a sequence of self-adjoint matrices to a sequence of self-adjoint Hilbert-Schmidt operators on a separable Hilbert space, and get that for any $t \ge \sqrt{\frac{\|\mathcal{V}\|}{m}} + \frac{B}{3m}$,

$$\Pr\left(\left\|\frac{1}{m}\sum_{i=1}^m \mathcal{X}_i\right\| \ge t\right) \le \frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|}\exp\left(\frac{-mt^2}{2\|\mathcal{V}\| + 2Bt/3}\right). \tag{82}$$

Rewriting

$$\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|}\exp\left(\frac{-mt^2}{2\|\mathcal{V}\|+2Bt/3}\right)=\delta,$$

as a quadratic equation with respect to the variable t, and then solving the quadratic equation, we get

$$t_0=\frac{B\beta}{3m}+\sqrt{\left(\frac{B\beta}{3m}\right)^2+\frac{2\beta\|\mathcal{V}\|}{m}}\leq\frac{2B\beta}{3m}+\sqrt{\frac{2\beta\|\mathcal{V}\|}{m}}:=t^*,$$

where we used $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}, \forall a,b>0$. Note that $\beta>1$, and thus $t_0\geq\sqrt{\frac{\|\mathcal{V}\|}{m}}+\frac{B}{3m}$. By

$$\Pr\left(\left\|\frac{1}{m}\sum_{i=1}^m\mathcal{X}_i\right\|\geq t_*\right)\leq\Pr\left(\left\|\frac{1}{m}\sum_{i=1}^m\mathcal{X}_i\right\|\geq t_0\right),$$

and applying (82) to bound the left-hand side, one can get the desire result. ■

Applying the above lemma, one can prove Lemma 18 as follows.

Proof of Lemma 18 The proof can be also found in (Rudi et al., 2015; Dicker et al., 2017; Hsu et al., 2014). Unlike the result in (Rudi et al., 2015) which requires the condition $\lambda\leq\|\mathcal{T}\|$, our results hold for any $\lambda>0$. We will use Lemma 26 to prove the result. Let $|\mathbf{x}|=m$ and $\mathcal{X}_i=\mathcal{T}_{\tilde{\lambda}}^{-1/2}(\mathcal{T}-\mathcal{T}_{x_i})\mathcal{T}_{\tilde{\lambda}}^{-1/2}$, for all $i\in[m]$. Then $\mathcal{T}_{\tilde{\lambda}}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\tilde{\lambda}}^{-1/2}=\frac{1}{m}\sum_{i=1}^m\mathcal{X}_i$. Obviously, for any $\mathcal{X}=\mathcal{X}_i$, $\mathbb{E}[\mathcal{X}]=0$, and

$$\|\mathcal{X}\|\leq\mathbb{E}\left[\|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\right]+\|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\leq 2\kappa^2/\tilde{\lambda},$$

where for the last inequality, we used Assumption (7) which implies

$$\|\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}\|\leq\operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_x\mathcal{T}_{\tilde{\lambda}}^{-1/2})=\operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T}_x)=\langle\mathcal{T}_{\tilde{\lambda}}^{-1}K_x,K_x\rangle_H\leq\kappa^2/\tilde{\lambda}.$$

Also, by $\mathbb{E}(A-\mathbb{E}A)^2\preccurlyeq\mathbb{E}A^2$,

$$\begin{aligned}\mathbb{E}\mathcal{X}^2 &\preccurlyeq \mathbb{E}(\mathcal{T}_{\tilde{\lambda}}^{-1/2}\mathcal{T}_{\tilde{x}}\mathcal{T}_{\tilde{\lambda}}^{-1/2})^2=\mathbb{E}[\langle\mathcal{T}_{\tilde{\lambda}}^{-1}K_x,K_x\rangle_H\mathcal{T}_{\tilde{\lambda}}^{-1/2}K_x\otimes K_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}]\\ &\preccurlyeq \frac{\kappa^2}{\tilde{\lambda}}\mathbb{E}[\mathcal{T}_{\tilde{\lambda}}^{-1/2}K_x\otimes K_x\mathcal{T}_{\tilde{\lambda}}^{-1/2}]=\frac{\kappa^2}{\tilde{\lambda}}\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T}=\mathcal{V},\end{aligned}$$

Note that $\|\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T}\|=\frac{\|\mathcal{T}\|}{\|\mathcal{T}\|+\tilde{\lambda}}\leq 1$. Therefore, $\|\mathcal{V}\|\leq\frac{\kappa^2}{\tilde{\lambda}}$ and

$$\frac{\operatorname{tr}(\mathcal{V})}{\|\mathcal{V}\|}=\frac{\mathcal{N}(\tilde{\lambda})\|\mathcal{T}\|+\operatorname{tr}(\mathcal{T}_{\tilde{\lambda}}^{-1}\mathcal{T})\tilde{\lambda}}{\|\mathcal{T}\|}\leq\frac{\mathcal{N}(\tilde{\lambda})\|\mathcal{T}\|+\operatorname{tr}(\mathcal{T})}{\|\mathcal{T}\|}\leq\frac{\kappa^2(\mathcal{N}(\tilde{\lambda})+1)}{\|\mathcal{T}\|},$$

where for the last inequality we used (24). Now, the result can be proved by applying Lemma 26. ■

B.8 Proof of Lemma 21

Note that from the independence of $\mathbf{z}_1, \cdots, \mathbf{z}_m$ and (76), we have

$$\mathbb{E}_{\bar{\mathbf{y}}}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho = \frac{1}{m^2}\sum_{s,l=1}^m \mathbb{E}_{\bar{\mathbf{y}}}\langle \mathcal{S}_\rho(g_{s,t} - h_{s,t}), \mathcal{S}_\rho(g_{l,t} - h_{l,t})\rangle_\rho = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}_{\mathbf{y}_s}\|\mathcal{S}_\rho(g_{s,t} - h_{s,t})\|_\rho^2.$$

