Submitted to the Annals of Statistics

THE STOCHASTIC GRADIENT DESCENT FROM A NONLINEAR TIME SERIES PERSPECTIVE

BY JIAQI LI[1,a], ZHIPENG LOU[2,b], STEFAN RICHTER[3,c] AND WEI BIAO WU[4,d]

[1]*Department of Statistics and Data Science, Washington University in St. Louis,* [a]*lijiaqi@wustl.edu*

[2]*Department of Statistics, University of Pittsburgh,* [b]*zhl318@pitt.edu*

[3]*Institute of Applied Mathematics, Heidelberg University,* [c]*stefan.richter@iwr.uni-heidelberg.de*

[4]*Department of Statistics, University of Chicago,* [d]*wbwu@galton.uchicago.edu*

This paper revisits the statistical behaviors of Stochastic Gradient Descent (SGD) through a novel perspective of time series analysis. Traditional approaches, mostly treating SGD as Markov chains, focused on convergence in probabilistic measures like the Wasserstein-2 distance. These approaches may face challenges when dealing with heavy-tailed noises and can fall short in handling non-stationary processes due to reliance on fixed initial points. To address these issues, we interpret the SGD as a nonlinear time series stemming from an iteration of random functions to establish convergence in Euclidean distance, specifically through Geometric Moment Contraction. This new perspective allows a deeper understanding of step size effects in the SGD procedure by revealing stationary solutions of derivative processes under mild conditions. Additionally, we extend this interpretation to the averaged SGD (ASGD) and provide refined statistical guarantees, including almost sure and moment convergence allowing heavy-tailed noises, quenched central limit theorems and invariance principles that hold with any initial points. Based on these asymptotic results, we introduce an innovative online inference method for ASGD with enhanced Richardson-Romberg extrapolation. We show that this estimator achieves the optimal mean squared error (MSE) rate, and we propose another bias-reduced variant. Numerical experiments demonstrate that our proposed empirical confidence intervals exhibit asymptotically precise coverage probabilities.

1. Introduction. We consider the optimization problem posed by a strongly convex objective function $G : \mathbb{R}^d \mapsto \mathbb{R}$, defined as below,

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} G(\theta), \quad \text{where } G(\theta) = \mathbb{E}_{X\sim\Pi}[g(\theta, X)]. \tag{1}$$

In (1), $g(\theta, X)$ represents the noise-perturbed measurement of $G(\theta)$ and X is a random variable sampled from the distribution Π. This optimization challenge has garnered tremendous attention across diverse fields, encompassing statistical learning [5, 8, 49, 53], optimization [24, 41, 43], and stochastic approximation [14, 31, 46, 54].

Traditional deterministic optimization techniques often come with high computational costs, making them less feasible for large-scale problems or streaming datasets. To tackle these issues, various online methods have been proposed. The Robbins-Monro algorithm [27, 47], also known as Stochastic Gradient Descent (SGD), was among the first and has become the most widely used method due to its simplicity and efficiency. Specifically, given a starting point $\theta_0 \in \mathbb{R}^d$, the k-th iteration of the SGD algorithm can be defined by the recursive

MSC2020 subject classifications: Primary 62F12, 62E20; secondary 62M10.

Keywords and phrases: stochastic gradient descent, iterated random function, derivative process, quenched central limit theorem, online inference.

procedure

$$\theta_k(\gamma) = \theta_{k-1}(\gamma) - \gamma \cdot \nabla g(\theta_{k-1}(\gamma), X_k), \quad k = 1, 2, \ldots, \tag{2}$$

where $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) observations sampled from some distribution Π, $\gamma > 0$ is a constant step size, $\nabla g(\theta, X)$ is a stochastic gradient vector of the objective function $g(\theta, X)$ with respect to the first argument $\theta \in \mathbb{R}^d$. We shall view $\theta_k(\gamma)$ as a random function of γ and θ_0 and it depends on $X_1, \ldots, X_k$. The primary goal of this paper is to provide a systematic asymptotic theory for $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ in the framework of iterative random functions.

The convergence of the SGD process has been widely investigated, such as in the pioneering work by [4, 16, 51] and the subsequent studies [17, 20, 32, 35, 36, 48, 58]. Many researchers interpreted the SGD process $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ as a homogeneous Markov chain. For example, [44] was among the first who investigated the stationary solutions of the constant step-size SGD, [14] highlighted convergence of the constant step-size SGD to a unique stationary distribution π_γ in the Wasserstein-2 distance, and more recently, [38] derived high-confidence estimation bounds. Note that most existing works established the convergence of $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ to π_γ in terms of probabilistic distances, such as the Wasserstein-2 distance [14, 38]. However, it can be quite challenging to generalize this type of measures to the cases with heavy-tailed noises, since their applied tools (e.g. Markov chain theory) require finite p-th central moments for $p \geq 2$ [38, 62]. Further, even in the case with finite variances, the convergence in a probabilistic measure is less effective in deriving an asymptotic theory for the non-stationary SGD process $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$.

To bridge these gaps, this paper puts the SGD in a nonlinear time series framework, specifically by conceptualizing the SGD process as an iterated random function [12, 60]. The new insight enables us to establish the geometric moment contraction (GMC) for the SGD, which then can be used to prove its convergence in the Euclidean norm. In particular, we show that for any two SGD sequences $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ and $\{\theta_k^\dagger(\gamma)\}_{k\in\mathbb{N}}$ from recursion (2) with different initial points $\theta_0(\gamma) = \theta_0, \theta_0^\dagger(\gamma) = \theta_0^\dagger \in \mathbb{R}^d$, respectively, if $0 < \gamma < \gamma(p)$,

$$\|\theta_k(\gamma) - \theta_k^\dagger(\gamma)\|_p := \left(\mathbb{E}|\theta_k(\gamma) - \theta_k^\dagger(\gamma)|^p\right)^{1/p} \leq \rho^k |\theta_0 - \theta_0^\dagger|, \quad p > 1, \tag{3}$$

where $\gamma(p) > 0$ and $\rho = \rho_{\gamma,p} \in (0, 1)$ are specified in Theorem 2.2, and $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. Such convergence ensures the existence and uniqueness of the stationary distribution π_γ. It is worth noticing that, in contrast to the theory for SGD convergence in the Wasserstein-2 distance, the GMC generalizes the analysis of SGD from $p \geq 2$ to all $p > 1$, strictly relaxing the moment assumptions on the noise in recent literature. Additionally, it allows to handle the non-stationarity of the SGD process due to fixed initial points. Thus, asymptotic results can be formulated for the complete observed SGD process starting from the first iteration.

In principle, our results for a constant step size γ can be extended to decaying step sizes $\gamma_n \asymp n^{-\beta}$ for some $1/2 < \beta < 1$. We shall focus on the constant step size in this paper, since it is a popular choice in practice [2]. It can benefit from faster convergence in over-parametrized situations [37, 40, 52, 57] and circumvent the problem of finding a suitable decay rate via β. Note that the SGD process with a constant step size does not converge, but oscillates around the mean of the stationary distribution with an average magnitude of $\gamma^{1/2}$ [44]. We provide a sharp bound for γ that ensures the contraction of SGD, offering valuable guidance for practical step-size selection. To gain a deeper insight into the variations in SGD as influenced by different step sizes, we use the GMC to show the existence of the first and second stationary derivative processes with respect to γ, i.e., $\{\partial_\gamma \theta_k(\gamma)\}_{k\in\mathbb{N}}$ and $\{\partial_\gamma^2 \theta_k(\gamma)\}_{k\in\mathbb{N}}$, under some mild conditions. Derivative processes are powerful tools in

measuring the deviation of a non-stationary sequence from stationarity; see for example [9, 10]. Motivated by this, we discuss the potential usage of derivative processes towards an explicit non-asymptotic expansion of the SGD process in terms of γ.

Besides the convergence and error margins of SGD-based estimators, it is also important to quantify the uncertainty of these point estimators by providing confidence intervals (CI). For example, [18, 19, 63] employed bootstrap strategies to create confidence intervals; [34, 56] alternated iterations of the SGD algorithm, drawing inferences from a non-asymptotic perspective; [6] introduced a cutting-edge class of regularized dual averaging to quantify uncertainties for online sparse methodologies. In this paper, we propose an online inference method for the well-known averaged SGD (ASGD), also referred to as Polyak-Ruppert averaging [45, 50]. ASGD can effectively reduce noise impact and has an asymptotically normal limiting distribution. Specifically, given an initial point $\theta_0 \in \mathbb{R}^d$, define the n-th ASGD iterate by

$$\bar{\theta}_n(\gamma) = \frac{1}{n}\sum_{k=1}^{n} \theta_k(\gamma), \quad n \geq 1. \tag{4}$$

Under some regularity conditions, $\{\bar{\theta}_n(\gamma)\}_{n\in\mathbb{N}}$ converges at rate $O_{\mathbb{P}}(n^{-1/2})$ to some point [11, 21]. See also other variants of ASGD formulations in [13, 15, 25, 33, 39]. However, since $\{\bar{\theta}_n(\gamma)\}_{n\in\mathbb{N}}$ is non-stationary due to the fixed initial point θ_0, its asymptotics heavily rely on the distance in which the SGD convergence is stated. For instance, to obtain $\mathcal{L}^p$-convergence of $\bar{\theta}_k(\gamma)$ from a Wasserstein-type convergence result for $p \geq 1$, one needs that the $2p$-th moment of $\nabla g(\theta, X)$ exists (see Remark 3), which seems to be an unnecessarily strong condition. To close this gap, we establish the asymptotic stationarity of ASGD using the GMC property and provide a quenched central limit theorem (CLT). Roughly speaking, for any arbitrary initial point $\theta_0 = v_0 \in \mathbb{R}^d$, we can achieve the following asymptotic normality

$$\mathbb{P}^{v_0}\big(\sqrt{n}\big(\bar{\theta}_n(\gamma) - \theta^*_\infty(\gamma)\big) \in A\big) \to \mathbb{P}\big(N(0, \Sigma^\circ(\gamma)) \in A\big), \quad \text{for measurable } A \subset \mathbb{R}^d, \tag{5}$$

as $n \to \infty$, where $\Sigma^\circ(\gamma)$ denotes the long-run covariance matrix

$$\Sigma^\circ(\gamma) = \sum_{k=-\infty}^{\infty} \operatorname{cov}(\theta^\circ_0(\gamma), \theta^\circ_k(\gamma)) \tag{6}$$

and $\theta^\circ_k(\gamma) \sim \pi_\gamma$ is the stationary SGD sequence following recursion (2) with

$$\theta^*_\infty(\gamma) = \lim_{k\to\infty} \mathbb{E}[\theta_k(\gamma)] = \mathbb{E}[\theta^\circ_1(\gamma)] = \int u\pi_\gamma(du). \tag{7}$$

This quenched CLT is particularly relevant and useful in practice because varying initial points are frequently employed in SGD, and a CLT formulated only for the stationary SGD sequence is insufficient for statistical inference. We can further establish an invariance principle (also known as, a functional CLT) for the process convergence, that is,

$$n^{-1/2}\Big\{\sum_{k=1}^{\lfloor nu\rfloor}\big(\theta_k(\gamma) - \theta^*_\infty(\gamma)\big), 0 \leq u \leq 1\Big\} \Rightarrow \{\Sigma^\circ(\gamma)^{1/2}\mathbb{B}(u), 0 \leq u \leq 1\}, \tag{8}$$

as $n \to \infty$, where $\mathbb{B}(u)$ is a standard d-dimensional Brownian motion. The weak convergence in (8) is also quenched, which means (8) is true for all initial points $\theta_0 \in \mathbb{R}^d$.

Before constructing confidence intervals for the ASGD, we shall note that the ASGD with a constant step size γ is biased. Therefore, we adopt a Richardson-Romberg extrapolation [55], firstly introduced to SGD by [14] to reduce the bias. We provide an enhanced non-asymptotic bias expansion of $\theta^*_\infty(\gamma)$ in terms of γ, i.e.,

$$\theta^*_\infty(\gamma) - \theta^* = \gamma\Delta_1 + \gamma^2\Delta_2 + \mathcal{R}^{(3)}_\gamma,$$

where Δ_1, Δ_2 are independent of γ, and the remaining term satisfies $\|\mathcal{R}_\gamma^{(3)}\|_p \le C\gamma^{5/2}$ for some constant $C > 0$ independent of γ. By this expansion, we can run three parallel SGD sequences $\theta_k(a_1\gamma)$, $\theta_k(a_2\gamma)$ and $\theta_k(a_3\gamma)$, $k \in \mathbb{N}$, with three different step sizes $a_1\gamma$, $a_2\gamma$ and $a_3\gamma$, where the constants $a_1, a_2, a_3 > 0$ are explicitly provided. Then, we define an extrapolated ASGD estimator as

$$\widehat{\theta}_n(\gamma) = b_1\bar{\theta}_n(a_1\gamma) + b_2\bar{\theta}_n(a_2\gamma) + b_3\bar{\theta}_n(a_3\gamma), \tag{9}$$

with some weighting constants b_1, b_2, b_3. By choosing the constants a_i, b_i $(i = 1, 2, 3)$ suitably, we can achieve a cancellation of the bias components $\gamma\Delta_1, \gamma^2\Delta_2$ (the detailed expressions are deferred to (28)).

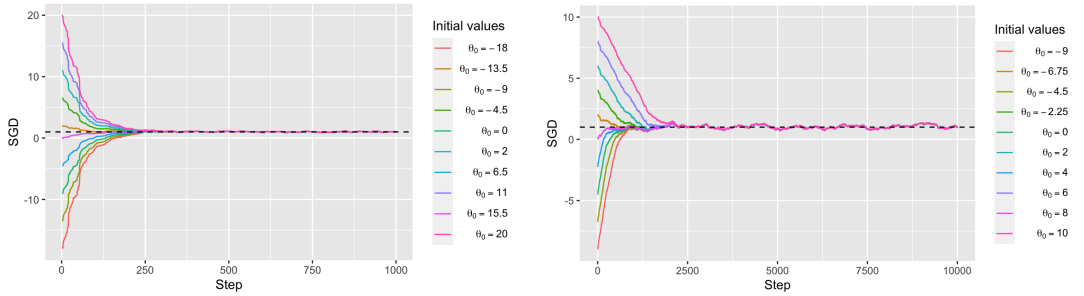
Subsequently, we propose a novel online inference method for $\widehat{\theta}_n(\gamma)$ by a blocking-based online estimator of the long-run covariance matrix $\Sigma(\gamma)$. The empirical CI is detailed in (50). Compared to other online inference methods such as [7, 64], we show that our proposed online estimator can achieve the optimal mean square error (MSE) rate in case of dependence (see Remark 5). Moreover, we introduce another variant of the online estimator to reduce the bias due to the dependency between SGD iterates. The bias-reduced CI exhibits more precise coverage probabilities in numerical experiments.

Contributions. This study introduces technical tools in nonlinear time series to the machine learning community by providing a new interpretation for the SGD process. Our primary contributions are three-fold: (i) conceptualize the evolution of SGD as an iterative random function and establish its convergence in the Euclidean norm with sharp conditions on γ, effectively addressing the challenges in heavy-tails and non-stationarity (Section 2); (ii) provide the quenched limiting distributions for the (extrapolated) ASGD with arbitrary initial points, facilitating a bias-reduced online inference method with optimal MSE (Sections 3 & 4); (iii) first time analyze step-size impacts on the SGD process by using derivative processes, which can be of independent interests (Section 5).

Notation. For a vector $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$ and $q > 0$, we denote $|v|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $|v| = |v|_2$. For a matrix A, denote the Frobenius norm by $|A|_2$. Denote the identity matrix in $\mathbb{R}^{d\times d}$ by $\boldsymbol{I}_d$ and the vector $(1, \ldots, 1)^\top \in \mathbb{R}^d$ by $\mathbf{1}_d$. For any $s > 0$ and a random vector X, we say $X \in \mathcal{L}^s$ if $\|X\|_s = (\mathbb{E}|X|_2^s)^{1/s} < \infty$. For two positive number sequences (a_n) and (b_n), we say $a_n = O(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n \asymp b_n$) if there exists $C > 0$ such that $a_n/b_n \le C$ (resp. $1/C \le a_n/b_n \le C$) for all large n, and write $a_n = o(b_n)$ or $a_n \ll b_n$ if $a_n/b_n \to 0$ as $n \to \infty$. Let (X_n) and (Y_n) be two sequences of random variables. Write $X_n = o_{\mathbb{P}}(Y_n)$ if $X_n/Y_n \to 0$ in probability as $n \to \infty$. Consider the finite dimensional Euclidean space $\mathbb{R}^d$ embedded with the canonical inner product $\langle\cdot,\cdot\rangle$. For two real vector spaces $F, G \subset \mathbb{R}^d$, we denote the tensor product of F and G by $F \otimes G$. For each d-dimensional vector $x \in F$ and $y \in G$, we denote the tensor product of x and y by $x \otimes y \in F \otimes G$, and denote the k-th tensor power of x by $x^{\otimes k} \in F^{\otimes k}$, where $F^{\otimes k}$ is the k-th tensor power of F.

2. Geometric Moment Contraction. Note that the SGD sequence $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ in (2) is a Markov chain because $\theta_k(\gamma)$ only depends on its past through $\theta_{k-1}(\gamma)$. Since $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ is typically non-stationary when the initial point $\theta_0 \in \mathbb{R}^d$ is fixed, we are interested in the existence of stationary distribution of $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ and the related convergence rate of $\theta_k(\gamma)$ towards this stationary distribution. The recursive form of the SGD imposes a complicated dependency structure on $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$, making it challenging to establish an asymptotic theory. To provide refined asymptotics, tools from iterated random functions [12, 60] are applied.

First, the geometric moment contracting property (see Definition 2.1) of the recursive function (2) is introduced. According to [60], this property enables us to establish the weak convergence of $\theta_k(\gamma)$ to a unique stationary distribution $\pi_\gamma \in \mathcal{L}^p(\mathbb{R}^d)$, for some $p \ge 1$, that is $\theta_k(\gamma) \Rightarrow \pi_\gamma$, as $k \to \infty$, at an exponential convergence rate (see Theorem 2.2). Unlike the



(a) Linear regression. (b) Logistic regression.

Fig 1: A simulation example for the convergence trace of SGD estimates relative to the step numbers, where the optimum $\theta^* = 1$ and the constant step size $\gamma = 0.05$. The total numbers of steps are 10^3 and 10^4 for linear regression and logistic regression, respectively.

Wasserstein distance utilized in previous literature such as [14], the geometric moment contraction directly provides convergence on the moments $\|\theta_k(\gamma) - \theta^*_\infty(\gamma)\|_p$ with $p > 1$, as $k \to \infty$. This property substantially facilitates the construction of an asymptotic theory for SGD iterates and it offers a natural base from which limiting distributions and concentration inequalities related to $\theta_k(\gamma)$ can be derived.

DEFINITION 2.1 (Geometric-moment contraction). Let X_i, X'_j, be i.i.d. random variables, $i, j \in \mathbb{Z}$. Consider the causal process

$$Y_k = H(X_k, \ldots, X_1, X_0, X_{-1}, \ldots), \tag{10}$$

where $H(\cdot)$ is a measurable function so that Y_k is well-defined and $\mathbb{E}|Y_k|^p < \infty$, $p \geq 1$. We say that Y_k is *geometric-moment contracting (GMC)* if there exist some constants $C > 0$ and $0 < r < 1$ such that for all $k \in \mathbb{N}$,

$$\mathbb{E}|Y_k - Y_k^\dagger|^p \leq Cr^k, \text{ where } Y_k^\dagger = H(X_k, \ldots, X_1, X'_0, X'_{-1}, \ldots) \tag{11}$$

is a coupled version of Y_k with $X_j, j \leq 0$, in the latter replaced by i.i.d. random variables $X'_j, j \leq 0$, which are also independent of X_i, $i \in \mathbb{Z}$.

In this paper, we shall show that under suitable conditions, GMC as defined above holds for $Y_k = \theta_k(\gamma)$ with $\theta_k(\gamma)$ generated from the SGD algorithm in (2) (see Theorem 2.2). We emphasize that we can allow infinite variance for Y_k, which was rarely investigated in the literature.

