Journal of Machine Learning Research 18 (2017) 1-47 Submitted 3/17; Revised 8/17; Published 10/17

Optimal Rates for Multi-pass Stochastic Gradient Methods

Junhong Lin* JUNHONG.LIN@IIT.IT
Laboratory for Computational and Statistical Learning
Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Lorenzo Rosasco LROSASCO@MIT.EDU
DIBRIS, Università di Genova
Via Dodecaneso, 35 — 16146 Genova, Italy
Laboratory for Computational and Statistical Learning
Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Editor: Leon Bottou

Abstract

We analyze the learning properties of the stochastic gradient method when multiple passes over the data and mini-batches are allowed. We study how regularization properties are controlled by the step-size, the number of passes and the mini-batch size. In particular, we consider the square loss and show that for a universal step-size choice, the number of passes acts as a regularization parameter, and optimal finite sample bounds can be achieved by early-stopping. Moreover, we show that larger step-sizes are allowed when considering mini-batches. Our analysis is based on a unifying approach, encompassing both batch and stochastic gradient methods as special cases. As a byproduct, we derive optimal convergence results for batch gradient methods (even in the non-attainable cases).

1. Introduction

Modern machine learning applications require computational approaches that are at the same time statistically accurate and numerically efficient (Bousquet and Bottou, 2008). This has motivated a recent interest in stochastic gradient methods (SGM), since on the one hand they enjoy good practical performances, especially in large scale scenarios, and on the other hand they are amenable to theoretical studies. In particular, unlike other learning approaches, such as empirical risk minimization or Tikhonov regularization, theoretical results on SGM naturally integrate statistical and computational aspects.

Most generalization studies on SGM consider the case where only one pass over the data is allowed and the step-size is appropriately chosen, see (Cesa-Bianchi et al., 2004; Nemirovski et al., 2009; Ying and Pontil, 2008; Tarres and Yao, 2014; Dieuleveut and Bach, 2016; Orabona, 2014) and references therein, possibly considering averaging (Poljak, 1987). In particular, recent works show how the step-size can be seen to play the role of a regularization parameter whose choice controls the bias and variance properties of the obtained solution (Ying and Pontil, 2008; Tarres and Yao, 2014; Dieuleveut and Bach, 2016; Lin et al., 2016a). These latter works show that balancing these contributions, it is possible to derive a step-size choice leading to optimal

*. J.L. is now with the Swiss Federal Institute of Technology in Lausanne (jhlin5@hotmail.com)

learning bounds. Such a choice typically depends on some unknown properties of the data generating distributions and it can be chosen by cross-validation in practice.

While processing each data point only once is natural in streaming/online scenarios, in practice SGM is often used to process large data-sets and multiple passes over the data are typically considered. In this case, the number of passes over the data, as well as the step-size, need then to be determined. While the role of multiple passes is well understood if the goal is empirical risk minimization (see e.g., Boyd and Mutapcic, 2007), its effect with respect to generalization is less clear. A few recent works have recently started to tackle this question. In particular, results in this direction have been derived in (Hardt et al., 2016) and (Lin et al., 2016a). The former work considers a general stochastic optimization setting and studies stability properties of SGM allowing to derive convergence results as well as finite sample bounds. The latter work, restricted to supervised learning, further develops these results to compare the respective roles of step-size and number of passes, and show how different parameter settings can lead to optimal error bounds. In particular, it shows that there are two extreme cases: while one between the step-size or the number of passes is fixed a priori, while the other one acts as a regularization parameter and needs to be chosen adaptively. The main shortcoming of these latter results is that they are for the worst case, in the sense that they do not consider the possible effect of benign assumptions on the problem (Zhang, 2005; Caponnetto and De Vito, 2007) that can lead to faster rates for other learning approaches such as Tikhonov regularization. Further, these results do not consider the possible effect on generalization of mini-batches, rather than a single point in each gradient step (Shalev-Shwartz et al., 2011; Dekel et al., 2012; Sra et al., 2012; Ng, 2016). This latter strategy is often considered especially for parallel implementation of SGM.

The study in this paper fills in these gaps in the case where the loss function is the least squares loss. We consider a variant of SGM for least squares, where gradients are sampled uniformly at random and mini-batches are allowed. The number of passes, the step-size and the mini-batch size are then parameters to be determined. Our main results highlight the respective roles of these parameters and show how can they be chosen so that the corresponding solutions achieve optimal learning errors in a variety of settings. In particular, we show for the first time that multi-pass SGM with early stopping and a universal step-size choice can achieve optimal convergence rates, matching those of ridge regression (Smale and Zhou, 2007; Caponnetto and De Vito, 2007). Further, our analysis shows how the mini-batch size and the step-size choice are tightly related. Indeed, larger mini-batch sizes allow considering larger step-sizes while keeping the optimal learning bounds. This result gives insights on how to exploit mini-batches for parallel computations while preserving optimal statistical accuracy. Finally, we note that a recent work (Rosasco and Villa, 2015) is related to the analysis in the paper. The generalization properties of a multi-pass incremental gradient are analyzed in (Rosasco and Villa, 2015), for a cyclic, rather than a stochastic, choice of the gradients and with no mini-batches. The analysis in this latter case appears to be harder and results in (Rosasco and Villa, 2015) give good learning bounds only in restricted setting and considering iterates rather than the excess risk. Compared to (Rosasco and Villa, 2015) our results show how stochasticity can be exploited to get fast rates and analyze the role of mini-batches. The basic idea of our proof is to approximate the SGM learning sequence in terms of the batch gradient descent sequence, see Subsection 3.7 for further details. This allows to study batch and stochastic gradient methods simultaneously, and may be also useful for analyzing other learning algorithms.

This paper is an extended version of a prior conference paper (Lin and Rosasco, 2016). In (Lin and Rosasco, 2016), we give convergence results with optimal rates for the attainable case (i.e., assuming the existence of at least one minimizer of the expected risk over the hypothesis space) in a fixed step-size setting. In this new version, we give convergence results with optimal rates, for both the attainable and non-attainable cases, and consider more general step-size choices. The extension from the attainable case to the non-attainable case is non-trivial. As will be seen from the proof, in contrast to the attainable case, a different and refined estimation is needed for the non-attainable case. Interestingly, as a byproduct of this paper, we also derived optimal rates for the batch gradient descent methods in the non-attainable case. To the best of our knowledge, such a result may be the first kind for batch gradient methods, without requiring any extra unlabeled data as that in (Caponnetto and Yao, 2010). Finally, we also add novel convergence results for the iterates showing that they converge to the minimal norm solution of the expected risk with optimal rates.

The rest of this paper is organized as follows. Section 2 introduces the learning setting and the SGM algorithm. Main results with discussions and proof sketches are presented in Section 3. Preliminary lemmas necessary for the proofs will be given in Section 4 while detailed proofs will be conducted in Sections 5 to 8. Finally, simple numerical simulations are given in Section 9 to complement our theoretical results.

Notation For any $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum of a and b. $\mathbb{N}$ is the set of all positive integers. For any $T \in \mathbb{N}$, $[T]$ denotes the set $\{1, \cdots, T\}$. For any two positive sequences $\{a_t\}_{t\in[T]}$ and $\{b_t\}_{t\in[T]}$, the notation $a_t \lesssim b_t$ for all $t \in [T]$ means that there exists a positive constant $C \geq 0$ such that C is independent of t and that $a_t \leq C b_t$ for all $t \in [T]$.

2. Learning with SGM

We begin by introducing the learning setting we consider, and then describe the SGM learning algorithm. Following (Rosasco and Villa, 2015), the formulation we consider is close to the setting of functional regression, and covers the reproducing kernel Hilbert space (RKHS) setting as a special case, see Appendix A. In particular, it reduces to standard linear regression for finite dimensions.

2.1 Learning Problems

Let H be a separable Hilbert space, with inner product and induced norm denoted by $\langle \cdot, \cdot \rangle_H$ and $\| \cdot \|_H$, respectively. Let the input space $X \subseteq H$ and the output space $Y \subseteq \mathbb{R}$. Let ρ be an unknown probability measure on $Z = X \times Y$, $\rho_X(\cdot)$ the induced marginal measure on X, and $\rho(\cdot|x)$ the conditional probability measure on Y with respect to $x \in X$ and ρ.

Considering the square loss function, the problem under study is the minimization of the *risk*,

$$\inf_{\omega \in H} \mathcal{E}(\omega), \quad \mathcal{E}(\omega) = \int_{X \times Y} (\langle \omega, x \rangle_H - y)^2 d\rho(x, y), \tag{1}$$

when the measure ρ is known only through a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ of size $m \in \mathbb{N}$, independently and identically distributed (i.i.d.) according to ρ. In the following, we measure the quality of an approximate solution $\hat{\omega} \in H$ (an estimator) considering *the excess risk*, i.e.,

$$\mathcal{E}(\hat{\omega}) - \inf_{\omega \in H} \mathcal{E}(\omega). \tag{2}$$

Throughout this paper, we assume that there exists a constant $\kappa \in [1, \infty[$, such that

$$\langle x, x' \rangle_H \leq \kappa^2, \quad \forall x, x' \in X. \tag{3}$$

2.2 Stochastic Gradient Method

We study the following variant of SGM, possibly with mini-batches. Unlike some of the variants studied in the literature, the algorithm we consider in this paper does not involve any explicit penalty term or any projection step, in which case one does not need to tune the penalty/projection parameter.

Algorithm 1 *Let $b \in [m]$. Given any sample $\mathbf{z}$, the b-minibatch stochastic gradient method is defined by $\omega_1 = 0$ and*

$$\omega_{t+1} = \omega_t - \eta_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} (\langle \omega_t, x_{j_i} \rangle_H - y_{j_i}) x_{j_i}, \qquad t = 1, \ldots, T, \tag{4}$$

where $\{\eta_t > 0\}$ is a step-size sequence. Here, $j_1, j_2, \cdots, j_{bT}$ are i.i.d. random variables from the uniform distribution on $[m]$ [1].

We add some comments on the above algorithm. First, different choices for the mini-batch size b can lead to different algorithms. In particular, for $b = 1$, the above algorithm corresponds to a simple SGM, while for $b = m$, it is a stochastic version of the batch gradient descent. In this paper, we are particularly interested in the cases of $b = 1$ and $b = \sqrt{m}$. Second, other choices on the initial value, rather than $\omega_1 = 0$, is possible. In fact, following from our proofs in this paper, the interested readers can see that the convergence results stated in the next subsections still hold for other choices of initial values. Finally, the number of total iterations T can be bigger than the number of sample points m. This indicates that we can use the sample 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 'passes' over the data is referred to $\lceil \frac{bt}{m} \rceil$ at t iterations of the algorithm.

The aim of this paper is to derive excess risk bounds for Algorithm 1. Throughout this paper, we assume that $\{\eta_t\}_t$ is non-increasing, and $T \in \mathbb{N}$ with $T \geq 3$. We denote by $\mathbf{J}_t$ the set $\{j_l : l = b(t-1)+1, \cdots, bt\}$ and by $\mathbf{J}$ the set $\{j_l : l = 1, \cdots, bT\}$.

3. Main Results with Discussions

In this section, we first state some basic assumptions. Then, we present and discuss our main results.

3.1 Assumptions

The following assumption is related to a moment assumption on $|y|^2$. It is weaker than the often considered bounded output assumption, such as the binary classification problems where $Y = \{-1, 1\}$.

Assumption 1 *There exists constants $M \in]0, \infty[$ and $v \in]1, \infty[$ such that*

$$\int_Y y^{2l} d\rho(y|x) \leq l! M^l v, \quad \forall l \in \mathbb{N}, \tag{5}$$

ρ_X-almost surely.

1. Note that, the random variables $j_1, \cdots, j_{bT}$ are conditionally independent given the sample $\mathbf{z}$.

To present our next assumption, we introduce the operator $\mathcal{L}_\rho : L^2(H, \rho_X) \to L^2(H, \rho_X)$, defined by $\mathcal{L}_\rho(f) = \int_X \langle x, \cdot \rangle_H f(x) \rho_X(x)$. Here, $L^2(H, \rho_X)$ is the Hilbert space of square integral functions from H to $\mathbb{R}$ with respect to ρ_X, with norm,

$$\|f\|_\rho = \left(\int_X |f(x)|^2 d\rho_X(x) \right)^{1/2}.$$

Under Assumption (3), $\mathcal{L}_\rho$ can be proved to be positive trace class operators (Cucker and Zhou, 2007), and hence $\mathcal{L}_\rho^\zeta$ with $\zeta \in \mathbb{R}$ can be defined by using the spectral theory.

It is well known (see e.g., Cucker and Zhou, 2007) that the function minimizing $\int_Z (f(x) - y)^2 d\rho(z)$ over all measurable functions $f : H \to \mathbb{R}$ is the regression function, given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X. \tag{6}$$

Define another Hilbert space $H_\rho = \{f : X \to \mathbb{R} | \exists \omega \in H \text{ with } f(x) = \langle \omega, x \rangle_H, \rho_X\text{-almost surely}\}$. Under Assumption (3), it is easy to see that H_ρ is a subspace of $L^2(H, \rho_X)$. Let $f_{\mathcal{H}}$ be the projection of the regression function f_ρ onto the closure of H_ρ in $L^2(H, \rho_X)$. It is easy to see that the search for a solution of Problem (1) is equivalent to the search of a linear function in H_ρ to approximate $f_{\mathcal{H}}$. From this point of view, bounds on the excess risk of a learning algorithm on H_ρ or H, naturally depend on the following assumption, which quantifies how well, the target function $f_{\mathcal{H}}$ can be approximated by H_ρ.

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

The above assumption is fairly standard in non-parametric regression (Cucker and Zhou, 2007; Rosasco and Villa, 2015). The bigger ζ is, the more stringent the assumption is, since

$$\mathcal{L}_\rho^{\zeta_1}(L^2(H, \rho_X)) \subseteq \mathcal{L}_\rho^{\zeta_2}(L^2(H, \rho_X)) \quad \text{when } \zeta_1 \geq \zeta_2.$$

In particular, for $\zeta = 0$, we are making no assumption, while for $\zeta = 1/2$, we are requiring $f_{\mathcal{H}} \in H_\rho$, since (Rosasco and Villa, 2015)

$$H_\rho = \mathcal{L}_\rho^{1/2}(L^2(H, \rho_X)). \tag{7}$$

In the case of $\zeta \geq 1/2$, $f_{\mathcal{H}} \in H_\rho$, which implies Problem (1) has at least one solution in the space H. In this case, we denote $\omega^\dagger$ as the solution with the minimal H-norm.

Finally, the last assumption relates to the capacity of the hypothesis space.

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

$$\operatorname{tr}(\mathcal{L}_\rho(\mathcal{L}_\rho + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \textit{for all } \lambda > 0. \tag{8}$$

The left hand-side of of (8) is called as the effective dimension (Caponnetto and De Vito, 2007), or the degrees of freedom (Zhang, 2005). It can be related to covering/entropy number conditions, see (Steinwart and Christmann, 2008) for further details. Assumption 3 is always true for $\gamma = 1$ and $c_\gamma = \kappa^2$, since $\mathcal{L}_\rho$ is a trace class operator which implies the eigenvalues of $\mathcal{L}_\rho$, denoted as σ_i, satisfy $\operatorname{tr}(\mathcal{L}_\rho) = \sum_i \sigma_i \leq \kappa^2$. This is referred to as the capacity independent setting. Assumption 3 with $\gamma \in]0, 1]$ allows to derive better error rates. It is satisfied, e.g., if the eigenvalues of $\mathcal{L}_\rho$ satisfy a polynomial decaying condition $\sigma_i \sim i^{-1/\gamma}$, or with $\gamma = 0$ if $\mathcal{L}_\rho$ is finite rank.

3.2 Optimal Rates for SGM and Batch GM: Simplified Versions

We start with the following corollaries, which are the simplified versions of our main results stated in the next subsections.

Corollary 1 (Optimal Rate for SGM) *Under Assumptions 2 and 3, let* $|y| \leq M$ *almost surely for some* $M > 0$. *Let* $p_* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$ *if* $2\zeta + \gamma > 1$, *or* $p_* = \lceil m^{1-\epsilon} \rceil$ *with* $\epsilon \in]0, 1[$ *otherwise. Consider the SGM with*
1) $b = 1$, $\eta_t \simeq \frac{1}{m}$ *for all* $t \in [(p_* m)]$, *and* $\tilde{\omega}_{p_*} = \omega_{p_* m+1}$.
If $\delta \in]0, 1]$ *and* $m \geq m_\delta$, *then with probability*[2] *at least* $1 - \delta$, *it holds*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\tilde{\omega}_{p_*})] - \inf_H \mathcal{E} \leq C \begin{cases} m^{-\frac{2\zeta}{2\zeta+\gamma}} & \text{when } 2\zeta + \gamma > 1; \\ m^{-2\zeta(1-\epsilon)} & \text{otherwise.} \end{cases} \tag{9}$$

Furthermore, the above also holds for the SGM with[3]
2) $b = \sqrt{m}$, $\eta_t \simeq \frac{1}{\sqrt{m}}$ *for all* $t \in [(p_* \sqrt{m})]$, *and* $\tilde{\omega}_{p_*} = \omega_{p_* \sqrt{m}+1}$.
In the above, m_δ *and* C *are positive constants depending on* $\kappa^2, \|\mathcal{T}_\rho\|, M, \zeta, R, c_\gamma, \gamma$, *a polynomial of* $\log m$ *and* $\log(1/\delta)$, *and* m_δ *also on* δ *(and also on* $\|f_{\mathcal{H}}\|_\infty$ *in the case that* $\zeta < 1/2$*).*

We add some comments on the above result. First, the above result asserts that, at p_* passes over the data, the SGM with two different fixed step-size and fixed mini-batch size choices, achieves optimal learning error bounds, matching (or improving) those of ridge regression (Smale and Zhou, 2007; Caponnetto and De Vito, 2007). Second, according to the above result, using mini-batch allows to use a larger step-size while achieving the same optimal error bounds. Finally, the above result can be further simplified in some special cases. For example, if we consider the capacity independent case, i.e., $\gamma = 1$, and assuming that $f_{\mathcal{H}} \in H_\rho$, which is equivalent to making Assumption 2 with $\zeta = 1/2$ as mentioned before, the error bound is $O(m^{-1/2})$, while the number of passes $p_* = \lceil \sqrt{m} \rceil$.

Remark 1 (Finite Dimensional Case) *With a simple modification of our proofs, we can derive similar results for the finite dimensional case, i.e.,* $H = \mathbb{R}^d$, *where in this case,* $\gamma = 0$. *In particular, letting* $\zeta = 1/2$, *under the same assumptions of Corollary 1, if one considers the SGM with* $b = 1$ *and* $\eta_t \simeq \frac{1}{m}$ *for all* $t \in [m^2]$, *then with high probability,* $\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{m^2+1})] - \inf_H \mathcal{E} \lesssim d/m$, *provided that* $m \gtrsim d \log d$.

Remark 2 *From the proofs, one can easily see that if* $f_{\mathcal{H}}$ *and* $\mathcal{E}(\tilde{\omega}_{p_*}) - \inf_H \mathcal{E}$ *are replaced respectively by* $f_* \in L^2(H, \rho_X)$ *and* $\|\langle \cdot, \tilde{\omega}_{p_*} \rangle_H - f_*\|_\rho^2$, *in both the assumptions and the error bounds, then all theorems and their corollaries of this paper are still true, as long as* f_* *satisfies* $\int_X (f_* - f_\rho)(x) K_x d\rho_X = 0$. *As a result, if we assume that* f_ρ *satisfies Assumption 2 (with* $f_{\mathcal{H}}$ *replaced by* f_ρ*), as typically done in (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Caponnetto and Yao, 2010) for the RKHS setting, we have that with high probability,*

$$\mathbb{E}_{\mathbf{J}} \|\langle \cdot, \tilde{\omega}_{p_*} \rangle_H - f_\rho\|_\rho^2 \leq C \begin{cases} m^{-\frac{2\zeta}{2\zeta+\gamma}} & \text{when } 2\zeta + \gamma > 1; \\ m^{-2\zeta(1-\epsilon)} & \text{otherwise.} \end{cases}$$

In this case, the factor $\|f_{\mathcal{H}}\|_\infty$ *from the upper bounds for the case* $\zeta < 1/2$ *is exactly* $\|f_\rho\|_\infty$ *and can be controlled by the condition* $|y| \leq M$ *(and more generally, by Assumption 1). Since*

2. Here, 'high probability' refers to the sample $\mathbf{z}$.
3. Here, we assume that $\sqrt{m}$ is an integer.

many common RKHSs are universally consistent (Steinwart and Christmann, 2008), making Assumption 2 on f_ρ is natural and moreover, deriving error bounds with respect to f_ρ seems to be more interesting in this case.

