An Analysis of Constant Step Size SGD in the Non-convex Regime: Asymptotic Normality and Bias

Lu Yu*   Krishnakumar Balasubramanian†   Stanislav Volgushev‡   Murat A. Erdogdu§

July 29, 2020

Abstract

Structured non-convex learning problems, for which critical points have favorable statistical properties, arise frequently in statistical machine learning. Algorithmic convergence and statistical estimation rates are well-understood for such problems. However, quantifying the uncertainty associated with the underlying training algorithm is not well-studied in the non-convex setting. In order to address this short-coming, in this work, we establish an asymptotic normality result for the constant step size stochastic gradient descent (SGD) algorithm—a widely used algorithm in practice. Specifically, based on the relationship between SGD and Markov Chains [DDB19], we show that the average of SGD iterates is asymptotically normally distributed around the expected value of their unique invariant distribution, as long as the non-convex and non-smooth objective function satisfies a dissipativity property. We also characterize the bias between this expected value and the critical points of the objective function under various local regularity conditions. Together, the above two results could be leveraged to construct confidence intervals for non-convex problems that are trained using the SGD algorithm.

1 Introduction

Non-convex learning problems are prevalent in modern statistical machine learning applications such as matrix and tensor completion [GHJY15, GLM16, XYZ19, CLC19, CLPC19], deep neural networks [GBC16, JK17, MIG+19], and robust empirical risk minimization [Loh17, LLM18, MBM18]. Developing theoretically principled approaches for tackling such non-convex problems depends critically on the interplay between two aspects. From a computational perspective, variants of stochastic gradient descent (SGD) converge to first-order critical points [GL13, FLLZ18] or local minimizers [NP06, GHJY15, JGN+17, TSJ+18] of the objective function. From a statistical perspective, *oftentimes* these critical points or local minimizers have nice statistical properties [Kaw16, GLM16, Loh17, MMMO17, EvdG18, CLC19]; see also [FD82] for a counterexample. For the purpose of uncertainty quantification in such non-convex learning paradigms, studying the fluctuations of iterative algorithms used for training becomes extremely important. In this work, we focus on the widely used constant step size SGD, and develop results for quantifying the uncertainty associated with this algorithm for a class of non-convex problems.

---

*Department of Statistical Sciences at the University of Toronto, and Vector Institute `luyu@utstat.toronto.edu`

†Department of Statistics, University of California, Davis `kbala@ucdavis.edu`

‡Department of Statistical Sciences at the University of Toronto `stanislav.volgushev@utoronto.ca`

§Department of Computer Science and Department of Statistical Sciences at the University of Toronto, and Vector Institute `erdogdu@cs.toronto.edu`

Specifically, we consider minimizing a non-smooth and non-convex objective function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\min_{\theta \in \mathbb{R}^d} f(\theta). \tag{1.1}$$

The iterations of SGD with a constant step size $\eta > 0$, initialized at $\theta_0^{(\eta)} \equiv \theta_0 \in \mathbb{R}^d$, are given by

$$\theta_{k+1}^{(\eta)} = \theta_k^{(\eta)} - \eta \big(\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}) \big), \quad k \geq 0, \tag{1.2}$$

where $\{\xi_k\}_{k \geq 1}$ is a sequence of random functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ corresponding to the stochasticity in the gradient estimate. Several problems in machine learning and statistics are naturally formulated as the optimization problem is (1.1), where the function $f(\theta)$ is given by

$$f(\theta) := \mathbb{E}_Z[F(\theta, Z)] = \int F(\theta, Z) \, dP(Z), \tag{1.3}$$

where the function $F(\theta, Z)$ is typically the loss function composed with functions from hypothesis class parametrized by $\theta \in \mathbb{R}^d$, and depends on the random variable $Z \in \mathbb{R}^p$. The distribution $P(Z)$ is typically unknown. Then the iterations of online SGD with a constant step size $\eta > 0$, and batch-size b_k, initialized at $\theta_0^{(\eta)} \equiv \theta_0 \in \mathbb{R}^d$, are given by

$$\theta_{k+1}^{(\eta)} = \theta_k^{(\eta)} - \frac{\eta}{b_k} \sum_{j=1}^{b_k} \nabla F(\theta_k^{(\eta)}, Z_j), \quad k \geq 0, \tag{1.4}$$

where independent samples $Z_j \sim P(Z)$, is used to estimate the true gradient in each iteration k. Furthermore, the samples Z_j used across all iterates k are also independent. Typically, we also have $\nabla F(\theta, Z)$ to be unbiased estimates of the true gradient $\nabla f(\theta)$, for all $\theta \in \mathbb{R}^d$. The above iterates are indeed a special case of the iterates in (1.2), with the noise sequence $\{\xi_k(\theta_k^{(\eta)})\}_{k \geq 1}$ given by

$$\xi_k(\theta_k^{(\eta)}) := \frac{1}{b_k} \sum_{j=1}^{b_k} \left[\nabla F(\theta_k^{(\eta)}, Z_j) - \nabla f(\theta_k^{(\eta)}) \right].$$

Although proposed in the 1950s by [RM51], SGD has been the algorithm of choice for training statistical models due to its simplicity, and superior performance in large-scale settings [FP99, DDB19, WRS$^+$17, BH18]. However, the fluctuations of this algorithm is well-understood only when the objective function f is strongly convex and smooth, and the step size η satisfies a specific decreasing schedule so that the iterates asymptotically converge to the *unique* minimizer [PJ92, DR18, ABE19]. On the other hand, it is well-known that the SGD iterates in (1.2) can be viewed as a Markov chain which allows them to converge to a random vector rather than a single critical point [DDB19]. Building on this analogy between SGD and Markov chains, the aforementioned shortcomings can be alleviated by simply relaxing the global smoothness as well as the strong convexity assumptions to the tails of the objective function f, which allows for non-convex structure around the region of interest. Similar kinds of tail relaxations have been successfully employed in the diffusion theory when the target potential is non-convex [RRT17, CCAY$^+$18, EMS18], but they are not studied in the context of non-convex optimization with the SGD algorithm. In this work, we study the fluctuations and the bias of the averaged SGD iterates in (1.2), around the first-order critical points of the minimization problem (1.1). Our contributions can be summarized as follows.

- For a non-convex and non-smooth objective function f with tails growing at least quadratically, we establish the uniqueness of the stationary distribution of the constant step size SGD iterates in Proposition 2.1, and the asymptotic normality of Polyak-Ruppert averaging in Theorem 2.1. To the best of our knowledge, these are the first uniqueness and normality results for the SGD algorithm when the objective function is non-convex (even not strongly convex) and non-smooth.

- We further show in Proposition 3.1 that, under the assumptions leading to the CLT, the asymptotic bias between the expectation of the Lipschitz test function ϕ under the stationary distribution of the SGD iterates and the value of ϕ at any first-order critical point is bounded by a constant depending on the tail growth properties of f.

- Finally, we show in Theorems 3.1 and 3.2 that with additional local smoothness assumptions on the function f that allow non-convexity, we can establish a control over the bias in terms of step size. We further characterize the bias when the objective is (not strongly) convex in Theorem 3.3, providing a thorough bias analysis for the constant step size SGD under various settings that are frequently encountered in statistics.

Our results provide algorithm-dependent guarantees for uncertainty quantification, and they could be potentially leveraged to obtain confidence intervals for non-convex and non-smooth learning problems. This is contrary to the majority of the existing results in statistics, which only establish normality results for the true stationary point of the non-convex objective function; see for example [Loh17, QCLP19]. While being useful, such results completely ignore the computational hardships associated with non-convex optimization; hence, their practical implications are limited. On the other hand, in the optimization and learning theory literature, a majority of the existing results establish the rate of convergence of an algorithm to a critical point, and do not quantify the fluctuations associated with that algorithm. Our work bridges these separate lines of thought by providing asymptotic normality results directly for the SGD algorithm used for minimizing non-convex and non-smooth functions.

More Related Works. Establishing asymptotic normality results for the SGD algorithm began with the works of [Chu54, Sac58, Fab68, Rup88, Sha89], with [PJ92] providing a definitive result for strongly convex objectives. In particular, [PJ92] and [Rup88] established that the averaged SGD iterates with an appropriately chosen decreasing step size is asymptotically normal with the variance achieving the Cramer-Rao lower bound for parameter estimation. Recent works, for example [TFBJ18, SZ18, DR18, TA17, FXY18], leverage the asymptotic normality analysis of [PJ92], and compute confidence intervals for SGD. The benefits of constant step size SGD for faster convergence under overparametrization has also be demonstrated in the works of [SR13, NWS14, MBB18, VBS19]. The use of Markov chain theory to study constant step size stochastic approximation algorithms has been considered in several works [Kif88, Ben96, PV98, FP99, AMP00, TV19]. Recently, [DDB19, CT18] investigated the asymptotic variance of constant step size SGD. We emphasize here that most of the above works assume strongly convex and smooth objective functions. Finally, there exists a vast literature on analyzing Markov chain Monte Carlo sampling algorithms based on discretizing diffusions. We refer the interested reader to [DK17, BDMP17, CCAY$^+$18, DM17, Dal17, CCBJ17, BEL18, DCWY18, DRD18, LWME19, SL19, EH20] and the references therein, for details.

Notation. For $a, b \in \mathbb{R}$, denote by $a \vee b$ and $a \wedge b$ the maximum and the minimum of a and b, respectively. We use $\|\cdot\|$ to denote the Euclidean norm in $\mathbb{R}^d$. We denote the largest eigenvalue of the matrix A as $\lambda_{\max}(A)$, and the smallest one as $\lambda_{\min}(A)$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ represent a probability

space, and denote by $\mathcal{B}(\mathbb{R}^d)$, the Borel σ-field of $\mathbb{R}^d$. Let $\mathcal{P}_k(\mathbb{R}^d) := \{\nu : \int_{\mathbb{R}^d} \|\theta\|^k \nu(d\theta) < \infty\}$ denote the set of probability measures with finite k-th moments. For a probability distribution π and a function g on $\mathcal{X}$, we define $\pi(g) := \int_{\mathcal{X}} g(x) d\pi(x)$, and $\mathcal{L}_2(\pi) := \{g : \mathcal{X} \to \mathbb{R} : \pi(g^2) < \infty\}$.

2 Central Limit Theorem for The Constant Step Size SGD

In this section, we establish an asymptotic central limit theorem (CLT) for the Polyak-Ruppert averaging of the constant step size SGD iterates given in (1.2) when the objective function is potentially non-convex, non-smooth, and has quadratically growing tails. More specifically, we first prove that there exists a unique stationary distribution $\pi_\eta \in \mathcal{P}_2(\mathbb{R}^d)$ for the Markov chain defined by the SGD algorithm when the objective function is dissipative (see Assumption 2.2) with gradient exhibiting at most linear growth (see Assumption 2.1). Furthermore, under the same conditions, we prove that a CLT holds for the Polyak-Ruppert averaging, and it is independent of the initialization. In what follows, we list and discuss the main assumptions required to establish a CLT for the SGD iterates, and compare them to those existing in the literature.

Assumption 2.1 (Linear growth). *The gradient of the objective function f has at most linear growth. That is, for some $L \geq 0$, we have*

$$\|\nabla f(\theta)\| \leq L(1 + \|\theta\|) \quad for \ all \quad \theta \in \mathbb{R}^d.$$

Majority of the results on SGD focus on smooth functions with gradients satisfying $\|\nabla f(\theta) - \nabla f(\theta')\| \leq \|\theta - \theta'\|$ for all $\theta, \theta' \in \mathbb{R}^d$; see e.g. [PJ92, DDB19]. The above condition allows for non-smooth objectives, and is a significant relaxation of the standard Lipschitz gradient condition.

