General Weighted Averaging in Stochastic Gradient Descent: CLT and Adaptive Optimality

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

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

Abstract

Stochastic Gradient Descent (SGD) is a cornerstone of machine learning, prized for its efficiency in large-scale optimization. This paper revisits SGD by introducing a general weighted averaging framework that significantly enhances its applicability. We establish asymptotic normality for a wide range of weighted averaged SGD solutions under minimal assumptions, providing a groundbreaking necessary condition for the central limit theorem in certain settings. This enables asymptotically valid online inference, empowering real-time confidence interval construction. Furthermore, we propose an adaptive averaging scheme, inspired by optimal weights for linear models, which achieves optimal superior non-asymptotic bounds. Our theoretical advances and empirical validations redefine SGD's capabilities, offering transformative insights for statistical learning and optimization.

1 INTRODUCTION

Stochastic Gradient Descent (SGD) is one of the most popular optimization algorithms. It (along with its variants) plays a crucial role in statistical and machine learning problems (Robbins and Monro, 1951; Lai, 2003; Bottou et al., 2018). To find the minimizer of a convex function F, SGD produces a sequence of iterates through an unbiased estimate of F's gradient/subgradient. The algorithm is easy to implement and popular in applications for its effectiveness, memory and computational efficiency, and online property.

Theoretical properties of SGD have been extensively studied in stochastic approximation theory and machine learning from asymptotic convergence to non-asymptotic analysis (Kushner and Yin, 2003; Moulines and Bach, 2011; Toulis and Airoldi, 2017; Harvey et al., 2019). After n iterations, optimal convergence rates of optimization error are $\mathcal{O}(1/n)$ and $\mathcal{O}(1/\sqrt{n})$ for strongly-convex and convex problems (Nemirovski et al., 2009; Lacoste-Julien et al., 2012). It is shown that SGD with a learning rate proportional to the inverse of the number of iterations is not optimal under the wrong setting of the proportionality constant or without strong convexity assumptions. The idea of using *averaging* to accelerate convergence is proposed by Ruppert (1988) and Polyak and Juditsky (1992). They demonstrated that using a learning rate with slower decays, combined with uniform averaging, robustly leads to information-theoretically optimal asymptotic variance. This averaging scheme, where all iterates are averaged, is known as Polyak-Ruppert averaging or averaged SGD (ASGD). However, it is not optimal from a non-asymptotic perspective (Moulines and Bach, 2011; Needell et al., 2014). Moreover, for non-smooth objective functions, neither the final iterate nor ASGD can achieve the optimal finite sample convergence rate (see, e.g., Section 4 in Rakhlin et al., 2012). To address these issues, various modified versions of ASGD have been proposed, such as suffix averaging (Rakhlin et al., 2012) and polynomial-decay averaging (Shamir and Zhang, 2013) for non-smooth problems, exponential weighted moving average (EWMA) for capturing time variation, elastic averaging in parallel computing environments (Zhang et al., 2015), and a simple weight proportional to $\mathcal{O}(n)$ for the projected stochastic subgradient method (Lacoste-Julien et al., 2012).

As mentioned earlier, employing a suitable averaging scheme in specific settings can help enhance conver-

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

gence and necessitates only a straightforward modification to the original SGD algorithm. This paper delves deeper into the characteristics of averaging by examining a comprehensive averaging scheme. Our main objectives include understanding the variability and statistical efficiency of a general weighted averaged SGD, thereby providing inference foundations for it. Under certain mild assumptions, we find the optimal condition for the central limit theorem (CLT) of weighted averaged SGD solutions. The condition is optimal in the sense that it is sufficient and cannot be weakened under certain settings. We also establish the asymptotic normality of a generic weighting scheme commonly seen in practice (with slightly stronger yet more easily verifiable conditions on weights). The theoretical results are applicable to a wide range of existing algorithms, including the last iterate of SGD, the polynomial-decay, and suffix averaged SGD. Moreover, statistical inference for weighted averaged SGD is a valuable byproduct owing to its close relation to ASGD. This means we can effectively utilize inference methods, such as covariance matrix estimation (Chen et al., 2020; Zhu et al., 2023) or asymptotic pivotal statistics (Lee et al., 2022; Zhu et al., 2024), originally designed for ASGD in this context. We also refine the functional CLT of SGD originally presented in Lee et al. (2022) under weaker assumptions.

Beyond asymptotic normality, we also investigate finite sample convergence. When considering finite sample MSE, it is challenging to identify a single averaging scheme that is optimal for all objective functions. To gain insights, we examine the linear model to derive adaptively weighted SGD iterates that minimize the finite sample MSE. Crucially, this adaptive weighting algorithm preserves SGD's online nature and retains the CLT guarantee with optimal asymptotic covariance, thereby enabling practitioners to conduct statistical inference efficiently. Our numerical study indicates that the adaptive weights derived from the linear model not only achieve the optimal statistical rate but also exhibit favorable non-asymptotic convergence across various models.

Related Works. There has recently been increasing interest in the statistical inference of SGD and other stochastic approximation procedures. One line of the work is to study the limiting distribution. Following the celebrated asymptotic normality results of ASGD (Ruppert, 1988; Polyak and Juditsky, 1992), similar asymptotic normality results are proposed for specific algorithms in different settings such as non-convex settings, linear systems, and online decision-making (Yu et al., 2021; Mou et al., 2020; Chen et al., 2021). Also, various methods based on covariance matrix estimation or bootstrap/subsampling are proposed to construct asymptotically valid confidence intervals (Chen et al., 2020; Zhu et al., 2023; Li et al., 2018; Fang et al., 2018; Liang and Su, 2019; Lee et al., 2022). In addition to asymptotic results, there are also many non-asymptotic works related to inference, for example, non-asymptotic normal approximation of SGD (Anastasiou et al., 2019) and concentration analysis (Davis et al., 2021; Harvey et al., 2019; Lou et al., 2022).

In the broader context of stochastic approximation, additional works have also highlighted the utility of weighted averaging. Dippon and Renz (1997) derived asymptotic normality for a fixed polynomial weighting form. Mokkadem and Pelletier (2006) studied averaging in two-time-scale stochastic approximation algorithms. More recently, Boyer and Godichon-Baggioni (2023) considered the Weighted Averaged Stochastic Newton Algorithm and introduced a condition on the weighted averaging sequence ratio.

The remainder of this paper is organized as follows. In Section 2, we introduce the main theorem for asymptotic normality along with statistical inference methods that can be implemented in an online fashion. In Section 3, we present two significant examples: polynomial-decay averaging and suffix averaging, both of which are applicable to our theorem. Additionally, we adapt the suffix averaging process for online implementation. In Section 4, we identify optimal weights for the linear model and propose a novel adaptive weighting scheme that achieves both optimal statistical rates and favorable finite sample performance. Numerical results are provided in Section 5. Finally, Section 6 concludes the paper and outlines directions for future research.

2 ASYMPTOTIC NORMALITY FOR WEIGHTED SGD

2.1 Overview

In many statistical estimation and machine learning problems, we need to estimate the minimizer of a convex objective function $F(x)$, mapping from $\mathbb{R}^d$ to $\mathbb{R}$, as $F(x) = \mathbb{E}_{\xi\sim\Pi} f(x, \xi)$, where $f(x, \xi)$ is a loss function and ξ is a random variable following the distribution Π. Assuming the minimizer

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

exists, the SGD updates the iterate as follows (initialized at x_0),

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

where ξ_i are $i.i.d$ from Π, η_i is the i-th step size or learning rate, $\nabla f(x, \xi)$ is the gradient of f with respect to the first argument, and x_0 is a starting point.

Let $\bar{x}_n = \sum_{i=1}^n x_i/n$ be the uniform average. Under suitable conditions, Polyak and Juditsky (1992) shows that

$$\sqrt{n}(\bar{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, V), \tag{2}$$

where

$$\begin{aligned} V = A^{-1}SA^{-1}, A = \nabla^2 F(x^*), \\ S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T). \end{aligned} \tag{3}$$

In this paper, one of our goals is to establish the asymptotic normality for the general weighted average

$$\tilde{x}_n = \sum_{i=1}^n w_{n,i} x_i, \tag{4}$$

where $w_{n,i}$, $1 \le i \le n$, denotes the weight of x_i after the n-th update with $\sum_{i=1}^n w_{n,i} = 1$. We propose a general condition on $\{w_{n,i}\}_{1\le i\le n}$ for the weighted averaged SGD (4) to be asymptotically normal. Additionally, we also provide easily verifiable conditions for the asymptotic normality result with a modified weighting condition. To be specific, we show that under certain mild assumptions,

$$\sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, wV), \text{ where } w = \lim_{n\to\infty} n \sum_{i=1}^n (w_{n,i})^2$$

is a prefactor caused by weighting, and V is defined in (3). We demonstrate that this result holds for many existing averaging schemes, as well as the adaptive averaging we present in Section 4.

We now introduce some notation. Throughout the paper, for $t \in \mathbb{R}$, $\lfloor t \rfloor = \max\{i \in \mathbb{Z} : i \le t\}$ and $\lceil t \rceil = \min\{i \in \mathbb{Z} : i \ge t\}$. We use $\mathscr{F}_i$ to denote the nested σ-algebra generated by $\{\xi_1, \dots, \xi_i\}$. Let $\mathbb{E}_i$ denote the conditional expectation $\mathbb{E}(\cdot|\mathscr{F}_i)$, and $\mathbb{P}_i$ denote the conditional probability $\mathbb{P}(\cdot|\mathscr{F}_i)$. We use $\xrightarrow{D}$ to denote convergence in distribution, and $\xrightarrow{P}$ to denote convergence in probability. For a vector $a = (a_1, \dots, a_p)^\top$ let the norm $|a| = (\sum_{i=1}^p a_i^2)^{1/2}$. Define $|A|_F$, $|A|$, $\lambda_{\max}(A)$, $\lambda_{\min}(A)$, and $(A)_{jj}$ as the Frobenius norm, operator norm, the largest eigenvalue, the smallest eigenvalue, and the j-th diagonal of a matrix A. For a random variable, vector, or matrix X, we write $X \in \mathcal{L}^p$, $p > 0$, if $\|X\|_p := (\mathbb{E}|X|^p)^{1/p} < \infty$, and write $\|X\|_2 = \|X\|$. For notation simplicity, we use C, $\tilde{C}$, K and $C_1, C_2, \dots$, to denote different constants, whose values may change in different equations.

2.2 Main Results

We begin by introducing several assumptions.

Assumption 2.1. *$F(x)$ is continuously differentiable with $\nabla^2 F(x^*)$ existing and strongly convex with parameter $\mu > 0$. That is, for any x_1 and x_2, the inner product $\langle \nabla F(x_2) - \nabla F(x_1), x_2 - x_1 \rangle \ge \mu |x_1 - x_2|^2$, or equivalently*

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

Assumption 2.2. *The function $f(x, \xi)$ is continuously differentiable w.r.t. x for any ξ, $\sigma(x) := \|\nabla f(x, \xi)\| < \infty$, and $\nabla f(x, \xi)$ is stochastic Lipschitz continuous with parameter L_2, i.e., for any x_1 and x_2,*

$$\|\nabla f(x_1, \xi) - \nabla f(x_2, \xi)\| \le L_2 |x_1 - x_2|.$$

Assumption 2.1 requires strong convexity and the existence of the Hessian at the true parameter, which are needed to derive the asymptotic normality of ASGD solutions. These properties are also important for obtaining the desired error bounds on SGD iterates and the asymptotic properties of the weighted averaged SGD.

For Assumption 2.2, recall that we use $\|\cdot\| = (\mathbb{E}|\cdot|^p)^{1/p}$ to denote the $\mathcal{L}^p$ norm of a random vector, and the expectation here is taken in terms of ξ. Assumption 2.2 ensures that Leibniz's integration rule holds. Consequently, the gradient noise $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$ is a martingale difference, i.e., $\mathbb{E}_{n-1}\epsilon_n = \mathbb{E}(\epsilon_n|\mathscr{F}_{n-1}) = 0$. This stochastic Lipschitz continuity condition of the gradient guarantees the L-smoothness of F and the boundedness of gradient noise moments. Easily verified examples include linear and logistic regression. Similar assumptions have been adopted in the literature on asymptotics and statistical inference of SGD (Polyak and Juditsky, 1992; Chen et al., 2020; Li et al., 2022a; Zhu et al., 2023; Li et al., 2024).

With the aforementioned assumptions in place, we will now present our main results regarding the very general condition of the asymptotic normality for weighted averaged SGD solutions. Theorem 2.3 allows general weights $w_{n,i}$ and any starting point x_0. Recall (3) for A and S. We first introduce the expression of the asymptotic covariance matrix of $\tilde{x}_n$:

$$\begin{aligned} V_n &= \sum_{i=1}^n \Phi_{n,i} S \Phi_{n,i}, \quad \text{where} \\ \Phi_{n,i} &= \eta_i \sum_{k=i}^n w_{n,k} \prod_{j=i+1}^k (\mathbf{I}_d - \eta_j A). \end{aligned} \tag{5}$$

Theorem 2.3. *Given SGD (1) with step size $\eta_i = \eta i^{-\alpha}$ for some $\eta > 0$ and $1/2 < \alpha < 1$, we consider the general averaging scheme (4) with $w_{n,i} \ge 0$.*
(i) Under Assumptions 2.1 and 2.2, the following condition

$$\max_{1\le i\le n} |V_n^{-\frac{1}{2}} \Phi_{n,i}| \to 0 \tag{6}$$

is sufficient for the central limit theorem: for any starting point x_0,

$$V_n^{-1/2}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{I}_d). \tag{7}$$

(ii) Consider the special case where $f(x,\xi) = \frac{1}{2}|\xi - A^{\frac{1}{2}}x|^2$ *and assume that* $\nabla f(x^*, \xi_i)$ *is not normally distributed. Then condition (6) is also necessary for the CLT (7).*

Theorem 2.3(ii) represents the first known necessary condition for the CLT to hold for SGD solutions. The proof relies on the Lévy-Cramér theorem, which characterizes the factor closeness of Gaussian distributions.

Remark 2.4. *In the appendix Section B.3 we provide a more general version of CLT, which further allows a broader range of the learning rate schedules* η_i.

As a specific example, by setting $w_{n,n} = 1$ and $w_{n,i} = 0$ for $1 \le i < n$, Theorem 2.3 implies the following Corollary 2.5, which asserts CLT for the last iterate of SGD when $1/2 < \alpha < 1$. This yields a novel and self-contained proof of the CLT for the last iterate. A key step involves demonstrating the convergence of the covariance matrix; the detailed argument is provided in the appendix. Early versions of such CLTs were first obtained in Chung (1954), while subsequent works such as Sacks (1958); Fabian (1968); McLeish (1976); Borkar (2008); Toulis and Airoldi (2017) cover the case $\alpha = 1$.

Corollary 2.5. *Under Assumptions 2.1 and 2.2, the last SGD iterate in (1) with step size* $\eta_i = \eta i^{-\alpha}$ *for some* $\eta > 0$ *and* $1/2 < \alpha < 1$ *is asymptotically normal:*

$$n^{\frac{\alpha}{2}}(x_n - x^*) \xrightarrow{D} \mathcal{N}(0, \eta V_0),$$

where V_0 *is the unique solution to the Lyapunov equation* $AV_0 + V_0 A = S$.

Remark 2.6. *For a detailed discussion on the Lyapunov equation and its connection to stochastic optimization, we refer the reader to Theorem 1 in Pflug (1986).*

Since it is usually implicit whether a specific weighting scheme meets condition (6), we introduce in Theorem 2.7 and Remark 2.9 two more explicit ways to ensure the CLT for a general weighted averaging scheme.

Theorem 2.7. *Given SGD iterates in (1) with step size* $\eta_i = \eta i^{-\alpha}$ *for some* $\eta > 0$ *and* $1/2 < \alpha < 1$, *we consider the general averaging scheme (4) where the weight* $w_{n,i}$ *satisfies the following conditions:*

1. $|w_{n,i}| \le Cn^{-1}$ *for some constant* C.
2. $w = \lim_{n\to\infty} n \sum_{i=1}^{n} (w_{n,i})^2$ *exists.*
3. *piecewise Lipschitz: there exist functions* $\{f_n\}$ *from* $[0,1]$ *to* $\mathbb{R}$, *a finite set* $\{t_1, t_2, ..., t_N\} \subseteq [0,1]$ *and a constant* $L_1 > 0$ *such that* $w_{n,k} = f_n(k/n)/(\sum_{k=1}^{n} f_n(k/n))$ *with* $\sum_{k=1}^{n} f_n(k/n) \asymp n$ *and* $|f_n(s_1) - f_n(s_2)| \le L_1|s_1 - s_2|$, *for all* $n \in \mathbb{N}^+$ *and all* s_1, $s_2 \in [t_j, t_{j+1}]$ *where* $1 \le j \le N-1$.

Then under Assumptions 2.1 and 2.2, we have the CLT $\sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, wV)$, *where* V *is defined in* (3)

Remark 2.8. *The asymptotic covariance of the weighted average* $\tilde{x}_n$ *here is composed of a prefactor* w *and the sandwich form* $A^{-1}SA^{-1}$. *In the context of ASGD, where* $w_{n,i} = 1/n$, *the prefactor* $w = 1$, *which aligns with our results. We also show in the appendix that, this asymptotic covariance is exactly the limit of* $nV_n = n\Phi_{n,i}S\Phi_{n,i}$ *as defined in Theorem 2.3, indicating the consistency of our main results.*

Remark 2.9. *The piecewise Lipschitz condition (condition 3 in Theorem 2.7) requires that the majority of the weights do not undergo drastic changes. It is satisfied by a variety of averaging schemes (see Sections 3 and 4), and a slightly stronger yet simpler alternative is* $|w_{n,i+1} - w_{n,i}| \le \tilde{C}n^{-2}$, $\tilde{C} > 0$.

2.3 Online Statistical Inference

The asymptotic normality result presented in Theorem 2.7 enables statistical inference for the weighted averaged SGD, which includes tasks such as constructing confidence intervals. We will explore two online methods, originally designed for ASGD, applicable in this context.