Taking the expectation with respect to $\bar{\mathbf{x}}$, we get

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_t - \bar{h}_t)\|_\rho = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}\|\mathcal{S}_\rho(g_{s,t} - h_{s,t})\|_\rho^2 = \frac{1}{m}\mathbb{E}\|\mathcal{S}_\rho(g_{1,t} - h_{1,t})\|_\rho^2.$$

The proof is complete. ■

B.9 Proof of Lemma 22

Note that by (75) and from the conditional independence of $\mathbf{J}_s, \cdots \mathbf{J}_m$ (given $\bar{\mathbf{z}}$), we have

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho(\bar{f}_t - \bar{g}_t)\|_\rho = \frac{1}{m^2}\sum_{s,l=1}^m \mathbb{E}_{\mathbf{J}}\langle \mathcal{S}_\rho(f_{s,t} - g_{s,t}), \mathcal{S}_\rho(f_{l,t} - g_{l,t})\rangle_\rho = \frac{1}{m^2}\sum_{s=1}^m \mathbb{E}_{\mathbf{J}_s}\|\mathcal{S}_\rho(f_{s,t} - g_{s,t})\|_\rho^2.$$

Taking the expectation with respect to $\bar{\mathbf{z}}$, we thus prove the desired result. ■

Appendix C. Proofs for Distributed Spectral Algorithms

The proof for distributed SGM in Section 3 involves the analysis for distributed GM. In this section, we will extend our analysis for distributed GM to distributed SA. The proof almost follows along the same lines as the proof for distributed GM in Subsections 6.3 and 6.4, but some of them need some delicate modifications, the reason for which lies in that the qualification τ for GM can be any positive number while it is a fixed constant for a general SA.

C.1 Error Decomposition

We begin with an error decomposition. To introduce the error decomposition, we define an auxiliary function, generated by pseudo-SA as follows.Given a spectral function $\widetilde{G}_\lambda$, for any $s \in [m]$, the function $h_\lambda^{\mathbf{z}_s}$ generated by the pseudo spectral algorithm over $\mathbf{x}_s$ is given by

$$h_\lambda^{\mathbf{z}_s} = \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}_s})\mathcal{L}_{\mathbf{x}_s} f_\rho. \tag{83}$$

The estimator generated by distributed pseudo-spectral algorithm is the averaging over these local estimators,

$$\bar{h}_\lambda^{\bar{\mathbf{z}}} = \frac{1}{m}\sum_{s=1}^m h_\lambda^{\mathbf{z}_s}. \tag{84}$$

We note that the above algorithm can not be implemented in practice as the regression function f_ρ is unknown. From the definition of the regression function, similar to (76), we can prove that

$$\mathbb{E}_{\mathbf{y}_s}[g_\lambda^{\mathbf{z}_s}] = h_\lambda^{\mathbf{z}_s}, \tag{85}$$

and thus

$$\mathbb{E}_{\bar{\mathbf{y}}}[\bar{g}_\lambda^{\bar{\mathbf{z}}}] = \bar{h}_\lambda^{\bar{\mathbf{z}}},$$

Using these basic properties, analogous to Proposition 1, we have the following error decomposition for distributed SA.

Proposition 8 *We have*

$$\mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 = \mathbb{E}\|\bar{h}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 + \mathbb{E}\|\mathcal{S}_\rho \bar{g}_\lambda^{\bar{\mathbf{z}}} - \bar{h}_\lambda^{\bar{\mathbf{z}}}\|_\rho^2.$$

The right-hand side is composed of two terms. The first term is called as bias, and the second term is called as sample variance. In what follows, we will estimate these two terms separably.

C.2 Estimating Bias

Analogous to Lemma 7, we can show that the bias term $\mathbb{E}\|\bar{h}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2$ can be upper bounded in terms of the local bias $\mathbb{E}\|h_\lambda^{\mathbf{z}_1} - f_\rho\|_\rho^2$.

Lemma 27 *We have* $\mathbb{E}\|\bar{h}_\lambda^{\bar{\mathbf{z}}} - f_\rho\|_\rho^2 \le \mathbb{E}\|h_\lambda^{\mathbf{z}_1} - f_\rho\|_\rho^2$.

Proof The proof is the same as that in Lemma 7 by using Hölder's inequality. ∎

In what follows, we will estimate the local bias $\mathbb{E}\|h_\lambda^{\mathbf{z}_1} - f_\rho\|_\rho^2$. Throughout the rest of this subsection, we shall drop the index $s = 1$ for the first local estimator whenever it shows up, i.e., we rewrite $h_\lambda^{\mathbf{z}_1}$ as $h_\lambda^{\mathbf{z}}$, $\mathbf{z}_1$ as $\mathbf{z}$, etc. To do so, we need to introduce a population function defined by

$$\tilde{r}_\lambda = \widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_\rho. \tag{86}$$

The function $\tilde{r}_\lambda$ is deterministic and it is independent from the samples. Since $\widetilde{G}_\lambda(\cdot)$ is a filter function with qualification $\tau > 0$ and constants E, F_τ, similar to Lemma 10, we have the following results for a filter function according to the spectral theorem.

Lemma 28 *Let L be a compact, positive operator on a separable Hilbert space H such that $\|L\| \le \kappa^2$. Then for any $\tilde{\lambda} \ge 0$,*
1) $\|(L + \tilde{\lambda})^\alpha \widetilde{G}_\lambda(L)\| \le E\lambda^{\alpha-1}(1 + (\tilde{\lambda}/\lambda)^\alpha), \ \forall \alpha \in [0, 1]$.
2) $\|(I - L\widetilde{G}_\lambda(L))(L + \tilde{\lambda})^\alpha\| \le F_\tau 2^{(\alpha-1)_+}\lambda^\alpha(1 + (\tilde{\lambda}/\lambda)^\alpha), \quad \forall \alpha \in [0, \tau]$.