Assumption 2.2 (Dissipativity). *The objective function f is (α, β)-dissipative. That is, there exists positive constants α, β such that*

$$\langle \theta, \nabla f(\theta) \rangle \geq \alpha \|\theta\|^2 - \beta \quad for \ all \ \ \theta \in \mathbb{R}^d.$$

The dissipativity assumption has its origins in the analysis of dynamical systems, and is used widely in the analysis of optimization and learning algorithms [MSH02, RRT17, EMS18, XCZG18]. It could be viewed as a relaxation of strong convexity since it restricts the quadratic growth assumption to the tails of the function f, enforcing no local growth around the first-order critical points. A canonical example for this condition is the sum of a quadratic and any non-convex function with bounded gradient. For example, consider the function $x \to x^2 + 10 \sin(x)$ which is clearly non-convex and $(1, 25)$-dissipative. It is worth mentioning that many statistical learning problems such as phase retrieval [TV19] satisfy Assumption 2.2.

Assumption 2.3 (Noise sequence). *Gradient noise sequence $\{\xi_k\}_{k \geq 1}$ is a collection of i.i.d. random fields satisfying*

$$\mathbb{E}[\xi_1(\theta)] = 0 \ \ and \ \ \mathbb{E}^{1/2}[\|\xi_1(\theta)\|^2] \leq L_\xi(1 + \|\theta\|),$$

for any $\theta \in \mathbb{R}^d$ and a positive constant L_ξ. Moreover, for each $\theta \in \mathbb{R}^d$ the distribution of the random variable $\xi_1(\theta)$ can be decomposed as $\mu_{1,\theta} + \mu_{2,\theta}$ where $\mu_{1,\theta}$ has a density, say p_θ, with respect to Lebesgue measure which satisfies $\inf_{\theta \in C} p_\theta(t) > 0$ for any bounded set C and any $t \in \mathbb{R}^p$.

Assumption 2.3 as formulated above is stronger than what is used in the proofs. It can easily be seen that the lower bound on the density p_θ is only required to hold for a specific set whose form

depends on η and various constants from Assumptions 2.1–2.3. The form of this set is complicated, and an exact expression is given in the Appendix – see equation (A.3). We also emphasize that Assumption 2.3 also does not specify any explicit parametric form for the distribution of the noise sequence contrary to recent works in non-convex settings where dissipitavity condition has been heavily utilized [RRT17, XCZG18, EMS18].

We now establish the existence and uniqueness of the stationary distribution of the SGD iterates (1.2).

Proposition 2.1 (Ergodicity of SGD). *Let the Assumptions 2.1-2.3 hold. For a step size satisfying*

$$\eta < \frac{\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi)) \vee 0}}{3L^2 + L_\xi},$$

the following statements hold for the SGD (1.2).

(a) *SGD iterates admit a unique stationary distribution $\pi_\eta \in \mathcal{P}_2(\mathbb{R}^d)$, depending on the choice of step size η.*

(b) *For a test function $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfying $|\phi(\theta)| \leq L_\phi(1 + \|\theta\|)\ \forall \theta \in \mathbb{R}^d$ and some $L_\phi > 0$, and for any initialization $\theta_0^{(\eta)} = \theta_0 \in \mathbb{R}^d$ of the SGD algorithm, there exists $\rho \in (0, 1)$ and κ (both depending on η) such that we have*

$$\left| \mathbb{E}\big[\phi\big(\theta_k^{(\eta)}\big)\big] - \pi_\eta(\phi) \right| \leq \kappa\, \rho^k (1 + \|\theta_0\|^2),$$

where $\pi_\eta(\phi) := \int \phi(x) d\pi_\eta(x)$.

The uniqueness of the stationary distribution of the constant step size SGD has been established in [DDB19] for strongly convex and smooth objectives. In Proposition 2.1, we relax both of these assumptions allowing for non-convex and non-smooth objectives. Our proof relies on V-uniform ergodicity [MT12], which is fundamentally different from the ergodicity analysis in [DDB19]. Under the dissipativity condition (quadratic growth of f), geometric ergodicity in Proposition 2.1 is not surprising; yet, it is worth highlighting that the function f as well as the noise sequence require significantly less structure than what was assumed in the literature. The above step size assumption is almost standard and it is required to obtain a uniform bound on the moments of SGD iterates. We highlight that similar to the gradient descent algorithm, the step size depends on a quantity that serves as a *surrogate* condition number in our setting, namely, L/α. For the purposes of establishing a CLT, it is sufficient to consider moments of order 4 (in fact any order larger than 2 suffices), but it is also worth noting that any order moments of SGD can be controlled under Assumption 2.2 as long as the noise has the same order finite moment.

Next, we state our first principal contribution, a central limit theorem for the averaged SGD iterates starting from any initial distribution for a non-convex objective. For a test function $\phi : \mathbb{R}^d \to \mathbb{R}$, we denote the centered partial sums of ϕ evaluated at the SGD iterates with $S_n(\phi)$, i.e., we define

$$S_n(\phi) := \sum_{k=0}^{n-1} \left[\phi\big(\theta_k^{(\eta)}\big) - \pi_\eta(\phi) \right] \quad \text{where} \quad \pi_\eta(\phi) := \int \phi(x) d\pi_\eta(x).$$

Theorem 2.1 (CLT). *Let the Assumptions 2.1-2.3 hold. For a step size η and a test function ϕ satisfying the conditions in Proposition 2.1, we define $\sigma_{\pi_\eta}^2(\phi) := \lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{\pi_\eta}\big[S_n^2(\phi)\big]$. Then,*

$$n^{-1/2} S_n(\phi) \xrightarrow{d} \mathcal{N}\big(0, \sigma_{\pi_\eta}^2(\phi)\big).$$

The above result characterizes the fluctuations of a test function ϕ averaged across SGD iterates, even when the objective function is both non-convex and non-smooth. The asymptotic variance in the above CLT can be equivalently stated in another compact form. If we define the centered test function as $h(\theta) = \phi(\theta) - \pi_\eta(\phi)$, the asymptotic variance can be written as

$$\sigma^2_{\pi_\eta}(\phi) = 2\pi_\eta(h\hat{h}) - \pi_\eta(h^2) \quad \text{where} \quad \hat{h} = \sum_{k=0}^\infty \mathbb{E}\Big[h\big(\theta_k^{(\eta)}\big)\Big].$$

Indeed, this is the variance we compute at the end of our proof in Section A. However, the expression in Theorem 2.1 is obtained by simply applying [DMPS18, Thm 21.2.6]. For the case of strongly convex functions with decreasing step size schedule, it is well-known from the works of [PJ92, Rup88] that the limiting variance of the averaged SGD iterates achieves the Cramer-Rao lower bound for parameter estimation; see also [MB11, ABE19] for non-asymptotic rates in various metrics. The question of providing lower bounds for the limiting variance of the critical points in the non-convex setting is extremely subtle, and is often handled on a case-by-case basis. We refer the interested reader to [Gey94, Sha00, Loh17].

There are several important implications of the above CLT result for constructing confidence intervals in practice. First note that, following the standard construction in inference, one can write the distribution of the sample mean approximately as $n^{-1}S_n(\phi) \approx \mathcal{N}\big(0, n^{-1}\sigma^2_{\pi_\eta}(\phi)\big)$. Here, one needs to estimate the population quantity, the asymptotic variance $\sigma^2_{\pi_\eta}(\phi)$, for the purpose of obtaining confidence intervals. In Section 5, we discuss three strategies for estimating this quantity, which could be eventually used for inference in practice. A theoretical analysis of the proposed approaches in Section 5 is beyond the scope of this work.

3 Bias of the Constant Step Size SGD

In this section, we present a thorough analysis of the bias of constant step size SGD algorithm. We first show in Section 3.1 that, in the non-convex and non-smooth case for which we established the CLT, the SGD algorithm converges to a ball that contains all the first-order critical points exponentially fast; nevertheless, the bias is not controllable with the step size. Motivated by this, we provide three types of bias analyses in Section 3.2 under different local growth assumptions on the objective f, characterizing the bias behavior in various non-convex and convex settings. For this, we strength the assumption of the noise sequence as follows.

Assumption 3.1 (Noise sequence). *Gradient noise sequence $\{\xi_k\}_{k\geq 1}$ satisfies Assumption 2.3, and it holds for any $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|\xi_1(\theta)\|^4\big] \leq L_\xi(1 + \|\theta\|^4),$$

where L_ξ is the same as in Assumption 2.3.

3.1 Bias without Local Regularity

Bias behavior of an algorithm is intimately related to the local properties of the objective at critical points. Therefore, under the mild assumptions that yield the CLT, one cannot expect a tight control over the bias. However, the tail growth condition is sufficient for a rough characterization, which is still important because even the points that are close to the local minimizers generally have favorable computational [BVB16, MMMO17, CLC19], and statistical properties [Loh17, EvdG18, QCLP19].

If Assumption 2.2 holds for an objective function f, all first-order critical points of f must lie inside a ball of radius $\sqrt{\beta/\alpha}$. Based on this, we show that the SGD iterates (1.2) will move towards this ball exponentially fast, which ultimately establishes a bound on the non-asymptotic bias, and in the limit case yields a bound on the asymptotic bias. The following result formalizes this statement.

Proposition 3.1. *Let Assumptions 2.1, 2.2, and 3.1 hold. For θ^* denoting an arbitrary critical point of the objective function f, define the constants $\bar{L} := L(1 + \|\theta^*\|)$, and*

$$
\begin{aligned}
c_{L,\alpha} &:= \left[\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi)) \vee 0} \right] / [3L^2 + L_\xi] \\
c_{L,\alpha}^\dagger &:= \left[\alpha - \sqrt{(\alpha^2 - 16L_\dagger) \vee 0} \right] / (64 L_\dagger)
\end{aligned}
\tag{3.1}
$$

with $L_\dagger := \bar{L}^2 + 16\Big(L_\xi^{3/4}\big(1 + (\beta/\alpha)^3\big) \vee L_\xi^{1/2}\big(1 + (\beta/\alpha)^2\big) \vee L_\xi\big(1 + (\beta/\alpha)^4\big)\Big)$. Then, for SGD iterates initialized at a fixed point $\theta_0 \in \mathbb{R}^d$ and a step size satisfying $\eta < 1 \wedge \frac{1}{10\bar{L}} \wedge c_{L,\alpha} \wedge c_{L,\alpha}^\dagger$, we have

$$
\mathbb{E}\big[\|\theta_k^{(\eta)} - \theta^*\|^4 \big]^{1/2} \le \rho^k \|\theta_0 - \theta^*\|^2 + D,
\tag{3.2}
$$

where constants are

$$
\begin{aligned}
D :=& \frac{64}{\alpha} \Big(\bar{L}^4 + L_\xi\big(1 + (\beta/\alpha)^4\big) + 512\bar{L}^6 + 23 L_\xi^{3/2}\big(1 + (\beta/\alpha)^6\big) \Big)^{1/2} \\
& \vee \frac{8}{\alpha} \Big(\beta + (\sqrt{\alpha} + 2L/\sqrt{\alpha})^2 \|\theta^*\| + L\|\theta^*\| + 6\bar{L}^2 + 9 L_\xi^{1/2}\big(1 + (\beta/\alpha)^2\big) + 16 \Big), \\
\rho :=& \sqrt{1 - 2\alpha\eta + 32 L_\dagger \eta^2} \in (0, 1).
\end{aligned}
$$

Consequently, for any test function ϕ that is L_ϕ-Lipschitz continuous, we have

$$
\big| \pi_\eta(\phi) - \phi(\theta^*) \big| \le L_\phi \sqrt{D}.
$$

The above theorem establishes that the SGD algorithm initialized far away from any critical point will converge (in the 4-th expectation) to the ball that contains all the first-order critical points exponentially fast. The first term in the upper bound (3.2) depends on the initialization, but decays to zero exponentially fast with the number of iterations, for a fixed step size. The second term in the bound (3.2) is a constant independent of the iteration number as well as the step size, which serves as the squared radius of the ball that contains all the critical points plus an additional offset to account for the randomness in the SGD iterates. In other words, SGD algorithm initialized at any point and with any sufficiently small step size will find this ball of interest exponentially fast.