Via covariance matrix estimation. One approach to construct confidence intervals involves estimating the limiting covariance matrix V and directly utilizing the asymptotic normality result outlined in Theorem 2.7. For example, the 95% confidence interval for the j-th element x_j^* can be constructed as: $\left[(\tilde{x}_n)_j \pm z_{0.975}\sqrt{\frac{w(\hat{V}_n)_{jj}}{n}}\right]$, where $z_{0.975}$ corresponds to the 97.5% percentile for a standard normal distribution, $(\tilde{x}_n)_j$ denotes the j-th coordinate of $\tilde{x}_n$, and $\hat{V}_n$ is the estimate for the sandwich form matrix $V = A^{-1}SA^{-1}$. The prefactor w can be readily computed for any specific averaging scheme, meaning that we only need to estimate the sandwich form matrix V. It is important to note that this sandwich form matrix V is the asymptotic covariance matrix of ASGD $\bar{x}_n$. There are several established approaches to estimating the asymptotic covariance matrix of ASGD in an online fashion, such as the plug-in method, which estimates A and S separately, and the batch-means method, which employs SGD iterates to directly construct $\hat{V}_n$ (Chen et al., 2020; Zhu et al., 2023). As we are utilizing the vanilla

SGD algorithm, these aforementioned approaches can be applied to obtain a consistent sandwich form matrix estimate $\hat{V}_n$. Then, we can construct the asymptotic valid confidence intervals.

Via asymptotically pivotal statistics. An alternative approach to constructing confidence intervals is to use asymptotically pivotal statistics. One method involves random scaling, where we studentize $\sqrt{n}(\tilde{x}_n - x^*)$ using the random scaling matrix $\hat{V}_{rs,n}$ as described in Lee et al. (2022), i.e.,

$$\hat{V}_{rs,n} = \frac{1}{n}\sum_{s=1}^{n}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{s}(x_i - \bar{x}_n)\right\}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{s}(x_i - \bar{x}_n)\right\}^T.$$

The following functional CLT in Theorem 2.10 concerns weak convergence for the partial sum process $S_k = \sum_{i=1}^{k}(x_i - x^*)$, which was originally proved in Lee et al. (2022). Note that the functional CLT only requires the existence of the $(2+\epsilon)$-th moment of the noise. This moment condition is relaxed compared to Assumption 1.(v) in Lee et al. (2022). By the continuous mapping theorem, (9) follows from (8).

Theorem 2.10. *Assume $\|\nabla f(x^*, \xi)\|_p < \infty$ for some $p > 2$. Let $\mathbb{B}(u) = (\mathbb{B}_1(u), \ldots, \mathbb{B}_d(u))$ be the standard d-dimensional Brownian motion, and $\Rightarrow$ denote the weak convergence in $l^\infty[0,1]$. Under Assumptions 2.1 and 2.2, we have the functional CLT:*

$$\{S_{\lfloor nu \rfloor}/\sqrt{n}, 0 \le u \le 1\} \Rightarrow V^{1/2}\{\mathbb{B}(u), 0 \le u \le 1\}, \tag{8}$$

Consequently, the asymptotic pivotal convergence holds:

$$\sqrt{n}((\tilde{x}_n)_j - x_j^*)/\sqrt{(\hat{V}_{rs,n})_{jj}} \xrightarrow{D} \sqrt{w}U \tag{9}$$

where

$$U = \mathbb{B}_1(1)/\sqrt{\int_0^1 \{\mathbb{B}_1(r) - r\mathbb{B}_1(1)\}^2 dr}.$$

With (9), one can construct confidence intervals based on Theorem 2.10. For instance, the 95% confidence interval for the j-th element x_j^* can be constructed as: $(\tilde{x}_n)_j \pm u_{rs,0.975}\{w(\hat{V}_{rs,n})_{jj}/n\}^{1/2}$, where $u_{rs,0.975}$ is the 97.5% quantile for U. Note that the random scaling matrix $\hat{V}_{rs,n}$ can be updated recursively, thereby enabling the construction of confidence intervals in an online fashion. Besides the random calling method, one can also employ the parallel run method to construct a t-type statistic using a small number of parallel runs of weighted averaged SGD; we refer to Zhu et al. (2024) for details.

3 EXAMPLES

In this section, we shall apply our results to two widely used specific examples of averaging schemes: polynomial-decay averaging (Shamir and Zhang, 2013) and suffix averaging (Rakhlin et al., 2012). The original suffix averaging scheme is not recursive. Here we shall modify it to an online fashion in which estimates can be computed recursively. We will show that they are not statistically optimal, with a constant prefactor that is strictly greater than unity; see Corollaries 3.1 and 3.2 below.

3.1 Polynomial-decay Averaging

With a small number $\gamma \ge 1$, the polynomial-decay averaging (Shamir and Zhang, 2013) is defined as follows: Given iterates $\{x_i\}_{i=1}^{\infty}$, $\tilde{x}_1 = x_1$, and for any $n \ge 1$,

$$\tilde{x}_n = (1 - \frac{\gamma+1}{\gamma+n})\tilde{x}_{n-1} + \frac{\gamma+1}{\gamma+n}x_n. \tag{10}$$

The recursion form in (10) can be rewritten as the weighted average $\tilde{x}_n = \sum_{i=1}^{n} w_{n,i}x_i$ with weights

$$\begin{aligned} w_{n,i} &= \frac{\gamma+1}{\gamma+i}\prod_{j=i+1}^{n}\frac{j-1}{j+\gamma} \\ &= \frac{\gamma+1}{n}\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}, \end{aligned}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. The weights $w_{n,i}$ satisfy conditions in Theorem 2.7. Therefore we have the following CLT in view of

$$\lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2 = \frac{(\gamma+1)^2}{2\gamma+1} =: \tau.$$

Corollary 3.1. *Consider SGD (1) with step size $\eta_i = \eta i^{-\alpha}$ for $\eta > 0$, $0.5 < \alpha < 1$, and polynomial-decay averaging $\tilde{x}_n$ defined in (10). Under Assumptions 2.1 and 2.2, we have $\sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, \tau V)$.*

3.2 Suffix Averaging

The κ-suffix averaging in Rakhlin et al. (2012) is defined as the average of the last $\lceil \kappa n \rceil$ iterates:

$$\tilde{x}_n = \frac{1}{\lceil \kappa n \rceil}\sum_{i=\lceil (1-\kappa)n \rceil}^{n} x_i,\ 0 < \kappa < 1 \tag{11}$$

For κ-suffix averaging, the weight $w_{n,i} = 1/\lceil \kappa n \rceil$ for $i > (1-\kappa)n$ otherwise 0. The weight satisfies conditions in Theorem 2.7 with $\lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2 = 1/\kappa$. Thus we have the following CLT:

Corollary 3.2. *Consider SGD iterates in* (1) *with step size* $\eta_i = \eta i^{-\alpha}$ *for* $\eta > 0$, $0.5 < \alpha < 1$, *and* $\tilde{x}_n$ *defined in* (11). *Under Assumptions 2.1 and 2.2, we have* $\sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, \kappa^{-1}V)$.

Remark 3.3 (Online algorithm for suffix averaging)**.** *Since* $(1-\kappa)n$ *depends on* n, *the* κ*-suffix averaging cannot be computed on-the-fly unless all iterations are stored and the stopping time* n *is known beforehand. To enable "on-the-fly" computation, we modify the suffix averaging procedure to an online method. We employ the concept of online batch scheme: divide the rounds into blocks and track iterations within the current block (or the most recent blocks). The block sizes are pre-defined based on various objectives and training parameters. For a pre-defined sequence* $(a_m)_{m\geq 0}$, *we treat* x_{a_m} *as the start of the* m*-th block. Let* m_t *denote the block index for the* t*-th iteration, satisfying* $a_{m_t} \leq t < a_{m_t+1}$.

In the online suffix averaging procedure, we partition the rounds into exponentially increasing blocks, and maintain the average of the last two blocks. In particular, we set $a_m = \lfloor 2^{m-1} \rfloor + 1, m \geq 0$. *Then* $B_0 = \{x_1\}, B_1 = \{x_2\}, B_2 = \{x_3, x_4\}, B_3 = \{x_5, x_6, x_7, x_8\}, \ldots,$ *and the end index of the* m*-th block is* 2^m. *We have* $m_t = \lceil \log_2 t \rceil$, *i.e.,* $\lfloor 2^{m_t - 1} \rfloor < t \leq 2^{m_t}$. *Given the SGD iterates* $x_1, x_2, \ldots$, *the online suffix averaging procedure is defined as follows:* $\hat{x}_1 = x_1, \hat{x}_2 = (x_1 + x_2)/2$, *and for* $t \geq 3$,

$$\hat{x}_t = \frac{1}{t - 2^{\lceil \log_2 t \rceil - 2}} \left(\sum_{k=2^{\lceil \log_2 t \rceil - 2}+1}^{2^{\lceil \log_2 t \rceil - 1}} x_k + \sum_{k=2^{\lceil \log_2 t \rceil - 1}+1}^{t} x_k \right). \tag{12}$$

Note that $1/2 < (t - 2^{\lceil \log_2 t \rceil - 2})/t \leq 3/4$ *for* $t \geq 3$. *Therefore, the online suffix averaging is a form of robust suffix averaging with* $1/2 < \kappa \leq 3/4$, *i.e., the average would always correspond to a constant-portion suffix of all iterates. Thus the online suffix average* $\hat{x}_t$ *in* (12) *can be updated recursively.*

4 NON-ASYMPTOTIC MEAN SQUARED ERROR (MSE)

In addition to the asymptotic distribution and statistical convergence rates, it is also important to consider finite sample performance when dealing with finite data problems or when early stopping is desired. In this section, we will examine the optimal weight for a linear model in terms of finite sample MSE, building upon the concept of *best linear unbiased estimation* (BLUE). Furthermore, we will introduce a novel adaptive averaging scheme based on insights from the mean estimation model. This particular scheme is both statistically efficient with optimal variance, and has a fast finite sample convergence rate, outperforming existing averaging schemes in the mean estimation model. For (4), we can evaluate its non-asymptotic performance through its $\text{MSE}(\tilde{x}_n) = \mathbb{E}|\tilde{x}_n - x^*|^2$. To obtain the MSE-optimal weights, we consider

$$\min_{c=(c_1,\,\cdots,\,c_n):\, c^T\mathbb{1}=1} \mathbb{E}\Big|\sum_{i=1}^{n} c_i x_i - x^*\Big|^2, \tag{13}$$

where $\mathbb{1}$ denotes the all-ones vector. Let $\Sigma = (\mathbb{E}((x_i - x^*)^T(x_j - x^*)))_{1\leq i,j\leq n}$. The solution to the above constrained optimization problem is

$$c = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}}. \tag{14}$$

The solution depends on the correlation between SGD iterates and can vary across different models. In this section, we shall study the linear regression model, which can have a closed-form solution of the BLUE weights. The latter can provide insights into the properties of an optimal weight in a generalized form.

Consider the following linear regression model:

$$b_i = a_i x^* + \epsilon_i, \tag{15}$$

where x^* denotes the unknown parameter of interest, ϵ_i are *i.i.d.* standard normal, and $\xi_i = (a_i, b_i)$ denote the observed streaming data. To solve the above linear regression problem, we consider the squared loss function $F(x) = \mathbb{E}(f(x, \xi_i)) = \mathbb{E}(a_i x - b_i)^2/2$, and SGD sequence with step size η_i at the i-th iteration:

$$x_i = x_{i-1} - \eta_i a_i (a_i x_{i-1} - b_i). \tag{16}$$

Proposition 4.1. *Consider the linear model in* (15) *and SGD sequence* x_i *defined in* (16) *with step size* $\eta_1 = a_1^{-2}$ *and general* $\eta_i > 0$, $i \geq 2$. *The unique solution to the optimization problem* (13) *is given by*

$$c_{n,i} = \frac{a_{i+1}^2 + \eta_i^{-1} - \eta_{i+1}^{-1}}{S_n}, 1 \leq i \leq n-1,$$
$$c_{n,n} = \frac{1}{\eta_n S_n}, \text{ where } S_n = \sum_{i=1}^{n} a_i^2.$$

The weights in Proposition 4.1 are adjusted by the learning rates η_i. However, one characteristic of these weights is that the last weight is significantly larger than the preceding ones. In comparison with the equal weights scheme, our $c_{n,i}$ is intuitively attractive by putting most weight on the last iterate.

A new averaging scheme: adaptive weighted averaging. In the special case of the mean estimation model, where $a_i = 1, \forall i \geq 1$, then the weights are given by:

$$c_{n,i} = \frac{1 + \eta_i^{-1} - \eta_{i+1}^{-1}}{n}, \, 1 \leq i \leq n-1, \, c_{n,n} = \frac{1}{n\eta_n}. \tag{17}$$

When the number of iteration increases from n to $n+1$, we have $c_{n+1,i} = (n+1)^{-1} n c_{n,i}, 1 \leq i \leq n-1$. Therefore the weighted averaged SGD $\tilde{x}_n$ with above optimal weights can also be recursively updated:

$$\tilde{x}_{n+1} = \frac{n}{n+1}\tilde{x}_n + \frac{1 - \eta_{n+1}^{-1}}{n+1} x_n + \frac{\eta_{n+1}^{-1}}{n+1} x_{n+1}. \tag{18}$$

The newly introduced averaging scheme, referred to as *adaptive weighted averaging*, effectively decreases the weight of earlier iterates in comparison to the newly arriving ones. From the definition of the optimal weights, the weights in (17) minimize the finite sample MSE among all possible weights for the mean estimation model. On the other hand, it also satisfies the condition in Theorem 2.3, which indicates that the corresponding weighted averaged SGD satisfies CLT with the asymptotic covariance matrix being the same as that of ASGD estimates; see the following Corollary 4.2. Thus, the proposed adaptive weighted average achieves both fast finite sample convergence rates and the optimal statistical rate.

Corollary 4.2. *Under the settings in Theorem 2.3, for weights $c_{n,i}$ in (17), we have $\sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, V)$.*

Connection with uniform averaging. It is interesting to compare the adaptive weighted average in (18) with uniform average (ASGD). The recursion of ASGD $\bar{x}_n$ takes the following form

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1} x_{n+1}. \tag{19}$$

To build the connection between (18) and (19), we rewrite (18) as

$$\tilde{x}_{n+1} = \frac{n}{n+1}\tilde{x}_n + \frac{1}{n+1} x_{n+1} + \frac{\eta_{n+1}^{-1} - 1}{n+1}(x_{n+1} - x_n). \tag{20}$$

Thus we can consider the proposed adaptive weighted average as a modified ASGD with a correction term, where the correction term reduces the weight of earlier iterates and increases the weight of the latest iteration. This modification bears a certain similarity with other existing variance-reduced modifications on SGD where a correction term is applied on stochastic gradients, such as SGD with momentum (Nesterov, 1983; Kingma and Ba, 2015; Cutkosky and Orabona, 2019).

5 NUMERICAL EXPERIMENT

In this section, we check the asymptotic normality property of the general weighted averaged SGD and investigate the non-asymptotic performance of various averaging schemes in different settings.

5.1 Asymptotic Normality for Different Averaging Schemes

To verify the asymptotic normality and the limiting covariance matrix derived in Theorem 2.7, we consider three averaging schemes: polynomial-decay, suffix averaging (as described in Section 3), and the adaptive averaging scheme proposed in (18). We focus on two classes of loss functions: squared loss $f(x, \xi_i = (a_i, b_i)) = (a_i^T x - b_i)^2/2$ for the linear regression model, and logit loss: $f(x, \xi_i = (a_i, b_i)) = \log(1 + \exp(-b_i a_i^T x))$ for the logistic regression model. In both models, we assume that the data $\xi_i = (a_i, b_i)$, are independent, where a_i represents the explanatory variable generated from $\mathcal{N}(0, \mathbf{I}_d)$, and b_i represents the response variable generated from two different distributions correspondingly. For linear regression, we assume $b_i \sim \mathcal{N}(a_i^T x^*, 1)$, while for logistic regression, $b_i \in \{1, -1\}$ is generated from a Bernoulli distribution, where $\mathbb{P}(b_i|a_i) = 1/(1+\exp(-b_i a_i^T x^*))$. For asymptotic normality we are going to verify in Theorem 2.7. For squared loss, it is easy to derive that $A = S = \mathbf{I}_d$ and therefore $V = \mathbf{I}_d$. For logit loss, since the explicit forms for A and S are difficult to obtain, we use Monte-Carlo simulation to numerically compute V.

In simulations, we set $d = 5$ and the true parameter $x^* = (1, -2, 0, 0, 4)^T$ for both models. We generate SGD sequences with $\eta_i = i^{-\alpha}, \alpha = 0.505$, and apply different averaging schemes. The number of iterations $n = 100000$. For the polynomial-decay averaging, we choose $\gamma = 3$ Shamir and Zhang (2013), and for the suffix averaging we choose $\kappa = 0.5$ Rakhlin et al. (2012). Then the prefactors w for polynomial-decay and suffix averaging schemes are $16/7$ and 2. For polynomial decay and suffix averaged SGD, we plot the density of the standardized error with and without prefactor w, i.e., $w^{-1}V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$ and $V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$. For adaptive weighted SGD, the prefactor is 1 according to Corollary 4.2, so we only plot the density of the standardized error $V^{-1}\sqrt{n}(\tilde{x}_n - x^*)$. As shown in Figure 1, under both the polynomial decay and suffix averaging schemes, the standardized error—after being scaled by the prefactor w from our calculations—approximately follows a standard normal distribution. The same behavior is observed for our proposed adaptive weighted SGD procedure. These findings support the conclusion of Theorem 2.7, affirming the asymptotic normality of the weighted SGD solutions and the correctness of the limiting covariance matrix we derived.