As a byproduct of our proofs in this paper, we derive the following optimal results for batch gradient methods (GM), defined by $\nu_1 = 0$ and

$$\nu_{t+1} = \nu_t - \eta_t \frac{1}{m} \sum_{i=1}^m (\langle \nu_t, x_i \rangle_H - y_i) x_i, \qquad t = 1, \ldots, T. \tag{10}$$

Corollary 2 (Optimal Rate for Batch GM) *Under the assumptions and notations of Corollary 1, consider batch GM (10) with $\eta_t \simeq 1$. If m is large enough, then with high probability, (9) holds for $\tilde{\omega}_{p_*} = \nu_{p_*+1}$.*

In the above corollary, the convergence rates are optimal for $2\zeta + \gamma > 1$. To the best of our knowledge, these results are the first ones with minimax rates (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2016) for the batch GM in the non-attainable case. Particularly, they improve the results in the previous literature, see Subsection 3.6 for more discussions.

Corollaries 1 and 2 cover the main contributions of this paper. In the following subsections, we will present the main theorems of this paper, following with several corollaries and simple discussions, from which one can derive the simplified versions stated in this subsection. In the next subsection, we present results for SGM in the attainable case while results in the non-attainable case will be given in Subsection 3.4, as the bounds for these two cases are different and particularly their proofs require different estimations. At last, results with more specific convergence rates for batch GM will be presented in Subsection 3.5.

3.3 Main Results for SGM: Attainable Case

In this subsection, we present convergence results in the attainable case, i.e., $\zeta \geq 1/2$, following with simple discussions. One of our main theorems in the attainable case is stated next, and provides error bounds for the studied algorithm. For the sake of readability, we only present results in a fixed step-size setting in this section. Results in a general setting ($\eta_t = \eta_1 t^{-\theta}$ with $0 \leq \theta < 1$ can be found in Section 7.

Theorem 1 *Under Assumptions 1, 2 and 3, let $\zeta \geq 1/2$, $\delta \in]0,1[$, $\eta_t = \eta\kappa^{-2}$ for all $t \in [T]$, with $\eta \leq \frac{1}{8(\log T + 1)}$. If $m \geq m_\delta$, then the following holds with probability at least $1-\delta$: for all $t \in [T]$,*

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \leq q_1 (\eta t)^{-2\zeta} + q_2 m^{-\frac{2\zeta}{2\zeta+\gamma}} (1 + m^{-\frac{1}{2\zeta+\gamma}} \eta t)^2 \log^2 T \log^2 \frac{1}{\delta} \\ + q_3 \eta b^{-1} (1 \vee m^{-\frac{1}{2\zeta+\gamma}} \eta t) \log T.\end{aligned} \tag{11}$$

Here, m_δ, q_1, q_2 and q_3 are positive constants depending on $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma, \gamma$, and m_δ also on δ (which will be given explicitly in the proof).

There are three terms in the upper bounds of (11). The first term depends on the regularity of the target function and it arises from bounding the bias, while the last two terms result from estimating the sample variance and the computational variance (due to the random choices of the points), respectively. To derive optimal rates, it is necessary to balance these three terms.

Solving this trade-off problem leads to different choices on η, T, and b, corresponding to different regularization strategies, as shown in subsequent corollaries.

The first corollary gives generalization error bounds for simple SGM, with a universal step-size depending on the number of sample points.

Corollary 3 *Under Assumptions 1, 2 and 3, let $\zeta \geq 1/2$, $\delta \in]0,1[$, $b = 1$ and $\eta_t \simeq \frac{1}{m}$ for all $t \in [m^2]$. If $m \geq m_\delta$, then with probability at least $1-\delta$, there holds*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim \left\{ \left(\frac{m}{t}\right)^{2\zeta} + m^{-\frac{2\zeta+2}{2\zeta+\gamma}} \left(\frac{t}{m}\right)^2 \right\} \cdot \log^2 m \log^2 \frac{1}{\delta}, \quad \forall t \in [m^2], \tag{12}$$

and in particular,

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{T^*+1})] - \inf_H \mathcal{E} \lesssim m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^2 m \log^2 \frac{1}{\delta}, \tag{13}$$

where $T^* = \lceil m^{\frac{2\zeta+\gamma+1}{2\zeta+\gamma}} \rceil$. Here, m_δ is exactly the same as in Theorem 1.

Remark 3 *Ignoring the logarithmic term and letting $t = pm$, Eq. (12) becomes*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{pm+1})] - \inf_H \mathcal{E} \lesssim p^{-2\zeta} + m^{-\frac{2\zeta+2}{2\zeta+\gamma}} p^2.$$

A smaller p may lead to a larger bias, while a larger p may lead to a larger sample error. From this point of view, p has a regularization effect.

The second corollary provides error bounds for SGM with a fixed mini-batch size and a fixed step-size (which depend on the number of sample points).

Corollary 4 *Under Assumptions 1, 2 and 3, let $\zeta \geq 1/2$, $\delta \in]0,1[$, $b = \lceil \sqrt{m} \rceil$ and $\eta_t \simeq \frac{1}{\sqrt{m}}$ for all $t \in [m^2]$. If $m \geq m_\delta$, then with probability at least $1-\delta$, there holds*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim \left\{ \left(\frac{\sqrt{m}}{t}\right)^{2\zeta} + m^{-\frac{2\zeta+2}{2\zeta+\gamma}} \left(\frac{t}{\sqrt{m}}\right)^2 \right\} \log^2 m \log^2 \frac{1}{\delta}, \quad \forall t \in [m^2], \tag{14}$$

and particularly,

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{T^*+1})] - \inf_H \mathcal{E} \lesssim m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^2 m \log^2 \frac{1}{\delta}, \tag{15}$$

where $T^* = \lceil m^{\frac{1}{2\zeta+\gamma} + \frac{1}{2}} \rceil$.

The above two corollaries follow from Theorem 1 with the simple observation that the dominating terms in (11) are the terms related to the bias and the sample variance, when a small step-size is chosen. The only free parameter in (12) and (14) is the number of iterations/passes. The ideal stopping rule is achieved by balancing the two terms related to the bias and the sample variance, showing the regularization effect of the number of passes. Since the ideal stopping rule depends on the unknown parameters ζ and γ, a hold-out cross-validation procedure is often used to tune the stopping rule in practice. Using an argument similar to that in Chapter 6 from (Steinwart and Christmann, 2008), it is possible to show that this procedure can achieve the same convergence rate.

We give some further remarks. First, the upper bound in (13) is optimal up to a logarithmic factor, in the sense that it matches the minimax lower rate in (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2016). Second, according to Corollaries 3 and 4, $\frac{bT^*}{m} \simeq m^{\frac{1}{2\zeta+\gamma}}$ passes over the data are needed to obtain optimal rates in both cases. Finally, in comparing the simple SGM and the mini-batch SGM, Corollaries 3 and 4 show that a larger step-size is allowed to use for the latter.

In the next result, both the step-size and the stopping rule are tuned to obtain optimal rates for simple SGM with multiple passes. In this case, the step-size and the number of iterations are the regularization parameters.

Corollary 5 *Under Assumptions 1, 2 and 3, let* $\zeta \geq 1/2$, $\delta \in]0,1[$, $b = 1$ *and* $\eta_t \simeq m^{-\frac{2\zeta}{2\zeta+\gamma}}$ *for all* $t \in [m^2]$. *If* $m \geq m_\delta$, *and* $T^* = \lceil m^{\frac{2\zeta+1}{2\zeta+\gamma}} \rceil$, *then (13) holds with probability at least* $1-\delta$.

The next corollary shows that for some suitable mini-batch sizes, optimal rates can be achieved with a constant step-size (which is nearly independent of the number of sample points) by early stopping.

Corollary 6 *Under Assumptions 1, 2 and 3, let* $\zeta \geq 1/2$, $\delta \in]0,1[$, $b = \lceil m^{\frac{2\zeta}{2\zeta+\gamma}} \rceil$ *and* $\eta_t \simeq \frac{1}{\log m}$ *for all* $t \in [m]$. *If* $m \geq m_\delta$, *and* $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$, *then (13) holds with probability at least* $1-\delta$.

According to Corollaries 5 and 6, around $m^{\frac{1-\gamma}{2\zeta+\gamma}}$ passes over the data are needed to achieve the best performance in the above two strategies. In comparisons with Corollaries 3 and 4 where around $m^{\frac{\zeta+1}{2\zeta+\gamma}}$ passes are required, the latter seems to require fewer passes over the data. However, in this case, one might have to run the algorithms multiple times to tune the step-size, or the mini-batch size.

Remark 4 *1) If we make no assumption on the capacity, i.e.,* $\gamma = 1$, *Corollary 5 recovers the result in (Ying and Pontil, 2008) for one pass SGM.*
2) If we make no assumption on the capacity and assume that $f_{\mathcal{H}} \in H_\rho$, *from Corollaries 5 and 6, we see that the optimal convergence rate* $O(m^{-1/2})$ *can be achieved after one pass over the data in both of these two strategies. In this special case, Corollaries 5 and 6 recover the results for one pass SGM in, e.g., (Shamir and Zhang, 2013; Dekel et al., 2012).*

The next result gives generalization error bounds for 'batch' SGM with a constant step-size (nearly independent of the number of sample points).

Corollary 7 *Under Assumptions 1, 2 and 3, let* $\zeta \geq 1/2$, $\delta \in]0,1[$, $b = m$ *and* $\eta_t \simeq \frac{1}{\log m}$ *for all* $t \in [m]$. *If* $m \geq m_\delta$, *and* $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$, *then (13) holds with probability at least* $1-\delta$.

Theorem 1 and its corollaries give convergence results with respect to the target function values. In the next theorem and corollary, we will present convergence results in H-norm.

Theorem 2 *Under the assumptions of Theorem 1, the following holds with probability at least* $1-\delta$: *for all* $t \in [T]$

$$\mathbb{E}_{\mathbf{J}}[\|\omega_t - \omega^\dagger\|_H^2] \leq q_1(\eta t)^{1-2\zeta} + q_2 m^{-\frac{2\zeta-1}{2\zeta+\gamma}}(1 + m^{-\frac{1}{2\zeta+\gamma}}\eta t)^2 \log^2 T \log^2 \frac{1}{\delta} + q_3 \eta^2 t b^{-1}. \tag{16}$$

Here, q_1, q_2 *and* q_3 *are positive constants depending on* $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma$, *and* γ *(which can be given explicitly in the proof).*

The proof of the above theorem is similar as that for Theorem 1, and will be given in Subsection 8. Again, the upper bound in (16) is composed of three terms related to bias, sample variance, and computational variance. Balancing these three terms leads to different choices on η, T, and b, as shown in the following corollary.

Corollary 8 *With the same assumptions and notations from any one of Corollaries 3 to 7, the following holds with probability at least* $1-\delta$:

$$\mathbb{E}_{\mathbf{J}}[\|\omega_{T^*+1} - \omega^\dagger\|_H^2] \lesssim m^{-\frac{2\zeta-1}{2\zeta+\gamma}} \log^2 m \log^2 \frac{1}{\delta}.$$

The convergence rate in the above corollary is optimal up to a logarithmic factor, as it matches the minimax rate shown in (Blanchard and Mücke, 2016).

In the next subsection, we will present convergence results in the non-attainable case, i.e., $\zeta < 1/2$.

3.4 Main Results for SGM: Non-attainable Case

Our main theorem in the non-attainable case is stated next, and provides error bounds for the studied algorithm. Here, we present results with a fixed step-size, whereas general results with a decaying step-size will be given in Section 7.

Theorem 3 *Under Assumptions 1, 2 and 3, let* $\zeta \le 1/2$, $\delta \in]0,1[$, $\eta_t = \eta\kappa^{-2}$ *for all* $t \in [T]$, *with* $0 < \eta \le \frac{1}{8(\log T+1)}$. *Then the following holds for all* $t \in [T]$ *with probability at least* $1-\delta$*:*
1) if $2\zeta + \gamma > 1$ *and* $m \ge m_\delta$, *then*

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \le \left(q_1(\eta t)^{-2\zeta} + q_2 m^{-\frac{2\zeta}{2\zeta+\gamma}}\right)(1 \vee m^{-\frac{1}{2\zeta+\gamma}}\eta t)^3 \log^4 T \log^2 \frac{1}{\delta} \\ + q_3 \eta b^{-1}(1 \vee m^{-\frac{1}{2\zeta+\gamma}}\eta t)\log T;\end{aligned} \tag{17}$$

2) if $2\zeta + \gamma \le 1$ *and for some* $\epsilon \in]0,1[$, $m \ge m_{\delta,\epsilon}$, *then*

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \le \left(q_1(\eta t)^{-2\zeta} + q_2 m^{\gamma(1-\epsilon)-1}\right)(1 \vee \eta m^{\epsilon-1} t)^3 \log^4 T \log^2 \frac{1}{\delta} \\ + q_3 \eta b^{-1}(1 \vee m^{\epsilon-1}\eta t)\log T.\end{aligned}$$

Here, m_δ *(or* $m_{\delta,\epsilon}$*),* q_1, q_2 *and* q_3 *are positive constants depending only on* $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma, \gamma$, $\|f_{\mathcal{H}}\|_\infty$, *and* m_δ *(or* $m_{\delta,\epsilon}$*) also on* δ *(and* ϵ*).*

The upper bounds in (11) (for the attainable case) and (17) (for the non-attainable case) are similar, whereas the latter has an extra logarithmic factor. Consequently, in the subsequent corollaries, we derive $O(m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^4 m)$ for the non-attainable case. In comparison with that for the attainable case, the convergence rate for the non-attainable case has an extra $\log^2 m$ factor.

Similar to Corollaries 3 and 4, and as direct consequences of the above theorem, we have the following generalization error bounds for the studied algorithm with different choices of parameters in the non-attainable case.

Corollary 9 *Under Assumptions 1, 2 and 3, let* $\zeta \le 1/2$, $\delta \in]0,1[$, $b = 1$ *and* $\eta_t \simeq \frac{1}{m}$ *for all* $t \in [m^2]$. *With probability at least* $1-\delta$, *the following holds:*
1) if $2\zeta + \gamma > 1$, $m \ge m_\delta$ *and* $T^* = \lceil m^{\frac{1+2\zeta+\gamma}{2\zeta+\gamma}} \rceil$, *then*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{T^*+1})] - \inf_H \mathcal{E} \lesssim m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^4 m \log^2 \frac{1}{\delta}; \tag{18}$$

2) if $2\zeta+\gamma \leq 1$*, and for some* $\epsilon \in]0,1[$*,* $m \geq m_{\delta,\epsilon}$*, and* $T^* = \lceil m^{2-\epsilon} \rceil$*, then*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{T^*+1})] - \inf_H \mathcal{E} \lesssim m^{-2\zeta(1-\epsilon)} \log^4 m \log^2 \frac{1}{\delta}. \tag{19}$$

Here, m_δ *and* $m_{\delta,\epsilon}$ *are given by Theorem 3.*

Corollary 10 *Under Assumptions 1, 2 and 3, let* $\zeta \leq 1/2$ *,* $\delta \in]0,1[$*,* $b \simeq \sqrt{m}$ *and* $\eta_t \simeq \frac{1}{\sqrt{m}}$ *for all* $t \in [m^2]$*. With probability at least* $1-\delta$*, there holds*
1) if $2\zeta+\gamma > 1$*,* $m \geq m_\delta$ *and* $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}+\frac{1}{2}} \rceil$*, then (18) holds;*
2) if $2\zeta+\gamma \leq 1$*, for some* $\epsilon \in]0,1[$*,* $m \geq m_{\delta,\epsilon}$*, and* $T^* = \lceil m^{\frac{3}{2}-\epsilon} \rceil$*, then (19) holds.*

The convergence rates in the above corollaries, i.e., $m^{-\frac{2\zeta}{2\zeta+\gamma}}$ if $2\zeta+\gamma > 1$ or $m^{-2\zeta(1-\epsilon)}$ otherwise, match those in (Dieuleveut and Bach, 2016) for one pass SGM with averaging, up to a logarithmic factor. Also, in the capacity independent case, i.e., $\gamma = 1$, the convergence rates in the above corollary read as $m^{-\frac{2\zeta}{2\zeta+1}}$ (since $2\zeta+\gamma$ is always bigger than 1), which are exactly the same as those in (Ying and Pontil, 2008) for one pass SGM.

Similar results to Corollaries 5–7 can be also derived for the non-attainable case by applying Theorem 3. Refer to Appendix B for more details.

3.5 Main Results for Batch GM

In this subsection, we present convergence results for batch GM. As a byproduct of our proofs in this paper, we have the following convergence rates for batch GM.

Theorem 4 *Under Assumptions 1, 2 and 3, set* $\eta_t \simeq 1$*, for all* $t \in [m]$*. Let* $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$ *if* $2\zeta+\gamma > 1$*, or* $T^* = \lceil m^{1-\epsilon} \rceil$ *with* $\epsilon \in]0,1[$ *otherwise. Then with probability at least* $1-\delta$ *(*$0 < \delta < 1$*), the following holds for the learning sequence generated by (10):*
1) if $\zeta > 1/2$ *and* $m \geq m_\delta$*, then*

$$\mathcal{E}(\nu_{T_*+1}) - \inf_H \mathcal{E} \lesssim m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^2 m \log^2 \frac{1}{\delta};$$

2) if $\zeta \leq 1/2$*,* $2\zeta+\gamma > 1$ *and* $m \geq m_\delta$*, then*

$$\mathcal{E}(\nu_{T_*+1}) - \inf_H \mathcal{E} \lesssim m^{-\frac{2\zeta}{2\zeta+\gamma}} \log^4 m \log^2 \frac{1}{\delta};$$

3) if $2\zeta+\gamma \leq 1$ *and* $m \geq m_{\delta,\epsilon}$*, then*

$$\mathcal{E}(\nu_{T_*+1}) - \inf_H \mathcal{E} \lesssim m^{-2\zeta(1-\epsilon)} \log^4 m \log^2 \frac{1}{\delta}.$$

Here, m_δ *(or* $m_{\delta,\epsilon}$*), and all the constants in the upper bounds are positive and depend only on* $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma, \gamma$*,* $\|f_{\mathcal{H}}\|_\infty$*, and* m_δ *(or* $m_{\delta,\epsilon}$*) also on* δ *(and* ϵ*).*

3.6 Discussions

We must compare our results with previous works. For non-parametric regression with the square loss, one pass SGM has been studied in, e.g., (Ying and Pontil, 2008; Shamir and Zhang, 2013; Tarres and Yao, 2014; Dieuleveut and Bach, 2016). In particular, Ying and Pontil (2008) proved capacity independent rate of order $O(m^{-\frac{2\zeta}{2\zeta+1}}\log m)$ with a fixed step-size $\eta \simeq m^{-\frac{2\zeta}{2\zeta+1}}$, and Dieuleveut and Bach (2016) derived capacity dependent error bounds of order $O(m^{-\frac{2\min(\zeta,1)}{2\min(\zeta,1)+\gamma}})$ (when $2\zeta+\gamma>1$) for the average. Note also that a regularized version of SGM has been studied in (Tarres and Yao, 2014), where the derived convergence rate is of order $O(m^{-\frac{2\zeta}{2\zeta+1}})$ assuming that $\zeta \in [\frac{1}{2},1]$. In comparison with these existing convergence rates, our rates from (13) are comparable, either involving the capacity condition, or allowing a broader regularity parameter ζ (which thus improves the rates). For finite dimensional cases, it has been shown in (Bach and Moulines, 2013) that one pass SGM with averaging with a constant step-size achieves the optimal convergence rate of $O(d/m)$. In comparisons, our results for multi-pass SGM with a smaller step-size seems to be suboptimal in the computational complexity, as we need m passes over the data to achieve the same rate. The reason for this may arise from "the computational error" that will be introduced later, or the fact that we do not consider an averaging step as done in (Bach and Moulines, 2013). We hope that in the future by considering a larger step-size and averaging, one can reduce the computational complexity of multi-pass SGM while achieving the same rate.

More recently, Rosasco and Villa (2015) studied multiple passes SGM with a fixed ordering at each pass, also called incremental gradient method. Making no assumption on the capacity, rates of order $O(m^{-\frac{\zeta}{\zeta+1}})$ (in $L^2(H,\rho_X)$-norm) with a universal step-size $\eta \simeq 1/m$ are derived. In comparisons, Corollary 3 achieves better rates, while considering the capacity assumption. Note also that Rosasco and Villa (2015) proved sharp rate in H-norm for $\zeta \geq 1/2$ in the capacity independent case. In comparisons, we derive optimal capacity-dependent rate, considering mini-batches.