With the above lemma, analogous to Lemma 14, we have the following properties for the population function.

Lemma 29 *Under Assumption 2, the following results hold.*
1) For any $\zeta - \tau \le a \le \zeta$, we have

$$\|\mathcal{L}^{-a}\left(\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho\right)\|_\rho \le F_\tau R \lambda^{\zeta - a}.$$

2) We have

$$\|\mathcal{T}^{a-1/2}\tilde{r}_\lambda\|_H \le ER \cdot \begin{cases} \lambda^{\zeta+a-1}, & \text{if } -\zeta \le a \le 1-\zeta, \\ \kappa^{2(\zeta+a-1)}, & \text{if } a \ge 1-\zeta. \end{cases} \tag{87}$$

Note that there is a subtle difference between Lemma 14.(1) and Lemma 29.(1). The latter requires $a \ge \zeta - \tau$ while the former does not, the reason for which is that, the qualification τ is fixed in the latter while it can be any positive constant in the former. This difference makes the proof for SA slightly different to the one for GM, when estimating the bias.
Proof 1) According to the spectral theory,

$$\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* = \mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{S}_\rho^*\mathcal{S}_\rho)\mathcal{S}_\rho^* = \widetilde{G}_\lambda(\mathcal{S}_\rho\mathcal{S}_\rho^*)\mathcal{S}_\rho\mathcal{S}_\rho^* = \widetilde{G}_\lambda(\mathcal{L})\mathcal{L}.$$

Combining with (86), we thus have

$$\mathcal{L}^{-a}(\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho) = \mathcal{L}^{-a}\left(\widetilde{G}_\lambda(\mathcal{L})\mathcal{L} - I\right) f_\rho.$$

Taking the ρ-norm, and applying Assumption 2, we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho)\|_\rho \le \|\mathcal{L}^{\zeta-a}(\widetilde{G}_\lambda(\mathcal{L})\mathcal{L} - I)\| R.$$

Note that the condition (7) implies (24). By a similar argument as that for 2) of Lemma 28, one can prove the first desired result.

2) By (86) and Assumption 2,

$$\|\mathcal{T}^{a-1/2}\tilde{r}_\lambda\|_H = \|\mathcal{T}^{a-1/2}\widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_\rho\|_H \le \|\mathcal{T}^{a-1/2}\widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^\zeta\| R.$$

Noting that

$$\begin{aligned}\|\mathcal{T}^{a-1/2}\widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^\zeta\| &= \|\mathcal{T}^{a-1/2}\widetilde{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\mathcal{L}^{2\zeta}\mathcal{S}_\rho\widetilde{G}_\lambda(\mathcal{T})\mathcal{T}^{a-1/2}\|^{1/2} \\ &= \|\widetilde{G}_\lambda^2(\mathcal{T})\mathcal{T}^{2\zeta+2a}\|^{1/2} = \|\widetilde{G}_\lambda(\mathcal{T})\mathcal{T}^{\zeta+a}\|,\end{aligned}$$

we thus have

$$\|\mathcal{T}^{a-1/2}\tilde{r}_\lambda\|_H \le \|\widetilde{G}_\lambda(\mathcal{T})\mathcal{T}^{\zeta+a}\| R.$$

If $0 \le \zeta + a \le 1$, i.e., $-\zeta \le a \le 1-\zeta$, then using 1) of Lemma 28, we get

$$\|\mathcal{T}^{a-1/2}\tilde{r}_\lambda\|_H \le \lambda^{\zeta+a-1} E R.$$

Similarly, when $a \ge 1-\zeta$, we have

$$\|\mathcal{T}^{a-1/2}\tilde{r}_\lambda\|_H \le \|\widetilde{G}_\lambda(\mathcal{T})\mathcal{T}\|\|\mathcal{T}\|^{\zeta+a-1} R \le \kappa^{2(\zeta+a-1)} E R,$$

where for the last inequality we used 1) of Lemma 28 and (24). This thus proves the second desired result. ■

With the above lemmas, similar to Lemma 15, we have the following analytic result, which enables us to estimate the bias term in terms of several random quantities.

Lemma 30 *Under Assumption 2, let*

$$\Delta_1^{\mathbf{z}} = \|\mathcal{T}_{\tilde\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\tilde\lambda}^{-1/2}\|^2 \vee 1, \qquad \Delta_3^{\mathbf{z}} = \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|$$

and

$$\Delta_2^{\mathbf{z}} = \|\mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}}\tilde{r}_\lambda + \mathcal{T}\tilde{r}_\lambda\|_H.$$

Then the following results hold for any $\tilde\lambda > 0$.
1) For $0 < \zeta \le 1$

$$\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho \le \left(1 \vee \left(\frac{\tilde\lambda}{\lambda}\right)^{\zeta\vee\frac{1}{2}}\right)(C_1' R (\Delta_1^{\mathbf{z}})^{\zeta\vee\frac{1}{2}}\lambda^\zeta + 2E\sqrt{\Delta_1^{\mathbf{z}}}\lambda^{-\frac{1}{2}}\Delta_2^{\mathbf{z}}). \tag{88}$$

2) For $\zeta > 1$,

$$\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho \le \Delta_1^{\mathbf{z}}\left(1 \vee \left(\frac{\tilde\lambda}{\lambda}\right)^{\zeta}\right)(C_2' R\lambda^\zeta + 2E\lambda^{-\frac{1}{2}}\Delta_2^{\mathbf{z}} + C_3' R\lambda^{\frac{1}{2}}(\Delta_3^{\mathbf{z}})^{(\zeta-\frac{1}{2})\wedge 1}). \tag{89}$$

Here, C_1', C_2' and C_3' are positive constants depending only on ζ, κ, E, and F_τ.