REMARK 1 (Exponential convergence of the SGD iterates). To motivate the geometric-moment contraction for the SGD iterates, we shall first present a simulation example to show that for varied initial points $\theta_0 \in \mathbb{R}$, the SGD iterates $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ forget the initial point exponentially fast for well-conditioned problems such as strongly convex objectives. In Figure 1, we show the convergence trace of SGD estimates for the least-square loss of the linear regression and the negative log-likelihood loss of the penalized logistic regression. We defer the details of the synthetic data to Section 6. As shown in Figure 1, both models demonstrate that SGD iterations converge to the global minimum at an exponential rate across various initial points. This inspires us to directly verify the SGD convergence in the Euclidean distance beyond probabilistic measures (e.g. Wasserstein distance), that is, to establish the geometric moment contraction for the SGD sequence.

By the example in Remark 1 and other literature on the SGD convergence rates, one can see that the Markov chain $\{\theta_k(\gamma)\}_{k\geq 1}$ converges exponentially fast for different constant step sizes, and the initial conditions (i.e. arbitrary starting points θ_0 and θ_0') are forgotten exponentially quickly. However, all the existing literature only provides the theoretical convergence in terms of probability measures such as the Wasserstein distance, while we show that the SGD iterates $\{\theta_k(\gamma)\}_{k\geq 1}$ satisfy the GMC property in (11). Based on this convergence in the Euclidean distance, we can further provide refined asymptotic properties of the SGD iterates in the subsequent sections.

Following the framework of iterated random functions [12, 60], we define

$$F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d, \quad (\theta, X) \mapsto F_X(\theta) = \theta - \gamma \cdot \nabla g(\theta, X), \tag{12}$$

which is a measurable function. Let $\theta_0, \theta_0^\dagger \in \mathbb{R}^d$ be some given initial points independent of $X_1, \ldots, X_n$. Then, we can write the SGD iterates in (2) based on θ_0 or $\theta_0^\dagger$, respectively, as

$$\theta_k = F_{X_k}(\theta_{k-1}) = F_{X_k} \circ \cdots \circ F_{X_1}(\theta_0) \quad \text{and} \quad \theta_k^\dagger = F_{X_k} \circ \cdots \circ F_{X_1}(\theta_0^\dagger). \tag{13}$$

We shall introduce some conditions on convexity and smoothness.

ASSUMPTION 1 (μ-strong convexity). Let $m(\theta) = \mathbb{E}[\nabla g(\theta, X)]$. Assume that for any $x \in \mathbb{R}$, $\theta \mapsto g(\theta, x)$ is a continuously differentiable function. Additionally, suppose that there exists a constant $\mu > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$,

$$\langle m(\theta) - m(\theta'), \theta - \theta' \rangle \geq \mu |\theta - \theta'|^2.$$

ASSUMPTION 2 (Stochastic Lipschitz continuity). Let $p > 1$. (i) For θ^* defined in (1), assume that $\|\nabla g(\theta^*, X)\|_p < \infty$; (ii) Assume that there exists some constant $L_p > 0$ such that for all θ, $\theta' \in \mathbb{R}^d$,

$$\|\nabla g(\theta, X) - \nabla g(\theta', X)\|_p \leq L_p |\theta - \theta'|.$$

Assumption 1 is equivalent to the following conditions that are commonly seen in the relevant literature [14, 23]: (a) if $g(\theta, X)$ is twice differentiable, then the Hessian matrix $\mathbb{E}\nabla^2 g(\theta, X)$ is positive definite with smallest eigenvalue of $\mathbb{E}\nabla^2 g(\theta, X)$ larger than μ; (b) for any $t \in (0, 1)$,

$$m(t\theta_1 + (1-t)\theta_2) \leq t m(\theta_1) + (1-t) m(\theta_2) - t(1-t)(\mu/2)|\theta_1 - \theta_2|^2;$$

and (c) $m(\theta_1) \geq m(\theta_2) + \nabla m(\theta_1)^\top (\theta_1 - \theta_2) + (\mu/2)|\theta_1 - \theta_2|^2$. Assumption 2(i) is a mild condition on the p-th moment of the stochastic gradient. Assumption 2(ii) requires stochastic Lipschitz continuity of the gradient, which also indicates that the function $m(\theta)$ is L_1-smooth, by noting that $\|\cdot\|_p \geq \|\cdot\|_1 = \mathbb{E}|\cdot|$. Similar assumptions have been adopted in the literature on SGD asymptotics; see for example [7] on parameter estimation and inference with the SGD. It is noteworthy that Assumption 2 only requires $p > 1$, which allows for infinite variance in random noises. This is a nontrivial extension that generalizes the previous studies which supposed $p \geq 2$ to the cases where $1 < p < 2$.

Theorem 2.2 below implies that for sufficiently small $\gamma > 0$, we have the weak convergence

$$\theta_k(\gamma) \Rightarrow \pi_\gamma \text{ as } k \to \infty, \tag{14}$$

where π_γ is the stationary distribution. We define the boundary $\gamma(p) > 0$ as the root to the equation

$$(1 + \gamma L_p)^p = 1 + p\gamma L_p + p\mu\gamma \text{ if } p \geq 2 \tag{15}$$

and

$$\gamma(p) = (p\mu/(2^{2-p}L_p^p))^{1/(p-1)} \text{ if } 1 < p < 2. \tag{16}$$

When $p \geq 2$, since the function $g(u) = ((1+u)^p - 1 - pu)/u = p\int_0^1((1+ut)^{p-1} - 1)dt$ is strictly increasing in $u \in (0,\infty)$, $\lim_{u\downarrow 0} g(u) = 0$, and $\lim_{u\uparrow\infty} g(u) = \infty$, there is a unique root $u = u(p)$ to the equation $g(u) = p\mu/L_p$, which implies that the root $\gamma(p)$ to equation (15) is $u(p)/L_p$. A particularly interesting special case is $p = 2$. In this case both (15) and (16) with $p \uparrow 2$ yield $\gamma(2) = 2\mu/L_2^2$, suggesting a smooth transition at the borderline $p = 2$, and a central limit theorem is given in Theorem 3.2.

THEOREM 2.2 (Geometric moment contraction). *Let $p > 1$. Recall (15) and (16) for $\gamma(p)$. Suppose that Assumptions 1 and 2 hold and $0 < \gamma < \gamma(p)$. Then, $F_X(\theta)$ defined in (12) satisfies a geometric moment contraction, that is, for all $\theta, \theta' \in \mathbb{R}^d$,*

$$\|F_X(\theta) - F_X(\theta')\|_p \leq \rho|\theta - \theta'|, \tag{17}$$

where

$$\rho = \rho_{\gamma,p} \quad \text{and} \quad \rho_{\gamma,p}^p = \begin{cases} 1 - p\mu\gamma + \gamma^p 2^{2-p}L_p^p < 1, & \text{if } 1 < p < 2; \\ (1+\gamma L_p)^p - p\gamma L_p - p\mu\gamma < 1, & \text{if } p \geq 2. \end{cases} \tag{18}$$

Consequently, for any initial point $\theta_0 \in \mathbb{R}^d$, there exists a unique stationary distribution π_γ which does not depend on θ_0, such that $\theta_k(\gamma) \Rightarrow \pi_\gamma$ as $k \to \infty$, and the distribution π_γ has a finite p-th moment $\int |u|^p \pi_\gamma(du) < \infty$. Additionally, let $\theta_0^\circ(\gamma) \sim \pi_\gamma$ follow the stationary distribution and define $\theta_k^\circ(\gamma)$, $k \geq 1$, according to the iteration (2). Then the following geometric moment contraction holds:

$$\|\theta_k(\gamma) - \theta_k^\circ(\gamma)\|_p \leq \rho^k \|\theta_0 - \theta_0^\circ(\gamma)\|_p = \rho^k \left[\int |\theta_0 - u|^p \pi_\gamma(du)\right]^{1/p}. \tag{19}$$

Theorem 2.2 provides an explicit sufficient condition for the existence of the p-th moment of the stationary distribution based on the step size γ. In general, the existence of higher-order moments asks for smaller values of γ. The geometric contraction (19) gives an explicit rate for the coupling of $\theta_k(\gamma)$ and $\theta_k^\circ(\gamma)$ and it shows that the initial point θ_0 is forgotten exponentially quickly. The latter property will be used in our quenched limit theorems in Section 3 which concerns the asymptotic properties for partial sums of $\theta_k(\gamma)$ with any initial point θ_0, a paradigm which is highly relevant in the study of SGD.

Note that to establish the convergence of the SGD to the stationary distribution π_γ in terms of the Wasserstein-2 distance, it is usually assumed that the gradient $\nabla g(\theta, X)$ is almost surely L-co-coercive (see for example Assumption A4(p) in [14]). However, this assumption is stronger than our Assumption 2 since it poses constraints beyond Lipschitz continuity directly on ∇g. For example, it requires the observed data to be a.s. bounded or with bounded kurtosis. We shall consider the following linear model as an example to show that our conditions are less restrictive.

REMARK 2 (Range of γ in linear regression). Let $X_k = (Z_k, Y_k) \in \mathbb{R}^d \times \mathbb{R}$, $k \geq 1$, be i.i.d. random samples following the linear regression model $Y_k = Z_k^\top \theta^* + \epsilon_k$, where $\theta^* \in \mathbb{R}^d$ is the population parameter vector of interest and $\epsilon_k \in \mathbb{R}$, $k \geq 1$, are i.i.d. random noise independent of $\{Z_k\}_{k\geq 1}$. Our Assumption 1 and Assumption 2 hold with $\mu = \lambda_{\min}\{\mathbb{E}(Z_1 Z_1^\top)\}$ and $L_2 = \sup_{\delta\in\mathbb{R}^d:|\delta|=1} \|Z_1 Z_1^\top \delta\|_2$, respectively, where $\lambda_{\min}\{\cdot\}$ denotes the smallest eigenvalue. Consequently, Theorem 2.2 ensures the GMC of the sequence $\{\theta_k(\gamma)\}_{k\geq 1}$ as long as

$$0 < \gamma < \frac{2\lambda_{\min}\{\mathbb{E}(Z_1 Z_1^\top)\}}{\sup_{\delta\in\mathbb{R}^d:|\delta|=1} \|Z_1 Z_1^\top \delta\|_2^2}. \tag{20}$$

It is worth mentioning that the range above for γ can not be improved in general. To see this, suppose now $d = 1$, $\theta^* = 0$, and $\mathbb{E}(Z_1^4) = \varpi\mathbb{E}(Z_1^2)$ for some $0 < \varpi < \infty$. Then (20) reduces to $0 < \gamma < 2/\varpi$. In particular, for any $\gamma \geq 2/\varpi$, the second moment of the stationary distribution π_γ is no longer finite. Assume otherwise, the second moment of the asymptotic distribution π_γ is finite for $\gamma \geq 2/\varpi$, that is

$$\varsigma^2 := \int u^2 \pi_\gamma(du) < \infty.$$

Then, in view of (2), for any $\gamma \geq 2/\varpi$, we have

$$\varsigma^2 = \varsigma^2\mathbb{E}(1 - 2\gamma Z_1^2 + \gamma^2 Z_1^4) + \gamma^2\mathbb{E}(Z_1^2)\mathbb{E}(\epsilon_1^2) \geq \varsigma^2 + \gamma^2\mathbb{E}(Z_1^2)\mathbb{E}(\epsilon_1^2) > \varsigma^2,$$

which leads to a contradiction.

REMARK 3 (Extension to heavy-tailed noises). Note that by providing the convergence of SGD in terms of the Wasserstein-2 distance, one requires at least the $2p$-th finite moment to bound the distance $\mathbb{E}|\bar{\theta}_k(\gamma) - \theta_\infty^*|^p$, and the moment contraction can only be established for $p \geq 2$. The reason is that the martingale decomposition is adopted in the previous literature, that is

$$F_X(\theta) = \theta - \gamma m(\theta) + \gamma D(\theta, X), \tag{21}$$

where

$$m(\theta) = \mathbb{E}[\nabla g(\theta, X)], \quad D(\theta, X) = m(\theta) - \nabla g(\theta, X). \tag{22}$$

This decomposition may lead to a loose upper bound of $\|F_X(\theta) - F_X(\theta')\|_p$ when applying the Cauchy-Schwarz inequality twice. This type of condition on higher-order finite moments (at least fourth moment) can be too restricted, especially if the input data has heavy tails or are influenced by outliers. Differently, we address this issue by directly providing the bound for $\|F_X(\theta) - F_X(\theta')\|_p$ without the martingale decomposition in (21). As such, we only require the existence of p-th moment instead of $2p$ for the GMC property of the SGD, and the moment contraction holds for all $p > 1$. This improvement can accommodate a wider variety of distributions of the observations, even those with an infinite variance.

The GMC property in Theorem 2.2 characterizes the dependencies of the SGD process $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ on the initial conditions $\theta_0 \in \mathbb{R}^d$, which decays exponentially fast as the step number $k \to \infty$. This theoretical guarantee can be especially useful in practical applications, since SGD is often implemented with a variety of starting points θ_0 to enhance the robustness and effectiveness of the optimization process. Based on GMC, we shall further introduce the convergence for the ASGD sequence $\{\bar{\theta}_n(\gamma)\}_{n\in\mathbb{N}}$ in Section 3. These convergence results shall facilitate more refined limiting behaviour of ASGD, such as the quenched central limit theorems, which provides a fundamental base for the statistical inference of ASGD estimates with any arbitrary initial conditions (see Section 4).

3. Asymptotic Theory. In this section, we first introduce the asymptotic properties of the ASGD estimate $\bar{\theta}_n(\gamma)$ defined in (4), for all $p > 1$, by using the geometric-moment contraction provided in Theorem 2.2. In particular, we give a bound on the moment $\|\bar{\theta}_n(\gamma) - \theta_\infty^*(\gamma)\|_p$ and provide an almost sure convergence result, for all $p > 1$. Then, for $p \geq 2$, we will provide the quenched central limit theorem for $\bar{\theta}_n(\gamma)$ and establish the quenched invariance principle with $p > 2$.

3.1. *Quenched Central Limit Theorem.* We shall start with the moment convergence and almost sure convergence of the ASGD estimate $\bar{\theta}_n(\gamma)$.

THEOREM 3.1 (Convergence of the ASGD). *Let $p > 1$. Recall $\theta^*_\infty(\gamma)$ defined in (7). Under the conditions of Theorem 2.2, we have the p-th order moment convergence*

$$\|\bar{\theta}_n(\gamma) - \theta^*_\infty(\gamma)\|_p = O(n^{1/(p\wedge 2)-1}),$$

and the following almost sure convergence:

(i) *for $1 < p < 2$, $|\bar{\theta}_n(\gamma) - \theta^*_\infty(\gamma)| = o_{a.s.}(n^{1/p-1})$;*
(ii) *for $p \geq 2$, $|\bar{\theta}_n(\gamma) - \theta^*_\infty(\gamma)| = O_{a.s.}\big((n^{-1}\log\log(n))^{1/2}\big)$.*

The above convergence results also hold for the sample average $\bar{\theta}^\circ_n(\gamma) = n^{-1}\sum_{k=1}^n \theta^\circ_k(\gamma)$ of the stationary SGD process $\{\theta^\circ_k(\gamma)\}_{k\in\mathbb{N}}$.

Next, we shall show the (quenched) central limit theorem for this Markov chain, that is, for any arbitrary initial point $\theta_0 \in \mathbb{R}^d$, the central limit theorem holds for $\bar{\theta}_n(\gamma)$. Since the SGD sequence with an arbitrary starting point is a non-stationary Markov chain, we need to derive an approximation by a stationary process by following (19) of Theorem 2.2, which is the key to the proof of the quenched limit theorems.

THEOREM 3.2 (Quenched central limit theorem). *For the recursion (2) with i.i.d. samples X_i, $i \geq 1$, suppose that Assumptions 1 and 2 hold with some $p \geq 2$. Let $\gamma_1, \ldots, \gamma_\ell \in (0, 2\mu/L_2^2)$ be $\ell \geq 1$ different step sizes and*

$$S^\circ_n = \sum_{k=1}^n \varepsilon^\circ_k, \text{ where } \varepsilon^\circ_k = \mathrm{vec}(\theta^\circ_k(\gamma_1) - \theta^*_\infty(\gamma_1), \ldots, \theta^\circ_k(\gamma_\ell) - \theta^*_\infty(\gamma_\ell)), \tag{23}$$

where $\mathrm{vec}(v_1, \ldots, v_l) = (v_1^\top, \ldots, v_l^\top)^\top$. Let $\Sigma^\circ = \sum_{k=-\infty}^\infty \mathrm{cov}(\varepsilon^\circ_0, \varepsilon^\circ_k)$ be the long run covariance matrix of the stationary $(d\ell)$-dimensional vector process $(\varepsilon^\circ_k)_{k\in\mathbb{Z}}$. Similarly we define S_n as S°_n with $\theta^\circ_k(\gamma)$ in the latter replaced by $\theta_k(\gamma)$. Then $n^{-1/2}S^\circ_n \Rightarrow N(0, \Sigma^\circ)$, and for any arbitrary initial point $\mathrm{vec}(\theta_0(\gamma_1), \ldots, \theta_0(\gamma_\ell)) \in \mathbb{R}^{dl}$, $n^{-1/2}S_n \Rightarrow N(0, \Sigma^\circ)$.

Based on the asymptotic normality results of the ASGD provided above, we can construct asymptotically valid confidence intervals for the model parameter θ. To facilitate the use of the result in practice, we propose a new online inference method in Section 4 for the estimation of the long-run covariance matrix $\Sigma(\gamma)$ of the bias-reduced ASGD (see (30)). The long run covariance matrix Σ° of the stationary SGD in Theorem 3.2 can be similarly estimated. Two concrete examples for constructing confidence intervals are provided on a linear regression model and a logistic regression model; see Section 6 for more details.

3.2. *Quenched Invariance Principles.* The seminal work of Komlós, Major, and Tusnády [29, 30], presents an optimal Wiener approximation (KMT approximation) for the partial sums of i.i.d. random vectors. Their findings have become indispensable tools in both probability and statistics. [26] extends this result to the dependent case, followed by [3] who further generalized the KMT approximation to the multiple time series. Building on these previous works, in this section, we adapt the KMT approximation towards the recursive procedures, specifically by providing an invariance principle and its quenched version for the vector-valued ASGD processes $\{\bar{\theta}_k(\gamma)\}_{k\in\mathbb{N}}$.

The rigorous statement of the quenched invariance principle is as follows.

THEOREM 3.3 (Quenched invariance principle). *Suppose that Assumptions 1 and 2 hold with some $p > 2$. Recall Theorem 3.2 for S_n and S_n° with step sizes $\gamma_1, \ldots, \gamma_\ell \in (0, \gamma(p))$. Then there exists a (richer) probability space $(\Omega_c, \mathcal{A}_c, \mathbb{P}_c)$ on which we can define random vectors $\varepsilon_k^c \in \mathbb{R}^{dl}$ with the partial sum process $S_i^c = \sum_{k=1}^i \varepsilon_k^c$, and a Gaussian process $G_i^c = \sum_{k=1}^i Z_k^c$, where $\{Z_k^c\}_{k\geq 1}$ are independent Gaussian $N(0, \boldsymbol{I}_{dl})$ random vectors, such that $(S_i^c)_{1\leq i\leq n} \overset{\mathcal{D}}{=} (S_i^\circ)_{1\leq i\leq n}$ and*

$$\max_{i\leq n} \left| S_i^c - \Sigma^{\circ 1/2} G_i^c \right| = o_{\mathbb{P}}(n^{1/p}), \quad \text{in } (\Omega_c, \mathcal{A}_c, \mathbb{P}_c). \tag{24}$$

Additionally, the above approximation also holds for S_n with any arbitrary initial point $\mathrm{vec}(\theta_0(\gamma_1), \ldots, \theta_0(\gamma_\ell)) \in \mathbb{R}^{dl}$.

With the functional CLT provided above, we can establish the Gaussian approximation for the asymptotic distribution of the entire path of the SGD estimates over iterations, not just their final or averaged values. For example, instead of a single confidence interval for the final estimate or at some specific step k, the FCLT enables the construction of sequential confidence intervals that cover the entire trajectory of the SGD estimates.

4. Online Inference. In this section, we propose an online statistical inference method for the population parameter θ^* based on the SGD algorithm [7]. To this end, we shall first provide an explicit non-asymptotic bias expansion for the ASGD estimates in terms of the step size γ (see Theorem 4.1) and introduce a Richardson-Romberg extrapolated ASGD estimator utilizing three different step sizes, which effectively cancels out the bias term on the order of γ and γ^2.