3.2 Bias with Local Regularity

In this part, we present algorithmically controllable bounds on the bias under local regularity conditions. Section 3.2.1 provides a direct control on $\mathbb{E}[\|\theta_k^{(\eta)} - \theta^*\|]$ under the assumption that the unique minimizer θ^* exists. In Sections 3.2.2 and 3.2.3, we characterize the degree of sub-optimality $\mathbb{E}[f(\theta_k^{(\eta)})] - f^*$ where f^* is the global minimum which is not necessarily attained at a unique point.

3.2.1 Localized dissipativity condition

We now introduce the generalized dissipativity condition which, in addition to the tail growth enforced in Assumption 2.2, imposes a local growth around the unique critical point θ^*.

Assumption 3.2 (Localized dissipativity). *The objective function f satisfies*

$$\langle \nabla f(\theta), \theta - \theta^* \rangle \geq \begin{cases} \alpha \|\theta - \theta^*\|^2 - \beta & \|\theta - \theta^*\| \geq R \\ g\big(\|\theta - \theta^*\|\big) & \|\theta - \theta^*\| < R, \end{cases}$$

where $\theta^* \in \mathbb{R}^d$ is the unique minimizer of f, $R := \frac{\delta}{\alpha} + \sqrt{\frac{\beta}{\alpha}}$ with $\delta \in (0, \infty)$, $g : [0, \infty) \to [0, \infty)$ is a convex function with $g(0) = 0$ whose inverse exists.

If $g(x) = x^2$, the objective function is *locally* strongly convex. However, the above assumption covers a wide range of objectives with different local growth rates depending on the function g. Next, we show that the above assumption along with the assumptions leading to the CLT is sufficient to establish an algorithmic control over the bias with a sufficiently small step size.

Theorem 3.1. *Let the Assumptions 2.1, 3.1, and 3.2 hold. Then SGD iterates with step size satisfying $\eta < c_{L,\alpha}$ for $c_{L,\alpha}$ in (3.1) admit the stationary distribution $\theta^{(\eta)} \sim \pi_\eta$ which satisfies*

$$\mathbb{E}\big[\|\theta^{(\eta)} - \theta^*\|\big] \leq \frac{C}{\delta}\eta + g^{-1}(C\eta),$$

where

$$C := 2\Big(3L^2 + 3L_\xi^{1/2}(1 + (\beta/\alpha)^2)\Big)\Big(\int \|\theta\|^2 \pi_\eta(d\theta) + \|\theta^*\|^2\Big) + 3L^2\|\theta^*\|^2 + 5L^2 + 2L_\xi^{1/2}\Big(1 + (\beta/\alpha)^2\Big).$$

Further, for a test function $\phi : \mathbb{R}^d \to \mathbb{R}$ that is L_ϕ-Lipschitz, the bias satisfies

$$\big|\pi_\eta(\phi) - \phi(\theta^*)\big| \leq L_\phi\big(C\eta/\delta + g^{-1}(C\eta)\big).$$

If the local growth is linear, i.e. $g(x) = x$, we obtain the bias $|\pi_\eta(\phi) - \phi(\theta^*)| \leq \mathcal{O}(\eta)$. If local growth is quadratic, i.e. $g(x) = x^2$, the growth is *locally* slower than the linear case; thus, we get bias control $|\pi_\eta(\phi) - \phi(\theta^*)| \leq \mathcal{O}(\eta^{1/2})$, which is worse in step size dependency, it reduces to the bound derived in [DDB19, Lemma 10].

We highlight that [DCLZ19] prove the following lower bound: $\liminf_{k \to \infty} \mathbb{E}[\|\theta_k^{(\eta)} - \theta^*\|^2]^{1/2} \geq c\eta^{1/2}$ for some $c > 0$ under the assumption of Lipschitz gradients. This is in line with our findings since Lipschitz gradients imply $g(x) \leq x^2$ for small x.

3.2.2 Generalized Łojasiewicz condition

In this section we work with a generalization of the Łojasiewicz condition.

Assumption 3.3 (Generalized Łojasiewicz condition). *The objective function f has a critical point θ^* and it satisfies*

$$\|\nabla f(\theta)\|^2 \geq \begin{cases} \gamma\{f(\theta) - f(\theta^*)\} & \|\theta - \theta^*\| \geq R, \\ g\big(f(\theta) - f(\theta^*)\big) & \|\theta - \theta^*\| < R, \end{cases}$$

where γ and R are positive constants, and $g : [0, \infty) \to [0, \infty)$ is a convex function with $g(0) = 0$ whose inverse exists.

In the case $g(x) = x^\kappa$ with $\kappa \in [1, 2)$, for example, the above condition is termed as the Łojasiewicz inequality [GLCY16], and for $\kappa = 1$, it reduces to the well-known Polyak-Łojasiewicz (PL) inequality [KNS16]. Note that this inequality implies that every critical point is a global minimizer; yet, it does not imply the existence of a unique critical point.

The following result establishes an algorithmically controllable bias bound in terms of the step size.

Theorem 3.2. *Let the Assumptions 2.1, 2.2, 3.1, and 3.3 hold, and the Hessian satisfies $\|\nabla^2 f(\theta)\| \leq \tilde{L}(1 + \|\theta\|)$, $\forall \theta \in \mathbb{R}^d$ and some $\tilde{L}$. Then, the SGD iterates with a step size satisfying $\eta < \frac{2}{\tilde{L}} \wedge c_{L,\alpha} \wedge c_{L,\alpha}^\dagger \wedge 1$ for $c_{L,\alpha}, c_{L,\alpha}^\dagger$ in (3.1) have the stationary distribution π_η,*

$$\pi_\eta(f) - f(\theta^*) \leq g^{-1}\Big(\frac{2M\eta}{2 - \tilde{L}\eta}\Big) + \frac{2M\eta}{2 - \tilde{L}\eta},$$

where

$$M := 12\tilde{L}\big(L + L_\xi^{1/2} + L_\xi^{1/4}\big)^2\Big(1 + m + m^{3/4} + \int \|\theta\|^2 \pi_\eta(d\theta)\Big) \quad with$$

$$m := \frac{8}{7\alpha}\Big[(\beta + 6L^2 + 3L_\xi^{1/2} + 16)\int \|\theta\|^2 \pi_\eta(d\theta) + 16L^4 + 2L_\xi + 128L^6 + 8L_\xi^{3/2}\Big].$$

Additionally, if the test function is given as $\phi = \tilde{\phi} \circ f$ for a function $\tilde{\phi}$ that is $L_{\tilde{\phi}}$-Lipschitz, it holds that

$$\big|\pi_\eta(\phi) - \phi(\theta^*)\big| \leq L_{\tilde{\phi}}\bigg\{g^{-1}\Big(\frac{2M\eta}{2 - \tilde{L}\eta}\Big) + \frac{2M\eta}{2 - \tilde{L}\eta}\bigg\}.$$

For smooth objectives with Lipschitz gradient, [KNS16] provide a linear rate under the PL-inequality (see also [DYJG17, Lemma 2]), which yields the asymptotic bias $|\pi_\eta(\phi) - \phi(\theta^*)| \leq \mathcal{O}(\eta)$. The above result recovers their findings as a special case, and provides a considerable generalization.

3.2.3 Convexity

To make the analysis of constant step size SGD complete, we digress from the main theme of this paper and consider the constant step size SGD in the non-strongly convex regime, for which there is no bias characterization known to authors. We show that, under the convexity assumption, one can achieve the same bias control as in the case of PL-inequality.

Theorem 3.3. *Let the Assumptions 2.1, 2.2, and 3.1 hold for a convex function f. Then, the SGD iterates with a step size satisfying $\eta < c_{L,\alpha}$ for $c_{L,\alpha}$ in (3.1) admit the stationary distribution π_η, which satisfies*

$$\pi_\eta(f) - f^* \leq C\eta,$$

where

$$C := 2\Big(3L^2 + 3L_\xi^{1/2}(1 + (\beta/\alpha)^2)\Big)\Big(\int \|\theta\|^2 \pi_\eta(d\theta) + \|\theta^*\|^2\Big) + 3L^2\|\theta^*\|^2 + 5L^2 + 2L_\xi^{1/2}\Big(1 + (\beta/\alpha)^2\Big).$$

Additionally, if the test function is given as $\phi = \tilde{\phi} \circ f$ for a function $\tilde{\phi}$ that is $L_{\tilde{\phi}}$-Lipschitz, then,

$$\big|\pi_\eta(\phi) - \phi(\theta^*)\big| \leq L_{\tilde{\phi}}C\eta.$$

9

Convexity implies that any critical point θ^* is a global minimizer, which is similar to the PL-inequality; yet, it does not imply a unique minimizer unlike strong convexity. The resulting step size dependency of the bias is the same as in the case of PL-inequality, which is because both of these conditions assert a similar gradient-based domination criterion on the sub-optimality. That is, we have in the convex case $\langle \nabla f(\theta), \theta - \theta^* \rangle \geq f(\theta) - f(\theta^*)$, and in the case of PL-inequality $\gamma^{-1} \|f(\theta)\|^2 \geq f(\theta) - f(\theta^*)$.

4 Examples and Numerical Studies

We now demonstrate the asymptotic normality and bias in non-convex optimization with two examples arising in robust statistics for which our assumptions can be verified. We consider the online SGD setting with the update rule (1.4) and also the semi-stochastic setting, where the noise sequence $\{\xi_k(\theta)\}_{k \geq 1}$ is independent of θ and is simply a sequence of i.i.d. random vectors – such a setting helps to demonstrate how to verify our assumptions explicitly.

4.1 Regularized MLE for heavy-tailed linear regression

While the least-squares loss function is common in the context of linear regression, it is well-documented that it suffers from robustness issues when the error distribution of the model is heavy-tailed [Hub04]. Indeed in fields like finance, oftentimes the Student's t-distribution is used to model the heavy-tailed error [FY17]. In this case, defining the random vector $Z := (X, Y)$, the stochastic optimization problem in (1.3) is given by the expectation of the function

$$F(Z, \theta) := \log\big(1 + (Y - \langle X, \theta \rangle)^2\big) + \frac{\lambda}{2} \|\theta\|^2,$$

which is non-convex (as a function of θ) for small penalty levels λ. Correspondingly, given n independent and identically distributed samples $(\mathbf{x}_i, y_i)$, the finite-sum version of the optimization problem corresponds to minimizing the following objective function

$$f(\theta) := \frac{1}{2m} \sum_{i=1}^{m} \log\big(1 + (y_i - \langle \mathbf{x}_i, \theta \rangle)^2\big) + \frac{\lambda}{2} \|\theta\|^2. \tag{4.1}$$

We consider the finite-sum setup as we would be able to verify our assumptions and empirically demonstrate the bias result in a clean manner in this setup, as we demonstrate next.

4.1.1 Semi-stochastic Gradient Descent

In the experiments, $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_m)^\top \in \mathbb{R}^{m \times d}$ represents a fixed design matrix generated from $\mathbf{X}_{ij} \sim \text{Bernoulli}(\pm 1)/\sqrt{d}$, and $\mathbf{y} := (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$ represents the response vector generated according to the linear model $y_i = \langle \mathbf{x}_i, \theta_{\text{true}} \rangle + \varepsilon$ with $(\theta_{\text{true}})_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$, and ε is Student-t (df $= 10$) noise. We choose $m = 5000$, $d = 10$, and the Lipshitz test function $\phi(\theta) = \|\theta\|$ unless stated otherwise.