The idea of using mini-batches (and parallel implements) to speed up SGM in a general stochastic optimization setting can be found, e.g., in (Shalev-Shwartz et al., 2011; Dekel et al., 2012; Sra et al., 2012; Ng, 2016). Our theoretical findings, especially the interplay between the mini-batch size and the step-size, can give further insights on parallelization learning. Besides, it has been shown in (Cotter et al., 2011; Dekel et al., 2012) that for one pass mini-batch SGM with a fixed step-size $\eta \simeq b/\sqrt{m}$ and a smooth loss function, assuming the existence of at least one solution in the hypothesis space for the expected risk minimization, the convergence rate is of order $O(\sqrt{1/m}+b/m)$ by considering an averaging scheme. When adapting to the learning setting we consider, this reads as that if $f_{\mathcal{H}} \in H_\rho$, i.e., $\zeta = 1/2$, the convergence rate for the average is $O(\sqrt{1/m}+b/m)$. Note that, $f_{\mathcal{H}}$ does not necessarily belong to H_ρ in general. Also, our derived convergence rate from Corollary 4 is better, when the regularity parameter ζ is greater than $1/2$, or γ is smaller than 1.

For batch GM in the attainable case, convergent results with optimal rates have been derived in, e.g, (Bauer et al., 2007; Caponnetto and Yao, 2010; Blanchard and Mücke, 2016; Dicker et al., 2017). In particular, Bauer et al. (2007) proved convergence rates $O(m^{-\frac{2\zeta}{2\zeta+1}})$ without considering Assumption 3, and Caponnetto and Yao (2010) derived convergence rates $O(m^{-\frac{2\zeta}{2\zeta+\gamma}})$. For the non-attainable case, convergent results with suboptimal rates $O(m^{\frac{-2\zeta}{2\zeta+2}})$ can be found in (Yao et al., 2007), and to the best of our knowledge, the only result with optimal rate $O(m^{\frac{-2\zeta}{2\zeta+\gamma}})$ is the one derived by Caponnetto and Yao (2010), but the result requires extra unlabeled data. In contrast, Theorem 4 of this paper does not require any extra unlabeled data, while achieving the same optimal rates (up to a logarithmic factor). To the best of our knowledge, Theorem 4 may be the first optimal result in the non-attainable case for batch GM.

We end this discussion with some further comments on batch GM and simple SGM. First, according to Corollaries 1 and 2, it seems that both simple SGM (with step-size $\eta_t \simeq m^{-1}$) and batch GM (with step-size $\eta_t \simeq 1$) have the same computational complexities (which are related to the number of passes) and the same orders of upper bounds. However, there is a subtle difference between these two algorithms. As we see from (22) in the coming subsection, every m iterations of simple SGM (with step-size $\eta_t \simeq m^{-1}$) corresponds to one iteration of batch GM (with step-size $\eta_t \simeq 1$). In this sense, SGM discretizes and refines the regularization path of batch GM, which thus may lead to smaller generalization errors. This phenomenon can be further understood by comparing our derived bounds, (11) and (73), for these two algorithms. Indeed, if one can ignore the computational error, one can easily show that the minimization (over t) of right hand-side of (11) with $\eta \simeq m^{-1}$ is always smaller than that of (73) with $\eta \simeq 1$. At last, by Corollary 6, using a larger step-size for SGM allows one to stop earlier (while sharing the same optimal rates), which thus reduces the computational complexity. This suggests that SGM may have some computational advantage over batch GM.

3.7 Proof Sketch (Error Decomposition)

The key to our proof is a novel error decomposition, which may be also used in analysing other learning algorithms. One may also use the approach in (Bousquet and Bottou, 2008; Lin et al., 2016b,a) which is based on the following error decomposition,

$$\mathbb{E}\mathcal{E}(\omega_t) - \inf_H \mathcal{E} = [\mathbb{E}(\mathcal{E}(\omega_t) - \mathcal{E}_{\mathbf{z}}(\omega_t)) + \mathbb{E}\mathcal{E}_{\mathbf{z}}(\tilde{\omega}) - \mathcal{E}(\tilde{\omega})] + \mathbb{E}(\mathcal{E}_{\mathbf{z}}(\omega_t) - \mathcal{E}_{\mathbf{z}}(\tilde{\omega})) + \mathcal{E}(\tilde{\omega}) - \inf_H \mathcal{E},$$

where $\tilde{\omega} \in H$ is some suitably intermediate element and $\mathcal{E}_{\mathbf{z}}$ denotes the empirical risk over $\mathbf{z}$, i.e.,

$$\mathcal{E}_{\mathbf{z}}(\cdot) = \frac{1}{m}\sum_{i=1}^{m} (\langle \cdot, x_i\rangle - y_i)^2 . \tag{20}$$

However, one can only derive a sub-optimal convergence rate, since the proof procedure involves upper bounding the learning sequence to estimate the sample error (the first term of right-hand side). Also, in this case, the 'regularity' of the regression function can not be fully utilized for estimating the bias (the last term). Thanks to the property of squares loss, we can exploit a different error decomposition leading to better results.

To describe the decomposition, we need to introduce two sequences. The *population iteration* is defined by $\mu_1 = 0$ and

$$\mu_{t+1} = \mu_t - \eta_t \int_X (\langle \mu_t, x\rangle_H - f_\rho(x)) x d\rho_X(x), \qquad t = 1, \ldots, T. \tag{21}$$

The above iterated procedure is ideal and can not be implemented in practice, since the distribution ρ_X is unknown in general. Replacing ρ_X by the empirical measure and $f_\rho(x_i)$ by y_i, we derive the *sample iteration* (associated with the sample $\mathbf{z}$), i.e., (10). Clearly, μ_t is deterministic and ν_t is a H-valued random variable depending on $\mathbf{z}$. Given the sample $\mathbf{z}$, the sequence $\{\nu_t\}_t$ has a natural relationship with the learning sequence $\{\omega_t\}_t$, since

$$\mathbb{E}_{\mathbf{J}}[\omega_t] = \nu_t. \tag{22}$$

Indeed, taking the expectation with respect to $\mathbf{J}_t$ on both sides of (4), and noting that ω_t depends only on $\mathbf{J}_1, \cdots, \mathbf{J}_{t-1}$ (given any $\mathbf{z}$), one has

$$\mathbb{E}_{\mathbf{J}_t}[\omega_{t+1}] = \omega_t - \eta_t \frac{1}{m}\sum_{i=1}^{m}(\langle \omega_t, x_i\rangle_H - y_i) x_i,$$

and thus,

$$\mathbb{E}_{\mathbf{J}}[\omega_{t+1}] = \mathbb{E}_{\mathbf{J}}[\omega_t] - \eta_t \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbb{E}_{\mathbf{J}}[\omega_t], x_i \rangle_H - y_i) x_i, \quad t = 1, \ldots, T,$$

which satisfies the iterative relationship given in (10). By an induction argument, (22) can then be proved.

Let $\mathcal{S}_\rho : H \to L^2(H, \rho_X)$ be the linear map defined by $(\mathcal{S}_\rho \omega)(x) = \langle \omega, x \rangle_H, \forall \omega, x \in H$. We have the following error decomposition.

Proposition 1 *We have*

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_t)] - \inf_H \mathcal{E} \le 2\|\mathcal{S}_\rho \mu_t - f_{\mathcal{H}}\|_\rho^2 + 2\|\mathcal{S}_\rho \nu_t - \mathcal{S}_\rho \mu_t\|_\rho^2 + \mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t\|_\rho^2]. \tag{23}$$

Proof For any $\omega \in H$, we have (Rosasco and Villa, 2015)

$$\mathcal{E}(\omega) - \inf_{\omega \in H} \mathcal{E}(\omega) = \|\mathcal{S}_\rho \omega - f_{\mathcal{H}}\|_\rho^2.$$

Thus, $\mathcal{E}(\omega_t) - \inf_H \mathcal{E} = \|\mathcal{S}_\rho \omega_t - f_{\mathcal{H}}\|_\rho^2$, and

$$\begin{aligned}
&\mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - f_{\mathcal{H}}\|_\rho^2] = \mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t + \mathcal{S}_\rho \nu_t - f_{\mathcal{H}}\|_\rho^2] \\
&= \mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t\|_\rho^2 + \|\mathcal{S}_\rho \nu_t - f_{\mathcal{H}}\|_\rho^2] + 2\mathbb{E}_{\mathbf{J}} \langle \mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t, \mathcal{S}_\rho \nu_t - f_{\mathcal{H}} \rangle_\rho.
\end{aligned}$$

Using (22) in the above equality, we get,

$$\mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - f_{\mathcal{H}}\|_\rho^2] = \mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t\|_\rho^2 + \|\mathcal{S}_\rho \nu_t - f_{\mathcal{H}}\|_\rho^2].$$

The proof is finished by considering,

$$\|\mathcal{S}_\rho \nu_t - f_{\mathcal{H}}\|_\rho^2 = \|\mathcal{S}_\rho \nu_t - \mathcal{S}_\rho \mu_t + \mathcal{S}_\rho \mu_t - f_{\mathcal{H}}\|_\rho^2 \le 2\|\mathcal{S}_\rho \nu_t - \mathcal{S}_\rho \mu_t\|_\rho^2 + 2\|\mathcal{S}_\rho \mu_t - \mathcal{S}_\rho f_{\mathcal{H}}\|_\rho^2.$$

■

There are three terms in the upper bound of the error decomposition (23). We refer to the deterministic term $\|\mathcal{S}_\rho \mu_t - f_{\mathcal{H}}\|_\rho^2$ as the *bias*, the term $\|\mathcal{S}_\rho \nu_t - \mathcal{S}_\rho \mu_t\|_\rho^2$ depending on $\mathbf{z}$ as the *sample variance*, and $\mathbb{E}_{\mathbf{J}}[\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t\|_\rho^2]$ as the *computational variance*. The bias term, which is deterministic, has been well studied in the literature, see e.g., (Yao et al., 2007) and also (Rosasco and Villa, 2015). The main novelties of this paper are the estimate of the sample and computational variances and the difficult part is the estimate of the computational variances. The proof of these results is quite lengthy and makes use of some ideas from (Yao et al., 2007; Smale and Zhou, 2007; Bauer et al., 2007; Ying and Pontil, 2008; Tarres and Yao, 2014; Rudi et al., 2015). These three error terms will be estimated in Sections 5 and 6. The bounds in Theorems 1 and 3 thus follow plugging these estimations in the error decomposition, see Section 7 for more details. The proof for Theorem 2 is similar, see Section 8 for the details.

4. Preliminary Analysis

In this section, we introduce some notation and preliminary lemmas that are necessary to our proofs.

4.1 Notation

We first introduce some notations. For $t \in \mathbb{N}$, $\Pi_{t+1}^T(L) = \prod_{k=t+1}^T (I - \eta_k L)$ for $t \in [T-1]$ and $\Pi_{T+1}^T(L) = I$, for any operator $L : \mathcal{H} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space and I denotes the identity operator on $\mathcal{H}$. $\mathbb{E}[\xi]$ denotes the expectation of a random variable ξ. For a given bounded operator $L : L^2(H, \rho_X) \to H$, $\|L\|$ denotes the operator norm of L, i.e., $\|L\| = \sup_{f \in L^2(H,\rho_X), \|f\|_\rho = 1} \|Lf\|_H$. We will use the conventional notations on summation and production: $\prod_{i=t+1}^t = 1$ and $\sum_{i=t+1}^t = 0$.

We next introduce some auxiliary operators. Let $\mathcal{S}_\rho : H \to L^2(H, \rho_X)$ be the linear map $\omega \to \langle \omega, \cdot \rangle_H$, which is bounded by κ under Assumption (3). Furthermore, we consider the adjoint operator $\mathcal{S}_\rho^* : L^2(H, \rho_X) \to H$, the covariance operator $\mathcal{T}_\rho : H \to H$ given by $\mathcal{T}_\rho = \mathcal{S}_\rho^* \mathcal{S}_\rho$, and the operator $\mathcal{L}_\rho : L^2(H, \rho_X) \to L^2(H, \rho_X)$ given by $\mathcal{S}_\rho \mathcal{S}_\rho^*$. It can be easily proved that $\mathcal{S}_\rho^* g = \int_X x g(x) d\rho_X(x)$ and $\mathcal{T}_\rho = \int_X \langle \cdot, x \rangle_H x d\rho_X(x)$. The operators $\mathcal{T}_\rho$ and $\mathcal{L}_\rho$ can be proved to be positive trace class operators (and hence compact). For any $\omega \in H$, it is easy to prove the following isometry property (Steinwart and Christmann, 2008)

$$\|\mathcal{S}_\rho \omega\|_\rho = \|\sqrt{\mathcal{T}_\rho} \omega\|_H. \tag{24}$$

We define the sampling operator $\mathcal{S}_{\mathbf{x}} : H \to \mathbb{R}^m$ by $(\mathcal{S}_{\mathbf{x}} \omega)_i = \langle \omega, x_i \rangle_H$, $i \in [m]$, where the norm $\|\cdot\|_{\mathbb{R}^m}$ in $\mathbb{R}^m$ is the Euclidean norm times $1/\sqrt{m}$. Its adjoint operator $\mathcal{S}_{\mathbf{x}}^* : \mathbb{R}^m \to H$, defined by $\langle \mathcal{S}_{\mathbf{x}}^* \mathbf{y}, \omega \rangle_H = \langle \mathbf{y}, \mathcal{S}_{\mathbf{x}} \omega \rangle_{\mathbb{R}^m}$ for $\mathbf{y} \in \mathbb{R}^m$ is thus given by $\mathcal{S}_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m} \sum_{i=1}^m y_i x_i$. Moreover, we can define the empirical covariance operator $\mathcal{T}_{\mathbf{x}} : H \to H$ such that $\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^* \mathcal{S}_{\mathbf{x}}$. Obviously,

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

With these notations, (21) and (10) can be rewritten as

$$\mu_{t+1} = \mu_t - \eta_t (\mathcal{T}_\rho \mu_t - \mathcal{S}_\rho^* f_\rho), \qquad t = 1, \ldots, T, \tag{25}$$

and

$$\nu_{t+1} = \nu_t - \eta_t (\mathcal{T}_{\mathbf{x}} \nu_t - \mathcal{S}_{\mathbf{x}}^* \mathbf{y}), \qquad t = 1, \ldots, T, \tag{26}$$

respectively.

Using the projection theorem, one can prove that

$$\mathcal{S}_\rho^* f_\rho = \mathcal{S}_\rho^* f_{\mathcal{H}}. \tag{27}$$

Indeed, since $f_{\mathcal{H}}$ is the projection of the regression function f_ρ onto the closure of H_ρ in $L^2(H, \rho_X)$, according to the projection theorem, one has

$$\langle f_{\mathcal{H}} - f_\rho, \mathcal{S}_\rho \omega \rangle_\rho = 0, \qquad \forall \omega \in H,$$

which can be written as

$$\langle \mathcal{S}_\rho^* f_{\mathcal{H}} - \mathcal{S}_\rho^* f_\rho, \omega \rangle_H = 0, \qquad \forall \omega \in H,$$

and thus leading to (27).

4.2 Concentration Inequality

We need 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 11 *Let $w_1, \cdots, w_m$ be i.i.d random variables in a Hilbert space with norm $\|\cdot\|$. Suppose that there are two positive constants B and σ^2 such that*

$$\mathbb{E}[\|w_1 - \mathbb{E}[w_1]\|^l] \leq \frac{1}{2} l! B^{l-2} \sigma^2, \quad \forall l \geq 2. \tag{28}$$

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

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

In particular, (28) holds if

$$\|w_1\| \leq B/2 \quad a.s., \quad \text{and} \quad \mathbb{E}[\|w_1\|^2] \leq \sigma^2. \tag{29}$$

4.3 Basic Estimates

Finally, we introduce the following three basic estimates, whose proofs can be found in Appendix C.

Lemma 12 *Let $\theta \in [0, 1[$, and $t \in \mathbb{N}$. Then*

$$\frac{t^{1-\theta}}{2} \leq \sum_{k=1}^{t} k^{-\theta} \leq \frac{t^{1-\theta}}{1-\theta}.$$

Lemma 13 *Let $\theta \in \mathbb{R}$ and $t \in \mathbb{N}$. Then*

$$\sum_{k=1}^{t} k^{-\theta} \leq t^{\max(1-\theta, 0)}(1 + \log t).$$

Lemma 14 *Let $q \in \mathbb{R}$ and $t \in \mathbb{N}$ with $t \geq 3$. Then*

$$\sum_{k=1}^{t-1} \frac{1}{t-k} k^{-q} \leq 2t^{-\min(q,1)}(1 + \log t).$$

In the next sections, we begin prooving the main results. The proofs are quite lengthy and they are divided into several steps. For the ease of readability, we list some of the notations and definitions in Appendix D. We also remark that we are particularly interested in developing error bounds in terms of the stepsize η_t $(= \eta_1 t^{-\theta})$, the number of iterations t or T, the 'regularization' parameter $\lambda > 0$, the sample size m, the minibatch size b, and the failing profitability δ. Other parameters such as $\kappa^2, \|\mathcal{T}_\rho\|, M, v, R, c_\gamma$ and $\|f_{\mathcal{H}}\|_\infty$ can be always viewed as some constants, which are less important in our error bounds.

5. Estimating Bias and Sample Variance

In this section, we estimate the bias and the sample variance.

5.1 Bias

In this subsection, we develop upper bounds for the bias, i.e., $\|\mathcal{S}_\rho \mu_t - f_{\mathcal{H}}\|_\rho^2$. Towards this end, we introduce the following lemma, whose proof borrows idea from (Ying and Pontil, 2008; Tarres and Yao, 2014).

Lemma 15 *Let L be a compact, positive operator on a separable Hilbert space H. Assume that $\eta_1 \|L\| \leq 1$. Then for $t \in \mathbb{N}$ and any non-negative integer $k \leq t-1$,*

$$\|\Pi_{k+1}^t(L) L^\zeta\| \leq \left(\frac{\zeta}{\mathrm{e} \sum_{j=k+1}^t \eta_j} \right)^\zeta. \tag{30}$$

Proof Let $\{\sigma_i\}$ be the sequence of eigenvalues of L. We have

$$\|\Pi_{k+1}^t(L) L^\zeta\| = \sup_i \prod_{l=k+1}^t (1 - \eta_l \sigma_i) \sigma_i^\zeta.$$

Using the basic inequality

$$1 + x \leq \mathrm{e}^x \qquad \text{for all } x \geq -1, \tag{31}$$

with $\eta_l \|L\| \leq 1$, we get

$$\begin{aligned} \|\Pi_{k+1}^t(L) L^\zeta\| &\leq \sup_i \exp\left\{ -\sigma_i \sum_{l=k+1}^t \eta_l \right\} \sigma_i^\zeta \\ &\leq \sup_{x \geq 0} \exp\left\{ -x \sum_{l=k+1}^t \eta_l \right\} x^\zeta. \end{aligned}$$

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

$$\sup_{x \geq 0} \mathrm{e}^{-cx} x^\zeta = \left(\frac{\zeta}{\mathrm{e}c} \right)^\zeta. \tag{32}$$

Using this inequality, one can get the desired result (30). ■

With the above lemma and Lemma 12, we can derive the following result for the bias.

Proposition 2 *Under Assumption 2, let $\eta_1 \kappa^2 \leq 1$. Then, for any $t \in \mathbb{N}$,*

$$\|\mathcal{S}_\rho \mu_{t+1} - f_{\mathcal{H}}\|_\rho \leq R \left(\frac{\zeta}{2 \sum_{j=1}^t \eta_j} \right)^\zeta. \tag{33}$$

In particular, if $\eta_t = \eta t^{-\theta}$ for all $t \in \mathbb{N}$, with $\eta \in]0, \kappa^{-2}]$ and $\theta \in [0, 1[$, then

$$\|\mathcal{S}_\rho \mu_{t+1} - f_{\mathcal{H}}\|_\rho \leq R \zeta^\zeta \eta^{-\zeta} t^{(\theta - 1)\zeta}. \tag{34}$$

The above result is essentially proved in (Yao et al., 2007), see also (Rosasco and Villa, 2015) when step-size is fixed. For the sake of completeness, we provide a proof in Appendix C. The following lemma gives upper bounds for the sequence $\{\mu_t\}_{t \in \mathbb{N}}$ in H-norm. It will be used for the estimation on the sample variance in the next section.

Lemma 16 *Under Assumption 2, let $\eta_1\kappa^2 \leq 1$. The following holds for all $t \in \mathbb{N}$:*
1) If $\zeta \geq 1/2$,

$$\|\mu_{t+1}\|_H \leq R\kappa^{2\zeta-1}. \tag{35}$$

2) If $\zeta \in]0, 1/2]$,

$$\|\mu_{t+1}\|_H \leq R\left\{\kappa^{2\zeta-1} \vee \left(\sum_{k=1}^{t}\eta_k\right)^{\frac{1}{2}-\zeta}\right\}. \tag{36}$$

Proof The proof can be found in Appendix C. The proof for a fixed step-size (i.e., $\eta_t = \eta$ for all t) can be also found in (Rosasco and Villa, 2015). For a general step-size, the proof is similar. Note also that our proof for the non-attainable case is simpler than that in (Rosasco and Villa, 2015). ■

5.2 Sample Variance

In this subsection, we estimate the sample variance, i.e., $\mathbb{E}[\|\mathcal{S}_\rho\mu_t - \mathcal{S}_\rho\nu_t\|_\rho^2]$. Towards this end, we need some preliminary analysis. We first introduce the following key inequality, which also provides the basic idea on estimating $\mathbb{E}[\|\mathcal{S}_\rho\mu_t - \mathcal{S}_\rho\nu_t\|_\rho^2]$.