The upper bound from (89) is a bit worser than the one in (44).

Proof We can estimate $\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho$ as

$$\begin{aligned}\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{L}_{\mathbf{x}} f_\rho - f_\rho\|_\rho \leq & \| \underbrace{\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})[\mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}}\tilde{r}_\lambda + \mathcal{T}\tilde{r}_\lambda]}_{\textbf{Bias.1}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})[\mathcal{S}_\rho^* f_\rho - \mathcal{T}\tilde{r}_\lambda]}_{\textbf{Bias.2}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho[I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}]\tilde{r}_\lambda}_{\textbf{Bias.3}} \|_\rho \\ & + \| \underbrace{\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho}_{\textbf{Bias.4}} \|_\rho. \end{aligned} \tag{90}$$

In the rest of the proof, we will estimate the four terms of the r.h.s separately.

Estimating Bias.4

Using 1) of Lemma 29 with $a = 0$, we get

$$\|\mathbf{Bias.4}\|_\rho \leq F_\tau R \lambda^\zeta. \tag{91}$$

Estimating Bias.1

By a simple calculation and (47), we know that for any $f \in H$ and any $b \in [0, \frac{1}{2}]$,

$$\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq \|\mathcal{T}_{\tilde{\lambda}}^{1/2} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{b}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-b} \ \mathcal{T}_{\tilde{\lambda}}^{b}\| \|\mathcal{T}_{\tilde{\lambda}}^{-b} f\|_H.$$

Note that by 1) of Lemma 28, with (28),

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2} \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{b}\| \leq E(1 + (\tilde{\lambda}/\lambda)^{b+\frac{1}{2}}) \lambda^{b-\frac{1}{2}},$$

and by Lemma 11, we get

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-b} \ \mathcal{T}_{\tilde{\lambda}}^{b}\| \leq \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}} \ \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|^{2b}.$$

Therefore, for any $f \in H$ and any $b \in [0, \frac{1}{2}]$,

$$\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}}) f\|_\rho \leq (\Delta_1^{\mathbf{z}})^{b+\frac{1}{2}} E(1 + (\tilde{\lambda}/\lambda)^{b+\frac{1}{2}}) \lambda^{b-\frac{1}{2}} \|\mathcal{T}_{\tilde{\lambda}}^{-b} f\|_H. \tag{92}$$

Letting $f = \mathcal{L}_{\mathbf{x}} f_\rho - \mathcal{S}_\rho^* f_\rho - \mathcal{T}_{\mathbf{x}}\tilde{r}_\lambda + \mathcal{T}\tilde{r}_\lambda$ and $b = \frac{1}{2}$ in the above, we get

$$\|\mathbf{Bias.1}\|_\rho \leq E(1 + \sqrt{\tilde{\lambda}/\lambda}) \lambda^{-\frac{1}{2}} \sqrt{\Delta_1^{\mathbf{z}}} \Delta_2^{\mathbf{z}}. \tag{93}$$

Estimating Bias.2

Thus, letting $f = \mathcal{T}\tilde{r}_\lambda - \mathcal{S}_\rho^* f_\rho$, in (92), we have

$$\begin{aligned}\|\mathbf{Bias.2}\|_\rho &\leq E\|\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}}\|^{2b+1} (1 + (\tilde{\lambda}/\lambda)^{b+\frac{1}{2}}) \lambda^{b-\frac{1}{2}} \|\mathcal{T}_{\tilde{\lambda}}^{-b}[\mathcal{T}\tilde{r}_\lambda - \mathcal{S}_\rho^* f_\rho]\|_H \\ &\leq E(\Delta_1^{\mathbf{z}})^{b+\frac{1}{2}} (1 + (\tilde{\lambda}/\lambda)^{b+\frac{1}{2}}) \lambda^{b-\frac{1}{2}} \|\mathcal{L}_{\tilde{\lambda}}^{-b+\frac{1}{2}}[\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho]\|_H.\end{aligned}$$

When $\zeta \leq \frac{1}{2}$, we have $\tau - \zeta \geq \frac{1}{2}$ since $\tau \geq 1$. Letting $b = 0$, and applying Lemma 29.(1) with $a = -\frac{1}{2}$, we get

$$\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})[\mathcal{T}\tilde{r}_\lambda - \mathcal{S}_\rho^* f_\rho]\|_\rho \leq E F_\tau R (\Delta_1^{\mathbf{z}})^{\frac{1}{2}} (1 + (\tilde{\lambda}/\lambda)^{\frac{1}{2}}) \lambda^\zeta.$$

Similarly, when $\frac{1}{2} \le \zeta \le 1$, we choose $b = \zeta - \frac{1}{2}$, and applying Lemma 29.(1) with $a = \zeta - 1$, we get

$$\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})[\mathcal{T}\tilde{r}_\lambda - \mathcal{S}_\rho^* f_\rho]]\|_\rho \le EF_\tau R(\Delta_1^{\mathbf{z}})^\zeta (1 + (\tilde{\lambda}/\lambda)^\zeta)\lambda^\zeta.$$

When $\zeta \ge 1$, we choose $b = \frac{1}{2}$, and applying Lemma 29.(1) with $a = 0$, we get

$$\|\mathcal{S}_\rho \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})[\mathcal{T}\tilde{r}_\lambda - \mathcal{S}_\rho^* f_\rho]]\|_\rho \le EF_\tau R\Delta_1^{\mathbf{z}}(1 + (\tilde{\lambda}/\lambda))\lambda^\zeta.$$

