Refining Covariance Matrix Estimation in Stochastic Gradient Descent Through Bias Reduction

Ziyang Wei
Department of Statistics
University of Chicago
Chicago, IL 60637, USA
ziyangw@uchicago.edu

Wanrong Zhu
Department of Statistics
University of California, Irvine
Irvine, CA 92617, USA
wanronz1@uci.edu

Jingyang Lyu
Department of Statistics
University of Wisconsin–Madison
Madison, WI 53706, USA
jlyu55@wisc.edu

Wei Biao Wu
Department of Statistics
University of Chicago
Chicago, IL 60637, USA
wbwu@uchicago.edu

Abstract

We study online inference and asymptotic covariance estimation for the stochastic gradient descent (SGD) algorithm. While classical methods—such as plug-in and batch-means estimators—are available, they either require inaccessible second-order (Hessian) information or suffer from slow convergence. To address these challenges, we propose a novel, fully online de-biased covariance estimator that eliminates the need for second-order derivatives while significantly improving estimation accuracy. Our method employs a bias-reduction technique to achieve a convergence rate of $n^{(\alpha-1)/2}\sqrt{\log n}$, outperforming existing Hessian-free alternatives.

1 INTRODUCTION

Stochastic gradient descent (SGD), also known as the Robbins–Monro algorithm (Robbins and Monro, 1951), has become a cornerstone of large-scale machine learning due to its simplicity and notable practical performance (Bottou et al., 2018). Consider the classic model-parameter estimation setting, where the true model parameter $x^* \in \mathbb{R}^d$ is characterized as the minimizer of a convex objective function $F(x)$ from $\mathbb{R}^d$ to $\mathbb{R}$, i.e.,

$$x^* = \arg\min_{x\in\mathbb{R}^d} F(x). \tag{1}$$

The objective function $F(x)$ is defined as the expectation of a random loss function $f(x,\xi)$, that is, $F(x) = \mathbb{E}_{\xi\sim\Pi} f(x,\xi)$, where ξ is a random variable representing data drawn from the distribution Π. With initial point x_0, the i-th iteration of the SGD algorithm takes the following form

$$x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \xi_i),\ i \geq 1, \tag{2}$$

where $\{\xi_i\}_{i\geq 1}$ is a sequence of $i.i.d$ samples from the distribution Π, ∇f is the gradient of $f(x,\xi)$ with respect to the first argument x, and η_i is the step size at the i-th step.

As the SGD algorithm progresses and converges, its iterates often resemble noisy approximations of the true optimum. Therefore, it is important not only to evaluate convergence but also to assess the reliability of these estimates by understanding the asymptotic distribution and variability of the iterates. In the foundational work of Polyak and Juditsky (1992), it is shown that, when the step size decays polynomially, i.e., $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$ and $\alpha \in (0.5, 1)$, the average of all past SGD iterates, $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$, exhibits asymptotic normality under suitable conditions:

$$\sqrt{n}(\bar{x}_n - x^*) \Rightarrow N(0, \Sigma), \tag{3}$$

where

$$\Sigma = \nabla^2 F(x^*)^{-1}\mathbb{E}\left([\nabla f(x^*,\xi)][\nabla f(x^*,\xi)]^\intercal\right)\nabla^2 F(x^*)^{-1}.$$

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

If the model is well-specified, this sandwich form limiting covariance matrix achieves the Cramér-Rao lower bound, with Σ^{-1} corresponding to the Fisher information matrix, as discussed in Chen et al. (2020). Similar results have been established for alternative variants of SGD—such as those using different weighting strategies or second-order methods—with appropriately adapted limiting covariance matrices (Li et al., 2022; Na and Mahoney, 2022; Wei et al., 2026).

In the references above, the analytical form of the limiting covariance is studied. However, in practice, the limiting covariance is unknown, as it depends on the underlying data and noise distribution, which are typically not known. Estimating this covariance is therefore essential for quantifying uncertainty in SGD-based estimates and enabling principled statistical inference, such as constructing confidence intervals. Moreover, to stay aligned with the spirit of SGD—namely, computational and memory efficiency, and online updates—it is especially important that the covariance estimation procedure also adheres to these principles. This makes the task of estimating the covariance matrix particularly challenging.

In this paper, we focus on the most fundamental setting: averaged SGD with a polynomially decaying learning rate, $\eta_i = \eta i^{-\alpha}$, where $\eta > 0$ and $\alpha \in (0.5, 1)$. Our goal is to estimate the limiting covariance matrix of $\sqrt{n}\bar{x}_n$, denoted by Σ in (3), in a fully online fashion—using only the SGD iterates and without requiring additional computations such as Hessian evaluations. Existing methods for online, Hessian-free covariance estimation—which we will discuss later in Section 2—achieve a best-known convergence rate of $n^{(\alpha-1)/4}$, which is relatively slow compared to the convergence rates of SGD: $\mathcal{O}(1/n)$ for strongly-convex and $\mathcal{O}(1/\sqrt{n})$ for convex problems (Nemirovski et al., 2009; Lacoste-Julien et al., 2012). A natural question arises: can we improve this result while relying solely on the information provided by a single-pass SGD sequence?

Our contributions: We give an affirmative response in this paper by presenting the de-biased estimator, a novel covariance estimator refined by the bias-reduction technique. It significantly enhances both the theoretical and practical convergence of the estimation error. Specifically, we show that the error rate of the de-biased estimator is $n^{(\alpha-1)/2}\sqrt{\log n}$, which represents the best known convergence rate to date. Moreover, we propose a single-pass algorithm that updates the estimator using only the SGD iterates, with an update cost of $\mathcal{O}(d^2)$—the minimal computational requirement for estimating a $d \times d$ matrix.

There are alternative inference methods that rely on asymptotic pivotal statistics (Lee et al., 2022; Su and Zhu, 2023; Luo et al., 2022; Zhu et al., 2024). Although these methods can sometimes perform better in terms of constructing confidence intervals, they do not yield consistent covariance estimators. Therefore, we do not discuss them in detail here. Beyond statistical inference for stochastic approximation, the estimation of covariance and spectral properties in time series is an independent and well-established topic in the literature (Liu and Wu, 2010; Flegal and Jones, 2010; Xiao and Wu, 2011, 2012; Chen et al., 2013; Zhang and Wu, 2017). The novel bias-reduction estimator proposed in our work offers fresh insights into this area, as it can estimate the dependence structure of time-inhomogeneous processes and is amenable to on-the-fly computation.

Throughout the paper, we use the following notation. For a vector $a = (a_1, \ldots, a_d)^\top$, let the norm $\|a\|_p = (\sum_{i=1}^d a_i^p)^{1/p}$. For a matrix $A \in \mathbb{R}^{d\times d}$, we use $\|A\|_2$ or $\|A\|$ to denote its operator norm and $\|A\|_F$ to denote its Frobenius norm. For $t \in \mathbb{R}$, $\lfloor t \rfloor = \max\{i \in \mathbb{Z} : i \le t\}$ and $\lceil t \rceil = \min\{i \in \mathbb{Z} : i \ge t\}$. For positive sequences $\{a_n\}$ and $\{b_n\}$, $n \in \mathbb{N}$, we write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \le C b_n$ for all $n \in \mathbb{N}$. We write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For a finite set B, we use $|B|$ to denote its cardinality.

The remainder of the paper is organized as follows. Section 2 reviews existing work on covariance matrix estimation in different settings. In Section 3, we present the formulation and algorithm for our recursive de-biased estimator. In Section 4, we establish theoretical guarantees for the consistency of our estimator. Section 5 demonstrates the superior finite-sample performance of the de-biased method through numerical experiments. Finally, Section 6 concludes the paper and outlines directions for future research.

2 BACKGROUND AND RELATED WORK

We begin by reviewing existing methods for online covariance matrix estimation related to stochastic gradient descent (SGD). Recall that the limiting covariance matrix Σ of the averaged SGD in (3) takes the so-called sandwich form: $\Sigma = A^{-1}SA^{-1}$, where

$$A = \nabla^2 F(x^*),\ S = \mathbb{E}\left([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^\intercal\right). \quad (4)$$

Three primary approaches for estimating the covariance structure have been developed in the stochastic approximation literature.

The first method approximates the distribution via bootstrap resampling, as explored in (Fang et al., 2018; Li et al., 2018; Zhong et al., 2023). To obtain a consis-

tent covariance estimate or achieve stable confidence intervals, a large number of bootstrap sequences are required, each of which modifies the original SGD sequence by adding perturbations and recomputing the gradients at every iteration. Consequently, the total cost becomes a substantial multiple of the original SGD run, which makes this approach computationally expensive and less suitable for SGD-related tasks. In this work, we aim to develop a method that operates entirely on a single SGD trajectory.

The second approach is the plug-in estimator introduced by Chen et al. (2020), which separately estimates A and S in the sandwich form of the asymptotic covariance. The key idea is to approximate these matrices using empirical averages. Specifically, the estimators are given by $\widehat{A}_n = \frac{1}{n}\sum_{i=1}^n \nabla^2 f(x_{i-1}, \xi_i)$, $\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n [\nabla f(x_{i-1}, \xi_i)][\nabla f(x_{i-1}, \xi_i)]^\mathsf{T}$. The resulting covariance estimator $\widehat{A}_n^{-1}\widehat{S}_n\widehat{A}_n^{-1}$ is consistent, with a convergence rate of $n^{-\alpha/2}$, and can be computed in an online manner. Similar ideas have been applied in various settings, including constrained stochastic optimization, online decision-making, and non-differentiable problems Na and Mahoney (2025); Chen et al. (2021b,a). However, the practical implementation of this estimator presents challenges: it requires access to the stochastic Hessian, which is often unavailable, and involves matrix computations with a computational complexity of $\mathcal{O}(d^3)$, making it inefficient for high-dimensional scenarios.

The third approach is the batch-means estimator, a Hessian-free method that relies solely on SGD iterates. Unlike the plug-in estimator, which is based on the sandwich formula, the batch-means method directly estimates the variance by analyzing the variability in the SGD sequence itself. The origins of batch-means methods with fixed batch size can be traced back to long-run variance estimation for time-homogeneous Markov chains (Glynn and Iglehart, 1990; Glynn and Whitt, 1991; Damerdji, 1991; Geyer, 1992). Chen et al. (2020) introduced a batch-means method with increasing batch sizes to account for the complex correlation structure inherent in SGD. However, the batch construction in their approach cannot be updated recursively, as it depends on the total number of iterations. This means one would need to store all past iterates and, when new data arrive, reconstruct the batches using all previous iterations and then compute the estimator. To address this limitation, Zhu et al. (2023) developed a fully online batch-means covariance estimator that can be updated on the fly. While the batch-means approach is computationally efficient and practically feasible, it suffers from a relatively slow convergence rate of $n^{(\alpha-1)/4}$, for both online and offline methods.

The same batching idea has been extended to broader contexts. For instance, Jiang et al. (2025) demonstrated that the online batch-means covariance estimator remains valid in nonsmooth and potentially nonmonotone (including certain nonconvex) settings, preserving the same convergence rate of $n^{(\alpha-1)/4}$. Meanwhile, Kuang et al. (2025) proposed a weighted sample covariance estimator for sketched Newton iterates and provided theoretical guarantees within the framework of second-order optimization.

In this paper, we modify the Hessian-free batch-means estimator and provide a de-biased version that retains its computational advantages while offering improved convergence.

3 METHOD

3.1 Motivation for Bias-Reduction

We now take a closer look at the batch-means estimator and provide some intuition for our proposed de-biasing idea. For simplicity, we consider a general one-dimensional sequence $\{X_i\}_{i\geq 1}$.

As discussed in the previous section, rather than relying on the sandwich form of Σ, the batch-means method aims to estimate the variability of the average $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ directly from the sequence $\{X_i\}_{i\geq 1}$. Specifically, it targets the quantity

$$\sigma_n = n\mathrm{Var}\bar{X}_n,$$

which captures the dispersion of the running average and converges to the asymptotic covariance matrix Σ in the SGD setting. In the independent and identically distributed (*i.i.d.*) setting, a natural choice for estimating the variance of the sample mean is the sample covariance:

$$\hat{\sigma}_n = \frac{\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}{n}.$$

However, in the case of SGD, the iterates X_i are highly correlated, which leads to substantial underestimation when using this simple form. To account for this correlation, the online batch-means estimator Zhu et al. (2023) modifies the sample covariance by incorporating local batching:

$$\hat{\sigma}_{n,bm} = \frac{\sum_{i=1}^n \left(\sum_{j=i-l_i+1}^i X_j - l_i\bar{X}_n\right)^2}{\sum_{i=1}^n l_i},$$

where $1 \leq l_i \leq i$ denotes the size of the i-th batch in their paper. The choice of batch sizes is guided by the correlation structure of the SGD iterates. While this estimator is computationally efficient and Hessian-free,

its convergence can be relatively slow due to bias introduced in its formulation. That is, the batch-means estimator $\hat{\sigma}_{n,bm}$ still builds on the structure of the biased sample variance, albeit with adjusted weights. This motivates our pursuit of a more principled, de-biased approach.

Instead of adjusting the biased sample covariance, we propose constructing an estimator that more directly reflects the theoretical definition of σ_n. In particular, observe the identity:

$$\sigma_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Bigg[2(X_i - \mathbb{E}(X_i))\sum_{k=1}^{i}(X_k - \mathbb{E}(X_k)) - (X_i - \mathbb{E}(X_i))^2\Bigg], \tag{5}$$

which offers an alternative way of characterizing the variance of the average. This motivates the following empirical estimator $\hat{\sigma}_{n,db}$:

$$\frac{1}{n}\sum_{i=1}^{n}\Bigg[2(X_i - \bar{X}_n)\Big(\sum_{k=i-\ell_i+1}^{i} X_k - \ell_i\bar{X}_n\Big) - (X_i - \bar{X}_n)^2\Bigg], \tag{6}$$

where ℓ_i again denotes the size of the i-th batch in this paper. Building on this intuition, the following subsection formally defines the de-biased estimator for the limiting covariance matrix Σ of averaged SGD. Our proposed estimator is constructed to mirror the expansion (5) as closely as possible, combined with a principled choice of the batch size ℓ_i to guarantee the theoretical convergence rate.

3.2 De-biased Estimator

Recall that the sequence of SGD iterates $\{x_i\}_{i\geq 1}$ is generated by the recursion in (2) at the i-th iteration, we define the batch

$$B_i = \{i - \ell_i + 1, \ldots, i\}, \quad |B_i| = \ell_i,$$

and propose the de-biased estimator: $\widehat{\Sigma}_n$ as:

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\Bigg[(x_i - \bar{x}_n)\Big(\sum_{k=i-\ell_i+1}^{i} x_k - \ell_i\bar{x}_n\Big)^{\mathsf{T}} + \Big(\sum_{k=i-\ell_i+1}^{i} x_k - \ell_i\bar{x}_n\Big)(x_i - \bar{x}_n)^{\mathsf{T}} - (x_i - \bar{x}_n)(x_i - \bar{x}_n)^{\mathsf{T}}\Bigg], \tag{7}$$

where $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$. Our theoretical analysis in Section 4 establishes that this estimator is consistent for Σ when the batch sizes satisfy $\ell_i = \mathcal{O}(i^\alpha \log i)$ with $\alpha \in (0.5, 1)$. In practice, general choices of ℓ_i often prevent online updates due to overlapping batches and variable batch sizes. To address this, we introduce a block-based batching scheme that ensures each batch has size $|B_i| = \mathcal{O}(i^\alpha \log i)$ while remaining fully compatible with online implementation.

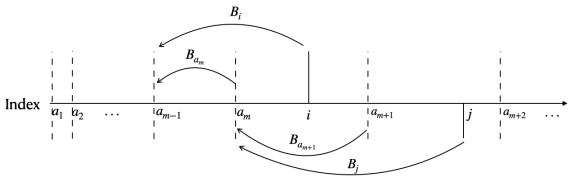

Figure 1: Illustration for Block-based Batching Scheme

Remark 1. *For stationary time series, a conceptually related bias-reduction method was proposed by Xiao and Wu (2011) to estimate the long-run covariance. Inspired by their work, we develop the de-biased estimator in the substantially more intricate non-stationary and time-inhomogeneous regime. The motivation in Xiao and Wu (2011) relied on the relationship between auto-covariance functions and the long-run covariance, which is intrinsically limited to stationary processes. In contrast, the intuition of our estimator comes from the non-asymptotic covariance of the sample average that applies to much broader settings.*

Our bias-reduction approach and selection of batch size also differ from those in Xiao and Wu (2011), where the authors used an increasing threshold sequence to determine which batches are retained in the estimator (see their Equation (12)). Our method is free of this additional threshold parameter. Instead, we analytically design the batch size ℓ_i to directly balance the bias-variance tradeoff.

3.3 Practical Implementation

Block-based Batching Scheme. We define a sequence of indices $\{a_m\}_{m\geq 1}$ with $a_1 = 1$, and subsequent values satisfy the condition:

$$a_m - a_{m-1} + 1 = \lfloor a_m^\alpha \log(a_m)\rfloor, m \geq 2. \tag{8}$$

That is to say, we start a new block when the current block length reaches $\lfloor i^\alpha \log(i)\rfloor$. At each iteration i, find the unique index m such that $a_m \leq i < a_{m+1}$, and define the batch

$$B_i = \{a_{m-1}, a_{m-1} + 1, ..., a_m, ..., i\},$$

which corresponds to $\ell_i = |B_i| = i - a_{m-1} + 1$. This batching strategy always incorporates the two most

recent blocks and satisfies the desired scaling for ℓ_i; see figure 1 for an illustration.

Proposition 1. *Let the batch B_i be constructed using the Block-based Batching Scheme with indices sequence $\{a_m\}_{m\geq 1}$ as described above in (8). Then the batch size satisfies*

$$\lfloor i^\alpha \log(i) \rfloor \leq |B_i| \leq 2\lfloor i^\alpha \log(i) \rfloor$$

and hence $|B_i| = \mathcal{O}(i^\alpha \log i)$.

Online Update. In practice, we can fix $a_1 = 1$, and then recursively determine each subsequent a_m by incrementing its value until the condition in (8) is satisfied. This enables the batch structure to be constructed in a fully online fashion, and thus supports the recursive update of the de-biased estimator. Specifically, we update the batch sum and batch length as follows. Recall that at the i-th step, the batch B_i contains of two parts: the previous block $\{a_{m-1}, \ldots, a_m - 1\}$, and the current block $\{a_m, \ldots, i\}$. Let S_0 denote the sum of previous block, and S_1 the sum of current block. Record the starting index of the current block a_m as a,

- If $i - a + 1 < \lfloor i^\alpha \log(i) \rfloor$, we simply add the new iterate x_i to the current batch and all we need to update is $S_1 = S_1 + x_i$
- If $i - a + 1 \geq \lfloor i^\alpha \log(i) \rfloor$, we reset the batch by discarding the previous block. In this case, we update: $S_0 = S_1$ and $S_1 = x_i$, $a = i$.

Then the batch sum $\sum_{k=i-\ell_i+1}^{i} x_k = S_0 + S_1$, and the batch size $\ell_i = i - a + 1$, can both be updated recursively. If we expand the de-biased estimator in (7), we obtain:

$$\begin{aligned} n\widehat{\Sigma}_n = &\sum_{i=1}^{n} x_i \left(\sum_{k=i-\ell_i+1}^{i} x_k \right)^\mathsf{T} + \sum_{i=1}^{n} \left(\sum_{k=i-\ell_i+1}^{i} x_k \right) x_i^\mathsf{T} \\ &- \sum_{i=1}^{n} \left(\ell_i x_i + \sum_{k=i-\ell_i+1}^{i} x_k \right) \bar{x}_n^\mathsf{T} \\ &- \bar{x}_n \sum_{i=1}^{n} \left(\ell_i x_i + \sum_{k=i-\ell_i+1}^{i} x_k \right)^\mathsf{T} \\ &+ \sum_{i=1}^{n} (2\ell_i + 1) \bar{x}_n \bar{x}_n^\mathsf{T} - \sum_{i=1}^{n} x_i x_i^\mathsf{T}. \end{aligned}$$

We observe that the estimator can be computed recursively, provided we maintain recursive updates of the averaged SGD, batch sum, and batch size. We summarize the details in Algorithm 1, which demonstrates that the covariance matrix estimator can indeed be computed recursively. This implies that, as the SGD algorithm progresses from the i-th to $(i+1)$-th iterate, the covariance matrix can be updated simultaneously with minimal computation. The update requires only the new SGD iterate and a few quantities—such as the batch sum and batch size—carried over from the previous step. The computational cost per iteration is $\mathcal{O}(d^2)$, which is minimal, given that we are estimating a $d \times d$ matrix.