Lemma 17 *For all $t \in [T]$, we have*

$$\|\mathcal{S}_\rho\nu_{t+1} - \mathcal{S}_\rho\mu_{t+1}\|_\rho \leq \sum_{k=1}^{t}\eta_k\left\|\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})N_k\right\|_H, \tag{37}$$

where

$$N_k = (\mathcal{T}_\rho\mu_k - \mathcal{S}_\rho^* f_\rho) - (\mathcal{T}_{\mathbf{x}}\mu_k - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}), \qquad \forall k \in [T]. \tag{38}$$

Proof Since ν_{t+1} and μ_{t+1} are given by (26) and (25), respectively,

$$\begin{aligned}\nu_{t+1} - \mu_{t+1} &= \nu_t - \mu_t + \eta_t\left\{(\mathcal{T}_\rho\mu_t - \mathcal{S}_\rho^* f_\rho) - (\mathcal{T}_{\mathbf{x}}\nu_t - \mathcal{S}_{\mathbf{x}}^*\mathbf{y})\right\}\\ &= (I - \eta_t\mathcal{T}_{\mathbf{x}})(\nu_t - \mu_t) + \eta_t\left\{(\mathcal{T}_\rho\mu_t - \mathcal{S}_\rho^* f_\rho) - (\mathcal{T}_{\mathbf{x}}\mu_t - \mathcal{S}_{\mathbf{x}}^*\mathbf{y})\right\},\end{aligned}$$

which is exactly

$$\nu_{t+1} - \mu_{t+1} = (I - \eta_t\mathcal{T}_{\mathbf{x}})(\nu_t - \mu_t) + \eta_t N_t.$$

Applying this relationship iteratively, with $\nu_1 = \mu_1 = 0$,

$$\nu_{t+1} - \mu_{t+1} = \Pi_1^t(\mathcal{T}_{\mathbf{x}})(\nu_1 - \mu_1) + \sum_{k=1}^{t}\eta_k\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})N_k = \sum_{k=1}^{t}\eta_k\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})N_k. \tag{39}$$

By (24), we have

$$\|\mathcal{S}_\rho\nu_{t+1} - \mathcal{S}_\rho\mu_{t+1}\|_\rho = \left\|\sum_{k=1}^{t}\eta_k\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})N_k\right\|_H,$$

which leads to the desired result (37). ■

The above lemma shows that in order to upper bound $\mathbb{E}[\|\mathcal{S}_\rho\mu_t - \mathcal{S}_\rho\nu_t\|_\rho^2]$, one may only need to bound $\left\|\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})N_k\right\|_H$. A detailed look at this latter term indicates that one may analyze the terms $\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})$ and N_k separately, since $\mathbb{E}_{\mathbf{z}}[N_k]=0$ and the properties of the deterministic sequence $\{\mu_k\}_k$ have been derived in Section 5.1. Moreover, to exploit the capacity condition from Assumption 3, we estimate $\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}N_k\|_H$ (with $\lambda>0$ properly chosen later), rather than $\|N_k\|_H$, as follows.

Lemma 18 *Under Assumptions 1, 2 and 3, let $\{N_t\}_t$ be as in (38). Then for any fixed $\lambda>0$, and $T\geq 2$,*
1) if $\zeta\geq 1/2$, with probability at least $1-\delta_1$, the following holds for all $k\in\mathbb{N}$:

$$\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}N_k\|_H \leq 4(R\kappa^{2\zeta}+\sqrt{M})\left(\frac{\kappa}{m\sqrt{\lambda}}+\frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right)\log\frac{4}{\delta_1}. \tag{40}$$

2) if $\zeta\in]0,1/2[$, with probability at least $1-\delta_1$, the following holds for all $k\in[T]$:

$$\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}N_k\|_H \leq 2\left(3\|f_{\mathcal{H}}\|_\infty+2\sqrt{M}+\kappa R\right)\left(\frac{\kappa}{m\sqrt{\lambda}}+\frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right)\log\frac{3T}{\delta_1}$$
$$+\frac{2\kappa^2R\left(\sum_{i=1}^k\eta_i\right)^{\frac{1}{2}-\zeta}}{m\sqrt{\lambda}}\log\frac{3T}{\delta_1}+\frac{2\kappa R}{\sqrt{m\lambda}}\left(\frac{1}{\sum_{i=1}^k\eta_i}\right)^{\zeta}\log\frac{3T}{\delta_1}. \tag{41}$$

Proof We will apply Bernstein inequality from Lemma 11 to prove the result.
Attainable Case: $\zeta\geq 1/2$. See Appendix C for the proof.
Non-attainable case: $0<\zeta<1/2$.
Let $w_i=(f_{\mathcal{H}}(x_i)-y_i)(\mathcal{T}_\rho+\lambda)^{-1/2}x_i$, for all $i\in[m]$. Noting that by (27), and taking the expectation with respect to the random variable (x,y) (from the distribution ρ),

$$\mathbb{E}[\omega]=\mathbb{E}[(f_{\mathcal{H}}(x)-f_\rho(x))(\mathcal{T}_\rho+\lambda)^{-1/2}x]=0.$$

Applying Hölder's inequality, for any $l\geq 2$,

$$\begin{aligned}&\mathbb{E}[\|w-\mathbb{E}[w]\|_H^l]=\mathbb{E}[\|w\|_H^l]\leq 2^{l-1}\mathbb{E}[(|f_{\mathcal{H}}(x)|^l+|y|^l)\|(\mathcal{T}_\rho+\lambda)^{-1/2}x\|_H^l]\\&\leq 2^{l-1}\int_X(\|f_{\mathcal{H}}\|_\infty^l+\int_Y|y|^ld\rho(y|x))\|(\mathcal{T}_\rho+\lambda)^{-1/2}x\|_H^l d\rho_X(x).\end{aligned}$$

Using Cauchy-Schwarz's inequality and Assumption 1 which implies,

$$\int_Y y^l d\rho(y|x)\leq\left(\int_Y|y|^{2l}d\rho(y|x)\right)^{\frac{1}{2}}\leq\sqrt{l!M^lv}\leq l!(\sqrt{M})^l\sqrt{v}, \tag{42}$$

we get

$$\mathbb{E}[\|w-\mathbb{E}[w]\|_H^l]\leq 2^{l-1}(\|f_{\mathcal{H}}\|_\infty^l+l!(\sqrt{M})^l\sqrt{v})\int_X\|(\mathcal{T}_\rho+\lambda)^{-1/2}x\|_H^l d\rho_X(x). \tag{43}$$

By Assumption (3),

$$\|(\mathcal{T}_\rho+\lambda I)^{-\frac{1}{2}}x\|_H\leq\frac{\|x\|_H}{\sqrt{\lambda}}\leq\frac{\kappa}{\sqrt{\lambda}}. \tag{44}$$

Besides, using the fact that $\mathbb{E}[\|\xi\|_H^2] = \mathbb{E}[\mathrm{tr}(\xi \otimes \xi)] = \mathrm{tr}(\mathbb{E}[\xi \otimes \xi])$ and $\mathbb{E}[x \otimes x] = \mathcal{T}_\rho$, we know that

$$\int_X \|(\mathcal{T}_\rho + \lambda I)^{-\frac{1}{2}} x\|_H^2 d\rho_X(x) = \mathrm{tr}((\mathcal{T}_\rho + \lambda I)^{-\frac{1}{2}} \mathcal{T}_\rho (\mathcal{T}_\rho + \lambda I)^{-\frac{1}{2}}) = \mathrm{tr}((\mathcal{T}_\rho + \lambda I)^{-1} \mathcal{T}_\rho),$$

and as a result of the above and Assumption 3,

$$\int_X \|(\mathcal{T}_\rho + \lambda I)^{-\frac{1}{2}} x\|_H^2 d\rho_X(x) \leq c_\gamma \lambda^{-\gamma}.$$

It thus follows that

$$\int_X \|(\mathcal{T}_\rho + \lambda)^{-1/2} x\|_H^l d\rho_X(x) \leq \left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2} \int_X \|(\mathcal{T}_\rho + \lambda)^{-1/2} x\|_H^2 d\rho_X(x) \leq \left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2} c_\gamma \lambda^{-\gamma}. \quad (45)$$

Plugging the above inequality into (43),

$$\begin{aligned} \mathbb{E}[\|w - \mathbb{E}[w]\|_H^l] &\leq 2^{l-1}(\|f_{\mathcal{H}}\|_\infty^l + l!(\sqrt{M})^l \sqrt{v}) \left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2} c_\gamma \lambda^{-\gamma} \\ &\leq \frac{1}{2} l! \left(\frac{2\kappa(\|f_{\mathcal{H}}\|_\infty + \sqrt{M})}{\sqrt{\lambda}}\right)^{l-2} 4 c_\gamma \sqrt{v} (\|f_{\mathcal{H}}\|_\infty + \sqrt{M})^2 \lambda^{-\gamma}. \end{aligned}$$

Therefore, using Lemma 11, we get that with probability at least $1 - \delta$,

$$\begin{aligned} \left\|(\mathcal{T}_\rho + \lambda)^{-\frac{1}{2}} \frac{1}{m} \sum_{i=1}^m (f_{\mathcal{H}}(x_i) - y_i) x_i \right\|_H &= \left\| \frac{1}{m} \sum_{i=1}^m (\mathbb{E}[w_i] - w_i) \right\|_H \\ &\leq 4(\sqrt{M} + \|f_{\mathcal{H}}\|_\infty) \left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{\sqrt{v} c_\gamma}}{\sqrt{m \lambda^\gamma}} \right) \log \frac{2}{\delta}. \end{aligned} \quad (46)$$

We next let $\xi_i = (\mathcal{T}_\rho + \lambda)^{-1/2}(\langle \mu_k, x_i \rangle - f_{\mathcal{H}}(x_i)) x_i$, for all $i \in [m]$. We assume that $k \geq 2$. (The proof for the case $k = 1$ is simpler as $\mu_1 = 0$.) It is easy to see that the expectation of each ξ_i with respect to the random variable (x_i, y_i) is

$$\mathbb{E}[\xi] = (\mathcal{T}_\rho + \lambda)^{-1/2}(\mathcal{T}_\rho \mu_k - \mathcal{S}_\rho^* f_{\mathcal{H}}) = (\mathcal{T}_\rho + \lambda)^{-1/2}(\mathcal{T}_\rho \mu_k - \mathcal{S}_\rho^* f_\rho),$$

and

$$\|\xi\|_H \leq (\|\mathcal{S}_\rho \mu_k\|_\infty + \|f_{\mathcal{H}}\|_\infty) \|(\mathcal{T}_\rho + \lambda)^{-1/2} x\|_H.$$

By Assumption (3), $\|\mathcal{S}_\rho \mu_k\|_\infty \leq \kappa \|\mu_k\|_H$. It thus follows from the above and (44) that

$$\|\xi\|_H \leq (\kappa \|\mu_k\|_H + \|f_{\mathcal{H}}\|_\infty) \frac{\kappa}{\sqrt{\lambda}}.$$

Besides,

$$\mathbb{E}\|\xi\|_H^2 \leq \frac{\kappa^2}{\lambda} \mathbb{E}(\mu_k(x) - f_{\mathcal{H}}(x))^2 = \frac{\kappa^2}{\lambda} \|\mathcal{S}_\rho \mu_k - f_{\mathcal{H}}\|_\rho^2 \leq \frac{\kappa^2 R^2}{\lambda} \left(\frac{\zeta}{2 \sum_{i=1}^{k-1} \eta_i} \right)^{2\zeta} \leq \frac{\kappa^2 R^2}{\lambda} \left(\frac{1}{\sum_{i=1}^k \eta_i} \right)^{2\zeta},$$

where for the last inequality, we used (33). Applying Lemma 11 and (36), we get that with probability at least $1-\delta$,

$$
\begin{aligned}
&\left\|(\mathcal{T}_\rho+\lambda)^{-1/2}[\frac{1}{m}\sum_{i=1}^m(\mu_k(x_i)-f_{\mathcal{H}}(x_i))x_i-(\mathcal{T}_\rho\mu_k-\mathcal{S}_\rho^* f_\rho)]\right\|_H \\
\leq\ & 2\kappa\left(\frac{\kappa\|\mu_k\|_H+\|f_{\mathcal{H}}\|_\infty}{m\sqrt{\lambda}}+\frac{R}{\sqrt{m\lambda}}\left(\frac{1}{\sum_{i=1}^k\eta_i}\right)^{\zeta}\right)\log\frac{2}{\delta} \\
\leq\ & 2\kappa\left(\frac{\kappa R+\|f_{\mathcal{H}}\|_\infty}{m\sqrt{\lambda}}+\frac{\kappa R\left(\sum_{i=1}^k\eta_i\right)^{\frac{1}{2}-\zeta}}{m\sqrt{\lambda}}+\frac{R}{\sqrt{m\lambda}}\left(\frac{1}{\sum_{i=1}^k\eta_i}\right)^{\zeta}\right)\log\frac{2}{\delta}.
\end{aligned}
$$

Introducing the above estimate and (46) into the following inequality

$$
\begin{aligned}
\|(\mathcal{T}_\rho+\lambda)^{-1/2}N_k\|_H \leq & \left\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}\frac{1}{m}\sum_{i=1}^m(f_{\mathcal{H}}(x_i)-y_i)\,x_i\right\|_H \\
&+\left\|(\mathcal{T}_\rho+\lambda)^{-1/2}[\frac{1}{m}\sum_{i=1}^m(\mu_k(x_i)-f_{\mathcal{H}}(x_i))x_i-(\mathcal{T}_\rho\mu_k-\mathcal{S}_\rho^* f_\rho)]\right\|_H,
\end{aligned}
$$

and then substituting with (36), by a simple calculation, one can prove the desired result by scaling δ. ■

The next lemma is from Rudi et al. (2015), and is derived applying a recent Bernstein inequality from (Tropp, 2012; Minsker, 2011) for a sum of random operators.

Lemma 19 *Let $\delta_2\in(0,1)$ and $\frac{9\kappa^2}{m}\log\frac{m}{\delta_2}\leq\lambda\leq\|\mathcal{T}_\rho\|$. Then the following holds with probability at least $1-\delta_2$,*

$$
\|(\mathcal{T}_{\mathbf{x}}+\lambda I)^{-\frac{1}{2}}\mathcal{T}_\rho^{\frac{1}{2}}\|\leq\|(\mathcal{T}_{\mathbf{x}}+\lambda I)^{-\frac{1}{2}}(\mathcal{T}_\rho+\lambda I)^{\frac{1}{2}}\|\leq 2. \tag{47}
$$

Now we are in a position to estimate the sample variance.

Proposition 3 *Under Assumptions 1, 2 and 3, let $\eta_1\kappa^2\leq 1$ and $0<\lambda\leq\|\mathcal{T}_\rho\|$. Assume that (47) holds. Then for all $t\in[T]$:*
1) if $\zeta\geq 1/2$, and (40) hold, then for $t\in\mathbb{N}$,

$$
\begin{aligned}
&\|\mathcal{S}_\rho\nu_{t+1}-\mathcal{S}_\rho\mu_{t+1}\|_\rho \\
\leq& 4(R\kappa^{2\zeta}+\sqrt{M})\left(\frac{\kappa}{m\sqrt{\lambda}}+\frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right)\left(\sum_{k=1}^{t-1}\frac{2\eta_k}{\sum_{i=k+1}^t\eta_i}+4\lambda\sum_{k=1}^{t-1}\eta_k+\sqrt{2}\kappa^2\eta_t\right)\log\frac{4}{\delta_1}.
\end{aligned}
\tag{48}
$$

2) if $\zeta < 1/2$, and (41) hold for any $t \in [T]$, then for $t \in [T]$:

$$\|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho \leq \left(\sum_{k=1}^{t-1} \frac{2\eta_k}{\sum_{i=k+1}^t \eta_i} + 4\lambda \sum_{k=1}^{t-1} \eta_k + \sqrt{2}\kappa^2 \eta_t \right)$$
$$\times \left(2\left(3\|f_{\mathcal{H}}\|_\infty + 3\sqrt{M} + \kappa R\right) \left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}} \right) + \frac{2\kappa^2 R \left(\sum_{i=1}^t \eta_i\right)^{\frac{1}{2}-\zeta}}{m\sqrt{\lambda}} \right) \log \frac{3T}{\delta_1}$$
$$+ \frac{2\kappa R}{\sqrt{m\lambda}} \log \frac{3T}{\delta_1} \left(\sum_{k=1}^{t-1} \frac{2\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta \sum_{i=k+1}^t \eta_i} + 4\lambda \sum_{k=1}^{t-1} \frac{\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta} + \frac{\sqrt{2}\kappa^2 \eta_t}{\left(\sum_{i=1}^t \eta_i\right)^\zeta} \right). \quad (49)$$

Proof For notational simplicity, we let $\mathcal{T}_{\rho,\lambda} = \mathcal{T}_\rho + \lambda I$ and $\mathcal{T}_{\mathbf{x},\lambda} = \mathcal{T}_{\mathbf{x}} + \lambda I$. Note that by Lemma 17, we have (37). When $k \in [t-1]$, by rewriting $\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k$ as

$$\mathcal{T}_\rho^{\frac{1}{2}} \mathcal{T}_{\mathbf{x},\lambda}^{-\frac{1}{2}} \mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}} \mathcal{T}_{\mathbf{x},\lambda}^{-\frac{1}{2}} \mathcal{T}_{\rho,\lambda}^{\frac{1}{2}} \mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k,$$

we can upper bound $\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\|_H$ as

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\|_H \leq \|\mathcal{T}_\rho^{\frac{1}{2}} \mathcal{T}_{\mathbf{x},\lambda}^{-\frac{1}{2}}\| \|\mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}}\| \|\mathcal{T}_{\mathbf{x},\lambda}^{-\frac{1}{2}} \mathcal{T}_{\rho,\lambda}^{\frac{1}{2}}\| \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k\|_H.$$

Applying (47), the above can be relaxed as

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\|_H \leq 4\|\mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \mathcal{T}_{\mathbf{x},\lambda}^{\frac{1}{2}}\| \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k\|_H,$$

which is equivalent to

$$\|\mathcal{T}_{\rho,\lambda}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\|_H \leq 4\|\mathcal{T}_{\mathbf{x},\lambda} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k\|_H.$$

Thus, following from $\eta_k \kappa^2 \leq 1$ which implies $\eta_k \|\mathcal{T}_{\mathbf{x}}\| \leq 1$,

$$\begin{aligned} \|\mathcal{T}_{\mathbf{x},\lambda} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| &\leq \|\mathcal{T}_{\mathbf{x}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| + \|\lambda \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \\ &\leq \|\mathcal{T}_{\mathbf{x}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| + \lambda. \end{aligned}$$

Applying Lemma 15 with $\zeta = 1$ to bound $\|\mathcal{T}_{\mathbf{x}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|$, we get

$$\|\mathcal{T}_{\mathbf{x},\lambda} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \leq \frac{1}{\mathrm{e} \sum_{j=k+1}^t \eta_j} + \lambda.$$

When $k = t$,

$$\begin{aligned} \|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\|_H = \|\mathcal{T}_\rho^{\frac{1}{2}} N_t\|_H &\leq \|\mathcal{T}_\rho^{\frac{1}{2}}\| \|\mathcal{T}_{\rho,\lambda}^{\frac{1}{2}}\| \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_t\|_H \\ &\leq \|\mathcal{T}_\rho\|^{\frac{1}{2}} (\|\mathcal{T}_\rho\| + \lambda)^{\frac{1}{2}} \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_t\|_H. \end{aligned}$$

Since $\lambda \leq \|\mathcal{T}_\rho\| \leq \mathrm{tr}(\mathcal{T}_\rho) \leq \kappa^2$, we derive

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_t\|_H \leq \sqrt{2}\kappa^2 \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_t\|_H.$$

From the above analysis, we see that $\sum_{k=1}^t \eta_k \left\| \mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k \right\|_H$ can be upper bounded by

$$\leq \left(\sum_{k=1}^{t-1} \frac{\eta_k/2 \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k\|_H}{\sum_{i=k+1}^t \eta_i} + \lambda \sum_{k=1}^{t-1} \eta_k \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_k\|_H + \sqrt{2}\kappa^2 \eta_t \|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} N_t\|_H \right).$$

Plugging (40) (or (41)) into the above, and then combining with (37), we get the desired bound (48) (or (49)). The proof is complete. ■

Setting $\eta_t = \eta_1 t^{-\theta}$ in the above proposition, with the basic estimates from Section 4, we get the following explicit bounds for the sample variance.