From the above estimate, we get

$$\|\mathbf{Bias.2}\|_\rho \le EF_\tau R\lambda^\zeta \times \begin{cases} (1 + (\tilde{\lambda}/\lambda)^{1/2})(\Delta_1^{\mathbf{z}})^{1/2} & \text{if } 0 < \zeta \le 1/2, \\ (1 + (\tilde{\lambda}/\lambda)^\zeta)(\Delta_1^{\mathbf{z}})^\zeta & \text{if } 1/2 < \zeta \le 1, \\ (1 + \tilde{\lambda}/\lambda)\Delta_1^{\mathbf{z}} & \text{if } \zeta > 1. \end{cases} \tag{94}$$

Estimating Bias.3
When $\zeta \le 1/2$, by a simple calculation and (47), we have

$$\begin{aligned}\|\mathbf{Bias.3}\|_\rho \le & \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\tilde{\lambda}}^{1/2}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-1/2}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\| \|\tilde{r}_\lambda\|_H \\ \le & \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\| \|\tilde{r}_\lambda\|_H,\end{aligned}$$

By 2) of Lemma 28, with (28),

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{1/2}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\| \le F_\tau(1 + \sqrt{\tilde{\lambda}/\lambda})\sqrt{\lambda}, \tag{95}$$

and by 2) of Lemma 29, $\|\tilde{r}_\lambda\|_H \le ER\lambda^{\zeta - 1/2}$. It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}}(1 + \sqrt{\tilde{\lambda}/\lambda})EF_\tau R\lambda^\zeta.$$

When $1/2 < \zeta \le 1$, by a simple computation, we have

$$\|\mathbf{Bias.3}\|_\rho \le \|\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-\frac{1}{2}}\| \|\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2} - \zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| \|\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2} - \zeta}\tilde{r}_\lambda\|_H.$$

Applying (47) and 2) of Lemma 29, we have

$$\|\mathbf{Bias.3}\|_\rho \le \sqrt{\Delta_1^{\mathbf{z}}} \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2} - \zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| ER.$$

By 2) of Lemma 28,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| \le F_\tau(1 + (\tilde{\lambda}/\lambda)^\zeta)\lambda^\zeta.$$

Besides, by $\zeta \le 1$ and Lemma 11,

$$\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2} - \zeta}\mathcal{T}_{\tilde{\lambda}}^{\zeta - \frac{1}{2}}\| = \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}(2\zeta - 1)}\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}(2\zeta - 1)}\| \le \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}}\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}}\|^{2\zeta - 1} \le (\Delta_1^{\mathbf{z}})^{\zeta - \frac{1}{2}}.$$

It thus follows that

$$\|\mathbf{Bias.3}\|_\rho \le (\Delta_1^{\mathbf{z}})^\zeta (1 + (\tilde{\lambda}/\lambda)^\zeta)EF_\tau R\lambda^\zeta.$$

When $\zeta > 1$, we rewrite **Bias.3** as

$$\mathcal{S}_\rho \mathcal{T}_{\tilde{\lambda}}^{-\frac{1}{2}} \cdot \mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}} \cdot \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}} (I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})(\mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}} + \mathcal{T}^{\zeta-\frac{1}{2}} - \mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}})\mathcal{T}^{\frac{1}{2}-\zeta}\tilde{r}_\lambda.$$

By a simple calculation and (47), we can upper bound $\|\mathbf{Bias.3}\|_\rho$ by

$$\leq \|\mathcal{T}_{\tilde{\lambda}}^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{-\frac{1}{2}}\|(\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})\|\|\mathcal{T}^{\zeta-\frac{1}{2}} - \mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}}\|)\|\mathcal{T}^{\frac{1}{2}-\zeta}\tilde{r}_\lambda\|_H.$$

Introducing with (95), and applying 2) of Lemma 29,

$$\|\mathbf{Bias.3}\|_\rho \leq \sqrt{\Delta_1^\mathbf{z}}(\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x}^{\zeta-1/2}\| + F_\tau(\sqrt{\tilde{\lambda}/\lambda}+1)\sqrt{\lambda}\|\mathcal{T}^{\zeta-1/2} - \mathcal{T}_\mathbf{x}^{\zeta-1/2}\|)ER.$$

By 2) of Lemma 28,

$$\begin{aligned}\|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\frac{1}{2}}(I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}}\| \leq& \|\mathcal{T}_{\mathbf{x}\tilde{\lambda}}^{\zeta}(I - \widetilde{G}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_\mathbf{x})\| \\ \leq& 2^{\zeta-1}F_\tau(1 + (\tilde{\lambda}/\lambda)^\zeta)\lambda^\zeta.\end{aligned}$$

Moreover, by Lemma 13 and $\max(\|\mathcal{T}\|, \|\mathcal{T}_\mathbf{x}\|) \leq \kappa^2$,

$$\|\mathcal{T}^{\zeta-\frac{1}{2}} - \mathcal{T}_\mathbf{x}^{\zeta-\frac{1}{2}}\| \leq (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}\|\mathcal{T} - \mathcal{T}_\mathbf{x}\|^{(\zeta-\frac{1}{2})\wedge 1}.$$

Therefore, when $\zeta > 1$, **Bias.3** can be estimated as

$$\|\mathbf{Bias.3}\|_\rho \leq \sqrt{\Delta_1^\mathbf{z}}\left(2^{\zeta-1}(1 + (\tilde{\lambda}/\lambda)^\zeta)\lambda^\zeta + (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(\sqrt{\tilde{\lambda}/\lambda}+1)\sqrt{\lambda}(\Delta_3^\mathbf{z})^{(\zeta-\frac{1}{2})\wedge 1}\right)EF_\tau R.$$