Algorithm 1 Recursive update for the de-biased covariance estimator

Input: Step sizes $\{\eta_i\}_{i\geq 1}$, initialization $x_0, a = 1, S_0 = S_1 = \bar{x} = P = W = Q = q = 0$

for $i = 1, 2, \ldots$ **do**
  Sample $\xi_i \sim \Pi$
  $x_i = x_{i-1} - \eta_i \nabla f(x_{i-1}, \xi_i)$
  $\bar{x} = ((i-1)\bar{x} + x_i)/i$
  if $i - a + 1 >= \lfloor i^\alpha \log(i) \rfloor$ **then**
    $a = i$
    $S_0 = S_1, S_1 = x_i$
  else
    $S_1 = S_1 + x_i$
  end if
  $l = i - a + 1$
  $S = S_0 + S_1$
  $P = P + x_i S^\mathsf{T}$; $W = W + l x_i + S$; $Q = Q + x_i x_i^\mathsf{T}$; $q = q + 2l + 1$
  $V = P + P^\mathsf{T} - W\bar{x}^\mathsf{T} - \bar{x}W^\mathsf{T} + q\bar{x}\bar{x}^\mathsf{T} - Q$
  Output (if necessary): $\widehat{\Sigma}_i = V/i$
end for

4 THEORETICAL GUARANTEE

The batch size ℓ_i plays a critical role in constructing an accurate estimator. As mentioned in Section 3, an effective choice is to let ℓ_i grow at the rate $\mathcal{O}(i^\alpha \log i)$, and we propose a practical, fully online batching scheme that satisfies this condition. In this section, we first build intuition for this choice by analyzing a simple mean estimation model as a motivating example in Section 4.1. Then, in Section 4.2, we establish that this batch size leads to a consistent estimator in the general setting. Specifically, we provide an upper bound on the mean squared error (MSE) $\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|^2$, and show that the bound is tight.

4.1 Intuition Behind the Batch Size Choice

Let $\{\xi_i\}_{i\in\mathbb{N}}$ be a sequence of *i.i.d.* random variables from the model $\xi = x^* + e$, where $x^* \in \mathbb{R}$ is the true population mean of interest, and e is a Gaussian random error with $\mathbb{E}[e] = 0$ and $\mathbb{E}[e^2] = \sigma < \infty$. Consider the squared loss function $f(x, \xi) = (\xi - x)^2/2$, and

generate the i-th SGD iterate by

$$x_i = x_{i-1} - \eta_i(x_{i-1} - \xi_i), \tag{9}$$

where $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$ and $\alpha \in (0.5, 1)$. For simplicity of illustration, we assume $x^* = x_0 = 0$ in this subsection, and consider an oracle estimator

$$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}[2x_i(\sum_{k=i-\ell_i+1}^{i} x_k) - x_i^2]. \tag{10}$$

Unlike the construction in (7), we did not subtract the sample mean $\bar{x}_n$ in the estimator $\hat{\sigma}_n$ because $\mathbb{E}x_i = 0$ for all i. Due to the Gaussianity and linearity of the model, we can derive closed-form expressions for the oracle de-biased estimator $\hat{\sigma}_n$. This tractability enables an explicit analysis of the batch size ℓ_i and the precise order of the resulting estimation error.

We begin by analyzing how the error between $\hat{\sigma}_n$ and the finite-sample variance $\sigma_n = \mathrm{Var}(\sqrt{n}(\bar{x}_n)$ depends on the choice of ℓ_i, as summarized in the following proposition. In practice, since we work with finite samples, a bound on the error between $\hat{\sigma}_n$ and σ_n not only implies convergence to the limiting variance Σ, but also provides a more practical assessment of the estimator's performance in finite-sample settings.

Proposition 2. *Consider the SGD iterates $\{x_i\}_{i=1}^n$ defined by (9), with step sizes $\eta_i = \eta i^{-\alpha}$ for some $\eta > 0$ and $\alpha \in (0.5, 1)$. The proposed oracle de-biased estimator $\hat{\sigma}_n$ defined in (10) converges to the true finite-sample variance $\sigma_n = \mathrm{Var}(\sqrt{n}\bar{x}_n)$ with the following bound:*

$$|\mathbb{E}(\hat{\sigma}_n - \sigma_n)| \lesssim \frac{1}{n}\sum_{i=1}^{n}\exp\{-\eta i^{-\alpha}\ell_i\}.$$

Moreover, if the batch size satisfies $\ell_i \lesssim i^\alpha$, this convergence rate is tight, i.e.,

$$|\mathbb{E}(\hat{\sigma}_n - \sigma_n)| \asymp \frac{1}{n}\sum_{i=1}^{n}\exp\{-\eta i^{-\alpha}\ell_i\}.$$

Remark 2 (Choice of batch size ℓ_i)**.** *Proposition 2 provides two important insights regarding the batch size.*

- *If the batch size grows too slowly, specifically if $\ell_i \lesssim i^\alpha$, the estimator incurs a non-negligible bias:*

$$|\mathbb{E}(\hat{\sigma}_n - \sigma_n)| \gtrsim \frac{1}{n}\sum_{i=1}^{n}\exp(-\eta) = \exp(-\eta),$$

  which does not vanish as $n \to \infty$.

- *As long as the batch size ℓ_i grow slightly faster than i^α, for example,*

$$\ell_i \geq C_\eta i^\alpha \log i \tag{11}$$

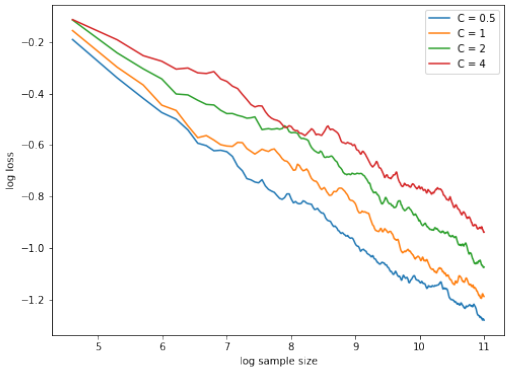

Figure 2: Log-log Plots of Estimation Errors for Different Values of C in the Batch Sequence $\{a_k\}$, under the Linear Model with $d = 1$.

for some universal constant C_η, the estimator becomes asymptotically unbiased since

$$|\mathbb{E}(\hat{\sigma}_n - \sigma_n)| \lesssim \frac{1}{n}\sum_{i=1}^{n}\exp(-\eta C_\eta \log i) \to 0.$$

This justifies the recommended batch size scaling of order $\mathcal{O}(i^\alpha \log i)$.

To align with the broader theoretical literature, we establish the consistency by bounding the error between $\hat{\sigma}_n$ and the limiting variance Σ in (3). By elementary calculations, one can show that the limiting variance equals σ, the variance of the noise, in the context of (9).

Theorem 3. *Consider the SGD iterates $\{x_i\}_{i=1}^n$ defined by (9), with step sizes $\eta_i = \eta i^{-\alpha}$ for some $\eta > 0$ and $\alpha \in (0.5, 1)$. Let $\hat{\sigma}_n$ be the proposed oracle de-biased estimator in (10) with $\ell_i = C_\eta i^\alpha \log i$ for some constant $C_\eta > \eta^{-1}$. Then the convergence rate of MSE satisfies*

$$\mathbb{E}(\hat{\sigma}_n - \sigma)^2 \asymp n^{-1+\alpha}\log n.$$

In Theorem 3, the order of MSE is $n^{-1+\alpha}\log n$ as long as $\eta C_\eta > 1$. Since η and C_η are free to choose in practice, this condition is easily achievable. We will assume this holds in the rest of our paper. For values of α close to $1/2$, the MSE rate approaches $n^{-1/2}\log n$, which is significantly sharper than the rates of existing online covariance estimators. For instance, the non-overlapping online estimator in Zhu et al. (2023) and the batch-means estimator in Chen et al. (2021a) achieve rates of $n^{-1/3}$ and $n^{-1/4}$, respectively.

Table 1: Comparison of De-biased and Online BM Estimators for Different Models with $d = 5$. Standard Deviations are Reported in Parentheses.

		$n = 15000$	$n = 30000$	$n = 60000$
Linear	De-biased	1.55 (0.36)	1.39 (0.35)	1.21 (0.26)
	Online BM	1.79 (0.45)	1.76 (0.48)	1.65 (0.42)
Logistic	De-biased	10.62 (0.92)	9.78 (0.89)	8.99 (0.92)
	Online BM	11.08 (1.07)	10.72 (1.30)	10.21 (1.64)
Expectile	De-biased	1.87 (0.34)	1.68 (0.28)	1.51 (0.24)
	Online BM	2.19 (0.40)	2.20 (0.49)	2.15 (0.49)

Table 2: Comparison of De-biased and Online BM Estimators for Different Models with $d = 20$. Standard Deviations are Reported in Parentheses.

		$n = 50000$	$n = 100000$	$n = 200000$
Linear	De-biased	4.96 (0.52)	4.34 (0.41)	3.82 (0.33)
	Online BM	6.51 (0.95)	5.95 (0.82)	5.66 (0.74)
Logistic	De-biased	38.19 (1.33)	36.01 (1.33)	33.92 (1.28)
	Online BM	43.50 (3.16)	42.57 (3.34)	41.98 (3.51)
Expectile	De-biased	5.10 (0.49)	4.53 (0.39)	4.02 (0.29)
	Online BM	6.89 (0.93)	6.45 (0.80)	6.18 (0.77)

4.2 General Convergence Analysis of the De-biased Estimator

The convergence in the general setting is significantly more difficult to analyze due to the nonlinearity and complex dependency structure of the iterates. Nevertheless, in this section, we will show that the upper bound on the MSE: $\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|^2$ matches the exact rate established in Theorem 3.

Before the main theorem, we introduce some basic assumptions concerning convexity, Lipschitz continuity, boundness, and other regularity conditions.

Assumption 4. *The objective function $F(x)$ is continuously differentiable and strongly convex with parameter $\mu > 0$. That is, for any x_1 and x_2,*

$$F(x_2) \geq F(x_1) + \langle \nabla F(x_1), x_2 - x_1 \rangle + \frac{\mu}{2}\|x_1 - x_2\|_2^2.$$

Further assume that $\nabla^2 F(x^*)$ exists.

Assumption 5. *The function $f(x, \xi)$ is continuous differentiable with respect to x for any ξ and $\|\nabla f(x, \xi)\|_2$ is uniformly integrable for any x.*

Assumption 6. *The gradient noise satisfies $\mathbb{E}_\xi \|\nabla f(x^*, \xi)\|_2^q < \infty$ for some $q \geq 8$, and $\nabla f(x, \xi)$ is stochastic Lipschitz continuous with parameter L, i.e., for any x_1 and x_2,*

$$(\mathbb{E}_\xi \|\nabla f(x_1, \xi) - \nabla f(x_2, \xi)\|_2^q)^{\frac{1}{q}} \leq L\|x_1 - x_2\|_2.$$

Assumptions 4-6 are relatively mild and commonly appear in the literature on convex optimization and statistical inference using SGD, as well as in prior work on estimating the limiting covariance matrix Chen et al. (2021a); Lee et al. (2022); Zhu et al. (2023, 2024).

Theorem 7. *Under Assumption 4-6, consider the SGD iterates $\{x_i\}_{i=1}^n$ with step sizes $\eta_i = \eta i^{-\alpha}$ for some $\eta > 0$ and $\alpha \in (0.5, 1)$. Let $\widehat{\Sigma}_n$ be the de-biased covariance estimator defined in (7), using batch sizes $\ell_i = C_\eta i^\alpha \log i$ for a constant $C_\eta > \eta^{-1}$. Then $\widehat{\Sigma}_n$ converges to the limiting covariance matrix Σ at the following rate:*

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|^2 \lesssim n^{-1+\alpha} \log n.$$

Remark 3. *The convergence rate established in Theorem 7 is notable for two key reasons.*

First, when combined with Theorem 3, which analyzes the simple mean estimation setting, it demonstrates that the rate $n^{-1+\alpha} \log n$ is tight — that is, it cannot be improved in general. Theorem 3 shows that this rate is achieved exactly in the simplest setting, implying that our upper bound in the general model is sharp for the proposed de-biased estimator.

Second, to the best of our knowledge, this rate is the best known in the literature for online covariance estimation under SGD without requiring information from the Hessian. Specifically, the batch-means estimators studied in Chen et al. (2021a) and Zhu et al. (2023) yield upper bounds of order $\mathcal{O}(n^{(-1+\alpha)/4})$ for $\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|$, which is is significantly slower than our bound:

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\| \leq \sqrt{\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|^2} \lesssim n^{(-1+\alpha)/2}\sqrt{\log n}.$$

Moreover, while the plug-in method in Chen et al.

Table 3: Comparison of De-biased and Online BM Estimators for Different Models with $d = 50$. Standard Deviations are Reported in Parentheses.

		$n = 125000$	$n = 250000$	$n = 500000$
Linear	De-biased	9.96 (0.52)	8.63 (0.37)	7.91 (0.32)
	Online BM	14.37 (1.22)	13.31 (0.98)	12.68 (0.91)
Logistic	De-biased	96.07 (3.69)	94.09 (2.90)	92.78(2.59)
	Online BM	124.77 (8.48)	122.90 (6.96)	122.14 (7.12)
Expectile	De-biased	9.94 (0.49)	8.76 (0.38)	8.12 (0.35)
	Online BM	15.10 (1.22)	14.15 (1.02)	13.62 (0.97)

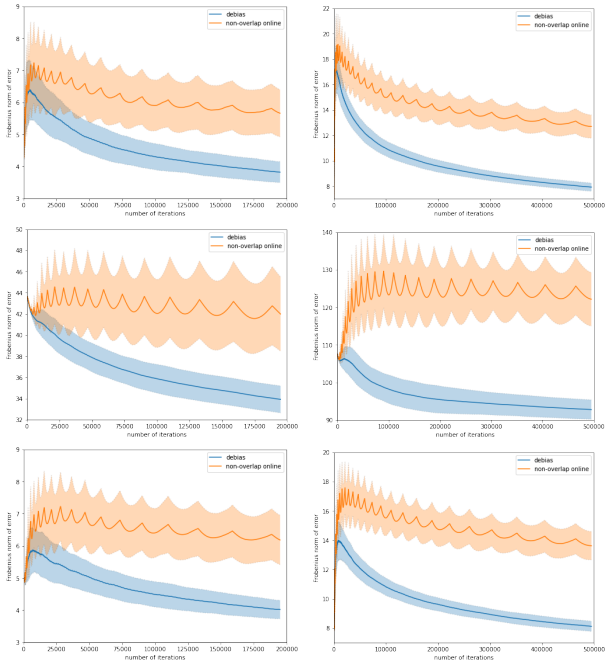

Figure 3: The Estimation Error at Each Iteration. Left: $d = 20$. Right: $d = 50$. Top: Linear Regression Model. Middle: Logistic Regression Model. Bottom: Expectile Regression Model. The Error Band Represents One Standard Deviation.

(2021a) achieves a faster rate of $\mathcal{O}(n^{-\alpha/2})$, it relies heavily on repeated Hessian computation and matrix inversion, making it computationally expensive. In contrast, our estimator achieves a comparable rate—particularly when $\alpha \approx 1/2$, while remaining fully online and computationally efficient, without requiring any access to second-order information.

5 NUMERICAL STUDY

In this section, we assess the empirical performance of the de-biased estimator across different settings. In the subsequent numerical experiments, $\xi_i = (a_i, b_i), i = 1, 2, ...$ are *i.i.d.* random vectors, where $a_i \sim \mathcal{N}(0, \mathbf{I}_d)$, and b_i are drawn from different distributions based on the linear, logistic and expectile regression models.

For the linear regression model, $b_i \sim \mathcal{N}(a_i^\mathsf{T} x^*, 1)$, and the loss function is defined as the squared loss

$$f(x, \xi_i = (a_i, b_i)) = (a_i^\mathsf{T} x - b_i)^2/2.$$

For the logistic regression model, $b_i \in \{1, -1\}$ is generated from a Bernoulli distribution with the probability given by $\mathbb{P}(b_i|a_i) = 1/(1 + \exp(-b_i a_i^\mathsf{T} x^*))$. We use the logit loss

$$f(x, \xi_i = (a_i, b_i)) = \log(1 + \exp(-b_i a_i^\mathsf{T} x)).$$

Notice that both loss functions are the corresponding negative log-likelihood.

For expectile regression (Newey and Powell, 1987), which has been extensively studied and applied in the statistical and economic literature (Efron, 1991; Taylor, 2008; Daouia et al., 2024), we specify the model as follows. Rewriting x as $(x, x_0) \in \mathbb{R}^{d+1}$ where x_0 is a univariate intercept term. The loss function for τ-th expectile regression is defined as

$$f(x, x_0, \xi) = |\tau - 1_{\{b < a^\mathsf{T} x + x_0\}}|(b - a^\mathsf{T} x - x_0)^2,\ 0 < \tau < 1,$$

and $b_i \sim \mathcal{N}(a_i^\mathsf{T} x^*, 1)$.

The true parameter x^* is an arithmetic sequence from 0 to 1 with length d. We evaluate and report the estimation error $\|\bar{\Sigma}_n - \Sigma\|_F$. All reported results are averaged over 500 independent runs. The learning rate in SGD is chosen as $\eta_i = 0.5i^{-0.505}$, following the choice in Zhu et al. (2023).

For the linear regression model, a straightforward derivation yields $A = S = \mathbf{I}_d$ and thus $\Sigma = \mathbf{I}_d$. For the logistic regression model, since an explicit form for A and S is challenging to derive, we empirically estimate the covariance matrix Σ using Monte Carlo simulations.

In the expectile regression setting, the unique optimizer of its objective function is (x^*, x_0^τ), where x_0^τ is the τ-th expectile of a univariate standard Gaussian distribution. Elementary calculus yields

$$A = 2[(1 - \tau) + (2\tau - 1)\,\Phi(-x_0^\tau)]\mathbf{I}_d,$$

$$S = 4(\tau^2\alpha_+ + (1-\tau)^2\alpha_-)\mathbf{I}_d,$$

with

$$\alpha_+ = \Phi(-x_0^\tau)\,(1+(x_0^\tau)^2) - x_0^\tau\phi(x_0^\tau)$$

$$\alpha_- = (1+\Phi(x_0^\tau))\,(1+(x_0^\tau)^2) + x_0^\tau\phi(x_0^\tau),$$

and Φ and ϕ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. Therefore, we have the analytic expression for $\Sigma = A^{-1}SA^{-1}$.

We begin by investigating the impact of batch size on the convergence rate of the de-biased estimator in a simple one-dimensional linear model, where the loss is defined as the absolute error of the estimated variance. For the online approach of the de-biased estimator, the batch sequence $\{a_k\}$ is chosen as $a_k = Ca_k^0$ for some constant C, where

$$a_m^0 - a_{m-1}^0 + 1 = \lfloor (a_m^0)^\alpha \log(a_m^0) \rfloor, m \geq 2. \qquad (12)$$

In the appendix, we will prove that this choice of the batch sequence also ensures $|B_i| = \mathcal{O}(i^\alpha \log i)$, similar to Proposition 1. Figure 2 shows that smaller values of C lead to a slight reduction in estimation error. Overall, the convergence rate—reflected in the slope—remains largely unaffected by variations in C within a reasonable range, demonstrating the robust convergence behavior of the de-biased estimator. Unless otherwise specified, we choose $C = 0.5$ in the following simulations.