4.1. *Richardson-Romberg Extrapolation.* In the context of the SGD, the iterates can exhibit a bias due to the constant step size when approaching the optimal solution. This bias arises because the noisy gradient estimates do not average out entirely, even if the iterations are performed infinitely often. To resolve this issue, [14] first time adopted the Richardson-Romberg extrapolation, which is a technique primarily used in numerical integration to improve the accuracy of a solution by exploiting the knowledge of its error's rate of convergence. For example, [23] provides a general bias expansion of Richardson-Romberg extrapolated estimators under mixing conditions. In this section, we establish a bias expansion to cancel out both γ and γ^2 terms in the extrapolated ASGD estimator expansion.

First, we provide a theoretical foundation, specifically by giving a non-asymptotic expansion of the bias between the ASGD solution in (4) and the global minimum θ^*, that is, $\mathbb{E}(\bar{\theta}_k(\gamma) - \theta^*)$. This bias can be further decomposed into two parts as follows:

$$\mathbb{E}[\bar{\theta}_k(\gamma) - \theta^*] = \mathbb{E}[\bar{\theta}_k(\gamma) - \theta_k^\circ(\gamma)] + \mathbb{E}[\theta_k^\circ(\gamma) - \theta^*]. \tag{25}$$

Here $\theta_k^\circ(\gamma)$ has the stationary distribution π_γ for all $k \in \mathbb{N}$. The first part $\mathbb{E}[\bar{\theta}_k(\gamma) - \theta_k^\circ(\gamma)]$ converges to zero at a rate of $O(1/k)$ as $k \to \infty$ by applying the GMC property shown in Theorem 2.2. Moreover, we provide an expansion of the second part $\mathbb{E}[\bar{\theta}_k^\circ(\gamma) - \theta^*]$ for small step sizes $\gamma > 0$, with an error of the order of $O(\gamma^{5/2})$. In addition, we shall show an explicit formula for the coefficients of γ and γ^2 in the expansion. To achieve these goals, the following assumption is needed.

ASSUMPTION 3 (Bounded variation of derivatives). Assume that the objective function $g(\theta, X)$ is six times continuously differentiable with respect to θ. Moreover, for $p \geq 2$, there exists some positive constant $M_p < \infty$ such that $\|g^{(j)}(\theta^*, X)\|_p \leq M_p$ for any $j \in \{2, \ldots, 6\}$.

To establish a refined non-asymptotic bias expansion in (25) in terms of γ, we further develop the residual term derived by Theorem 4 in [14]. To this end, the smoothness condition we pose in Assumption 3 requires the existence of the sixth order derivative of the objective function $g(\theta, x)$, which is slightly higher than the one in [14]. However, it is worth noticing that despite the similar methodology we used in this study compared to existing works, which is mainly based on the Taylor expansion of the iterative function $F_X(\theta)$ around the global optimum θ^*, we also provide insights into another potential approach that relies on the derivative processes of the SGD iterates with respect to γ. We refer to a detailed discussion in Section 5.

THEOREM 4.1 (Bias expansion). *Recall the tensor product denoted by $\otimes$. Suppose that Assumptions 1–3 hold and that the constant step size $\gamma > 0$ is small enough. Recall $m(\theta) = \mathbb{E}[\nabla g(\theta, X)]$. Denote the Hessian matrix and the third-order tensor by $\boldsymbol{M}_1 \in \mathbb{R}^{d\times d}$ and $\boldsymbol{M}_2 \in \mathbb{R}^{d\times d\times d}$ respectively with*

$$\boldsymbol{M}_1 = \nabla m(\theta^*) \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \nabla m(\theta^*),$$
$$\boldsymbol{M}_2 = \nabla m(\theta^*) \otimes \boldsymbol{I}_d \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \nabla m(\theta^*) \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \boldsymbol{I}_d \otimes \nabla m(\theta^*).$$

Then, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are invertible. Also, for the SGD iterates $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ in (2) with any initial point $\theta_0 \in \mathbb{R}^d$, we have

$$\theta^*_\infty(\gamma) - \theta^* = \gamma \cdot \Delta_1 + \gamma^2 \cdot \Delta_2 + \mathcal{R}^{(3)}_\gamma, \tag{26}$$

where Δ_1 and Δ_2 are two vectors in $\mathbb{R}^d$ with

$$\Delta_1 = [\nabla m(\theta^*)]^{-1} \nabla^2 m(\theta^*) \boldsymbol{M}_1^{-1} \mathbb{E}\big\{[\nabla^2 g(\theta^*, X_n)]^{\otimes 2}\big\},$$
$$\Delta_2 = [\nabla m(\theta^*)]^{-1} \big(\nabla^3 m(\theta^*)/3\big) \boldsymbol{M}_2^{-1} \mathbb{E}\big\{[\nabla^2 g(\theta^*, X_n)]^{\otimes 3}\big\},$$

which are both independent of γ, and $\mathcal{R}^{(3)}_\gamma \in \mathbb{R}^d$ with $\|\mathcal{R}^{(3)}_\gamma\|_p \le C\gamma^{5/2}$ for some constant $C > 0$ independent of γ.

Now we proceed with the basic idea of the Richardson-Romberg extrapolation in SGD. In particular, we compute the trajectory of the SGD at multiple step-sizes and then combine these trajectories in a specific manner to produce a new, bias-reduced trajectory. This approach allows for a more accurate convergence to the solution, leveraging the information from multiple step-sizes to correct the inherent bias in constant step-size SGD. Specifically, we consider three different constant step sizes $a_1\gamma$, $a_2\gamma$ and $a_3\gamma$ with $\gamma > 0$ and the constants $a_1, a_2, a_3 > 0$. Then, we have three sequences of stochastic gradient descent iterates $\{\theta_k(a_1\gamma)\}_{k\in\mathbb{N}}$, $\{\theta_k(a_2\gamma)\}_{k\in\mathbb{N}}$ and $\{\theta_k(a_3\gamma)\}_{k\in\mathbb{N}}$ respectively, with the initial point $\theta_0 \in \mathbb{R}^d$. Motivated by Theorem 4.1, we propose to estimate θ^* via the Richardson-Romberg extrapolation as follows [14, 22],

$$\widehat{\theta}_n(\gamma) = b_1\bar{\theta}_n(a_1\gamma) + b_2\bar{\theta}_n(a_2\gamma) + b_3\bar{\theta}_n(a_3\gamma), \tag{27}$$

where

$$b_1 = \frac{a_2 a_3}{a_1^2 - a_1a_2 - a_1a_3 + a_2a_3},$$
$$b_2 = -\frac{a_1 a_3}{a_1a_2 - a_1a_3 - a_2^2 + a_2a_3},$$
$$b_3 = \frac{a_1 a_2}{a_1a_2 - a_1a_3 - a_2a_3 + a_3^2}. \tag{28}$$

One can see that the constants b_1, b_2, b_3 defined above satisfy $b_1 + b_2 + b_3 = 1$, $a_1 b_1 + a_2 b_2 + a_3 b_3 = 0$ and $a_1^2 b_1 + a_2^2 b_2 + a_3^2 b_3 = 0$, which leads to a cancellation of the terms $\gamma\Delta_1$ and $\gamma^2\Delta_2$ in (26).

Before presenting the asymptotic distribution of $\widehat{\theta}_n(\gamma)$, we shall introduce some basic definitions and notation. Denote the stationary SGD sequence with the step size $a\gamma$ by $\{\theta_k^\circ(a\gamma)\}_{k\in\mathbb{Z}}$, for some constant $a > 0$, which follows the stationary distribution $\pi_{a\gamma}$. Moreover, for the same constants a_i, b_i, $i = 1, 2, 3$, specified in (27), we define the stationary process $\{\vartheta_k^\circ(\gamma)\}_{k\in\mathbb{Z}}$, where

$$\vartheta_k^\circ(\gamma) = b_1\theta_k^\circ(a_1\gamma) + b_2\theta_k^\circ(a_2\gamma) + b_3\theta_k^\circ(a_3\gamma). \tag{29}$$

Consequently, the long run covariance matrix of the stationary sequence $\{\vartheta_k^\circ(\gamma)\}_{k\in\mathbb{Z}}$ is given by

$$\Sigma(\gamma) = \sum_{k\in\mathbb{Z}} \mathrm{cov}\{\vartheta_0^\circ(\gamma), \vartheta_k^\circ(\gamma)\}. \tag{30}$$

COROLLARY 4.2 (Quenched CLT for the extrapolated ASGD). *Consider the extrapolated ASGD estimator $\widehat{\theta}_n(\gamma)$ in (27). Under the conditions of Theorem 3.2, for any initial point $\theta_0 \in \mathbb{R}^d$,*

$$\sqrt{n}[\widehat{\theta}_n(\gamma) - \vartheta_\infty^*(\gamma)] \Rightarrow N(0, \Sigma(\gamma)), \quad \text{as } n \to \infty$$

where the center $\vartheta_\infty^*(\gamma) = b_1\theta_\infty^*(a_1\gamma) + b_2\theta_\infty^*(a_2\gamma) + b_3\theta_\infty^*(a_3\gamma)$ with $\theta_\infty^*(\gamma)$ defined in (7) and the long-run covariance matrix $\Sigma(\gamma)$ is defined in (30).

As a direct application of Theorem 3.2, by the Crámer-Wold device, the central limit theorem also holds for the Richardson-Romberg bias corrected estimate (27) as stated in Corollary 4.2, which is a linear combination of $\theta_k^\circ(\gamma_s)$, $k \in \mathbb{N}$, $s = 1, \ldots, \ell$.

REMARK 4 (Refined bias expansion). We establish the theoretical properties of the SGD estimator with three different step sizes utilized in the Richardson-Romberg extrapolation in Corollary 4.2, which can be extended to even more step sizes on a simple side. However, these results can be hardly achieved by using the second-moment Wasserstein distance applied in [14], since the SGD sequence can be non-stationary and the bound in the $\mathcal{L}^2$-norm may not be sufficient and is nontrivial to be generalized to higher orders. Therefore, we can see the advantage of using the Euclidean distance (i.e., the GMC property) to characterize how the SGD iterates depend on the initial conditions. Here again, we shall emphasize the benefits from leveraging the technical tools in nonlinear time series to study the heavily-dependent random processes.

4.2. *Online Estimator for Long-Run Covariance Matrices.* For simplicity of notation, throughout Sections 4.2 and 4.3, we consider the estimator $\widehat{\theta}_n(\gamma)$ in (27) with $a_1 = 1, a_2 = 2, a_3 = 0$ and $b_1 = 2, b_2 = -1, b_3 = 0$. In particular, we write $\vartheta_k(\gamma) = 2\theta_k(\gamma) - \theta_k(2\gamma)$ for each $k \in \mathbb{N}$. Then we can rewrite this bias-corrected ASGD estimator $\widehat{\theta}_n(\gamma)$ as

$$\widehat{\theta}_n(\gamma) = \frac{1}{n}\sum_{k=1}^{n} \vartheta_k(\gamma).$$

Throughout this paper, we assume that $\lambda_{\min}(\Sigma(\gamma)) \geq c_0 > 0$ for some positive constant $c_0 < \infty$. In this section, we construct recursive consistent estimator for the long-run covariance matrix $\Sigma(\gamma)$ defined in (30). It has been studied in the literature for time series [59, 61, 64].

Let $\{\eta_m\}_{m\in\mathbb{N}}$ be a strictly increasing positive integer-valued sequence. For each $m \in \mathbb{N}$, we define the consecutive blocks $\{B_m\}_{m\in\mathbb{N}}$ as

$$B_m = \{\eta_m, \eta_m + 1, \ldots, \eta_{m+1} - 1\}.$$

Throughout, we assume that $\eta_{m+1} - \eta_m \to \infty$ as $m \to \infty$. Define $\psi(n) = \max\{m \in \mathbb{N} : \eta_m \le n\}$ and $\phi(n) = \eta_{\psi(n)}$ for each $n \in \mathbb{N}$. Then the nonoverlapping recursive estimator of the long-run covariance matrix $\Sigma(\gamma)$ is defined by $V_n(\gamma)/n$, where

$$V_n(\gamma) = \sum_{m=1}^{\psi(n)-1} \left(\sum_{k\in B_m} \{\vartheta_k(\gamma) - \widehat{\theta}_n(\gamma)\} \right)^{\otimes 2} + \left(\sum_{k=\phi(n)}^{n} \{\vartheta_k(\gamma) - \widehat{\theta}_n(\gamma)\} \right)^{\otimes 2}. \tag{31}$$

For simplicity of notation, we write $\delta_\eta(n) = n - \phi(n) + 1$,

$$\mathcal{S}_m(\gamma) = \sum_{k\in B_m} \vartheta_k(\gamma) \quad \text{and} \quad \mathcal{R}_n(\gamma) = \sum_{k=\phi(n)}^{n} \vartheta_k(\gamma).$$

To facilitate the recursive computation of $V_n(\gamma)$, we rewrite it as follows,

$$\begin{aligned} V_n(\gamma) &= \left(\sum_{m=1}^{\psi(n)-1} \mathcal{S}_m(\gamma)^{\otimes 2} + \mathcal{R}_n(\gamma)^{\otimes 2} \right) + \left(\sum_{m=1}^{\psi(n)-1} |B_m|^2 + |\delta_\eta(n)|^2 \right) \widehat{\theta}_n(\gamma)^{\otimes 2} \\ &\quad - \left(\sum_{m=1}^{\psi(n)-1} |B_m| \mathcal{S}_m(\gamma) + \delta_\eta(n) \mathcal{R}_n(\gamma) \right) \widehat{\theta}_n(\gamma)^\top \\ &\quad - \widehat{\theta}_n(\gamma) \left(\sum_{m=1}^{\psi(n)-1} |B_m| \mathcal{S}_m(\gamma) + \delta_\eta(n) \mathcal{R}_n(\gamma) \right)^\top \\ &=: \mathcal{V}_n(\gamma) + K_n \widehat{\theta}_n(\gamma)^{\otimes 2} - H_n(\gamma) \widehat{\theta}_n(\gamma)^\top - \widehat{\theta}_n(\gamma) H_n(\gamma)^\top. \end{aligned}$$

Then, it reduces to recursively compute $\{\mathcal{V}_n(\gamma), K_n, H_n(\gamma), \widehat{\theta}_n(\gamma)\}$ for any $n \in \mathbb{N}$. To this end, we propose a recursive algorithm which only requires $O(1)$ storage and establish the convergence rate of the proposed estimator.

The rational behind the recursive computation for $V_n(\gamma)$ is as follows: if $n + 1 < \eta_{\psi(n)+1}$, then $n + 1$ still belongs to the block $B_{\psi(n)}$ and $\psi(n + 1) = \psi(n)$. Also we have $\mathcal{R}_{n+1}(\gamma) = \mathcal{R}_n(\gamma) + \vartheta_{n+1}(\gamma)$ and $\delta_\eta(n+1) = \delta_\eta(n) + 1$. Consequently $\{K_{n+1}, \mathcal{V}_{n+1}(\gamma), H_{n+1}(\gamma)\}$ can be recursively updated via $K_{n+1} = K_n - |\delta_\eta(n)|^2 + |\delta_\eta(n + 1)|^2$,

$$\begin{aligned} \mathcal{V}_{n+1}(\gamma) &= \mathcal{V}_n(\gamma) - \mathcal{R}_n(\gamma)^{\otimes 2} + \mathcal{R}_{n+1}(\gamma)^{\otimes 2}, \\ H_{n+1}(\gamma) &= H_n(\gamma) - \delta_\eta(n) \mathcal{R}_n(\gamma) + \delta_\eta(n + 1) \mathcal{R}_{n+1}(\gamma). \end{aligned}$$

Otherwise, if $n + 1 = \eta_{\psi(n)}$, we have $\psi(n + 1) = \psi(n) + 1$. Hence $\mathcal{R}_{n+1}(\gamma) = \vartheta_{n+1}(\gamma)$ and $\delta_\eta(n + 1) = 1$. In this case, $\{K_{n+1}, \mathcal{V}_{n+1}(\gamma), H_{n+1}(\gamma)\}$ can be recursively updated via $K_{n+1} = K_n + 1$,

$$\mathcal{V}_{n+1}(\gamma) = \mathcal{V}_n(\gamma) + \mathcal{R}_{n+1}(\gamma)^{\otimes 2} \text{ and } H_{n+1}(\gamma) = H_n(\gamma) + \mathcal{R}_{n+1}(\gamma).$$

Consequently, given $\vartheta_1(\gamma), \ldots, \vartheta_n(\gamma)$, our estimator for the long-run covariance matrix $\Sigma(\gamma)$ is given by

$$\widehat{\Sigma}_n(\gamma) = \frac{1}{n} V_n(\gamma).$$

In summary, our recursive algorithm for $V_n(\gamma)$ is stated in Algorithm 1 below.

Algorithm 1: Recursive estimation of the long-run covariance matrix

Data: Observations $X_1, X_2, \ldots, X_n$; objective function $g(\cdot)$; constant step size γ; predefined sequence $\{\eta_m\}_{m\in\mathbb{N}}$

Result: Extrapolated ASGD estimator $\widehat{\theta}_{n+1}(\gamma)$; estimated long-run covariance matrix $\widehat{\Sigma}_{n+1}(\gamma)$

Initialize $\theta_0(\gamma) = \theta_0(2\gamma) = \widehat{\theta}_0(\gamma) = \vartheta_0(\gamma) = \mathcal{R}_0(\gamma) \leftarrow 0$,
$\psi(0) \leftarrow 1$, $\delta_\eta(0) \leftarrow 1$, $K_0 = H_0(\gamma) \leftarrow 1$, $\mathcal{V}_0(\gamma) = V_0(\gamma) \leftarrow 0$

for $n = 0, 1, 2, 3, \ldots$ **do**

$\theta_{n+1}(\gamma) \leftarrow \theta_n(\gamma) - \gamma\nabla g(\theta_n(\gamma), X_{n+1})$; `/* SGD estimate */`

$\theta_{n+1}(2\gamma) \leftarrow \theta_n(2\gamma) - 2\gamma\nabla g(\theta_n(\gamma), X_{n+1})$;

$\vartheta_{n+1}(\gamma) \leftarrow 2\theta_{n+1}(\gamma) - \theta_{n+1}(2\gamma)$;

$\widehat{\theta}_{n+1}(\gamma) \leftarrow \{n\widehat{\theta}_n(\gamma) + \vartheta_{n+1}(\gamma)\}/(n+1)$; `/* ASGD estimate */`

if $n + 1 < \eta_{\psi(n)+1}$ **then**

$\mathcal{R}_{n+1}(\gamma) \leftarrow \mathcal{R}_n(\gamma) + \vartheta_{n+1}(\gamma)$, $\delta_\eta(n+1) \leftarrow \delta_\eta(n) + 1$;

$K_{n+1} \leftarrow K_n - \delta_\eta^2(n) + \delta_\eta^2(n+1)$, $\psi(n+1) \leftarrow \psi(n)$;

$H_{n+1}(\gamma) \leftarrow H_n(\gamma) - \delta_\eta(n)\mathcal{R}_n(\gamma) + \delta_\eta(n+1)\mathcal{R}_{n+1}(\gamma)$;

$\mathcal{V}_{n+1}(\gamma) \leftarrow \mathcal{V}_n(\gamma) - \mathcal{R}_n(\gamma)^{\otimes 2} + \mathcal{R}_{n+1}(\gamma)^{\otimes 2}$;

else

$\mathcal{R}_{n+1}(\gamma) \leftarrow \vartheta_{n+1}(\gamma)$, $\delta_\eta(n+1) \leftarrow 1$;

$\psi(n+1) \leftarrow \psi(n) + 1$;

$K_{n+1} \leftarrow K_n + 1$, $H_{n+1}(\gamma) \leftarrow H_n(\gamma) + \mathcal{R}_{n+1}(\gamma)$;

$\mathcal{V}_{n+1}(\gamma) \leftarrow \mathcal{V}_n(\gamma) + \mathcal{R}_{n+1}(\gamma)^{\otimes 2}$;

end

$V_{n+1}(\gamma) \leftarrow \mathcal{V}_{n+1}(\gamma) + K_{n+1}\widehat{\theta}_{n+1}(\gamma)^{\otimes 2} - H_{n+1}(\gamma)\widehat{\theta}_{n+1}(\gamma)^\top - \widehat{\theta}_{n+1}(\gamma)H_{n+1}(\gamma)^\top$;

$\widehat{\Sigma}_{n+1}(\gamma) \leftarrow V_{n+1}(\gamma)/(n+1)$; `/* Recursive long-run covariance matrix estimate */`

end

THEOREM 4.3 (Precision of $\widehat{\Sigma}_n(\gamma)$). *Let $\eta_m = \lfloor c_1 m^\beta \rfloor$ for some $c_1 > 0$ and $\beta > 1$. Let Assumption 1 and 2 hold with $p = 4$. Then, we have*

$$\mathbb{E}|\widehat{\Sigma}_n(\gamma) - \Sigma(\gamma)|_2^2 \lesssim n^{(2/\beta-2)\vee(-1/\beta)}.$$