Asymptotic normality: Fig. 1-(a,b,c,d) demonstrates the normality and the bias of SGD with heavy-tailed gradient noise distributed as Student-t (df $= 5$). Each plot has two density curves where red and blue curves in Fig. 1-(a,b) respectively correspond to initializations with $\theta_0 = (1, \ldots, 1)$ and $\theta_0' = (1.5, \ldots, 1.5)$ with step size $\eta = 0.3$; green and orange curves in Fig. 1-c correspond to step sizes $\eta = 0.2$ and $\eta' = 0.3$ with initialization θ_0. All experiments are based on
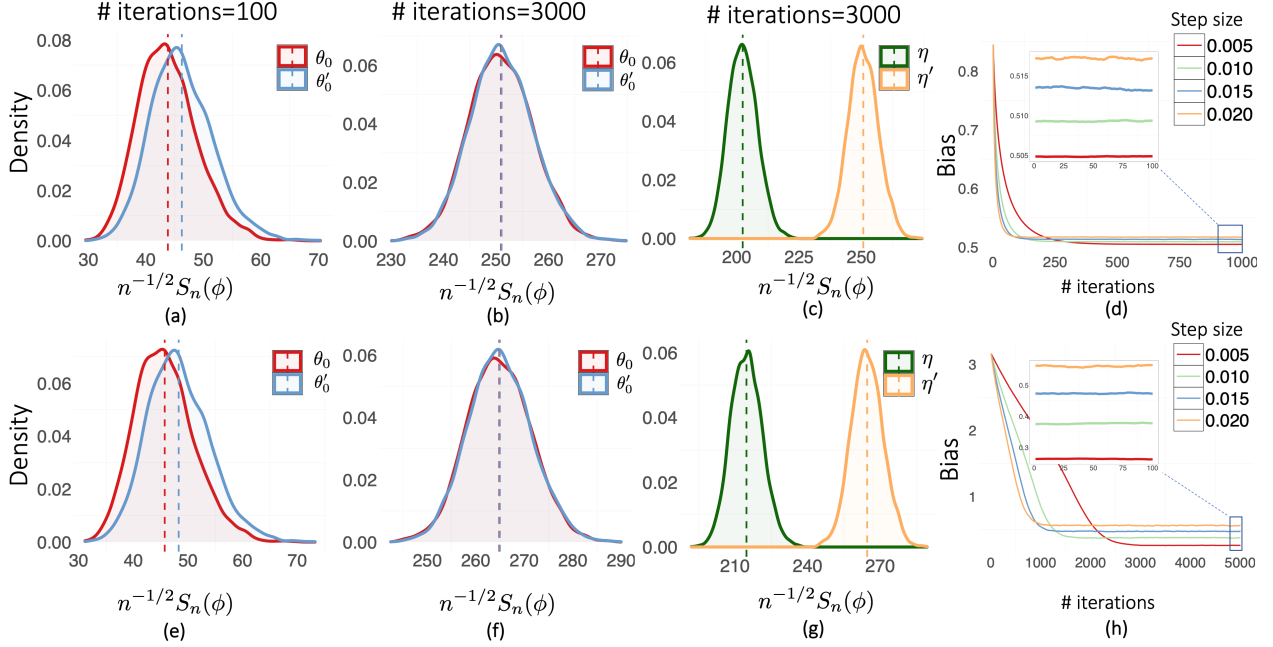
Figure 1: First and second rows correspond to non-convex examples in Sections 4.1.1 and 4.2.1, respectively. Figures (a,b), (e,f) show the density of $n^{-1/2}S_n(\phi) = n^{-1/2}\sum_{k=1}^{n}\phi(\theta_k^{(\eta)})$ with different initializations (red, blue) for different number of iterations. Figures (c,g) show the same density with different step sizes. Figures (d,h) show the evolution of bias against the number of iterations.

4000 Monte Carlo runs. We observe in Fig. 1-a that different initializations have an early impact on the normality when the number of iterations is moderate. However, when SGD is run for a longer time, this effect is removed as in Fig. 1-b. Lastly, Fig.1-c demonstrates the effect of step size on the normality, where the means are different for different step sizes as they depend on the stationary distribution π_η. Indeed, the above results are not surprising. One can verify that the objective function (4.1) satisfies Assumptions 2.1 and 2.2. The above objective has the following gradient

$$\nabla f(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbf{x}_i(\langle \mathbf{x}_i, \theta\rangle - y_i)}{1 + (y_i - \langle \mathbf{x}_i, \theta\rangle)^2} + \lambda\theta.$$

Because $\|\nabla f(\theta)\| \leq \left(\lambda_{\max}(\frac{1}{m}\mathbf{X}^\top\mathbf{X}) + \lambda\right)\|\theta\| + \frac{1}{m}\|\mathbf{X}^\top\mathbf{y}\|$ by the triangle inequality and the fact that the denominator is lower bounded by 1, Assumption 2.1 holds. For Assumption 2.2, we write

$$\langle\nabla f(\theta), \theta\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{(\langle\mathbf{x}_i, \theta\rangle)^2 - y_i\langle\mathbf{x}_i, \theta\rangle}{1 + (y_i - \langle\mathbf{x}_i, \theta\rangle)^2} + \lambda\|\theta\|^2 \geq -\left\|\frac{1}{m}\mathbf{X}^\top\mathbf{y}\right\|\|\theta\| + \lambda\|\theta\|^2,$$

by Cauchy-Schwartz inequality. Next, using Young's inequality $-\left\|\frac{1}{m}\mathbf{X}^\top\mathbf{y}\right\|\|\theta\| \geq -\frac{1}{\lambda}\left\|\frac{1}{m}\mathbf{X}^\top\mathbf{y}\right\|^2 - \frac{\lambda}{4}\|\theta\|^2$, Assumption 2.2 holds for $\alpha = \lambda/4$ and $\beta = \frac{1}{\lambda}\left\|\frac{1}{m}\mathbf{X}^\top\mathbf{y}\right\|^2$. Finally, the gradient noise has finite 4-th moment with support on $\mathbb{R}^d$; thus, Assumption 2.3 is satisfied, and Theorem 2.1 is applicable.

Bias: In order to demonstrate the bias behavior without speculation, one needs the global minimum θ^* of the non-convex problem. Therefore, we simplify the problem (4.1) to another non-convex
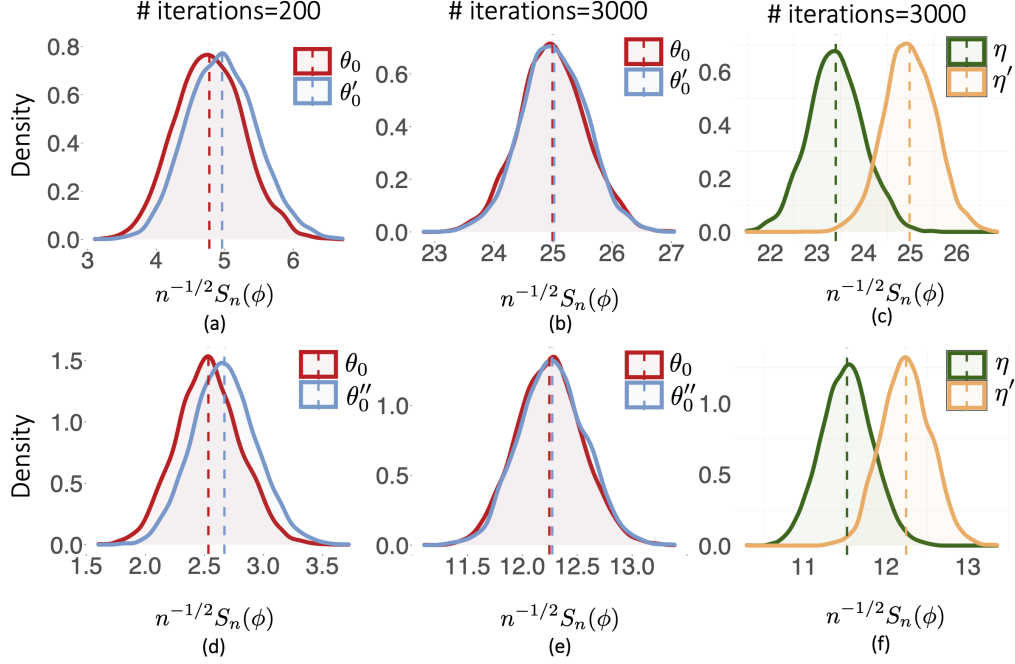
11

Figure 2: First and second rows correspond to non-convex examples in Sections 4.1.2 and 4.2.2, respectively. Figures (a,b), (e,f) show the density of $n^{-1/2}S_n(\phi) = n^{-1/2}\sum_{k=1}^{n}\phi(\theta_k^{(\eta)})$ with different initializations (red, blue) for different number of iterations. Figures (c,g) show the same density with different step sizes.

problem

$$f(\theta) := \frac{1}{2}\log\left(1 + \|\theta\|^2\right) + \frac{\lambda}{2}\|\theta\|^2 \,.$$

Notice that the general structure is the same, with no data, and θ^* is known, i.e. $\theta^* = 0$.

We choose the test function $\phi(\theta) = \tilde{\phi} \circ f(\theta)$, where $\tilde{\phi}(x) = 1/(1 + e^{-x})$ is Lipschitz. Fig. 1-(d) demonstrates how the bias $\pi_\eta(\phi) - \phi(\theta^*)$ changes over iterations, where different curves correspond to different step sizes. We notice that larger step size provides fast initial decrease; yet the resulting asymptotic bias is larger which aligns with our theory. To verify assumptions, we compute the gradient and the Hessian respectively as

$$\nabla f(\theta) = \frac{\theta}{1 + \|\theta\|^2} + \lambda\theta, \qquad \text{and} \qquad \nabla^2 f(\theta) = \frac{I}{1 + \|\theta\|^2} - \frac{2\theta\theta^\top}{(1 + \|\theta\|^2)^2} + I\lambda,$$

with I denoting the identity matrix. For small λ the above function is clearly non-convex. To see this, choose $\lambda = 0.1, u = \theta/\|\theta\|$ and note that $\langle u, \nabla^2 f(\theta)u \rangle < 0$ whenever $1.5 \leq \|\theta\| \leq 2$. Also, note that

$$\|\nabla f(\theta)\|^2 = \|\theta\|^2\left(\lambda + 1/(1 + \|\theta\|^2)\right)^2 \geq \frac{2\lambda^2}{1 + \lambda}\left\{f(\theta) - f(\theta^*)\right\}.$$

Thus, Assumption 3.3 is satisfied for $\gamma = \frac{2\lambda^2}{1+\lambda}$ and $g(x) = \gamma x^2$. Following the same steps in the regression setting, one can also verify Assumptions 2.1-2.3; hence, Theorem 3.2 can be applied.

4.1.2 Online Stochastic Gradient Descent

For our online SGD experiments, we use $b_k = 2$, for all iterations k to obtain the stochastic gradient. We also experimented with $m_k = 10, 50$ and observed similar behavior. The distribution of the random vector $Z = (X, Y) \in \mathbb{R}^{d+1}$, is as follows: Each coordinate of the vector $X \in \mathbb{R}^d$, is generated as Bernoulli$(\pm 1)/\sqrt{d}$ and given vector X, the response $Y \in \mathbb{R}$ is generated according to the linear model $Y = \langle X, \theta_{\text{true}} \rangle + \varepsilon$ with each coordinate of $\theta_{\text{true}} \in \mathbb{R}^d$ generated from Unif$(0, 1)$, and fixed, and $\varepsilon \in \mathbb{R}$ is Student-t (df $= 10$) noise. We choose $d = 10$, and set a burn-in period of size 100.

Asymptotic normality: Fig. 2-(a,b,c) demonstrates the normality of online SGD. Each plot has two density curves where red and blue curves in Fig. 2-(a,b) respectively correspond to initializations with $\theta_0 = (1, \ldots, 1)$ and $\theta_0' = (2.5, \ldots, 2.5)$ with step size $\eta = 0.3$; green and orange curves in Fig. 2-c correspond to step sizes $\eta = 0.2$ and $\eta' = 0.3$ with initialization θ_0. All experiments are based on 4000 Monte Carlo runs. We observe in Fig. 2-a that different initializations have an early impact on the normality when the number of iterations is moderate. However, when SGD is run for a longer time, this effect is removed as in Fig. 2-b. Lastly, Fig.2-c demonstrates the effect of step size on the normality, where the means are different for different step sizes as they depend on the stationary distribution π_η.