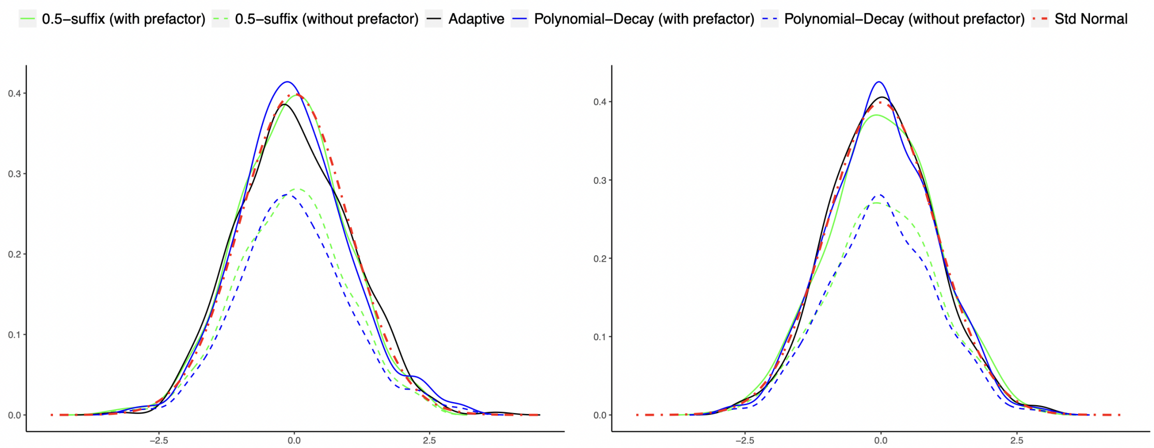

Figure 1: Density Plot for the Standardized Error. Left: Linear Model. Right: Logistic Model.

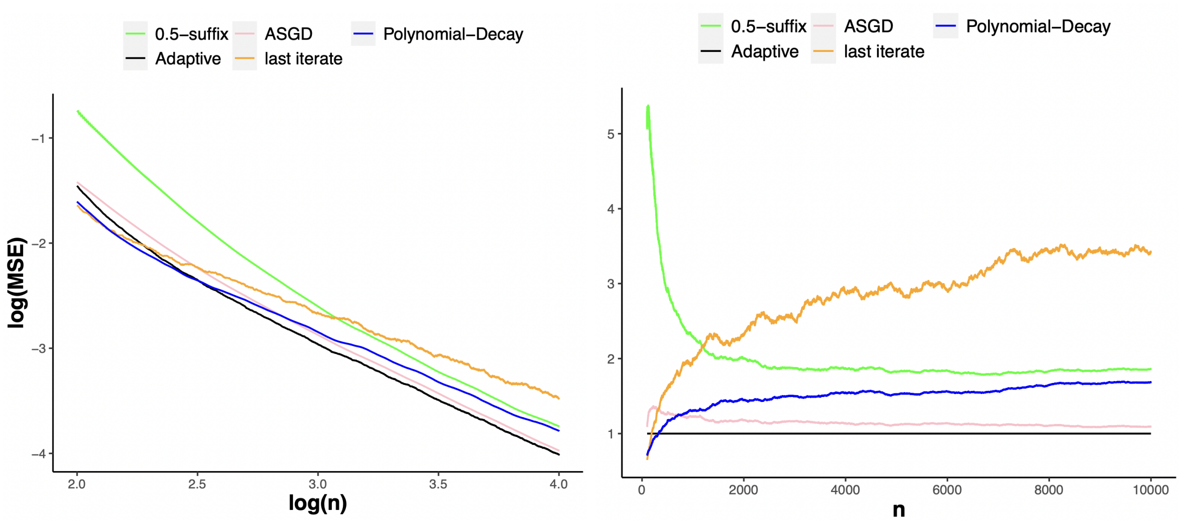

Figure 2: Left: Log-log Plots for MSE. Right: The Ratio of MSE Between Different Averaging Schemes and Adaptive Weighted Averaging at Each Step.

5.2 Non-asymptotic Performance of Different Averaging Schemes

In this section, we will show that the adaptive weighted averaging scheme in (18) has a good non-asymptotic performance in different cases.

Linear model: optimal MSE. We first validate the optimality in terms of finite sample MSE in the linear regression model, which is a generalization of the mean estimation model. The linear regression model we examine employs identical simulation settings as described in Section 5.1. Here the step size $\eta_i = i^{-0.8}$, and all the measurements are averaged over 400 independent runs. . We present coordinate-averaged MSE at certain steps in Figure 2 and include more detailed results in the appendix. We can see that the adaptive weighted averaging outperforms other averaging schemes. It indicates that adaptive weighted SGD has the potential to beat ASGD in not only the mean estimation model but also a more general optimization problem.

Expectile regression: trend of optimal weight. Expectiles, proposed by Newey and Powell (1987), is closely associated with the commonly adopted measures: Value at Risk (VaR) and Conditional Expected Shortfall (CES), which have important applications in finance and risk management (Efron, 1991; Taylor, 2008). We discuss expectile regression as it has a non-smooth loss function, $F(x) = \mathbb{E}_{y\sim\Pi}(|\rho - 1_{\{y<x\}}|(y - x)^2), 0 < \rho < 1$. For the expectile estimation problem, the optimal weights in Section 4 do not have a closed form solution. So we use Monte-Carlo simulation to numerically compute the inverse of the covariance matrix

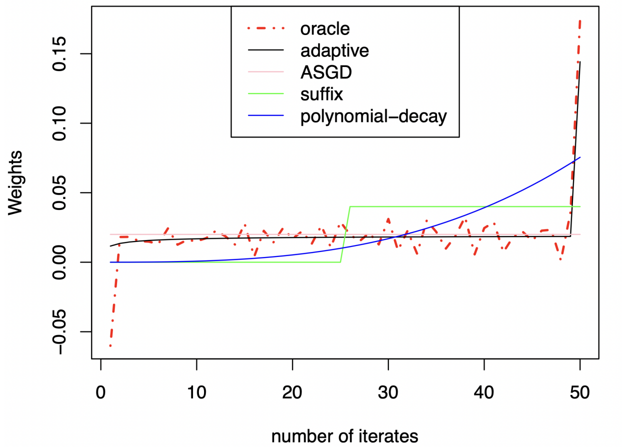

Figure 3: Comparison of Different Weighting Schemes under the Expectile Regression Model.

and obtain the oracle weights based on (14). We then compare these oracle weights with all the weighting schemes that we studied previously with $\rho = 0.8$ and $\eta_i = i^{-0.505}$. The weights for each SGD iterate are plotted in Figure 3 with the total iteration $n = 50$. The most remarkable feature of the oracle weight is its *highest* weight assigned to the last iterate. The adaptive weight we proposed in Section 4 is able to capture this characteristic, while the other averaging schemes fail to recover it. This observation shows that our adaptive weighted averaging scheme is also promising for the non-smooth optimization problem as it aligns the closest with the trend of the oracle weights.

5.3 Performance in High-dimensional Settings

We conduct further experiments on linear regression with $d = 100$ and $d = 500$ to assess the finite sample performance of weighted averaging in the high-dimensional regime. Each coordinate of a_i is drawn independently from a zero-mean normal distribution with standard deviation $\sigma = 0.1$, and x^* is set as an arithmetic sequence from 0 to 1 with length d. Other experimental settings and parameters identical to those described previously. We evaluate the mean squared error (MSE) at step $n = 5000$ for $d = 100$ and $n = 100000$ for $d = 500$, averaging the results over 500 independent replications. As demonstrated in Table 1, the adaptive weighting achieves the lowest estimation error and exhibits enhanced stability compared to alternative weighting schemes.

In the appendix, we also present the simulation results of statistical inference, where the empirical coverage rates of confidence intervals constructed by different weighting methods are reported. See Section D for details.

Table 1: Comparison of MSE Across Different Averaging Schemes. Results are Presented as MSE (Standard Deviation).

Method	$d = 100$	$d = 500$
Adaptive	**0.0451** (0.018)	**0.074** (0.102)
ASGD	0.0930 (0.077)	0.5977 (1.285)
Suffix	0.0505 (0.042)	0.1152 (0.294)
Poly-decay	0.0458 (0.033)	0.1156 (0.336)

6 DISCUSSION

This article establishes the foundational asymptotic theory for the general weighted averaged SGD algorithm with a novel sufficient condition for the CLT. We provide a comprehensive study of the limiting distribution of the weighted averaged SGD, systematically bridging uncertainty quantification with online learning. Beyond asymptotic results, we examine the linear model to derive an adaptively weighted averaging scheme designed to minimize the finite-sample MSE, and propose a recursive weighting algorithm that maintains the CLT guarantee with optimal asymptotic covariance.

Several promising directions remain for future research. Quantifying the non-asymptotic Gaussian approximation of the weighted averaged SGD is highly desirable. This could be achieved by applying existing frameworks of Berry-Esseen type theorems (Shao and Zhang, 2022; Wei et al., 2025; Sheshukova et al., 2025) and mean-squared Gaussian approximation (Zhu et al., 2024). In addition, the recent work of Davis et al. (2024) proved the CLT for standard ASGD in certain non-smooth settings. It would be valuable to leverage their approach to generalize our results for the weighted averaged SGD.

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

Anastasiou, A., Balasubramanian, K., and Erdogdu, M. A. (2019). Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Springer.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Boyer, C. and Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972.

Chen, H., Lu, W., and Song, R. (2021). 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. *Annals of Statistics*, 48(1):251–273.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*. Springer Texts in Statistics. Springer, 3rd edition.

Chung, K. L. (1954). On a Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 25(3):463 – 483.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*.

Davis, D., Drusvyatskiy, D., and Jiang, L. (2024). Asymptotic normality and optimality in nonsmooth stochastic approximation. *The Annals of Statistics*, 52(4):1485–1508.

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. (2021). From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49).

Dippon, J. and Renz, J. (1997). Weighted means in stochastic approximation of minima. *SIAM Journal on Control and Optimization*, 35(5):1811–1827.

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

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332.

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.

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. (2019). Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

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.

Lai, T. L. (2003). Stochastic approximation. *Annals of Statistics*, 31(2):391–406.

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*.

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

Li, J., Lou, Z., Richter, S., and Wu, W. B. (2024). The stochastic gradient descent from a nonlinear time series perspective. *preprint*.

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

Li, X., Liang, J., Chang, X., and Zhang, Z. (2022b). Statistical estimation and online inference via local sgd. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1613–1661. PMLR.

Liang, T. and Su, W. J. (2019). Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):431–456.

Lou, Z., Zhu, W., and Wu, W. B. (2022). Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22.

McLeish, D. L. (1976). Functional and random central limit theorems for the robbins-munro process. *Journal of Applied Probability*, 13(1):148–154.

Mokkadem, A. and Pelletier, M. (2006). Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. (2020). On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*.

Needell, D., Ward, R., and Srebro, N. (2014). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*.

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.

Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547.

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

Pflug, G. C. (1986). Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666.

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

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405.

Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*.

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*.

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

Toulis, P. and Airoldi, E. M. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics*, 45(4):1694–1727.

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*.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294 – 2320.

Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. (2021). An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. In *Advances in Neural Information Processing Systems*.

Zhang, S., Choromanska, A. E., and LeCun, Y. (2015). Deep learning with elastic averaging sgd. *Advances in Neural Information Processing Systems*.

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

1. 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]

2. 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]

3. 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]

4. 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.

5. 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.

General Weighted Averaging in Stochastic Gradient Descent: CLT and Adaptive Optimality: Supplementary Materials

In the appendix, we provide the notations and technical lemmas in Section A, the proofs of all main paper results in Section B—in particular, the proof of Theorem 2.3 in Section B.1, the proof of Corollary 2.5 in Section B.2, and the proof of Theorem 2.7 in Section B.4. Supporting lemmas required for the proofs of the main results are given in Section C, and additional experimental results are presented in Section D.

A NOTATIONS AND TECHNICAL LEMMAS

We first introduce some notation. For positive sequences $a_{n,i}$ and $b_{n,i}$ with $0 \le i \le n$ (or simply a_n and b_n), we write $a_{n,i} \lesssim b_{n,i}$ if there exists a constant C such that $a_{n,i} \le C b_{n,i}$ for all $0 \le i \le n$. We write and $a_{n,i} \gtrsim b_{n,i}$ if $b_{n,i} \lesssim a_{n,i}$, and $a_{n,i} \asymp b_{n,i}$ if both $a_{n,i} \lesssim b_{n,i}$ and $a_{n,i} \gtrsim b_{n,i}$.

Before the proof of main results, we present some technical lemmas and defer the proof to section C. The first one is a lemma in Polyak and Juditsky (1992):

Lemma A.1. *Choose the step size as $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$ and $0.5 < \alpha < 1$. For a real symmetric positive definite matrix A, define a matrices sequence Y_i^k: $Y_i^i = \boldsymbol{I}$ and for any $k > i$:*

$$Y_i^k = \prod_{j=i+1}^{k} (\boldsymbol{I} - \eta_j A).$$

We also define $\bar{Y}_i^n$ and ϕ_i^n as follows,

$$\bar{Y}_i^n = \eta_i \sum_{k=i}^{n} Y_i^k, \ \ n \ge i,$$

$$\phi_i^n = A^{-1} - \bar{Y}_i^n.$$

Then $\exists\ 0 < K < \infty$ such that $\forall\ j$ and $i \ge j$

$$|\phi_i^n| \le K,$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} |\phi_i^n| = 0.$$

Lemma A.1 is a simple reduction of Lemma 1 in Polyak and Juditsky (1992). The term Y_i^k appears frequently in the explicit form of weighted SGD solutions. The next lemma further investigates the order of some quantities related to it. We defer the proof to section C.

Lemma A.2. *Let $\lambda > 0$ be a constant. Define a real sequence $\{Y_{(\lambda)}{}_i^k\}$ with*

$$Y_{(\lambda)}{}_i^k = \begin{cases} 1 & \text{if } i = k, \\ \displaystyle\prod_{j=i+1}^{k} (1 - \lambda\eta_j) & \text{if } i < k. \end{cases} \tag{21}$$

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

1. $$|Y_{(\lambda)}{}_i^k| \asymp \exp\left\{\frac{\lambda\eta}{1-\alpha}\left(i^{1-\alpha} - k^{1-\alpha}\right)\right\} \le \exp\left\{-\lambda\eta i^{-\alpha}(k-i)\right\}. \tag{22}$$

2. *For $\beta, \gamma > 0$,*

$$\sum_{i=1}^{k} \exp\left(\beta i^{1-\alpha}\right) i^{-\gamma\alpha} \asymp \exp\left(\beta k^{1-\alpha}\right) k^{-(\gamma-1)\alpha}, \tag{23}$$

which implies

$$\sum_{i=1}^{k} |Y_{(\lambda)}{}_i^k|^\beta |i^{-\alpha}|^\gamma \asymp k^{-(\gamma-1)\alpha}. \tag{24}$$

3. *For any $\beta > 0$,*

$$\sum_{i=j}^{k} \exp(-\beta i^{1-\alpha}) \lesssim \exp(-\beta j^{1-\alpha}) j^\alpha.$$

Here and in the sequel, we treat x_0 as fixed. We also invoke Lemma 3.1 in Zhu et al. (2023), which have the following conclusion: for some constant C,

$$\|x_n - x^*\| \le C n^{-\alpha/2}, \ |x_n - x^*\|^2 \le C n^{-\alpha}.$$

B PROOF OF MAIN RESULTS

B.1 Proof of Theorem 2.3

Proof. We decompose the SGD iterates as follows,

$$x_i - x^* = x_{i-1} - x^* - \eta_i A(x_{i-1} - x^*) - \eta_i R_i - \eta_i D_i - \eta_i \nabla f(x^*, \xi_i), \tag{25}$$

where

$$R_i = \nabla F(x_{i-1}) - A(x_{i-1} - x^*), \ \|R_i\| = \mathcal{O}(|x_{i-1} - x^*|^2),$$
$$D_i = \nabla f(x_{i-1}, \xi_i) - \nabla F(x_{i-1}) - \nabla f(x^*, \xi_i), \ \|D_i\|^2 = \mathcal{O}(|x_{i-1} - x^*|^2)$$

due to Taylor's expansion and stochastic Lipschitz property. Notice that $\nabla F(x^*) = 0$ here. Denote $\delta_i = x_i - x^*$ the error sequence. The weighted averaged error sequence $\tilde{\delta}_n = \sum_{i=1}^n w_{n,i}\delta_i$ takes the following form,

$$\tilde{\delta}_n = \sum_{i=1}^{n} w_{n,i} \prod_{j=1}^{i} (\mathbf{I}_d - \eta_j A)\delta_0 - \sum_{i=1}^{n} \Phi_{n,i}(R_i + D_i + \nabla f(x^*, \xi_i)) \tag{26}$$

$$= \eta_1^{-1}(\mathbf{I}_d - \eta_1 A)\Phi_{n,1}\delta_0 - \sum_{i=1}^{n} \Phi_{n,i}(R_i + D_i + \nabla f(x^*, \xi_i)). \tag{27}$$

It is clear that $|V_n^{-1/2}\eta_1^{-1}(\mathbf{I}_d - \eta_1 A)\Phi_{n,1}| \to 0$ by Condition (6). Notice that $\{D_i\}$ is a martingale difference sequence. For any $\nu > 0$, there exists an integer K such that $2(K+1)^{-\alpha} < \nu$. For any $n > K$, we decompose the martingle difference error term by K,

$$\begin{aligned}
\|V_n^{-1/2} \sum_{i=1}^{n} \Phi_{n,i} D_i\|^2 &= \|V_n^{-1/2} \sum_{i=1}^{K} \Phi_{n,i} D_i\|^2 + \|V_n^{-1/2} \sum_{i=K+1}^{n} \Phi_{n,i} D_i\|^2 \\
&\le \sum_{i=1}^{K} |V_n^{-1/2}\Phi_{n,i}|^2 \|D_i\|^2 + \sum_{i=K+1}^{n} |V_n^{-1/2}\Phi_{n,i}|^2 \|D_i\|^2 \\
&\lesssim \sum_{i=1}^{K} |V_n^{-1/2}\Phi_{n,i}|^2 i^{-\alpha} + \sum_{i=K+1}^{n} |V_n^{-1/2}\Phi_{n,i}|^2 (K+1)^{-\alpha} \\
&\lesssim \sum_{i=1}^{K} |V_n^{-1/2}\Phi_{n,i}|^2 i^{-\alpha} + (K+1)^{-\alpha} \sum_{i=K+1}^{n} |V_n^{-1/2}\Phi_{n,i} S^{1/2}|^2.
\end{aligned}$$

Since K is fixed, Condition (6) implies that there exists $N > 0$ such that for all $n \geq N$, the first term is smaller than ν. For the second term, notice that

$$\begin{aligned}\sum_{i=K+1}^{n} |V_n^{-1/2}\Phi_{n,i}S^{1/2}|^2 &= \sum_{i=K+1}^{n} |V_n^{-1/2}\Phi_{n,i}S\Phi_{n,i}V_n^{-1/2}| \\ &= \sum_{i=K+1}^{n} |V_n^{-1}\Phi_{n,i}S\Phi_{n,i}| \leq \mathbf{trace}(\mathbf{I}_d) = d\end{aligned}$$

because $\sum_{i=1}^{n} V_n^{-1}\Phi_{n,i}S\Phi_{n,i} = \mathbf{I}_d$ and the largest eigenvalue of a positive definite matrix is bounded by its trace. To see this, consider positive-definite matrices M_i with $\sum_{i=1}^{n} M_i = I_d$, since $|M_i|$ is its largest eigenvalue and $trace(M_i)$ equals the sum of all eigenvalues of M_i, we have $|M_i| \leq trace(X_i)$. Summing both sides $\sum_{i=1}^{n} |M_i| \leq \sum_{i=1}^{n} trace(M_i) = trace(\sum_{i=1}^{n} M_i) = trace(I_d) = d$.
Finally, we have

$$\|V_n^{-1/2}\sum_{i=1}^{n}\Phi_{n,i}D_i\|^2 \lesssim \nu + d\nu.$$

As a result, the martingale difference error term converges to 0.