REMARK 5 (Optimal convergence rate). Theorem 4.3 establishes the mean squared error (MSE) of the recursive estimator for the long-run covariance matrix $\Sigma(\gamma)$. In particular, when $\beta = 3/2$, we have

$$\mathbb{E}|\widehat{\Sigma}_n(\gamma) - \Sigma(\gamma)|_2^2 \lesssim n^{-2/3}. \tag{32}$$

As far as we know, (32) reveals that our recursive estimator $\widehat{\Sigma}_n(\gamma)$ attains the optimal convergence rate of long-run covariance matrix estimation. Recently, [64] proposed a similar recursive estimator for $\Sigma(\gamma)$ in the context of SGD estimation with decaying step size $\gamma_t = c_0 t^{-\alpha}$ for some $\alpha \in (1/2, 1)$. Under some regularity conditions, the convergence rate of their recursive estimator is $n^{-(1-\alpha)/4}$ which is much slower than our bound $n^{-1/3}$ in (32). □

4.3. *Bias-Reduced Online Estimator.* In this section, we introduce a new estimator for the long-run variance based on an idea for bias reduction. To illustrate the idea, we first consider the recursive estimator based on the stationary sequence $\{\vartheta_k^\circ(\gamma)\}_{k\in\mathbb{Z}}$ in (29). Recall that $\phi(k) = \eta_m$ for any $k \in \{1, ..., n\}$ such that $\eta_m \le k < \eta_{m+1}$. Also, recall in (7), we write $\theta_\infty^*(\gamma) = \mathbb{E}\{\theta_k^\circ(\gamma)\}$ for $\gamma > 0$. Based on $\{\vartheta_k^\circ(\gamma)\}_{k\in\mathbb{Z}}$, define $V_n^\circ(\gamma) = \sum_{k=1}^n Q_k^\circ(\gamma)$, where for each $k \in \{1, \ldots, n\}$,

$$Q_k^\circ(\gamma) = (\vartheta_k^\circ(\gamma) - \vartheta_\infty^*(\gamma))^{\otimes 2} + \sum_{\ell=\phi(k)}^{k-1} (\vartheta_k^\circ(\gamma) - \vartheta_\infty^*(\gamma))(\vartheta_\ell^\circ(\gamma) - \vartheta_\infty^*(\gamma))^\top$$

$$+\sum_{\ell=\phi(k)}^{k-1}(\vartheta_\ell^\circ(\gamma)-\vartheta_\infty^*(\gamma))(\vartheta_k^\circ(\gamma)-\vartheta_\infty^*(\gamma))^\top.$$

In view of (30), the bias of $Q_k^\circ(\gamma)$ is

$$\mathbb{E}\{Q_k^\circ(\gamma)\}-\Sigma(\gamma)=\sum_{|\ell|>k-\phi(k)}\operatorname{cov}\{\vartheta_0^\circ(\gamma),\vartheta_\ell^\circ(\gamma)\}.$$

which reveals that $Q_k^\circ(\gamma)$ can have a larger bias for smaller value of $k-\phi(k)$. This motivates us to propose a new estimator by excluding these terms with small value of $k-\phi(k)$. More specifically, we define a sequence of positive integers $\{\tau(m)\}_{m\in\mathbb{N}}$ such that $\tau(m)\le(\eta_{m+1}-\eta_m)$ for all $n\in\mathbb{N}$. Then, the bias reduced estimator of $V_n^\circ(\gamma)$ is defined by excluding all these $Q_k^\circ(\gamma)$ with $k-\phi(k)$ smaller than the threshold $\tau(\psi(k))$, that is,

$$U_n^\circ(\gamma)=\sum_{k=1}^n Q_k^\circ(\gamma)\mathbb{I}\{k-\phi(k)\ge\tau(\psi(k))\}.$$

This motivates us to propose the bias-reduced counterpart of $V_n(\gamma)$ as follows:

$$U_n(\gamma)=\sum_{k=1}^n\Bigg((\vartheta_k(\gamma)-\widehat{\theta}_n(\gamma))^{\otimes2}+\sum_{\ell=\phi(k)}^{k-1}\{\vartheta_k(\gamma)-\widehat{\theta}_n(\gamma)\}\{\vartheta_\ell(\gamma)-\widehat{\theta}_n(\gamma)\}^\top$$
$$+\sum_{\ell=\phi(k)}^{k-1}\{\vartheta_\ell(\gamma)-\widehat{\theta}_n(\gamma)\}\{\vartheta_k(\gamma)-\widehat{\theta}_n(\gamma)\}^\top\Bigg)\mathbb{I}\{k-\phi(k)\ge\tau(\psi(k))\}.$$

For simplicity of notation, we write $\mathcal{E}_k(\tau)=\{k-\phi(k)\ge\tau(\psi(k))\}$ for each $k\in\mathbb{N}$. For each block B_m, we use $\mathcal{S}_m^\tau(\gamma)=\sum_{k=\eta_m}^{\eta_m+\tau(m)-1}\vartheta_k(\gamma)$ to denote the part excluded for bias reduction from $\mathcal{S}_m(\gamma)$. To facilitate the recursive computation of $U_n(\gamma)$, we first rewrite it as

$$U_n(\gamma)=\mathcal{U}_n(\gamma)+\mathcal{K}_n\widehat{\theta}_n(\gamma)^{\otimes2}-\mathcal{H}_n(\gamma)\widehat{\theta}_n(\gamma)^\top-\widehat{\theta}_n(\gamma)\mathcal{H}_n(\gamma)^\top,$$

where

$$\mathcal{U}_n(\gamma)=\sum_{k=1}^n\Bigg(\vartheta_k(\gamma)^{\otimes2}+\sum_{\ell=\phi(k)}^{k-1}\vartheta_k(\gamma)\vartheta_\ell(\gamma)^\top+\sum_{\ell=\phi(k)}^{k-1}\vartheta_\ell(\gamma)\vartheta_k(\gamma)^\top\Bigg)\mathbb{I}\{\mathcal{E}_k(\tau)\},$$

$$\mathcal{K}_n=\sum_{m=1}^{\psi(n)-1}\{|B_m|^2-\tau^2(m)\}+\{|\delta_\eta(n)|^2-\tau^2(\psi(n))\}\times\mathbb{I}\{\mathcal{E}_n(\tau)\},$$

$$\mathcal{H}_n(\gamma)=\sum_{m=1}^{\psi(n)-1}\{|B_m|\mathcal{S}_m(\gamma)-\tau(m)\mathcal{S}_m^\tau(\gamma)\}$$
$$+\{\delta_\eta(n)\mathcal{R}_n(\gamma)-\tau(\psi(n))\mathcal{S}_{\psi(n)}^\tau(\gamma)\}\times\mathbb{I}\{\mathcal{E}_n(\tau)\}.$$

Following the idea of Algorithm 1, our algorithm for recursively computing $U_n(\gamma)$ is summarized in Algorithm 2. Given $\{\vartheta_k(\gamma)\}_{k\in\{1,\ldots,n\}}$, our estimator for $\Sigma(\gamma)$ with bias reduction is then defined by

$$\widetilde{\Sigma}_n(\gamma)=\frac{1}{\kappa(n)}U_n(\gamma),\quad\text{where }\kappa(n)=n-\sum_{m=1}^{\psi(n)-1}\tau(m)-\min\{n-\phi(n),\tau(\psi(n))\}.$$

We provide the consistency rate of this estimator $\widetilde{\Sigma}_n(\gamma)$ in Theorem 4.4.

Algorithm 2: Recursive estimation of the bias-reduced long-run covariance matrix

Data: Observations $X_1, X_2, \ldots, X_n$; objective function $g(\cdot)$; constant step size γ; predefined sequences $\{\eta_m\}_{m\in\mathbb{N}}$ and $\{\tau(m)\}_{m\in\mathbb{N}}$

Result: Extrapolated ASGD estimator $\widehat{\theta}_{n+1}(\gamma)$; estimated long-run covariance matrix $\widetilde{\Sigma}_{n+1}(\gamma)$

Initialize $\theta_0(\gamma)=\theta_0(2\gamma)=\widehat{\theta}_0(\gamma)=\vartheta_0(\gamma)=\mathcal{R}_0(\gamma)=\mathcal{U}_0(\gamma)=\mathcal{S}^{\tau}_{\psi(0)}(\gamma)\leftarrow 0$,
$\psi(0)\leftarrow 1$, $\phi(0)\leftarrow \eta_1$, $\delta_\eta(0)\leftarrow 1$, $\mathcal{K}_0=\mathcal{H}_0(\gamma)\leftarrow 1$, $\mathcal{V}_0(\gamma)=V_0(\gamma)\leftarrow 0$, $\kappa(0)\leftarrow 1$

for $n=0,1,2,3,\ldots$ **do**

$\theta_{n+1}(\gamma)\leftarrow \theta_n(\gamma)-\gamma\nabla g(\theta_n(\gamma),X_{n+1})$; `/* SGD estimate */`

$\theta_{n+1}(2\gamma)\leftarrow \theta_n(2\gamma)-2\gamma\nabla g(\theta_n(\gamma),X_{n+1})$;

$\vartheta_{n+1}(\gamma)\leftarrow 2\theta_{n+1}(\gamma)-\theta_{n+1}(2\gamma)$;

$\widehat{\theta}_{n+1}(\gamma)\leftarrow\{n\widehat{\theta}_n(\gamma)+\vartheta_{n+1}(\gamma)\}/(n+1)$; `/* ASGD estimate */`

if $n+1<\eta_{\psi(n)+1}$ **then**

$\psi(n+1)\leftarrow\psi(n)$, $\phi(n+1)\leftarrow\phi(n)$, $\delta(n+1)\leftarrow\delta(n)+1$;

$\mathcal{R}_{n+1}(\gamma)\leftarrow\mathcal{R}_n(\gamma)+\vartheta_{n+1}(\gamma)$;

if $n+1\geq\phi(n+1)+\tau(\psi(n+1))$ **then**

if $n\geq\phi(n)+\tau(\psi(n))$ **then**

$\mathcal{U}_{n+1}(\gamma)\leftarrow\mathcal{U}_n(\gamma)+\vartheta_{n+1}(\gamma)^{\otimes 2}+\vartheta_{n+1}(\gamma)\mathcal{R}_n(\gamma)^{\top}+\mathcal{R}_n(\gamma)\vartheta_{n+1}(\gamma)^{\top}$;

$\mathcal{K}_{n+1}\leftarrow\mathcal{K}_n-\delta_\eta^2(n)+\delta_\eta^2(n+1)$, $\mathcal{S}^{\tau}_{\psi(n+1)}(\gamma)\leftarrow\mathcal{S}^{\tau}_{\psi(n)}(\gamma)$;

$\mathcal{H}_{n+1}(\gamma)\leftarrow\mathcal{H}_n(\gamma)-\delta_\eta(n)\mathcal{R}_n(\gamma)+\delta_\eta(n+1)\mathcal{R}_{n+1}(\gamma)$;

else

$\mathcal{U}_{n+1}(\gamma)\leftarrow\mathcal{U}_n(\gamma)+\vartheta_{n+1}(\gamma)^{\otimes 2}+\vartheta_{n+1}(\gamma)\mathcal{R}_n(\gamma)^{\top}+\mathcal{R}_n(\gamma)\vartheta_{n+1}(\gamma)^{\top}$;

$\mathcal{K}_{n+1}\leftarrow\mathcal{K}_n+\delta_\eta^2(n+1)-\tau^2(\psi(n+1))$;

$\mathcal{H}_{n+1}(\gamma)\leftarrow\mathcal{H}_n(\gamma)+\delta_\eta(n+1)\mathcal{R}_{n+1}(\gamma)-\tau(\psi(n+1))\mathcal{S}^{\tau}_{\psi(n+1)}(\gamma)$;

end

else

$\mathcal{K}_{n+1}\leftarrow\mathcal{K}_n$, $\mathcal{S}^{\tau}_{\psi(n+1)}(\gamma)\leftarrow 0$;

$\mathcal{U}_{n+1}(\gamma)\leftarrow\mathcal{U}_n(\gamma)$, $\mathcal{H}_{n+1}\leftarrow\mathcal{H}_n$;

end

else

$\psi(n+1)\leftarrow\psi(n)+1$, $\phi(n+1)=n+1$, $\delta_\eta(n+1)=1$;

$\mathcal{R}_{n+1}(\gamma)\leftarrow\vartheta_{n+1}(\gamma)$, $\mathcal{S}^{\tau}_{\psi(n+1)}(\gamma)\leftarrow 0$;

$\mathcal{U}_{n+1}(\gamma)\leftarrow\mathcal{U}_n(\gamma)$, $\mathcal{K}_{n+1}\leftarrow\mathcal{K}_n$, $\mathcal{H}_{n+1}(\gamma)\leftarrow\mathcal{H}_n(\gamma)$;

end

$U_{n+1}(\gamma)\leftarrow\mathcal{U}_{n+1}(\gamma)+\mathcal{K}_{n+1}\widehat{\theta}_{n+1}(\gamma)^{\otimes 2}-\mathcal{H}_{n+1}(\gamma)\widehat{\theta}_{n+1}(\gamma)^{\top}-\widehat{\theta}_{n+1}(\gamma)\mathcal{H}_{n+1}(\gamma)^{\top}$;

$\kappa(n+1)\leftarrow n+1-\sum_{m=1}^{\psi(n+1)-1}\tau(m)-\min\{n+1-\phi(n+1),\tau(\psi(n+1))\}$;

$\widetilde{\Sigma}_{n+1}(\gamma)\leftarrow U_{n+1}(\gamma)/\kappa(n+1)$; `/* Recursive long-run covariance matrix estimate with bias reduction */`

end

THEOREM 4.4 (Precision of $\widetilde{\Sigma}_n(\gamma)$). *Suppose that the conditions of Theorem 4.3 hold.*

(1) *Let* $\eta_m=\lfloor c_1 m^{\beta}\rfloor$ *and* $\tau(m)=\lfloor c_2\log m\rfloor$ *for some* $c_1>0$, $\beta>1$ *and* $c_2>0$. *We have*

$$\mathbb{E}|\widetilde{\Sigma}_n(\gamma)-\Sigma(\gamma)|_2^2\lesssim n^{-1/\beta}. \tag{33}$$

(2) *Let* $\eta_m=\lfloor \bar{c}_1 m\log m\rfloor$ *and* $\tau(m)=\lfloor \bar{c}_2\log m\rfloor$ *for some* $\bar{c}_1>\bar{c}_2>0$. *Then*

$$\mathbb{E}|\widetilde{\Sigma}_n(\gamma)-\Sigma(\gamma)|_2^2\lesssim \frac{\log n}{n}. \tag{34}$$

We remark that the convergence rate in (34) is quite sharp, by noting that the mean squared error for a parametric model is typically of order $O(1/n)$.

5. Derivative Analysis of Step-Size Effects on SGD. Notably, the magnitude of the SGD update is controlled by the step size, since a small learning rate can lead to slow convergence, while a large learning rate can cause oscillations or even divergence from the minimum. The ideal learning rate can lie somewhere in between and is typically found through empirical tuning. In Theorem 2.2, we have shown that the GMC holds for $\{\theta_k(\gamma)\}_{k\in\mathbb{N}}$ when $0<\gamma<\gamma(p)$ with $\gamma(p)$ defined in (15) and (16). However, the theory on the optimal step size is quite limited in the literature. We here provide a new insight by investigating the derivative process of SGD with respect to γ using the GMC property.

Recall that the stationary SGD process $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ can be written as

$$\theta_k^\circ(\gamma)=\theta_{k-1}^\circ(\gamma)-\gamma\nabla g(\theta_{k-1}^\circ(\gamma),X_k),\quad k\geq 1. \tag{35}$$

We shall respectively show the existence of the stationary solutions for the first and second-order derivative processes under some mild conditions.

5.1. *Existence of the First Derivative Processes.* First, we consider the first-order derivative process. Assume that $g(\theta,X)$ is second-differentiable with respect to the first argument. Then, by taking the derivative of both sides of $\theta_k^\circ(\gamma)$ in (35) with respect to γ, we have

$$\partial_\gamma\theta_k^\circ(\gamma)=\big[\boldsymbol{I}_d-\gamma\nabla^2 g(\theta_{k-1}^\circ(\gamma),X_n)\big]\partial_\gamma\theta_{k-1}^\circ(\gamma)-\nabla g(\theta_{k-1}^\circ(\gamma),X_k), \tag{36}$$

where $\nabla^2 g(\theta,X)$ is a Hessian matrix in $\mathbb{R}^{d\times d}$ with respect to θ. Based on this expression, we consider a new recursive sequence $\{\dot\theta_k(\gamma)\}_{k\in\mathbb{N}}$ defined by

$$\dot\theta_k(\gamma)=\boldsymbol{A}(\theta_{k-1}^\circ(\gamma),X_k)\dot\theta_{k-1}(\gamma)-\nabla g(\theta_{k-1}^\circ(\gamma),X_k), \tag{37}$$

where the random coefficient matrix $\boldsymbol{A}(\theta,X)\in\mathbb{R}^{d\times d}$ is defined by

$$\boldsymbol{A}(\theta,X)=\boldsymbol{I}_d-\gamma\nabla^2 g(\theta,X). \tag{38}$$

THEOREM 5.1 (Existence of stationary first derivative process). *Assume that $g(\theta,X)$ in (1) is a second-differentiable function with respect to θ. Denote a unit vector by $\delta\in\mathbb{R}^d$, i.e., $|\delta|=1$. For an arbitrary $p\geq 2$, we define the random variable*

$$A_p(\theta):=\sup_{\{\delta\in\mathbb{R}^d,|\delta|=1\}}\mathbb{E}_{X\sim\Pi}\big[|\boldsymbol{A}(\theta,X)\delta|^p\big],$$

where the random matrix $\boldsymbol{A}(\theta,X)$ is defined in (38). Consider the stationary SGD iterates $\theta^\circ(\gamma)\sim\pi_\gamma$ as defined in (35). Assume that we can choose some step size $\gamma\in(0,\gamma(p))$ with $\gamma(p)$ defined in Theorem 2.2 such that

$$\mathbb{E}_{\theta^\circ(\gamma)\sim\pi_\gamma}[A_p(\theta^\circ(\gamma))]<1. \tag{39}$$

Then, a stationary solution of the first derivative procedure $\{\dot\theta_k(\gamma)\}_{k\in\mathbb{N}}$ in (37) exists.

REMARK 6 (Uniform moment bound). Note that as a special case, when $p=2$, the assumption on the uniform moment bound for the coefficient matrix $\boldsymbol{A}(\theta,X)$ in (39) can also imply

$$\big\|\mathbb{E}\big[\boldsymbol{A}(\theta_{k-1}^\circ(\gamma),X_k)^\top\boldsymbol{A}(\theta_{k-1}^\circ(\gamma),X_k)\big]\big\|_{\mathrm{op}}<1,$$

for all $k\geq 1$, where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. This is a mild assumption that is frequently adopted in the literature for a strong convergence result; see for instance Assumption 1 in [8]. Many practical applications can fulfill the condition (39). For example, the linear regression with the least-square loss as the objective function satisfies the moment condition in Theorem 5.1.

5.2. *Existence of the Second Derivative Processes.* Recall the bias expansion for the ASGD estimate in Theorem 4.1, where we have provided an explicit form for the coefficient of γ^2 term and bound the remaining term by $O(\gamma^{5/2})$ under some additional smoothness conditions in Assumption 3. Therefore, beyond the first derivative process with respect to γ, we would like to further establish the existence of a stationary solution for the second-order derivative process, which can potentially provide a more refined bias expansion for the SGD estimates.