We compare the estimation error of our de-biased method with that of the online batch-means method (non-overlap version) proposed by Zhu et al. (2023), which is, to the best of our knowledge, the only Hessian-free online estimator available in the literature. Figure 3 displays the Frobenius norm of the estimation error plotted against the number of iterations for linear, logistic, and expectile regression models. The results highlight a noticeably sharper convergence rate and substantially lower estimation error of the de-biased estimator. Additional plots for other settings are provided in the supplementary material. As shown in Tables 1-3, the de-biased approach consistently achieves higher accuracy and stability than the non-overlap online method across different model types and dimensional settings as the number of iterations increases.

In the appendix, we also present numerical experiment results to demonstrate the application of confidence interval construction. We evaluate the performances under the same settings. Post the estimation of the limiting covariance matrix with $\widehat{\Sigma}_n$, we construct the $(1-q)100\%$ confidence interval for the j-th coordinates of x^* as

$$\left[(\bar{x}_n)_j - z_{1-q/2}\sqrt{\widehat{\Sigma}_{n,jj}/n}, (\bar{x}_n)_j + z_{1-q/2}\sqrt{\widehat{\Sigma}_{n,jj}/n}\right],$$

where $(\bar{x}_n)_j$ represents the j-th coordinates of the averaged SGD after n iterations, $\widehat{\Sigma}_{n,jj}$ is the j-th diagonal of $\widehat{\Sigma}_n$, and $z_{1-q/2}$ is the $1-q/2$-th percentile of the standard Gaussian distribution.

6 DISCUSSION

In this article, we introduce a novel, fully online de-biased covariance estimator for averaged SGD. The bias-reduction technique significantly improves the estimation accuracy over existing Hessian-free approaches, leading to a convergence rate of $n^{(\alpha-1)/2}\sqrt{\log n}$. Our approach requires substantially less computation than the bootstrap and plug-in estimators while operating entirely on a single SGD trajectory. Beyond its computational efficiency, our method remains applicable in constrained settings where the Hessian or raw data are inaccessible.

Several promising directions remain for future work. First, a non-asymptotic evaluation of confidence intervals could be obtained via advanced non-asymptotic Gaussian approximations (Shao and Zhang, 2022; Wei et al., 2025; Sheshukova et al., 2025) when combined with the de-biased covariance estimator. Our refined convergence result would yield a sharper guarantee for coverage rates in finite-sample regimes. Second, since all matrix norms are equivalent for a fixed dimension, our main theorem naturally applies to other matrix metrics, such as the L_1 and Frobenius norms. However, explicitly characterizing the dimension-dependent convergence rate under specific matrix norms remains a highly valuable extension. Finally, since our proposed method can be viewed as a de-biased adaptation of the online batch-means estimator, we anticipate that similar generalizations to nonsmooth problems—as recently explored by Jiang et al. (2025)—should also apply to the bias-reduction approach.

Acknowledgements

We sincerely thank the program chair, senior area chair, area chair, and the four reviewers for their constructive feedback and involved discussion, which has greatly improved the clarity of our paper. Wei Biao Wu's research is partially supported by the NSF (Grant NSF/DMS-2311249).

References

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning.

SIAM review, 60(2):223–311.

Chen, H., Lu, W., and Song, R. (2021a). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719.

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1).

Chen, X., Liu, W., and Zhang, Y. (2021b). First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994 – 3021.

Damerdji, H. (1991). Strong consistency and other properties of the spectral variance estimator. *Management Science*, 37(11):1424–1440.

Daouia, A., Stupfler, G., and Usseglio-Carleve, A. (2024). An expectile computation cookbook. *Statistics and Computing*, 34(3):103.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, pages 93–125.

Fang, Y., Xu, J., and Yang, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034 – 1070.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, pages 473–483.

Glynn, P. W. and Iglehart, D. L. (1990). Simulation output analysis using standardized time series. *Mathematics of Operations Research*, 15(1):1–16.

Glynn, P. W. and Whitt, W. (1991). Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435.

Jiang, L., Roy, A., Balasubramanian, K., Davis, D., Drusvyatskiy, D., and Na, S. (2025). Online covariance estimation in nonsmooth stochastic approximation. *arXiv preprint arXiv:2502.05305*.

Kuang, W., Anitescu, M., and Na, S. (2025). Online covariance matrix estimation in sketched newton methods. *arXiv preprint arXiv:2502.07114*.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method. Preprint. Available at arXiv:1212.2002.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2022). Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389.

Li, C. J., Mou, W., Wainwright, M., and Jordan, M. (2022). Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, pages 909–981. PMLR.

Li, T., Liu, L., Kyrillidis, A., and Caramanis, C. (2018). Statistical inference using sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory*, 26(4):1218–1245.

Luo, Y., Huo, X., and Mei, Y. (2022). Covariance estimators for the root-sgd algorithm in online learning. *arXiv preprint arXiv:2212.01259*.

Na, S. and Mahoney, M. (2025). Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *Journal of Machine Learning Research*, 26(33):1–75.

Na, S. and Mahoney, M. W. (2022). Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv: 2205.13687 v1*.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

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

Rio, E. (2009). Moment Inequalities for Sums of Dependent Random Variables under Projective Conditions. *Journal of Theoretical Probability*, 22(1):146–163.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 22(4):400–407.

Shao, Q.-M. and Zhang, Z.-S. (2022). Berry–esseen bounds for multivariate nonlinear statistics with applications to m-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576.

Sheshukova, M., Samsonov, S., Belomestny, D., Moulines, E., Shao, Q.-M., Zhang, Z.-S., and Naumov, A. (2025). Gaussian approximation and multiplier bootstrap for stochastic gradient descent. *arXiv preprint arXiv:2502.06719*.

Su, W. J. and Zhu, Y. (2023). Higrad: Uncertainty quantification for online learning and stochastic approximation. *Journal of Machine Learning Research*, 24(124):1–53.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252.

Wei, Z., Li, J., Lou, Z., and Wu, W. B. (2025). Gaussian approximation and concentration of constant learning-rate stochastic gradient descent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wei, Z., Zhu, W., and Wu, W. B. (2026). General weighted averaging in stochastic gradient descent: Clt and adaptive optimality. In *The 29th International Conference on Artificial Intelligence and Statistics*.

Xiao, H. and Wu, W. B. (2011). A single-pass algorithm for spectrum estimation with fast convergence. *IEEE transactions on information theory*, 57(7):4720–4731.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1):466 – 493.

Zhang, D. and Wu, W. B. (2017). Asymptotic theory for estimators of high-order statistics of stationary processes. *IEEE Transactions on Information Theory*, 64(7):4907–4922.

Zhong, Y., Kuffner, T., and Lahiri, S. (2023). Online bootstrap inference with nonconvex stochastic gradient descent estimator. *arXiv preprint arXiv:2306.02205*.

Zhu, W., Chen, X., and Wu, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404.

Zhu, W., Lou, Z., Wei, Z., and Wu, W. B. (2024). High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*.

Checklist

(i) For all models and algorithms presented, check if you include:

- (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
- (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
- (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

(ii) For any theoretical claim, check if you include:

- (a) Statements of the full set of assumptions of all theoretical results. [Yes]
- (b) Complete proofs of all theoretical results. [Yes]
- (c) Clear explanations of any assumptions. [Yes]

(iii) For all figures and tables that present empirical results, check if you include:

- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
- (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
- (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

(iv) If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

- (a) Citations of the creator If your work uses existing assets. [Not Applicable]
- (b) The license information of the assets, if applicable. [Not Applicable]
- (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
- (d) Information about consent from data providers/curators. [Not Applicable]
- (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

We are not using existing assets or releasing new assets.

(v) If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

We did not use crowdsourcing or conduct research with human subjects.

Refining Covariance Matrix Estimation in Stochastic Gradient Descent Through Bias Reduction: Supplementary Materials

The supplement material is organized as follows: In section A, we demonstrate the high-level idea behind the proof of the main theorem. In section B, we introduce some useful technical lemmas. In section C, we prove the consistency of our proposed de-biased estimator under the mean estimation model, as well as Proposition 2 and Theorem 3. In section D we prove the main theorem in general cases, i.e., Theorem 7. Additional details and results of numerical experiments are provided in Section F.

A PROOF SKETCH OF THE MAIN THEOREM

For a vector $x \in \mathbb{R}^d$, we use $|x|$ to denote its Euclidean norm. For a random variable $X \in \mathbb{R}^d$, we use $\|X\|_p = (\mathbb{E}\,|X|^p)^{1/p}$ to denote its L_p norm, and write $\|X\| = \|X\|_2$. For a (random) matrix $A \in \mathbb{R}^{d\times d}$, we use $|A|$ or $|A|_2$ to denote its operator norm, use $|A|_F$ to denote its Frobenius norm, and use $\lambda_{max}(A)$, $\lambda_{min}(A)$ to denote its maximum and minimum eigenvalue. Write $\|A\|_p = (\mathbb{E}\,|A|^p)^{1/p}$ and $\|A\| = \|A\|_2$. For a finite set B, we use $|B|$ to denote its cardinality.

Here and in the sequel, let the SGD iterates take the following form

$$\begin{aligned} X_i &= X_{i-1} - \eta_i \nabla f(X_{i-1}, \xi_i) \\ &= X_{i-1} - \eta_i \nabla F(X_{i-1}) + \eta_i \epsilon_i, \end{aligned} \tag{13}$$

where $\epsilon_i = \nabla F(X_{i-1}) - \nabla f(X_{i-1}, \xi_i)$ is a martingale difference. Define the error sequence

$$\delta_i = X_i - x^*$$

which satisfies the recursion

$$\delta_i = \delta_{i-1} - \eta_i \nabla F(X_{i-1}) + \eta_i \epsilon_i. \tag{14}$$

By expressing X_i in terms of δ_i, the covariance estimator $\widehat{\Sigma}_n$ can be written as:

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\left[(\delta_i - \bar{\delta}_n)\Big(\sum_{k=t_i}^{i}\delta_k - \ell_i\bar{\delta}_n\Big)^{\mathsf{T}} + \Big(\sum_{k=t_i}^{i}\delta_k - \ell_i\bar{\delta}_n\Big)(\delta_i - \bar{\delta}_n)^{\mathsf{T}} - (\delta_i - \bar{\delta}_n)(\delta_i - \bar{\delta}_n)^{\mathsf{T}}\right],$$

where $t_i = i - \ell_i + 1$. The goal is to bound the mean squared error:

$$\mathrm{MSE}(\widehat{\Sigma}_n) = \|\widehat{\Sigma}_n - \Sigma\|^2 \lesssim O(n^{-1+\alpha}\log n).$$

Our strategy is to first approximate the error sequence δ_i using a linear approximation, and then apply a martingale decomposition involving *i.i.d* approximation. Specifically, we define

$$U_i = (1 - \eta_i A)U_{i-1} + \eta_i\epsilon_i, U_0 = \delta_0,$$

as the linear approximation sequence, derived from a first-order Taylor expansion. Letting $r_i = \nabla F(X_{i-1}) - A\delta_{i-1}$ for $i \geq 1$ and $r_0 = 0$, the error due to linear approximation satisfies the recursion

$$s_i = (1 - \eta_i A)s_{i-1} - \eta_i r_i, s_0 = 0.$$

Similarly, let $\epsilon_i^* = -\nabla f(x^*, \xi_i)$, which forms a mean-zero i.i.d. sequence. Then, the *i.i.d* approximation sequence is given by

$$\widetilde{U}_i = (1 - \eta_i A)\widetilde{U}_{i-1} + \eta_i\epsilon_i^*, \widetilde{U}_0 = \delta_0.$$

The error sequence due to *i.i.d* approximation is

$$\Delta_i = (1 - \eta_i A)\Delta_{i-1} + \eta_i v_i, \Delta_0 = 0,$$

where $v_i = \epsilon_i - \epsilon_i^*$.

We next define the estimator without centering:

$$\widehat{\Sigma}_{n,\delta} = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\Big(\sum_{k=t_i}^{i}\delta_k\Big)^{\mathsf{T}} + \Big(\sum_{k=t_i}^{i}\delta_k\Big)\delta_i^{\mathsf{T}} - \delta_i\delta_i^{\mathsf{T}}\right], \tag{15}$$

and the corresponding estimators based on the linear and i.i.d. approximations:

$$\widehat{\Sigma}_{n,U} = \frac{1}{n}\sum_{i=1}^{n}\left[U_iU_i^{\mathsf{T}} + U_i\Big(\sum_{k=t_i}^{i-1}U_k\Big)^{\mathsf{T}} + \Big(\sum_{k=t_i}^{i-1}U_k\Big)U_i^{\mathsf{T}}\right],$$
$$\widehat{\Sigma}_{n,\widetilde{U}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{U}_i\widetilde{U}_i^{\mathsf{T}} + \widetilde{U}_i\Big(\sum_{k=t_i}^{i-1}\widetilde{U}_k\Big)^{\mathsf{T}} + \Big(\sum_{k=t_i}^{i-1}\widetilde{U}_k\Big)\widetilde{U}_i^{\mathsf{T}}\right].$$

To bound the estimation error, we decompose it as follows:

$$\|\widehat{\Sigma}_n - \Sigma\| \le \|\widehat{\Sigma}_n - \widehat{\Sigma}_{n,\delta}\| + \|\widehat{\Sigma}_{n,\delta} - \widehat{\Sigma}_{n,U}\| + \|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\| + \|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\|,$$

We will use these approximations, both in the mean estimation and general cases, to establish the convergence of our proposed estimator in the subsequent sections.

B TECHNICAL SUPPLEMENT

In this section, we introduce some fundamental technical results, which are frequently applied in the following analysis of SGD iterates. We defer the proof to Section E. We also prove Proposition 1 in the paper, which quantifies the batch size.

Lemma 8. *Let $\lambda > 0$ be a constant. For any $i \in \mathbb{N}^+$, define a real sequence $\{{Y_{(\lambda)}}_j^i\}$ with*

$${Y_{(\lambda)}}_j^i = \begin{cases} 1 & \text{if } j = i, \\ \prod_{k=j+1}^{i}(1 - \lambda\eta_k) & \text{if } j < i. \end{cases} \tag{16}$$

Then for all $0 \le j \le i$, we have

(i)

$$|{Y_{(\lambda)}}_j^i| \asymp \exp\left\{\frac{\lambda\eta}{1-\alpha}\left(j^{1-\alpha} - i^{1-\alpha}\right)\right\} \le \exp\left\{-\lambda\eta i^{-\alpha}(i-j)\right\}. \tag{17}$$

and $|{Y_{(\lambda)}}_j^i| \asymp \exp\{-\lambda\eta i^{-\alpha}(i-j)\}$ if $j \ge i - Ci^{\alpha}\log i$ for some constant C.

(ii) For $\beta, \gamma > 0$,

$$\sum_{j=1}^{i}\exp\left(\beta j^{1-\alpha}\right)j^{-\gamma\alpha} \asymp \exp\left(\beta i^{1-\alpha}\right)i^{-(\gamma-1)\alpha}, \tag{18}$$

which implies

$$\sum_{j=1}^{i}|{Y_{(\lambda)}}_j^i|^{\beta}|j^{-\alpha}|^{\gamma} \asymp i^{-(\gamma-1)\alpha}. \tag{19}$$

(iii) For any $n > i$,

$$S_{(\lambda)_i^n} := \sum_{k=i+1}^{n} Y_{(\lambda)_i^k} \lesssim (i+1)^{\alpha}. \tag{20}$$

Lemma 9. *For $\gamma > 0$,*

$$\sum_{j=t_i}^{i} j^{-\gamma} \lesssim i^{-\gamma}\ell_i \asymp i^{\alpha-\gamma}\log i. \tag{21}$$

Lemma 10 (Rosenthal inequality)**.** *If $\xi_1, \cdots, \xi_n$ are independent zero mean random variables and $p \geq 2$, then there exist constant C_p only depends on p such that*

$$\left\| \sum_{i=1}^{n} \xi_i \right\|_p \leq C_p \max\left\{ \left(\sum_{i=1}^{n} \mathbb{E}\left|\xi\right|^2 \right)^{1/2}, \left(\sum_{i=1}^{n} \mathbb{E}\left|\xi\right|^p \right)^{1/p} \right\}. \tag{22}$$

Lemma 11 (Burkholder inequality)**.** *If $D_1, \cdots, D_n$ are martingale differences and $p \geq 2$, then*

$$\left\| \sum_{i=1}^{n} D_i \right\|_p^2 \leq (p-1) \sum_{i=1}^{n} \left\| D_i \right\|_p^2. \tag{23}$$

B.1 Proof of Proposition 1

Proof. We consider two cases

(i) Case $i = a_m$: By the definition of the sequence $\{a_m\}$, we have

$$|B_i| = a_m - a_{m-1} + 1 = \lfloor a_m^{\alpha} \log a_m \rfloor = \lfloor i^{\alpha} \log i \rfloor.$$

So the claimed bounds hold exactly in this case.

(ii) Case $i > a_m$: We write the batch size as

$$|B_i| = i - a_{m-1} + 1 = (a_m - a_{m-1} + 1) + (i - a_m).$$

The first term satisfies $a_m - a_{m-1} + 1 = \lfloor a_m^{\alpha} \log a_m \rfloor$ by construction. To analyze the second term, note that:

$$\lfloor (k+1)^{\alpha} \log(k+1) \rfloor - \lfloor k^{\alpha} \log k \rfloor \leq 1, \quad \forall k \geq 1.$$

By applying a telescoping sum from $k = a_m$ to $i - 1$, we obtain

$$i - a_m \geq \lfloor i^{\alpha} \log i \rfloor - \lfloor a_m^{\alpha} \log a_m \rfloor.$$

Therefore,

$$|B_i| = \lfloor a_m^{\alpha} \log a_m \rfloor + (i - a_m) \geq \lfloor i^{\alpha} \log i \rfloor.$$

For the upper bound, note that

$$\begin{aligned} & i - a_m \\ =& a_{m+1} - a_m - (a_{m+1} - i) \\ \leq& \lfloor a_{m+1}^{\alpha} \log a_{m+1} \rfloor - \left(\lfloor a_{m+1}^{\alpha} \log a_{m+1} \rfloor - \lfloor i^{\alpha} \log i \rfloor \right) \\ =& \lfloor i^{\alpha} \log i \rfloor. \end{aligned}$$

This gives

$$\begin{aligned} |B_i| &= \lfloor a_m^{\alpha} \log a_m \rfloor + (i - a_m) \\ &\leq \lfloor i^{\alpha} \log i \rfloor + \lfloor i^{\alpha} \log i \rfloor = 2\lfloor i^{\alpha} \log i \rfloor. \end{aligned}$$

Combining both bounds yields the desired result:

$$\lfloor i^{\alpha} \log i \rfloor \leq |B_i| \leq 2\lfloor i^{\alpha} \log i \rfloor.$$

□

Remark 4. *In practice, we choose the batch sequence $\tilde{a}_m = Ca_m$ (or the closest integer to Ca_m) to obtain some flexibility. Based on Proposition 1, we can show that the batch size $|\tilde{B}_i|$ decided by $\tilde{a}_m$ also meets the condition $|\tilde{B}_i| \asymp i^\alpha \log i$: for sufficiently large i and m such that $\tilde{a}_m \le i < \tilde{a}_{m+1}$, we find an integer k such that $|k - i/C| \le 1$ and $a_m \le k < a_{m+1}$. Proposition 1 shows $k - a_{m-1} \asymp k^\alpha \log k$, which further implies*

$$i - \tilde{a}_{m-1} \asymp Ck^\alpha \log k \asymp C(\frac{i}{C})^\alpha(\log i - \log C) \asymp i^\alpha \log i.$$

C MEAN ESTIMATION MODEL

In this subsection, we consider the data $\{y_i\}_{i\in\mathbb{N}^+}$ generated by a linear model

$$y_i = x^* + e_i,$$

where $x^* \in \mathbb{R}$ is the true mean to be estimated, and e_i's are i.i.d. random errors with $\mathbb{E}[e_i] = 0$, $\mathbb{E}[e_i^2] < \infty$. Consider the squared loss function at x: $f(x, y) = (y - x)^2/2$. Clearly,

$$F(x) = \mathbb{E}_y[f(x, y)] = \frac{1}{2}\mathbb{E}[(x - x^* - e)^2] = \frac{(x - x^*)^2 + \mathbb{E}[e^2]}{2},$$

and $x^* = \arg\min_x F(x)$ holds. Then the i-th SGD iterate takes the form

$$\begin{aligned} X_i &= X_{i-1} - \eta_i \nabla f(X_{i-1}, y_i) \\ &= X_{i-1} + \eta_i(y_i - X_{i-1}), \end{aligned}$$

and the error $\delta_i = X_i - x^*$ takes the form

$$\begin{aligned} \delta_i &= \delta_{i-1} + \eta_i(e_i - \delta_{i-1}) \\ &= (1 - \eta_i)\delta_{i-1} + \eta_i e_i, \end{aligned}$$

where $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$ and $\alpha \in (0.5, 1)$.