The other error term also vanishes since

$$\begin{aligned}\|V_n^{-1/2}\sum_{i=1}^{n}\Phi_{n,i}R_i\| &\leq \sum_{i=1}^{n}|V_n^{-1/2}\Phi_{n,i}|\|R_i\| \\ &\lesssim \sum_{i=1}^{n}|V_n^{-1/2}\Phi_{n,i}|i^{-\alpha} \\ &\leq \sum_{i=1}^{K}|V_n^{-1/2}\Phi_{n,i}|i^{-\alpha} + \sum_{i=K+1}^{n}|V_n^{-1/2}\Phi_{n,i}|i^{-\alpha}.\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(\sum_{i=K+1}^{n}|V_n^{-1/2}\Phi_{n,i}|i^{-\alpha})^2 \leq (\sum_{i=K+1}^{n}|V_n^{-1/2}\Phi_{n,i}|^2)\sum_{i=K+1}^{n} i^{-2\alpha} \lesssim d(K+1)^{1-2\alpha}.$$

Then we can use the identical argument as before to show that $\mathbb{E}|V_n^{-1/2}\sum_{i=1}^{n}\Phi_{n,i}R_i| \to 0$.

The main term that does not vanish is $V_n^{-1/2}\sum_{i=1}^{n}\Phi_{n,i}\nabla f(x^*,\xi_i)$. To show it converges to the standard normal distribution, it suffices to prove that the sequence satisfies the Lindeberg condition, i.e., for any $\nu > 0$,

$$\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\mathbb{1}_{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\geq\nu}\} = 0. \tag{28}$$

By Assumption 2.2, $\|\nabla f(x^*,\xi_i)\|^2 < \infty$. Let $M_n = \max_{1\leq i\leq n}|V_n^{-1/2}\Phi_{n,i}|^2$, we have

$$\begin{aligned}&\mathbb{E}\{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\mathbb{1}_{|V_n^{-1/2}\Phi_{n,i}|^2|\nabla f(x^*,\xi_i)|^2\geq\nu}\} \\ \leq\ &\mathbb{E}\{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\mathbb{1}_{M_n|\nabla f(x^*,\xi_i)|^2\geq\nu}\} \\ \leq\ &|V_n^{-1/2}\Phi_{n,i}|^2\mathbb{E}\{|\nabla f(x^*,\xi_i)|^2\mathbb{1}_{M_n|\nabla f(x^*,\xi_i)|^2\geq\nu}\},\end{aligned}$$

and $\mathbb{E}\{|\nabla f(x^*,\xi_i)|^2\mathbb{1}_{M_n|\nabla f(x^*,\xi_i)|^2\geq\nu}\} \to 0$. The reason is that by Condition (6), the indicators in (28) go to 0.

And the Dominated convergence theorem guarantees the expectation's convergence to 0. Then it yields

$$\begin{aligned}
&\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\mathbb{1}_{|V_n^{-1/2}\Phi_{n,i}\nabla f(x^*,\xi_i)|^2\geq\nu}\}\\
\leq&\sum_{i=1}^{n}|V_n^{-1/2}\Phi_{n,i}|^2\lim_{n\to\infty}\mathbb{E}\{\nabla f(x^*,\xi_i)|^2\mathbb{1}_{M_n|\nabla f(x^*,\xi_i)|^2\geq\nu}\}\\
\leq& d\lim_{n\to\infty}\mathbb{E}\{|\nabla f(x^*,\xi_i)|^2\mathbb{1}_{M_n|\nabla f(x^*,\xi_i)|^2\geq\nu}\}=0.
\end{aligned}$$

So (28) holds and the conclusion is proved.

To show that (7) implies Condition (6), we construct a counterexample such that CLT does not hold when Condition (6) fails. Consider

$$f(x,\xi)=\frac{1}{2}|\xi-A^{\frac{1}{2}}x|^2=\frac{1}{2}x^TAx-\xi^TA^{\frac{1}{2}}x+\frac{1}{2}\xi^T\xi.$$

where $\xi\in\mathbb{R}^d$ is a mean zero random variable with $\mathbb{E}(\xi\xi^T)=A^{-\frac{1}{2}}SA^{-\frac{1}{2}}$. We have $F(x)=(x^TAx+\mathbb{E}\xi^T\xi)/2$ and $x^*=0$. The SGD iterates take the form

$$x_i=x_{i-1}-\eta_i(Ax_{i-1}-A^{\frac{1}{2}}\xi_i)=(\mathbf{I}_d-\eta_iA)x_{i-1}+\eta_iA^{\frac{1}{2}}\xi_i. \tag{29}$$

Let $z_i=A^{\frac{1}{2}}\xi_i$. The standardized weighted averaged SGD is formulated as

$$\tilde{x}_n=\eta_1^{-1}V_n^{-1/2}(\mathbf{I}_d-\eta_1A)\Phi_{n,1}x_0+V_n^{-\frac{1}{2}}\sum_{i=1}^{n}\Phi_{n,i}z_i,$$

where x_0 is fixed. So we only need to consider the distribution of $V_n^{-\frac{1}{2}}\sum_{i=1}^n\Phi_{n,i}z_i$. When Condition (6) fails, there exists a subsequence n_k such that

$$\lim_{k\to\infty}\max_{1\leq i\leq n_k}|V_{n_k}^{-\frac{1}{2}}\Phi_{n_k,i}|=c>0.$$

Without loss of generality, we can assume the original sequence has this limit and the maximum is taken on $i=1$. Since the matrix sequence $V_n^{-1/2}\Phi_{n,1}$ is uniformly bounded in the operator norm, it also has a convergence subsequence. So we can further assume that $V_n^{-1/2}\Phi_{n,1}$ converges to some matrix Γ. Since matrix norms are continuous, we have $|\Gamma|=c$. We can choose the unit eigenvector of the largest eigenvalue of Γ and denote it as v. Since the CLT in condition 7 holds, we have

$$v^TV_n^{-\frac{1}{2}}\Phi_{n,1}z_1+\Theta_n\to\mathcal{N}(0,1),\ \text{where}\ \Theta_n=v^T\sum_{i=2}^{n}V_n^{-\frac{1}{2}}\Phi_{n,i}z_i.$$

Moreover, $v^TV_n^{-\frac{1}{2}}\Phi_{n,1}z_1\perp\!\!\!\perp\Theta_n$ and $v^TV_n^{-\frac{1}{2}}\Phi_{n,1}z_1\to cv^Tz_1$. Denote the characteristic function of $v^TV_n^{-\frac{1}{2}}\Phi_{n,1}z_1$ and Θ_n as $\phi_n(t)$ and $\psi_n(t)$, we have

$$\phi_n(t)\psi_n(t)=e^{\frac{-t^2}{2}},$$

and hence

$$\psi_n(t)=\frac{e^{\frac{-t^2}{2}}}{\phi_n(t)}\to\frac{e^{\frac{-t^2}{2}}}{\mathbb{E}e^{itcv^Tz_1}}.$$

The limit of characteristic functions is still a characteristic function, which implies that Θ_n has a limiting distribution Θ_∞. Finally, we have that the independent sum $cv^Tz_1+\Theta_\infty$ follows a standard Gaussian distribution. By the Lévy-Cramér theorem (cf Chow and Teicher (1997)), cv^Tz_1 must also be a Gaussian variable as well. This contradicts condition 7 which requires CLT to hold for non-Gaussian z_1. So we have proved the necessity. $\square$

B.2 Proof of Corollary 2.5

Proof. Define the polynomial $p_{n,i}(x) = \eta_i \prod_{j=i+1}^{n}(1-\eta_j x)$. To prove

$$\max_{1\le i\le n} |V_n^{-\frac{1}{2}}\Phi_{n,i}| \to 0,$$

where $V_n = \sum_{i=1}^{n} \Phi_{n,i} S \Phi_{n,i}$ and $\Phi_{n,i} = \eta_i \prod_{j=i+1}^{n}(\mathbf{I}_d - \eta_j A) = p_{n,i}(A)$, it suffices to show that

$$\max_{1\le i\le n} \frac{p_{n,i}(\lambda_{\min}(A))^2}{\sum_{i=1}^{n} p_{n,i}(\lambda_{\max}(A))^2} \to 0,$$

because

$$|\Phi_{n,i}| = \lambda_{\max}(\Phi_{n,i}) = p_{n,i}(\lambda_{\min}(A)),$$

and

$$\lambda_{\min}(V_n) \ge \lambda_{\min}(S)\lambda_{\min}\left(\sum_{i=1}^{n}\Phi_{n,i}^2\right) \ge \lambda_{\min}(S)\sum_{i=1}^{n} p_{n,i}(\lambda_{\max}(A))^2,$$

hence

$$|V_n^{-\frac{1}{2}}| \le \frac{1}{\sqrt{\lambda_{\min}(S)\sum_{i=1}^{n} p_{n,i}(\lambda_{\max}(A))^2}}.$$

By Lemma A.2, we have

$$\begin{aligned}
\frac{p_{n,i}(\lambda_{\min}(A))^2}{\sum_{i=1}^{n} p_{n,i}(\lambda_{\max}(A))^2} &\asymp i^{-2\alpha} n^{\alpha} \exp\left\{\frac{2\lambda_{\min}(A)\eta}{1-\alpha}\left(i^{1-\alpha} - n^{1-\alpha}\right)\right\} \\
&\le i^{-2\alpha} n^{\alpha} \exp\left\{-\lambda_{\min}(A)\eta n^{-\alpha}(n-i)\right\}.
\end{aligned}$$

It is clear that $\max_{1\le i\le n} i^{-2\alpha} n^{\alpha} \exp\{-\lambda_{\min}(A)\eta n^{-\alpha}(n-i)\} \to 0$. So the CLT for the last iterate of SGD holds. Then it suffices to show that $\eta_n^{-1}V_n \to V$. Notice that

$$\begin{aligned}
V_{n+1} &= \eta_{n+1}^2 S + \sum_{k=1}^{n} \Phi_{n,i}(\mathbf{I}_d - \eta_{n+1}A) S \Phi_{n,i}(\mathbf{I}_d - \eta_{n+1}A) \\
&= (\mathbf{I}_d - \eta_{n+1}A) V_n (\mathbf{I}_d - \eta_{n+1}A) + \eta_{n+1}^2 S.
\end{aligned}$$

Let $\Gamma_n = \eta_n^{-1}V_n$. Plug Γ_n and $AV + VA = S$ into the formula above, we have

$$\begin{aligned}
&\Gamma_{n+1} - V \\
&= (\mathbf{I}_d - \eta_{n+1}A)(\Gamma_n - V)(\mathbf{I}_d - \eta_{n+1}A) + \eta_{n+1}^2 AVA + (\eta_{n+1}^{-1} - \eta_n^{-1})(\mathbf{I}_d - \eta_{n+1}A)V_n(\mathbf{I}_d - \eta_{n+1}A).
\end{aligned}$$

For n large enough, the matrix norm of the last term of the left-hand side can be bounded as

$$\begin{aligned}
|(\eta_{n+1}^{-1} - \eta_n^{-1})(\mathbf{I}_d - \eta_{n+1}A)V_n(\mathbf{I}_d - \eta_{n+1}A)| &\le \eta[(n+1)^{\alpha} - n^{\alpha}]|V_n| \\
&\lesssim \eta\alpha n^{\alpha-1}\sum_{i=1}^{n}|\Phi_{n,i}|^2|S| \\
&\lesssim n^{\alpha-1}\sum_{i=1}^{n}\eta_i^2\prod_{j=i+1}^{n}|(\mathbf{I}_d - \eta_j A)|^2 \\
&\lesssim n^{\alpha-1}\sum_{i=1}^{n}\eta_i^2\prod_{j=i+1}^{n}(1-\eta_j\lambda_{\min}(A))^2 \\
&\asymp n^{\alpha-1}n^{-\alpha} = n^{-1},
\end{aligned}$$

where the last step is from Lemma A.2. Notice that $\eta_{n+1}^2 \lesssim n^{-1}$. Hence there exists a universal constant C such that

$$|\Gamma_{n+1} - V| \le (1-\eta_{n+1}\lambda_{\min}(A))^2|\Gamma_n - V| + \frac{C}{n}.$$

Recursively updating the inequality we get

$$
\begin{aligned}
|\Gamma_n - V| &\le \sum_{i=1}^{n} \frac{C}{i} \prod_{j=i+1}^{n} (1 - \frac{\eta\lambda_{\min}(A)}{j^\alpha})^2 + \prod_{j=1}^{n} (1 - \frac{\eta\lambda_{\min}(A)}{j^\alpha})^2 |\Gamma_0 - V| \\
&\lesssim \sum_{i=1}^{n} (i^{-\alpha})^{1/\alpha} \prod_{j=i+1}^{n} (1 - \frac{\eta\lambda_{\min}(A)}{j^\alpha})^2 + \prod_{j=1}^{n} (1 - \frac{\eta\lambda_{\min}(A)}{j^\alpha})^2 \\
&\asymp n^{\alpha-1} + \prod_{j=1}^{n} (1 - \frac{\eta\lambda_{\min}(A)}{j^\alpha})^2 \to 0.
\end{aligned}
$$

The last step is also from Lemma A.2. So we have proved that $\eta_n^{-1} V_n \to V$. As a result,

$$
n^{\frac{\alpha}{2}} (x_n - x^*) \xrightarrow{D} \mathcal{N}(0, \eta V),
$$

□

B.3 A CLT with General Weights and General Form of Learning Rates

In this section, we shall present a CLT with general weights and a general form of learning rates. Proposition B.1 below provides an explicit bound for the MSE of x_n.

Proposition B.1. *Let Assumptions 2.1 and 2.2 be satisfied. Define*

$$
v_n = (1 - 2\eta_n\mu + 2\eta_n^2 L_2^2) v_{n-1} + 2\eta_n^2 \sigma^2(x^*),\ n \ge 1,\ v_0 = |x_0 - x^*|^2. \tag{30}
$$

Then $\mathbb{E}|x_n - x^*|^2 \le v_n$*. In particular, if* $\eta_n \asymp n^{-\beta}$*,* $\beta \in (0,1)$*, then* $v_n \lesssim n^{-\beta}$*.*

Proof. Recall the decomposition of SGD (25) and take the norm on both sides,

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, since $\mathbb{E}[\epsilon_i | \delta_{i-1} - \eta_i \nabla F(x_{i-1})] = 0$, we have

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

Further by Assumption 2.1 and 2.2, we have

$$
\begin{aligned}
\|\delta_i\|^2 &\le \|\delta_{i-1} - \eta_i \nabla F(x_{i-1})\|^2 + \eta_i^2 \|\epsilon_i\|^2 \\
&\le (1 - 2\eta_i\mu) \|\delta_{i-1}\|^2 + 2\eta_i^2 (\|\epsilon_i^*\|^2 + L_2^2 \|\delta_{i-1}\|^2) \\
&\le (1 - 2\eta_i\mu + 2\eta_i^2 L_2^2) \|\delta_{i-1}\|^2 + 2\eta_i^2 \sigma^2(x^*).
\end{aligned}
$$

Then by definition of v_n we have $\mathbb{E}|\delta_n|^2 = \mathbb{E}|x_n - x^*|^2 \le v_n$. If $\eta_i \asymp i^{-\beta}$, we have

$$
\begin{aligned}
v_i &\lesssim \prod_{k=1}^{i} (1 - ck^{-\beta}) v_0 + 2\sigma^2(x^*) \sum_{j=1}^{i} j^{-2\beta} \prod_{k=j+1}^{i} (1 - ck^{-\beta}) \\
&\asymp {Y_{(c)}}_0^i + \sum_{j=1}^{i} {Y_{(c)}}_j^i j^{-2\beta} \asymp i^{-\beta}.
\end{aligned}
$$

The last two steps are from Lemma A.2. □

Let $\lambda_1 \leq \ldots \leq \lambda_d$ be eigenvalues of A. For $j \geq 1$ define

$$b_j = \lambda_{\max}(\mathbf{I}_d - \eta_j A) = \max_{1 \leq l \leq d} |1 - \eta_j \lambda_j| = \max(|1 - \eta_j \lambda_1|, |1 - \eta_j \lambda_d|). \tag{31}$$

For small η_j, we have $b_j = 1 - \eta_j \lambda_1 < 1$, which plays the role of contraction factor.