From the above analysis, we know that $\|\mathbf{Bias.3}\|_\rho$ can be upper bounded by

$$EF_\tau R\begin{cases}\sqrt{\Delta_1^\mathbf{z}}(\sqrt{\tilde{\lambda}/\lambda}+1)\lambda^\zeta, & \text{if } \zeta \in]0, 1/2], \\ (\Delta_1^\mathbf{z})^\zeta((\tilde{\lambda}/\lambda)^\zeta + 1)\lambda^\zeta, & \text{if } \zeta \in]1/2, 1], \\ \sqrt{\Delta_1^\mathbf{z}}\left(2^{\zeta-1}(1 + (\tilde{\lambda}/\lambda)^\zeta)\lambda^\zeta + (2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}(\sqrt{\tilde{\lambda}/\lambda}+1)\sqrt{\lambda}(\Delta_3^\mathbf{z})^{(\zeta-\frac{1}{2})\wedge 1}\right), & \text{if } \zeta \in]1, \infty[.\end{cases} \tag{96}$$

Introducing (91), (93) (94) and (96) into (90), and by a simple calculation, one can prove the desired results with

$$C_1' = F_\tau(1 + 4E),$$
$$C_2' = F_\tau\left(1 + 2E + 2^\zeta E\right),$$
$$\text{and} \quad C_3' = 2EF_\tau(2\zeta\kappa^{2\zeta-3})^{\mathbf{1}_{\{2\zeta\geq 3\}}}.$$

■

The rest of the proofs parallelize as those for distributed GM.

Proposition 9 *Under Assumptions 2 and 3, we let $\tilde{\lambda} = n^{-1+\theta}$ for some $\theta \in [0, 1]$. Then the following results hold.*
1) For $0 < \zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_\lambda^\mathbf{z} - f_\rho\|_\rho^2 \leq C_5'\left(R + \mathbf{1}_{\{2\zeta<1\}}\|f_\rho\|_\infty\right)^2\left(1 \vee [\gamma(\theta^{-1}\wedge\log n)]^{2\zeta\vee 1} \vee \frac{\tilde{\lambda}^2}{\lambda^2}\right)\lambda^{2\zeta}$$

2)For $\zeta > 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho^2 \le C_6' R^2 \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda^{2\zeta}} \vee \lambda^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1} \vee [\gamma(\theta^{-1}\wedge \log n)]^2\right)\lambda^{2\zeta}.$$

Here, C_5' and C_6' are positive constants depending only on $\kappa, \zeta, E, F_\tau, c_\gamma, \|\mathcal{T}\|$ and can be given explicitly in the proof.

Proof We will use Lemma 30 to prove the results. To do so, we need to estimate $\Delta_1^{\mathbf{z}}$, $\Delta_2^{\mathbf{z}}$ and $\Delta_3^{\mathbf{z}}$.

By Lemma 19, we have that with probability at least $1-\delta$, (58) holds, where $a_{n,\delta,\gamma}(1-\theta) = a_{n,\delta,\gamma}(2/3, 1-\theta)$ is given by (54). By Lemma 16, we have that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \le 2\kappa\left(\frac{2\|\tilde{r}_\lambda - f_\rho\|_\infty}{n} + \frac{\|\mathcal{S}_\rho \tilde{r}_\lambda - f_\rho\|_\rho}{\sqrt{n}}\right)\log\frac{2}{\delta}.$$

Applying Lemma 29 with $a=0$ to estimate $\|\mathcal{S}_\rho\tilde{r}_\lambda - f_\rho\|_\rho$, we get that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \le 2\kappa\left(2\|\tilde{r}_\lambda - f_\rho\|_\infty/n + F_\tau R\lambda^\zeta/\sqrt{n}\right)\log\frac{2}{\delta}.$$

When $\zeta \ge 1/2$, we know that there exists some $f_H = \mathcal{T}^{\zeta-1}\mathcal{S}_\rho^*\mathcal{L}^{-\zeta} f_\rho \in H$ such that $\mathcal{S}_\rho f_H = f_\rho$ (Steinwart and Christmann, 2008) and

$$\|\tilde{r}_\lambda - f_\rho\|_\infty \le \kappa\|\tilde{r}_\lambda - f_H\|_H \le \kappa F_\tau R\lambda^{\zeta-1/2},$$

where for the last inequality, we used Lemma 28. When $\zeta < 1/2$, by 2) of Lemma 29, $\|\tilde{r}_\lambda\|_H \le ER\lambda^{\zeta-1/2}$, which thus lead to

$$\|\tilde{r}_\lambda - f_\rho\|_\infty \le \kappa\|\tilde{r}_\lambda\|_H + \|f_\rho\|_\infty \le \kappa E R\lambda^{\zeta-1/2} + \|f_\rho\|_\infty.$$

From the above analysis, we get that with probability at least $1-\delta$,

$$\Delta_2^{\mathbf{z}} \le \log\frac{2}{\delta}\begin{cases} 2\kappa F_\tau R\big(2\kappa/(\lambda n) + 1/\sqrt{\lambda n}\big)\lambda^{\zeta+1/2}, & \text{if } \zeta \ge 1/2,\\ 2\kappa\big(2\kappa E R/(\lambda n) + 2\|f_\rho\|_\infty(n\lambda)^{-\zeta-1/2} + F_\tau R/\sqrt{n\lambda}\big)\lambda^{\zeta+1/2}, & \text{if } \zeta < 1/2,\end{cases}$$

which can be further relaxed as

$$\Delta_2^{\mathbf{z}} \le C_4'\widetilde{R}(1\vee(\lambda n)^{-1})\lambda^{\zeta+1/2}\log\frac{2}{\delta}, \quad \widetilde{R} = R + \mathbf{1}_{\{2\zeta<1\}}\|f_\rho\|_\infty, \tag{97}$$

where

$$C_4' \le \begin{cases} 2\kappa F_\tau\big(2\kappa+1\big), & \text{if } \zeta \ge 1/2,\\ 2\kappa\big(2\kappa E + 2 + F_\tau\big), & \text{if } \zeta < 1/2.\end{cases}$$

Applying Lemma 17, we have that with probability at least $1-\delta$, (60) holds.