Proposition 4 *Under Assumptions 1, 2 and 3, let $\eta_t = \eta_1 t^{-\theta}$ with $\eta_1 \in]0, \kappa^{-2}]$ and $\theta \in [0,1[$. Assume that (47) holds. Then the following holds for all $t \in [T]$ and any $0 < \lambda \leq \|\mathcal{T}_\rho\|$:*
1) If $\zeta \geq 1/2$, and (40) holds for all $t \in [T]$,

$$\begin{aligned} &\|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho \\ &\leq 4(R\kappa^{2\zeta} + \sqrt{M}) \left(\frac{8\lambda\eta_1 t^{1-\theta}}{1-\theta} + 4\log t + 4 + \sqrt{2}\eta_1\kappa^2 \right) \left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}} \right) \log \frac{4}{\delta_1}. \end{aligned} \tag{50}$$

2) If $\zeta < 1/2$, and (41) holds for all $t \in [T]$,

$$\begin{aligned} \|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho \leq &\left(\frac{8\lambda\eta_1 t^{1-\theta}}{1-\theta} + 4\log t + 4 + \sqrt{2}\eta_1\kappa^2 \right) \log \frac{3T}{\delta_1} \\ \times &\left(2\left(3\|f_{\mathcal{H}}\|_\infty + 3\sqrt{M} + \kappa R\right) \left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}} \right) + \left(\frac{\kappa}{1-\theta} \sqrt{\frac{\eta_1 t^{1-\theta}}{m}} + 1 \right) \frac{4\kappa R}{\sqrt{m\lambda}} \frac{1}{(\eta_1 t^{1-\theta})^\zeta} \right). \end{aligned} \tag{51}$$

Proof By Proposition 3, we have (48) or (49). Note that

$$\sum_{k=1}^{t-1} \frac{\eta_k}{\sum_{i=k+1}^t \eta_i} = \sum_{k=1}^{t-1} \frac{k^{-\theta}}{\sum_{i=k+1}^t i^{-\theta}} \leq \sum_{k=1}^{t-1} \frac{k^{-\theta}}{(t-k)t^{-\theta}}.$$

Applying Lemma 14, we get

$$\sum_{k=1}^{t-1} \frac{\eta_k}{\sum_{i=k+1}^t \eta_i} \leq 2 + 2\log t,$$

and by Lemma 12,

$$\sum_{k=1}^{t-1} \eta_k = \eta_1 \sum_{k=1}^{t-1} k^{-\theta} \leq \frac{2\eta_1 t^{1-\theta}}{1-\theta}.$$

Introducing the last two estimates into (48) and (49), one can get (50) and that

$$\|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho \leq \left(\frac{8\lambda\eta_1 t^{1-\theta}}{1-\theta} + 4\log t + 4 + \sqrt{2}\eta_1\kappa^2 \right)$$

$$\times \left(2\left(3\|f_{\mathcal{H}}\|_\infty + 3\sqrt{M} + \kappa R\right)\left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right) + \frac{4\kappa^2 R\left(\eta_1 t^{1-\theta}\right)^{\frac{1}{2}-\zeta}}{(1-\theta)m\sqrt{\lambda}} \right) \log\frac{3T}{\delta_1}$$

$$+ \frac{2\kappa R}{\sqrt{m\lambda}}\log\frac{3T}{\delta_1}\left(\sum_{k=1}^{t-1} \frac{2\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta \sum_{i=k+1}^t \eta_i} + 4\lambda \sum_{k=1}^{t-1} \frac{\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta} + \frac{\sqrt{2}\kappa^2\eta_t}{\left(\sum_{i=1}^t \eta_i\right)^\zeta} \right).$$

To prove (51), it remains to estimate the last term of the above. Again, using Lemmas 12, 13 and 14, we get

$$\begin{aligned} \sum_{k=1}^{t-1} \frac{\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta \sum_{i=k+1}^t \eta_i} &= \frac{1}{\eta_1^\zeta}\sum_{k=1}^{t-1} \frac{k^{-\theta}}{\left(\sum_{i=1}^k i^{-\theta}\right)^\zeta \sum_{i=k+1}^t i^{-\theta}} \\ \leq \quad & \frac{1}{\eta_1^\zeta}\sum_{k=1}^{t-1}\frac{k^{-\theta}}{(k^{1-\theta}/2)^\zeta (t-k)t^{-\theta}} = \frac{2^\zeta}{\eta_1^\zeta} t^\theta \sum_{k=1}^{t-1} \frac{k^{-(\theta+\zeta(1-\theta))}}{t-k} \\ \leq \quad & \frac{2^\zeta}{\eta_1^\zeta} t^\theta 2 t^{-(\theta+\zeta(1-\theta))}(1+\log t) \leq \frac{4(1+\log t)}{(\eta_1 t^{1-\theta})^\zeta}, \end{aligned}$$

$$\sum_{k=1}^{t-1} \frac{\eta_k}{\left(\sum_{i=1}^k \eta_i\right)^\zeta} = \eta_1^{1-\zeta}\sum_{k=1}^{t-1} \frac{k^{-\theta}}{\left(\sum_{i=1}^k i^{-\theta}\right)^\zeta} \leq 2^\zeta \eta_1^{1-\zeta}\sum_{k=1}^{t-1} k^{-(\theta+\zeta(1-\theta))} \leq \frac{2(\eta_1 t^{1-\theta})^{1-\zeta}}{(1-\theta)}, \quad \text{and}$$

$$\frac{\eta_t}{\left(\sum_{i=1}^t \eta_i\right)^\zeta} = \frac{\eta_1 t^{-\theta}}{\left(\sum_{i=1}^t \eta_1 i^{-\theta}\right)^\zeta} \leq 2^\zeta \frac{\eta_1 t^{-\theta}}{(\eta_1 t^{1-\theta})^\zeta} \leq \frac{\sqrt{2}\eta_1}{(\eta_1 t^{1-\theta})^\zeta}.$$

Therefore,

$$\|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho \leq \left(\frac{8\lambda\eta_1 t^{1-\theta}}{1-\theta} + 4\log t + 4 + \sqrt{2}\eta_1\kappa^2 \right)$$

$$\times \left(2\left(3\|f_{\mathcal{H}}\|_\infty + 3\sqrt{M} + \kappa R\right)\left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right) + \frac{4\kappa^2 R\left(\eta_1 t^{1-\theta}\right)^{\frac{1}{2}-\zeta}}{(1-\theta)m\sqrt{\lambda}} \right) \log\frac{3T}{\delta_1}$$

$$+ \frac{2\kappa R}{\sqrt{m\lambda}}\log\frac{3T}{\delta_1}\left(8 + 8\log t + \frac{8\lambda\eta_1 t^{1-\theta}}{1-\theta} + 2\kappa^2\eta_1\right)\frac{1}{(\eta_1 t^{1-\theta})^\zeta}.$$

Rearranging terms, we can prove the second part. ■

In conclusion, we get the following result for the sample variance.

Theorem 5 *Under Assumptions 1, 2 and 3, let $\delta_1, \delta_2 \in]0,1[$ and $\frac{9\kappa^2}{m}\log\frac{m}{\delta_2} \leq \lambda \leq \|\mathcal{T}_\rho\|$. Let $\eta_t = \eta_1 t^{-\theta}$ for all $t \in [T]$, with $\eta_1 \in]0, \kappa^{-2}]$ and $\theta \in [0,1[$. Then with probability at least $1-\delta_1-\delta_2$, the following holds for all $t \in [T]$:*
1) if $\zeta \geq 1/2$, we have (50).
2) if $\zeta < 1/2$, we have (51).

6. Estimating Computational Variance

In this section, we estimate the computational variance, $\mathbb{E}[\|\mathcal{S}_\rho\omega_t - \mathcal{S}_\rho\nu_t\|_\rho^2]$. For this, a series of lemmas is introduced.

6.1 Cumulative Error

We have the following lemma, which shows that the computational variance can be controlled by a sum of weighted empirical risks.

Lemma 20 *We have*

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_{t+1} - \mathcal{S}_\rho\nu_{t+1}\|_\rho^2 \leq \frac{\kappa^2}{b}\sum_{k=1}^{t}\eta_k^2\left\|\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_k)]. \tag{52}$$

Proof Since ω_{t+1} and ν_{t+1} are given by (4) and (26), respectively,

$$\begin{aligned}\omega_{t+1} - \nu_{t+1} &= (\omega_t - \nu_t) + \eta_t\left\{(\mathcal{T}_{\mathbf{x}}\nu_t - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - \frac{1}{b}\sum_{i=b(t-1)+1}^{bt}(\langle\omega_t, x_{j_i}\rangle_H - y_{j_i})x_{j_i}\right\}\\ &= (I - \eta_t\mathcal{T}_{\mathbf{x}})(\omega_t - \nu_t) + \frac{\eta_t}{b}\sum_{i=b(t-1)+1}^{bt}\left\{(\mathcal{T}_{\mathbf{x}}\omega_t - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - (\langle\omega_t, x_{j_i}\rangle_H - y_{j_i})x_{j_i}\right\}.\end{aligned}$$

Applying this relationship iteratively,

$$\omega_{t+1} - \nu_{t+1} = \Pi_1^t(\mathcal{T}_{\mathbf{x}})(\omega_1 - \nu_1) + \frac{1}{b}\sum_{k=1}^{t}\sum_{i=b(k-1)+1}^{bk}\eta_k\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})M_{k,i},$$

where we denote

$$M_{k,i} = (\mathcal{T}_{\mathbf{x}}\omega_k - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - (\langle\omega_k, x_{j_i}\rangle_H - y_{j_i})x_{j_i}. \tag{53}$$

Since $\omega_1 = \nu_1 = 0$, then

$$\omega_{t+1} - \nu_{t+1} = \frac{1}{b}\sum_{k=1}^{t}\sum_{i=b(k-1)+1}^{bk}\eta_k\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})M_{k,i}.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_{t+1} - \mathcal{S}_\rho\nu_{t+1}\|_\rho^2 &= \frac{1}{b^2}\mathbb{E}_{\mathbf{J}}\left\|\sum_{k=1}^{t}\sum_{i=b(k-1)+1}^{bk}\eta_k\mathcal{S}_\rho\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})M_{k,i}\right\|_\rho^2\\ &= \frac{1}{b^2}\sum_{k=1}^{t}\sum_{i=b(k-1)+1}^{bk}\eta_k^2\mathbb{E}_{\mathbf{J}}\left\|\mathcal{S}_\rho\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})M_{k,i}\right\|_\rho^2,\end{aligned} \tag{54}$$

where for the last equality, we use the fact that if $k \neq k'$, or $k = k'$ but $i \neq i'$[4], then

$$\mathbb{E}_{\mathbf{J}}\langle\mathcal{S}_\rho\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})M_{k,i}, \mathcal{S}_\rho\Pi_{k'+1}^t(\mathcal{T}_{\mathbf{x}})M_{k',i'}\rangle_\rho = 0.$$

4. This is possible only when $b \geq 2$.

Indeed, if $k \neq k'$, without loss of generality, we consider the case $k < k'$. Recalling that $M_{k,i}$ is given by (53) and that given any $\mathbf{z}$, ω_k is depending only on $\mathbf{J}_1, \cdots, \mathbf{J}_{k-1}$, we thus have

$$\begin{aligned}&\mathbb{E}_{\mathbf{J}} \langle \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i}, \mathcal{S}_\rho \Pi_{k'+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k',i'} \rangle_\rho \\ &= \mathbb{E}_{\mathbf{J}_1, \cdots, \mathbf{J}_{k'-1}} \langle \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i}, \mathcal{S}_\rho \Pi_{k'+1}^t(\mathcal{T}_{\mathbf{x}}) \mathbb{E}_{\mathbf{J}_{k'}}[M_{k',i'}] \rangle_\rho = 0.\end{aligned}$$

If $k = k'$ but $i \neq i'$, without loss of generality, we assume $i < i'$. By noting that ω_k is depending only on $\mathbf{J}_1, \cdots, \mathbf{J}_{k-1}$ and $M_{k,i}$ is depending only on ω_k and z_{j_i} (given any sample $\mathbf{z}$),

$$\begin{aligned}&\mathbb{E}_{\mathbf{J}} \langle \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i}, \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i'} \rangle_\rho \\ &= \mathbb{E}_{\mathbf{J}_1, \cdots, \mathbf{J}_{k-1}} \langle \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \mathbb{E}_{j_i}[M_{k,i}], \mathcal{S}_\rho \Pi_{k'+1}^t(\mathcal{T}_{\mathbf{x}}) \mathbb{E}_{j_{i'}}[M_{k,i'}] \rangle_\rho = 0.\end{aligned}$$

Using the isometry property (24) to (54),

$$\mathbb{E}_{\mathbf{J}} \left\| \mathcal{S}_\rho \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i} \right\|_\rho^2 = \mathbb{E}_{\mathbf{J}} \left\| \mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) M_{k,i} \right\|_H^2 \leq \left\| \mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \right\|^2 \mathbb{E}_{\mathbf{J}} \|M_{k,i}\|_H^2 ,$$

and by applying the inequality $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_H^2] \leq \mathbb{E}[\|\xi\|_H^2]$,

$$\mathbb{E}_{\mathbf{J}} \|M_{k,i}\|_H^2 \leq \mathbb{E}_{\mathbf{J}} \|(\langle \omega_k, x_{j_i} \rangle_H - y_{j_i}) x_{j_i}\|_H^2 \leq \kappa^2 \mathbb{E}_{\mathbf{J}}[(\langle \omega_k, x_{j_i} \rangle_H - y_{j_i})^2] = \kappa^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_k)],$$

where for the last inequality we use (3). Therefore, we can get the desired result. ■

To estimate the computational variance from (52), we need to further develop upper bounds for the empirical risks and the weighted factors, which will be given in the following two subsections.

6.2 Bounding the Empirical Risk

This subsection is devoted to upper bounding $\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)]$. The process relies on some tools from convex analysis and a decomposition related to the weighted averages and the last iterates from (Shamir and Zhang, 2013; Lin et al., 2016b). We begin by introducing the following lemma, a fact based on the square loss' special properties.

Lemma 21 *Given any sample* $\mathbf{z}$, *and* $l \in \mathbb{N}$, *let* $\omega \in H$ *be independent from* $\mathbf{J}_l$, *then*

$$\eta_l \left(\mathcal{E}_{\mathbf{z}}(\omega_l) - \mathcal{E}_{\mathbf{z}}(\omega) \right) \leq \|\omega_l - \omega\|_H^2 - \mathbb{E}_{\mathbf{J}_l} \|\omega_{l+1} - \omega\|_H^2 + \eta_l^2 \kappa^2 \mathcal{E}_{\mathbf{z}}(\omega_l). \quad (55)$$

Proof Since ω_{t+1} is given be (4), subtracting both sides of (4) by ω, taking the square H-norm, and expanding the inner product,

$$\begin{aligned}\|\omega_{l+1} - \omega\|_H^2 = \|\omega_l - \omega\|_H^2 + \frac{\eta_l^2}{b^2} \left\| \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i}) x_{j_i} \right\|_H^2 \\ + \frac{2\eta_l}{b} \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i}) \langle \omega - \omega_l, x_{j_i} \rangle_H.\end{aligned}$$

By Assumption (3), $\|x_{j_i}\|_H \leq \kappa$, and thus

$$\left\| \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i}) x_{j_i} \right\|_H^2 \leq \left(\sum_{i=b(l-1)+1}^{bl} |\langle \omega_l, x_{j_i} \rangle_H - y_{j_i}| \kappa \right)^2 \leq \kappa^2 b \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i})^2,$$

where for the last inequality, we used Cauchy-Schwarz inequality. Thus,

$$\|\omega_{l+1} - \omega\|_H^2 \leq \|\omega_l - \omega\|_H^2 + \frac{\eta_l^2 \kappa^2}{b} \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i})^2 + \frac{2\eta_l}{b} \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i})(\langle \omega, x_{j_i} \rangle_H - \langle \omega_l, x_{j_i} \rangle_H).$$

Using the basic inequality $a(b-a) \leq (b^2 - a^2)/2, \forall a, b \in \mathbb{R}$,

$$\|\omega_{l+1} - \omega\|_H^2 \leq \|\omega_l - \omega\|_H^2 + \frac{\eta_l^2 \kappa^2}{b} \sum_{i=b(l-1)+1}^{bl} (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i})^2 + \frac{\eta_l}{b} \sum_{i=b(l-1)+1}^{bl} \left((\langle \omega, x_{j_i} \rangle_H - y_{j_i})^2 - (\langle \omega_l, x_{j_i} \rangle_H - y_{j_i})^2 \right).$$

Noting that ω_l and ω are independent from $\mathbf{J}_l$, and taking the expectation on both sides with respect to $\mathbf{J}_l$,

$$\mathbb{E}_{\mathbf{J}_l} \|\omega_{l+1} - \omega\|_H^2 \leq \|\omega_l - \omega\|_H^2 + \eta_l^2 \kappa^2 \mathcal{E}_{\mathbf{z}}(\omega_l) + \eta_l \left(\mathcal{E}_{\mathbf{z}}(\omega) - \mathcal{E}_{\mathbf{z}}(\omega_l) \right),$$

which leads to the desired result by rearranging terms. The proof is complete. ■

Using the above lemma and a decomposition related to the weighted averages and the last iterates from (Shamir and Zhang, 2013; Lin et al., 2016b), we can prove the following relationship.

Lemma 22 *Let $\eta_1 \kappa^2 \leq 1/2$ for all $t \in \mathbb{N}$. Then*

$$\eta_t \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)] \leq 4\mathcal{E}_{\mathbf{z}}(0) \frac{1}{t} \sum_{l=1}^{t} \eta_l + 2\kappa^2 \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)]. \tag{56}$$

Proof For $k = 1, \cdots, t-1$,

$$\begin{aligned}
&\frac{1}{k} \sum_{i=t-k+1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] - \frac{1}{k+1} \sum_{i=t-k}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] \\
= &\frac{1}{k(k+1)} \left\{ (k+1) \sum_{i=t-k+1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] - k \sum_{i=t-k}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] \right\} \\
= &\frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} (\eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] - \eta_{t-k} \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_{t-k})]).
\end{aligned}$$

Summing over $k = 1, \cdots, t-1$, and rearranging terms, we get (Lin et al., 2016b)

$$\eta_t \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)] = \frac{1}{t}\sum_{i=1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] + \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} (\eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] - \eta_{t-k}\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_{t-k})]).$$

Since $\{\eta_t\}_t$ is decreasing and $\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_{t-k})]$ is non-negative, the above can be relaxed as

$$\eta_t \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)] \leq \frac{1}{t}\sum_{i=1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] + \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i) - \mathcal{E}_{\mathbf{z}}(\omega_{t-k})]. \tag{57}$$

In the rest of the proof, we will upper bound the last two terms of the above.