Theorem B.2. *Consider the SGD (1). Let Assumptions 2.1 and 2.2 be satisfied. Assume that weights $w_{n,i}$ in (4) satisfy $\sum_{i=1}^n w_{n,i} = 1$. Let v_n be defined in (30) and $\chi_k = \prod_{i=1}^k b_i$. Assume that*

$$\lim_{n \to \infty} \lambda_{\max}(V_n^{-1}) \sum_{i=1}^n v_{i-1} \left(\sum_{k=i}^n |w_{n,k}| \frac{\chi_k}{\chi_i} \eta_j \right)^2 = 0 \tag{32}$$

and

$$\lim_{n \to \infty} \lambda_{\max}(V_n^{-1/2}) \sum_{j=1}^n |w_{n,j}| \sum_{i=1}^j \frac{\chi_j}{\chi_i} \eta_i v_{i-1} = 0. \tag{33}$$

Further assume (6). Then we have the following central limit theorem: for any starting point x_0,

$$V_n^{-1/2}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{I}_d).$$

Proof. Recall the decomposition in 26. The same argument in B.1 can be applied here to show that $V_n^{-1/2} \sum_{i=1}^n \Phi_{n,i} \nabla f(x^*, \xi_i)$ converges to the standard normal distribution. And we still have $|V_n^{-1/2} \eta_1^{-1} (\mathbf{I}_d - \eta_1 A) \Phi_{n,1}| \to 0$ by Condition (6). Recall the definition of

$$\Phi_{n,i} = \eta_i \sum_{k=i}^n w_{n,k} \prod_{j=i+1}^k (\mathbf{I}_d - \eta_j A).$$

We can bound its operator norm by

$$|\Phi_{n,i}| \leq \eta_i \sum_{k=i}^n |w_{n,k}| \frac{\chi_k}{\chi_j}.$$

As a result,

$$\begin{aligned} \|V_n^{-1/2} \sum_{i=1}^n \Phi_{n,i} D_i\|^2 &\leq \sum_{i=1}^n |V_n^{-1/2} \Phi_{n,i}|^2 \|D_i\|^2 \\ &\lesssim \lambda_{\max}(V_n^{-1}) \sum_{i=1}^n \eta_i^2 \Big(\sum_{k=i}^n |w_{n,k}| \frac{\chi_k}{\chi_j} \Big)^2 v_{i-1} \to 0 \end{aligned}$$

by condition 32. Moreover,

$$\begin{aligned} \|V_n^{-1/2} \sum_{i=1}^n \Phi_{n,i} R_i\| &\leq \sum_{i=1}^n |V_n^{-1/2} \Phi_{n,i}| \|R_i\| \\ &\lesssim \sum_{i=1}^n \lambda_{\max}(V_n^{-1/2}) |\Phi_{n,i}| v_{i-1} \\ &\leq \sum_{i=1}^n \lambda_{\max}(V_n^{-1/2}) \eta_i \sum_{k=i}^n |w_{n,k}| \frac{\chi_k}{\chi_j} v_{i-1} \\ &= \lambda_{\max}(V_n^{-1/2}) \sum_{k=1}^n |w_{n,k}| \sum_{i=1}^k \frac{\chi_k}{\chi_j} \eta_i v_{i-1} \to 0 \end{aligned}$$

by condition 33. So the conclusion follows. $\square$

B.4 Proof of Theorem 2.7

B.4.1 High level summary

We first present a sketch of the proof of Theorem 2.7.

The error sequence $\delta_i = x_i - x^*$ takes the following form

$$\delta_i = \delta_{i-1} - \eta_i \nabla F(x_{i-1}) + \eta_i \epsilon_i, \; i \geq 1, \tag{34}$$

where $\epsilon_i = \nabla F(x_{i-1}) - \nabla f(x_{i-1}, \xi_i)$. Since $\nabla F(x^*) = 0$, by Taylor's expansion of F around x^* we have $\nabla F(x_n) \approx A\delta_n$, which inspires the idea to approximate the general SGD sequence with a corresponding linear sequence as follows:

$$\delta'_i = (I - \eta_i A)\delta'_{i-1} + \eta_i \epsilon_i, \; \delta'_0 = \delta_0. \tag{35}$$

The following lemma shows that the weighted average of the linear sequence δ'_i exhibits the asymptotic normality.

Lemma B.3. *Let δ'_i be defined in (35). Then under the settings in Theorem 2.7, for $\tilde{\delta}'_n = \sum_{i=1}^n w_{n,i}\delta'_i$ we have*

$$\sqrt{n}\tilde{\delta}'_n \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}),$$

where $w = \lim_{n\to\infty} n \sum_{i=1}^n (w_{n,i})^2$, $A = \nabla^2 F(x^*)$, and $S = \mathbb{E}([\nabla f(x^*, \xi)][\nabla f(x^*, \xi)]^T)$.

To prove Lemma B.3, we decompose $\sqrt{n}\tilde{\delta}'_n$ into four terms:

$$\begin{aligned} \sqrt{n}\tilde{\delta}'_n = {} & \sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\epsilon_i + \sqrt{n}\sum_{i=1}^n w_{n,i}Y_0^i\delta_0 \\ & + \sqrt{n}\sum_{i=1}^n w_{n,i}a_i^n\epsilon_i + \sqrt{n}\sum_{i=1}^n b_i^n\epsilon_i, \end{aligned} \tag{36}$$

where $a_i^n = \sum_{k=i}^n Y_i^k\eta_i - A^{-1}$, $b_i^n = \sum_{k=i+1}^n (w_{n,k} - w_{n,i})Y_i^k\eta_i$ and

$$Y_i^k = \prod_{j=i+1}^k (\mathbf{I} - \eta_j A), k > i, Y_i^i = \mathbf{I}.$$

The last three terms in (36) vanish as n goes to infinity. The first term $\sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\epsilon_i$ is a linear combination of martingale differences and the following lemma shows that it is asymptotically normal.

Lemma B.4 (Martingale difference asymptotic normality)**.** *Under the settings in Theorem 2.7,*

$$\sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\epsilon_i \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}).$$

Once Lemma B.3 is established, we can prove Theorem 2.7 using the linear approximation technique.

B.4.2 Detailed proof of theorem 2.7

Proof. Recall the error sequence of SGD iterates $\delta_n = x_n - x^*$. It also takes the form

$$\delta_n = \delta_{n-1} - \eta_n \nabla F(x_{n-1}) + \eta_n \epsilon_n.$$

The weighted averaged error sequence is $\tilde{\delta}_n = \sum_{i=1}^n w_{n,i}\delta_i$. Since $\sum_{i=1}^n w_{n,i} = 1$, we have $\tilde{\delta}_n = \tilde{x}_n - x^*$. We have also defined the linear error sequence

$$\delta'_n = \delta'_{n-1} - \eta_n A\delta'_{n-1} + \eta_n\epsilon_n, \; \delta'_0 = x_0 - x^*,$$

$$\tilde{\delta}'_n = \sum_{i=1}^n w_{n,i}\delta'_i.$$

We claim that Lemma B.3 is true, i.e.,

$$\sqrt{n}\tilde{\delta}'_n \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}),$$

then it suffices to prove that $\sqrt{n}\tilde{\delta}'_n$ and $\sqrt{n}\tilde{\delta}_n$ are asymptotically equally distributed. Let s_n be the difference between the nonlinear and linear sequence. It also takes the following recursion form:

$$\begin{aligned} s_n = \delta_n - \delta'_n &= \delta_{n-1} - \eta_n \nabla F(x_{n-1}) - (\mathbf{I} - \eta_n A)\delta'_{n-1} \\ &= (\mathbf{I} - \eta_n A)(\delta_{n-1} - \delta'_{n-1}) - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}) \\ &= (\mathbf{I} - \eta_n A)s_{n-1} - \eta_n(\nabla F(x_{n-1}) - A\delta_{n-1}). \end{aligned} \tag{37}$$

Recall the definition of Y_i^n:

$$Y_i^k = \prod_{j=i+1}^{k} (\mathbf{I} - \eta_j A), k > i, Y_i^i = \mathbf{I}.$$

We can use Y_i^n to rewrite s_n as

$$s_n = \sum_{i=1}^{n} Y_i^n \eta_i [A\delta_{i-1} - \nabla F(x_{i-1})].$$

Define the weighted average difference between the nonlinear and linear sequence:

$$\tilde{s}_n = \sum_{i=1}^{n} w_{n,i} s_i = \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{i} Y_j^i \eta_j [A\delta_{j-1} - \nabla F(x_{j-1})].$$

Note that $\sqrt{n}\tilde{\delta}_n = \sqrt{n}\tilde{s}_n + \sqrt{n}\tilde{\delta}'_n$ and $\sqrt{n}\tilde{\delta}'_n \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1})$. To prove

$$\sqrt{n}\tilde{\delta}_n = \sqrt{n}(\tilde{x}_n - x^*) \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}),$$

it is suffice to prove $\sqrt{n}\tilde{s}_n$ converges to 0 in probability.

$$\begin{aligned} |\tilde{s}_n| &\leq \sum_{i=1}^{n} w_{n,i} \sum_{j=1}^{i} |Y_j^i| \eta_j |A\delta_{j-1} - \nabla F(x_{j-1})| \\ &\lesssim \frac{1}{n} \sum_{j=1}^{n} |A\delta_{j-1} - \nabla F(x_{j-1})| \eta_j \Big(\sum_{i=j}^{n} |Y_j^i| \Big) \\ &\lesssim \frac{1}{n} \sum_{j=1}^{n} |A\delta_{j-1} - \nabla F(x_{j-1})| j^{-\alpha} \big(1 + (j+1)^{\alpha}\big) \\ &\lesssim \frac{1}{n} \sum_{j=1}^{n} |A\delta_{j-1} - \nabla F(x_{j-1})| \\ &\lesssim \frac{1}{n} \sum_{j=1}^{n} |\delta_{j-1}|^2. \end{aligned} \tag{38}$$

The second inequality is obtained by upper bounding $w_{n,i}$ and exchanging the order of summations. The third inequality comes from Lemma A.2 in Zhu et al. (2023). The last inequality is from Taylor's expansion around x^*. From Lemma 3.2 in Zhu et al. (2023) we know that

$$\mathbb{E}|\delta_{j-1}|^2 \lesssim (j-1)^{-\alpha}.$$

So there exists a constant $C > 0$ such that

$$\sum_{j=1}^{n} \frac{1}{\sqrt{j}} \mathbb{E}|\delta_{j-1}|^2 \lesssim \sum_{j=1}^{n} \frac{1}{\sqrt{j}} (j-1)^{-\alpha} \lesssim \sum_{j=1}^{n} j^{-0.5-\alpha} \leq C.$$

By Kronecker's lemma,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}|\delta_{j-1}|^2 \to 0.$$

As a result, for any fixed $h>0$,

$$\mathbb{P}(\sqrt{n}|\tilde{s}_n|>h) \le \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}|\delta_{j-1}|^2>h\Big) \le \frac{1}{\sqrt{n}h}\mathbb{E}\sum_{j=1}^{n}|\delta_{j-1}|^2 \to 0.$$

Thus we proved that $\sqrt{n}\tilde{s}_n$ converges to 0 in probability, and the theorem is proved.

□

B.4.3 Proof of Lemma B.3 and Lemma B.4

Proof of Lemma B.3. By definition of the linear error term, we have

$$\delta_n' = \prod_{i=1}^{n}(\mathbf{I}-\eta_i A)\delta_0 + \sum_{i=1}^{n}\prod_{j=i+1}^{n}(\mathbf{I}-\eta_j A)\eta_i\epsilon_i,$$

where the matrix sequence $Y_i^k,\ k\ge i$ is defined as

$$Y_i^k = \prod_{j=i+1}^{k}(\mathbf{I}-\eta_j A), k>i, Y_i^i = \mathbf{I}.$$

Here we also use the convention that $\prod_{j=n+1}^{n}(\mathbf{I}-\eta_j A)=I$. Then the weighted averaged error sequence $\tilde{\delta}_n'$ takes the form:

$$\begin{aligned}
\tilde{\delta}_n' &= \sum_{i=1}^{n}w_{n,i}\prod_{j=1}^{i}(\mathbf{I}-\eta_j A)\delta_0 + \sum_{k=1}^{n}w_{n,k}\sum_{i=1}^{k}\prod_{j=i+1}^{k}(\mathbf{I}-\eta_j A)\eta_i\epsilon_i \\
&= \sum_{i=1}^{n}w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n}\sum_{k=i}^{n}w_{n,k}Y_i^k\eta_i\epsilon_i \\
&= \sum_{i=1}^{n}w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n}w_{n,i}\sum_{k=i}^{n}Y_i^k\eta_i\epsilon_i + \sum_{i=1}^{n}\sum_{k=i+1}^{n}(w_{n,k}-w_{n,i})Y_i^k\eta_i\epsilon_i \\
&= \sum_{i=1}^{n}w_{n,i}A^{-1}\epsilon_i + \sum_{i=1}^{n}w_{n,i}Y_0^i\delta_0 + \sum_{i=1}^{n}w_{n,i}(\sum_{k=i}^{n}Y_i^k\eta_i - A^{-1})\epsilon_i + \sum_{i=1}^{n}\sum_{k=i+1}^{n}(w_{n,k}-w_{n,i})Y_i^k\eta_i\epsilon_i \\
&\triangleq I+II+III+IV
\end{aligned} \tag{39}$$

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

$$\sum_{k=i+1}^{n}|Y_i^k| \lesssim (i+1)^{\alpha}.$$

So we have,

$$\lim_{n\to\infty}|\sqrt{n}\sum_{i=1}^{n}w_{n,i}Y_0^i\delta_0| \lesssim \lim_{n\to\infty}\frac{1}{n}|\sqrt{n}\sum_{i=1}^{n}Y_0^i| \lesssim \lim_{n\to\infty}\frac{1}{\sqrt{n}}=0.$$

Recall $\phi_i^n=\sum_{k=i}^{n}Y_i^k\eta_i - A^{-1}$. Then by Lemma A.1 and the fact that $|\phi_i^n|\le K$,

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{E}|\sqrt{n}\sum_{i=1}^{n}w_{n,i}(\sum_{k=i}^{n}Y_i^k\eta_i - A^{-1})\epsilon_i|^2 &= \lim_{n\to\infty}\mathbb{E}|\sqrt{n}\sum_{i=1}^{n}w_{n,i}\phi_i^n\epsilon_i|^2 \\
&\lesssim \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|\phi_i^n|^2 \\
&\le \lim_{n\to\infty}\frac{K}{n}\sum_{i=1}^{n}|\phi_i^n| = 0.
\end{aligned}$$

The result shows that $\sqrt{n}II$ and $\sqrt{n}III$ converge to 0 in L^2 norm. For $\sqrt{n}IV$, let $a_i^n = \sum_{k=i+1}^{n}(w_{n,k} - w_{n,i})Y_i^k\eta_i$. By Lemma A.1 in Zhu et al. (2023), we have

$$|Y_i^k| \leq \exp(-\lambda \sum_{t=i+1}^{k} \eta_t),$$

and

$$|a_i^n| \lesssim \frac{1}{n} \sum_{k=i+1}^{n} |Y_i^k|\eta_i \lesssim \frac{1}{n}$$

We also have

$$\lim_{n\to\infty} \sum_{i=1}^{n} |a_i^n| \leq \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}|\eta_i \exp(-\lambda \sum_{t=i+1}^{k} \eta_t).$$

Now we claim that the term above goes to 0, i.e.,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}|\eta_i \exp(-\lambda \sum_{t=i+1}^{k} \eta_t) = 0.$$

As a result,

$$\begin{aligned}
\mathbb{E}|\sqrt{n}IV|^2 &= \mathbb{E}|\sqrt{n} \sum_{i=1}^{n} a_i^n \epsilon_i|^2 \\
&\lesssim \frac{1}{n} \sum_{i=1}^{n} |na_i^n|^2 \\
&\lesssim \frac{1}{n} \sum_{i=1}^{n} |na_i^n| = \sum_{i=1}^{n} |a_i^n| \to 0.
\end{aligned}$$

So the conclusion follows from the asymptotic normality of the first term, which is implied by Lemma B.4:

$$\sqrt{n} \sum_{i=1}^{n} w_{n,i} A^{-1} \epsilon_i \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}).$$

Finally, we only need to prove the claim that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}|\eta_i \exp(-\lambda \sum_{t=i+1}^{k} \eta_t) = 0.$$

By the piecewise Lipschitz condition, we have that except for finite jumping points,

$$|w_{n,k} - w_{n,k-1}| \leq L_1 \frac{1}{n^2}$$

We first assume that there are no jumping points and the inequality always holds. Then we need the following 3 steps:

Step 1: Define $m_i^i = 0$ and for any $k > 1$,

$$m_i^k = \sum_{t=i+1}^{k} \eta_t.$$

Then our goal is to prove $\sum_{i=1}^{n} \sum_{k=i+1}^{n} |w_{n,k} - w_{n,i}|\eta_i \exp(-\lambda m_i^k) \to 0$. Recall that $\lambda = \min(\lambda_{\min}(A), 1/(2\eta))$. Let $\mu = \lfloor 1/\lambda \rfloor + 1$. Choose an $N \in \mathbb{N}^+$ such that $\forall k > i \geq N$,

$$m_i^k \geq \mu \log \frac{k}{i}.$$

Since $m_i^k \geq \eta(k^{1-\alpha} - i^{1-\alpha})/(1-\alpha)$, we can always find such an N.
Step 2: Let $b_i^n = \sum_{k=i+1}^n |w_{n,k} - w_{n,i}|\eta_i \exp(-\lambda m_i^k)$. We decompose $\sum_{i=1}^n b_i^n$ into two parts:

$$\begin{aligned}
\sum_{i=1}^n b_i^n &= \sum_{i=1}^n \eta_i \sum_{k=i+1}^n (w_{n,k} - w_{n,i}) \exp(-\lambda m_i^k) \\
&\leq \sum_{i=1}^N \eta_i \sum_{k=i+1}^n |w_{n,k} - w_{n,i}| \exp(-\lambda m_i^k) + \sum_{i=N+1}^n \eta_i \sum_{k=i+1}^n |w_{n,k} - w_{n,i}| \exp(-\lambda m_i^k) \\
&\triangleq I_1 + I_2
\end{aligned} \tag{40}$$