For $0 < \zeta \le 1$, by Lemma 30, (58) and (97), we have that with probability at least $1-2\delta$,

$$\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho \le \left(3^{\zeta\vee\frac{1}{2}}C_1' R a_{n,\delta,\gamma}^{\zeta\vee\frac{1}{2}}(1-\theta) + 2\sqrt{3}EC_4'\widetilde{R}a_{n,\delta,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{2}{\delta}\right)\left(1\vee\left(\frac{\tilde{\lambda}}{\lambda}\right)^{\zeta\vee\frac{1}{2}}\vee\frac{1}{n\lambda}\right)\lambda^\zeta.$$

Rescaling δ, and then combining with Lemma 20, we get

$$\mathbb{E}\|\mathcal{S}_\rho h_{t+1} - f_\rho\|_\rho^2$$
$$\leq \widetilde{R}^2 \int_0^1 \left(3^{\zeta\vee\frac{1}{2}} C_1' a_{n,\delta/2,\gamma}^{\zeta\vee\frac{1}{2}}(1-\theta) + 2\sqrt{3}EC_4' a_{n,\delta/2,\gamma}^{\frac{1}{2}}(1-\theta)\log\frac{4}{\delta} \right)^2 d\delta \left(1 \vee \left(\frac{\tilde{\lambda}}{\lambda}\right)^{2\zeta\vee 1} \vee \frac{1}{n^2\lambda^2} \right) \lambda^{2\zeta}.$$

By a direct computation and noting that $\tilde{\lambda} \geq n^{-1}$ and $\zeta \leq 1$, one can prove the first desired result with $A = \log\frac{8\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|}$, and

$$C_5' = 2[C_1'^2(48\kappa^2)^{2\zeta\vee 1}(A^{2\zeta\vee 1} + \Gamma(3)) + 192\kappa^2 C_4'^2 E^2(A(\log^2 4 + 2 + 2\log 4) + \log^2 4 + 4\log 4 + 6)].$$

For $\zeta > 1$, by Lemma 30, (58), (97) and (60), we know that with probability at least $1 - 3\delta$,

$$\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho$$
$$\leq 3R(C_2' + 2EC_4' + 6\kappa^2 C_3')a_{n,\delta,\gamma}(1-\theta)\log\frac{2}{\delta}\left(1 \vee \frac{\tilde{\lambda}^\zeta}{\lambda^\zeta} \vee \frac{1}{n\lambda} \vee \lambda^{\frac{1}{2}-\zeta}\left(\frac{1}{n}\right)^{\frac{(\zeta-\frac{1}{2})\wedge 1}{2}} \right)\lambda^\zeta.$$

Rescaling δ, and applying Lemma 20, we get

$$\mathbb{E}\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho^2$$
$$\leq 9R^2(C_2' + 2EC_4' + 6\kappa^2 C_3')^2 \int_0^1 a_{n,\delta/3,\gamma}^2(1-\theta)\log^2\frac{6}{\delta} d\delta \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda^{2\zeta}} \vee \frac{1}{n^2\lambda^2} \vee \lambda^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1} \right)\lambda^{2\zeta},$$

which leads to the second desired result with

$$C_6' = 24^3\kappa^4(C_2' + 2EC_4' + 6\kappa^2 C_3')^2(A+1)^2(\log 6 + 1)^2, \; A = \log\frac{12\kappa^2(c_\gamma+1)\mathrm{e}}{\|\mathcal{T}\|},$$

by noting that $\tilde{\lambda} \geq n^{-1}$ and $\zeta \geq 1$. The proof is complete. ■

Combining Proposition 9 with Lemma 27, we get the following results for the bias of the fully averaged estimators.

Proposition 10 *Under Assumptions 2 and 3, for any $\tilde{\lambda} = n^{-1+\theta}$ with $\theta \in [0,1]$, the following results hold.*
1) For $\zeta \leq 1$,

$$\mathbb{E}\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho^2 \leq C_5'\left(R + \mathbf{1}_{2\zeta<1}\|f_\rho\|_\infty\right)^2 \left(1 \vee [\gamma(\theta^{-1}\wedge\log n)]^{2\zeta\vee 1} \vee \frac{\tilde{\lambda}^2}{\lambda^2} \right)\lambda^{2\zeta}. \tag{98}$$

2) For $1 < \zeta \leq \tau$,

$$\mathbb{E}\|\mathcal{S}_\rho h_\lambda^{\mathbf{z}} - f_\rho\|_\rho^2 \leq C_6' R^2 \left(1 \vee \frac{\tilde{\lambda}^{2\zeta}}{\lambda^{2\zeta}} \vee \lambda^{1-2\zeta}\left(\frac{1}{n}\right)^{(\zeta-\frac{1}{2})\wedge 1} \vee [\gamma(\theta^{-1}\wedge\log n)]^2 \right)\lambda^{2\zeta} \tag{99}$$

Here, C_5' and C_6' are given by Proposition 9.

C.3 Estimating Sample Variance

In this section, we estimate sample variance $\|\mathcal{S}_\rho(\bar{g}_\lambda^{\bar{\mathbf{z}}} - \bar{h}_\lambda^{\bar{\mathbf{z}}})\|_\rho$. We first introduce the following lemma.

Lemma 31 *We have*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar{g}_\lambda^{\bar{\mathbf{z}}} - \bar{h}_\lambda^{\bar{\mathbf{z}}})\|_\rho = \frac{1}{m}\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}_1} - h_\lambda^{\mathbf{z}_1})\|_\rho^2. \tag{100}$$

Proof The proof is the same as that in Lemma 21 by applying (85). ■

According to Lemma 31, we know that the sample variance of the averaging over m local estimators can be well controlled in terms of the sample variance of a local estimator. In what follows, we will estimate the local sample variance, $\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}_1} - h_\lambda^{\mathbf{z}_1})\|_\rho^2$. Throughout the rest of this subsection, we shall drop the index $s = 1$ and write $\mathbf{z}_1$ as $\mathbf{z}$, $\mathbf{x}_1$ as $\mathbf{x}$.