4.2 Regularized Blake-Zisserman MLE for corrupted linear regression

While the above example was based on linear-regression with heavy-tailed noise, we now consider the case of heavy-tailed regression with corrupted noise. In this setup, the noise model in linear regression is assumed to be Gaussian, but a fraction of the noise vectors are assumed to be corrupted in the sense that they are drawn from a uniform distribution. Such a scenario arises in visual reconstruction problems; see for example [BZ87] for details. In this case, defining the random vector $Z := (X, Y)$, the stochastic optimization problem in (1.3) is given by the expectation of the function

$$F(Z, \theta) := \log\left(\nu + e^{-(Y - \langle X, \theta \rangle)^2}\right) + \frac{\lambda}{2}\|\theta\|^2, \quad \nu > 0.$$

Similar the previous case, we also consider the finite-sum version: Given n independent and identically distributed samples $(\mathbf{x}_i, y_i)$, it corresponds to minimizing the following objective function

$$f(\theta) = -\frac{1}{2m}\sum_{i=1}^m \log\left(\nu + e^{-(y_i - \langle \mathbf{x}_i, \theta \rangle)^2}\right) + \frac{\lambda}{2}\|\theta\|^2, \quad \nu > 0.$$

4.2.1 Semi-stochastic Gradient Descent

In the experiments, we use the same setup and parameters as in Section 4.1.1.

Asymptotic normality: Fig 1-(e,f,g) demonstrates the asymptotic normality of the SGD with heavy-tailed gradient noise Student-t(df $= 6$). The experimental setup is the same as the previous example with the same values for $\theta_0, \theta_0', \eta, \eta'$. We observe the early impact of initialization in Fig 1-a, the clear normality in Fig. 1-b, and the effect of step size on CLT in Fig.1-c. These observations also align with our theory since this objective also satisfies our assumptions. Indeed, it has the gradient

$$\nabla f(\theta) = -\frac{1}{m}\sum_{i=1}^m \frac{\mathbf{x}_i\left(y_i - \langle \mathbf{x}_i, \theta \rangle\right)e^{-(y_i - \langle \mathbf{x}_i, \theta \rangle)^2}}{\nu + e^{-(y_i - \langle \mathbf{x}_i, \theta \rangle)^2}} + \lambda\theta.$$

The triangle inequality yields

$$\|\nabla f(\theta)\| \leq \frac{1}{1+\nu}\|\tfrac{1}{m}\mathbf{X}^\top\mathbf{y}\| + \left(\frac{1}{1+\nu}\lambda_{\max}(\tfrac{1}{m}\mathbf{X}^\top\mathbf{X}) + \lambda\right)\|\theta\|,$$

which verifies Assumption 2.1. To verify the dissipativity assumption, we can write

$$\langle\nabla f(\theta), \theta\rangle = \langle-\frac{1}{m}\sum_{i=1}^m \frac{\mathbf{x}_i\big(y_i - \langle\mathbf{x}_i, \theta\rangle\big)e^{-(y_i-\langle\mathbf{x}_i,\theta\rangle)^2}}{\nu + e^{-(y_i-\langle\mathbf{x}_i,\theta\rangle)^2}} + \lambda\theta, \theta\rangle \geq -\frac{1}{1+\nu}\|\tfrac{1}{m}\mathbf{X}^\top\mathbf{y}\|\|\theta\| + \lambda\|\theta\|^2.$$

The inequality follows from the triangle and Cauchy-Schwartz inequalities. Using Young's inequality, we obtain

$$-\frac{1}{1+\nu}\|\tfrac{1}{m}\mathbf{X}^\top\mathbf{y}\|\|\theta\| \geq -\frac{1}{\lambda(1+\nu)}\|\tfrac{1}{m}\mathbf{X}^\top\mathbf{y}\|^2 - \frac{\lambda}{4(1+\nu)}\|\theta\|^2,$$

which shows that the above function is dissipative for $\alpha = \lambda/2$ and $\beta = \frac{1}{2\lambda(1+\nu)^2}\|\tfrac{1}{m}\mathbf{X}^\top\mathbf{y}\|^2$; thus, Assumption 2.2 holds.

Bias: Similar to the previous example, we simplify the problem so that we can compute the bias $\pi_\eta(\phi) - \phi(\theta^*)$. We consider the function

$$f(\theta) := -\frac{1}{2}\log\big(\nu + e^{-\|\theta\|^2}\big) + \frac{\lambda}{2}\|\theta\|^2, \quad \nu > 0.$$

We observe in Fig.1-h that smaller step sizes lead to smaller asymptotic bias. To verify that this can be predicted from our theory, we write the gradient and the Hessian respectively, as

$$\nabla f(\theta) = \frac{\theta}{1 + \nu e^{\|\theta\|^2}} + \lambda\theta \quad \text{and} \quad \nabla^2 f(\theta) = \frac{I}{1 + \nu e^{\|\theta\|^2}} - \frac{2\nu e^{\|\theta\|^2}}{(1 + \nu e^{\|\theta\|^2})^2}\theta\theta^\top + \lambda I.$$

First, note that the Hessian can have negative eigenvalues for small values of λ. For example, for $\nu = 1$, $\lambda = 0.1$, and the unit direction $u = \theta/\|\theta\|$, we have $\langle u, \nabla^2 f(\theta)u\rangle < 0$ for $1 \leq \|\theta\|^2 \leq 2$; thus the function is non-convex. But we also have

$$\langle\nabla f(\theta), \theta\rangle = \|\theta\|^2\big(\lambda + 1/\big(1 + \nu e^{\|\theta\|^2}\big)\big) \geq \big(\lambda + 1/\big(1 + \nu e^{R^2}\big)\big)\|\theta\|^2$$

for $\|\theta\| \leq R$ and $\langle\nabla f(\theta), \theta\rangle \geq \lambda\|\theta\|^2$ for $\|\theta\|^2 > R$; thus, Assumption 3.2 is satisfied for $\alpha = \lambda$, and any $\beta \geq 0$ and $g(x) = \big(\lambda + 1/\big(1 + \nu e^{R^2}\big)\big)x^2$. Following the same steps in the previous example, one can also verify Assumptions 2.1-2.3; therefore, Theorem 3.1 follows.

4.2.2 Online Stochastic Gradient Descent

In the experiments, we use the same setup as in Section 4.1.2.

Asymptotic normality: Fig. 2-(d,e,f) demonstrates the normality of online SGD. Each plot has two density curves where red and blue curves in Fig. 2-(d,e) respectively correspond to initializations with $\theta_0 = (1, \ldots, 1)$ and $\theta_0'' = (1.5, \ldots, 1.5)$ with step size $\eta = 0.3$; green and orange curves in Fig. 2-c correspond to step sizes $\eta = 0.2$ and $\eta' = 0.3$ with initialization θ_0. All experiments are based on 4000 Monte Carlo runs. We observe in Fig. 2-d that different initializations have an early impact on the normality when the number of iterations are moderate. However, when SGD is run for a longer time, this effect is removed as in Fig. 2-e. Lastly, Fig.2-f demonstrates the effect of step size on the normality, where the means are different for different step sizes as they depend on the stationary distribution π_η.

5 Discussions

By leveraging the connection between constant step size SGD and Markov chains [DDB19], we provided theoretical results characterizing the bias and the fluctuations of constant step size SGD for non-convex and non-smooth optimization which arises frequently in modern statistical learning.

Estimating the Asymptotic Variance: As discussed in Section 2, in order for using the established CLT to compute confidence intervals in practice, the population expectation $\pi_\eta(\phi)$ and asymptotic variance $\sigma^2_{\pi_\eta}(\phi)$ have to be estimated. We suggest the following three ways to do so:

- Estimate them based on sample average of a single trajectory of SGD iterates, i.e., the mean $\pi_\eta(\phi)$ is estimated as $n^{-1}\sum_{k=0}^{n-1}\phi(\theta_k^{(\eta)})$ and the variance $\sigma^2_{\pi_\eta}(\phi)$ by adopting the online approach of [ZCW20] to the constant step size setting.

- First run N parallel SGD trajectories and compute the average of each trajectory, to obtain N independent observations from the stationary distribution π_η. Next, use the N observations to compute the sample mean and the sample variance estimators for $\pi_\eta(\phi)$ and $\sigma^2_{\pi_\eta}(\phi)$.

- Leverage the online bootstrap and variance estimation approaches proposed in [FXY18, SZ18, CLT$^+$20] for the constant step size SGD setting in order to obtain estimates for $\pi_\eta(\phi)$ and $\sigma^2_{\pi_\eta}(\phi)$.

A theoretical investigation on the relative merits of the above approaches is left as future work.

Acknowledgements

References

[ABE19]   Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu, *Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt*, Conference on Learning Theory, 2019, pp. 115–137.

[AMP00]   Rafik Aguech, Eric Moulines, and Pierre Priouret, *On a perturbation approach for the analysis of stochastic tracking algorithms*, SIAM Journal on Control and Optimization **39** (2000), no. 3, 872–899.

[BDMP17]   Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra, *Sampling from a log-concave distribution with compact support with proximal langevin monte carlo*, COLT, 2017.

[BEL18]   Sébastien Bubeck, Ronen Eldan, and Joseph Lehec, *Sampling from a log-concave distribution with projected langevin monte carlo*, Discrete & Computational Geometry (2018).

[Ben96]   Michel Benaim, *A dynamical system approach to stochastic approximations*, SIAM Journal on Control and Optimization **34** (1996), no. 2, 437–472.

[BH18]   Lukas Balles and Philipp Hennig, *Dissecting adam: The sign, magnitude and variance of stochastic gradients*, International Conference on Machine Learning, 2018, pp. 404–413.

[BVB16]   Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira, *The non-convex burer-monteiro approach works on smooth semidefinite programs*, Advances in Neural Information Processing Systems, 2016, pp. 2757–2765.

[BZ87]   Andrew Blake and Andrew Zisserman, *Visual reconstruction*, 1987.

[CCAY+18]   Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan, *Sharp convergence rates for langevin dynamics in the nonconvex setting*, arXiv preprint arXiv:1805.01648 (2018).

[CCBJ17]   Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan, *Underdamped langevin mcmc: A non-asymptotic analysis*, arXiv preprint arXiv:1707.03663 (2017).

[Chu54]   Kai Lai Chung, *On a stochastic approximation method*, The Annals of Mathematical Statistics (1954), 463–483.

[CLC19]   Yuejie Chi, Yue M Lu, and Yuxin Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Transactions on Signal Processing **67** (2019), no. 20, 5239–5269.

[CLPC19]   Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen, *Nonconvex low-rank tensor completion from noisy data*, Advances in Neural Information Processing Systems, 2019, pp. 1861–1872.

[CLT+20]   Xi Chen, Jason D Lee, Xin T Tong, Yichen Zhang, et al., *Statistical inference for model parameters in stochastic gradient descent*, The Annals of Statistics **48** (2020), no. 1, 251–273.

[CT18]   Jerry Chee and Panos Toulis, *Convergence diagnostics for stochastic gradient descent with constant learning rate*, International Conference on Artificial Intelligence and Statistics, 2018, pp. 1476–1485.

[Dal17]   Arnak S Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79** (2017), no. 3, 651–676.

[DCLZ19]   Zhiyan Ding, Yiding Chen, Qin Li, and Xiaojin Zhu, *Error lower bounds of constant step-size stochastic gradient descent*, arXiv preprint arXiv:1910.08212 (2019).