Step 3: Show that each term goes to 0 when $n \to \infty$. For the first term we have

$$I_1 \leq \sum_{i=1}^N \eta_i \frac{2C}{n} \sum_{k=i+1}^n \exp(-\lambda m_i^k) \lesssim \frac{1}{n} \sum_{i=1}^N (i+1)^\alpha i^{-\alpha} \lesssim \frac{1}{n}$$

The second inequality is due to Lemma A.1 and Lemma A.2 in Zhu et al. (2023). Now there exists a constant $\tilde{C}$ such that $|w_{n,t+1} - w_{n,t}| \leq L_1/n^2$. So for the second term we have

$$\begin{aligned}
I_2 &= \sum_{i=N+1}^n \eta_i \sum_{k=i+1}^n \{\sum_{t=i}^{k-1} |w_{n,t+1} - w_{n,t}|\} e^{-\lambda m_i^k} \\
&\lesssim \sum_{i=N+1}^n \eta_i \sum_{k=i+1}^n \sum_{t=i}^{k-1} \frac{1}{n^2} e^{-\lambda m_i^k} \\
&\lesssim n^{\alpha-2} \sum_{i=N+1}^n \eta_i \sum_{k=i+1}^n \sum_{t=i}^{k-1} \frac{1}{t^\alpha} e^{-\lambda m_i^k} \\
&\lesssim n^{\alpha-2} \sum_{i=N+1}^n \eta_i \sum_{k=i+1}^n m_i^k e^{-\lambda m_i^k} \\
&= n^{\alpha-2} \sum_{i=N+1}^n \sum_{k=i+1}^n \frac{m_i^k \eta_i}{\eta_k} e^{-\lambda m_i^k} (m_i^k - m_i^{k-1})
\end{aligned} \tag{41}$$

The second inequality is because $t^\alpha \leq n^\alpha$. Notice that $\frac{\eta_i}{\eta_k} \leq \frac{k}{i} \leq e^{\frac{m_i^k}{\mu}}$ by step 2,

$$\begin{aligned}
I_2 &\lesssim n^{\alpha-2} \sum_{i=N+1}^n \sum_{k=i+1}^n m_i^k e^{(\frac{1}{\mu} - \lambda) m_i^k} (m_i^k - m_i^{k-1}) \\
&\lesssim n^{\alpha-2} \sum_{i=N+1}^n \int_0^{+\infty} m e^{(\frac{1}{\mu} - \lambda) m} dm \\
&\lesssim n^{\alpha-1} \to 0.
\end{aligned} \tag{42}$$

We have shown that

$$\sum_{i=1}^n b_i^n \lesssim n^{\alpha-1} \to 0.$$

So the claim is proved under the condition of no jumping points. Then we deal with the finite jumping points case. Without loss of generality, we assume there is one point $0 < \kappa < 1$ such that the Lipschitz condition holds on $[0, \kappa]$ and $[\kappa, 1]$. The case of multiple but finite points follows the same argument.

Let $d_n = w_{n,\lceil \kappa n \rceil} - w_{n,\lfloor \kappa n \rfloor}$ and $w'_{n,i} = -d_n$ for $i > \kappa n$ and 0 otherwise. Notice that $|w'_{n,i}| \lesssim 1/n$.We have

$$|w_{n,k} - w_{n,i}| \leq |w_{n,k} + w'_{n,k} - w_{n,i} - w'_{n,i}| + |w'_{n,k} - w'_{n,i}|,$$

and the term in the claim is a linear combination of $|w_{n,k}-w_{n,i}|$. So we can plug in the triangle inequality where $|w_{n,k}+w'_{n,k}-w_{n,i}-w'_{n,i}|$ has no jumping points, for which

$$\lim_{n\to\infty}\sum_{i=1}^{n}\eta_i\sum_{k=i+1}^{n}|w_{n,k}+w'_{n,k}-w_{n,i}-w'_{n,i}|\exp(-\lambda\sum_{t=i+1}^{k}\eta_t)=0$$

is proved. So we only need to verify the claim for $|w'_{n,k}-w'_{n,i}|$. By Lemma A.1. in Zhu et al. (2023) we have

$$\exp(-\lambda\sum_{t=i+1}^{k}\eta_t)\le\exp\Big(-\frac{\lambda\eta}{1-\alpha}\big(k^{1-\alpha}-(i+1)^{1-\alpha}\big)\Big).$$

By Lemma A.2,

$$\begin{aligned}
&\sum_{i=1}^{n}\eta_i\sum_{k=i+1}^{n}|w'_{n,k}-w'_{n,i}|\exp(-\lambda\sum_{t=i+1}^{k}\eta_t)\\
\le&\sum_{i=1}^{\lfloor\kappa n\rfloor}\eta_i\sum_{k=\lceil\kappa n\rceil}^{n}\frac{1}{d_n}\exp\Big(-\frac{\lambda\eta}{1-\alpha}\big(k^{1-\alpha}-(i+1)^{1-\alpha}\big)\Big)\\
\lesssim&\frac{1}{n}\sum_{i=1}^{\lfloor\kappa n\rfloor}i^{-\alpha}\exp\Big(\frac{\lambda\eta}{1-\alpha}(i+1)^{1-\alpha}\Big)\sum_{k=\lceil\kappa n\rceil}^{n}\exp\Big(-\frac{\lambda\eta}{1-\alpha}k^{1-\alpha}\Big)\\
\lesssim&\frac{1}{n}\exp\Big(\frac{\lambda\eta}{1-\alpha}\lfloor\kappa n\rfloor^{1-\alpha}\Big)\exp\Big(-\frac{\lambda\eta}{1-\alpha}\lceil\kappa n\rceil^{1-\alpha}\Big)\lceil\kappa n\rceil^{\alpha}\\
\lesssim&\frac{\lceil\kappa n\rceil^{\alpha}}{n}
\end{aligned}\tag{43}$$

So the claim $\lim_{n\to\infty}\sum_{i=1}^{n}\eta_i\sum_{k=i+1}^{n}|w'_{n,k}-w'_{n,i}|\exp(-\lambda\sum_{t=i+1}^{k}\eta_t)=0$ holds, and we complete the proof of Lemma B.3. □

Proof of Lemma B.4. We further decompose the weighted sum of the martingale difference sequence as

$$\sqrt{n}\sum_{i=1}^{n}w_{n,i}A^{-1}\epsilon_i=\sqrt{n}\sum_{i=1}^{n}w_{n,i}A^{-1}(\epsilon_i-\nabla f(x^*,\xi_i))+\sqrt{n}\sum_{i=1}^{n}w_{n,i}A^{-1}\nabla f(x^*,\xi_i).$$

By Assumption 2.2,

$$\mathbb{E}|\nabla f(x_{i-1},\xi_i)-\nabla f(x^*,\xi_i)|^2\le L^2|\delta_{i-1}|^2$$

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

$$\begin{aligned}
|\nabla F(x_{i-1})-\nabla F(x^*)|^2&=|\mathbb{E}_{i-1}(\nabla f(x_{i-1},\xi_i)-\nabla f(x^*,\xi_i))|^2\\
&\le\mathbb{E}_{i-1}|\nabla f(x_{i-1},\xi_i)-\nabla f(x^*,\xi_i)|^2\\
&\le L^2|\delta_{i-1}|^2.
\end{aligned}$$

Since $\nabla F(x^*)=0$, we have

$$(\mathbb{E}_{i-1}|\epsilon_i-\nabla f(x^*,\xi_i)|^2)^{\frac{1}{2}}\le|\nabla F(x_{i-1})|+(\mathbb{E}_{i-1}|\nabla f(x_{i-1},\xi_i)-\nabla f(x^*,\xi_i)|^2)^{\frac{1}{2}}\le L|\delta_{i-1}|.$$

We know that $\mathbb{E}|\delta_{i-1}|^2\lesssim(i-1)^{-\alpha}$, so

$$\mathbb{E}|\epsilon_i-\nabla f(x^*,\xi_i)|^2\lesssim L^2(i-1)^{-\alpha}.$$

Since $|w_{n,i}|\le Cn^{-1}$,

$$\begin{aligned}
\mathbb{E}|\sqrt{n}\sum_{i=1}^{n}w_{n,i}A^{-1}(\epsilon_i-\nabla f(x^*,\xi_i))|^2&\lesssim\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\epsilon_i-\nabla f(x^*,\xi_i)|^2\\
&\lesssim\frac{1}{n}(\sum_{i=2}^{n}(i-1)^{-\alpha}+|\delta_0|^2)\asymp n^{-\alpha}
\end{aligned}$$

which vanishes as $n \to \infty$. Then it suffices to prove the CLT for $\sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}\nabla f(x^*, \xi_i)$. Let $Z_i = A^{-1}\nabla f(x^*, \xi_i)$ be a sequence of mean-zero variables with finite second moments, and denote the covariance matrix as $\Sigma_n = \sum_{i=1}^n (\sqrt{n}w_{n,i})^2 A^{-1}SA^{-1}$. Since $\sum_{i=1}^n (\sqrt{n}w_{n,i})^2 \to w$ and $|w_{n,i}| \le Cn^{-1}$, we have

$$|\Sigma_n^{-\frac{1}{2}}\sqrt{n}w_{n,i}Z_i|^2 \le \frac{4}{w}nw_{n,i}^2|(A^{-1}SA^{-1})^{-1}||Z_i| \lesssim \frac{|Z_i|^2}{n}.$$

Hence for any $\nu > 0$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\{|\Sigma_n^{-\frac{1}{2}}\sqrt{n}w_{n,i}Z_i|^2 \mathbb{1}_{|\Sigma_n^{-\frac{1}{2}}\sqrt{n}w_{n,i}Z_i|^2 \ge \nu}\} &\lesssim \lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\{\frac{|Z_i|^2}{n}\mathbb{1}_{|Z_i|^2 \ge n\nu}\} \\ &= \mathbb{E}\{|Z_i|^2\mathbb{1}_{|Z_i|^2 \ge n\nu}\} \to 0,\end{aligned}$$

which converges to 0 because of the fact that $\mathbb{1}_{|Z_i|^2 \ge n\nu} \to 0$ and Lebesgue's dominated convergence theorem. We have verified the Lindeberg condition. As a result,

$$\Sigma_n^{-\frac{1}{2}}\sqrt{n}\sum_{i=1}^n w_{n,i}Z_i \xrightarrow{D} \mathcal{N}(0, \mathbf{I}_d).$$

and since $\Sigma_n \to wA^{-1}SA^{-1}$,

$$\sqrt{n}\sum_{i=1}^n w_{n,i}Z_i \xrightarrow{D} \mathcal{N}(0, wA^{-1}SA^{-1}).$$

Here we also illustrate that $nV_n \to A^{-1}SA^{-1}$ under the setting of Theorem 2.7. Recall the SGD iterates

$$x_i = x_{i-1} - (Ax_{i-1} - A^{\frac{1}{2}}\xi_i) = (\mathbf{I}_d - A)x_{i-1} + z_i.$$

where $\mathbb{E}z_i = 0$ and $\mathrm{Cov}(z_i) = S$. The weighted averaged SGD is formulated as

$$\tilde{x}_n = \eta_1^{-1}(\mathbf{I}_d - \eta_1 A)\Phi_{n,1}x_0 + \sum_{i=1}^n \Phi_{n,i}z_i,$$

with its covariance $V_n = \sum_{i=1}^n \Phi_{n,i}S\Phi_{n,i}$. On the other hand,

$$\begin{aligned}\sqrt{n}\tilde{x}_n &= \sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}z_i + \sqrt{n}\sum_{i=1}^n w_{n,i}Y_0^i\delta_0 \\ &\quad + \sqrt{n}\sum_{i=1}^n w_{n,i}a_i^n z_i + \sqrt{n}\sum_{i=1}^n b_i^n z_i,\end{aligned} \tag{44}$$

and we have shown that the last three terms on the left-hand side L^2 converge to 0. Hence

$$\lim_{n\to\infty}\mathrm{Cov}(\sqrt{n}\tilde{x}_n) = \lim_{n\to\infty} nV_n = \lim_{n\to\infty}\mathrm{Cov}(\sqrt{n}\sum_{i=1}^n w_{n,i}A^{-1}z_i) = wA^{-1}SA^{-1}.$$

□

B.5 Proof of Theorem 2.10

B.5.1 High level summary and remark

We first present a sketch of the proof which outlines the main techniques. The high-level idea is to decompose the partial sum process S_n as

$$S_n = S_{L,n} + S_{D,n} + S_{R,n} + \mathcal{R}_n.$$

Step 1 Prove the FCLT of the "linear partial sum process" $S_{L,i}$. It suffices to verify the finite-dimensional convergence and the tightness condition.

Step 2 Establish a uniform approximation of $S_{L,n}$ to the target process S_n. In the decomposition, $S_{D,n}$, $S_{R,n}$ and $\mathcal{R}_n$ are the error terms of martingale differences, Taylor expansion's remainder, and the initial point. We show that the maximum of each standardized error process converges to 0, and complete the proof.

In both steps, we leverage a sharp maximal inequality, Proposition 1 in (Wu, 2007), to bound the maximum of the partial sum process.

Remark B.5. *In Lee et al. (2022), Assumption 1.(iv) and (v) require that $\alpha > 1/2$ and $\mathbb{E}\|\xi_t\|^{2p} < \infty$ for some $p \geq 1/(1-\alpha)$. So the condition is $2p \geq 2/(1-\alpha) > 4$. In their proof of Theorem 1, the following expansion is used:*

$$(\sum_{j=1}^{m-1} \|w_j^m\|^2 \|\xi_j\|^2)^p = \sum_{j_1,\ldots,j_p=1}^{m-1} \prod_{l=1}^{p} \|w_{j_l}^m\|^2 \|\xi_{j_l}\|^2.$$

Therefore, it is clear that $p \in \mathbb{N}$. In conclusion, their moment condition is $p \geq 3$ and $2p \geq 6$.

Our paper uses p instead of $2p$ in the moment condition. In this context, the condition in Lee et al. (2022) becomes $p \geq 6$, while we greatly reduce it to $p > 2$ without requiring $p \in \mathbb{N}$. Another work Li et al. (2022b) also investigates the FCLT in the setting of local SGD. However, their Assumption 3.2 requires $\sup_x \mathbb{E}|\nabla f(x,\xi) - \nabla F(x)|^p < \infty$ for some $p > 2$. This uniformly bounded p-th moment condition is also much stronger than our assumption, since we only assume it is bounded at one point x^*.

B.5.2 Detailed Proof

Proof. Recall (25) for the decomposition of SGD iterates. Without loss of generality, we can assume $x^* = 0$ here. Let

$$y_n = y_{n-1} - \eta_n A y_{n-1} - \eta_n \nabla f(x^*, \xi_n), \quad y_0 = 0, \quad S_{L,n} = \sum_{i=1}^{n} y_i,$$

$$\Delta_{D,n} = \Delta_{D,n-1} - \eta_n A \Delta_{D,n-1} - \eta_n D_n, \quad \Delta_{D,0} = 0, \quad S_{D,n} = \sum_{i=1}^{n} \Delta_{D,i},$$

$$\Delta_{R,n} = \Delta_{R,n-1} - \eta_n A \Delta_{R,n-1} - \eta_n R_n, \quad \Delta_{R,0} = 0, \quad S_{R,n} = \sum_{i=1}^{n} \Delta_{R,i}.$$

Then the error term due to the starting point goes to 0 as

$$\frac{1}{\sqrt{n}} \max_{1\leq i\leq n} |S_i - S_{L,i} - S_{D,i} - S_{R,i}| \lesssim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |\prod_{j=2}^{i} (\mathbf{I}_d - \eta_j A)|) \lesssim \frac{1}{\sqrt{n}}.$$

We prove the functional CLT in two steps. First, we prove the functional CLT for $S_{L,n}$. To this end, we need to verify the following tightness condition: for any $r_1, r_2 > 0$, there exists $\delta > 0$ and $n_0 \in \mathbb{N}^+$ such that for all $n \geq n_0$,

$$\mathbb{P}\Big(\max_{|i_1 - i_2| \leq n\delta} \frac{|S_{L,i_1} - S_{L,i_2}|}{\sqrt{n}} \geq r_1 \Big) \leq r_2. \tag{45}$$

Without loss of generality, we take $r_1 = r_2 = r$. Notice that the probability in (45) can be upper bounded by

$$\mathbb{P}\Big(\max_{|i_1 - i_2| \leq n\delta} \frac{|S_{L,i_1} - S_{L,i_2}|}{\sqrt{n}} \geq r \Big) \leq \sum_{l=0}^{\lceil \frac{1}{\delta} \rceil} \mathbb{P}\Big(\max_{t_l \leq i \leq t_{l+1}} \frac{|S_{L,i} - S_{L,,t_l}|}{\sqrt{n}} \geq \frac{r}{2} \Big), \tag{46}$$

Let $t_1 = 1$, $t_l = t_{l-1} + \lfloor n\delta \rfloor$ for $1 < l \leq \lceil 1/\delta \rceil$ and $t_{\lceil 1/\delta \rceil + 1} = n$. By Markov's inequality,

$$\sum_{l=0}^{\lceil \frac{1}{\delta} \rceil} \mathbb{P}\Big(\max_{t_l \leq i \leq t_{l+1}} \frac{|S_{L,i} - S_{L,,t_l}|}{\sqrt{n}} \geq \frac{r}{2} \Big) \leq \sum_{l=0}^{\lceil \frac{1}{\delta} \rceil} \mathbb{E} \max_{t_l \leq i \leq t_{l+1}} \frac{2^q |S_{L,i} - S_{L,t_l}|^q}{(r\sqrt{n})^q} \leq \frac{2^q}{(r\sqrt{n})^q} \sum_{l=0}^{\lceil \frac{1}{\delta} \rceil} \mathbb{E} \max_{0 \leq i \leq \lfloor n\delta \rfloor} |S_{L,t_l+i} - S_{L,t_l}|^q. \tag{47}$$

We apply the powerful maximal inequality Proposition 1 in Wu (2007) to bound the expectation of the maximum as

$$\begin{aligned}\big(\mathbb{E}\max_{0\le i\le\lfloor n\delta\rfloor}|S_{L,t_l+i}-S_{L,t_l}|^q\big)^{1/q}&\le\big(\mathbb{E}\max_{0\le i\le 2^d}|S_{L,t_l+i}-S_{L,t_l}|^q\big)^{1/q}\\&\le\sum_{k=0}^{d}\Big[\sum_{m=1}^{2^{d-k}}\mathbb{E}|S_{L,t_l+2^km}-S_{L,t_l+2^k(m-1)}|^q\Big]^{1/q},\end{aligned}$$

where $d=\lceil\log_2(\lfloor n\delta\rfloor)\rceil$. Now we prove the following lemma:

Lemma B.6. $\|y_{m+1}+...+y_{m+j}\|_q\lesssim\sqrt{j}$.