Proposition 11 *Under Assumption 3, let $\tilde{\lambda} = n^{\theta-1}$ for some $\theta \in [0,1]$. Then*

$$\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}} - h_\lambda^{\mathbf{z}})\|_\rho^2 \leq C_8' \frac{\sigma^2}{n\tilde{\lambda}^\gamma}\left(1 \vee \frac{\tilde{\lambda}}{\lambda} \vee [\gamma(\theta^{-1} \wedge \log n])\right).$$

Here, C_8' is a positive constant depending only on $\kappa, c_\gamma, \|\mathcal{T}\|, E$ and will be given explicitly in the proof.

Proof For notational simplicity, we let $\epsilon_i = y_i - f_\rho(x_i)$ for all $i \in [n]$ and $\boldsymbol{\epsilon} = (\epsilon_i)_{1\leq i\leq n}$. Then from the definitions of $h_\lambda^{\mathbf{z}_s}$ and $g_\lambda^{\mathbf{z}_s}$

$$g_\lambda^{\mathbf{z}} - h_\lambda^{\mathbf{z}} = \widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}.$$

Using the above relationship and the isometric property (25), we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 &= \mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_\rho^2 \\
&= \mathbb{E}_{\mathbf{y}}\|\mathcal{T}^{1/2}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\boldsymbol{\epsilon}\|_H^2 \\
&= \frac{1}{n^2}\sum_{l,k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k]\operatorname{tr}\left(\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})K_{x_l}\otimes K_{x_k}\right).
\end{aligned}$$

From the definition of f_ρ and the independence of z_l and z_k when $l \neq k$, we know that $\mathbb{E}_{\mathbf{y}}[\epsilon_l\epsilon_k] = 0$ whenever $l \neq k$. Therefore,

$$\mathbb{E}_{\mathbf{y}}\|\mathcal{S}_\rho(g_{t+1} - h_{t+1})\|_\rho^2 = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}_{\mathbf{y}}[\epsilon_k^2]\operatorname{tr}\left(\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right).$$

Using Assumption 1,

$$\begin{aligned}\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}} - h_\lambda^{\mathbf{z}})\|_\rho^2 \leq & \frac{\sigma^2}{n^2}\sum_{k=1}^n \operatorname{tr}\left(\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})K_{x_k}\otimes K_{x_k}\right)\\ = & \frac{\sigma^2}{n}\operatorname{tr}\left(\mathcal{T}(\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}}))^2\mathcal{T}_{\mathbf{x}}\right)\\ \leq & \frac{\sigma^2}{n}\operatorname{tr}(\mathcal{T}_{\tilde\lambda}^{-1/2}\mathcal{T}\mathcal{T}_{\tilde\lambda}^{1/2})\|\mathcal{T}_{\tilde\lambda}^{1/2}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\tilde\lambda}^{1/2}\|\\ \leq & \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\Delta_1^{\mathbf{z}}\|\mathcal{T}_{\mathbf{x}\tilde\lambda}^{1/2}\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})^2\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\tilde\lambda}^{1/2}\|\\ \leq & \frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\Delta_1^{\mathbf{z}}\|\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}\|(\|\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}\| + \tilde\lambda\|\widetilde{G}_\lambda(\mathcal{T}_{\mathbf{x}})\|)\\ \leq & E^2\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}\Delta_1^{\mathbf{z}}(1+\tilde\lambda/\lambda),\end{aligned}$$

where for the last inequality, we used 1) of Lemma 28. Taking the expectation with respect to $\mathbf{x}$, this leads to

$$\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}} - h_\lambda^{\mathbf{z}})\|_\rho^2 \leq E^2\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1+\tilde\lambda/\lambda)\mathbb{E}[\Delta_1^{\mathbf{z}}].$$

Applying Lemmas 19 and 20, we get

$$\begin{aligned}\mathbb{E}\|\mathcal{S}_\rho(g_\lambda^{\mathbf{z}} - h_\lambda^{\mathbf{z}})\|_\rho^2 \leq & 6E^2\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1\vee(\tilde\lambda/\lambda))\int_0^1 a_{n,\delta,\gamma}(2/3, 1-\theta)d\delta\\ \leq & C_7'\frac{\sigma^2\mathcal{N}(\tilde\lambda)}{n}(1\vee(\tilde\lambda/\lambda)\vee[\gamma(\theta^{-1}\wedge\log n)]),\end{aligned}$$

where $C_7' = 48E^2\kappa^2\log\frac{4\kappa^2(c_\gamma+1)\mathrm{e}^2}{\|\mathcal{T}\|}$. Using Assumption 3, we get the desired result with

$$C_8' = c_\gamma C_7.$$

■

Using the above proposition and Lemma 31, we derive the following results for sample variance.

Proposition 12 *Under Assumption 3, let $\tilde\lambda = n^{\theta-1}$ for some $\theta\in[0,1]$. Then for any $t\in[T]$,*

$$\mathbb{E}\|\mathcal{S}_\rho(\bar g_\lambda^{\bar{\mathbf{z}}} - \bar h_\lambda^{\bar{\mathbf{z}}})\|_\rho^2 \leq C_8'\frac{\sigma^2}{N\tilde\lambda^\gamma}\left(1\vee\left(\frac{\tilde\lambda}{\lambda}\right)\vee[\gamma(\theta^{-1}\wedge\log n)]\right), \tag{101}$$

where C_8' is given by Proposition 11.

C.4 Deriving Total Error Bounds

Proof of Theorem 2 The proof can be finished by simply applying Propositions 12 and 10 into Proposition 8. ■

Proof of Corollary 6 The results are direct consequences of Theorem 2. ■