[DCWY18]   Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu, *Log-concave sampling: Metropolis-hastings algorithms are fast!*, arXiv preprint arXiv:1801.02309 (2018).

[DDB19]   Aymeric Dieuleveut, Alain Durmus, and Francis Bach, *Bridging the gap between constant step size stochastic gradient descent and markov chains*, The Annals of Statistics (to appear) (2019).

[DK17]   Arnak S Dalalyan and Avetik G Karagulyan, *User-friendly guarantees for the langevin monte carlo with inaccurate gradient*, arXiv preprint arXiv:1710.00095 (2017).

[DM17]   Alain Durmus and Eric Moulines, *Nonasymptotic convergence analysis for the unadjusted langevin algorithm*, The Annals of Applied Probability **27** (2017), no. 3, 1551–1587.

[DMPS18]   Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier, *Markov chains*, Springer, 2018.

[DR18]   John Duchi and Feng Ruan, *Asymptotic optimality in stochastic optimization*, Arxiv Preprint (2018).

[DRD18]   Arnak S Dalalyan and Lionel Riou-Durand, *On sampling from a log-concave density using kinetic langevin diffusions*, arXiv preprint arXiv:1807.09382 (2018).

[DYJG17]   Soham De, Abhay Yadav, David Jacobs, and Tom Goldstein, *Automated inference with adaptive batches*, Artificial Intelligence and Statistics, 2017, pp. 1504–1513.

[EH20]   Murat A Erdogdu and Rasa Hosseinzadeh, *On the convergence of langevin monte carlo: The interplay between tail growth and smoothness*, arXiv preprint arXiv:2005.13097 (2020).

[EMS18]   Murat A Erdogdu, Lester Mackey, and Ohad Shamir, *Global non-convex optimization with discretized diffusions*, Advances in Neural Information Processing Systems, 2018, pp. 9671–9680.

[EvdG18]   Andreas Elsener and Sara van de Geer, *Sharp oracle inequalities for stationary points of nonconvex penalized m-estimators*, IEEE Transactions on Information Theory **65** (2018), no. 3, 1452–1472.

[Fab68]   Vaclav Fabian, *On asymptotic normality in stochastic approximation*, The Annals of Mathematical Statistics **39** (1968), no. 4, 1327–1332.

[FD82]   DA Freedman and P Diaconis, *On inconsistent m-estimators*, The Annals of Statistics **10** (1982), no. 2, 454–461.

[FLLZ18]   Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang, *Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator*, Advances in Neural Information Processing Systems, 2018, pp. 689–699.

[FP99]   Jean-Claude Fort and Gilles Pages, *Asymptotic behavior of a markovian stochastic algorithm with constant step*, SIAM journal on control and optimization **37** (1999), no. 5, 1456–1482.

[FXY18]   Yixin Fang, Jinfeng Xu, and Lei Yang, *Online bootstrap confidence intervals for the stochastic gradient descent estimator*, The Journal of Machine Learning Research **19** (2018), no. 1, 3053–3073.

[FY17]   Jianqing Fan and Qiwei Yao, *The elements of financial econometrics*, Cambridge University Press, 2017.

[GBC16]   Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.

[Gey94]   Charles J Geyer, *On the asymptotics of constrained m-estimation*, The Annals of Statistics **22** (1994), no. 4, 1993–2010.

[GHJY15]   Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points - online stochastic gradient for tensor decomposition*, Conference on Learning Theory, 2015, pp. 797–842.

[GL13]   Saeed Ghadimi and Guanghui Lan, *Stochastic first-and zeroth-order methods for non-convex stochastic programming*, SIAM Journal on Optimization **23** (2013), no. 4, 2341–2368.

[GLCY16]   Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan, *On the {\L} ojasiewicz exponent of the quadratic sphere constrained optimization problem*, arXiv preprint arXiv:1611.08781 (2016).

[GLM16]   Rong Ge, Jason D Lee, and Tengyu Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, 2016, pp. 2973–2981.

[Hub04]   Peter J Huber, *Robust statistics*, vol. 523, John Wiley & Sons, 2004.

[JGN$^+$17]   Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan, *How to escape saddle points efficiently*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1724–1732.

[JK17]   Prateek Jain and Purushottam Kar, *Non-convex optimization for machine learning*, Foundations and Trends® in Machine Learning **10** (2017), no. 3-4, 142–336.

[Kaw16]   Kenji Kawaguchi, *Deep learning without poor local minima*, Advances in neural information processing systems, 2016, pp. 586–594.

[Kif88]   Yuri Kifer, *Random perturbations of dynamical systems*, Nonlinear Problems in Future Particle Accelerators. World Scientific (1988), 189.

[KNS16]   Hamed Karimi, Julie Nutini, and Mark Schmidt, *Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 795–811.

[LLM18]   Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu, *Robust classification via mom minimization*, arXiv preprint arXiv:1808.03106 (2018).

[Loh17]   Po-Ling Loh, *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, The Annals of Statistics **45** (2017), no. 2, 866–896.

[LWME19] Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu, *Stochastic runge-kutta accelerates langevin monte carlo and beyond*, Advances in Neural Information Processing Systems, 2019, pp. 7748–7760.

[MB11]   Eric Moulines and Francis R Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in Neural Information Processing Systems, 2011, pp. 451–459.

[MBB18]   Siyuan Ma, Raef Bassily, and Mikhail Belkin, *The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning*, International Conference on Machine Learning, 2018, pp. 3325–3334.

[MBM18]   Song Mei, Yu Bai, and Andrea Montanari, *The landscape of empirical risk for nonconvex losses*, The Annals of Statistics **46** (2018), no. 6A, 2747–2774.

[MIG$^+$19] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson, *A simple baseline for bayesian uncertainty in deep learning*, Advances in Neural Information Processing Systems, 2019, pp. 13132–13143.

[MMMO17] Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira, *Solving sdps for synchronization and maxcut problems via the grothendieck inequality*, Conference on Learning Theory, 2017, pp. 1476–1515.

[MSH02]   Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham, *Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise*, Stochastic processes and their applications **101** (2002), no. 2, 185–232.

[MT12]   Sean P Meyn and Richard L Tweedie, *Markov chains and stochastic stability*, Springer Science & Business Media, 2012.

[NP06]   Yurii Nesterov and Boris T Polyak, *Cubic regularization of newton method and its global performance*, Mathematical Programming **108** (2006), no. 1, 177–205.

[NWS14]   Deanna Needell, Rachel Ward, and Nati Srebro, *Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm*, Advances in neural information processing systems, 2014, pp. 1017–1025.

[PJ92]   Boris T Polyak and Anatoli B Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization **30** (1992), no. 4, 838–855.

[PV98]   P Priouret and A Yu Veretenikov, *A remark on the stability of the lms tracking algorithm*, Stochastic analysis and applications **16** (1998), no. 1, 119–129.

[QCLP19]   Zhengling Qi, Ying Cui, Yufeng Liu, and Jong-Shi Pang, *Statistical analysis of stationary solutions of coupled nonconvex nonsmooth empirical risk minimization*, arXiv preprint arXiv:1910.02488 (2019).

[RM51]   Herbert Robbins and Sutton Monro, *A stochastic approximation method*, The annals of mathematical statistics (1951), 400–407.

[RRT17]   Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky, *Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis*, Conference on Learning Theory, 2017, pp. 1674–1703.

[Rup88]   David Ruppert, *Efficient estimations from a slowly convergent robbins-monro process*, Tech. report, Cornell University Operations Research and Industrial Engineering, 1988.

[Sac58]   Jerome Sacks, *Asymptotic distribution of stochastic approximation procedures*, The Annals of Mathematical Statistics **29** (1958), no. 2, 373–405.

[Sha89]   Alexander Shapiro, *Asymptotic properties of statistical estimators in stochastic programming*, The Annals of Statistics **17** (1989), no. 2, 841–858.

[Sha00] \rule{2cm}{0.4pt}, *On the asymptotics of constrained local m-estimators*, Annals of statistics (2000), 948–960.

[SL19] Ruoqi Shen and Yin Tat Lee, *The randomized midpoint method for log-concave sampling*, Advances in Neural Information Processing Systems, 2019, pp. 2098–2109.

[SR13] Mark Schmidt and Nicolas Le Roux, *Fast convergence of stochastic gradient descent under a strong growth condition*, arXiv preprint arXiv:1308.6370 (2013).

[SZ18] Weijie Su and Yuancheng Zhu, *Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent*, arXiv preprint arXiv:1802.04876 (2018).

[TA17] Panos Toulis and Edoardo M Airoldi, *Asymptotic and finite-sample properties of estimators based on stochastic gradients*, The Annals of Statistics **45** (2017), no. 4, 1694–1727.

[TFBJ18] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan, *Averaging stochastic gradient descent on riemannian manifolds*, arXiv preprint arXiv:1802.09128 (2018).

[TSJ+18] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan, *Stochastic cubic regularization for fast nonconvex optimization*, Advances in neural information processing systems, 2018, pp. 2899–2908.

[TV19] Yan Shuo Tan and Roman Vershynin, *Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval*, arXiv preprint arXiv:1910.12837 (2019).

[VBS19] Sharan Vaswani, Francis Bach, and Mark Schmidt, *Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron*, The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1195–1204.

[WRS+17] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht, *The marginal value of adaptive gradient methods in machine learning*, Advances in Neural Information Processing Systems, 2017, pp. 4148–4158.

[XCZG18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu, *Global convergence of langevin dynamics based algorithms for nonconvex optimization*, Advances in Neural Information Processing Systems, 2018, pp. 3122–3133.

[XYZ19] Dong Xia, Ming Yuan, and Cun-Hui Zhang, *Statistically optimal and computationally efficient low rank tensor completion from noisy entries*, The Annals of Statistics (to appear) (2019).

[ZCW20] Wanrong Zhu, Xi Chen, and Wei Biao Wu, *A fully online approach for covariance matrices estimation of stochastic gradient descent solutions*, arXiv preprint arXiv:2002.03979 (2020).

A Proofs for Sections 2 and 3

A.1 Preliminaries and Additional Notations

Note that the sequence of iterates $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is a homogeneous Markov chain [DDB19]. We denote the (sub-)σ-algebra (of $\mathcal{F}$) of events up to and including the k-th iteration as $\mathcal{F}_k$. By definition, the discrete-time stochastic process defined in (1.2) is adapted to its natural filtration $\{\mathcal{F}_k\}_{k\geq 0}$. We denote the Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ associated with SGD iterates (1.2) by P with

$$P(\theta_k^{(\eta)}, A) = \mathbb{P}(\theta_{k+1}^{(\eta)} \in A | \theta_k^{(\eta)}) \ \ \mathbb{P} - a.s., \quad \forall A \in \mathcal{B}(\mathbb{R}^d), k \geq 0\,.$$

Define the k-th power of this kernel iteratively: define $P^1 := P$, and for $k \geq 1$, for all $\tilde{\theta} \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$, define

$$P^{k+1}(\tilde{\theta}, A) := \int_{\mathbb{R}^d} P(\tilde{\theta}, d\theta) P^k(\theta, A)\,.$$

For any function $\phi : \mathbb{R}^d \to \mathbb{R}$ and $k \geq 0$, define the measurable function $P^k\phi(\theta) : \mathbb{R}^d \to \mathbb{R}$ for all $\theta \in \mathbb{R}^d$ via

$$P^k\phi(\theta) = \int \phi(\tilde{\theta}) P^k(\theta, d\tilde{\theta})\,.$$

Given the L_ϕ-Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$ and the expectation of ϕ under the stationary measure π_η, define the function h as

$$h : \mathbb{R}^d \to \mathbb{R}$$
$$\theta \mapsto \phi(\theta) - \pi_\eta(\phi)\,.$$

Note that $\pi_\eta(h) = 0$ and h is L_ϕ-Lipschitz. Define the partial sum $S_n(\phi) := \sum_{k=0}^{n-1} h(\theta_k^{(\eta)})$. Moreover, we define

$$\bar{\theta}_\eta := \int_{\mathbb{R}^d} \theta d\pi_\eta(\theta)\,.$$

A.2 Proofs of Proposition 2.1 and Theorem 2.1

We start with some preliminary results required to prove the CLT.