To bound the first term of the right side of (57), we apply Lemma 21 with $\omega = 0$ to get

$$\eta_l \mathbb{E}_{\mathbf{J}}\left(\mathcal{E}_{\mathbf{z}}(\omega_l) - \mathcal{E}_{\mathbf{z}}(0)\right) \leq \mathbb{E}_{\mathbf{J}}[\|\omega_l\|_H^2 - \|\omega_{l+1}\|_H^2] + \eta_l^2\kappa^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)].$$

Rearranging terms,

$$\eta_l(1-\eta_l\kappa^2)\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)] \leq \mathbb{E}_{\mathbf{J}}[\|\omega_l\|_H^2 - \|\omega_{l+1}\|_H^2] + \eta_l \mathcal{E}_{\mathbf{z}}(0).$$

It thus follows from the above and $\eta_l\kappa^2 \leq 1/2$ that

$$\eta_l \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)]/2 \leq \mathbb{E}_{\mathbf{J}}[\|\omega_l\|_H^2 - \|\omega_{l+1}\|_H^2] + \eta_l \mathcal{E}_{\mathbf{z}}(0).$$

Summing up over $l = 1, \cdots, t$,

$$\sum_{l=1}^{t} \eta_l \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)]/2 \leq \mathbb{E}_{\mathbf{J}}[\|w_1\|_H^2 - \|\omega_{t+1}\|_H^2] + \mathcal{E}_{\mathbf{z}}(0)\sum_{l=1}^{t}\eta_l.$$

Introducing with $\omega_1 = 0$, $\|\omega_{t+1}\|_H^2 \geq 0$, and then multiplying both sides by $2/t$, we get

$$\frac{1}{t}\sum_{l=1}^{t} \eta_l \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_l)] \leq 2\mathcal{E}_{\mathbf{z}}(0)\frac{1}{t}\sum_{l=1}^{t}\eta_l. \tag{58}$$

It remains to bound the last term of (57). Let $k \in [t-1]$ and $i \in \{t-k, \cdots, t\}$. Note that given the sample $\mathbf{z}$, ω_i is depending only on $\mathbf{J}_1, \cdots, \mathbf{J}_{i-1}$ when $i > 1$ and $\omega_1 = 0$. Thus, we can apply Lemma 21 with $\omega = \omega_{t-k}$ to derive

$$\eta_i\left(\mathcal{E}_{\mathbf{z}}(\omega_i) - \mathcal{E}_{\mathbf{z}}(\omega_{t-k})\right) \leq \|\omega_i - \omega_{t-k}\|_H^2 - \mathbb{E}_{\mathbf{J}_i}\|\omega_{i+1} - \omega_{t-k}\|_H^2 + \eta_i^2\kappa^2\mathcal{E}_{\mathbf{z}}(\omega_i).$$

Therefore,

$$\eta_i \mathbb{E}_{\mathbf{J}}\left[\mathcal{E}_{\mathbf{z}}(\omega_i) - \mathcal{E}_{\mathbf{z}}(\omega_{t-k})\right] \leq \mathbb{E}_{\mathbf{J}}[\|\omega_i - \omega_{t-k}\|_H^2 - \|\omega_{i+1} - \omega_{t-k}\|_H^2] + \eta_i^2\kappa^2\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)].$$

Summing up over $i = t-k, \cdots, t$,

$$\sum_{i=t-k}^{t} \eta_i \mathbb{E}_{\mathbf{J}}\left[\mathcal{E}_{\mathbf{z}}(\omega_i) - \mathcal{E}_{\mathbf{z}}(\omega_{t-k})\right] \leq \kappa^2 \sum_{i=t-k}^{t} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)].$$

Note that the left hand side is exactly $\sum_{i=t-k+1}^{t} \eta_i \mathbb{E}_{\mathbf{J}}\left[\mathcal{E}_{\mathbf{z}}(\omega_i) - \mathcal{E}_{\mathbf{z}}(\omega_{t-k})\right]$. We thus know that the last term of (57) can be upper bounded by

$$\begin{aligned} & \kappa^2 \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] \\ = \quad & \kappa^2 \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] + \kappa^2 \eta_t^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)] \sum_{k=1}^{t-1} \frac{1}{k(k+1)}. \end{aligned}$$

Using the fact that

$$\sum_{k=1}^{t-1} \frac{1}{k(k+1)} = \sum_{k=1}^{t-1} \left(\frac{1}{k} - \frac{1}{k+1} \right) = 1 - \frac{1}{t} \leq 1,$$

and $\kappa^2 \eta_t \leq 1/2$, we get that the last term of (57) can be bounded as

$$\begin{aligned} & \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} \eta_i (\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] - \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_{t-k})]) \\ \leq \quad & \kappa^2 \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)] + \eta_t \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)]/2. \end{aligned}$$

Plugging the above and (58) into the decomposition (57), and rearranging terms

$$\eta_t \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)]/2 \leq 2\mathcal{E}_{\mathbf{z}}(0) \frac{1}{t} \sum_{l=1}^{t} \eta_l + \kappa^2 \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)],$$

which leads to the desired result by multiplying both sides by 2. The proof is complete. ∎

We also need the following lemma, whose proof can be done using an induction argument.

Lemma 23 *Let $\{u_t\}_{t=1}^T$, $\{A_t\}_{t=1}^T$ and $\{B_t\}_{t=1}^T$ be three sequences of non-negative numbers such that $u_1 \leq A_1$ and*

$$u_t \leq A_t + B_t \sup_{i \in [t-1]} u_i, \qquad \forall t \in \{2, 3, \cdots, T\}. \tag{59}$$

Let $\sup_{t \in [T]} B_t \leq B < 1$. Then for all $t \in [T]$,

$$\sup_{k \in [t]} u_k \leq \frac{1}{1-B} \sup_{k \in [t]} A_k. \tag{60}$$

Proof When $t = 1$, (60) holds trivially since $u_1 \leq A_1$ and $B < 1$. Now assume for some $t \in \mathbb{N}$ with $2 \leq t \leq T$,

$$\sup_{i \in [t-1]} u_i \leq \frac{1}{1-B} \sup_{i \in [t-1]} A_i.$$

Then, by (59), the above hypothesis, and $B_t \leq B$, we have

$$u_t \leq A_t + B_t \sup_{i \in [t-1]} u_i \leq A_t + \frac{B_t}{1-B} \sup_{i \in [t-1]} A_i \leq \sup_{i \in [t]} A_i \left(1 + \frac{B_t}{1-B} \right) \leq \sup_{i \in [t]} A_i \frac{1}{1-B}.$$

Consequently,

$$\sup_{k\in[t]} u_k \leq \frac{1}{1-B} \sup_{k\in[t]} A_k,$$

thereby showing that indeed (60) holds for t. By mathematical induction, (60) holds for every $t \in [T]$. The proof is complete. ■

Now we can bound $\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_k)]$ as follows.

Lemma 24 *Let $\eta_1\kappa^2 \leq 1/2$ and for all $t \in [T]$ with $t \geq 2$,*

$$\frac{1}{\eta_t}\sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \leq \frac{1}{4\kappa^2}. \tag{61}$$

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

$$\sup_{k\in[t]} \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_k)] \leq 8\mathcal{E}_{\mathbf{z}}(0) \sup_{k\in[t]} \left\{ \frac{1}{\eta_k k} \sum_{l=1}^{k} \eta_l \right\}. \tag{62}$$

Proof By Lemma 22, we have (56). Dividing both sides by η_t, we can relax the inequality as

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)] \leq 4\mathcal{E}_{\mathbf{z}}(0)\frac{1}{\eta_t t}\sum_{l=1}^{t}\eta_l + 2\kappa^2\frac{1}{\eta_t}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t-1}\eta_i^2 \sup_{i\in[t-1]} \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_i)].$$

In Lemma 23, we let $u_t = \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_t)]$, $A_t = 4\mathcal{E}_{\mathbf{z}}(0)\frac{1}{\eta_t t}\sum_{l=1}^{t}\eta_l$ and

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

Condition (61) guarantees that $\sup_{t\in[T]} B_t \leq 1/2$. Thus, (60) holds, and the desired result follows by plugging with $B = 1/2$. The proof is complete. ■

Finally, we need the following lemma to bound $\mathcal{E}_{\mathbf{z}}(0)$, whose proof follows from applying the Bernstein inequality from Lemma 11.

Lemma 25 *Under Assumption 1, with probability at least $1-\delta_3$ ($\delta_3 \in]0,1[$), there holds*

$$\mathcal{E}_{\mathbf{z}}(0) \leq Mv + 2Mv\left(\frac{1}{m} + \frac{\sqrt{2}}{\sqrt{m}}\right)\log\frac{2}{\delta_3}.$$

In particular, if $m \geq 32\log^2\frac{2}{\delta_3}$, then

$$\mathcal{E}_{\mathbf{z}}(0) \leq 2Mv. \tag{63}$$

Proof Following from (5),

$$\int_Z y^{2l} d\rho \leq \frac{1}{2} l! M^{l-2} \cdot (2M^2 v), \qquad \forall l \in \mathbb{N},$$

and

$$\int_Z y^2 d\rho \leq Mv.$$

Therefore,

$$\begin{aligned}
\int_Z |y^2 - \mathbb{E}y^2|^l d\rho &\leq \int_Z \max(|y|^{2l}, (\mathbb{E}y^2)^l) d\rho \\
&\leq \int_Z (|y|^{2l} + (\mathbb{E}y^2)^l) d\rho \\
&\leq \frac{1}{2} l! M^{l-2} \cdot (2M^2 v) + (Mv)^l \\
&\leq \frac{1}{2} l! (Mv)^{l-2} (2Mv)^2,
\end{aligned}$$

where for the last inequality we used $v \geq 1$. Applying Lemma 11, with $\omega_i = y_i^2$ for all $i \in [m]$, $B = Mv$ and $\sigma = 2Mv$, we know that with probability at least $1 - \delta_3$, there holds

$$\frac{1}{n} \sum_{i=1}^n y_i^2 - \int_Z y^2 d\rho \leq 2Mv \left(\frac{1}{m} + \frac{2}{\sqrt{m}} \right) \log \frac{2}{\delta_3}.$$

The proof is complete. ∎

6.3 Bounding $\left\| \mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \right\|$

We bound the weighted factor $\left\| \mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \right\|$ as follows.

Lemma 26 *Assume (47) holds for some $\lambda > 0$ and $\eta_1 \kappa^2 \leq 1$. Then*

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2 \leq \frac{1}{\sum_{i=k+1}^t \eta_i} + 4\lambda.$$

Proof Note that we have

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \leq \|\mathcal{T}_\rho^{\frac{1}{2}} (\mathcal{T}_{\mathbf{x}} + \lambda I)^{-\frac{1}{2}}\| \|(\mathcal{T}_{\mathbf{x}} + \lambda I)^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|.$$

Using (47), we can relax the above as

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \leq 2\|(\mathcal{T}_{\mathbf{x}} + \lambda I)^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|,$$

which leads to

$$\|\mathcal{T}_\rho^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2 \leq 4\|(\mathcal{T}_{\mathbf{x}} + \lambda I)^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2.$$

Since

$$\begin{aligned}
\|(\mathcal{T}_{\mathbf{x}} + \lambda I)^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2 &= \|(\mathcal{T}_{\mathbf{x}} + \lambda I) \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| \\
&\leq \|\mathcal{T}_{\mathbf{x}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\| + \lambda \\
&= \|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2 + \lambda,
\end{aligned}$$

and with $\eta_t\kappa^2 \leq 1$, $\|\mathcal{T}_{\mathbf{x}}\| \leq \mathrm{tr}(\mathcal{T}_{\mathbf{x}}) \leq \kappa^2$, by Lemma 15,

$$\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\|^2 \leq \frac{1}{2\mathrm{e}\sum_{i=k+1}^t \eta_i} \leq \frac{1}{4\sum_{i=k+1}^t \eta_i},$$

we thus derive the desired result. The proof is complete. ∎

6.4 Deriving Error Bounds

With Lemmas 20–26, we are ready to estimate the computational variance , $\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_{t+1} - \mathcal{S}_\rho\nu_{t+1}\|_\rho^2$, as follows.

Proposition 5 *Under Assumption 1, assume (47) holds for some $\lambda > 0$, $\eta_1\kappa^2 \leq 1/2$, (61) and (63). Then, we have for all $t \in [T]$,*

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_{t+1} - \mathcal{S}_\rho\nu_{t+1}\|_\rho^2 \leq \frac{16Mv\kappa^2}{b}\sup_{k\in[t]}\left\{\frac{1}{\eta_k k}\sum_{l=1}^k \eta_l\right\}\left(\sum_{k=1}^{t-1}\frac{\eta_k^2}{\sum_{i=k+1}^t \eta_i} + 4\lambda\sum_{k=1}^{t-1}\eta_k^2 + \eta_t^2\kappa^2\right). \tag{64}$$

Proof According to Lemmas 20 and 24, we have (52) and (62). It thus follows that

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

Now the proof can be finished by applying Lemma 26 which tells us that

$$\begin{aligned}\sum_{k=1}^t \eta_k^2\left\|\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2 &= \sum_{k=1}^{t-1}\eta_k^2\left\|\mathcal{T}_\rho^{\frac{1}{2}}\Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}})\right\|^2 + \eta_t^2\left\|\mathcal{T}_\rho^{\frac{1}{2}}\right\|^2 \\ &\leq \sum_{k=1}^{t-1}\frac{\eta_k^2}{\sum_{i=k+1}^t \eta_i} + 4\lambda\sum_{k=1}^{t-1}\eta_k^2 + \eta_t^2\kappa^2,\end{aligned}$$

and (63) to the above inequality. The proof is complete. ∎

Setting $\eta_t = \eta_1 t^{-\theta}$ for some appropriate η_1 and θ in the above proposition, we get the following explicitly upper bounds for $\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_t - \mathcal{S}_\rho\omega_t\|_\rho^2$.

Proposition 6 *Under Assumption 1, assume (47) holds for some $\lambda > 0$ and (63). Let $\eta_t = \eta_1 t^{-\theta}$ for all $t \in [T]$, with $\theta \in [0,1[$ and*

$$0 < \eta_1 \leq \frac{t^{\min(\theta,1-\theta)}}{8\kappa^2(\log t + 1)}, \qquad \forall t \in [T]. \tag{65}$$

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

$$\mathbb{E}_{\mathbf{J}}\|\omega_{t+1} - \nu_{t+1}\|_\rho^2 \leq \frac{16Mv\kappa^2}{b(1-\theta)}\left(5\eta_1 t^{-\min(\theta,1-\theta)} + 8\lambda\eta_1^2 t^{(1-2\theta)_+}\right)(1 \vee \log t). \tag{66}$$

Proof We will use Proposition 5 to prove the result. Thus, we need to verify the condition (61). Note that

$$\sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 = \sum_{i=1}^{t-1} \eta_i^2 \sum_{k=t-i}^{t-1} \frac{1}{k(k+1)} = \sum_{i=1}^{t-1} \eta_i^2 \left(\frac{1}{t-i} - \frac{1}{t} \right) \leq \sum_{i=1}^{t-1} \frac{\eta_i^2}{t-i}.$$

Substituting with $\eta_i = \eta i^{-\theta}$, and by Lemma 14,

$$\sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \leq \eta_1^2 \sum_{i=1}^{t-1} \frac{i^{-2\theta}}{t-i} \leq 2\eta_1^2 t^{-\min(2\theta,1)} (\log t + 1).$$

Dividing both sides by η_t ($= \eta_1 t^{-\theta}$), and then using (65),

$$\frac{1}{\eta_t} \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t-1} \eta_i^2 \leq 2\eta_1 t^{-\min(\theta,1-\theta)} (\log t + 1) \leq \frac{1}{4\kappa^2}.$$

This verifies (61). Note also that by taking $t = 1$ in (65), for all $t \in [T]$,

$$\eta_t \kappa^2 \leq \eta_1 \kappa^2 \leq \frac{1}{8\kappa^2} \leq \frac{1}{2}.$$

We thus can apply Proposition 5 to derive (64). What remains is to control the right hand side of (64). Since

$$\sum_{k=1}^{t-1} \frac{\eta_k^2}{\sum_{i=k+1}^t \eta_i} = \eta_1 \sum_{k=1}^{t-1} \frac{k^{-2\theta}}{\sum_{i=k+1}^t i^{-\theta}} \leq \eta_1 \sum_{k=1}^{t-1} \frac{k^{-2\theta}}{(t-k)t^{-\theta}},$$

combining with Lemma 14,

$$\sum_{k=1}^{t-1} \frac{\eta_k^2}{\sum_{i=k+1}^t \eta_i} \leq 2\eta_1 t^{-\min(\theta,1-\theta)} (\log t + 1).$$

Also, by Lemma 12,

$$\frac{1}{\eta_k k} \sum_{l=1}^{k} \eta_l = \frac{1}{k^{1-\theta}} \sum_{l=1}^{k} l^{-\theta} \leq \frac{1}{1-\theta},$$

and by Lemma 13,

$$\sum_{k=1}^{t-1} \eta_k^2 = \eta_1^2 \sum_{k=1}^{t-1} k^{-2\theta} \leq \eta_1^2 t^{\max(1-2\theta,0)} (\log t + 1).$$

Introducing the last three estimates into (64) and using that $\eta_t^2 \kappa^2 \leq \eta_1 t^{-\theta}$ by (65), we get the desired result. The proof is complete. ■

Collect some of the above analysis, we get the following result for the computational variance.

Theorem 6 *Under Assumptions 1, let* $\delta_2 \in]0,1[$, $\frac{9\kappa^2}{m} \log \frac{m}{\delta_2} \leq \lambda \leq \|\mathcal{T}_\rho\|$, $\delta_3 \in]0,1[$, $m \geq 32 \log^2 \frac{2}{\delta_3}$, *and* $\eta_t = \eta_1 t^{-\theta}$ *for all* $t \in [T]$, *with* $\theta \in [0,1[$ *and* η_1 *such that (65). Then, with probability at least* $1 - \delta_2 - \delta_3$, *(66) holds for all* $t \in [T]$.

7. Deriving Total Error Bounds

The purpose of this section is to derive total error bounds.

7.1 Attainable Case

We have the following general theorem for $\zeta \geq 1/2$, with which we prove our main results stated in Section 3.

Theorem 7 *Under Assumptions 1, 2 and 3, let $\zeta \geq 1/2$, $T \in \mathbb{N}$ with $T \geq 3$, $\delta \in]0,1[$, $\eta_t = \eta\kappa^{-2}t^{-\theta}$ for all $t \in [T]$, with $\theta \in [0,1[$ and η such that*

$$0 < \eta \leq \frac{t^{\min(\theta,1-\theta)}}{8(\log t + 1)}, \qquad \forall t \in [T]. \tag{67}$$

If for some $\epsilon \in]0,1]$,

$$m \geq \left(\frac{18\kappa^2}{\epsilon\|\mathcal{T}_\rho\|} \log\left(\frac{27\kappa^2}{\epsilon\|\mathcal{T}_\rho\|\delta} \right)\right)^{1/\epsilon}, \tag{68}$$

then the following holds with probability at least $1-\delta$: for all $t \in [T]$,

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \leq q_1(\eta t^{1-\theta})^{-2\zeta} + q_2 m^{\gamma(1-\epsilon)-1}(1 \vee \eta^2 m^{2\epsilon-2}t^{2-2\theta})(\log T)^2 \log^2 \frac{12}{\delta} \\ + q_3\eta b^{-1}(t^{-\min(\theta,1-\theta)} \vee m^{\epsilon-1}\eta t^{(1-2\theta)_+}) \log T.\end{aligned} \tag{69}$$

Here, $q_1 = 2R^2\zeta^{2\zeta}$, $q_2 = \frac{10^4(R\kappa^{2\zeta}+\sqrt{M})^2(\kappa/\sqrt{\|\mathcal{T}_\rho\|}+\sqrt{2\sqrt{v}c_\gamma/\|\mathcal{T}_\rho\|^\gamma})^2}{(1-\theta)^2}$, and $q_3 = \frac{208Mv}{1-\theta}$.

Proof Let $\lambda = \|\mathcal{T}_\rho\| m^{\epsilon-1}$. Clearly, $\lambda \leq \|\mathcal{T}_\rho\|$. For any $A \geq 0$ and $B \geq 1$, by applying (32) with $\zeta = 1, x = (Bm)^\epsilon$ and $c = \frac{\epsilon}{2AB^\epsilon}$,

$$A\log(Bm) = \frac{A}{\epsilon}\log((Bm)^\epsilon) \leq \frac{A}{\epsilon}\log\left(\frac{2AB^\epsilon}{e\epsilon}\right) + \frac{1}{2}m^\epsilon \leq \frac{A}{\epsilon}\log\left(\frac{AB}{\epsilon}\right) + \frac{1}{2}m^\epsilon. \tag{70}$$

Using the above inequality with $A = \frac{9\kappa^2}{\|\mathcal{T}_\rho\|}$ and $B = \frac{1}{\delta_2}$, one can prove that the condition (68) ensures that $\frac{9\kappa^2}{m}\log\frac{m}{\delta_2} \leq \lambda$ is satisfied with $\delta_2 = \frac{\delta}{3}$, Therefore, by Lemma 19, (47) holds with probability at least $1-\delta_2$. Similarly the condition (68) implies that $m \geq 32\log^2\frac{2}{\delta_3}$ is satisfied with $\delta_3 = \frac{\delta}{3}$, and thus by Lemma 25, (63) holds with probability at least $1-\delta_3$. Combining with Lemma 18, by taking the union bound, we know that with probability at least $1-\delta_1-\delta_2-\delta_3$, (47), (63) and (40) hold for all $k \in [T]$. Now, we can apply Propositions 4 and 6 to get (50) and (66). Noting that by (67), $\sqrt{2}\eta \leq 1$, and by a simple calculation, we derive from (50) that

$$\begin{aligned}&\|\mathcal{S}_\rho\nu_{t+1} - \mathcal{S}_\rho\mu_{t+1}\|_\rho^2 \\ \leq& \frac{4624(R\kappa^{2\zeta}+\sqrt{M})^2(\kappa/\sqrt{\|\mathcal{T}_\rho\|}+\sqrt{2\sqrt{v}c_\gamma/\|\mathcal{T}_\rho\|^\gamma})^2}{(1-\theta)^2} m^{\gamma(1-\epsilon)-1}(1 \vee \lambda^2\eta^2\kappa^{-4}t^{2-2\theta} \vee \log^2 t)\log^2\frac{4}{\delta_1} \\ \leq& \frac{4624(R\kappa^{2\zeta}+\sqrt{M})^2(\kappa/\sqrt{\|\mathcal{T}_\rho\|}+\sqrt{2\sqrt{v}c_\gamma/\|\mathcal{T}_\rho\|^\gamma})^2}{(1-\theta)^2} m^{\gamma(1-\epsilon)-1}(1 \vee \eta^2 m^{2\epsilon-2}t^{2-2\theta})(\log T)^2\log^2\frac{4}{\delta_1},\end{aligned}$$

where for the last inequality, we used $\|\mathcal{T}_\rho\| \leq \kappa^2$. Similarly, by a simple calculation, we get from (66) that

$$\begin{aligned}\mathbb{E}_{\mathbf{J}}\|\mathcal{S}_\rho\omega_{t+1} - \mathcal{S}_\rho\nu_{t+1}\|_\rho^2 &\leq \frac{208Mv}{b(1-\theta)}(\eta t^{-\min(\theta,1-\theta)} \vee \lambda\eta^2\kappa^{-2}t^{(1-2\theta)_+})(1 \vee \log t) \\ &\leq \frac{208Mv}{b(1-\theta)}(\eta t^{-\min(\theta,1-\theta)} \vee m^{\epsilon-1}\eta^2 t^{(1-2\theta)_+})\log T.\end{aligned}$$

Letting $\delta_1 = \frac{\delta}{3}$, and introducing the above estimates and (34) into (23), we get (69). The proof is complete. ■

Proof [of Theorem 1] By choosing $\epsilon = 1 - \frac{1}{2\zeta+\gamma}$ and $\theta = 0$ in Theorem 7, then the condition (68) reduces to $m \geq m_\delta$, where

$$m_\delta = \left(\frac{18\kappa^2 p}{\|\mathcal{T}_\rho\|} \log\left(\frac{27\kappa^2 p}{\|\mathcal{T}_\rho\|\delta} \right) \right)^p, \quad p = \frac{2\zeta+\gamma}{2\zeta+\gamma-1}. \tag{71}$$

The desired result thus follows by applying Theorem 7. ■

7.2 Non Attainable Case

For the non-attainable case, we have the following general results on generalization errors for SGM.