Without loss of generality, assume $x^* = 0$. Then $\delta_i = X_i$ and the SGD iterate takes the form

$$X_i = (1 - \eta_i)X_{i-1} + \eta_i e_i, \quad i \ge 1; \qquad X_0 = x_0, \tag{24}$$

where x_0 is deterministic. We first provide some useful technical lemmas in Section C.1. Then, in Section C.2, we prove Proposition 2 and Theorem 3, where e_i is assumed to be a mean-zero Gaussian random error. Finally, we generalize the results to the case where e_i is non-Gaussian and $\{X_i\}$ is multi-dimensional.

C.1 Technical Lemmas for the Mean Estimation Model

Lemma 12. *With the definition of $\{Y_{(\lambda)}{}_j^i\}$ in Lemma 8, sequence $\{X_i\}$ can be rewritten as*

$$X_i = Y_j^i X_j + \sum_{k=j+1}^{i} Y_k^i \eta_k e_k, \quad 0 \le j < i, \tag{25}$$

where $Y_k^i := Y_{(1)}{}_k^i$. Therefore for all $1 \le j < i$, we have

(i) $\mathrm{var}(X_i) \asymp \mathbb{E}[X_i^2] \asymp i^{-\alpha}$.

(ii) $0 \le \mathrm{cov}(X_i, X_j) \asymp \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} j^{-\alpha} \le \exp\left\{-\eta i^{-\alpha}(i - j)\right\} j^{-\alpha}$.

(iii) $\mathrm{cov}(X_i, \sum_{k=1}^{j} X_k) \asymp \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} \le \exp\left\{-\eta i^{-\alpha}(i - j)\right\}$.

(iv) $\mathrm{cov}(\sum_{k=i}^{\infty} X_k, X_j) \lesssim \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} i^\alpha j^{-\alpha} \le \exp\left\{-\eta i^{-\alpha}(i - j)\right\} i^\alpha j^{-\alpha}$.

Proof.

(i) Take $j = 0$ in (25),

$$X_i = Y_0^i x_0 + \sum_{k=1}^{i} Y_k^i \eta_k e_k, \quad i \geq 1. \tag{26}$$

Recall that e_k's are i.i.d., then by Lemma 8, we have

$$\operatorname{var}(X_i) = \mathbb{E}[X_i^2] = |Y_0^i|^2 x_0^2 + \sum_{k=1}^{i} |Y_k^i|^2 |\eta_k|^2 \mathbb{E}[e_k^2] \asymp \exp\left\{\frac{\lambda\eta}{1-\alpha} i^{1-\alpha}\right\} + i^{-\alpha} \asymp i^{-\alpha}.$$

(ii) Use (25) and apply Lemma 8 (i), we have

$$\begin{aligned} 0 \leq \operatorname{cov}(X_i, X_j) = Y_j^i \operatorname{var}(X_j) &\asymp \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} j^{-\alpha} \\ &\leq \exp\left\{-\eta i^{-\alpha}(i-j)\right\} j^{-\alpha}. \end{aligned} \tag{27}$$

(iii) Using result in (ii), and noticing that

$$\int \exp(\beta u^{1-\alpha}) u^{-\alpha} \, \mathrm{d}u = \frac{1}{\beta(1-\alpha)} \exp(\beta u^{1-\alpha}) + C, \tag{28}$$

we have

$$\begin{aligned} \operatorname{cov}\left(X_i, \sum_{k=1}^{j} X_k\right) &\asymp \sum_{k=1}^{j} \exp\left\{\frac{\eta}{1-\alpha}(k^{1-\alpha} - i^{1-\alpha})\right\} k^{-\alpha} \\ &= \exp\left(-\frac{\eta}{1-\alpha} i^{1-\alpha}\right) \sum_{k=1}^{j} \exp\left(\frac{\eta}{1-\alpha} k^{1-\alpha}\right) k^{-\alpha} \\ &\asymp \exp\left(-\frac{\eta}{1-\alpha} i^{1-\alpha}\right) \int_1^{j+1} \exp\left(\frac{\eta}{1-\alpha} u^{1-\alpha}\right) u^{-\alpha} \, \mathrm{d}u \qquad \text{(eventually increasing in } k\text{)} \\ &= \exp\left(-\frac{\eta}{1-\alpha} i^{1-\alpha}\right) \eta^{-1} \left[\exp\left(\frac{\eta}{1-\alpha}(j+1)^{1-\alpha}\right) - \exp(\beta)\right] \\ &\asymp \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\}. \end{aligned}$$

The next bound is again obtained by (27).

(iv) Using result in (ii), we have

$$\begin{aligned} \operatorname{cov}\left(\sum_{k=i}^{\infty} X_k, X_j\right) &\asymp \sum_{k=i}^{\infty} \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - k^{1-\alpha})\right\} j^{-\alpha} \\ &= \exp\left(\frac{\eta}{1-\alpha} j^{1-\alpha}\right) j^{-\alpha} \sum_{k=i}^{\infty} \exp\left(-\frac{\eta}{1-\alpha} k^{1-\alpha}\right) \\ &\asymp \exp\left(\frac{\eta}{1-\alpha} j^{1-\alpha}\right) j^{-\alpha} \int_{i-1}^{\infty} \exp\left(-\frac{\eta}{1-\alpha} u^{1-\alpha}\right) \mathrm{d}u \qquad \text{(decreasing in } k\text{)} \\ &\asymp \exp\left(\frac{\eta}{1-\alpha} j^{1-\alpha}\right) j^{-\alpha} \int_{\frac{\eta}{1-\alpha}(i-1)^{1-\alpha}}^{\infty} \mathrm{e}^{-t} t^{\frac{\alpha}{1-\alpha}} \, \mathrm{d}t \qquad \left(t = \tfrac{\eta}{1-\alpha} u^{1-\alpha}\right) \\ &\lesssim \exp\left(\frac{\eta}{1-\alpha} j^{1-\alpha}\right) j^{-\alpha} (i-1)^{\alpha} \exp\left(-\frac{\eta}{1-\alpha}(i-1)^{1-\alpha}\right) \\ &\asymp \exp\left(\frac{\eta}{1-\alpha} j^{1-\alpha}\right) j^{-\alpha} i^{\alpha} \exp\left(-\frac{\eta}{1-\alpha} i^{1-\alpha} + \mathcal{O}(1)\right) \\ &\asymp \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} i^{\alpha} j^{-\alpha}, \end{aligned}$$

where the following bound is used for incomplete gamma function

$$\int_a^\infty \mathrm{e}^{-x} x^\beta \,\mathrm{d}x \le a^\beta \mathrm{e}^{-a} C_\beta, \qquad (a \ge 1,\, \beta > 1)$$

by noticing $\frac{\eta}{1-\alpha}(i-1)^{1-\alpha} \ge 1$ for all i large enough and $\frac{\alpha}{1-\alpha} > 1$.
The next bound is again obtained by (27).

□

Recall the debiased estimator (in the univariate case, with sample mean $\bar{X}_n$ omitted)

$$\widehat{\sigma}_n = \frac{1}{n}\sum_{i=1}^n \left[X_i^2 + 2X_i(X_{i-1} + \cdots + X_{t_i})\right] = \frac{1}{n}\sum_{i=1}^n \left[X_i^2 + 2X_i W_i\right], \tag{29}$$

with overlapping batch length: $\ell_i = \lfloor C\, i^\alpha \log i \rfloor$, where

$$W_i := \sum_{j=t_i}^{i-1} X_j. \qquad (t_i = i - \ell_i + 1)$$

Lemma 13. *For all $1 \le j < i$, we have*

(i) $\mathrm{var}(W_i) \lesssim \ell_i$.

(ii) $\mathrm{cov}(X_i, W_j) \lesssim \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} \le \exp\{-\eta i^{-\alpha}(i-j)\}$.

(iii) $\mathrm{cov}(X_j, W_i) \lesssim \mathcal{O}(1)$.

(iv) $\mathrm{cov}(W_i, W_j) \lesssim \ell_j$.

Proof.

(i) Note that $\mathrm{cov}(X_i, X_j) \ge 0$ for all i, j. Using Lemma 12 (i, iii), we have

$$\begin{aligned}
\mathrm{var}(W_i) &= \sum_{j=t_i}^{i-1} \mathrm{var}(X_j) + 2\sum_{j=t_i+1}^{i-1}\sum_{k=t_i}^{j-1} \mathrm{cov}(X_j, X_k) \\
&\le \sum_{j=t_i}^{i-1} \mathrm{var}(X_j) + 2\sum_{j=t_i}^{i-1}\sum_{k=1}^{j-1} \mathrm{cov}(X_j, X_k) \\
&\lesssim \sum_{j=t_i}^{i-1} j^{-\alpha} + \sum_{j=t_i}^{i-1} \exp\left\{\frac{\eta}{1-\alpha}\left((j-1)^{1-\alpha} - j^{1-\alpha}\right)\right\} \\
&\le \sum_{j=t_i}^{i-1} j^{-\alpha} + \sum_{j=t_i}^{i-1} \exp\left(-\eta j^{-\alpha}\right) \\
&\lesssim \sum_{j=t_i}^{i-1} \mathcal{O}(1) \lesssim i - t_i = \ell_i - 1 \asymp \ell_i.
\end{aligned}$$

(ii) Using Lemma 12 (iii), we have

$$\mathrm{cov}(X_i, W_j) \le \sum_{k=1}^{j} \mathrm{cov}(X_i, X_k) \lesssim \exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} \le \exp\left\{-\eta i^{-\alpha}(i-j)\right\}.$$

(iii) Using Lemma 12 (ii, iii, iv), we have

$$\begin{aligned}
\operatorname{cov}(X_j, W_i) &\le \sum_{k=1}^{\infty} \operatorname{cov}(X_j, X_k) \\
&= \sum_{k=1}^{j-1} \operatorname{cov}(X_j, X_k) + \operatorname{var}(X_j) + \sum_{k=j+1}^{\infty} \operatorname{cov}(X_j, X_k) \\
&\lesssim \exp\left\{ \frac{\eta}{1-\alpha} \left((j-1)^{1-\alpha} - j^{1-\alpha} \right) \right\} + j^{-\alpha} \\
&\quad + \exp\left\{ \frac{\eta}{1-\alpha} \left(j^{1-\alpha} - (j+1)^{1-\alpha} \right) \right\} \left(\frac{j+1}{j} \right)^{\alpha} \\
&\lesssim \exp\left(-\eta j^{-\alpha} \right) + j^{-\alpha} + \exp\left\{ -\eta (j+1)^{-\alpha} \right\} \\
&= \mathcal{O}(1).
\end{aligned}$$

(iv) Using result in (iii), we simply have

$$\operatorname{cov}(W_i, W_j) = \sum_{k=t_j}^{j-1} \operatorname{cov}(X_k, W_i) \lesssim \sum_{k=t_j}^{j-1} \mathcal{O}(1) \lesssim j - t_j = \ell_j - 1 \asymp \ell_j.$$

□

In the following Section C.2, we consider the case where e_i's are Gaussian random variables and assume $x_0 = 0$, same as the setting of Proposition 2 and Theorem 3. Then the true non-asymptotic variance for the mean estimation model is

$$\sigma_n := \operatorname{var}(\sqrt{n}\bar{X}_n) = \frac{1}{n}\mathbb{E}\left| \sum_{i=1}^{n} X_i \right|^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2 + 2X_i(X_{i-1} + \cdots + X_1) \right]. \tag{30}$$

C.2 Proof of Proposition 2 and Theorem 3

Proof of Proposition 2. by using Lemma 12 (ii-iii), we get

$$\begin{aligned}
0 \le n\left(\sigma_n - \mathbb{E}[\widehat{\sigma}_n]\right) &= \sum_{i=1}^{n} 2\mathbb{E}\left[X_i(X_{i-\ell_i} + \cdots + X_1) \right] \\
&= \sum_{i=1}^{n} 2 \cdot \sum_{j=1}^{i-\ell_i} \mathbb{E}[X_i X_j] \\
&= 2\sum_{i=1}^{n} \sum_{j=1}^{i-\ell_i} \operatorname{cov}(X_i, X_j) \\
&\asymp \sum_{i=1}^{n} \exp\left\{ \frac{\eta}{1-\alpha} \left((i-\ell_i)^{1-\alpha} - i^{1-\alpha} \right) \right\} \\
&\le \sum_{i=1}^{n} \exp\left(-\eta i^{-\alpha} \ell_i \right).
\end{aligned}$$

Therefore

$$|\mathbb{E}(\widehat{\sigma}_n - \sigma_n)| \lesssim \frac{\sum_{i=1}^{n} \exp\left(-\eta i^{-\alpha} \ell_i \right)}{n}. \tag{31}$$

Moreover, if $\ell_i \lesssim i^{\alpha}$, by the same argument of Lemma 8 (i), the last inequality becomes $\asymp$ and

$$|\mathbb{E}(\widehat{\sigma}_n - \sigma_n)| \asymp \frac{\sum_{i=1}^{n} \exp\left(-\eta i^{-\alpha} \ell_i \right)}{n}. \tag{32}$$

□

Before the proof of Theorem 3, we present some preliminary results for the moments of normal distribution.

Lemma 14 (results under normality)**.** *Let X, Y, Z, W be joint Gaussian random variables with zero mean, then we have*

(i) $\operatorname{cov}(XY, ZW) = \operatorname{cov}(X, Z)\operatorname{cov}(Y, W) + \operatorname{cov}(X, W)\operatorname{cov}(Y, Z)$.

(ii) $\operatorname{cov}(X^2, Z^2) = 2\operatorname{cov}(X, Z)^2$, $\operatorname{var}(X^2) = 2\operatorname{var}(X)^2$.

(iii) $\operatorname{var}(XY) \asymp \operatorname{var}(X)\operatorname{var}(Y)$. .

Proof.

(i) By definition of joint cumulant for zero mean random variables,

$$\begin{aligned}\operatorname{cum}(X, Y, Z, W) &= \mathbb{E}[XYZW] - \mathbb{E}[XY]\mathbb{E}[ZW] - \mathbb{E}[XZ]\mathbb{E}[YW] - \mathbb{E}[XW]\mathbb{E}[YZ] \\ &= \operatorname{cov}(XY, ZW) - \operatorname{cov}(X, Z)\operatorname{cov}(Y, W) - \operatorname{cov}(X, W)\operatorname{cov}(Y, Z).\end{aligned}$$

In addition, the cumulants of degree higher than 2 of Gaussian distribution are zero, which leads to the conclusion.

(ii) By setting $X = Y$ and $Z = W$ in (i), we get the first conclusion. The second one is obtained by further letting $X = Z$.

(iii) By setting $Z = X$ and $W = Y$ in (i), we get

$$\operatorname{var}(XY) = \operatorname{var}(X)\operatorname{var}(Y) + \operatorname{cov}(X, Y)^2.$$

Note that

$$|\operatorname{cov}(X, Y)| \le \sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)},$$

therefore

$$\operatorname{var}(X)\operatorname{var}(Y) \le \operatorname{var}(XY) \le 2\operatorname{var}(X)\operatorname{var}(Y).$$

□

Theorem 15 (Gaussian mean estimation model)**.** *If $e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$ and $\ell_i = \lfloor Ci^\alpha \log(i) \rfloor$ with $\eta C > 1$, then*

$$\operatorname{MSE}(\widehat{\sigma}_n) := \mathbb{E}\left[(\widehat{\sigma}_n - \sigma_n)^2\right] \asymp n^{-1+\alpha}\log n. \tag{33}$$

Proof. Recall the bias-variance decomposition of mean squared error,

$$\mathbb{E}\left[(\widehat{\sigma}_n - \sigma_n)^2\right] = (\sigma_n - \mathbb{E}[\widehat{\sigma}_n])^2 + \operatorname{var}(\widehat{\sigma}_n). \tag{34}$$

For the bias part, by the same argument as the proof of Proposition 4.1, we have

$$|\sigma_n - \mathbb{E}[\widehat{\sigma}_n]| \lesssim \frac{\sum_{i=1}^n \exp(-\eta C \log i)}{n} \asymp \frac{\sum_{i=1}^n i^{-\eta C}}{n} \asymp n^{-1}. \tag{35}$$

For the variance part, applying Cauchy-Schwarz inequality $\operatorname{var}(A + B) \le 2(\operatorname{var}(A) + \operatorname{var}(B))$, we have

$$\operatorname{var}(n\widehat{\sigma}_n) = \operatorname{var}\left(\sum_{i=1}^n \left[X_i^2 + 2X_iW_i\right]\right) \lesssim \operatorname{var}\left(\sum_{i=1}^n X_i^2\right) + \operatorname{var}\left(\sum_{i=1}^n X_iW_i\right). \tag{36}$$

For the first term, using Lemma 12 (i, ii), we have

$$
\begin{aligned}
&\operatorname{var}\Big(\sum_{i=1}^{n} X_i^2\Big) \\
&= \sum_{i=1}^{n} \operatorname{var}(X_i^2) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{cov}(X_i^2, X_j^2) \\
&= 2\sum_{i=1}^{n} \operatorname{var}(X_i)^2 + 4\sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{cov}(X_i, X_j)^2 && \text{(under Gaussian)} \\
&\lesssim \sum_{i=1}^{n} i^{-2\alpha} + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \exp\left\{\frac{2\eta}{1-\alpha}(j^{1-\alpha} - i^{1-\alpha})\right\} j^{-2\alpha} \\
&\asymp \mathcal{O}(1) + \sum_{i=2}^{n} \exp\left(-\frac{2\eta}{1-\alpha} i^{1-\alpha}\right) \sum_{j=1}^{i-1} \exp\left(\frac{2\eta}{1-\alpha} j^{1-\alpha}\right) j^{-2\alpha} && (\alpha > 1/2) \\
&\lesssim \mathcal{O}(1) + \sum_{i=2}^{n} \exp\left(-\frac{2\eta}{1-\alpha} i^{1-\alpha}\right) \int_0^i \exp\left(\frac{2\eta}{1-\alpha} u^{1-\alpha}\right) u^{-\alpha}\, \mathrm{d}u && \text{(drop a } j^{-\alpha}\text{)} \\
&= \mathcal{O}(1) + \sum_{i=2}^{n} \exp\left(-\frac{2\eta}{1-\alpha} i^{1-\alpha}\right) (2\eta)^{-1} \left[\exp\left(\frac{2\eta}{1-\alpha} i^{1-\alpha}\right) - 1\right] && \text{(by (28))} \\
&\asymp \mathcal{O}(1) + \sum_{i=2}^{n} \mathcal{O}(1) \asymp n. && (37)
\end{aligned}
$$

For the second term,

$$
\operatorname{var}\Big(\sum_{i=1}^{n} X_i W_i\Big) = \sum_{i=1}^{n} \operatorname{var}(X_i W_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{cov}(X_i W_i, X_j W_j) := I + II,
$$

where, using Lemma 12 (i) and Lemma 13 (i), we have

$$
\begin{aligned}
I &\asymp \sum_{i=1}^{n} \operatorname{var}(X_i)\operatorname{var}(W_i) && \text{(under Gaussian)} \\
&\lesssim \sum_{i=1}^{n} i^{-\alpha} \cdot \ell_i \\
&\asymp \sum_{i=1}^{n} \log i \asymp n \log n, && (\ell_i = \lfloor C i^{\alpha} \log i \rfloor)
\end{aligned}
$$

and (under Gaussian)

$$
II \asymp \sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{cov}(X_i, X_j)\operatorname{cov}(W_i, W_j) + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{cov}(X_i, W_j)\operatorname{cov}(X_j, W_i) := IIA + IIB.
$$