Recall the definition of $\boldsymbol{A}(\theta, X)$ in (38). For notation simplicity, we shall omit the dependence of $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ on the step size γ in this section and denote them by $\{\theta_k^\circ\}_{k\in\mathbb{N}}$ when no confusion should be caused. We take the derivative on the both sides of the first derivative process $\{\dot{\theta}_k(\gamma)\}_{k\in\mathbb{N}}$ in (37) and obtain

$$\begin{aligned}\partial_\gamma^2\theta_k^\circ = &\big[-\nabla^2 g(\theta_{k-1}^\circ, X_k) - \gamma\partial_\gamma\nabla^2 g(\theta_{k-1}^\circ, X_k)\big]\partial_\gamma\theta_{k-1}^\circ \\ &+ \boldsymbol{A}(\theta_{k-1}^\circ, X_k)\partial_\gamma^2\theta_{k-1}^\circ - \partial_\gamma\nabla g(\theta_{k-1}^\circ, X_k).\end{aligned}$$

Therefore, we can denote the second derivative recursive procedure by $\{\ddot{\theta}_k(\gamma)\}_{k\in\mathbb{N}}$ with

$$\ddot{\theta}_k(\gamma) = \boldsymbol{A}(\theta_{k-1}^\circ, X_k)\ddot{\theta}_{k-1}(\gamma) + [\partial_\gamma\boldsymbol{A}(\theta_{k-1}^\circ, X_k)]\dot{\theta}_{k-1}(\gamma) - \partial_\gamma\nabla g(\theta_{k-1}^\circ, X_k). \tag{40}$$

THEOREM 5.2 (Existence of stationary second derivative process). *Assume that $g(\theta, X)$ in (1) is a second-differentiable function with respect to θ. Also, suppose that we can choose some step size $\gamma \in (0, \gamma(p))$ with $\gamma(p)$ defined in Theorem 2.2 such that, for any unit vector $\delta \in \mathbb{R}^d$ and some constant $p \geq 2$,*

$$A_p^* := \sup_{\theta\in\mathbb{R}^d}\sup_{\{\delta\in\mathbb{R}^d, |\delta|=1\}} \mathbb{E}\big[\big|\boldsymbol{A}(\theta, X)\delta\big|^p\big] < 1, \tag{41}$$

where the random matrix $\boldsymbol{A}(\theta, X)$ is defined in (38). Then, the following two results hold.

(i) *Under the conditions in Theorem 2.2, for the first derivative process $\{\dot{\theta}_k(\gamma)\}_{k\in\mathbb{N}}$ defined in (37), the $p/2$-th moment of $\dot{\theta}_k(\gamma)$ exists for $p \geq 2$, i.e., $\|\dot{\theta}_k(\gamma)\|_{p/2} < \infty$.*

(ii) *Furthermore, if in addition, $\nabla^2 g(\theta(\gamma), X)$ is differentiable with respect to γ, i.e., $\partial_\gamma\nabla^2 g(\theta(\gamma), X)$ exists, then, there exists a stationary solution for the second derivative process $\{\ddot{\theta}_k(\gamma)\}_{k\in\mathbb{N}}$ in (40) and fulfills $\|\ddot{\theta}_k(\gamma)\|_{p/2} < \infty$.*

6. Numerical Experiments. This section is devoted to the experiments on simulated data to demonstrate the validity of our proposed online inference methods. Specifically, we consider two classes of examples: linear regression and logistic regression.

6.1. *Linear regression and logistic regression.* We generate two sequences of i.i.d. observation pairs $\{(\boldsymbol{z}_{1,t}, y_{1,t})\}_{t\geq 1}$ and $\{(\boldsymbol{z}_{2,t}, y_{2,t})\}_{t\geq 1}$ for the two regression models, respectively. Let $\boldsymbol{x}_{1,t} = (\boldsymbol{z}_{1,t}, y_{1,t})$ and $\boldsymbol{x}_{2,t} = (\boldsymbol{z}_{2,t}, y_{2,t})$. We denote the true unknown parameter in the models by θ_1^* and θ_2^*.

First, we generate $\boldsymbol{z}_{1,t}$, $\boldsymbol{z}_{2,t} \in \mathbb{R}^d$ from the standard normal distribution, i.e.,

$$\boldsymbol{z}_{1,t} \sim N(0, \boldsymbol{I}_d), \quad \boldsymbol{z}_{2,t} \sim N(0, \boldsymbol{I}_d). \tag{42}$$

Then, we simulate $y_{1,t}$, $y_{2,t} \in \mathbb{R}$ following

$$y_{1,t} = \boldsymbol{z}_{1,t}^\top\theta_1^* + \epsilon_t, \quad \text{where } \epsilon_t \sim N(0, 1), \tag{43}$$

$$y_{2,t} = \text{Bernoulli}\Big\{\frac{1}{1+\exp\left(-\boldsymbol{z}_{2,t}^\top\theta_2^*\right)}\Big\}. \tag{44}$$

Denoted by $g_1(\cdot)$ and $g_2(\cdot)$ respectively are the loss functions of the two models, which are defined as the negative log-likelihood, that is

$$g_1\big(\theta, \boldsymbol{z}_{1,t}, y_{1,t}\big) = (\boldsymbol{z}_{1,t}^\top \theta - y_{1,t})^2/2, \tag{45}$$

$$g_2\big(\theta, \boldsymbol{z}_{2,t}, y_{2,t}\big) = (1 - y_{2,t})\boldsymbol{z}_{2,t}^\top \theta + \log\big(1 + \exp\{-\boldsymbol{z}_{2,t}^\top \theta\}\big). \tag{46}$$

The true parameter $\theta^* \in \mathbb{R}^d$ is linearly spaced between 0 and 1. Since the likelihood loss of the logistic regression in (46) is strictly convex but not strongly convex [1], we add a small regularization term $0.005\|\theta\|_2^2$ to make it strongly convex.

6.2. *Empirical Performance of the Online Estimators.* We first evaluate the empirical performance of our proposed online inference method. Note that in most general cases, the true long-run covariance matrix $\Sigma(\gamma)$ does not have a closed form solution. Therefore, in this section, we focus on the least square loss of the linear regression, because the limiting covariance matrix of this quadratic case can be easily achieved. Consider the linear regression model in (43) and we have

$$\nabla m(\theta_1^*) = \mathbb{E}[\boldsymbol{z}_1 \boldsymbol{z}_1^\top] = \boldsymbol{I}_d. \tag{47}$$

and

$$\mathbb{E}\big[\nabla g(\theta_1^*, \boldsymbol{x}_1) \nabla g(\theta_1^*, \boldsymbol{x}_1)^\top\big] = \mathbb{E}[\epsilon^2]\mathbb{E}[\boldsymbol{z}_1 \boldsymbol{z}_1^\top] = \boldsymbol{I}_d, \tag{48}$$

which directly indicates that the limiting covariance matrix

$$\Sigma = \nabla m(\theta_1^*)^{-1} \mathbb{E}\big[\nabla g(\theta_1^*, \boldsymbol{x}_1) \nabla g(\theta_1^*, \boldsymbol{x}_1)^\top\big] \nabla m(\theta_1^*) = \boldsymbol{I}_d. \tag{49}$$

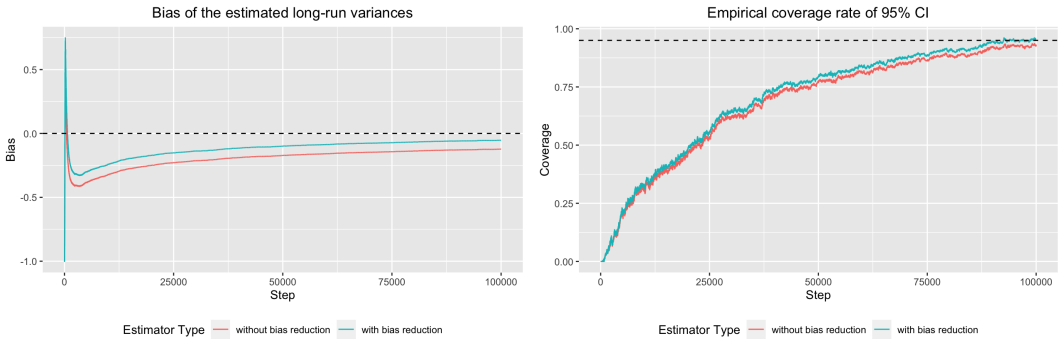
To compare our two online estimators, without and with bias reduction, we consider the one-dimensional case to report the bias of the two estimators relative to the iteration steps. All of our measurements are averaged over 500 independent runs. As shown in Figure 2a, the results suggest that both the non-debiased (*in red*) and the debiased (*in green*) estimators converge fast to the true value (*zero bias at the dashed line*) as the step number increases. Moreover, the advanced version with bias reduction yields a better estimated value which is closer to the true limiting variance. Here we set the constant step size to be 0.025. We also investigate other scenarios with different step sizes and we observe the similar results that the debiased estimator robustly outperform the non-debiased version. We defer the details to Figure 3 in Supplement B.

6.3. *Coverage Probabilities of Confidence Intervals.* In this section, we construct the online confidence intervals for the one-dimensional projection $\mathbf{1}_d^\top \theta^*$ of θ^* using our two online estimators. Both linear regression and logistic regression are investigated. In particular, at step k, using our online estimator $\widehat{\Sigma}$ of Σ, we can construct the $(1-\alpha)\%$ confidence intervals for θ_k for some given $\alpha \in (0, 1)$ as follows:

$$\Big[\widehat{\theta}_k \mathbf{1}_d^\top - z_{1-\alpha/2}\sqrt{(\mathbf{1}^\top \widehat{\Sigma}_k \mathbf{1}_d)/k},\ \widehat{\theta}_k \mathbf{1}_d^\top + z_{1-\alpha/2}\sqrt{(\mathbf{1}^\top \widehat{\Sigma}_k \mathbf{1}_d)/k}\Big], \tag{50}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-percentile of the standard normal distribution. The bias-reduced confidence interval can be similarly constructed with $\widehat{\Sigma}$ in (50) replaced by $\widetilde{\Sigma}$.

To evaluate the coverage probability, we take the average of 500 independent repetitions. In Figure 2b, we present the convergence trace of the empirical coverage rate of 95% CI with the constant step size $\gamma = 0.025$ in linear regression. For the cases with other constant step sizes or the logistic regression, we report the coverage rates in Table 1 with the standard errors shown in the brackets. Our results show that for both the non-debiased (*in red*) and the debiased (*in green*), the empirical coverage rates converge to the nominal probability 95%, while the non-debiased version exhibits an enhanced coverage rate.



(a) Bias of estimators.

(b) Empirical coverage probability.

Fig 2: Recursively estimated long-run variance averaged over 500 independent runs (Linear regression $\gamma = 0.025$). (Left) Bias of the two estimators. (Right) Empirical coverage probability of 95% confidence interval. The black dashed line indicates the nominal probability = 0.95.

TABLE 1

Empirical coverage probability of 95% confidence interval for linear regression and logistic regression over 500 independent runs. "w/o" represents the estimator without bias reduction, and "w/" represents the estimator with bias reduction. The standard deviation is reported in the brackets.

Linear regression					
step size	correction	$n = 5000$	$n = 10^4$	$n = 5 * 10^4$	$n = 10^5$
0.0125	w/o	0.194 (6.831e-07)	0.332 (9.507e-07)	0.796 (1.111e-06)	0.904 (6.976e-07)
	w/	0.198 (5.839e-07)	0.364 (9.720e-07)	0.854 (1.074e-06)	0.948 (4.345e-07)
0.025	w/o	0.278 (2.056e-07)	0.402 (7.553e-08)	0.844 (7.553e-08)	0.912 (8.940e-07)
	w/	0.286 (1.577e-07)	0.408 (1.053e-06)	0.854 (8.623e-07)	0.954 (1.976e-07)
0.05	w/o	0.266 (1.543e-07)	0.380 (1.652e-07)	0.810 (8.912e-08)	0.910 (8.623e-08)
	w/	0.280 (2.251e-07)	0.396 (1.811e-07)	0.832 (1.068e-07)	0.946 (2.428e-07)
Logistic regression					
step size	correction	$n = 5 * 10^4$	$n = 10^5$	$n = 5 * 10^5$	$n = 10^6$
0.0125	w/o	0.174 (3.899e-07)	0.268 (4.942e-07)	0.718 (5.002e-07)	0.894 (5.690e-07)
	w/	0.176 (2.362e-07)	0.272 (3.614e-07)	0.720 (2.251e-07)	0.906 (6.688e-07)
0.025	w/o	0.180 (2.794e-07)	0.324 (2.523e-07)	0.720 (5.031e-07)	0.902 (8.649e-07)
	w/	0.184 (3.490e-07)	0.328 (5.704e-07)	0.722 (3.157e-07)	0.914 (6.472e-07)
0.05	w/o	0.252 (1.369e-07)	0.380 (2.350e-07)	0.774 (2.091e-08)	0.908 (2.684e-08)
	w/	0.254 (2.001e-07)	0.386 (2.293e-08)	0.780 (1.329e-07)	0.922 (2.023e-07)

7. Conclusion and Discussion. In this paper, we bring nonlinear time-series tools to the machine learning community by providing new interpretations for the SGD iterates. We introduce the concept of geometric-moment contraction from iterated random functions, providing a new way to view the evolution of SGD process. Building on this, we establish the convergence of ASGD in challenging heavy-tailed scenarios. For noise with finite variance, we provide a quenched CLT and invariance principle for ASGD, allowing for effective statistical inference regardless of the starting point. By the limiting distributions in quenched CLT, we propose an efficient online method for estimating the long-run covariance matrix of ASGD iterates. This approach enables the construction of empirical confidence intervals, enhancing the quantification of uncertainty in ASGD predictions.

Another contribution of our study is the identification of a precise range for the constant step-size parameter γ that guarantees the contraction of SGD. Additionally, we refine the

bias expansion of ASGD in terms of the step size by providing theory for an improved Richardson-Romberg extrapolation. Furthermore, we show the existence of stationary solutions for derivative SGD processes, which can lead to a deeper understanding of the SGD's behavior under varying step sizes.

We would like to emphasize the potential of applying time-series technical tools beyond the SGD procedure to a broad spectrum of machine learning challenges where statistical guarantees are in demand. The proof strategies we have developed for SGD theory have promising applications across other machine learning algorithms. In future research, we can extend geometric-moment contraction to SGD processes with iteration-dependent step sizes, specifically with $\gamma \asymp n^{-\beta}$, for some $1/2 < \beta < 1$. Similarly, for mini-batch gradient descent, which balances the features of batch gradient descent and SGD, we propose investigating its asymptotic behavior using time series coupling concepts to manage intricate dependency structures.

Moreover, our online inference method has potential for generalization to other machine learning algorithms, particularly those involving recursive procedures or exhibiting temporal dynamics and non-stationarity, such as Nesterov's accelerated gradient [42], an advanced form of momentum gradient descent, and some adaptive algorithms such as Adam [28] that demonstrates better convergence than the SGD in certain deep learning applications. This paper initiates the promising future for integrating time-series analysis with machine learning, enhancing the theoretical understanding and practical application of these modern algorithms in complex data streams.

SUPPLEMENTARY MATERIAL

Supplement to "The Stochastic Gradient Descent from a Nonlinear Time Series Perspective".
This supplementary material contains all the technical proofs for the theorems presented in the manuscript. Additional information of the simulation studies is also included.

REFERENCES

[1] BACH, F. (2014). Adaptivity of Averaged Stochastic Gradient Descent to Local Strong Convexity for Logistic Regression. *Journal of Machine Learning Research* **15** 595–627.

[2] BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Proceedings of the 26th International Conference on Neural Information Processing Systems* 773–781.

[3] BERKES, I., LIU, W. and WU, W. B. (2014). Komlós–Major–Tusnády approximation under dependence. *The Annals of Probability* **42** 794–817.

[4] BLUM, J. R. (1954). Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics* **25** 382–386.

[5] BOTTOU, L. (1998). Online Algorithms and Stochastic Approximations. In *Online Learning and Neural Networks* (D. Saad, ed.) Cambridge University Press, Cambridge, UK.

[6] CHAO, S.-K. and CHENG, G. (2019). A generalization of regularized dual averaging and its dynamics. *arXiv preprint*. arXiv:1909.10072.

[7] CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics* **48** 251–273.

[8] CLARA, G., LANGER, S. and SCHMIDT-HIEBER, J. (2023). Dropout Regularization Versus ℓ_2-Penalization in the Linear Model. *arXiv prerpint*. arXiv:2306.10529.

[9] DAHLHAUS, R. and RAO, S. S. (2006). Statistical inference for time-varying ARCH processes. *The Annals of Statistics* **34** 1075–1114.

[10] DAHLHAUS, R., RICHTER, S. and WU, W. B. (2019). Towards a general theory for nonlinear locally stationary processes. *Bernoulli* **25** 1013–1044.

[11] DEFOSSEZ, A. and BACH, F. (2015). Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* 205–213.

[12] DIACONIS, P. and FREEDMAN, D. (1999). Iterated Random Functions. *SIAM Review* **41** 45–76.

[13] DIEULEVEUT, A. and BACH, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics* **44** 1363–1399.

[14] DIEULEVEUT, A., DURMUS, A. and BACH, F. (2020). Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics* **48** 1348–1382.

[15] DIEULEVEUT, A., FLAMMARION, N. and BACH, F. (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research* **18** 3520–3570.

[16] DVORETZKY, A. (1956). On Stochastic Approximation. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **3** 39–56.

[17] FABIAN, V. (1968). On Asymptotic Normality in Stochastic Approximation. *The Annals of Mathematical Statistics* **39** 1327–1332.

[18] FANG, Y. (2019). Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics* **46** 987–1002.

[19] FANG, Y., XU, J. and YANG, L. (2018). Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator. *Journal of Machine Learning Research* **19** 1–21.

[20] GANDIKOTA, V., KANE, D., MAITY, R. K. and MAZUMDAR, A. (2022). vqSGD: Vector Quantized Stochastic Gradient Descent. *IEEE Transactions on Information Theory* **68** 4573–4587.

[21] GYÖRFI, L. and WALK, H. (1996). On the Averaged Stochastic Approximation for Linear Regression. *SIAM Journal on Control and Optimization* **34** 31–61.

[22] HÄRDLE, W. (1986). A note on jackknifing kernel regression function estimators. *IEEE Transactions on Information Theory* **32** 298–300.

[23] HUO, D., CHEN, Y. and XIE, Q. (2023). Bias and Extrapolation in Markovian Linear Stochastic Approximation with Constant Stepsizes. *Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* **51** 81—82.

[24] IMAIZUMI, M. and SCHMIDT-HIEBER, J. (2023). On Generalization Bounds for Deep Networks Based on Loss Surface Implicit Regularization. *IEEE Transactions on Information Theory* **69** 1203–1223.

[25] JAIN, P., KAKADE, S. M., KIDAMBI, R., NETRAPALLI, P. and SIDFORD, A. (2018). Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification. *Journal of Machine Learning Research* **18** 1–42.

[26] KARMAKAR, S. and WU, W. B. (2020). Optimal Gaussian Approximation for Multiple Time Series. *Statistica Sinica* **30** 1399–1417.

[27] KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics* **23** 462–466.

[28] KINGMA, D. P. and BA, J. (2014). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

[29] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **32** 111–131.

[30] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **34** 33–58.

[31] KONAKOV, V. and MAMMEN, E. (2023). Local Limit Theorems and Strong Approximations for Robbins-Monro Procedures. *arXiv prerpint*. arXiv:2304.10673.

[32] LAI, T. L. (2003). Stochastic approximation: invited paper. *The Annals of Statistics* **31** 391–406.

[33] LAKSHMINARAYANAN, C. and SZEPESVARI, C. (2018). Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go? In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* 1347–1355.

[34] LIANG, T. and SU, W. (2019). Statistical Inference for the Population Landscape via Moment-Adjusted Stochastic Gradients. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **81** 431–456.

[35] LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control* **22** 551–575.

[36] 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.

[37] MA, S., BASSILY, R. and BELKIN, M. (2018). The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning. In *Proceedings of the 35th International Conference on Machine Learning* 3325–3334.

[38] MERAD, I. and GAÏFFAS, S. (2023). Convergence and concentration properties of constant step-size SGD through Markov chains. *arXiv preprint*. arXiv:2306.11497.

[39] MOULINES, E. and BACH, F. (2011). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems* 856–864.

[40] NEEDELL, D., WARD, R. and SREBRO, N. (2014). Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* 1017–1025.

[41] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2009). Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization* **19** 1574–1609.

[42] NESTEROV, Y. (1983). A method of solving the convex programming problem with convergence rate O(1/k^2). *Soviet Mathematics Doklady* **27** 372–376.

[43] NESTEROV, Y. and VIAL, J. (2008). Confidence level solutions for stochastic programming. *Automatica J.IFAC* **44** 1559–1568.

[44] PFLUG, G. C. (1986). Stochastic Minimization with Constant Step-Size: Asymptotic Laws. *SIAM Journal on Control and Optimization* **24** 655–666.

[45] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* **30** 838–855.

[46] RAKHLIN, A., SHAMIR, O. and SRIDHARAN, K. (2011). Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *Proceedings of the 29th International Conference on Machine Learning*.