Theorem 8 *Under Assumptions 1, 2 and 3, let $\zeta \leq 1/2$, $T \in \mathbb{N}$ with $T \geq 3$, $\delta \in]0,1[$, $\eta_t = \eta\kappa^{-2}t^{-\theta}$ for all $t \in [T]$, with $\theta \in [0,1[$ and η such that (67) and for some $\epsilon \in]0,1]$, (68) holds. Then the following holds with probability at least $1-\delta$: for all $t \in [T]$,*

$$\begin{aligned} \mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim \left((\eta t^{1-\theta})^{-2\zeta} + m^{\gamma(1-\epsilon)-1} \right) (1 \vee \eta m^{\epsilon-1} t^{1-\theta})^3 \log^4 T \log^2 \frac{1}{\delta} \\ + \eta b^{-1} (t^{-\min(\theta,1-\theta)} \vee m^{\epsilon-1} \eta t^{(1-2\theta)_+}) \log T. \end{aligned} \tag{72}$$

Here, the constant in the upper bounds is positive and depends only on $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma, \gamma$ and $\|f_{\mathcal{H}}\|_\infty$.

Proof The proof is similar to that for Theorem 7. We include the sketch only and omit the constants appeared. Similar to the proof of Theorem 7, with $\lambda = \|\mathcal{T}_\rho\| m^{\epsilon-1}$, one can prove that with probability at least $1 - \delta_1 - \delta_2 - \delta_3$, (47), (63) and (41) hold for all $k \in [T]$. Now, we can apply Propositions 4 and 6 to get (51) and (66). Noting that by (65), $\sqrt{2}\eta \leq 1$, and by a simple calculation, we derive from (51) that

$$\begin{aligned} &\|\mathcal{S}_\rho \nu_{t+1} - \mathcal{S}_\rho \mu_{t+1}\|_\rho^2 \\ \lesssim & m^{\gamma(1-\epsilon)-1} (1 \vee \eta^2 m^{2\epsilon-2} t^{2-2\theta}) \log^4 T \log^2 \frac{1}{\delta} + (\eta t^{1-\theta})^{-2\zeta} (1 \vee \eta t^{1-\theta} m^{-1})(1 \vee \eta^2 m^{2\epsilon-2} t^{2-2\theta}) \log^4 T \log^2 \frac{1}{\delta} \\ \lesssim & m^{\gamma(1-\epsilon)-1} (1 \vee \eta^2 m^{2\epsilon-2} t^{2-2\theta}) \log^4 T \log^2 \frac{1}{\delta} + (\eta t^{1-\theta})^{-2\zeta} (1 \vee \eta m^{\epsilon-1} t^{1-\theta})^3 \log^4 T \log^2 \frac{1}{\delta}. \end{aligned}$$

The rest of the proof parallelizes to that for Theorem 7. ■

Now, we are in a position to prove Theorem 3.

Proof [of Theorem 3] The second part of the theorem follows directly from applying Theorem 8 with $\theta = 0$. The first part can be proved by applying Theorem 8 with $\theta = 0$ and $\epsilon = 1 - \frac{1}{2\zeta+\gamma}$, combining with the same argument from the proof of Theorem 1 to verify the condition (68). We omit the details. ■

7.3 Batch GM

Following the proof of Theorems 1 and 3, we know that the following results hold for batch GM, from which one can prove Theorem 4.

Theorem 9 *Under Assumptions 1, 2 and 3, set $\eta_t = \eta\kappa^{-2}$ with $\eta \leq 1$, for all $t \in [m]$. With probability at least $1-\delta$ $(0<\delta<1)$, the following holds for the learning sequence generated by (10):*
1) if $\zeta > 1/2$ and $m \geq m_\delta$ with m_δ given by (71), then

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim (\eta t)^{-2\zeta} + m^{-\frac{2\zeta}{2\zeta+\gamma}}(1+m^{-\frac{1}{2\zeta+\gamma}}\eta t)^2 \log^2 T \log^2 \frac{1}{\delta}; \tag{73}$$

2) if $\zeta \leq 1/2$, $2\zeta+\gamma>1$ and $m \geq m_\delta$ with m_δ given by (71), then

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim \left((\eta t)^{-2\zeta} + m^{-\frac{2\zeta}{2\zeta+\gamma}}\right)(1 \vee m^{-\frac{1}{2\zeta+\gamma}}\eta t)^3 \log^4 T \log^2 \frac{1}{\delta};$$

3) if $2\zeta+\gamma \leq 1$ and for some $\epsilon \in]0,1]$, (68) hold, then

$$\mathbb{E}_{\mathbf{J}}[\mathcal{E}(\omega_{t+1})] - \inf_H \mathcal{E} \lesssim \left((\eta t)^{-2\zeta} + m^{\gamma(1-\epsilon)-1}\right)(1 \vee \eta m^{\epsilon-1} t)^3 \log^4 T \log^2 \frac{1}{\delta}.$$

Here, all the constants in the upper bounds are positive and depend only on $\kappa^2, \|\mathcal{T}_\rho\|, M, v, \zeta, R, c_\gamma$ and γ (and also on $\|f_{\mathcal{H}}\|_\infty$ when $\zeta < 1/2$) .

8. Convergence in H-norm

In this section, we will give convergence results in H-norm for Algorithm 1 in the attainable case. For the sake of simplicity, we will only consider a fixed step-size sequence, i.e, $\eta_t = \eta$ for all t.

Using a similar procedure as that for (23), we can prove the following error decomposition,

$$\mathbb{E}_{\mathbf{J}}[\|\omega_t - \omega^\dagger\|_H^2] \lesssim \|\mu_t - \omega^\dagger\|_H^2 + \|\mu_t - \nu_t\|_H^2 + \mathbb{E}_{\mathbf{J}}[\|\omega_t - \nu_t\|_H^2]. \tag{74}$$

To estimate the bias term, $\|\mu_t - \omega^\dagger\|_H^2$, we introduce the following lemma from (Yao et al., 2007; Rosasco and Villa, 2015). Its proof is similar as that for (34) and will be given in Appendix C for the sake of completeness.

Lemma 27 *Under Assumption 2, let $\zeta \geq 1/2$ and $\eta_t = \eta$ for all $t \in \mathbb{N}$, with $\eta \in]0, \kappa^{-2}]$, then*

$$\|\mu_{t+1} - \omega^\dagger\|_H \leq R\left(\frac{\zeta - 1/2}{\eta t}\right)^{\zeta-1/2}. \tag{75}$$

To estimate the sample variance term, $\|\mu_t - \nu_t\|_H^2$, we use (39) and get that

$$\begin{aligned}\|\nu_{t+1} - \mu_{t+1}\|_H &= \left\|\sum_{k=1}^t \eta_k \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\right\|_H \\ \leq & \sum_{k=1}^t \eta_k \left\|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}}\right\| \left\|\mathcal{T}_{\rho,\lambda}^{-\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\right\|_H \leq \frac{1}{\sqrt{\lambda}} \sum_{k=1}^t \eta_k \left\|\mathcal{T}_{\rho,\lambda}^{\frac{1}{2}} \Pi_{k+1}^t(\mathcal{T}_{\mathbf{x}}) N_k\right\|_H.\end{aligned}$$

From the proof of Theorem 5, we know that $\sum_{k=1}^{t} \eta_k \left\| \mathcal{T}_{\rho,\lambda}^{\frac{1}{2}} \Pi_{k+1}^{t}(\mathcal{T}_{\mathbf{x}}) N_k \right\|_H$ is upper bounded by the right-hand side of (50). With $\eta_t = \eta$ and $\lambda = \|\mathcal{T}_\rho\| m^{-\frac{1}{2\zeta+\gamma}}$, we thus have

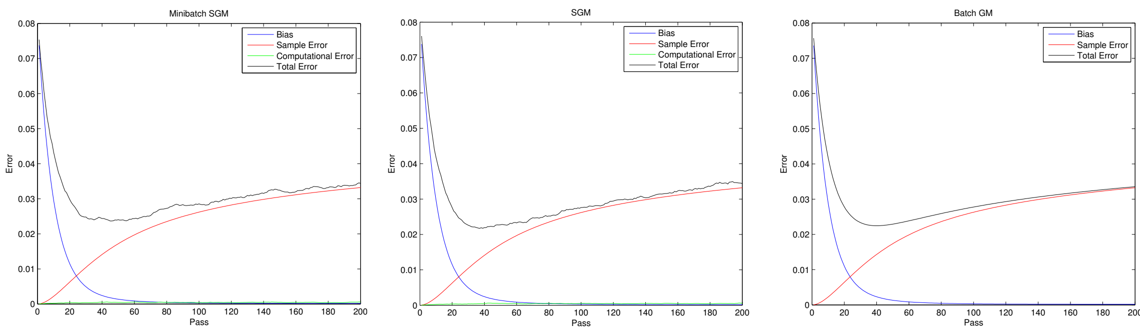
$$\|\nu_{t+1} - \mu_{t+1}\|_H \lesssim m^{-\frac{\zeta-1/2}{2\zeta+\gamma}} (1 + m^{-\frac{1}{2\zeta+\gamma}} \eta t) \log t \log \frac{1}{\delta}. \tag{76}$$

Finally, for the computational variance term, $\mathbb{E}_{\mathbf{J}}[\|\omega_t - \nu_t\|_H^2]$, we use a same procedure as that for (52) to get

$$\mathbb{E}_{\mathbf{J}} \|\omega_{t+1} - \nu_{t+1}\|_H^2 \leq \frac{\kappa^2}{b} \sum_{k=1}^{t} \eta^2 \left\| \Pi_{k+1}^{t}(\mathcal{T}_{\mathbf{x}}) \right\|^2 \mathbb{E}_{\mathbf{J}}[\mathcal{E}_{\mathbf{z}}(\omega_k)] \lesssim \frac{\eta^2 t}{b}, \tag{77}$$

where we used (62) and (63) in the last inequality. Introducing (75), (76) and (77) into the error decomposition (74), we can prove Theorem 2.

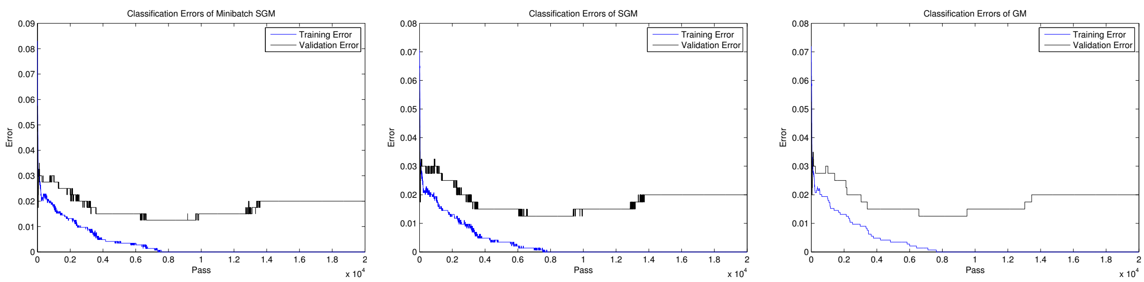
9. Numerical Simulations



(a) Minibatch SGM (b) SGM (c) Batch GM

Figure 1: Error decompositions for gradient-based learning algorithms on *synthesis data*, where *m = 100.*

In order to illustrate our theoretical results and the error decomposition, we first performed some simulations on a simple problem. We constructed $m = 100$ i.i.d. training examples of the form $y = f_\rho(x_i) + \omega_i$. Here, the regression function is $f_\rho(x) = |x - 1/2| - 1/2$, the input point x_i is



(a) Minibatch SGM (b) SGM (c) Batch GM

Figure 2: Misclassification Errors for gradient-based learning algorithms on *BreastCancer* dataset.

uniformly distributed in $[0,1]$, and ω_i is a Gaussian noise with zero mean and standard deviation 1, for each $i \in [m]$. We perform three experiments with the same H, a RKHS associated with a Gaussian kernel $K(x,x') = \exp(-(x-x')^2/(2\sigma^2))$ where $\sigma = 0.2$. In the first experiment, we run mini-batch SGM, where the mini-batch size $b = \sqrt{m}$, and the step-size $\eta_t = 1/(8\sqrt{m})$. In the second experiment, we run simple SGM where the step-size is fixed as $\eta_t = 1/(8m)$, while in the third experiment, we run batch GM using the fixed step-size $\eta_t = 1/8$. For mini-batch SGM and SGM, the total error $\|\mathcal{S}_\rho \omega_t - f_\rho\|^2_{L^2_{\hat\rho}}$, the bias $\|\mathcal{S}_\rho \hat\mu_t - f_\rho\|^2_{L^2_{\hat\rho}}$, the sample variance $\|\mathcal{S}_\rho \nu_t - \mathcal{S}_\rho \hat\mu_t\|^2_{L^2_{\hat\rho}}$ and the computational variance $\|\mathcal{S}_\rho \omega_t - \mathcal{S}_\rho \nu_t\|^2_{L^2_{\hat\rho}}$, averaged over 50 trials, are depicted in Figures 1a and 1b, respectively. For batch GM, the total error $\|\mathcal{S}_\rho \nu_t - f_\rho\|^2_{L^2_{\hat\rho}}$, the bias $\|\mathcal{S}_\rho \hat\mu_t - f_\rho\|^2_{L^2_{\hat\rho}}$ and the sample variance $\|\mathcal{S}_\rho \nu_t - \hat\mu_t\|^2_{L^2_{\hat\rho}}$, averaged over 50 trials are depicted in Figure 1c. Here, we replace the unknown marginal distribution ρ_X by an empirical measure $\hat\rho = \frac{1}{2000}\sum_{i=1}^{2000} \delta_{\hat x_i}$, where each $\hat x_i$ is uniformly distributed in $[0,1]$. From Figure 1a or 1b, we see that as the number of passes increases[5], the bias decreases, while the sample error increases. Furthermore, we see that in comparisons with the bias and the sample error, the computational error is negligible. In all these experiments, the minimal total error is achieved when the bias and the sample error are balanced. These empirical results show the effects of the three terms from the error decomposition, and complement the derived bound (11), as well as the regularization effect of the number of passes over the data. Finally, we tested the simple SGM, mini-batch SGM, and batch GM, using similar step-sizes as those in the first simulation, on the *BreastCancer* data-set[6]. The classification errors on the training set and the testing set of these three algorithms are depicted in Figure 2. We see that all of these algorithms perform similarly, which complement the bounds in Corollaries 3, 4 and 7.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC.

Appendices

A. Learning with Kernel Methods

Let the input space Ξ be a closed subset of Euclidean space $\mathbb{R}^d$, the output space $Y \subseteq \mathbb{R}$. Let μ be an unknown but fixed Borel probability measure on $\Xi \times Y$. Assume that $\{(\xi_i, y_i)\}_{i=1}^m$ are i.i.d. from the distribution μ. A reproducing kernel K is a symmetric function $K : \Xi \times \Xi \to \mathbb{R}$ 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 Ξ. The kernel K defines a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ as the completion of the linear span of the set $\{K_\xi(\cdot) := K(\xi,\cdot) : \xi \in \Xi\}$ with respect to the inner product $\langle K_\xi, K_u\rangle_K := K(\xi,u)$. For any $f \in \mathcal{H}_K$, the reproducing property holds: $f(\xi) = \langle K_\xi, f\rangle_K$.

5. Note that the terminology 'running the algorithm with p passes' means 'running the algorithm with $\lceil mp/b \rceil$ iterations', where b is the mini-batch size.
6. `https://archive.ics.uci.edu/ml/datasets/`

Example 1 (Sobolev Spaces) *Let $X = [0, 1]$ and the kernel*

$$K(x, x') = \begin{cases} (1-y)x, & x \le y; \\ (1-x)y, & x \ge y. \end{cases}$$

Then the kernel induces a Sobolev Space $H = \{f : X \to \mathbb{R} | f$ is absolutely continuous $, f(0) = f(1) = 0, f \in L^2(X)\}$.

In learning with kernel methods, one considers the following minimization problem

$$\inf_{f \in \mathcal{H}_K} \int_{\Xi \times Y} (f(\xi) - y)^2 d\mu(\xi, y).$$

Since $f(\xi) = \langle K_\xi, f \rangle$ by the reproducing property, the above can be rewritten as

$$\inf_{f \in \mathcal{H}_K} \int_{\Xi \times Y} (\langle f, K_\xi \rangle - y)^2 d\mu(\xi, y).$$

Letting $X = \{K_\xi : \xi \in \Xi\}$ and defining another probability measure $\rho(K_\xi, y) = \mu(\xi, y)$, the above reduces to the learning setting in Section 2.

B. Further Corollaries for SGM in the non-attainable case

In this section, we state the convergence results for the SGM with different parameter choices similar as those in Corollaries 5–7, in the non-attainable case. These results are direct consequences of Theorem 3.

Corollary 28 *Under Assumptions 1, 2 and 3, let $\zeta \le 1/2$, $\delta \in]0, 1[$, $b = 1$, and $\eta_t \simeq m^{-\frac{2\zeta}{(2\zeta+\gamma)\vee 1}}$ for all $t \in [m^2]$. With probability at least $1 - \delta$, the following holds:*
1) if $2\zeta + \gamma > 1$, $m \ge m_\delta$ and $T^* = \lceil m^{\frac{2\zeta+1}{2\zeta+\gamma}} \rceil$, then we have (18);
2) if $2\zeta + \gamma \le 1$, and for some $\epsilon \in]0, 1[$, $m \ge m_{\delta,\epsilon}$, and $T^* = \lceil m^{1+2\zeta-\epsilon} \rceil$, then we have (19).

Corollary 29 *Under Assumptions 1, 2 and 3, let $\zeta \le 1/2$, $\delta \in]0, 1[$, $b \simeq m^{\frac{2\zeta}{(2\zeta+\gamma)\vee 1}}$, and $\eta_t \simeq \frac{1}{\log m}$ for all $t \in [m]$. With probability at least $1 - \delta$, the following holds:*
1) if $2\zeta + \gamma > 1$, $m \ge m_\delta$ and $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$, then we have (18);
2) if $2\zeta + \gamma \le 1$, and for some $\epsilon \in]0, 1[$, $m \ge m_{\delta,\epsilon}$, and $T^* = \lceil m^{1-\epsilon} \rceil$, then we have (19).

Corollary 30 *Under Assumptions 1, 2 and 3, let $\zeta \le 1/2$, $\delta \in]0, 1[$, $b = m$ and $\eta_t \simeq \frac{1}{\log m}$ for all $t \in [m]$. With probability at least $1 - \delta$, the following holds:*
1) if $2\zeta + \gamma > 1$, $m \ge m_\delta$ and $T^* = \lceil m^{\frac{1}{2\zeta+\gamma}} \rceil$, then we have (18);
2) if $2\zeta + \gamma \le 1$, and for some $\epsilon \in]0, 1[$, $m \ge m_{\delta,\epsilon}$, and $T^* = \lceil m^{1-\epsilon} \rceil$, then we have (19).