For the first part, using Lemma 12 (i) and Lemma 13 (iv),

$$
\begin{aligned}
\mathit{IIA} &\lesssim \sum_{i=1}^{n}\sum_{j=2}^{i-1}\exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha}-i^{1-\alpha})\right\}j^{-\alpha}\ell_j && (38)\\
&\lesssim \sum_{i=1}^{n}\sum_{j=2}^{i-1}\exp\left\{-\eta i^{-\alpha}(i-j)\right\}\log j && (\ell_j=\lfloor Cj^{\alpha}\log j\rfloor)\\
&\le \sum_{i=1}^{n}\log i\sum_{k=1}^{\infty}\exp\left(-\eta i^{-\alpha}k\right) && (k=i-j)\\
&= \sum_{i=1}^{n}\log i\cdot\frac{\exp\left(-\eta i^{-\alpha}\right)}{1-\exp\left(-\eta i^{-\alpha}\right)}\\
&= \sum_{i=1}^{n}\log i\left(\exp\left(\eta i^{-\alpha}\right)-1\right)^{-1}\\
&\le \sum_{i=1}^{n}\eta^{-1}i^{\alpha}\log i && (\exp(x)-1\ge x)\\
&\asymp n^{1+\alpha}\log n.
\end{aligned}
$$

Similarly, for the second part, using Lemma 13 (ii, iii),

$$
\mathit{IIB}\lesssim\sum_{i=1}^{n}\sum_{j=2}^{i-1}\exp\left\{\frac{\eta}{1-\alpha}(j^{1-\alpha}-i^{1-\alpha})\right\},
$$

which is bounded by (38) in *IIA*. So $\mathit{II}\asymp\mathit{IIA}+\mathit{IIB}\lesssim n^{1+\alpha}\log n$, and therefore

$$
\operatorname{var}\left(\sum_{i=1}^{n}X_iW_i\right)=I+\mathit{II}\lesssim n\log n+n^{1+\alpha}\log n\asymp n^{1+\alpha}\log n. \tag{39}
$$

Substitute (37, 39) into (36), we have

$$
\operatorname{var}(n\widehat{\sigma}_n)\lesssim\operatorname{var}\left(\sum_{i=1}^{n}X_i^2\right)+\operatorname{var}\left(\sum_{i=1}^{n}X_iW_i\right)\lesssim n+n^{1+\alpha}\log n\asymp n^{1+\alpha}\log n.
$$

That is,

$$
\operatorname{var}(\widehat{\sigma}_n)\lesssim n^{-1+\alpha}\log n. \tag{40}
$$

For the lower bound, it suffices to show that

$$
\operatorname{var}(\sum_{i=1}^{n}X_iW_i)\gtrsim n^{\alpha+1}\log n.
$$

To this end, we use the fact that all covariances $\operatorname{cov}(X_iX_j)\ge 0$. Hence

$$
\begin{aligned}
\operatorname{var}(\sum_{i=1}^{n}X_iW_i) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_iW_i,X_jW_j)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}[\operatorname{cov}(X_i,X_j)\operatorname{cov}(W_i,W_j)+\operatorname{cov}(X_i,W_j)\operatorname{cov}(W_i,X_j)]\\
&\ge \sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_i,X_j)\operatorname{cov}(W_i,W_j)\\
&\ge \sum_{i=\mathcal{B}n}^{n-\mathcal{B}n}\ \sum_{j=i}^{i+\tau n^{\alpha}\log n}\operatorname{cov}(X_i,X_j)\operatorname{cov}(W_i,W_j)
\end{aligned}
$$

for some constant $0 < \tau < \mathcal{B} < 1/2$, and $\tau < C\mathcal{B}^\alpha$. Notice that for $\mathcal{B}n \le i \le j \le i + \tau n^\alpha \log n \le n$, the set $\{X_{t_i}, ..., X_i\}$ and $\{X_{t_j}, ..., X_j\}$ have at least

$$Ci^\alpha \log i - \tau n^\alpha \log n \ge (C\mathcal{B}^\alpha - \tau) n^\alpha \log n + C\mathcal{B}^\alpha n^\alpha \log \mathcal{B} \gtrsim \widetilde{C} n^\alpha \log n.$$

consecutive common terms for some constant $\widetilde{C}$, and at most $Cn^\alpha \log n$ common terms. Denote $\mathcal{A}_{ij}$ as the intersection of $\{X_{t_i}, ..., X_i\}$ and $\{X_{t_j}, ..., X_j\}$. When $\mathcal{B}n \le i \le j \le i + \tau n^\alpha \log n \le n$, we have

$$\operatorname{cov}(W_i, W_j) \ge \sum_{k,l \in \mathcal{A}_{ij}} \operatorname{Cov}(X_k, X_l) = \sum_{k \in \mathcal{A}_{ij}} \sum_{l \in \mathcal{A}_{ij}} \operatorname{cov}(X_k, X_l). \tag{41}$$

Then we claim that for any batch $\{X_t, ..., X_{t+m}\}$ with $t \asymp n$ and $m \asymp n^\alpha \log n$,

$$\sum_{i=t}^{t+m} \operatorname{cov}(X_t, X_i) \gtrsim 1.$$

The proof of this claim leverages Lemma 12 (i) and the argument of Lemma 8 (i):

$$\begin{aligned}
\sum_{i=t}^{t+m} \operatorname{cov}(X_t, X_i) &\asymp \sum_{i=t}^{t+m} t^{-\alpha} \exp\{\frac{\eta}{1-\alpha}(t^{1-\alpha} - i^{1-\alpha})\} \\
&= \sum_{i=t}^{t+m} t^{-\alpha} \exp\left\{-\eta t^{-\alpha}(i-t) + \frac{1}{2}\eta\alpha u^{-\alpha-1}(i-t)^2\right\} && \text{(for some } u \in (t,i)) \\
&\asymp \sum_{i=t}^{t+m} t^{-\alpha} \exp\left\{-\eta t^{-\alpha}(i-t)\right\} && (u^{-\alpha-1}(i-t)^2 \to 0) \\
&= t^{-\alpha} \frac{1 - \exp(-\eta t^{-\alpha}(m+1))}{1 - \exp(-\eta t^{-\alpha})} \gtrsim 1.
\end{aligned}$$

The claim and (41) immediately imply $\operatorname{cov}(W_i, W_j) \gtrsim n^\alpha \log n$ when $\mathcal{B}n \le i \le j \le i + \tau n^\alpha \log n \le n$. Moreover, we have

$$\begin{aligned}
\operatorname{var}(\sum_{i=1}^{n} X_i W_i) &\ge \sum_{i=\mathcal{B}n}^{n-\mathcal{B}n} \sum_{j=i}^{i+\tau n^\alpha \log n} \operatorname{cov}(X_i, X_j) \operatorname{cov}(W_i, W_j) \\
&\gtrsim \sum_{i=\mathcal{B}n}^{n-\mathcal{B}n} \sum_{j=i}^{i+\tau n^\alpha \log n} \operatorname{cov}(X_i, X_j) n^\alpha \log n \\
&\gtrsim n^\alpha \log n \sum_{i=\mathcal{B}n}^{n-\mathcal{B}n} 1 \ge (1 - 2\mathcal{B}) n^{1+\alpha} \log(n).
\end{aligned}$$

This completes the proof of the lower bound, and we have $\operatorname{var}(\widehat{\sigma}_n) \asymp n^{-1+\alpha} \log n$. Finally, by (34), we get

$$\mathbb{E}\left[(\widehat{\sigma}_n - \sigma_n)^2\right] = (\sigma_n - \mathbb{E}[\widehat{\sigma}_n])^2 + \operatorname{var}(\widehat{\sigma}_n) \asymp n^{-1+\alpha} \log n.$$

□

Now we are ready to prove Theorem 3 by a slight modification of the proof of Theorem 15.

Proof of Theorem 3. Notice that the only difference between Theorem 3 and Theorem 15 is the target variance, σ and σ_n. The variance of the estimator remains unchanged, and we have obtained that $\operatorname{var}(\widehat{\sigma}_n) \asymp n^{-1+\alpha} \log n$. For the bias part, in the proof of Proposition 20 (the fourth term in (48)), we show that $|\sigma_n - \sigma| \lesssim n^{-1+\alpha}$. Together with (35) and triangle inequality, we have $|\mathbb{E}\widehat{\sigma}_n - \sigma| \lesssim n^{-1+\alpha}$. Finally, the bias-variance decomposition implies

$$\mathbb{E}\left[(\widehat{\sigma}_n - \sigma)^2\right] = (\sigma - \mathbb{E}[\widehat{\sigma}_n])^2 + \operatorname{var}(\widehat{\sigma}_n) \asymp n^{-1+\alpha} \log n. \tag{42}$$

□

C.3 Generalization to Non-Gaussian Error Case

Theorem 15 can be generalized to the non-Gaussian error case by the following result.

Lemma 16 (normal comparison)**.** *Consider the following two SGD iterates*

$$X_i = (1-\eta_i)X_{i-1} + \eta_i e_i, \quad i \geq 1; \qquad X_0 = 0,$$
$$\widetilde{X}_i = (1-\eta_i)\widetilde{X}_{i-1} + \eta_i \widetilde{e}_i, \quad i \geq 1; \qquad \widetilde{X}_0 = 0,$$

with $e_i \overset{\text{i.i.d.}}{\sim} \Pi$, $\mathbb{E}[e_i] = 0$, $\mathbb{E}[e_i^2] = \sigma_e^2$, $\mathbb{E}[e_i^4] < \infty$, $\widetilde{e}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$. *The corresponding estimators are*

$$\widehat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\left[X_i^2 + 2X_i(X_{i-1} + \cdots + X_{t_i})\right],$$
$$\widetilde{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{X}_i^2 + 2\widetilde{X}_i(\widetilde{X}_{i-1} + \cdots + \widetilde{X}_{t_i})\right].$$

Then there exists constants $0 \leq C_1 \leq C_2 < \infty$, *such that* $C_1 \operatorname{var}(\widetilde{\sigma}_n) \leq \operatorname{var}(\widehat{\sigma}_n) \leq C_2 \operatorname{var}(\widetilde{\sigma}_n)$.

Proof. Note that $\widehat{\sigma}_n$ can be expressed in the following form

$$\widehat{\sigma}_n = \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{i,j}e_i e_j = \sum_{i=1}^{n}\beta_{i,i}e_i^2 + \sum_{1\leq i<j\leq n}\beta_{i,j}e_i e_j.$$

It's easy to check that all $n + \binom{n}{2}$ terms in the above summation are uncorrelated with each other. Then we have

$$\begin{aligned}\operatorname{var}(\widehat{\sigma}_n) &= \sum_{i=1}^{n}\beta_{i,i}^2 \operatorname{var}(e_i^2) + \sum_{1\leq i<j\leq n}\beta_{i,j}^2 \operatorname{var}(e_i e_j) \\ &= \sum_{i=1}^{n}\beta_{i,i}^2 \cdot (\kappa-1)\sigma_e^4 + \sum_{1\leq i<j\leq n}\beta_{i,j}^2\sigma_e^4.\end{aligned}$$

where $\mathbb{E}[e_i^4] = \kappa\sigma_e^4$ for some $1 \leq \kappa < \infty$. Similarly,

$$\operatorname{var}(\widetilde{\sigma}_n^2) = \sum_{i=1}^{n}\beta_{i,i}^2 \cdot 2\sigma_e^4 + \sum_{1\leq i<j\leq n}\beta_{i,j}^2\sigma_e^4.$$

As long as $0 \leq C_1 \leq \frac{\kappa-1}{2} \leq C_2 < \infty$, we have $C_1 \operatorname{var}(\widetilde{\sigma}_n) \leq \operatorname{var}(\widehat{\sigma}_n) \leq C_2 \operatorname{var}(\widetilde{\sigma}_n)$. □

Note that the bound for the bias term in the proof of Theorem 15 is invariant under the distribution of e_i. Therefore to derive the MSE of the de-biased estimator beyond Gaussian noise, it suffices to bound the variance term, which is shown to be the same order as that in Gaussian noise cases by Lemma 16. Consequently, we have the following conclusion for the general mean estimation model.

Corollary 17 (general mean estimation model)**.** *Suppose* $x_0 = 0$, $e_i \overset{\text{i.i.d.}}{\sim} \Pi$ *with* $\mathbb{E}[e_i] = 0$ *and* $\mathbb{E}[e_i^4] < \infty$. *Then*

$$\operatorname{MSE}(\widehat{\sigma}_n) := \mathbb{E}\left[(\widehat{\sigma}_n - \sigma_n)^2\right] \lesssim n^{-1+\alpha}\log n. \tag{43}$$

Remark 5. *The conclusion also holds if the SGD has the iteration*

$$X_i = (1-\eta_i A)X_{i-1} + \eta_i e_i = (1-\eta_i A)X_{i-1} + \eta_i A(A^{-1}e_i).$$

for some $A > 0$ *and* $e_i \overset{\text{i.i.d.}}{\sim} \Pi$, $\mathbb{E}[e_i] = 0$, $\mathbb{E}[e_i^4] < \infty$. *Because we can treat* $\eta_i A$ *as the new step size, and* $A^{-1}e_i$ *as the new i.i.d noise term. Then the setting is exactly the same as the setting of Corollary 17.*

C.4 Generalize to Multi-dimension

We present the above results and proofs in the one-dimensional case. However, the generalization to the multi-dimensional setting is natural. Before the proof of the multivariate case, we first introduce an ordering Lemma that helps with the error decomposition of random matrices.

Lemma 18. *Denote $V(M) = \mathbb{E}\|\mathbb{E}M - M\|_F^2$, the sum of variances of all entries of a $d \times d$ random matrix M. Similarly, define $U(M_1, M_2)$ to be the sum of the covariances of all corresponding entries of two $d \times d$ random matrices M_1 and M_2. For constant positive-definite matrices P_1, P_2, P_3, P_4 and a random matrix M, we have*

$$|U(P_1MP_2, P_3MP_4)| \le \|P_1\|_2\|P_2\|_2\|P_3\|_2\|P_4\|_2 V(M),$$

and when P_1, P_2, P_3, P_4 are all scalar matrices, the equality holds.

Proof. Define $\mathrm{vec}(M)$ to be the vectorization of a matrix M, i.e., it stacks the columns of M into a single, long column vector. Let $\Sigma = \mathrm{Cov}(\mathrm{vec}(M))$ be the covariance matrix of $\mathrm{vec}(M)$ which is positive semi-definite. Then we have

$$\mathrm{vec}(P_1MP_2) = (P_2^\top \otimes P_1)\ \mathrm{vec}(M).$$

Let $\mathrm{vec}(M) = v$, $B = P_2^\top \otimes P_1$ and $C = P_4^\top \otimes P_3$, then $\mathrm{vec}(P_1MP_2) = Bv$ and $\mathrm{vec}(P_3MP_4) = Cv$. Thereby

$$U(P_1MP_2, P_3MP_4) = \mathrm{trace}(\mathrm{Cov}(Bv, Cv)) = \mathrm{trace}\left(B\Sigma C^\top\right) = \mathrm{trace}\left(C^\top B\Sigma\right),$$

and similarly $\mathrm{trace}(\Sigma) = V(M)$. For a positive semi-definite Σ, by its spectral decomposition, $\Sigma = \sum_{k=1}^{d^2} \lambda_k u_k u_k^\top$ where $\lambda_k \ge 0$. As a result,

$$|\,\mathrm{trace}\left(C^\top B\Sigma\right)| = |\sum_{k=1}^{d^2} \mathrm{trace}(\lambda_k C^\top B u_k u_k^\top)| = |\sum_{k=1}^{d^2} \lambda_k u_k^\top C^\top B u_k| \le \sum_{k=1}^{d^2} \lambda_k \|C^\top B\|_2 \|u_k\|_2^2 = \|C^\top B\|_2\,\mathrm{trace}(\Sigma),$$

Then the conclusion is implied by

$$\|C^\top B\|_2 \le \|C\|_2\|B\|_2 = \|P_1\|_2\|P_2\|_2\|P_3\|_2\|P_4\|_2,$$

due to the property of the Kronecker product: $\|P_2^\top \otimes P_1\|_2 = \|P_1\|_2\|P_2\|_2$ and $\|P_4^\top \otimes P_3\|_2 = \|P_3\|_2\|P_4\|_2$. When P_1, P_2, P_3, P_4 are all scalar matrices, the equality clearly holds.

□

The high-level idea to prove the multivariate convergence rate is to control the norm of a matrix by its quadratic form, which is stated in the following lemma.

Lemma 19. *Let $u_i \in \mathbb{R}^d$ be the unit vector with the i-th coordinate equal 1, and for $i \ne j$ let*

$$u_{ij+} = \frac{u_i + u_j}{\sqrt{2}},\ \ u_{ij-} = \frac{u_i - u_j}{\sqrt{2}}.$$

Then for any real symmetric $d \times d$ matrix M,

$$\|M\|_F^2 \le \sum_{i=1}^{d} (u_i^\mathsf{T} M u_i)^2 + \sum_{1\le i<j\le d} (u_{ij+}^\mathsf{T} M u_{ij+})^2 + \sum_{1\le i<j\le d} (u_{ij-}^\mathsf{T} M u_{ij-})^2.$$

Proof. Write $M = (M_{ij})_{1\le i,j\le d}$. It is clear that $M_{ii} = u_i^\mathsf{T} M u_i$. Further notice that

$$(u_{ij+}^\mathsf{T} M u_{ij+})^2 = \frac{1}{4}(M_{ii} + M_{jj} + 2M_{ij})^2,\ \ (u_{ij-}^\mathsf{T} M u_{ij-})^2 = \frac{1}{4}(M_{ii} + M_{jj} - 2M_{ij})^2,$$

we have

$$(u_{ij+}^\mathsf{T} M u_{ij+})^2 + (u_{ij-}^\mathsf{T} M u_{ij-})^2 = \frac{1}{2}(M_{ii} + M_{jj})^2 + 2(M_{ij})^2 \ge 2(M_{ij})^2,$$

and the conclusion follows. □

Now we generalize the results to the multivariate case and show that

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|_F^2 \lesssim n^{-1+\alpha}\log n, \text{ where } \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\left[X_iX_i^\mathsf{T} + X_i(X_{i-1}+\cdots+X_{t_i})^\mathsf{T} + (X_{i-1}+\cdots+X_{t_i})X_i^\mathsf{T}\right].$$

We focus on the Frobenius norm of the error, which is larger than the operator norm. First, consider the multivariate iteration $X_i = (1-\eta_i)X_{i-1} + \eta_i e_i$, i.e., $X_i \in \mathbb{R}^d$ follows the same form of iteration as in (24). For any $a \in \mathbb{R}^d$ with $\|a\| = 1$, define $Y_i = a^\mathsf{T}X_i$. Then

$$Y_i = a^\mathsf{T}X_i = (1-\eta_i)a^\mathsf{T}X_{i-1} + \eta_i a^\mathsf{T}e_i = (1-\eta_i)Y_{i-1} + \eta_i\widetilde{e}_i, \qquad Y_0 = 0 \tag{44}$$

where $\widetilde{e}_i = a^\mathsf{T}e_i$, $\mathrm{var}(\widetilde{e}_i) \le \lambda_{\max}(\Sigma_e)$ and $\mathbb{E}|\widetilde{e}_i|^4 \le \mathbb{E}|e_i|^4$.