[47] ROBBINS, H. and MONRO, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics* **22** 400–407.

[48] ROBBINS, H. and SIEGMUND, D. (1971). A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications. In *Optimizing Methods in Statistics* 233–257. Academic Press.

[49] RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1986). Learning representations by back-propagating errors. *Nature* **323** 533–536.

[50] RUPPERT, D. (1988). Efficient Estimations from a Slowly Convergent Robbins-Monro Process. In *Technical report* Cornell University Operations Research and Industrial Engineering.

[51] SACKS, J. (1958). Asymptotic Distribution of Stochastic Approximation Procedures. *The Annals of Mathematical Statistics* **29** 373–405.

[52] SCHMIDT, M. and ROUX, N. L. (2013). Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. *arXiv prerpint*. arXiv:1308.6370.

[53] SHALEV-SHWARTZ, S., SHAMIR, O., SREBRO, N. and SRIDHARAN, K. (2009). Stochastic Convex Optimization. *Proceedings of the International Conference on Learning Theory*.

[54] SPALL, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control* **45** 1839–1853.

[55] STOER, J. and BULIRSCH, R. (2002). *Introduction to Numerical Analysis*. Springer, New York.

[56] SU, W. and ZHU, Y. (2023). HiGrad: Uncertainty Quantification for Online Learning and Stochastic Approximation. *Journal of Machine Learning Research* **24** 1–53.

[57] VASWANI, S., BACH, F. and SCHMIDT, M. (2019). Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*.

[58] WANG, X. and GAO, N. (2010). Stochastic Resource Allocation Over Fading Multiple Access and Broadcast Channels. *IEEE Transactions on Information Theory* **56** 2382–2391.

[59] WU, W. B. (2009). Recursive estimation of time-average variance constants. *The Annals of Applied Probability* **19** 1529–1552.

[60] WU, W. B. and SHAO, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability* **41** 425–436.

[61] XIAO, H. and WU, W. B. (2011). A single-pass algorithm for spectrum estimation with fast convergence. *IEEE Transactions on Information Theory* **57** 4720–4731.

[62] 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 *Proceedings of the 35th International Conference on Neural Information Processing Systems,* 4234–4248.

[63] ZHONG, Y., KUFFNER, T. and LAHIRI, S. (2023). Online Bootstrap Inference with Nonconvex Stochastic Gradient Descent Estimator. *arXiv preprint*. arXiv:2306.02205.

[64] ZHU, W., CHEN, X. and WU, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association* **118** 393–404.

reference not indexed here, supplement to submit is in another file: aos-supp.tex

SUPPLEMENT TO "THE STOCHASTIC GRADIENT DESCENT FROM A NONLINEAR TIME SERIES PERSPECTIVE"

APPENDIX A: POSTPONED PROOFS

A.1. Proofs of Theorems 2.2 & 3.1. The proof of Theorem 2.2 is essentially motivated by Lemma A.1 below, which itself can be of independent interests in the bounds of the power functions $|\cdot|^p$, for $p > 1$.

LEMMA A.1. *For any two vectors $x, y \in \mathbb{R}^d$, the following two inequalities hold:*
(i) for $p \geq 2$, $||x+y|^p - |x|^p - p|x|^{p-2}x^\top y| \leq (|x|+|y|)^p - |x|^p - p|x|^{p-1}|y|$;
(ii) for $1 < p < 2$, $||x+y|^p - |x|^p - p|x|^{p-2}x^\top y| \leq 2^{2-p}|y|^p$.

The proof of this lemma and its application to Theorem 2.2 are provided in the Proof of Theorem 2.2 below.

PROOF OF THEOREM 2.2. We first show the contraction in (17) for two cases with $p \geq 2$ and $1 < p < 2$ respectively.

(i) Case $p \geq 2$:

Let $x = \theta - \theta'$, $y = -\gamma[\nabla g(\theta, X) - \nabla g(\theta', X)]$. Since the contraction in (17) holds when $|\theta - \theta'| = 0$, w.l.o.g., we assume that $|x| > 0$. Let $\theta = \theta' + \delta v$, where v is a unit vector in $\mathbb{R}^d$ and $\delta \neq 0$. This gives us $x = \delta v$ and $|x| = |\theta - \theta'| = \delta$.

We first show the following analytical inequality:

$$\big||x+y|^p - |x|^p - p|x|^{p-2}x^\top y\big| \leq \big(|x|+|y|\big)^p - |x|^p - p|x|^{p-1}|y|. \tag{51}$$

Let $y = a\delta v + be$, where e is a unit vector in $\mathbb{R}^d$ which is orthogonal to v, and a, b are some constants in $\mathbb{R}$. We define $R^2 = (a\delta)^2 + b^2$ and thus $|y| = R$. Then, since $x^\top y = a\delta^2$, we have

$$\begin{aligned}\big||x+y|^p - |x|^p - p|x|^{p-2}x^\top y\big| &= \big|[(1+a)^2\delta^2 + b^2]^{p/2} - \delta^p - pa\delta^p\big| \\ &= \big|(\delta^2 + 2a\delta^2 + R^2)^{p/2} - \delta^p - pa\delta^p\big|. \end{aligned} \tag{52}$$

Let $a\delta = Ru$ with some constant $|u| \leq 1$. Then,

$$(\delta^2 + 2a\delta^2 + R^2)^{p/2} - \delta^p - pa\delta^p = (\delta + 2\delta Ru + R^2)^{p/2} - \delta^p - p\delta^{p-1}Ru =: f(u). \tag{53}$$

By taking the derivative of the function $f(u)$, we obtain

$$\begin{aligned}\frac{d}{du}f(u) &= \frac{p}{2}(\delta^2 + 2\delta Ru + R^2)^{p/2-1} \cdot 2\delta R - p\delta^{p-1}R \\ &= p\delta R\big[(\delta^2 + 2\delta Ru + R^2)^{p/2-1} - \delta^{p-2}\big], \end{aligned} \tag{54}$$

which indicates that $f(u)$ is decreasing in $[-1, -R/(2\delta)]$ and increasing in $(-R/(2\delta), 1]$. Since $f(1) > f(-1)$, for $p \geq 2$,

$$\max_{|u|\leq 1}|f(u)| = f(1) = (\delta + R)^p - \delta^p - p\delta^{p-1}R. \tag{55}$$

MSC2020 subject classifications: Primary 62F12, 62E20; secondary 62M10.
Keywords and phrases: stochastic gradient descent, iterated random function, derivative process, quenched central limit theorem, online inference.

This, along with the fact that $|x| = \delta$ and $|y| = R$ leads to the inequality in (51).

Further, we take the expectation on the both sides of the inequality (51) and achieve

$$\mathbb{E}|x+y|^p - |x|^p - p|x|^{p-2}\mathbb{E}(x^\top y) \le \mathbb{E}(|x|+|y|)^p - |x|^p - p|x|^{p-1}\mathbb{E}|y|. \tag{56}$$

Recall that $x = \theta - \theta'$, $y = -\gamma[\nabla g(\theta, X) - \nabla g(\theta', X)]$. We aim to bound $\mathbb{E}|x+y|^p$, which can be decomposed into two parts as follows:

$$\begin{aligned}\mathbb{E}|x+y|^p &= \big[\mathbb{E}|x+y|^p - |x|^p - p|x|^{p-2}\mathbb{E}(x^\top y)\big] + \big[|x|^p + p|x|^{p-2}\mathbb{E}(x^\top y)\big] \\ &=: \mathbb{I}_1 + \mathbb{I}_2.\end{aligned}$$

For the part $\mathbb{I}_2$, it follows from Assumption 1 that

$$\mathbb{E}(x^\top y) = -\gamma\langle \theta - \theta', m(\theta) - m(\theta')\rangle \le -\mu\gamma|\theta - \theta'|^2. \tag{57}$$

Next, for the part $\mathbb{I}_1$, we note that, for any $p > 1$ (p is not necessarily to be an integer), we can write the right-hand side of (56) into

$$\mathbb{E}(|x|+|y|)^p - |x|^p - p|x|^{p-1}\mathbb{E}|y| = \int_{s=0}^1 \int_{t=0}^s p(p-1)\mathbb{E}\big[(|x|+|y|t)^{p-2}|y|^2\big]dtds.$$

By Hölder's inequality and the triangle inequality, we have

$$\begin{aligned}\mathbb{E}\big[(|x|+|y|t)^{p-2}|y|^2\big] &\le \big(\mathbb{E}(|x|+|y|t)^p\big)^{(p-2)/p}\big(\mathbb{E}|y|^p\big)^{2/p} \\ &\le (|x| + \|y\|_p t)^{p-2}\|y\|_p^2.\end{aligned}$$

Furthermore, for each $p > 1$, it follows from Assumption 2 that

$$\|y\|_p = \big\| -\gamma[\nabla g(\theta, X) - \nabla g(\theta', X)]\big\|_p \le \gamma L_p|\theta - \theta'| = \gamma L_p|x|.$$

Hence, we can achieve

$$\mathbb{E}(|x|+|y|)^p - |x|^p - p|x|^{p-1}\mathbb{E}|y| \le (1+\gamma L_p)^p|x|^p - |x|^p - p\gamma L_p|x|^p. \tag{58}$$

Finally, by combining the results for $\mathbb{I}_1$ and $\mathbb{I}_2$, we can obtain

$$\begin{aligned}&\mathbb{E}|\theta - \theta' - \gamma(\nabla g(\theta, X) - \nabla g(\theta', X))|^p \\ &\le \big[1 - p\mu\gamma + (1+\gamma L_p)^p - 1 - p\gamma L_p\big]|\theta - \theta'|^p.\end{aligned} \tag{59}$$

As a direct consequence, the upper bound for γ is the solution to the following equation:

$$(1+\gamma L_p)^p - 1 - p\gamma(\mu + L_p) = 0. \tag{60}$$

(ii) Case $1 < p < 2$:

First, for real d-dimensional vectors x, y the analytic inequality

$$|x+y|^p \le |x|^p + p|x|^{p-2}x^\top y + 2^{2-p}|y|^p \tag{61}$$

is shown. For $x = 0$ or $y = 0$ this inequality holds obviously. For $|x|, |y| > 0$, define

$$\omega = \frac{|x|}{|y|} > 0 \ \text{ and } \ \varrho = \frac{x^\top y}{|x|\cdot|y|}.$$

Then $|\varrho| \le 1$ and

$$\frac{|x+y|^p - |x|^p - p|x|^{p-2}x^\top y}{|y|^p} = (\omega^2 + 1 + 2\omega\varrho)^{p/2} - \omega^p - p\omega^{p-1}\varrho =: \psi(\omega, \varrho).$$

Note that

$$\frac{\partial}{\partial \varrho}\psi(\omega,\varrho) = p\omega^{p-1}\left\{\left(1+\frac{2\varrho}{\omega}+\frac{1}{\omega^2}\right)^{p/2-1}-1\right\}.$$

If $\omega \geq 1/2$,

$$\sup_{\varrho\in[-1,1]}\psi(\omega,\varrho) = \psi\left(\omega,-\frac{1}{2\omega}\right) = \frac{p\omega^{p-2}}{2} \leq p2^{1-p} \leq 2^{2-p}.$$

If $0<\omega<1/2$, we define a constant factor

$$\widetilde{C}_p := \max_{\omega\in(0,1)}\left[(1-\omega)^p - \omega^p + p\omega^{p-1}\right]. \tag{62}$$

By Proposition 1.8 in [**?**], it follows that

$$\sup_{\varrho\in[-1,1]}\psi(\omega,\varrho) = \psi(\omega,-1) = (1-\omega)^p - \omega^p + p\omega^{p-1} \leq 2^{2-p}.$$

In total, we have that for all $\rho\in[-1,1]$, $\omega>0$ that

$$\psi(\omega,\rho) \leq 2^{2-p},$$

leading to

$$\frac{|x+y|^p - |x|^p - p|x|^{p-2}x^\top y}{|y|^p} \leq 2^{2-p}$$

and thus to (61). Next, recall that $x=\theta-\theta'$, $y=-\gamma[g(\theta,x)-g(\theta',x)]$. By taking the expectation on the both sides of inequality (61), we have

$$\begin{aligned}
&\mathbb{E}|\theta-\theta'-\gamma(\nabla g(\theta,X)-\nabla g(\theta',X))|^p \\
&\leq |\theta-\theta'|^p - p|\theta-\theta'|^{p-2}(\theta-\theta')^\top\mathbb{E}\left[\gamma\big(\nabla g(\theta,X)-\nabla g(\theta',X)\big)\right] \\
&\quad + 2^{2-p}\mathbb{E}\left|\gamma\big(\nabla g(\theta,X)-\nabla g(\theta',X)\big)\right|^p \\
&\leq \left(1-p\mu\gamma+2^{2-p}L_p^p\gamma^p\right)|\theta-\theta'|^p. \qquad (63)
\end{aligned}$$

This completes the proof of the geometric contraction condition.

Next, we shall show the existence and uniqueness of the stationary distribution π_γ based on the GMC in (11) by using Theorem 2 in [60]. To this end, recall the recursive function $F_X(\theta)$ defined in (12) and the global optimum θ^* defined in (1). Note that, for $p>1$, it follows from Assumption 2(i) that

$$\mathbb{E}|\theta^* - F_X(\theta^*)|^p = \mathbb{E}|\gamma\nabla g(\theta^*,X)|^p < \infty. \tag{64}$$

In addition, it follows from Theorem 2.2 that for $p>1$, there exist $\theta_0\in\mathbb{R}^d$ and $\rho_{\gamma,p}\in(0,1)$ such that

$$\mathbb{E}|F_{X_n}\circ\cdots\circ F_{X_1}(\theta) - F_{X_n}\circ\cdots\circ F_{X_1}(\theta_0)|^p \leq (\rho_{\gamma,p}^p)^n|\theta-\theta_0|^p \tag{65}$$

holds for all $\theta\in\mathbb{R}^d$, $n\in\mathbb{N}$. In fact, expressions (64) and (65) are also valid for all $p\leq 1$ by Hölder's inequality. To see this, we let $p'\in(0,p)$ and $\rho_{\gamma,p'} = \rho_{\gamma,p}^{p'/p}\in(0,1)$. Then, we have

$$\begin{aligned}
&\mathbb{E}|F_{X_n}\circ\cdots\circ F_{X_1}(\theta) - F_{X_n}\circ\cdots\circ F_{X_1}(\theta_0)|^{p'} \\
&\leq \left(\mathbb{E}|F_{X_n}\circ\cdots\circ F_{X_1}(\theta) - F_{X_n}\circ\cdots\circ F_{X_1}(\theta_0)|^{p}\right)^{p'/p} \\
&\leq \left[(\rho_{\gamma,p}^p)^n|\theta-\theta_0|^p\right]^{p'/p} = (\rho_{\gamma,p'}^p)^n|\theta-\theta_0|^{p'}. \qquad (66)
\end{aligned}$$

Therefore, both expressions (64) and (65) hold for all $p > 0$, which indicates that Conditions 1 and 2 in [60] are satisfied for our recursive function $F_X(\theta)$.

Denote the forward iteration process by $A_n(\theta) = F_{X_n} \circ \cdots \circ F_{X_1}(\theta)$. Now, we introduce the backward iteration process $B_n(\theta) = F_{X_0} \circ F_{X_{-1}} \cdots \circ F_{X_{1-n}}(\theta)$ ([**?**]). Note that for all $\theta \in \mathbb{R}^d$, we have the distributional equality $A_n(\theta) \stackrel{\mathcal{D}}{=} B_n(\theta)$. By Theorem 2 in [60], there exists a random variable B_∞ which is $\sigma(X_0, X_{-1}, \ldots)$ measurable such that for all $\theta \in \mathbb{R}^d$, $B_n(\theta) \to B_\infty$ almost surely as $n \to \infty$. Thus, $A_n(\theta)$ converges to B_∞ in distribution. Then, as shown by Remark 1 in [60], Theorem 1 in [12] is also implied. Hence, there exists a unique stationary distribution π with $B_\infty \sim \pi$, which does not depend θ.

Note that $\theta_n(\gamma) = A_n(\theta_0)$. Let $B_\infty = \theta_0^\circ(\gamma)$. Then, there exists a unique stationary distribution π_γ such that $\theta_0^\circ(\gamma) \sim \pi_\gamma$ and therefore, $\theta_k(\gamma) \Rightarrow \pi_\gamma$. As a direct consequence of (11), π_γ has finite p-th moment. Moreover, by applying the induction to (11), we can obtain the geometric moment contraction in (19), which completes the proof. □

PROOF OF THEOREM 3.1. First, we shall provide an upper bound for the p-th moment $\|\bar{\theta}_n(\gamma) - \theta_\infty^*(\gamma)\|_p$ for $p > 1$. If $p \geq 2$, by similar arguments as in expressions (69) and (70), we can obtain $\|\bar{\theta}_n(\gamma) - \theta_\infty^*(\gamma)\|_p = O(n^{-1/2})$. When $1 < p < 2$, by applying Theorem 1 in [**?**], it directly follows that $\|\bar{\theta}_n(\gamma) - \theta_\infty^*(\gamma)\|_p = O(n^{1/p-1})$.

Regarding the almost sure bounds, by the geometric moment contraction in Theorem 2.2, the results for the cases $1 < p < 2$ and $p \geq 2$ can be achieved by applying Corollary 2 (iii) and Theorem 2 (i) in [**?**].

□

A.2. Proof of Theorem 3.2.

PROOF OF THEOREM 3.2. Let $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ be the sequence of solutions from the SGD recursive function (2) with the stationary distribution π_γ. As a result, $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ is a uniformly ergodic Markov chain. Recall the recursive function $F_X(\theta)$ in (12) and the partial sum $S_n^\circ = \sum_{k=1}^n \varepsilon_k^\circ \in \mathbb{R}^{dl}$ in (23). Note that for each $0 < \gamma_s < \gamma(p)$, $s = 1, \ldots, \ell$, we have

$$\begin{aligned}\sum_{i=1}^n \left[\theta_i^\circ(\gamma_s) - \theta_\infty^*(\gamma_s)\right] &= F_{X_1}(\theta_0^\circ(\gamma_s)) - \mathbb{E}[F_{X_1}(\theta_0^\circ(\gamma_s))] + \cdots \\ &\quad + F_{X_n} \circ \cdots \circ F_{X_1}(\theta_0^\circ(\gamma_s)) - \mathbb{E}[F_{X_n} \circ \cdots \circ F_{X_1}(\theta_0^\circ(\gamma_s))]. \end{aligned} \tag{67}$$

By Theorem 2.1 in [**?**], since GMC ensures the short-range dependence, we can obtain that any fixed linear combination of the coordinates of S_n° converges to corresponding linear combination of normal vectors. Then, the multivariate CLT holds for the vector process $\{\varepsilon_k^\circ\}_{k\in\mathbb{N}}$ via the Cramér-Wold device, that is,

$$n^{-1/2} S_n^\circ \Rightarrow N(0, \Sigma^\circ). \tag{68}$$

Similarly, we recall the partial sum S_n with any arbitrary initial point $\theta_0 \in \mathbb{R}^d$ defined in Theorem 3.2. Since θ_0 may not follow the stationary distribution π_{γ_s}, the sequence $\{\theta_k(\gamma_s)\}_{k\in\mathbb{N}}$ can be non-stationary. Therefore, we need to use the GMC property in Theorem 2.2 to show that $\{\theta_k(\gamma_s)\}_{k\in\mathbb{N}}$ is asymptotically stationary and the quenched CLT holds. Let $p \geq 2$. It follows from Theorem 2.2 and the triangle inequality that, for each $0 < \gamma_s < \gamma(p)$, $s = 1, \ldots, \ell$,

$$\left\| \sum_{i=1}^n \left[\theta_i^\circ(\gamma_s) - \theta_i(\gamma_s)\right] \right\|_p$$

$$
\begin{aligned}
&= \big\| F_{X_1}(\theta_0^\circ(\gamma_s)) + F_{X_2} \circ F_{X_1}(\theta_0^\circ(\gamma_s)) + \cdots + F_{X_n} \circ \cdots \circ F_{X_1}(\theta_0^\circ(\gamma_s)) \\
&\quad - \big[F_{X_1}(\theta_0(\gamma_s)) + F_{X_2} \circ F_{X_1}(\theta_0(\gamma_s)) + \cdots + F_{X_n} \circ \cdots \circ F_{X_1}(\theta_0(\gamma_s)) \big] \big\|_p \\
&\leq \sum_{i=1}^{n} [\rho_{\gamma,p}]^i |\theta_0^\circ(\gamma_s) - \theta_0(\gamma_s)|, \tag{69}
\end{aligned}
$$

where $\rho_{\gamma,p}$ is defined in Theorem 2.2. We write $\theta_0(\gamma) = \theta_0$. In fact, we have

$$
\sum_{i=1}^{n} [\rho_{\gamma,p}]^i |\theta_0^\circ - \theta_0| = o_{\mathbb{P}}(n^{1/2}).
$$

To see this, note that since $0 < \rho_{\gamma,p} < 1$, we can derive the limit

$$
\lim_{n\to\infty} \sum_{i=1}^{n} [\rho_{\gamma,p}]^i = \lim_{n\to\infty} \frac{\rho_{\gamma,p}(1 - [\rho_{\gamma,p}]^n)}{1 - \rho_{\gamma,p}} = \frac{\rho_{\gamma,p}}{1 - \rho_{\gamma,p}}, \tag{70}
$$

which together with expression (69) gives

$$
\lim_{n\to\infty} \frac{\|S_n^\circ - S_n\|_p}{n^{1/2}} \leq \frac{\rho_{\gamma,p}}{1 - \rho_{\gamma,p}} \lim_{n\to\infty} \frac{|\theta_0^\circ - \theta_0|}{\sqrt{n}} \to 0.
$$

This, along with expression (68) and Markov's inequality yields the quenched CLT

$$
n^{-1/2} S_n \Rightarrow N(0, \Sigma^\circ), \tag{71}
$$

for any arbitrary starting point $\theta_0 \in \mathbb{R}^d$, which completes the proof. □

A.3. Proof of Theorem 3.3. Before showing the proof for Theorem 3.3, we first introduce the functional dependence measure following [**?**], which is the key towards the proof. In particular, we provide a detailed form of the functional dependence measure tailored for the stationary SGD process.