Lemma A.1. *Under Assumptions 2.1-2.3, it holds for any $\eta \in \left(0, \frac{\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi))\vee 0}}{3L^2 + L_\xi}\right)$ and any fixed initial point $\theta_0^{(\eta)} = \theta_0 \in \mathbb{R}^d$ that*

$$\mathbb{E}[\,\|\theta_{k+1}^{(\eta)}\|^2 + 1|\mathcal{F}_k] \leq \alpha_\dagger(\,\|\theta_k^{(\eta)}\|^2 + 1) + \beta_\dagger\,.$$

Here, $\alpha_\dagger \in (0,1)$ and $\beta_\dagger \in (0,\infty)$ are constants depending on η. The explicit formulas of $\alpha_\dagger, \beta_\dagger$ are given in the proof.

Proof of Lemma A.1. Define $U_\eta := \frac{\alpha - \sqrt{\max\{\alpha^2 - (3L^2 + L_\xi), 0\}}}{3L^2 + L_\xi}$. Given $\eta \in (0, U_\eta)$, define

$$\alpha_\dagger = 1 + \eta^2(3L^2 + L_\xi) - 2\eta\alpha$$

and note that with this definition $\alpha_\dagger \in (0,1)$ whenever $\eta \in (0, U_\eta)$. Then, with $\eta, \alpha_\dagger$, and the fixed initial point $\theta_0^{(\eta)} = \theta_0 \in \mathbb{R}^d$, we set

$$\beta_\dagger := \kappa(\alpha_\dagger^{1/2} - \alpha_\dagger)\,,$$

where

$$\kappa := \frac{4\eta(\alpha + \beta) + 12\eta^2 L^2}{\alpha_\dagger^{1/2} - \alpha_\dagger} \bigvee 1 \,.$$

It follows that $\beta_\dagger > 0$. Note that

$$\mathbb{E}[1 + \|\theta_{k+1}^{(\eta)}\|^2 | \mathcal{F}_k]$$
$$= \mathbb{E}[1 + \|\theta_k^{(\eta)} - \eta(\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}))\|^2 | \mathcal{F}_k]$$
$$= 1 + \mathbb{E}\big[\|\theta_k^{(\eta)}\|^2 + \eta^2 \|\nabla f(\theta_k^{(\eta)})\|^2 + \eta^2 \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 - 2\eta\langle\theta_k^{(\eta)}, \nabla f(\theta_k^{(\eta)})\rangle | \mathcal{F}_k\big] \,.$$

The last step follows from the Assumption 2.3. By Assumption 2.1, we have

$$\|\nabla f(\theta_k^{(\eta)})\|^2 \le L^2 (1 + \|\theta_k^{(\eta)}\|)^2 \,.$$

Squaring both sides and using the fact that $(1 + \|\theta_k^{(\eta)}\|)^2 \le 3(\|\theta_k^{(\eta)}\|^2 + 3)$ gives

$$\|\nabla f(\theta_k^{(\eta)})\|^2 \le 3L^2 (\|\theta_k^{(\eta)}\|^2 + 3) \,.$$

By Assumption 2.2, we obtain

$$\langle\theta_k^{(\eta)}, \nabla f(\theta_k^{(\eta)})\rangle \ge \alpha \|\theta_k^{(\eta)}\|^2 - \beta \,.$$

By Assumption 2.3, it holds that

$$\mathbb{E}[\|\xi_{k+1}(\theta_k^{(\eta)})\|^2 | \mathcal{F}_k] \le L_\xi (1 + \|\theta_k^{(\eta)}\|^2) \,.$$

Plugging the previous three inequalities into the first display provides us with

$$\mathbb{E}[1 + \|\theta_{k+1}^{(\eta)}\|^2 | \mathcal{F}_k] \le 1 + 9\eta^2 L^2 + \eta^2 L_\xi + 2\eta\beta + (1 + 3\eta^2 L^2 + \eta^2 L_\xi - 2\eta\alpha)\|\theta_k^{(\eta)}\|^2 \,. \tag{A.1}$$

Recall that $\alpha_\dagger = 1 + \eta^2(3L^2 + L_\xi) - 2\eta\alpha$. Plugging $\alpha_\dagger$ back into the previous display yields

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^2 + 1 | \mathcal{F}_k] \le \alpha_\dagger(\|\theta_k^{(\eta)}\|^2 + 1) + 2\eta(\alpha + \beta) + 6\eta^2 L^2 \,.$$

Note that $\beta_\dagger = \kappa(\alpha_\dagger^{1/2} - \alpha_\dagger)$, where

$$\kappa \ge \frac{4\eta(\alpha + \beta) + 12\eta^2 L^2}{\alpha_\dagger^{1/2} - \alpha_\dagger} \,.$$

It then follows that $\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^2 + 1 | \mathcal{F}_k] \le \alpha_\dagger(\|\theta_k^{(\eta)}\|^2 + 1) + \beta_\dagger$ as desired. $\square$

Corollary A.1 (Bounded second moment). *Under the assumptions stated in Lemma A.1, with the constant step size* $\eta \in \left(0, \frac{\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi)) \vee 0}}{3L^2 + L_\xi}\right)$ *the stationary distribution* π_η *satisfies*

$$\mu_{2,\eta} := \int \|\theta\|^2 \pi_\eta(d\theta) \le 3 + \frac{2\beta}{\alpha} \,.$$

Proof of Corollary A.1. Consider the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ starting from the stationary distribution π_η. By display (A.1), it holds that

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^2] \leq 9\eta^2 L^2 + \eta^2 L_\xi + 2\eta\beta + (1 + 3\eta^2 L^2 + \eta^2 L_\xi - 2\eta\alpha)\|\theta_k^{(\eta)}\|^2\,.$$

Using the fact that by stationarity $\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^2] = \mathbb{E}[\|\theta_k^{(\eta)}\|^2]$ and rearranging the previous display gives

$$\mathbb{E}[\|\theta_k^{(\eta)}\|^2] \leq \frac{9\eta L^2 + \eta L_\xi + 2\beta}{2\alpha - \eta(3L^2 + L_\xi)} \leq 3 + \frac{2\beta}{\alpha}\,.$$

$\square$

Corollary A.2 (Lyapunov condition)**.** *Under the assumptions stated in Lemma A.1, given the step size specified in Lemma A.1, it holds that*

$$\mathbb{E}[V(\theta_{k+1}^{(\eta)})|\mathcal{F}_k] \leq \alpha_\dagger V(\theta_k^{(\eta)}) + \beta_\dagger\,,$$

where the Lyapunov function $V(\theta)$ is defined via

$$V(\theta) := \|\theta\|^2 + 1\,. \tag{A.2}$$

Observe that by the proof of Lemma 15.2.8 in [MT12] this also implies that the drift condition (V4) in [MT12] holds with V defined above, $b = \beta_\dagger, \beta = (1 - \alpha_\dagger)/2$ and the following set $\mathcal{C}$

$$\mathcal{C} := \left\{\theta \in \mathbb{R}^d : V(\theta) \leq \frac{2\beta_\dagger}{\gamma - \alpha_\dagger}\right\}\,, \tag{A.3}$$

for an arbitrary but fixed $\gamma \in (\alpha_\dagger^{1/2}, 1)$.

Corollary A.3 (Minorization condition)**.** *Under Assumptions 2.1-2.3, given the step size specified in Lemma A.1, there exists a constant $\zeta > 0$, and a probability measure $\nu^\dagger$ (depending on η which is suppressed in the notation) with $\nu^\dagger(\mathcal{C}) = 1$ and $\nu^\dagger(\mathcal{C}^c) = 0$, such that*

$$P(\theta, A) \geq \zeta\nu^\dagger(A)$$

holds for any $A \in \mathcal{B}(\mathbb{R}^d)$ and $\theta \in \mathcal{C}$ for the set $\mathcal{C}$ defined in (A.3).

Proof of Corollary A.3. Recall the definition of the markov chain (1.2), we have

$$\xi_{k+1}(\theta_k^{(\eta)}) = \frac{\theta_k^{(\eta)} - \theta_{k+1}^{(\eta)}}{\eta} - \nabla f(\theta_k^{(\eta)})\,.$$

Recall that the distribution of $\xi_1(\theta)$ can be decomposed as $\mu_{1,\theta} + \mu_{2,\theta}$ where $\mu_{1,\theta}$ has density p_θ. It then holds for any $\theta \in \mathbb{R}^d$ that

$$P(\theta, \mathcal{C}) = \mathbb{P}(\theta_{k+1}^{(\eta)} \in \mathcal{C}|\theta_k^{(\eta)} = \theta) \geq \int_{t \in \mathcal{C}} \frac{1}{\eta^d} p_\theta\left(\frac{\theta - t}{\eta} - \nabla f(\theta)\right)dt > 0\,. \tag{A.4}$$

This implies every state in the state space is within reach of any other state over the set $\mathcal{C}$. Define the probability measure $\nu^\dagger$ with density

$$p_{\nu^\dagger}(t) := I\{\theta \in \mathcal{C}\}\frac{\inf_{\theta \in \mathcal{C}} p(t|\theta)}{\int_{t \in \mathcal{C}} \inf_{\theta \in \mathcal{C}} p(t|\theta)dt}\,,$$

and set the constant $\zeta := \int_{t \in \mathcal{C}} \inf_{\theta \in \mathcal{C}} p(t|\theta) dt$. By Assumption 2.3 and the display (A.4), it holds that $\zeta > 0$, $\nu^\dagger(\mathcal{C}) = 1$ and $\nu^\dagger(\mathcal{C}^c) = 0$. Moreover, it holds that any $A \in \mathcal{B}(\mathbb{R}^d)$ and $\theta \in \mathcal{C}$ that

$$P(\theta, A) \geq \zeta \nu^\dagger(A).$$

This implies the minorization condition is met for all choices of η given by Lemma A.1. $\qquad\square$

Lemma A.2. *Under Assumptions 2.1-2.3, the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is an aperiodic, ψ-irreducible, and Harris recurrent chain, with an invariant measure π_η.*

Remark A.1. *This lemma implies the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is positive.*

Proof of Lemma A.2. **Step 1**: We show that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is aperiodic. By Assumption 2.3, there does not exist $d \geq 2$ and a partition of size $d+1$ such that $\mathcal{B}(\mathbb{R}^d) = (\dot{\cup}_{i=1}^d D_i) \dot{\cup} N$, where $\dot{\cup}$ denotes the disjoint union, and N is a ψ-null (transient) set, such that $P(\theta, D_{i+1}) = 1$ holds for ψ-a.e. $\theta \in D_i$. Thus, the largest *period* of the chain defined in (1.2) is 1, which implies the chain is aperiodic.

Step 2: We show that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is ψ-irreducible, and recurrent with an invariant probability measure. We note that by Assumption 2.3, there exists some non-zero σ-finite measure ψ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that for any $\theta \in \mathbb{R}^d$ and any $A \in \mathcal{B}(\mathbb{R}^d)$ with $\psi(A) > 0$, it holds that

$$\mathbb{P}(\theta_{k+1}^{(\eta)} \in A | \theta_k^{(\eta)} = \theta) \geq \int_{\tilde{\theta} \in A} \frac{1}{\eta^d} p_\theta \left(\frac{\theta - \tilde{\theta}}{\eta} - \nabla f(\theta) \right) d\tilde{\theta} > 0,$$

where p_θ was defined in Assumption 2.3. This implies the Markov chain defined in (1.2) is ψ-irreducible. By the Lyapunov condition established in Corollary A.2, part (iii) of Theorem 15.0.1 in [MT12] holds. It then follows by condition (i) of this theorem that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is recurrent with an invariant probability measure π_η.