C. Proofs for Lemmas

Proof [of Lemma 12] Note that

$$\sum_{k=1}^t k^{-\theta} \le 1 + \sum_{k=2}^t \int_{k-1}^k u^{-\theta} du = 1 + \int_1^t u^{-\theta} du = \frac{t^{1-\theta} - \theta}{1 - \theta},$$

which leads to the first part of the desired result. Similarly,

$$\sum_{k=1}^{t} k^{-\theta} \geq \sum_{k=1}^{t} \int_{k}^{k+1} u^{-\theta} du = \int_{1}^{t+1} u^{-\theta} du = \frac{(t+1)^{1-\theta} - 1}{1-\theta},$$

and by mean value theorem, $(t+1)^{1-\theta} - 1 \geq (1-\theta)t(t+1)^{-\theta} \geq (1-\theta)t^{1-\theta}/2$. This proves the second part of the desired result. The proof is complete. ■

Proof [of Lemma 13] Note that

$$\sum_{k=1}^{t} k^{-\theta} = \sum_{k=1}^{t} k^{-1} k^{1-\theta} \leq t^{\max(1-\theta,0)} \sum_{k=1}^{t} k^{-1},$$

and

$$\sum_{k=1}^{t} k^{-1} \leq 1 + \sum_{k=2}^{t} \int_{k-1}^{k} u^{-1} du = 1 + \log t.$$

■

Proof [of Lemma 14] Note that

$$\sum_{k=1}^{t-1} \frac{1}{t-k} k^{-q} = \sum_{k=1}^{t-1} \frac{k^{1-q}}{(t-k)k} \leq t^{\max(1-q,0)} \sum_{k=1}^{t-1} \frac{1}{(t-k)k},$$

and that by Lemma 13,

$$\sum_{k=1}^{t-1} \frac{1}{(t-k)k} = \frac{1}{t} \sum_{k=1}^{t-1} \left(\frac{1}{t-k} + \frac{1}{k} \right) = \frac{2}{t} \sum_{k=1}^{t-1} \frac{1}{k} \leq \frac{2}{t}(1 + \log t).$$

■

Proof [of Proposition 2] Since μ_{t+1} is given by (25), introducing with (27),

$$\mu_{t+1} = \mu_t - \eta_t(\mathcal{T}_\rho \mu_t - \mathcal{S}_\rho^* f_{\mathcal{H}}). \tag{78}$$

Thus,

$$\mathcal{S}_\rho \mu_{t+1} = \mathcal{S}_\rho \mu_t - \eta_t \mathcal{S}_\rho(\mathcal{T}_\rho \mu_t - \mathcal{S}_\rho^* f_{\mathcal{H}}) = \mathcal{S}_\rho \mu_t - \eta_t \mathcal{L}_\rho(\mathcal{S}_\rho \mu_t - f_{\mathcal{H}}). \tag{79}$$

Subtracting both sides by $f_{\mathcal{H}}$,

$$\mathcal{S}_\rho \mu_{t+1} - f_{\mathcal{H}} = (I - \eta_t \mathcal{L}_\rho)(\mathcal{S}_\rho \mu_t - f_{\mathcal{H}}).$$

Using this equality iteratively, with $\mu_1 = 0$,

$$\mathcal{S}_\rho \mu_{t+1} - f_{\mathcal{H}} = -\Pi_1^t(\mathcal{L}_\rho) f_{\mathcal{H}}.$$

Taking the $L^2(H, \rho_X)$-norm, by Assumption 2,

$$\|\mathcal{S}_\rho \mu_{t+1} - f_{\mathcal{H}}\|_\rho = \|\Pi_1^t(\mathcal{L}_\rho) f_{\mathcal{H}}\|_\rho \leq \|\Pi_1^t(\mathcal{L}_\rho) \mathcal{L}_\rho^\zeta\| R.$$

By applying Lemma 15, we get (33). Combining (33) with Lemma 12, we get (34). The proof is complete. ■

Proof [of Lemma 16] From (78), we have

$$\mu_{t+1} = (I - \eta_t \mathcal{T}_\rho)\mu_t + \eta_t \mathcal{S}_\rho^* f_{\mathcal{H}}.$$

Applying this relationship iteratively, and using $\mu_1 = 0$, we get

$$\mu_{t+1} = \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\mathcal{T}_\rho)\mathcal{S}_\rho^* f_{\mathcal{H}} = \sum_{k=1}^t \eta_k \mathcal{S}_\rho^* \Pi_{k+1}^t(\mathcal{L}_\rho) f_{\mathcal{H}}.$$

Therefore, using Assumption 2 and spectral theory,

$$\|\mu_{t+1}\|_H \le \left\| \sum_{k=1}^t \eta_k \mathcal{S}_\rho^* \Pi_{k+1}^t(\mathcal{L}_\rho)\mathcal{L}_\rho^\zeta \right\| R \le R \max_{\sigma \in]0,\kappa^2]} \sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma).$$

Case $\zeta \ge 1/2$. For any $\sigma \in]0, \kappa^2]$,

$$\sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) \le \kappa^{2\zeta-1} \sigma \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) \le \kappa^{2\zeta-1},$$

where for the last inequality, we used

$$\sum_{k=1}^t \eta_k \sigma \Pi_{k+1}^t(\sigma) = \sum_{k=1}^t (1 - (1-\eta_k\sigma))\Pi_{k+1}^t(\sigma) = \sum_{k=1}^t \Pi_{k+1}^t(\sigma) - \sum_{k=1}^t \Pi_k^t(\sigma) = 1 - \Pi_1^t(\sigma). \quad (80)$$

Thus,

$$\|\mu_{t+1}\|_H \le R\kappa^{2\zeta-1}.$$

Case $\zeta < 1/2$. If $\sum_{k=1}^t \eta_k \le \kappa^{-2}$, then for any $\sigma \le \kappa^2$,

$$\sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) \le \sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \le \kappa^{2\zeta-1}.$$

If $\sum_{k=1}^t \eta_k > \kappa^{-2}$, then for any $\sigma \le (\sum_{k=1}^t \eta_k)^{-1}$,

$$\sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) \le \sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \le \left(\sum_{k=1}^t \eta_k \right)^{1/2-\zeta},$$

while for $\kappa^2 \ge \sigma \ge (\sum_{k=1}^t \eta_k)^{-1}$, by (80),

$$\sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) = \sigma^{\zeta-1/2} \sum_{k=1}^t \eta_k \sigma \Pi_{k+1}^t(\sigma) \le \sigma^{\zeta-1/2} \le \left(\sum_{k=1}^t \eta_k \right)^{1/2-\zeta}.$$

From the above analysis, we get that

$$\max_{\sigma \in]0,\kappa^2]} \sigma^{1/2+\zeta} \sum_{k=1}^t \eta_k \Pi_{k+1}^t(\sigma) \le \kappa^{2\zeta-1} \vee \left(\sum_{k=1}^t \eta_k \right)^{1/2-\zeta},$$

and thus

$$\|\mu_{t+1}\|_H \le R\left\{\kappa^{2\zeta-1} \vee \left(\sum_{k=1}^t \eta_k\right)^{1/2-\zeta}\right\}.$$

The proof is complete. ■

Proof [of Lemma 18 (1)] *Bounding* $\left\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}\left(\mathcal{S}_\rho^* f_\rho - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}\right)\right\|_H$*:*
For all $i \in [m]$, let $w_i = y_i(\mathcal{T}_\rho+\lambda I)^{-\frac{1}{2}}x_i$. Obviously, from the definitions of f_ρ (see (6)) and $\mathcal{S}_\rho$,

$$\mathbb{E}[w_1] = \mathbb{E}_{x_1}[f_\rho(x_1)(\mathcal{T}_\rho+\lambda I)^{-\frac{1}{2}}x_1] = (\mathcal{T}_\rho+\lambda I)^{-\frac{1}{2}}\mathcal{S}_\rho^* f_\rho.$$

Thus,

$$(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}\left(\mathcal{S}_\rho^* f_\rho - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}\right) = \frac{1}{m}\sum_{i=1}^m (\mathbb{E}[w_i]-w_i).$$

We next estimate the constants B and $\sigma^2(w_1)$ in (28). Note that for any $l \ge 2$,

$$\mathbb{E}[\|w_1-\mathbb{E}[w_1]\|_H^l] \le \mathbb{E}[(\|w_1\|_H + \mathbb{E}[\|w_1\|_H])^l].$$

By using Hölder's inequality twice,

$$\mathbb{E}[\|w_1-\mathbb{E}[w_1]\|_H^l] \le 2^{l-1}\mathbb{E}[\|w_1\|_H^l + (\mathbb{E}[\|w_1\|_H])^l] \le 2^{l-1}\mathbb{E}[\|w_1\|_H^l + \mathbb{E}[\|w_1\|_H^l]].$$

The right-hand side is exactly $2^l\mathbb{E}[\|w_1\|_H^l]$. Therefore, by recalling the definition of w_1 and expanding the integration,

$$\mathbb{E}[\|w_1-\mathbb{E}[w_1]\|_H^l] \le 2^l\int_X \|(\mathcal{T}_\rho+\lambda I)^{-\frac{1}{2}}x\|_H^l \int_Y y^l d\rho(y|x)d\rho_X(x). \tag{81}$$

Introducing (42) and (45) into the above inequality, we have

$$\mathbb{E}[\|w_1-\mathbb{E}[w_1]\|_H^l] \le l!(2\sqrt{M})^l\sqrt{v}\left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2} c_\gamma\lambda^{-\gamma} = \frac{1}{2}l!\left(\frac{2\kappa\sqrt{M}}{\sqrt{\lambda}}\right)^{l-2} 8M\sqrt{v}c_\gamma\lambda^{-\gamma}.$$

Applying Bernstein inequality with $B = \frac{2\kappa\sqrt{M}}{\sqrt{\lambda}}$ and $\sigma = \sqrt{8M\sqrt{v}c_\gamma\lambda^{-\gamma}}$, we get that with probability at least $1-\frac{\delta_1}{2}$, there holds

$$\left\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}\left(\mathcal{S}_\rho^* f_\rho - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}\right)\right\|_H = \left\|\frac{1}{m}\sum_{i=1}^m(\mathbb{E}[w_i]-w_i)\right\|_H \le 4\sqrt{M}\left(\frac{\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{2\sqrt{v}c_\gamma}}{\sqrt{m\lambda^\gamma}}\right)\log\frac{4}{\delta_1}. \tag{82}$$

Bounding $\|(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}(\mathcal{T}_\rho-\mathcal{T}_{\mathbf{x}})\|$*:*
Let $\xi_i = (\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}x_i\otimes x_i$, for all $i\in[m]$. It is easy to see that $\mathbb{E}[\xi_i] = (\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}\mathcal{T}_\rho$, and that $(\mathcal{T}_\rho+\lambda)^{-\frac{1}{2}}(\mathcal{T}_\rho-\mathcal{T}_{\mathbf{x}}) = \frac{1}{m}\sum_{i=1}^m(\mathbb{E}[\xi_i]-\xi_i)$. Denote the Hilbert-Schmidt norm of a bounded operator from H to H by $\|\cdot\|_{HS}$. Note that

$$\|\xi_1\|_{HS}^2 = \|x_1\|_H^2\mathrm{Trace}((\mathcal{T}_\rho+\lambda)^{-1/2}x_1\otimes x_1(\mathcal{T}_\rho+\lambda)^{-1/2}) = \|x_1\|_H^2\mathrm{Trace}((\mathcal{T}_\rho+\lambda)^{-1}x_1\otimes x_1).$$

By Assumption (3),

$$\|\xi_1\|_{HS} \le \sqrt{\kappa^2 \mathrm{Trace}((\mathcal{T}_\rho + \lambda)^{-1} x_1 \otimes x_1)} \le \sqrt{\kappa^2 \mathrm{Trace}(x_1 \otimes x_1)/\lambda} \le \kappa^2/\sqrt{\lambda},$$

and furthermore, by Assumption 3,

$$\mathbb{E}[\|\xi_1\|_{HS}^2] \le \kappa^2 \mathbb{E}\mathrm{Trace}((\mathcal{T}_\rho + \lambda)^{-1} x_1 \otimes x_1) = \kappa^2 \mathrm{Trace}((\mathcal{T}_\rho + \lambda)^{-1} \mathcal{T}_\rho) \le \kappa^2 c_\gamma \lambda^{-\gamma}.$$

According to Lemma 11, we get that with probability at least $1 - \frac{\delta_1}{2}$, there holds

$$\|(\mathcal{T}_\rho + \lambda)^{-\frac{1}{2}} (\mathcal{T}_\rho - \mathcal{T}_{\mathbf{x}})\|_{HS} \le 2\kappa \left(\frac{2\kappa}{m\sqrt{\lambda}} + \frac{\sqrt{c_\gamma}}{\sqrt{m\lambda^\gamma}} \right) \log \frac{4}{\delta_1}. \tag{83}$$

Finally, using the triangle inequality, we have,

$$\|(\mathcal{T}_\rho + \lambda)^{-\frac{1}{2}} N_k\|_H \le \|(\mathcal{T}_\rho + \lambda)^{-\frac{1}{2}} (\mathcal{T}_\rho - \mathcal{T}_{\mathbf{x}})\| \|\mu_k\|_H + \left\| (\mathcal{T}_\rho + \lambda)^{-\frac{1}{2}} \left(\mathcal{S}_\rho^* f_\rho - \mathcal{S}_{\mathbf{x}}^* \mathbf{y} \right) \right\|_H.$$

Applying (35) to the above, introducing with (82) and (83), and then noting that $\kappa \ge 1$ and $v \ge 1$, one can prove the first part of the lemma.
■

Proof [of Lemma 27] Obviously, $f_{\mathcal{H}} = \mathcal{S}_\rho \omega^\dagger$ and thus $\mathcal{T}_\rho \omega^\dagger = \mathcal{S}_\rho^* f_{\mathcal{H}}$. Combining with Assumption 2, $\mathcal{T}_\rho \omega^\dagger = \mathcal{S}_\rho^* \mathcal{L}_\rho^\zeta \mathcal{L}_\rho^{-\zeta} f_{\mathcal{H}} = \mathcal{T}_\rho^\zeta \mathcal{S}_\rho^* \mathcal{L}_\rho^{-\zeta} f_{\mathcal{H}}$, and $\omega^\dagger = \mathcal{T}_\rho^\dagger \mathcal{T}_\rho^\zeta \mathcal{S}_\rho^* \mathcal{L}_\rho^{-\zeta} f_{\mathcal{H}}$. Subtracting $\omega^\dagger$ from both sides of (78), and using $\mathcal{S}_\rho^* f_{\mathcal{H}} = \mathcal{T}_\rho \omega^\dagger$, we know that

$$\mu_{t+1} - \omega^\dagger = (I - \eta_t \mathcal{T}_\rho)(\mu_t - \omega^\dagger).$$

Applying this relationship iteratively, with $\mu_1 = 0$,

$$\mu_{t+1} - \omega^\dagger = -\Pi_1^t(\mathcal{T}_\rho)\omega^\dagger = -\Pi_1^t(\mathcal{T}_\rho) \mathcal{T}_\rho^\dagger \mathcal{T}_\rho^\zeta \mathcal{S}_\rho^* \mathcal{L}_\rho^{-\zeta} f_{\mathcal{H}}.$$

Therefore,

$$\|\mu_{t+1} - \omega^\dagger\|_H \le \|\Pi_1^t(\mathcal{T}_\rho) \mathcal{T}_\rho^\dagger \mathcal{T}_\rho^\zeta \mathcal{S}_\rho^*\| R \le \|\Pi_1^t(\mathcal{T}_\rho) \mathcal{T}_\rho^{\zeta - 1/2}\| R.$$

Applying Lemma 15, one can get the desired result. ■

D. List of Some Notations

Notation	Meaning
H	the hypothesis space
X, Y, Z	the input space, the output space and the sample space ($Z = X \times Y$)
ρ	the fixed probability measure on Z
ρ_X	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
$\mathbf{z}$	the sample $\{z_i = (x_i, y_i)\}_{i=1}^m$ of size $m \in \mathbb{N}$, where each z_i is i.i.d. according to ρ.
m	the sample size of the sample $\mathbf{z}$
$\mathcal{E}$	the expected risk defined by (1)
$\mathcal{E}_{\mathbf{z}}$	the empirical risk w.r.t the sample $\mathbf{z}$ defined by (20)
κ^2	the constant from the bounded assumption (3) on the hypothesis space H
$\{\omega_t\}_t$	the sequence generated by the SGM
θ	the decaying rate on step-sizes
b	the minibatch size of the SGM
T	the maximal number of iterations for the SGM
j_i (j_t etc.)	the random index from the uniform distribution on $[m]$ for the SGM
$\mathbf{J}_t$	the set of random indices at t-th iteration of the SGM
$\mathbf{J}$	the set of all random indices for the SGM after T iterations
$\mathbb{E}_{\mathbf{J}}$	the expectation with respect to the random variables $\mathbf{J}$ (conditional on $\mathbf{z}$)
$\{\eta_t\}_t$	the sequence of step-sizes
M, v	the positive constants from the moment (bounded) assumption on the output
$L^2(H, \rho_X)$	the Hilbert space of square integral functions from H to $\mathbb{R}$ with respect to ρ_X
f_ρ	the regression function defined (6)
H_ρ	$\{f : X \to \mathbb{R} \| \exists \omega \in H \text{ with } f(x) = \langle \omega, x \rangle_H, \rho_X\text{-almost surely}\}$
ζ, R	the parameters related to the 'regularity' of $f_{\mathcal{H}}$ (see Assumption 2)
$\omega^\dagger$	the solution of Problem (1) with the minimal norm in the attainable case
γ, c_γ	the parameters related to the effective dimension (see Assumption 3)
$\{\sigma_i\}_i$	the sequence of eigenvalues of $\mathcal{L}_\rho$
$\{\nu_t\}_t$	the sequence generated by the batch GM (10)
$\{\mu_k\}_k$	the sequence defined by the population iteration (21)
$\mathcal{S}_\rho$	the linear map from $H \to L^2(H, \rho_X)$ defined by $\mathcal{S}_\rho \omega = \langle \omega, \cdot \rangle$
$\mathcal{S}_\rho^*$	the adjoint operator of $\mathcal{S}_\rho$, $\mathcal{S}_\rho^* f = \int_X f(x) x d\rho_X(x)$
$\mathcal{L}_\rho$	the operator from $L^2(H, \rho_X)$ to $L^2(H, \rho_X)$, $\mathcal{L}_\rho(f) = \mathcal{S}_\rho \mathcal{S}_\rho^* f = \int_X \langle x, \cdot \rangle_H f(x) \rho_X(x)$
$\mathcal{T}_\rho$	the covariance operator from H to H, $\mathcal{T}_\rho = \mathcal{S}_\rho^* \mathcal{S}_\rho = \int_X \langle \cdot, x \rangle x d\rho_X(x)$
$\mathcal{S}_{\mathbf{x}}$	the sampling operator from H to $\mathbb{R}^m$, $(\mathcal{S}_{\mathbf{x}} \omega)_i = \langle \omega, x_i \rangle_H, i \in [m]$
$\mathcal{S}_{\mathbf{x}}^*$	the adjoint operator of $\mathcal{S}_{\mathbf{x}}$, $\mathcal{S}_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m} \sum_{i=1}^m y_i x_i$
$\mathcal{T}_{\mathbf{x}}$	the empirical covariance operator, $\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^* \mathcal{S}_{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^m \langle \cdot, x_i \rangle x_i$
$\Pi_{t+1}^T(L)$	$= \Pi_{k=t+1}(I - \eta_k L)$ when $t \in [T-1]$ and $\Pi_{T+1}^T = I$
$\sum_{i=t+1}^t \eta_i$	$= 0$
λ	a 'regularization' parameter, $\lambda > 0$
$\mathcal{T}_{\rho,\lambda}$,	$\mathcal{T}_{\rho,\lambda} = \mathcal{T}_\rho + \lambda$
$\mathcal{T}_{\mathbf{x},\lambda}$,	$\mathcal{T}_{\mathbf{x},\lambda} = \mathcal{T}_{\mathbf{x}} + \lambda$
$\{N_k\}_k$	the sequence defined by (38).
$M_{k,i}$	defined by (53)

References

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 Mücke. Optimal rates for regularization of statistical inverse learning problems. *arXiv preprint arXiv:1604.04054*, 2016.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Stephen Boyd and Almir Mutapcic. Stochastic subgradient methods. Notes for EE364b, Standford University, Winter 2007.

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

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

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. ISSN 0018-9448.

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems*, pages 1647–1655, 2011.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: an Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202, 2012.

Lee H Dicker, Dean P Foster, Daniel Hsu, et al. 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. Non-parametric stochastic approximation with large step sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances In Neural Information Processing Systems 29*, pages 4556–4564. 2016.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization of multiple passes SGM. In *International Conference on Machine Learning*, 2016a.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(77):1–38, 2016b.

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.

Andrew Ng. Machine learning. Coursera, Standford University, 2016.

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

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

Boris T Poljak. *Introduction to Optimization*. Optimization Software, 1987.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1621–1629, 2015.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1648–1656, 2015.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

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

Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for Machine Learning*. MIT Press, 2012.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science Business Media, 2008.

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

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.

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

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.

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