Proof. When $j\le m^\alpha$, $\|\sum_{i=m+1}^{m+j}y_i\|_q\le\sum_{i=m+1}^{m+j}\|y_i\|_q\lesssim\sum_{i=m+1}^{m+j}i^{-\alpha/2}\lesssim j(m+1)^{-\alpha/2}\lesssim j^{1-1/2}=\sqrt{j}$. When $j>m^\alpha$, the high level idea is to decompose the sum to a linear combination of $Z_k=\nabla f(x^*,\xi_k)$, which is also the sum of martingale differences. The relationship between y_i and y_j can be written as

$$y_i=\prod_{k=j+1}^{i}(\mathbf{I}-\eta_kA)y_j+\sum_{k=j+1}^{i}\prod_{l=k+1}^{i}(\mathbf{I}-\eta_lA)\eta_k\nabla f(x^*,\xi_k).\quad 0\le j<i,\tag{48}$$

So $y_{m+1}+...+y_{m+j}$ can be expressed as the following linear combination of independent random variables,

$$\sum_{i=m+1}^{m+j}\prod_{k=m+2}^{i}(\mathbf{I}-\eta_kA)y_{m+1}+\sum_{i=m+2}^{m+j}\prod_{k=m+3}^{i}(\mathbf{I}-\eta_kA)\eta_{m+2}Z_{m+2}+...+\eta_{m+j}Z_{m+j}.$$

We then use Burkholder's inequality since independent mean zero sequences are also martingale differences,

$$\|y_{m+1}+...+y_{m+j}\|_q^2\le(q-1)(\|\psi_{m,j,1}\|_q^2+...+\|\psi_{m,j,j}\|_q^2).$$

where $\psi_{m,j,1}=\sum_{i=m+1}^{m+j}\prod_{k=m+2}^{i}(\mathbf{I}-\eta_kA)y_{m+1}$ and

$$\psi_{m,j,l}=\sum_{i=m+l}^{m+j}\prod_{k=m+l+1}^{i}(\mathbf{I}-\eta_kA)\eta_{m+l}Z_{m+l},\ 2\le l\le j.$$

Recall that $\|y_{m+1}\|_q^2\lesssim(m+1)^{-\alpha}$, so we have

$$\|\psi_{m,j,1}\|_q^2\lesssim(m+1)^{-\alpha}\Big(\sum_{i=m+1}^{m+j}\prod_{k=m+2}^{i}|\mathbf{I}-\eta_kA|\Big)^2\lesssim(m+1)^{\alpha}\le 2j.$$

Since $\|Z_n\|_q^2$ is finite, we also have

$$\|\psi_{m,j,k}\|_q^2\lesssim(m+l)^{-2\alpha}\Big(\sum_{i=m+l}^{m+j}\prod_{k=m+l+1}^{i}|\mathbf{I}-\eta_kA|\Big)^2=\mathcal{O}(1).$$

As a result,

$$\|y_{m+1}+...+y_{m+j}\|_q^2\lesssim 2j+(j-1)\mathcal{O}(1)\lesssim j$$

and the conclusion is proven. $\square$

By the lemma, $\mathbb{E}|S_{L,t_l+2^km}-S_{L,t_l+2^k(m-1)}|^q\lesssim 2^{kq/2}$. Hence there exists a constant C such that

$$\begin{aligned}\mathbb{P}\Big(\max_{|i_1-i_2|\le n\delta}\frac{|S_{L,i_1}-S_{L,i_2}|}{\sqrt{n}}\ge r\Big)&\le C\frac{2^q}{\delta(r\sqrt{n})^q}(\sum_{k=0}^{d}(\sum_{m=1}^{2^{d-k}}2^{kq/2})^{1/q})^q\\&=C\frac{2^q}{\delta(r\sqrt{n})^q}2^{qd/2}(\sum_{k=0}^{d}(2^{1/q-1/2})^{d-k})^q.\end{aligned}$$

Since $q > 2$, the geometric series $\sum_{k=0}^{d}(2^{1/q-1/2})^{d-k}$ converges. We also have $2^d \lesssim n\delta$. So there exists a universal constant $\tilde{C}$ such that

$$\mathbb{P}\Big(\max_{|i_1-i_2|\le n\delta} \frac{|S_{L,i_1}-S_{L,i_2}|}{\sqrt{n}} \ge r\Big) \le \frac{\tilde{C}(n\delta)^{q/2}}{\delta(r\sqrt{n})^q} = \frac{\tilde{C}\delta^{q/2}}{\delta r^q}.$$

We can choose

$$\delta \le \big(\frac{r^{1+q}}{\tilde{C}}\big)^{2/(q-2)}$$

to ensure that

$$\mathbb{P}\Big(\max_{|i_1-i_2|\le n\delta} \frac{|S_{L,i_1}-S_{L,i_2}|}{\sqrt{n}} \ge r\Big) \le r.$$

Hence the tightness condition holds.

Next, we prove the finite-dimensional convergence. In the proof of Lemma B.3 and Lemma B.4, we have shown that

$$\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A^{-1}\nabla f(x^*,\xi_i)\| \to 0.$$

Let $S_n^* = \sum_{i=1}^{n}\nabla f(x^*,\xi_i)$. For all finite $\{t_j\}_{j=1}^{k} \in (0,1)$, it entails

$$\sum_{j=1}^{k}\|S^*_{\lfloor nt_j\rfloor}/\sqrt{nt_j} - S_{L,\lfloor nt_j\rfloor}/\sqrt{nt_j}\| \to 0.$$

So the finite-dimensional convergence follows from the Donsker's invariance principle on i.i.d sequence $\nabla f(x^*,\xi_i)$. The tightness condition and the finite-dimensional convergence ensure the FCLT for $S_{L,n}$.

Next, we show that

$$\frac{1}{\sqrt{n}}\|\max_{1\le i\le n}|S_{D,i}|\| \to 0, \quad \frac{1}{\sqrt{n}}\|\max_{1\le i\le n}|S_{R,i}|\| \to 0.$$

Recall that $\Delta_{D,n} = \sum_{i=1}^{n}\eta_i\prod_{j=i+1}^{n}(\mathbf{I}_d-\eta_j A)D_i$ where $D_i = \nabla f(x_{i-1},\xi_i) - \nabla F(x_{i-1}) - \nabla f(x^*,\xi_i)$. Again by Proposition 1 in Wu (2007), for $n \le 2^d < 2n$, we have

$$\begin{aligned}\|\max_{1\le i\le n}|S_{D,i}|\| &\le \|\max_{1\le i\le 2^d}|S_{D,i}|\| \\ &\le \sum_{k=0}^{d}\Big[\sum_{m=1}^{2^{d-k}}\|S_{D,2^k m} - S_{D,2^k(m-1)}\|^2\Big]^{1/2}.\end{aligned}$$

Notice that

$$\|S_{D,2^k m} - S_{D,2^k(m-1)}\| \le \sum_{i=1}^{2^k}\|\Delta_{D,2^k(m-1)+i}\| \lesssim \sum_{i=1}^{2^k}(2^k(m-1)+i)^{-\alpha}.$$

Here we use the fact that $\Delta_{D,n}$ is the linear combination of a martingale difference sequence, and hence

$$\|\Delta_{D,n}\|^2 \le \sum_{i=1}^{n}|\eta_i\prod_{j=i+1}^{n}(\mathbf{I}_d-\eta_j A)|^2\mathbb{E}|D_i|^2 \lesssim \sum_{i=1}^{n} i^{-3\alpha}\prod_{j=i+1}^{n}|(\mathbf{I}_d-\eta_j A)|^2 \asymp n^{-2\alpha}.$$

Now we can further bound the polynomial series as

$$\sum_{i=1}^{2^k}(2^k(m-1)+i)^{-\alpha} \lesssim (2^k m)^{1-\alpha} - (2^k(m-1))^{1-\alpha}.$$

As a result,

$$\begin{aligned}
\|\max_{1\le i\le n}|S_{D,i}|\| &\lesssim \sum_{k=0}^{d}\Big[\sum_{m=1}^{2^{d-k}}[(2^k m)^{1-\alpha}-(2^k(m-1))^{1-\alpha}]^2\Big]^{1/2}\\
&\le \sum_{k=0}^{d}\Big[\sum_{m=1}^{2^{d-k}}(2^k m)^{2-2\alpha}-(2^k(m-1))^{2-2\alpha}\Big]^{1/2}\\
&= \sum_{k=0}^{d}\sqrt{(2^k 2^{d-k})^{2-2\alpha}} = \sum_{k=0}^{d}2^{d(1-\alpha)}.
\end{aligned}$$

Finally, we have

$$\frac{1}{\sqrt{n}}\|\max_{1\le i\le n}|S_{D,i}|\| \lesssim \frac{1}{\sqrt{n}}\sum_{k=1}^{d}2^{d(1-\alpha)} \lesssim n^{1-\alpha-1/2}\log_2 n = n^{1/2-\alpha}\log_2 n \to 0.$$

Recall that the second term

$$\Delta_{R,n} = \sum_{i=1}^{n}\eta_i\prod_{j=i+1}^{n}(\mathbf{I}_d-\eta_j A)R_i.$$

So

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|\max_{1\le i\le n}|S_{R,i}|\| &\le \frac{1}{\sqrt{n}}\|\max_{1\le i\le n}\sum_{j=1}^{i}|\Delta_{R,j}|\|\\
&\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\Delta_{R,i}|\\
&\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{i}\eta_k\|R_k\|\|\prod_{j=k+1}^{i}(\mathbf{I}_d-\eta_j A)|\\
&\lesssim \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{i}i^{-2\alpha}|\prod_{j=k+1}^{i}(\mathbf{I}_d-\eta_j A)|\\
&\lesssim \frac{1}{\sqrt{n}}\sum_{i=1}^{n}i^{-\alpha} \asymp n^{1/2-\alpha}\to 0.
\end{aligned}$$

□

B.6 Proof of Corollary 3.1

Proof. Recall the definition of $\theta_{n,i}$:

$$\begin{aligned}
\theta_{n,i} &= \frac{\gamma+1}{\gamma+i}\prod_{j=i+1}^{n}\frac{j-1}{j+\gamma}\\
&= \frac{\gamma+1}{n}\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}.
\end{aligned}$$

We propose another Lemma here and defer the proof to section C.

Lemma B.7. *The weight $w_{n,i}=\theta_{n,i}$ satisfies $\sum_{i=1}^n w_{n,i}=1$, $w_{n,i}\le(\gamma+1)/n$ and*

$$\lim_{n\to\infty} n\sum_{i=1}^{n}(w_{n,i})^2 = \frac{(\gamma+1)^2}{2\gamma+1}.$$

Now we only have to verify the smoothness condition. We first show that there exists a constant $\tilde{C} = \gamma(\gamma+1)$ such that for all $1 \le i < n$,

$$|w_{n,i+1} - w_{n,i}| \le \tilde{C} n^{-2}.$$

Notice for any $n \in \mathbb{N}^+$,

$$|\theta_{n+1,n+1} - \theta_{n,n}| = \frac{(\gamma+1)(\gamma+n) - n(\gamma+1)}{(\gamma+n+1)(\gamma+n)} = \frac{\gamma(\gamma+1)}{(\gamma+n+1)(\gamma+n)} \le \frac{\gamma(\gamma+1)}{(n+1)^2},$$

and

$$\begin{aligned} |\theta_{n+1,n} - \theta_{n+1,n-1}| &= (1 - \frac{\gamma+1}{\gamma+n})(\theta_{n,n} - \theta_{n,n-1}) \\ &= \frac{\gamma(\gamma+1)}{(\gamma+n)(\gamma+n-1)} \frac{n}{\gamma+n+1}. \end{aligned} \tag{49}$$

Since $n \ge 1$, we have $|\theta_{n+1,n} - \theta_{n+1,n-1}| \le |\theta_{n+1,n+1} - \theta_{n,n}|$. Similarly we can prove that $|\theta_{n+1,i+1} - \theta_{n+1,i}| \le |\theta_{n+1,n+1} - \theta_{n,n}|$ for any $1 \le i \le n$. So $|\theta_{n+1,i+1} - \theta_{n+1,i}| \le \gamma(\gamma+1)/(n+1)^2$ for any $1 \le i \le n$, or equivalently, $|w_{n,i+1} - w_{n,i}| \le \tilde{C} n^{-2}$ for all $1 \le i < n$.

Let $f_n(k/n) = n\theta_{n,i}$, it is clear that this f_n meets the piecewise Lipschitz condition with $\sum_{i=1}^n f_n(k/n) = n$ and $L_1 = \gamma(\gamma+1)$. It is actually global Lipschitz with no exception. So all assumptions for Theorem 2.7 are verified and the CLT holds.

□

B.7 Proof of Corollary 3.2

Proof. It is obvious that for $\kappa-$suffix averaging $|w_{n,i}| \lesssim 1/n$, $w = 1/\kappa$ and the piecewise Lipschitz condition holds with only one exception. So the CLT holds by Theorem 2.7.

□

B.8 Proof of Proposition 4.1

We should prove the following Proposition B.8, a more detailed version of Proposition 4.1. Instead of requiring $\eta_1 = a_1^{-2}$ in Proposition 4.1, we first consider a general step size. Recall that we have the squared loss function $f(x, \xi_i = (a_i, b_i)) = (a_i x - b_i)^2/2$, and SGD iterates

$$x_i = x_{i-1} - \eta_i a_i (a_i x_{i-1} - b_i). \tag{50}$$

Proposition B.8. *The unique solution to the optimization problem*

$$\min_{c=(c_0, \cdots, c_n): c^T \mathbb{1} = 1} \mathbb{E}\left|\sum_{i=0}^n c_i (x_i - x^*)\right|^2$$

with x_i defined in (50) is given by

$$c = \frac{\Theta^T D^{-1} \Theta \mathbb{1}}{\mathbb{1}^T \Theta^T D^{-1} \Theta \mathbb{1}},$$

where

$$D = \begin{pmatrix} (x_0 - x^*)^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2 a_1^2 \eta_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 a_2^2 \eta_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 a_n^2 \eta_n^2 \end{pmatrix},$$

$$\Theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \eta_1 a_1^2 - 1 & 1 & 0 & \cdots & 0 \\ 0 & \eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \eta_n a_n^2 - 1 & 1 \end{pmatrix}.$$

More explicitly,

$$c_{n,0} = \frac{(\frac{\sigma}{x_0 - x^*})^2 + a_1^2 - \eta_1^{-1}}{S_n},$$

$$c_{n,i} = \frac{\eta_i^{-1} + a_{i+1}^2 - \eta_{i+1}^{-1}}{S_n}, 1 \le i \le n-1,$$

$$c_{n,n} = \frac{1}{\eta_n S_n},$$

where $S_n = (\frac{\sigma}{x_0 - x^*})^2 + \sum_{i=1}^n a_i^2$.

Proof. The SGD error sequence $x_i - x^*$ takes the recursion form

$$x_i - x^* = (1 - \eta_i a_i^2)(x_{i-1} - x^*) + \eta_i a_i (b_i - a_i x^*)$$

Let

$$\Theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \eta_1 a_1^2 - 1 & 1 & 0 & \cdots & 0 \\ 0 & \eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \eta_n a_n^2 - 1 & 1 & . \end{pmatrix}$$

Then we have

$$\Theta \begin{pmatrix} x_0 - x^* \\ x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix} = \begin{pmatrix} x_0 - x^* \\ \eta_1(b_1 - a_1 x^*)a_1 \\ \eta_2(b_2 - a_2 x^*)a_2 \\ \vdots \\ \eta_n(b_n - a_n x^*)a_n \end{pmatrix}. \tag{51}$$

We further treat a_i as fixed and denote Θ as the matrix in the left hand side above. Similar as the mean estimation model, here the optimal weights solution is also determined by $\Sigma = (\mathbb{E}(x_i x_j))_{i,j\ge 0}$, the "covariance" matrix of $(x_0,\ x_1,\ x_2,\ \cdots,\ x_n)$.

We further define

$$\Phi = \begin{pmatrix} x_0 - x^* \\ \eta_1(b_1 - a_1 x^*)a_1 \\ \eta_2(b_2 - a_2 x^*)a_2 \\ \vdots \\ \eta_n(b_n - a_n x^*)a_n \end{pmatrix},\ X = \begin{pmatrix} x_0 - x^* \\ x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix},$$

and

$$D = \begin{pmatrix} (x_0 - x^*)^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2 a_1^2 \eta_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 a_2^2 \eta_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 a_n^2 \eta_n^2 \end{pmatrix}.$$

Then $\Theta X = \Phi$ implies that $\mathbb{E}\Theta X X^T \Theta^T = \Theta \mathbb{E} X X^T \Theta^T = \Theta\Sigma\Theta^T = \mathbb{E}\Phi\Phi^T$. By the fact that $\mathbb{E}(x_i - x^*)(x_j - x^*) = 0$ for all $i \neq j$, we have $\mathbb{E}\Phi\Phi^T = D$. Thus we have a diagonalization of Σ as $\Theta\Sigma\Theta^T = D$.