Let $\widehat{\sigma}^2_{n,Y}, \sigma^2_{n,Y}, \sigma_Y$ denote, respectively, the covariance estimator (as in (30)), the true non-asymptotic variance (as in (29)) and the limiting variance (sandwich form) associated with the sequence $\{Y_i\}$. Similarly, define $\widehat{\Sigma}_n$, Σ_n, Σ as as the corresponding quantities in the multi-dimensional setting, based on the sequence $\{X_i\}$. Then

$$a^\mathsf{T}\widehat{\Sigma}_n a = \widehat{\sigma}^2_{n,Y}, a^\mathsf{T}\Sigma_n a = \sigma^2_{n,Y}, a^\mathsf{T}\Sigma a = \sigma_Y.$$

Since $\{Y_i\}$ is a one-dimensional SGD iterate of the same form as defined in (24), we can apply the one-dimensional bound:

$$\mathbb{E}\Big[(\widehat{\sigma}^2_{n,Y} - \sigma^2_{n,Y})^2\Big] \lesssim n^{-1+\alpha}\log n,$$

or

$$\mathbb{E}\Big[(\widehat{\sigma}^2_{n,Y} - \sigma^2_{Y})^2\Big] \lesssim n^{-1+\alpha}\log n.$$

In other words, for any $a \in \mathbb{R}^d$ with $\|a\| = 1$, since the noise term $\widetilde{e}_i$ the iteration (44) has uniformly bounded variances and 4-th moments, we have

$$\mathbb{E}\Big[a^\mathsf{T}(\widehat{\Sigma}_n - \Sigma_n)a\Big]^2 \lesssim n^{-1+\alpha}\log n. \tag{45}$$

By Lemma 19, we can choose a as u_i, u_{ij+} and u_{ij-} for $1 \le i < j \le d$, and the Frobenius norm can be bounded by a finite sum of such quadratic forms. As a result, (45) directly implies

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma_n\|_F^2 \lesssim n^{-1+\alpha}\log n.$$

Similarly we have

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma\|_F^2 \lesssim n^{-1+\alpha}\log n.$$

Finally, we prove the convergence rate for the iteration with an arbitrary positive-definite matrix A, i.e.,

$$X_i = (\mathbf{I}_d - \eta_i A)X_{i-1} + \eta_i e_i, \qquad X_0 = 0.$$

The high-level idea is to apply Lemma 18 to get an upper bound of $V(\widehat{\Sigma}_n)$ as a linear combination of $V(e_ie_j^\top)$, and show that this upper bound reduces to the case that has been discussed and solved. We decompose the estimation error into the bias and variance parts,

$$\mathbb{E}\|\widehat{\Sigma}_n - \Sigma_n\|_F^2 = \|\mathbb{E}\widehat{\Sigma}_n - \Sigma_n\|_F^2 + \mathbb{E}\|\mathbb{E}\widehat{\Sigma}_n - \widehat{\Sigma}_n\|_F^2.$$

The first term, i.e., the bias part, can be bounded following the same argument in the proof of Proposition 2 (Section C.2). For the variance part, we define another SGD iteration in $\mathbb{R}^d$,

$$\widetilde{X}_i = (1-\eta_i\lambda)\widetilde{X}_{i-1} + \eta_i e_i, \qquad \widetilde{X}_0 = 0,$$

where λ is the smallest eigenvalue of A. We denote the corresponding covariance estimator as $\widetilde{\Sigma}_n$. Notice that we have shown the convergence rate for $\widetilde{\Sigma}_n$, i.e., we have $\mathbb{E}\|\mathbb{E}\widetilde{\Sigma}_n - \widetilde{\Sigma}_n\|_F^2 \lesssim n^{-1+\alpha}\log n$. Then it suffices to show that

$$\mathbb{E}\|\mathbb{E}\widehat{\Sigma}_n - \widehat{\Sigma}_n\|_F^2 \le \mathbb{E}\|\mathbb{E}\widetilde{\Sigma}_n - \widetilde{\Sigma}_n\|_F^2, \tag{46}$$

i.e., $V(\widehat{\Sigma}_n) \le V(\widetilde{\Sigma}_n)$. Recall that $X_i = \sum_{k=1}^{i} Y_{(A)k}^{i}\eta_k e_k$, similar as the proof of 16, we express $\widehat{\Sigma}_n$ as $n + \binom{n}{2}$ uncorrelated summation:

$$\widehat{\Sigma}_n = \sum_{i=1}^{n}\sum_{j=1}^{n} h_{i,j}(e_i e_j^{\mathsf{T}}) = \sum_{i=1}^{n} h_{i,i}(e_i e_i^{\mathsf{T}}) + \sum_{1\le i<j\le n} h_{i,j}(e_i e_j^{\mathsf{T}}),$$

where $h_{i,j}$ are linear functions of $d\times d$ matrices such that for some finite index sets $\{S_{i,j} \in \{1,2,\dots,n\}\times\{1,2,\dots,n\}, 1\le i\le j\le n\}$,

$$h_{i,j}(e_i e_j^{\mathsf{T}}) = \eta_i\eta_j \sum_{t,k\in\mathcal{S}_{i,j}} Y_{(A)i}^{t} e_i e_j^{\mathsf{T}} Y_{(A)j}^{k}.$$

Here we define

$$Y_{(A)j}^{i} = \begin{cases} \mathbf{I}_d & \text{if } j = i, \\ \prod_{k=j+1}^{i} (\mathbf{I}_d - \eta_k A) & \text{if } j < i. \end{cases} \tag{47}$$

Leveraging Lemma 18, we have an upper bound of $V(\widehat{\Sigma}_n)$ as

$$V(\widehat{\Sigma}_n) \le \sum_{i=1}^{n} \eta_i^2 V(e_i e_i^{\mathsf{T}})\Big(\sum_{t,k\in\mathcal{S}_{i,i}} Y_{(\lambda)i}^{t} Y_{(\lambda)i}^{k}\Big)^2 + \sum_{1\le i<j\le n} \eta_i\eta_j V(e_i e_j^{\mathsf{T}})\Big(\sum_{t,k\in\mathcal{S}_{i,j}} Y_{(\lambda)i}^{t} Y_{(\lambda)j}^{k}\Big)^2.$$

The key observation is that the structures of $\widetilde{\Sigma}_n$ and $\widehat{\Sigma}_n$ are exactly the same, except that the matrix A should be replaced by $\lambda\mathbf{I}_d$. Following an identical argument,

$$V(\widetilde{\Sigma}_n) = \sum_{i=1}^{n} \eta_i^2 V(e_i e_i^{\mathsf{T}})\Big(\sum_{t,k\in\mathcal{S}_{i,i}} Y_{(\lambda)i}^{t} Y_{(\lambda)i}^{k}\Big)^2 + \sum_{1\le i<j\le n} \eta_i\eta_j V(e_i e_j^{\mathsf{T}})\Big(\sum_{t,k\in\mathcal{S}_{i,j}} Y_{(\lambda)i}^{t} Y_{(\lambda)j}^{k}\Big)^2$$

since by Lemma 18, the equality holds for scalar multiplier matrices. So we have proved that $V(\widehat{\Sigma}_n) \le V(\widetilde{\Sigma}_n) \lesssim n^{-1+\alpha}\log n$, and the proof for multivariate case is completed.

In the remainder of the proof, we will use the one-dimensional results established in this section, while all results remain valid in the multi-dimensional setting.

D PROOF OF THEOREM 7

As mentioned in the proof sketch, to bound the estimation error, we decompose it as follows:

$$\|\widehat{\Sigma}_n - \Sigma\| \le \|\widehat{\Sigma}_n - \widehat{\Sigma}_{n,\delta}\| + \|\widehat{\Sigma}_{n,\delta} - \widehat{\Sigma}_{n,U}\| + \|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\| + \|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\|. \tag{48}$$

D.1 With i.i.d Approximation: the Fourth Term

Recall the definition of *i.i.d* approximation sequence,

$$\widetilde{U}_i = (1-\eta_i A)\widetilde{U}_{i-1} + \eta_i\epsilon_i^*; \quad \widetilde{U}_0 = \delta_0,$$

where $\epsilon_i^* = -\nabla f(x^*,\xi_i)$. The corresponding estimator based on this sequence is given by

$$\widehat{\Sigma}_{n,\widetilde{U}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{U}_i^2 + 2\widetilde{U}_i\Big(\sum_{k=t_i}^{i-1}\widetilde{U}_k\Big)\right]$$

We consider the starting point δ_0 to be deterministic. The next proposition bound the fourth term in (48).

Proposition 20. *Assume* $\mathbb{E}[\nabla f(x^*,\xi_i)^4] < \infty$, *then*

$$\|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\|_2 = \sqrt{\mathbb{E}\left|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\right|^2} \lesssim \sqrt{n^{-1+\alpha}\log n}. \tag{49}$$

Proof. Define a sequence with starting point 0,

$$X_i^\circ = (1-\eta_i A)X_{i-1}^\circ + \eta_i \epsilon_i^*, \quad i \geq 1; \qquad X_0^\circ = 0, \tag{50}$$

and the corresponding estimator

$$\widehat{\Sigma}_n^\circ = \frac{1}{n}\sum_{i=1}^n \left[|X_i^\circ|^2 + 2X_i^\circ(X_{i-1}^\circ + \cdots + X_{t_i}^\circ)\right]. \tag{51}$$

Decompose

$$\|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\|_2 \leq \|\widehat{\Sigma}_{n,\widetilde{U}} - \widehat{\Sigma}_n^\circ\|_2 + \|\widehat{\Sigma}_n^\circ - \Sigma_n^\circ\|_2 + \|\Sigma_n^\circ - \Sigma_{n,\widetilde{U}}\|_2 + \|\Sigma_{n,\widetilde{U}} - \Sigma\|_2. \tag{52}$$

The bound for the second term is given by Corollary 17:

$$\|\widehat{\Sigma}_n^\circ - \Sigma_n^\circ\|_2 \lesssim \sqrt{n^{-1+\alpha}\log n}.$$

By Lemma 8 (i) and Lemma 12, we have $\max\{\left\|\widetilde{U}_i\right\|_2, \|X_i^\circ\|_2\} \leq i^{-\alpha/2}$, and

$$\zeta_i := \widetilde{U}_i - X_i^\circ = Y_{(A)}{}_0^i\delta_0 \lesssim \exp\left(-\eta A i^{1-\alpha}\right),$$

which implies $|\zeta_i| \lesssim i^{-K}$ for any $K > 0$. Therefore for the third term, we have

$$\begin{aligned}
|\Sigma_n^\circ - \Sigma_n| &= \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\widetilde{U}_i^2 + 2\widetilde{U}_i\Big(\sum_{j=1}^{i-1} X_j\Big)\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|X_i^\circ|^2 + 2X_i^\circ\Big(\sum_{j=1}^{i-1} X_j^\circ\Big)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\widetilde{U}_i\zeta_i + 2\widetilde{U}_i\Big(\sum_{j=1}^{i-1}\zeta_j\Big)\right]\right| + \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\zeta_i X_i^\circ + 2\zeta_i\Big(\sum_{j=1}^{i-1} X_j^\circ\Big)\right]\right| \\
&\lesssim \frac{1}{n}\sum_{i=1}^n \left\|\widetilde{U}_i\right\|_1 \sum_{j=1}^i |\zeta_j| + \frac{1}{n}\sum_{i=1}^n |\zeta_i| \sum_{j=1}^i \|X_j^\circ\|_1 \\
&\lesssim 2\cdot\left(\frac{1}{n}\sum_{i=1}^n i^{-\alpha/2}\right)\left(\sum_{j=1}^n j^{-K}\right) \lesssim n^{-K_1},
\end{aligned}$$

for any $K_1 > 0$. Similarly, for the first term,

$$\begin{aligned}
\|\widehat{\Sigma}_{n,\widetilde{U}} - \widehat{\Sigma}_n^\circ\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^n \left[\widetilde{U}_i^2 + 2\widetilde{U}_i\Big(\sum_{j=t_i}^{i-1} X_j\Big)\right] - \frac{1}{n}\sum_{i=1}^n \left[|X_i^\circ|^2 + 2X_i^\circ\Big(\sum_{j=t_i}^{i-1} X_j^\circ\Big)\right]\right\|_2 \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n \left[\widetilde{U}_i\zeta_i + 2\widetilde{U}_i\Big(\sum_{j=t_i}^{i-1}\zeta_j\Big)\right]\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^n \left[\zeta_i X_i^\circ + 2\zeta_i\Big(\sum_{j=t_i}^{i-1} X_j^\circ\Big)\right]\right\|_2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^n \left\|\widetilde{U}_i\right\|_2 \sum_{j=t_i}^i |\zeta_j| + \frac{1}{n}\sum_{i=1}^n |\zeta_i| \sum_{j=t_i}^i \|X_j^\circ\|_2 \lesssim n^{-K_1}.
\end{aligned}$$

For the fourth term, recall that ϵ_i^*'s are i.i.d. with $\mathbb{E}[\epsilon_i^*] = 0$, $\mathbb{E}[\epsilon_i^{*2}] = S$, then

$$\begin{aligned}
\Sigma_{n,\widetilde{U}} &= \frac{1}{n}\mathbb{E}\left[\Big(\sum_{i=1}^n \widetilde{U}_i\Big)^2\right] = \frac{1}{n}\mathbb{E}\left[\Big(S_{(A)}{}_0^n\delta_0 + \sum_{i=1}^n (1 + S_{(A)}{}_i^n)\eta_i\epsilon_i^*\Big)^2\right] \\
&= \frac{1}{n}\Big((S_{(A)}{}_0^n\delta_0)^2 + \sum_{i=1}^n (1 + S_{(A)}{}_i^n)^2\eta_i^2 S\Big).
\end{aligned}$$

Therefore

$$\begin{aligned}|\Sigma_{n,\widetilde{U}}-\Sigma| &\le \frac{1}{n}(S_{(A)_0}^{\ n}\delta_0)^2+\frac{1}{n}\sum_{i=1}^{n}\left[(1+S_{(A)_i}^{\ n})^2\eta_i^2-A^{-2}\right]S\\ &\lesssim \frac{1}{n}\cdot\mathcal{O}(1)+\frac{1}{n}\sum_{i=1}^{n}\left[(1+S_{(A)_i}^{\ n})\eta_i-A^{-1}\right]\cdot\mathcal{O}(1)\\ &\lesssim \frac{1}{n}+\frac{1}{n}\sum_{i=1}^{n}(S_{(A)_i}^{\ n}\eta_i-A^{-1}).\end{aligned}$$

To bound $(S_{(A)_i}^{\ n}\eta_i-A^{-1})$, by Lemma 8 (i),

$$S_{(A)_i}^{\ n}=\sum_{k=i+1}^{n}Y_{(A)_i}^{\ k}\asymp\sum_{k=i+1}^{n}\exp\left\{\frac{\eta A}{1-\alpha}\left(i^{1-\alpha}-k^{1-\alpha}\right)\right\},$$

and then

$$\begin{aligned}\sum_{k=i+1}^{n}\exp\left(-\frac{\eta A}{1-\alpha}k^{1-\alpha}\right)&\le\int_i^{\infty}\exp\left(-\frac{\eta A}{1-\alpha}u^{1-\alpha}\right)\mathrm{d}u\\ &=\frac{1}{\eta A}\left(\frac{1-\alpha}{\eta A}\right)^{\frac{\alpha}{1-\alpha}}\int_{\frac{\eta A}{1-\alpha}i^{1-\alpha}}^{\infty}\mathrm{e}^{-t}t^{\frac{\alpha}{1-\alpha}}\,\mathrm{d}t.\end{aligned}$$

Using integration by parts, for any fixed $a\ge 1$, $\beta>1$,

$$\begin{aligned}\int_a^{\infty}\mathrm{e}^{-x}x^{\beta}\,\mathrm{d}x&=\mathrm{e}^{-a}a^{\beta}+\beta\int_a^{\infty}\mathrm{e}^{-x}x^{\beta-1}\,\mathrm{d}x\\ &=\mathrm{e}^{-a}a^{\beta}\left(1+\beta a^{-1}\right)+\beta(\beta-1)\int_a^{\infty}\mathrm{e}^{-x}x^{\beta-2}\,\mathrm{d}x=\cdots\\ &=\mathrm{e}^{-a}a^{\beta}\left(1+\beta a^{-1}+\cdots\widetilde{\beta}!a^{-\lfloor\beta\rfloor}\right)+\beta!\int_a^{\infty}\mathrm{e}^{-x}x^{\beta-\lfloor\beta\rfloor-1}\,\mathrm{d}x\\ &\le\mathrm{e}^{-a}a^{\beta}\left(1+\beta a^{-1}+\cdots\widetilde{\beta}!a^{-\lfloor\beta\rfloor}+\beta!a^{-\beta}\right),\end{aligned}$$

where $\beta!=\beta(\beta-1)\cdots(\beta-\lfloor\beta\rfloor)$, $\widetilde{\beta}!=\beta!/(\beta-\lfloor\beta\rfloor)$.

Therefore, for i large enough,

$$\begin{aligned}&\int_{\frac{\eta A}{1-\alpha}i^{1-\alpha}}^{\infty}\mathrm{e}^{-t}t^{\frac{\alpha}{1-\alpha}}\,\mathrm{d}t\\ =\ &\exp\left(-\frac{\eta A}{1-\alpha}i^{1-\alpha}\right)\left(\frac{\eta A}{1-\alpha}i^{1-\alpha}\right)^{\frac{\alpha}{1-\alpha}}\left(1+(\eta A)^{-1}i^{-1+\alpha}+o(i^{-1+\alpha})\right),\end{aligned}$$

which implies

$$\begin{aligned}S_{(A)_i}^{\ n}&\lesssim\exp\left(\frac{\eta A}{1-\alpha}i^{1-\alpha}\right)\sum_{k=i+1}^{n}\exp\left(-\frac{\eta A}{1-\alpha}k^{1-\alpha}\right)\\ &=(\eta A)^{-1}i^{\alpha}\left(1+(\eta A)^{-1}i^{-1+\alpha}+o(i^{-1+\alpha})\right)\end{aligned}$$

and hence

$$\begin{aligned}S_{(A)_i}^{\ n}\eta_i-A^{-1}&\lesssim A^{-1}\left(1+(\eta A)^{-1}i^{-1+\alpha}+o(i^{-1+\alpha})\right)-A^{-1}\\ &\lesssim i^{-1+\alpha}.\end{aligned}$$

Finally, it gives

$$\begin{aligned}|\Sigma_{n,\widetilde{U}} - \Sigma| &\lesssim \frac{1}{n} + \frac{1}{n}\sum_{i=1}^{n}(S_{(A)}{}_i^n \eta_i - A^{-1}) \\ &\lesssim \frac{1}{n} + \frac{1}{n}\sum_{i=1}^{n} i^{-1+\alpha} \asymp n^{-1+\alpha}.\end{aligned}$$

Hence, putting together,

$$\begin{aligned}\|\widehat{\Sigma}_{n,\widetilde{U}} - \Sigma\|_2 &\leq \|\widehat{\Sigma}_{n,\widetilde{U}} - \widehat{\Sigma}_n^{\circ}\|_2 + \|\widehat{\Sigma}_n^{\circ} - \Sigma_n^{\circ}\|_2 + \|\Sigma_n^{\circ} - \Sigma_{n,\widetilde{U}}\|_2 + \|\Sigma_{n,\widetilde{U}} - \Sigma\|_2 \\ &\lesssim n^{-K_1} + \sqrt{n^{-1+\alpha}\log n} + n^{-K_1} + n^{-1+\alpha} \\ &\lesssim \sqrt{n^{-1+\alpha}\log n}. \qquad (\alpha > 1/2)\end{aligned}$$

□

D.2 With Linear Approximation: the Third Term

Recall the definition of the linear approximation sequence,

$$U_i = (1 - \eta_i A)U_{i-1} + \eta_i \epsilon_i; \quad U_0 = \delta_0,$$

where $\epsilon_i = \nabla F(X_{i-1}) - \nabla f(X_{i-1}, \xi_i)$. We consider the starting point δ_0 to be deterministic. The corresponding estimator based on this sequence is given by

$$\widehat{\Sigma}_{n,U} = \frac{1}{n}\sum_{i=1}^{n}\left[U_i^2 + 2U_i\Big(\sum_{k=t_i}^{i-1} U_k\Big)\right]$$

The next proposition bound the third term in (48).