Let X_i, X_j', $i, j \in \mathbb{Z}$ be i.i.d. random variables following the distribution Π in (1). Define the filtration $\mathcal{F}_i = (\ldots, X_{i-1}, X_i)$ and the coupled version $\mathcal{F}_{i,\{k\}} = (\ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_i)$. Then, there exists some measurable function H_γ such that the stationary SGD sequence $\{\theta_i^\circ(\gamma)\}_{i\in\mathbb{N}}$ can be written as the following causal process

$$
\theta_i^\circ(\gamma) = H_\gamma(\ldots, X_{i-1}, X_i) = H_\gamma(\mathcal{F}_i). \tag{72}
$$

Let $p > 1$. We define the functional dependence measure of $\theta_i^\circ(\gamma)$ as

$$
\delta_{k,p}(\gamma) = \|\theta_i^\circ(\gamma) - \theta_{i,\{k\}}^\circ(\gamma)\|_p, \quad \text{where } \theta_{i,\{k\}}^\circ(\gamma) = H_\gamma(\mathcal{F}_{i,\{k\}}). \tag{73}
$$

Further, if $\sum_{i=0}^{n} \delta_{i,p} < \infty$, we define the tail of cumulative dependence measure as

$$
\Theta_{m,p} = \sum_{i=m}^{\infty} \delta_{i,p}. \tag{74}
$$

Essentially, since the geometric moment contraction of the SGD iterates indicates the exponential decay of the functional dependence measure, one can quantify the error bound of the Gaussian approximation using the sample size n and the p-th finite moment condition in Assumption 2 with $p > 2$. We can obtain the best case rate of $n^{1/p}$ for any arbitrary initial point $\theta_0 \in \mathbb{R}^d$.

PROOF OF THEOREM 3.3. We shall first consider the random process $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ following the stationary distribution π_γ. Recall the recursive function $F_X(\theta)$ defined in (12) and we can thus rewrite $\theta_k^\circ(\gamma)$ into

$$\theta_k^\circ(\gamma) = F_{X_k} \circ \cdots \circ F_{X_1}(\theta_0^\circ).$$

Recall the contraction constant $\rho_{\gamma,p} \in (0,1)$ given by the geometric-moment contraction (GMC) in Theorem 2.2. For any $k \geq 1$ and some constant $p > 2$, we have

$$\sup_{\theta_0^\circ \neq \theta_0^{\circ\prime}} \frac{\|F_{X_k} \circ \cdots \circ F_{X_1}(\theta_0^\circ) - F_{X_k} \circ \cdots \circ F_{X_1}(\theta_0^{\circ\prime})\|_p}{|\theta_0^\circ - \theta_0^{\circ\prime}|} = [\rho_{\gamma,p}]^k < 1.$$

As a direct consequence, the functional dependence measure $\delta_{k,p}(\gamma)$ converges at an exponential rate as $k \to \infty$, that is, for some constant $0 < \rho_\gamma < 1$,

$$\delta_{k,p}(\gamma) = O(\rho_\gamma^k), \tag{75}$$

where the constant step size $\gamma \in (0,1)$ satisfies the conditions in Theorem 2.2. Therefore, there exists a constant $A > 0$ such that the tail cumulative dependence measure of the stationary SGD process $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ can be bounded by

$$\Theta_{i,p}(\gamma) = \sum_{k=i}^{\infty} \delta_{k,p}(\gamma) = O\big\{i^{-\chi}(\log(i))^{-A}\big\}, \tag{76}$$

where $\chi > 0$ is some constant that can go to infinity. Recall the stationary partial sum sequence $S_i^\circ = \sum_{k=1}^i \varepsilon_k^\circ$ defined in (23). Then, it follows from Theorem 2 in [26] that there exists a (richer) probability space $(\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{\mathbb{P}})$ on which we can define random vectors $\widetilde{\varepsilon}_k^\circ \in \mathbb{R}^{dl}$ with the partial sum process $\widetilde{S}_i^\circ = \sum_{k=1}^n \widetilde{\varepsilon}_k^\circ$, and a Gaussian process $\widetilde{G}_i^\circ = \sum_{k=1}^n \widetilde{Z}_k^\circ$, where $\widetilde{Z}_k^\circ$ is a mean zero independent Gaussian vector in $\mathbb{R}^{dl}$ with an identity covariance matrix, such that $(\widetilde{S}_i^\circ)_{i\geq 0} \overset{\mathcal{D}}{=} (S_i^\circ)_{i\geq 0}$ and

$$\max_{i\leq n} \big|\widetilde{S}_i^\circ - \Sigma^{\circ 1/2}\widetilde{G}_i^\circ\big| = o_{\mathbb{P}}(n^{1/p}), \quad \text{in } (\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{\mathbb{P}}). \tag{77}$$

Next, recall the partial sum sequence $S_i = \sum_{k=1}^i \varepsilon_k$ defined in Theorem 3.2. We shall bound the difference $\|\max_{i\leq n}|S_i - S_i^\circ|\|_p$, for $p > 2$. Note that, for each $0 < \gamma_s < \gamma(p)$, $s = 1, \ldots, \ell$, we write $S_i(\gamma_s) := \sum_{k=1}^i \theta_k(\gamma_s)$ and $S_i^\circ(\gamma_s) := \sum_{k=1}^i \theta_k^\circ(\gamma_s)$ for simplicity. Then, we have

$$\begin{aligned}
\widetilde{S}_i(\gamma_s) &:= S_i(\gamma_s) - S_i^\circ(\gamma_s) \\
&= \big[F_{X_1}(\theta_0(\gamma_s)) - F_{X_1}(\theta_0^\circ(\gamma_s))\big] + \big[F_{X_2} \circ F_{X_1}(\theta_0) - F_{X_2} \circ F_{X_1}(\theta_0^\circ(\gamma_s))\big] + \cdots \\
&\quad + \big[F_{X_i} \circ \cdots \circ F_{X_1}(\theta_0(\gamma_s)) - F_{X_i} \circ \cdots \circ F_{X_1}(\theta_0^\circ(\gamma_s))\big] \\
&=: \sum_{k=1}^i \widetilde{Y}_i(\theta_0(\gamma_s), \theta_0^\circ(\gamma_s)).
\end{aligned} \tag{78}$$

Recall the geometric-moment contraction in Theorem 2.2 which holds for any $\gamma_s \in (0, \gamma(p))$. Given the contraction constant $0 < \rho_{\gamma,p} < 1$ defined in Theorem 2.2, we have

$$\|\widetilde{Y}_i(\theta_0(\gamma_s), \theta_0^\circ(\gamma_s))\|_p \leq [\rho_{\gamma,p}]^i |\theta_0 - \theta_0^\circ|.$$

Write $\widetilde{Y}_i = \widetilde{Y}_i(\theta_0(\gamma), \theta_0^\circ(\gamma))$. Also note that by the triangle inequality, it follows that $|\widetilde{S}_i(\gamma)| \leq \sum_{k=1}^i |\widetilde{Y}_i|$, which is a non-decreasing sequence as i grows. Therefore, $\max_{1\leq i\leq n} |\widetilde{S}_i(\gamma)| \leq \max_{1\leq i\leq n} \sum_{k=1}^i |\widetilde{Y}_i| = \sum_{k=1}^n |\widetilde{Y}_i|$, which along with the triangle inequality further yields

$$\Big\| \max_{1\leq i\leq n} |\widetilde{S}_i(\gamma)| \Big\|_p \leq \Big\| \sum_{k=1}^n |\widetilde{Y}_i| \Big\|_p \leq \sum_{i=1}^n \|\widetilde{Y}_i\|_p \leq \frac{\rho_{\gamma,p}(1-[\rho_{\gamma,p}]^n)}{1-\rho_{\gamma,p}} |\theta_0 - \theta_0^\circ|. \tag{79}$$

Hence, when the step n is large, with probability tending to 1, for $p > 2$, we can obtain

$$\frac{\big\| \max_{1\leq i\leq n} |S_i(\gamma) - S_i^\circ(\gamma)| \big\|_p}{n^{1/p}} \to 0, \quad \text{as } n \to \infty. \tag{80}$$

Combining expressions (77) and (80), we can achieve the desired result. □

A.4. Proof of Theorem 4.1.

PROOF OF THEOREM 4.1. It follows directly from Theorem 4 in [14] that under Assumptions 1–3, $\boldsymbol{M}_1 = \nabla m(\theta^*) \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \nabla m(\theta^*)$ is invertible. By the similar arguments in the proof of Lemma 18 in [14], it can be shown that

$$\boldsymbol{M}_2 = \nabla m(\theta^*) \otimes \boldsymbol{I}_d \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \nabla m(\theta^*) \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \boldsymbol{I}_d \otimes \nabla m(\theta^*)$$

is also invertible. In addition, the coefficient of the γ term is

$$\nabla m(\theta^*)^{-1} \nabla^2 m(\theta^*) \boldsymbol{M}_1^{-1} \mathbb{E}[\nabla g(\theta^*, X)^{\otimes 2}].$$

We only need to show the explicit form of Δ_2 and bound the residual term of order $O(\gamma^{5/2})$.

Recall the stationary SGD process $\{\theta_k^\circ(\gamma)\}_{k\in\mathbb{N}}$ distributed according to π_γ. W.l.o.g., we assume that $\theta^* = 0$. Here, we show the case with $d = 1$ for simplicity. It can be extended to multivariate case using the similar reasoning and writing the k-th order derivative $\nabla^k g(\theta, X)$ below as the k-th order tensor in $\mathbb{R}^{d^k}$. Given the stationary SGD updates

$$\theta_n^\circ(\gamma) = \theta_{n-1}^\circ(\gamma) - \gamma \nabla g(\theta_{n-1}^\circ(\gamma), X_n), \quad n \geq 1, \tag{81}$$

we shall calculate $\mathbb{E}\theta_n^{\circ 2}(\gamma)$, $\mathbb{E}\theta_n^{\circ 3}(\gamma)$ and $\mathbb{E}\theta_n^{\circ 4}(\gamma)$ to obtain Δ_1 and Δ_2. By taking the expectation on both sides of (81), for all $n \geq 1$, we have

$$\mathbb{E}[\nabla g(\theta_{n-1}^\circ(\gamma), X_n)] = 0. \tag{82}$$

For conciseness, we denote the expectation of the k-th order derivative at $\theta^* = 0$ by

$$g_k = \mathbb{E}[\nabla^k g(0, X_n)].$$

By Assumption 3, we can perform the fifth order Taylor expansion on $\nabla g(\theta_{n-1}^\circ(\gamma), X_n)$ at $\theta^* = 0$, which gives

$$\begin{aligned} 0 = {} & g_1 + g_2 \mathbb{E}[\theta_{n-1}^\circ(\gamma)] + \frac{1}{2} g_3 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^2\} + \frac{1}{6} g_4 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^3\} \\ & + \frac{1}{24} g_5 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^4\} + r_\gamma^{(5)}, \end{aligned} \tag{83}$$

where $r_\gamma^{(5)} \in \mathbb{R}^d$ satisfying $\|r_\gamma^{(5)}\|_p = O(\gamma^{5/2})$ for $p \geq 2$ by the similar arguments in Lemma 13 in [14]. Since $\theta^* = 0$ is the global optimum, it follows that $g_1 = 0$, and therefore, we have the bias expansion

$$\begin{aligned} \mathbb{E}[\theta_{n-1}^\circ(\gamma)] = {} & -\frac{1}{2} g_2^{-1} g_3 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^2\} + \frac{1}{6} g_2^{-1} g_4 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^3\} \\ & + \frac{1}{24} g_2^{-1} g_5 \mathbb{E}\{[\theta_{n-1}^\circ(\gamma)]^4\} + g_2^{-1} r_\gamma^{(5)}. \end{aligned} \tag{84}$$

Based on the proof of Theorem 4 in [14], we shall develop further on the remaining term of $\mathbb{E}[\theta_{n-1}^{\circ 2}(\gamma)]$ therein which was on the order of $O(\gamma^2)$. We aim to provide an explicit form of the coefficient in front of γ^2.

Next, we only need to express $\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^3\}$ and $\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^4\}$ in terms of the second moment $\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^2\}$. To this end, we first square both sides of (81) and take the expectation on both sides as well, which yields

$$\begin{aligned}\mathbb{E}\{[\theta_n^{\circ}(\gamma)]^2\} &= \mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^2\} - 2\gamma\mathbb{E}[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)\theta_{n-1}^{\circ}(\gamma)] \\ &\quad + \gamma^2\mathbb{E}\{[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^2\}. \end{aligned} \tag{85}$$

We simplify the equation above and obtain

$$0 = -2\mathbb{E}[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)\theta_{n-1}^{\circ}(\gamma)] + \gamma\mathbb{E}\{[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^2\}. \tag{86}$$

Again, we apply the fourth order Taylor expansions to $\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)$ at $\theta^* = 0$ on both sides, and obtain

$$\begin{aligned} 0 &= -2\mathbb{E}\Big\{ g_2[\theta_{n-1}^{\circ}(\gamma)]^2 + \frac{1}{2}g_3[\theta_{n-1}^{\circ}(\gamma)]^3 + \frac{1}{6}g_4[\theta_{n-1}^{\circ}(\gamma)]^4 \Big\} \\ &\quad + \gamma\mathbb{E}\Big\{ \Big[g_2\theta_{n-1}^{\circ}(\gamma) + \frac{1}{2}g_3\theta_{n-1}^{\circ 2}(\gamma) + \frac{1}{6}g_4\theta_{n-1}^{\circ 3}(\gamma) \Big]^2 \Big\} + o(\gamma^2), \end{aligned} \tag{87}$$

which implies that

$$\begin{aligned} 0 &= (-2g_2 + \gamma g_2^2)\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^2\} + g_3(-1 + \gamma g_2)\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^3\} \\ &\quad + \Big(-\frac{1}{3}g_4 + \frac{1}{4}g_3^2 + \frac{1}{3}g_2 g_4 \Big)\mathbb{E}\{[\theta_{n-1}^{\circ}(\gamma)]^4\} + o(\gamma^2). \end{aligned} \tag{88}$$

Next, we take the cube of both sides of (81) and take the expectation, which gives

$$\begin{aligned} \mathbb{E}[\theta_n^{\circ 3}(\gamma)] &= \mathbb{E}[\theta_{n-1}^{\circ 3}(\gamma)] - 3\gamma\mathbb{E}[\theta_{n-1}^{\circ 2}(\gamma)\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)] \\ &\quad + 3\gamma^2\mathbb{E}\{\theta_{n-1}^{\circ}(\gamma)[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^2\} - \gamma^3\mathbb{E}\{[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^3\}. \end{aligned} \tag{89}$$

We simplify it and get

$$\begin{aligned} 0 &= -3\mathbb{E}[\theta_{n-1}^{\circ 2}(\gamma)\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)] + 3\gamma\mathbb{E}\{\theta_{n-1}^{\circ}(\gamma)[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^2\} \\ &\quad - \gamma^2\mathbb{E}\{[\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)]^3\}. \end{aligned} \tag{90}$$

Again, we apply the third order Taylor expansion to $\nabla g(\theta_{n-1}^{\circ}(\gamma), X_n)$ at 0 and obtain

$$\begin{aligned} 0 &= -3\mathbb{E}\big[\theta_{n-1}^{\circ 2}(\gamma)\big(g_2\theta_{n-1}^{\circ}(\gamma) + \frac{1}{2}g_3\theta_{n-1}^{\circ 2}(\gamma)\big)\big] \\ &\quad + 3\gamma\mathbb{E}\big[\theta_{n-1}^{\circ}(\gamma)\big(g_2\theta_{n-1}^{\circ}(\gamma) + \frac{1}{2}g_3\theta_{n-1}^{\circ 2}(\gamma)\big)^2\big] \\ &\quad - \gamma^2\mathbb{E}\big[\big(g_2\theta_{n-1}^{\circ}(\gamma) + \frac{1}{2}g_3\theta_{n-1}^{\circ 2}(\gamma)\big)^3\big] + o(\gamma^2), \end{aligned} \tag{91}$$

which leads to

$$0 = g_2(-3 + 3\gamma g_2 - \gamma^2 g_2^2)\mathbb{E}[\theta_{n-1}^{\circ 3}(\gamma)] + g_3\Big(-\frac{3}{2} + 3\gamma - \frac{3}{2}\gamma^2 g_2^2 g_3 \Big)\mathbb{E}[\theta_{n-1}^{\circ 4}(\gamma)] + o(\gamma^2). \tag{92}$$

Thus, by combing expressions (84), (88) and (92), we can achieve the exact term in front of γ^2 which is independent of γ. This completes the proof.

□

A.5. Proofs of Theorems 4.3 & 4.4.