Using the Lagrangian multiplier method, we can obtain the closed-form solution as

$$\frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}},$$

and it remains to show that the solution

$$\frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma^{-1}\mathbb{1}} = \frac{\Theta^T D^{-1}\Theta\mathbb{1}}{\mathbb{1}^T\Theta^T D^{-1}\Theta\mathbb{1}}$$

is the form of Proposition B.8. Here we give the closed form of the matrix $\Theta^T D^{-1}\Theta$,

$$\Theta^T D^{-1}\Theta =$$
$$\begin{pmatrix}
\frac{1}{(x_0-x^*)^2}+\frac{(\eta_1 a_1^2-1)^2}{\sigma^2 a_1^2\eta_1^2} & \frac{\eta_1 a_1^2-1}{\sigma^2 a_1^2\eta_1^2} & 0 & 0 & \cdots & 0 \\
\frac{\eta_1 a_1^2-1}{\sigma^2 a_1^2\eta_1^2} & \frac{1}{\sigma^2 a_1^2\eta_1^2}+\frac{(\eta_2 a_2^2-1)^2}{\sigma^2 a_2^2\eta_2^2} & \frac{\eta_2 a_2^2-1}{\sigma^2 a_2^2\eta_2^2} & 0 & \cdots & 0 \\
0 & \frac{\eta_2 a_2^2-1}{\sigma^2 a_2^2\eta_2^2} & \frac{1}{\sigma^2 a_2^2\eta_2^2}+\frac{(\eta_3 a_3^2-1)^2}{\sigma^2 a_3^2\eta_3^2} & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \frac{\eta_n a_n^2-1}{\sigma^2 a_n^2\eta_n^2} \\
0 & 0 & \cdots & 0 & \frac{\eta_n a_n^2-1}{\sigma^2 a_n^2\eta_n^2} & \frac{1}{\sigma^2 a_n^2\eta_n^2}
\end{pmatrix},$$

and the conclusion can be easily verified. □

The optimal weight of the initialization term $c_{n,0}$ in Proposition B.8 depends on σ^2 and the initial error $x_0 - x^*$, both of which can not be observed. To solve this problem, we may consider the two-step estimation: first estimate σ and x^* using ASGD or other averaged schemes with a small batch of SGD iterates, then plug them in to obtain the optimal weights. Another approach is to modify the structure of Θ and equation 51. Choosing $\eta_1 = a_1^{-2}$, we can exclude $x_0 - x^*$ from (51) and reduce it to

$$\begin{pmatrix}
1 & 0 & \cdots & 0 \\
\eta_2 a_2^2 - 1 & 1 & \cdots & 0 \\
\ddots & \ddots & \ddots & \vdots \\
0 & 0 & \eta_n a_n^2 - 1 & 1
\end{pmatrix}
\begin{pmatrix} x_1 - x^* \\ x_2 - x^* \\ \vdots \\ x_n - x^* \end{pmatrix}
=
\begin{pmatrix} \eta_1(b_1 - a_1 x^*)a_1 \\ \eta_2(b_2 - a_2 x^*)a_2 \\ \vdots \\ \eta_n(b_n - a_n x^*)a_n \end{pmatrix}.$$

In other words, if we plug $\eta_1 = a_1^{-2}$ in Proposition B.8, we have $x_1 = \eta_i b_i a_i$ and all SGD iterates will be free of $x_0 - x^*$.

Denote $\tilde{\Theta}$ as this reduced matrix in the left hand side above, and we can perform a diagonalization of $\Sigma_{-0} = (\mathbb{E}x_i x_j)_{i,j\geq 1}$, the "covariance" matrix of SGD sequence without x_0, as follows

$$\tilde{\Theta}\Sigma_{-0}\tilde{\Theta}^T = D_{-0} = \begin{pmatrix}
\sigma^2 a_1^2\eta_1^2 & 0 & \cdots & 0 \\
0 & \sigma^2 a_2^2\eta_2^2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2 a_n^2\eta_n^2
\end{pmatrix}.$$

Instead of the optimization problem in proposition 4.1, the decomposition of Σ_{-0} enables us to solve a reduced problem excluding the weight on x_0

$$\min_{c=(c_1,\,\cdots,\,c_n):c^T\mathbb{1}=1} \mathbb{E}|\sum_{i=1}^n c_i(x_i - x^*)|^2,$$

and get the minimizer in the form of

$$c = \frac{\Sigma_{-0}^{-1}\mathbb{1}}{\mathbb{1}^T\Sigma_{-0}^{-1}\mathbb{1}} = \frac{\tilde{\Theta}^T D_{-0}^{-1}\tilde{\Theta}\mathbb{1}}{\mathbb{1}^T\tilde{\Theta}^T D_{-0}^{-1}\tilde{\Theta}\mathbb{1}}.$$

The σ^2 terms in D_{-0} cancel out. Finally, the weighting scheme with $\eta_1 = a_1^{-2}$ is

$$c_{n,i} = \frac{\eta_i^{-1} + a_{i+1}^2 - \eta_{i+1}^{-1}}{S_n}, 1 \leq i \leq n-1,$$

$$c_{n,n} = \frac{1}{\eta_n S_n},$$

where $S_n = \sum_{i=1}^n a_i^2$.

B.9 Proof of Corollary 4.2

Proof. We define $\tilde{x}'_n = \sum_{i=1}^{n-1} c_{n,i} x_i$. From Corollary 2.5 we know

$$n^{\frac{\alpha}{2}}(x_n - x^*) = O_p(1).$$

Notice that

$$\sqrt{n}(\tilde{x}_n - x^*) - \sqrt{n}(\tilde{x}'_n - (1 - \frac{1}{n\eta_n})x^*) = \frac{1}{\sqrt{n}\eta_n}(x_n - x^*) = o_p(1)$$

because

$$n^{\alpha - \frac{1}{2}}(x_n - x^*) = n^{\frac{\alpha}{2} - \frac{1}{2}} n^{\frac{\alpha}{2}}(x_n - x^*) = o_p(1).$$

So $\sqrt{n}(\tilde{x}_n - x^*)$ and $\sqrt{n}(\tilde{x}'_n - (1 - 1/n\eta_n)x^*) = \sqrt{n}\sum_{i=1}^{n-1} c_{n,i}(x_i - x^*)$ has the same asymptotic distribution. Moreover, consider

$$\sqrt{n}\mathbb{E}|\sum_{i=1}^{n-1} c_{n,i}(x_i - x^*) - \sum_{i=1}^{n-1} \frac{1}{n}(x_i - x^*)| \leq \sqrt{n}\sum_{i=1}^{n-1} \frac{\eta_i^{-1} - \eta_{i+1}^{-1}}{n}\mathbb{E}|x_i - x^*| \lesssim \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} i^{\alpha-1} i^{-\alpha/2} \lesssim n^{\alpha/2 - 1/2}$$

which vanishes as $n \to \infty$. So $\tilde{x}'_n = \sum_{i=1}^{n-1} c_{n,i} x_i$ has the same asymptotic distribution as ASGD. Therefore the conclusion is proved. □

C PROOF OF AUXILIARY LEMMAS

C.1 Proof of Lemma A.2

Proof.

1. 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

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

and

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

which implies

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

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

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

Further, by Taylor series,

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

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

$$\begin{aligned}
\sum_{j=1}^{i}\exp\left(\beta j^{1-\alpha}\right) j^{-\gamma\alpha} &\asymp \int_1^{i+1}\exp\left(\beta t^{1-\alpha}\right) t^{-\gamma\alpha}dt \\
&\asymp \int_\beta^{\beta(i+1)^{1-\alpha}} e^u u^{-\frac{(\gamma-1)\alpha}{1-\alpha}}du. \qquad (52)
\end{aligned}$$

Using integration by parts, we have the following:

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

Therefore by (1), we have

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

3. Let $\psi(x)=\exp(-\beta x^{1-\alpha})$. It's a decreasing function of x. So we can bound the target term as

$$\begin{aligned}
\sum_{k=j}^{n}\exp(-\beta k^{1-\alpha}) &\lesssim \int_{j-1}^{n}\exp(-\beta x^{1-\alpha}))dx \\
&= \int_{\beta(j-1)^{1-\alpha}}^{\infty}\frac{s^{\frac{\alpha}{1-\alpha}}}{\beta(1-\alpha)}e^{-s}ds \qquad (53)\\
&\lesssim (j-1)^\alpha\exp(-\beta(j-1)^{1-\alpha}) \\
&\asymp \exp(-\beta j^{1-\alpha})j^\alpha.
\end{aligned}$$

□

C.2 Proof of Lemma B.7

Proof. Since $\Gamma(\gamma+i+1)\Gamma(n+1)<\Gamma(\gamma+n+1)\Gamma(i+1)$, we have

$$|\theta_{n,i}|\leq\frac{\gamma+1}{n}.$$

From the recursive form of polynomial decay averaged SGD, it is easy to see $\sum_{i=1}^{n}\theta_{n,i}=1$. The last step is to prove the limitation holds. Define

$$\Gamma_n(x)=\int_0^n t^{x-1}(1-\frac{t}{n})^n dt=\frac{n^x n!}{z(z+1)(z+2)\cdots(z+n)},$$

where the last equality is from integration by parts. It's well known that $(1-\frac{t}{n})^n\le(1-\frac{t}{n+1})^{n+1}$ and $\lim_{n\to\infty}(1-\frac{t}{n})^n=e^{-t}$ for any t. So $\Gamma_n(x)\le\Gamma(x)$. Meanwhile we have an equivalent definition of $\Gamma(x)$:

$$\Gamma(x)=\lim_{n\to\infty}\frac{n^x n!}{x(x+1)(x+2)\cdots(x+n)}=\lim_{n\to\infty}\Gamma_n(x).$$

So for any $\tau>0$, there exists an $N>0$ such that for all $n\ge N$,

$$0\le\Gamma(\gamma)-\frac{n^\gamma n!}{\gamma(\gamma+1)(\gamma+2)\cdots(\gamma+n)}\le\tau.$$

As a result, for $n\ge i\ge N$, we have $0\le\frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma}-1\le\Gamma(\gamma)\tau$ and $0\le\frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma}-1\le\Gamma(\gamma)\tau$, which implies

$$\left|\frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma}-\frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma}\right|\le 2\Gamma(\gamma)\tau.$$

Furthermore we have

$$\begin{aligned}\left|\frac{i^\gamma}{n^\gamma}-\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}\right|&=\left|\frac{t^\gamma\Gamma(n+1)}{\Gamma(\gamma+n+1)}\right|\left|\frac{\Gamma(\gamma+n+1)}{\Gamma(n+1)i^\gamma}-\frac{\Gamma(\gamma+i+1)}{\Gamma(i+1)i^\gamma}\right|\\&\le 2\Gamma(\gamma)\tau\left|\frac{n^\gamma\Gamma(n+1)}{\Gamma(\gamma+n+1)}\right|\\&\le 2\Gamma(\gamma)\tau.\end{aligned}\tag{54}$$

Now we estimate the following summation

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})]\\\le&\frac{1}{n}\sum_{i=1}^{N}|(\gamma+1)(\frac{i}{n})^\gamma|+\sum_{i=1}^{N}|(\theta_{n,i})|+\frac{\gamma+1}{n}\sum_{i=N+1}^{n}\left|(\frac{i}{n})^\gamma-\frac{\Gamma(\gamma+i+1)\Gamma(n+1)}{\Gamma(\gamma+n+1)\Gamma(i+1)}\right|\\\le&\frac{N(\gamma+1)}{n}+\frac{N}{n}\theta_{n,N}+\frac{n-N}{n}\tau(\gamma+1)\\\le&(\gamma+1)(\frac{N}{2n}+\tau).\end{aligned}\tag{55}$$

Let $\tau\to 0$ and $n\to\infty$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})]=0.\tag{56}$$

Since

$$\begin{aligned}[(\gamma+1)^2(\frac{i}{n})^{2\gamma}-n^2(\theta_{n,i})^2]&=[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})][(\gamma+1)(\frac{i}{n})^\gamma+n(\theta_{n,i})]\\&\le 2(\gamma+1)[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})],\end{aligned}\tag{57}$$

we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)^2(\frac{i}{n})^{2\gamma}-n^2(\theta_{n,i})^2]\le 2(\gamma+1)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(\gamma+1)(\frac{i}{n})^\gamma-n(\theta_{n,i})]=0.\tag{58}$$

Finally, notice that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\gamma+1)^2(\frac{i}{n})^{2\gamma} = \int_0^1 (\gamma+1)^2 x^{2\gamma} dx = \frac{(\gamma+1)^2}{2\gamma+1},$$

We have proved the limitation in Theorem 2.7 holds with

$$\lim_{n\to\infty} n\sum_{i=1}^{n}\theta_{n,i}^2 = \frac{(\gamma+1)^2}{2\gamma+1}.$$

□

D ADDITIONAL EXPERIMENT RESULTS

We conducted the experiments in R version 4.1.1 (2021-08-10) 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.

The choice of $\alpha = 0.505$ follows the work of (Zhu et al., 2023). A similar choice can be found in (Chen et al., 2020) to make α slightly larger than 0.5. To verify our result when α is near 1, we also set $\alpha = 0.8$. The choices of parameters of polynomial decay averaging ($\gamma = 3$) and suffix averaging ($\kappa = 0.5$) are the same as the settings in (Shamir and Zhang, 2013) and (Rakhlin et al., 2012).

We report MSE (Table 2 and 3) and standard deviations of squared errors (Table 4 and 5) in the simulation of the mean estimation model and the squared loss model. As is shown in Tables 4 and 5, the standard deviation of adaptive weighted averaging is also the smallest among all averaging schemes.

Finally, we compare the empirical coverage rates of confidence intervals constructed by different averaging schemes in Table 6. The nominal coverage level is 95%. For $d = 5$ linear and logistic regression, we perform statistical inference for each coordinate of x^*, an arithmetic sequence from 0 to 1 with length 5, and report the overall averaged empirical coverage.

Table 2: MSE for Mean Estimation Model.

$\alpha = 0.505$				
Average Scheme	$n = 400$	$n = 800$	$n = 1200$	$n = 1600$
adaptive weights	0.00251625	0.00116611	0.00080357	0.00064877
ASGD	0.00254422	0.00118549	0.00080872	0.00066366
κ-suffix	0.00450813	0.00232837	0.00170125	0.00143947
Polynomial-Decay	0.00464877	0.00192599	0.00150453	0.00010830
Last Iterate	0.02771139	0.01643044	0.01495235	0.01226139
$\alpha = 0.8$				
Average Scheme	$n = 400$	$n = 800$	$n = 1200$	$n = 1600$
adaptive weights	0.00251625	0.00116611	0.00080357	0.00064877
ASGD	0.00308790	0.00139040	0.00091637	0.00071408
κ-suffix	0.00366752	0.00168517	0.00119971	0.00108589
Polynomial-Decay	.00328634	0.00154893	0.00119466	0.00103280
Last Iterate	0.00472179	0.00250589	0.00214805	0.00168590

Table 3: MSE for Squared Loss Model. Here $\alpha = 0.8$.

squared loss					
Average Scheme	$n = 2000$	$n = 4000$	$n = 6000$	$n = 8000$	$n = 10000$
adaptive weights	0.00052560	0.00025011	0.00016126	0.00012099	0.00009732
ASGD	0.00061791	0.00028463	0.00018354	0.00013391	0.00010643
κ-suffix	0.00104944	0.00046774	0.00029666	0.00022174	0.00018105
Polynomial-Decay	0.00075921	0.00038657	0.00025183	0.00019944	0.00016392
Last Iterate	0.00121235	0.00071972	0.00046712	0.00041441	0.00033334

Table 4: Standard Deviations of Squared Errors for the Mean Estimation Model.

$\alpha = 0.505$				
Average Scheme	$n = 400$	$n = 800$	$n = 1200$	$n = 1600$
adaptive weights	0.00378955	0.00179229	0.00121118	0.00103080
ASGD	0.00387207	0.00185897	0.00119596	0.00106196
κ-suffix	0.00592971	0.00372515	0.00247292	0.00227080
Polynomial-Decay	0.00609187	0.00360348	0.00289309	0.00231915
Last Iterate	0.04230295	0.02266438	0.02185713	0.01771681
$\alpha = 0.8$				
Average Scheme	$n = 400$	$n = 800$	$n = 1200$	$n = 1600$
adaptive weights	0.00378955	0.00179229	0.00121118	0.00103080
ASGD	0.00486709	0.00217177	0.00139634	0.00108259
κ-suffix	0.00538974	0.00313332	0.00169102	0.00169910
Polynomial-Decay	0.00465715	0.00276466	0.00164636	0.00166781
Last Iterate	0.00672886	0.00371866	0.00338224	0.00253484

Table 5: Standard Deviations of Squared Errors for the Squared Loss Model. Here $\alpha = 0.8$.

squared loss					
Average Scheme	$n = 2000$	$n = 4000$	$n = 6000$	$n = 8000$	$n = 10000$
adaptive weights	0.00033900	0.00015772	0.00010264	0.00007903	0.00006308
ASGD	0.00046388	0.00018273	0.00011797	0.00008705	0.00007048
κ-suffix	0.00088597	0.00035209	0.00019249	0.00014681	0.00012083
Polynomial-Decay	0.00047245	0.00025178	0.00015650	0.00013407	0.00010312
Last Iterate	0.00069180	0.00046476	0.00031687	0.00025734	0.00019574

Table 6: Empirical Coverage under the Linear and Logistic Models. Here $\alpha = 0.505$ Following the Choice in Zhu et al. (2024), and $n = 500$.

Method	Linear Model	Logistic Model
Adaptive weights	0.966	0.951
ASGD	0.964	0.949
κ-Suffix	0.947	0.943
Polynomial-decay	0.956	0.937