Proposition 21 (linear case)**.** *Suppose Assumptions 4-6 hold, then we have*

$$\|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\|_2 = \sqrt{\mathbb{E}\left|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\right|^2} \lesssim \sqrt{n^{-1+\alpha}\log n}. \tag{53}$$

Moreover, combining this with Proposition 20, we conclude that the intermediate estimator based on the linear sequence is consistent, with

$$\|\widehat{\Sigma}_{n,U} - \Sigma\|_2 = \sqrt{\mathbb{E}\left|\widehat{\Sigma}_{n,U} - \Sigma\right|^2} \lesssim \sqrt{n^{-1+\alpha}\log n}.$$

Proof. By Lemma 24 (i), Lemma 25,

$$
\begin{aligned}
&\|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\|_2 \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n}\left[U_i^2 + 2U_i\Big(\sum_{j=t_i}^{i-1} U_j\Big)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{U}_i^2 + 2\widetilde{U}_i\Big(\sum_{j=t_i}^{i-1} \widetilde{U}_j\Big)\right] \right\|_2 \\
&\le \left\| \frac{1}{n}\sum_{i=1}^{n}\left[U_i\Delta_i + 2U_i\Big(\sum_{j=t_i}^{i-1} \Delta_j\Big)\right] \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\widetilde{U}_i + 2\Delta_i\Big(\sum_{j=t_i}^{i-1} \widetilde{U}_j\Big)\right] \right\|_2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[\|U_i\Delta_i\|_2 + \left\| U_i\Big(\sum_{j=t_i}^{i} \Delta_j\Big)\right\|_2 + \|\Delta_i\widetilde{U}_i\|_2 + \left\| \Delta_i\Big(\sum_{j=t_i}^{i} \widetilde{U}_j\Big)\right\|_2 \right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left[\|U_i\|_4\,\|\Delta_i\|_4 + \|U_i\|_4\left\|\sum_{j=t_i}^{i}\Delta_j\right\|_4 + \|\Delta_i\|_4\|\widetilde{U}_i\|_4 + \|\Delta_i\|_4\left\|\sum_{j=t_i}^{i}\widetilde{U}_j\right\|_4 \right] \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[i^{-\alpha/2}\cdot i^{-\alpha} + i^{-\alpha/2}\sqrt{\log i} + i^{-\alpha}\cdot i^{-\alpha/2} + i^{-\alpha}\sqrt{i^{\alpha}\log i}\right] \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n} i^{-\alpha/2}\sqrt{\log i} \lesssim \sqrt{n^{-\alpha}\log n}.
\end{aligned}
$$

Since $\alpha > 1/2$, we finally have

$$\|\widehat{\Sigma}_{n,U} - \widehat{\Sigma}_{n,\widetilde{U}}\|_2 \lesssim \sqrt{n^{-\alpha}\log n} \lesssim \sqrt{n^{-1+\alpha}\log n}.$$

□

D.3 General Case: the First and Second Term

Recall the definition of covariance estimators with and without the sample average,

$$
\begin{aligned}
\widehat{\Sigma}_n &= \frac{1}{n}\sum_{i=1}^{n}\left[2(\delta_i - \bar{\delta}_n)\Big(\sum_{k=t_i}^{i}\delta_k - \ell_i\bar{\delta}_n\Big) - (\delta_i - \bar{\delta}_n)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[2\delta_i\Big(\sum_{k=t_i}^{i}\delta_k\Big) - \delta_i^2\right] + \frac{1}{n}\sum_{i=1}^{n}(2\ell_i - 1)\bar{\delta}_n^2 - \frac{2}{n}\sum_{i=1}^{n}\Big(\sum_{k=t_i}^{i-1}\delta_k + \ell_i\delta_i\Big)\bar{\delta}_n. \\
\widehat{\Sigma}_{n,\delta} &= \frac{1}{n}\sum_{i=1}^{n}\left[2\delta_i\Big(\sum_{k=t_i}^{i}\delta_k\Big) - \delta_i^2\right].
\end{aligned}
$$

We now complete the proof by bounding the first and second terms in (48), i.e.,

$$\|\widehat{\Sigma}_n - \widehat{\Sigma}_{n,\delta}\|_2, \text{ and } \|\widehat{\Sigma}_{n,\delta} - \widehat{\Sigma}_{n,U}\|_2.$$

Complete proof of the Theorem 7. For the first term, by Lemma 27 (ii, iii), we have

$$\begin{aligned}
\|\widehat{\Sigma}_n - \widehat{\Sigma}_{n,\delta}\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^{n}(2\ell_i - 1)\bar{\delta}_n^2 - \frac{2}{n}\sum_{i=1}^{n}\Big(\sum_{k=t_i}^{i-1}\delta_k + \ell_i\delta_i\Big)\bar{\delta}_n \right\|_2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\ell_i\|\bar{\delta}_n^2\|_2 + \frac{1}{n}\left\|\sum_{i=1}^{n}\Big(\sum_{k=t_i}^{i-1}\delta_k + \ell_i\delta_i\Big)\bar{\delta}_n\right\|_2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\ell_i\|\bar{\delta}_n\|_4^2 + \frac{1}{n}\left\|\sum_{i=1}^{n}\Big(\sum_{k=t_i}^{i-1}\delta_k + \ell_i\delta_i\Big)\right\|_4\|\bar{\delta}_n\|_4 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\ell_i\|\bar{\delta}_n\|_4^2 + \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{k=t_i}^{i}\delta_k\right\|_4\|\bar{\delta}_n\|_4 + \frac{1}{n}\left\|\sum_{i=1}^{n}\ell_i\delta_i\right\|_4\|\bar{\delta}_n\|_4 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\ell_i\|\bar{\delta}_n\|_4^2 + \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{k=t_i}^{i}\delta_k\right\|_4\|\bar{\delta}_n\|_4 + \ell_n\|\bar{\delta}_n\|_4^2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}i^{\alpha}\log i\cdot n^{-1} + \frac{1}{n}\sum_{i=1}^{n}\sqrt{i^{\alpha}\log i}\cdot n^{-1/2} + n^{\alpha}\log n\cdot n^{-1} \\
&\asymp n^{-1+\alpha}\log n + \sqrt{n^{-1+\alpha}\log n} + n^{-1+\alpha}\log n \\
&\asymp \sqrt{n^{-1+\alpha}\log n}.
\end{aligned}$$

For the second term, by Lemma 23, 25 and 26(ii),

$$\begin{aligned}
&\|\widehat{\Sigma}_{n,\delta} - \widehat{\Sigma}_{n,U}\|_2 \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i^2 + 2\delta_i\Big(\sum_{j=t_i}^{i-1}\delta_j\Big)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[U_i^2 + 2U_i\Big(\sum_{j=t_i}^{i-1}U_j\Big)\right]\right\|_2 \\
&\le \left\| \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i s_i + 2\delta_i\Big(\sum_{j=t_i}^{i-1}s_j\Big)\right]\right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n}\left[s_i U_i + 2s_i\Big(\sum_{j=t_i}^{i-1}U_j\Big)\right]\right\|_2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[\|\delta_i s_i\|_2 + \left\|\delta_i\Big(\sum_{j=t_i}^{i-1}s_j\Big)\right\|_2 + \|s_i U_i\|_2 + \left\|s_i\Big(\sum_{j=t_i}^{i-1}U_j\Big)\right\|_2\right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left[\|\delta_i\|_4\|s_i\|_4 + \|\delta_i\|_4\left\|\sum_{j=t_i}^{i-1}s_j\right\|_4 + \|s_i\|_4\|U_i\|_4 + \|s_i\|_4\left\|\sum_{j=t_i}^{i-1}U_j\right\|_4\right] \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[\|\delta_i\|_4\sum_{j=t_i}^{i}\|s_j\|_4 + \|s_i\|_4\sum_{j=t_i}^{i}\|U_j\|_4\right] \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\left[i^{-\alpha/2}\ell_i i^{-\alpha} + i^{-\alpha}\ell_i i^{-\alpha/2}\right] \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}i^{-\alpha/2}\log i \lesssim n^{-\alpha/2}\log n.
\end{aligned}$$

Since $\alpha > 1/2$,

$$n^{-\alpha/2}\log n \lesssim \sqrt{n^{-1+\alpha}\log n}.$$

Now combined with bound on the third and the fourth term, as shown in Propositions 21 and 20, we finally have

$$\|\widehat{\Sigma}_n - \Sigma\|_2 \lesssim \sqrt{n^{-1+\alpha}\log n}.$$

□

D.4 Technical Lemmas

Lemma 22 (Lipschitz continuity of ∇F)**.** *Under Assumptions 4-6, we have*

$$\|\nabla F(X_1) - \nabla F(X_2)\| \le L|X_1 - X_2|.$$

Proof. By the stochastic Lipschitz continuity,

$$\mathbb{E}\|\nabla f(X_1,\xi) - \nabla f(X_2,\xi)\|_2^2 \le L^2|X_1 - X_2|^2, \quad \text{for all } \ X_1, X_2 \in \mathbb{R}^d.$$

By the convexity of $|\cdot|^2$ and Jensen's inequality,

$$\begin{aligned}\|\nabla F(X_1) - \nabla F(X_2)\|^2 &= |\mathbb{E}_\xi(\nabla f(X_1,\xi) - \nabla f(X_2,\xi))|^2 \\ &\le \mathbb{E}_\xi|\nabla f(X_1,\xi) - \nabla f(X_2,\xi)|^2 \\ &\le L^2|X_1 - X_2|^2.\end{aligned}$$

□

Lemma 23 (moment bounds for the error sequence)**.** *Under Assumptions 4-6, for $1 \le q \le 8$, the error sequence $\delta_n = X_n - x^*$ satisfies $\|\delta_n\|_q \le n^{-\alpha/2}$.*

Proof. The error sequence can be written as,

$$\begin{aligned}\delta_i &= \delta_{i-1} - \eta_i \nabla f(X_{i-1},\xi_i) \\ &= \delta_{i-1} - \eta_i \nabla F(X_{i-1}) + \eta_i \epsilon_i.\end{aligned}$$

By Rio's inequality (Rio, 2009), since $\mathbb{E}[\epsilon_i|\delta_{i-1} - \eta_i\nabla F(X_{i-1})] = 0$, we have

$$\|\delta_i\|_q^2 \le \|\delta_{i-1} - \eta_i \nabla F(X_{i-1})\|_q^2 + (q-1)\eta_i^2\|\epsilon_i\|_q^2.$$

Further by strong convexity and stochastic Lipschitz continuity, we have

$$\begin{aligned}\|\delta_i\|_q^2 &\le \|\delta_{i-1} - \eta_i \nabla F(X_{i-1})\|_q^2 + (q-1)\eta_i^2\|\epsilon_i\|_q^2 \\ &\le (1 - \eta_i c_1)\|\delta_{i-1}\|_q^2 + 2(q-1)\eta_i^2(\|\epsilon_i^*\|_q^2 + L_\epsilon^2\|\delta_{i-1}\|_q^2) \\ &\le (1 - \eta_i c_2)\|\delta_{i-1}\|_q^2 + 2(q-1)\eta_i^2\|\epsilon_i^*\|_q^2,\end{aligned}$$

where c_1 and c_2 are positive constants depending only on L_ϵ, q and η. Finally, by Lemma 8,

$$\begin{aligned}\|\delta_i\|_q^2 &\le \prod_{k=1}^{i}(1 - \eta_k c_2)\delta_0^2 + 2(q-1)\|\epsilon_i^*\|_q^2 \sum_{j=1}^{i}\eta_j^2 \prod_{k=j+1}^{i}(1 - \eta_k c_2) \\ &\asymp {Y_{(c_2)}}_0^{\,i} + \sum_{j=1}^{i} {Y_{(c_2)}}_j^{\,i} j^{-2\alpha} \asymp i^{-\alpha}.\end{aligned}$$

□

Lemma 24 (moment bounds for increments)**.** *Recall that*

- $\epsilon_k^* = -\nabla f(x^*,\xi_k)$ *is i.i.d.*
- $v_k = \epsilon_k - \epsilon_k^*$ *is a martingale differences sequence.*

(i) For all $i \in \mathbb{N}^+$ and $m = 1, 2, 4$, we have

$$\|\epsilon_i\|_m = \mathcal{O}(1), \quad \|\epsilon_i^*\|_m = \mathcal{O}(1), \quad \|v_i\|_m \lesssim i^{-\alpha/2}. \tag{54}$$

(ii) For all $i \neq j$, we have $\mathbb{E}[\epsilon_i \epsilon_j^*] = \mathbb{E}[\epsilon_i v_j] = \mathbb{E}[\epsilon_i^* v_j] = 0$.

Proof.

(i) Due to Lyapunov inequality, it suffices to consider $m = 4$.
The bound for ϵ_i is obtained by stochastic Lipschitz continuity:

$$(\mathbb{E}_{n-1}|\epsilon_n|^4)^{1/4} \leq (\mathbb{E}_{n-1}|v_n|^4)^{1/4} + (\mathbb{E}_{n-1}|\epsilon_n^*|^4)^{1/4} \lesssim \delta_{n-1} + 1,$$

and $\|\delta_{n-1}\|_4 \lesssim (n-1)^{-\alpha/2}$. The bound for v_i is obtained by Lipschitz continuity,

$$\begin{aligned}\|v_i\|_4 = \|\epsilon_i - \epsilon_i^*\|_4 &= \|\nabla F(X_{i-1}) - \nabla F(x^*) - (\nabla f(X_{i-1}, \xi_i) - \nabla f(x^*, \xi_i))\|_4 \\ &\leq \|\nabla F(X_{i-1}) - \nabla F(x^*)\|_4 + \|\nabla f(X_{i-1}, \xi_i) - \nabla f(x^*, \xi_i)\|_4 \\ &\lesssim \|X_{i-1} - x^*\|_4 = \|\delta_{i-1}\|_4 \lesssim i^{-\alpha/2}.\end{aligned}$$

Then a bound for ϵ_i^* is derived by $\|\epsilon_i^*\|_4 = \|\epsilon_i - v_i\|_4 \leq \|\epsilon_i\|_4 + \|v_i\|_4 \lesssim \mathcal{O}(1)$.

(ii) Recall that ξ_i's are i.i.d., ϵ_i's are martingale difference, and ϵ_i^*'s are i.i.d., then we have

$$\mathbb{E}[\epsilon_i \epsilon_j^*] = \begin{cases} \mathbb{E}\left[\mathbb{E}_{i-1}[\epsilon_i \epsilon_j^*]\right] = \mathbb{E}\left[\mathbb{E}_{i-1}[\epsilon_i]\epsilon_j^*\right] = 0 & \text{if } i > j, \\ \mathbb{E}\left[\mathbb{E}_i[\epsilon_i \epsilon_j^*]\right] = \mathbb{E}\left[\epsilon_i \mathbb{E}_i[\epsilon_j^*]\right] = 0 & \text{if } i < j. \end{cases}$$

Hence $\mathbb{E}[\epsilon_i v_j] = \mathbb{E}[\epsilon_i(\epsilon_j - \epsilon_j^*)] = 0$, $\mathbb{E}[\epsilon_i^* v_j] = \mathbb{E}[\epsilon_i^*(\epsilon_j - \epsilon_j^*)] = 0$

□

- Recall that

$$\begin{aligned} U_n &= (1 - \eta_n A) U_{n-1} + \eta_n \epsilon_n \\ \widetilde{U}_n &= (1 - \eta_n A) \widetilde{U}_{n-1} + \eta_n \epsilon_n^* \\ \Delta_n &= (1 - \eta_n A) \Delta_{n-1} + \eta_n v_n \end{aligned}$$

Lemma 25 (moment bounds for iterates)**.**

(i) For all $n \in \mathbb{N}^+$ and $m = 1, 2, 4$, we have

$$\|U_n\|_m \lesssim n^{-\alpha/2}, \quad \|\widetilde{U}_n\|_m \lesssim n^{-\alpha/2}, \quad \|\Delta_n\|_m \lesssim n^{-\alpha}. \tag{55}$$

(ii) For all $i \in \mathbb{N}^+$ and $m = 1, 2, 4$, we have

$$\left\| \sum_{j=t_i}^{i} U_j \right\|_m \lesssim t_i^{\alpha/2} + \sqrt{\ell_i} \asymp \sqrt{i^\alpha \log i}, \tag{56}$$

$$\left\| \sum_{j=t_i}^{i} \widetilde{U}_j \right\|_m \lesssim \sqrt{\ell_i} \asymp \sqrt{i^\alpha \log i}, \tag{57}$$

$$\left\| \sum_{j=t_i}^{i} \Delta_j \right\|_m \lesssim \mathcal{O}(1) + \sqrt{i^{-\alpha} \ell_i} = \sqrt{\log i}. \tag{58}$$

Proof. Due to Lyapunov inequality, it suffices to consider $m = 4$.

(i) For simplicity, write $Y_i^j := Y_{(A)i}^{\ j}$. Notice that U_n, $\widetilde{U}_n$, Δ_n can be written as

$$U_n = Y_0^n \delta_0 + \sum_{i=1}^n Y_i^n \eta_i \epsilon_i, \quad \widetilde{U}_n = Y_0^n \delta_0 + \sum_{i=1}^n Y_i^n \eta_i \epsilon_i^*, \quad \Delta_n = \sum_{i=1}^n Y_i^n \eta_i v_i.$$

Since ϵ_i's are martingale difference, by Lemma 8 (i, ii), Lemma 11 (Burkholder inequality), Lemma 24 (i), we have

$$\begin{aligned}
\|U_n\|_4 &\le Y_0^n \delta_0 + \left\| \sum_{i=1}^n Y_i^n \eta_i \epsilon_i \right\|_4 \\
&\lesssim Y_0^n \delta_0 + \sqrt{\sum_{i=1}^n \|Y_i^n \eta_i \epsilon_i\|_4^2} \\
&\lesssim Y_0^n + \sqrt{\sum_{i=1}^n |Y_i^n|^2 \left|i^{-\alpha}\right|^2} \\
&\lesssim \exp\left(-\eta A i^{1-\alpha}\right) + \sqrt{n^{-\alpha}} \\
&\asymp n^{-\alpha/2}.
\end{aligned}$$

$\widetilde{U}_n$ has the same bound as U_n since both $\|\epsilon_i\|_4, \|\epsilon_i^*\|_4 = \mathcal{O}(1)$. For Δ_n, similarly,

$$\begin{aligned}
\|\Delta_n\|_4 &= \left\| \sum_{i=1}^n Y_i^n \eta_i v_i \right\|_4 \\
&\lesssim \sqrt{\sum_{i=1}^n \|Y_i^n \eta_i v_i\|_4^2} \\
&\lesssim \sqrt{\sum_{i=1}^n |Y_i^n|^2 \left|i^{-\alpha}\right|^2 \cdot i^{-\alpha}} \\
&\lesssim \sqrt{n^{-2\alpha}} \asymp n^{-\alpha}.
\end{aligned}$$

(ii) For simplicity, write $S_i^j := S_{(A)i}^{\ j}$. Notice that the following sums can be written as

$$\begin{aligned}
\sum_{j=t_i}^i U_j &= \sum_{j=t_i}^i \left(Y_{t_i-1}^j U_{t_i-1} + \sum_{p=t_i}^j Y_p^j \eta_p \epsilon_p \right) = S_{t_i-1}^i U_{t_i-1} + \sum_{p=t_i}^i (1 + S_p^i) \eta_p \epsilon_p, \\
\sum_{j=t_i}^i \widetilde{U}_j &= \sum_{j=t_i}^i \left(Y_{t_i-1}^j \widetilde{U}_{t_i-1} + \sum_{p=t_i}^j Y_p^j \eta_p \epsilon_p^* \right) = S_{t_i-1}^i \widetilde{U}_{t_i-1} + \sum_{p=t_i}^i (1 + S_p^i) \eta_p \epsilon_p^*, \\
\sum_{j=t_i}^i \Delta_j &= \sum_{j=t_i}^i \left(Y_{t_i-1}^j \Delta_{t_i-1} + \sum_{p=t_i}^j Y_p^j \eta_p v_p \right) = S_{t_i-1}^i \Delta_{t_i-1} + \sum_{p=t_i}^i (1 + S_p^i) \eta_p v_p.
\end{aligned}$$

Again, by Lemma 8 (iii), Lemma 11 (Burkholder inequality), Lemma 24 (i), we have

$$\begin{aligned}
\left\|\sum_{j=t_i}^{i} U_j\right\|_4 &\le \left|S^i_{t_i-1}\right| \|U_{t_i-1}\|_4 + \left\|\sum_{p=t_i}^{i}(1+S^i_p)\eta_p\epsilon_p\right\|_4 \\
&\lesssim \left|S^i_{t_i-1}\right| \|U_{t_i-1}\|_4 + \sqrt{\sum_{p=t_i}^{i}\left\|(1+S^i_p)\eta_p\epsilon_p\right\|_4^2} \\
&= \left|S^i_{t_i-1}\right| \|U_{t_i-1}\|_4 + \sqrt{\sum_{p=t_i}^{i}(1+S^i_p)^2\eta_p^2\left\|\epsilon_p\right\|_4^2} \\
&\lesssim t_i^{\alpha}(t_i-1)^{-\alpha/2} + \sqrt{\sum_{p=t_i}^{i}(1+(p+1)^{\alpha})^2 p^{-2\alpha}\cdot\mathcal{O}(1)} \\
&\asymp t_i^{\alpha/2} + \sqrt{\ell_i}.
\end{aligned}$$

Notice that $t_i^{\alpha/2} \le i^{\alpha/2}$ and $\ell_i \asymp i^{\alpha}\log i$, so the above is bounded by $\sqrt{\ell_i} \asymp \sqrt{i^{\alpha}\log i}$.