PROOF OF THEOREM 4.3. W.l.o.g., we show the one-dimensional case here. The multi-dimensional case can follow the similar arguments and thus the proof is omitted. Let $\widehat{\sigma}_n^2(\gamma)$ (resp. $\sigma_n^2(\gamma)$) be $\widehat{\Sigma}_n(\gamma)$ (resp. $\Sigma_n(\gamma)$) with $d=1$. It is straightforward to verify that as $M\to\infty$, we have $\eta_{M+1}/\eta_M\to 1$ and

$$\sum_{m=1}^{M}(\eta_{m+1}-\eta_m)^2 \asymp \eta_{M+1}(\eta_{M+1}-\eta_M).$$

Recall (29) for the stationary process $\{\vartheta_k^\circ(\gamma)\}_{k\in\mathbb{Z}}$. Following the construction of $V_n(\gamma)$ in (31), we define

$$V_n^\circ(\gamma)=\sum_{m=1}^{\psi(n)-1}\left|\sum_{k\in B_m}\vartheta_k^\circ(\gamma)\right|^2+\left|\sum_{k=\phi(n)}^{n}\vartheta_k^\circ(\gamma)\right|^2=:\sum_{m=1}^{\psi(n)-1}|\mathcal{S}_m^\circ(\gamma)|^2+|\mathcal{R}_n^\circ(\gamma)|^2.$$

By Theorem 1 and Theorem 2 in [61], it follows that

$$\mathbb{E}\left|\frac{V_n^\circ(\gamma)}{n\sigma^2(\gamma)}-1\right|^2\lesssim n^{(2/\beta-2)\vee(-1/\beta)}.$$

Then it suffices to upper bound $\mathbb{E}|V_n(\gamma)-V_n^\circ(\gamma)|^2$. To this end,

$$\|V_n(\gamma)-V_n^\circ(\gamma)\|_2\le\|\mathcal{V}_n(\gamma)-V_n^\circ(\gamma)\|_2+2\|H_n(\gamma)\widehat{\theta}_n(\gamma)\|_2+K_n\|\widehat{\theta}_n(\gamma)\|_4^2. \tag{93}$$

Denote $\widetilde{L}_\gamma=\max\{\rho_{\gamma,p},L(2\gamma)\}$. By Theorem 2.2, it follows that

$$\begin{aligned}\|\mathcal{V}_n(\gamma)-V_n^\circ(\gamma)\|_2&\le\sum_{m=1}^{\psi(n)-1}\|\mathcal{S}_m(\gamma)-\mathcal{S}_m^\circ(\gamma)\|_4\|\mathcal{S}_m(\gamma)+\mathcal{S}_m^\circ(\gamma)\|_4\\&\quad+\|\mathcal{R}_n(\gamma)-\mathcal{R}_n^\circ(\gamma)\|_4\|\mathcal{R}_n(\gamma)+\mathcal{R}_n^\circ(\gamma)\|_4\\&\lesssim\sum_{m=1}^{\psi(n)-1}\widetilde{L}_\gamma^{\eta_m}(\eta_{m+1}-\eta_m)^{1/2}+\widetilde{L}_\gamma^{\phi(n)}(n-\phi(n)+1)^{1/2}\le C.\end{aligned}$$

Similarly, we have $\|\widehat{\theta}_n(\gamma)\|_4^2\lesssim\|n^{-1}\sum_{k=1}^n\vartheta_k^\circ(\gamma)\|_4^2+\|n^{-1}\sum_{k=1}^n(\vartheta_k(\gamma)-\vartheta_k^\circ(\gamma))\|_4^2\asymp n^{-1}$ and $\|H_n(\gamma)\|_4^2\lesssim n(n-\phi(n))^2$. Combining these with the fact that $K_n\asymp n(n-\phi(n))$ yields $K_n\|\widehat{\theta}_n(\gamma)\|_4^2\asymp(n-\phi(n))$ and $\|H_n(\gamma)\widehat{\theta}_n(\gamma)\|_2\lesssim(n-\phi(n))$. Putting all these pieces together, we obtain $\|V_n(\gamma)-V_n^\circ(\gamma)\|_2\lesssim(n-\phi(n))\asymp n^{1-1/\beta}$ in view of (93). Consequently, it follows that

$$\mathbb{E}\left|\frac{\widehat{\sigma}_n^2(\gamma)}{\sigma^2(\gamma)}-1\right|^2\lesssim\mathbb{E}\left|\frac{V_n^\circ(\gamma)}{n\sigma^2(\gamma)}-1\right|^2+\mathbb{E}\left|\frac{V_n^\circ(\gamma)-V_n(\gamma)}{n\sigma^2(\gamma)}\right|^2\lesssim n^{(2/\beta-2)\vee(-1/\beta)}.$$

□

PROOF OF THEOREM 4.4. The proof of Theorem 4.4 is similar to that of Theorem 4.3 and thus omitted.

□

A.6. Proofs of Theorems 5.1 & 5.2.

PROOF OF THEOREM 5.1. Recall the i.i.d. observations $X_1, \ldots, X_n$ and the filtration $\mathcal{F}_k = (\ldots, X_{k-1}, X_k)$. We shall omit the dependence of $\theta_k^\circ(\gamma)$ on the step size γ throughout the rest of the proof for the notation simplicity. By induction, the first-order derivative process in (37) can be further written into

$$\begin{aligned}\dot{\theta}_k &= \boldsymbol{A}(\theta_{k-1}^\circ, X_k)\dot{\theta}_{k-1}(\gamma) - \nabla g(\theta_{k-1}^\circ, X_k) \\ &= \Big(\prod_{i=1}^{k} \boldsymbol{A}(\theta_{i-1}^\circ, X_i)\Big)\dot{\theta}_0 - \sum_{i=1}^{k}\Big(\prod_{t=1}^{k-i} \boldsymbol{A}(\theta_{k-t}^\circ, X_{k-t+1})\Big)\nabla g(\theta_{i-1}^\circ, X_i), \end{aligned} \tag{94}$$

where the matrix product $\prod_{t=1}^{k} \boldsymbol{A}(\theta_{t-1}^\circ, X_t) = \boldsymbol{A}(\theta_{k-1}^\circ, X_k)\cdots\boldsymbol{A}(\theta_0^\circ, X_1)$. Note that the random matrices $\boldsymbol{A}(\theta_{k-1}^\circ, X_k)$ are dependent over k. Since $\{\theta_k^\circ\}_{k\in\mathbb{N}}$ is a stationary process, we shall apply the ergodic theorem as shown in [**?**] and we need to prove that, for any unit vector $\delta \in \mathbb{R}^d$,

$$\prod_{t=1}^{k} |\boldsymbol{A}(\theta_{t-1}^\circ, X_t)\delta| \to 0.$$

To this end, we shall note that by the condition in (39) and Jensen's inequality, we have, for any unit vector $\delta \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\big[\log\big|\boldsymbol{A}(\theta_{k-1}^\circ, X_k)\delta\big|\big] &= \mathbb{E}\Big[\mathbb{E}\big[\log\big|\boldsymbol{A}(\theta_{k-1}^\circ, X_k)\delta\big| \mid \mathcal{F}_{k-1}\big]\Big] \\ &\le \mathbb{E}\Big[\frac{1}{p}\log\mathbb{E}\big[\big|\boldsymbol{A}(\theta_{k-1}^\circ, X_k)\delta\big|^p \mid \mathcal{F}_{k-1}\big]\Big] \\ &\le \frac{1}{p}\log\mathbb{E}_{\theta^\circ\sim\pi_\gamma}[A_p(\theta_{k-1}^\circ)] < 0. \end{aligned} \tag{95}$$

Hence, by the strong law of larger numbers, we can achieve

$$\frac{1}{k}\sum_{i=1}^{k}\log\big|\boldsymbol{A}(\theta_{i-1}^\circ, X_i)\delta\big| \to \mathbb{E}\big[\log\big|\boldsymbol{A}(\theta_0^\circ, X_1)\delta\big|\big], \quad \text{a.s.} \tag{96}$$

which further gives

$$\prod_{i=1}^{k}\big|\boldsymbol{A}(\theta_{i-1}^\circ, X_i)\delta\big| = \exp\Big\{-k\cdot\frac{1}{k}\log\big|\boldsymbol{A}(\theta_{i-1}^\circ, X_i)\delta\big|\Big\} \to 0. \quad \text{a.s.} \tag{97}$$

Finally, we have, almost surely,

$$\dot{\theta}_k \to \sum_{i=1}^{k}\Big(\prod_{t=1}^{k-i}\boldsymbol{A}(\theta_{k-t}^\circ, X_{k-t+1})\Big). \tag{98}$$

This proves the existence of a stationary solution of the recursion (37). □

PROOF OF THEOREM 5.2. Recall the explicit form of the first-order derivative process $\dot{\theta}_k(\gamma)$ derived in expression (94). First, we show that, by assuming the existence of the Hessian matrix $\nabla^2 g(\theta, X)$ with respect to θ,

$$\mathbb{E}\big[|\dot{\theta}_k(\gamma)|^{p/2}\big] < \infty, \quad \text{for } p \ge 2. \tag{99}$$

Recall the i.i.d. observations $X_1, \ldots, X_n$ and the filtration $\mathcal{F}_k = (\ldots, X_{k-1}, X_k)$. Write $\boldsymbol{A}_k = \boldsymbol{A}(\theta^\circ_{k-1}, X_k)$ for convenience. Denote the operator norm by $\|\cdot\|_{\text{op}}$. When $p = 2$, for any unit vector $\delta \in \mathbb{R}^d$, we have

$$\begin{aligned}\mathbb{E}[|\boldsymbol{A}_k \cdots \boldsymbol{A}_1 \delta|^2] &= \mathbb{E}\big[\mathbb{E}[(\delta^\top \boldsymbol{A}_1^\top \ldots \boldsymbol{A}_k^\top \boldsymbol{A}_k \ldots \boldsymbol{A}_1 \delta) \mid \mathcal{F}_{k-1}]\big] \\ &= \mathbb{E}\big[\delta^\top \boldsymbol{A}_1^\top \ldots \boldsymbol{A}_{k-1}^\top \mathbb{E}(\boldsymbol{A}_k^\top \boldsymbol{A}_k) \boldsymbol{A}_{k-1} \ldots \boldsymbol{A}_1 \delta\big] \\ &\le \|\mathbb{E}[\boldsymbol{A}_k^\top \boldsymbol{A}_k]\|_{\text{op}} \cdot \mathbb{E}[\delta^\top \boldsymbol{A}_1^\top \ldots \boldsymbol{A}_{k-1}^\top \boldsymbol{A}_{k-1} \ldots \boldsymbol{A}_1 \delta],\end{aligned}$$

where $\|\mathbb{E}[\boldsymbol{A}_k^\top \boldsymbol{A}_k]\|_{\text{op}} < 1$ as implied by the condition in (39). Therefore, by induction, we have $\mathbb{E}[|\boldsymbol{A}_k \cdots \boldsymbol{A}_1 \delta|^2] < 1$. For $p > 2$, recall A_p^* defined in (41). Then, it similarly follows that

$$\begin{aligned}\mathbb{E}[|\boldsymbol{A}_k \cdots \boldsymbol{A}_1 \delta|^p] &= \mathbb{E}\big[|\delta^\top \boldsymbol{A}_1^\top \ldots \boldsymbol{A}_k^\top \boldsymbol{A}_k \ldots \boldsymbol{A}_1 \delta|^{p/2}\big] \\ &= \mathbb{E}\big[\mathbb{E}[|\delta^\top \boldsymbol{A}_1^\top \ldots \boldsymbol{A}_k^\top \boldsymbol{A}_k \ldots \boldsymbol{A}_1 \delta|^{p/2} \mid \mathcal{F}_{k-1}]\big] \\ &\le \prod_{i=1}^{k} \sup_{\delta \in \mathbb{R}^d, |\delta|=1} \mathbb{E}|\boldsymbol{A}_i \delta|^p \le [A_p^*]^k < 1.\end{aligned}$$

Note that by Fatou's Lemma and the Hölder inequality, we can achieve, for all large $k \in \mathbb{N}$,

$$\begin{aligned}\|\dot\theta_k(\gamma)\|_{p/2} &= \big\|\limsup_{k\to\infty} \dot\theta_k(\gamma)\big\|_{p/2} \\ &\le \limsup_{k\to\infty} \sum_{i=1}^{k} \Big[\prod_{t=1}^{k-i} \big(\sup_{\delta \in \mathbb{R}^d, |\delta|=1} \mathbb{E}|\boldsymbol{A}_{k-t+1}\delta|^p\big)^{1/p} \|\nabla g(\theta_{i-1}, X_i)\|_p\Big] \\ &\le \sum_{i=1}^{\infty} [A_p^*]^{i/p} \|\nabla g(\theta_{i-1}, X_i)\|_p, \qquad (100)\end{aligned}$$

which is bounded since the constant $A_p^* < 1$ and the p-th moment $\|\nabla g(\theta_{i-1}, X_i)\|_p < \infty$ as indicated by Assumption 2. Thus, we achieve the desired result in (i).

Next, we aim to show that the existence of a stationary solution for the second-order derivative process with respect to γ. Recall the recursion procedure in (40) that

$$\begin{aligned}\ddot\theta_k(\gamma) &= \boldsymbol{A}(\theta^\circ_{k-1}, X_k)\ddot\theta_{k-1}(\gamma) + [\partial_\gamma \boldsymbol{A}(\theta^\circ_{k-1}, X_k)]\dot\theta_{k-1}(\gamma) - \partial_\gamma \nabla g(\theta^\circ_{k-1}, X_k) \\ &=: \boldsymbol{A}(\theta^\circ_{k-1}, X_k)\ddot\theta_{k-1}(\gamma) + \boldsymbol{c}(\theta^\circ_{k-1}, X_k), \qquad (101)\end{aligned}$$

where

$$\boldsymbol{c}(\theta^\circ_{k-1}, X_k) = [\partial_\gamma \boldsymbol{A}(\theta^\circ_{k-1}, X_k)]\dot\theta_{k-1}(\gamma) - \partial_\gamma \nabla g(\theta^\circ_{k-1}, X_k). \qquad (102)$$

Due to the same structure of (101) compared to (37), the same arguments apply regarding the contraction of the random coefficient matrix $\boldsymbol{A}(\theta, X)$ which assures the existence of a stationary solution for the recursive procedure $\{\ddot\theta_k(\gamma)\}_{k\in\mathbb{N}}$. Further, notice that the structure of the recursive function $\boldsymbol{c}(\theta, X)$ in (102) is also similar to the one in (37) and we have already shown $\|\dot\theta_k(\gamma)\|_{p/2} < \infty$. Therefore, we only need to prove $\|\boldsymbol{c}(\theta, X)\|_p < \infty$ to obtain $\|\ddot\theta_k(\gamma)\|_{p/2} < \infty$. To see this, we note that by the definition of $\boldsymbol{A}(\theta, X)$ in (38), we have

$$\partial_\gamma \boldsymbol{A}(\theta^\circ_{k-1}, X_k) = -\nabla g(\theta^\circ_{k-1}, X_k) - \gamma \partial_\gamma \nabla g(\theta^\circ_{k-1}, X_k), \qquad (103)$$

which exists and has bounded p-th moment by assumptions. Therefore, we have $\|\ddot\theta_k(\gamma)\|_{p/2} < \infty$ for some fixed $\gamma \in (0, \gamma(p))$. □

APPENDIX B: ADDITIONAL INFORMATION OF SIMULATION STUDIES

B.1. Improved ASGD Estimates by Multiple Step Sizes. We consider the SGD procedure with three different constant step sizes: $\gamma = 0.0125$, 0.025 and 0.05. For simplicity, we denote respectively the extrapolated estimators using two and three different step sizes by

$$\widehat{\theta}_n^{(2)}(\gamma) = 2\bar{\theta}_n(\gamma) - \bar{\theta}_n(2\gamma),$$

and

$$\widehat{\theta}_n^{(3)}(\gamma) = \frac{8}{3}\bar{\theta}_n(\gamma) - 2\bar{\theta}_n(2\gamma) + \frac{1}{3}\bar{\theta}_n(4\gamma).$$

TABLE 2
Linear regression. Absolute value of the average bias over 100 independent runs.

step size	estimator	$n = 100$	$n = 1000$	$n = 5000$	$n = 10^4$	$n = 10^5$	$n = 10^6$
$\gamma = 0.025$	$\bar{\theta}_n(\gamma)$	1.942e-01	1.782e-02	4.520e-03	1.923e-03	1.823e-04	9.154e-05
	$\widehat{\theta}_n^{(2)}(\gamma)$	1.375e-01	1.462e-02	2.915e-03	1.355e-03	6.670e-05	**8.489e-06**
	$\widehat{\theta}_n^{(3)}(\gamma)$	8.638e-02	8.363e-03	2.823e-03	9.776e-04	2.470e-05	**1.725e-06**
$\gamma = 0.05$	$\bar{\theta}_n(\gamma)$	1.145e-01	1.911e-02	5.352e-03	2.632e-03	5.348e-04	1.327e-04
	$\widehat{\theta}_n^{(2)}(\gamma)$	8.051e-02	7.328e-03	3.207e-03	1.066e-03	3.083e-04	**8.604e-05**
	$\widehat{\theta}_n^{(3)}(\gamma)$	5.274e-02	4.713e-03	2.490e-03	8.798e-04	1.410e-04	**3.584e-06**

B.2. Comparison of Two Online Estimators of Long-Run Covariance. We report the convergence trace of our proposed online estimators with different constant step sizes. In Figure 3, we show the difference between the bias of the non-debiased and the bias-reduced estimators. One can observe that our bias reduction works robustly across different step sizes.

Further, we report more cases of the estimated long-run variances in both linear regression and logistic regression. Since we do not have a closed-form solution for the true long-run variance of the logistic regression, we simply choose the estimated long-run but with an extremely large iteration step number, i.e., 0.1 billion, as the baseline. As shown in Figure 4–7, our two proposed online estimators can converge to the true long-run as the step number increases.

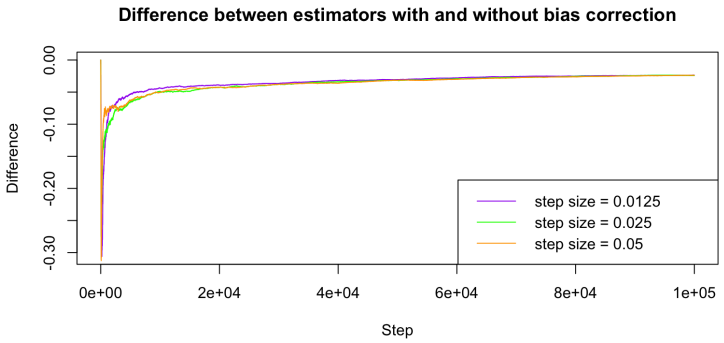
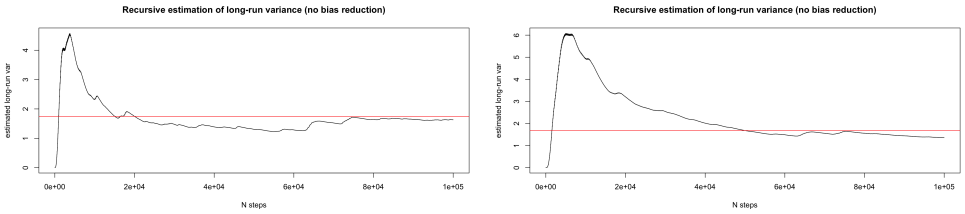
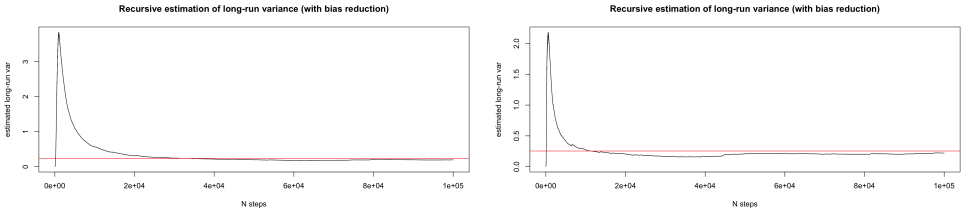


Fig 3: Difference between the estimators with and without bias correction, i.e., $\widehat{\sigma} - \widetilde{\sigma}$. The accuracy of the bias-reduced estimator outperforms the non-debiased one regardless of different step sizes.



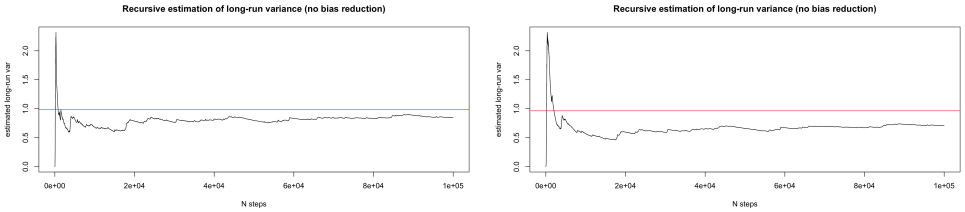
(a) Linear regression, $\gamma = 0.01$ (b) Linear regression, $\gamma = 0.005$

Fig 4: Recursive estimation of the long-run variance **without** bias reduction. The red line is the oracle value with $n = 0.1$ billion.



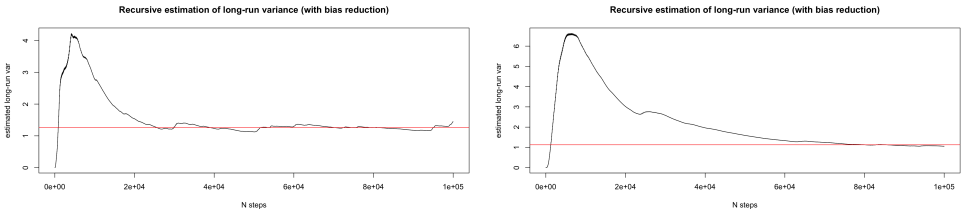
(a) Linear regression, $\gamma = 0.01$ (b) Linear regression, $\gamma = 0.005$

Fig 5: Recursive estimation of the long-run variance **with** bias reduction. The red line is the oracle value with $n = 0.1$ billion.



(a) Logistic regression, $\gamma = 0.01$ (b) Logistic regression, $\gamma = 0.005$

Fig 6: Recursive estimation of the long-run variance **without** bias reduction. The red line is the oracle value with $n = 0.1$ billion.



(a) Logistic regression, $\gamma = 0.01$ (b) Logistic regression, $\gamma = 0.005$

Fig 7: Recursive estimation of the long-run variance **with** bias reduction. The red line is the oracle value with $n = 0.1$ billion.