$\|\sum_{j=t_i}^{i}\widetilde{U}_j\|_4$ has the same bound as $\|\sum_{j=t_i}^{i}U_j\|_4$ since $\|\widetilde{U}_{t_i-1}\|_4, \|U_{t_i-1}\|_4 \lesssim (t_i-1)^{-\alpha/2}$ and $\|\epsilon^*_p\|_4, \|\epsilon_p\|_4 = \mathcal{O}(1)$.

For Δ_i, similarly, also by Lemma 9, we have

$$\begin{aligned}
\left\|\sum_{j=t_i}^{i} \Delta_j\right\|_4 &\le \left|S^i_{t_i-1}\right| \|\Delta_{t_i-1}\|_4 + \left\|\sum_{p=t_i}^{i}(1+S^i_p)\eta_p v_p\right\|_4 \\
&\lesssim \left|S^i_{t_i-1}\right| \|\Delta_{t_i-1}\|_4 + \sqrt{\sum_{p=t_i}^{i}\left\|(1+S^i_p)\eta_p v_p\right\|_4^2} \\
&= \left|S^i_{t_i-1}\right| \|\Delta_{t_i-1}\|_4 + \sqrt{\sum_{p=t_i}^{i}(1+S^i_p)^2\eta_p^2\left\|v_p\right\|_4^2} \\
&\lesssim t_i^{\alpha}(t_i-1)^{-\alpha} + \sqrt{\sum_{p=t_i}^{i}(1+(p+1)^{\alpha})^2 p^{-2\alpha}\cdot p^{-\alpha}} \\
&\asymp \mathcal{O}(1) + \sqrt{i^{-\alpha}\ell_i} \asymp \sqrt{\log i}.
\end{aligned}$$

Notes D.1. *For (ii), applying triangular inequality directly will give a loose bound (but still tight enough to use). For example, by Lemma 9,*

$$\left\|\sum_{j=t_i}^{i} \Delta_j\right\|_4 \le \sum_{j=t_i}^{i}\|\Delta_j\|_4 \lesssim \sum_{j=t_i}^{i} j^{-\alpha} \lesssim i^{-\alpha}\ell_i \asymp \log i.$$

□

Recall that $\delta_i = U_i + s_i$:

- $\delta_i = \delta_{i-1} - \eta_i \nabla f(X_{i-1}, \xi_i)$
- $U_i = (1-\eta_i A)U_{i-1} + \eta_i\epsilon_i, \quad U_0 = \delta_0$
- $s_i = (1-\eta_i A)s_{i-1} - \eta_i r_i, \quad s_0 = 0$

where

- $\epsilon_i = \nabla F(X_{i-1}) - \nabla f(X_{i-1}, \xi_i)$
- $r_i = \nabla F(X_{i-1}) - A\delta_{i-1}$

Lemma 26 (difference sequence)**.**

(i) For all $i \in \mathbb{N}^+$ and $m = 1, 2, 4$, we have

$$\|r_i\|_m \lesssim i^{-\alpha}. \tag{59}$$

(ii) For all $n \in \mathbb{N}^+$ and $m = 1, 2, 4$, we have

$$\|s_n\|_m \lesssim n^{-\alpha}. \tag{60}$$

Proof.

(i) Recall that by Taylor's expansion

$$\begin{aligned}\nabla F(X_{i-1}) &= \nabla F(x^*) + \nabla^2 F(x^*)(X_{i-1} - x^*) + O(|X_{i-1} - x^*|^2) \\ &= A\delta_{i-1} + O(\delta_{i-1}^2),\end{aligned}$$

therefore

$$\|r_i\|_4 = \|\nabla F(X_{i-1}) - A\delta_{i-1}\|_4 \lesssim \|\delta_{i-1}^2\|_4 = \|\delta_{i-1}\|_8^2 \lesssim i^{-\alpha}.$$

(ii) For simplicity, write $Y_i^j := Y_{(A)i}^{\ j}$. Notice that s_n can be written as

$$s_n = -\sum_{i=1}^n Y_i^n \eta_i r_i,$$

therefore

$$\|s_n\|_4 \le \sum_{i=1}^n |Y_i^n| \eta_i \, \|r_i\|_4 \lesssim \sum_{i=1}^n |Y_i^n| i^{-2\alpha} \lesssim n^{-\alpha}.$$

□

Recall that

$$\widehat{\Sigma}_{n,U} = \frac{1}{n}\sum_{i=1}^n \left[U_i^2 + 2U_i \Big(\sum_{k=t_i}^{i-1} U_k \Big) \right].$$
$$\widehat{\Sigma}_{n,\delta} = \frac{1}{n}\sum_{i=1}^n \left[\delta_i^2 + 2\delta_i \Big(\sum_{k=t_i}^{i-1} \delta_k \Big) \right].$$

Lemma 27 (moments for ASGD)**.** *For all $n \in \mathbb{N}^+$, we have*

(i) $\mathbb{E}[\bar{\delta}_n] \lesssim n^{-\alpha}$.

(ii) $\|\bar{\delta}_n\|_4 \lesssim n^{-1/2}$.

(iii) For all $i \in \mathbb{N}^+$,

$$\left\| \sum_{k=t_i}^{i} \delta_k \right\|_4 \lesssim t_i^{\alpha/2} + \sqrt{\ell_i} \asymp \sqrt{i^\alpha \log i}. \tag{61}$$

Proof.

(i) By Lemma 26 (ii), we have $\mathbb{E}[s_n] \leq \|s_n\|_1 \lesssim n^{-\alpha}$. Also recall that

$$\mathbb{E}[U_n] = \mathbb{E}\left[Y_{(A)}{}_0^n \delta_0 + \sum_{i=1}^n Y_{(A)}{}_i^n \eta_i \epsilon_i \right] = Y_{(A)}{}_0^n \delta_0 \lesssim \exp\left(-\eta A n^{1-\alpha}\right).$$

Therefore $\mathbb{E}[\delta_n] = \mathbb{E}[U_n] + \mathbb{E}[s_n] \lesssim n^{-\alpha}$, which implies

$$\mathbb{E}[\bar{\delta}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\delta_i] \lesssim n^{-\alpha}.$$

(ii) Recall that

$$\bar{U}_n = \frac{1}{n}\left(S_{(A)}{}_0^n \delta_0 + \sum_{i=1}^n (1 + S_{(A)}{}_i^n)\eta_i \epsilon_i \right),$$

therefore by Lemma 8 (iii), Lemma 11 (Burkholder's inequality), Lemma 24,

$$\begin{aligned}
\|\bar{U}_n\|_4 &\leq \frac{1}{n}\left| S_{(A)}{}_0^n \delta_0 \right| + \frac{1}{n}\left\| \sum_{i=1}^n (1 + S_{(A)}{}_i^n)\eta_i \epsilon_i \right\|_4 \\
&\lesssim \frac{1}{n}\left| S_{(A)}{}_0^n \right| + \frac{1}{n}\sqrt{\sum_{i=1}^n \left\| (1 + S_{(A)}{}_i^n)\eta_i \epsilon_i \right\|_4^2} \\
&\leq \frac{1}{n}\left| S_{(A)}{}_0^n \right| + \frac{1}{n}\sqrt{\sum_{i=1}^n (1 + S_{(A)}{}_i^n)^2 \eta_i^2 \left\| \epsilon_i \right\|_4^2} \\
&\lesssim \frac{1}{n}\cdot \mathcal{O}(1) + \frac{1}{n}\sqrt{\sum_{i=1}^n (i+1)^{2\alpha} \cdot i^{-2\alpha} \cdot \mathcal{O}(1)} \\
&\lesssim n^{-1} + n^{-1/2} \asymp n^{-1/2}.
\end{aligned}$$

Also by Lemma 26 (ii), we have

$$\|\bar{s}_n\|_4 \leq \frac{1}{n}\sum_{i=1}^n \|s_i\|_4 \lesssim \frac{1}{n}\sum_{i=1}^n i^{-\alpha} \asymp n^{-\alpha}.$$

Therefore $\|\bar{\delta}_n\|_4 \leq \|\bar{U}_n\|_4 + \|\bar{s}_n\|_4 \lesssim n^{-1/2} + n^{-\alpha} \asymp n^{-1/2}$.

(iii) By Lemma 9, Lemma 26 (ii), we have

$$\left\| \sum_{k=t_i}^i s_k \right\|_4 \leq \sum_{k=t_i}^i \|s_k\|_4 \lesssim \sum_{k=t_i}^i k^{-\alpha} \lesssim i^{-\alpha}\ell_i \asymp \log i.$$

In combination with Lemma 25 (ii), we get

$$\left\| \sum_{k=t_i}^i \delta_k \right\|_4 \leq \left\| \sum_{k=t_i}^i U_k \right\|_4 + \left\| \sum_{k=t_i}^i s_k \right\|_4 \lesssim t_i^{\alpha/2} + \sqrt{\ell_i} + i^{-\alpha}\ell_i \asymp \sqrt{i^\alpha \log i}.$$

□

E PROOF OF AUXILIARY RESULTS

Proof of Lemma 8.

(i) Clearly it holds for $j = i$. If $j < i$, noticing that $-x \geq \log(1-x) \geq -x - x^2/2$ for all $x \in [0, 1/2]$, and the set $\{k : \lambda\eta_k > 1/2\}$ is finite, then we have

$$|Y_{(\lambda)}{}_j^i| = \prod_{k=j+1}^{i} |1 - \lambda\eta_k| \lesssim \exp\bigg(-\lambda \sum_{k=j+1}^{i} \eta_k \bigg),$$

and

$$\log|Y_{(\lambda)}{}_j^i| = \sum_{k=j+1}^{i} \log|1 - \lambda\eta_k| \geq -\lambda \sum_{k=j+1}^{i} \eta_k - \lambda^2 \sum_{k=j+1}^{i} \eta_k^2 + \mathcal{O}(1)$$

which implies

$$|Y_{(\lambda)}{}_j^i| = \exp\bigg(\sum_{k=j+1}^{i} \log|1 - \lambda\eta_k| \bigg) \gtrsim \exp\bigg(-\lambda \sum_{k=j+1}^{i} \eta_k - \lambda^2 \sum_{k=j+1}^{i} \eta_k^2 \bigg) \gtrsim \exp\bigg(-\lambda \sum_{k=j+1}^{i} \eta_k \bigg)$$

since $\sum_{k=j+1}^{i} \eta_k^2$ is uniformly bounded for all i and j. As a result,

$$\begin{aligned}
|Y_{(\lambda)}{}_j^i| = \prod_{k=j+1}^{i} |1 - \lambda\eta_k| &\asymp \exp\bigg(-\lambda \sum_{k=j+1}^{i} \eta_k \bigg) \\
&= \exp\bigg(-\lambda\eta \sum_{k=j+1}^{i} k^{-\alpha} \bigg) \asymp \exp\bigg(-\lambda\eta \int_{j+1}^{i} u^{-\alpha}\,\mathrm{d}u \bigg) \\
&= \exp\bigg\{ -\frac{\lambda\eta}{1-\alpha} \left(i^{1-\alpha} - (j+1)^{1-\alpha} \right) \bigg\} \\
&= \exp\bigg\{ \frac{\lambda\eta}{1-\alpha} \left(j^{1-\alpha} - i^{1-\alpha} \right) + \mathcal{O}(1) \bigg\} \\
&\asymp \exp\bigg\{ \frac{\lambda\eta}{1-\alpha} \left(j^{1-\alpha} - i^{1-\alpha} \right) \bigg\}.
\end{aligned}$$

Further, by Taylor series,

$$\begin{aligned}
& \exp\bigg\{ \frac{\lambda\eta}{1-\alpha} (j^{1-\alpha} - i^{1-\alpha}) \bigg\} \\
=\ & \exp\bigg\{ -\lambda\eta i^{-\alpha}(i-j) - \frac{1}{2}\lambda\eta\alpha u^{-\alpha-1}(i-j)^2 \bigg\} && \text{(for some } u \in (j,i)\text{)} \\
\leq\ & \exp\left\{ -\lambda\eta i^{-\alpha}(i-j) \right\}.
\end{aligned}$$

The inequality becomes $\asymp$ if $Ci^\alpha \log i \geq i - j$ since for sufficiently large i, $Ci^\alpha \log i \leq i/2$, and in this case

$$u^{-\alpha-1}(i-j)^2 \leq C^2 i^{2\alpha} (\log i)^2 (i - Ci^\alpha \log i)^{-\alpha-1} \lesssim 2^{\alpha+1} C^2 i^{\alpha-1} (\log i)^2 \to 0.$$

(ii) Since $\exp\left(\beta j^{1-\alpha}\right) j^{-\gamma\alpha}$ is ultimately increasing in j, we have

$$\begin{aligned}
\sum_{j=1}^{i} \exp\left(\beta j^{1-\alpha}\right) j^{-\gamma\alpha} &\asymp \int_1^{i+1} \exp\left(\beta t^{1-\alpha}\right) t^{-\gamma\alpha}\,\mathrm{d}t \\
&\asymp \int_\beta^{\beta(i+1)^{1-\alpha}} \mathrm{e}^u u^{-\frac{(\gamma-1)\alpha}{1-\alpha}}\,\mathrm{d}u. \qquad (62)
\end{aligned}$$

Using integration by parts, we have the following:

$$\begin{aligned}
&\int_{\beta}^{\beta(i+1)^{1-\alpha}} \mathrm{e}^u u^{-\frac{(\gamma-1)\alpha}{1-\alpha}}\,\mathrm{d}u \\
&\asymp \mathrm{e}^u u^{-\frac{(\gamma-1)\alpha}{1-\alpha}}\Big|_{\beta}^{\beta(i+1)^{1-\alpha}} + \int_{\beta}^{\beta(i+1)^{1-\alpha}} \mathrm{e}^u u^{-\frac{(\gamma-1)\alpha}{1-\alpha}-1}\,\mathrm{d}u \\
&\asymp \exp\left\{\beta(i+1)^{1-\alpha}\right\}(i+1)^{-(\gamma-1)\alpha} + (\beta(i+1)^{1-\alpha}-\beta)\exp\left\{\beta(i+1)^{1-\alpha}\right\}(i+1)^{-(\gamma-1)\alpha-(1-\alpha)} \\
&\asymp \exp\left(\beta i^{1-\alpha}\right) i^{-(\gamma-1)\alpha}.
\end{aligned}$$

Therefore by (i), we have

$$\begin{aligned}
\sum_{j=1}^{i} |Y_{(\lambda)}{}_j^i|^{\beta}|j^{-\alpha}|^{\gamma} &\asymp \sum_{j=1}^{i} \exp\left\{\frac{\lambda\eta\beta}{1-\alpha}\left(j^{1-\alpha}-i^{1-\alpha}\right)\right\} j^{-\gamma\alpha} \\
&= \exp\left(-\frac{\lambda\eta\beta}{1-\alpha} i^{1-\alpha}\right) \sum_{j=1}^{i} \exp\left(\frac{\lambda\eta\beta}{1-\alpha} j^{1-\alpha}\right) j^{-\gamma\alpha} \\
&\asymp i^{-(\gamma-1)\alpha}.
\end{aligned}$$

(iii) (See Lemma A.2. in (Zhu et al., 2023))

□

Proof of Lemma 9. Recall that $\ell_i = i - t_i + 1 \asymp i^{\alpha}\log i$, then by Taylor series,

$$\begin{aligned}
\sum_{j=t_i}^{i} j^{-\gamma} &\asymp i^{1-\gamma} - t_i^{1-\gamma} \\
&= (1-\gamma)i^{-\gamma}(\ell_i - 1) + \frac{1}{2}\gamma(1-\gamma)u^{-\gamma-1}(\ell_i-1)^2 && (u \in (t_i, i)) \\
&\lesssim i^{-\gamma}\ell_i + (i - i^{\alpha}\log i)^{-\gamma-1}\ell_i^2 \\
&\asymp i^{-\gamma}\ell_i + i^{-\gamma-1}\ell_i \cdot i^{\alpha}\log i \\
&\asymp i^{-\gamma}\ell_i \asymp i^{\alpha-\gamma}\log i. && (\text{since } \alpha < 1)
\end{aligned}$$

□

F ADDITIONAL EXPERIMENT RESULTS

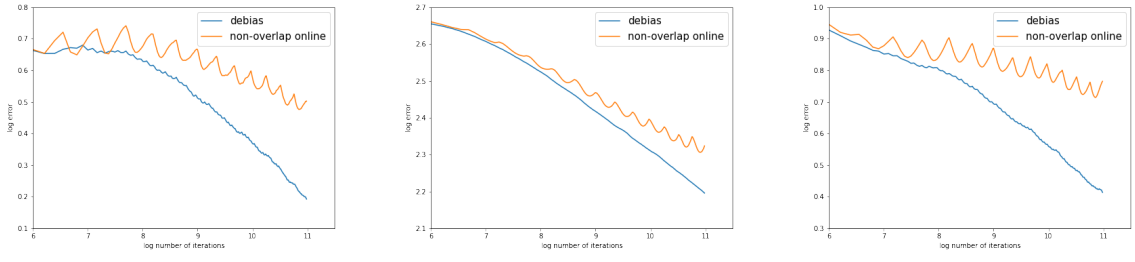

Figure 4: Log-log Plots for Estimation Error in Frobenius Norm ($d = 5$). Left: Linear Regression. Middle: Logistic Regression with. Right: Expectile Regression.

We conducted the experiments in Python version 3.8.8 (2021-02-19) on a MacBook Air with a GPU Apple M1, 4 performance and 4 efficiency cores, and 8 GB LPDDR4 memory, equipped with macOS Big Sur version 11.5.1.

We present additional Log-log plots across different models and dimensional settings in Figure 4. They clearly indicate that the de-biased estimator outperforms the non-overlap online method, achieving both sharper convergence rates (reflected in the slope) and lower estimation errors.

Table 4: Comparison of Empirical Coverage (Nominal Level 95%) across Different Models and Dimensions ($d = 5, 20, 50$).

		$d = 5$	$d = 20$	$d = 50$
Linear	De-biased	0.9236	0.9321	0.9390
	Online BM	0.8796	0.8946	0.9127
Logistic	De-biased	0.8872	0.8514	0.8517
	Online BM	0.8536	0.8178	0.8305
Expectile	De-biased	0.9033	0.9127	0.9203
	Online BM	0.8580	0.8809	0.8949

In Table 4, we compare the empirical coverage rates of confidence intervals constructed using de-biased and Online BM estimators averaged over all coordinates of x^*. The nominal coverage level is 95%. For linear and expectile regression, results are recorded at $n = 60000$ for $d = 5$, $n = 200000$ for $d = 20$, and $n = 500000$ for $d = 50$. For logistic regression, due to its highly non-linear and non-strongly convex nature, we set the batch size constant $C = 4$ and record the coverage at $n = 20000$ for $d = 5$ to ensure convergence, while keeping the sample sizes for $d = 20$ and $d = 50$ the same as above. As shown in Table 4, the de-biased estimator surpasses the Online BM method by achieving empirical coverage rates closer to the nominal 95% level across all scenarios, making it a preferable option for practitioners conducting statistical inference on model parameters.