Step 3: We show that the chain is Harris recurrent. Define the hitting time $\tau_\mathcal{C} := \inf\{n > 0 : \theta_n^{(\eta)} \in \mathcal{C}\}$, where the set $\mathcal{C}$ is defined in (A.3). By Corollary A.4 in [MSH02], it holds for any fixed $\theta_0^{(\eta)} = \theta_0 \in \mathbb{R}^d$ that

$$\mathbb{P}(\tau_\mathcal{C} < \infty) = 1.$$

By Proposition 10.2.4 in [DMPS18], the chain is Harris recurrent. $\qquad\square$

Now, we are ready to prove Proposition 2.1.

Proof of Proposition 2.1. (a) By Lemma A.2, the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is an aperiodic Harris recurrent chain, with an invariant measure π_η. Note that the chain is also positive. Thus condition (i) of Theorem 13.0.1 in [MT12] is satisfied and this implies the existence of a unique invariant measure π_η. The fact that this stationary distribution has a finite second moment was established in Corollary A.1.

(b) By Lemma A.2, the iterates $\{\theta_k^{(\eta)}\}_{k\geq 0}$ are realiztions from a ψ-irreducible and aperiodic chain. Note that

$$
\begin{aligned}
|\phi(\theta)| &\leq \kappa_\phi (1 + \|\theta\|) \\
&\leq 2\kappa_\phi \sqrt{1 + \|\theta\|^2} \\
&\leq 2\kappa_\phi V(\theta).
\end{aligned}
$$

24

By Corollary A.2, the condition (iv) of Theorem 16.0.1 in [MT12] with $V(\theta) = 2\kappa_\phi(1 + \|\theta\|^2)$ is fulfilled. By part (ii) in that theorem, it holds that for fixed $\theta_0^{(\eta)} = \theta_0 \in \mathbb{R}^d$

$$|P^k\phi(\theta_0) - \pi_\eta(\phi)| \leq \kappa\rho^k V(\theta_0),$$

where $\rho \in (0, 1), \kappa > 0$ are constants depending on ϕ. $\qquad \square$

We now prove Theorem 2.1. In order to do so, we first derive the central limit theorem for the function h when the Markov chain starting from its stationary distribution π_η.

Lemma A.3 (CLT with stationary initial distribution). *Assume Assumptions 2.1-2.3 hold. For any step size* $\eta \in \left(0, \frac{\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi))\vee 0}}{3L^2 + L_\xi}\right)$, *it holds that*

$$n^{-1/2}S_n(\phi) \underset{\mathbb{P}_{\pi_\eta}}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\pi_\eta}(\phi)),$$

where $\sigma^2_{\pi_\eta}(\phi) = 2\pi_\eta(h\hat{h}) - \pi_\eta(h^2)$ *with* $\hat{h} = \sum_{k=0}^\infty P^k h$.

Proof of Lemma A.3. We prove the claim by appealing to Theorem 17.0.1 in [MT12]. In order to do so, we first show that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is V-uniformly ergodic, where the function V is defined in (A.2). Then, we establish the CLT by employing Theorem 17.0.1 in [MT12].

Step 1: We show that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is V-uniformly ergodic. By Lemma A.2 and Proposition 2.1, the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is positive Harris recurrent with a unique stationary distribution π_η. Note that the chain $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is also ψ-irreducible and aperiodic. By Corollary A.2, condition (iv) of Theorem 16.0.1 in [MT12] is satisfied. Then, it follows from part (i) of this theorem that the iterates $\{\theta_k^{(\eta)}\}_{k\geq 0}$ is V-uniformly ergodic.

Step 2: We now establish the CLT for the averaged SGD iterates starting from the stationary distribution π_η. Note that for the test function $\phi(\theta)$, it holds for any $\theta \in \mathbb{R}^d$ that

$$|\phi(\theta)| \leq \kappa_\phi(1 + \|\theta\|) \leq 2\kappa_\phi\sqrt{1 + \|\theta\|^2},$$

which implies

$$|\phi(\theta)|^2 \leq 4\kappa_\phi^2 V(\theta).$$

Thus the conditions required to leverage Theorem 17.0.1 (ii), (iv) with $g(\theta) = \phi(\theta)$ in [MT12] are satisfied. Hence, by Theorem 17.0.1 in [MT12], we obtain

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} h(\theta_k^{(\eta)}) \underset{\mathbb{P}_{\pi_\eta}}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\pi_\eta}(\phi)),$$

where $\sigma^2_{\pi_\eta}(\phi) = 2\pi_\eta(h\hat{h}) - \pi_\eta(h^2) > 0$. $\qquad \square$

Proof of Theorem 2.1. By Lemma A.2 and Lemma A.3, the desired result follows readily from Proposition 17.1.6 in [MT12]. $\qquad \square$

A.3   Proofs of Proposition 3.1, Theorems 3.1, 3.2, and 3.3

We need the following auxiliary lemma.

Lemma A.4. *Assumptions 2.1 and 2.2 implies*

$$\langle \nabla f(\theta), \theta - \theta^* \rangle \geq \alpha' \|\theta - \theta^*\|^2 - \beta',$$

where $\theta^* \in \mathbb{R}^d$ is any critical point of function f, and α', β' are positive constants.

Proof of Lemma A.4. When $\theta^* = \mathbf{0}$, the result follows trivially from Assumption 2.2. Assume $\|\theta^*\| > 0$. Note that

$$\langle \nabla f(\theta), \theta - \theta^* \rangle = \langle \nabla f(\theta), \theta \rangle - \langle \nabla f(\theta), \theta^* \rangle.$$

By Assumption 2.2, it holds that

$$\langle \nabla f(\theta), \theta \rangle \geq \alpha \|\theta\|^2 - \beta$$
$$\geq \alpha(\|\theta - \theta^*\|^2 + \|\theta^*\|^2 - 2\|\theta^*\|\|\theta - \theta^*\|) - \beta.$$

By Assumption 2.1, Cauchy-Schwarz inequality and triangular inequality, it holds that

$$\langle \nabla f(\theta), \theta^* \rangle \leq \|\nabla f(\theta)\|\|\theta^*\| \leq L\|\theta^*\|(1 + \|\theta - \theta^*\| + \|\theta^*\|).$$

Combing the previous two displays yields

$$\langle \nabla f(\theta), \theta - \theta^* \rangle$$
$$\geq \alpha(\|\theta - \theta^*\|^2 + \|\theta^*\|^2 - 2\|\theta^*\|\|\theta - \theta^*\|) - \beta - L\|\theta^*\|(1 + \|\theta - \theta^*\| + \|\theta^*\|)$$
$$\geq \frac{\alpha}{2}\|\theta - \theta^*\|^2 - \beta - L\|\theta^*\|^2 - L\|\theta^*\|.$$

The desired result follows by setting $\alpha' := \frac{\alpha}{2}$ and $\beta' := \beta + \left(\sqrt{\alpha} + \frac{2L}{\sqrt{\alpha}}\right)^2\|\theta^*\|^2 + L\|\theta^*\|$. $\square$

Lemma A.5. *Under Assumptions 2.2 and 3.1, it holds for any $k \geq 1$ and $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}[\|\xi_{k+1}(\theta)\|^r] \leq L_\xi'^{r/4}(1 + \|\theta - \theta^*\|^r), \quad for \quad r \in \{2, 3, 4\},$$

where $\theta^* \in \mathbb{R}^d$ is any critical point of function f, and $L_\xi' := 8L_\xi(1 + (\beta/\alpha)^4)$.

Proof of Lemma A.5. By Assumptions 2.2 and 3.1, it holds that

$$\mathbb{E}[\|\xi_{k+1}(\theta)\|^4] \leq L_\xi(1 + \|\theta\|^4)$$
$$\leq L_\xi(1 + 8\|\theta - \theta^*\|^4 + 8\|\theta^*\|^4)$$
$$\leq L_\xi(1 + 8\|\theta - \theta^*\|^4 + 8(\beta/\alpha)^4)$$
$$\leq L_\xi'(1 + \|\theta - \theta^*\|^4),$$

where $L_\xi' := 8L_\xi(1 + (\beta/\alpha)^4)$. Similarly, for $r \in \{2, 3\}$ we have

$$\mathbb{E}[\|\xi_{k+1}(\theta)\|^r] \leq \mathbb{E}[\|\xi_{k+1}(\theta)\|^4]^{r/4}$$
$$\leq L_\xi^{r/4}(1 + \|\theta\|^4)^{r/4}$$
$$\leq L_\xi^{r/4}(1 + \|\theta\|^r)$$
$$\leq L_\xi^{r/4}(1 + 2^{r-1}\|\theta - \theta^*\|^r + 2^{r-1}\|\theta^*\|^r)$$
$$\leq L_\xi^{r/4}(1 + 2^{r-1}\|\theta - \theta^*\|^r + 2^{r-1}(\beta/\alpha)^r)$$
$$\leq L_\xi'^{r/4}(1 + \|\theta - \theta^*\|^r),$$

where L_ξ' is defined above. $\square$

Lemma A.6. *Under Assumptions 2.1, 2.2, and 3.1, with step size $\eta < 1 \wedge \frac{1}{10\bar{L}}$, it holds for any $k \geq 0$ that*

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)} - \theta^*\|^4 | \mathcal{F}_k]$$
$$\leq (1 - 4\eta\alpha' + 32L_\dagger\eta^2)\|\theta_k^{(\eta)} - \theta^*\|^4 + \eta(4\beta' + 24\bar{L}^2 + 12L_\xi'^{1/2} + 64)\|\theta_k^{(\eta)} - \theta^*\|^2$$
$$+ \eta^2(64\bar{L}^4 + 8L_\xi' + 32(4\bar{L}^3)^2 + 32(L_\xi')^{3/2}). \tag{A.5}$$

where $L_\dagger := \bar{L}^2 + 16\left(L_\xi^{3/4}(1 + (\beta/\alpha)^3) \vee L_\xi^{1/2}(1 + (\beta/\alpha)^2) \vee L_\xi(1 + (\beta/\alpha)^4)\right)$ with $\bar{L} := L(1 + \|\theta^*\|)$, L_ξ' is from Lemma A.5, and θ^* is any critical points of fuction f.

Proof of Lemma A.6. Define $\Delta_k := \|\theta_k^{(\eta)} - \theta^*\|$. It holds by Assumption 2.1 that

$$\|\nabla f(\theta_k^{(\eta)})\| \leq \bar{L}\Delta_k + \bar{L},$$

where $\bar{L} = L(\|\theta^*\| + 1)$. Note that

$$\Delta_{k+1}^4 = (\Delta_k^2 + \eta^2 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2 - 2\eta\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle)^2$$
$$= \Delta_k^4 + \eta^4 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^4 + 4\eta^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle^2$$
$$+ 2\eta^2\Delta_k^2 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2 - 4\eta\Delta_k^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle$$
$$- 4\eta^3 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle$$
$$= \Delta_k^4 + \text{I} + \text{II} + \text{III} + \text{IV} + \text{V},$$

where

$$\text{I} := \eta^4 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^4$$
$$\text{II} := 4\eta^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle^2$$
$$\text{III} := 2\eta^2\Delta_k^2 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2$$
$$\text{IV} := -4\eta\Delta_k^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle$$
$$\text{V} := -4\eta^3 \|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle.$$

To obtain the expectation $\mathbb{E}[\Delta_{k+1}^4]$, we first calculate the conditional expectation $\mathbb{E}[\Delta_{k+1}^4 | \mathcal{F}_k]$. For this, we proceed the conditional expectation of the above five terms separately. Note that

$$\mathbb{E}[\text{I}|\mathcal{F}_k] = \eta^4\mathbb{E}[\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^4|\mathcal{F}_k]$$
$$\leq \eta^4\mathbb{E}[8\|\nabla f(\theta_k^{(\eta)})\|^4 + 8\|\xi_{k+1}(\theta_k^{(\eta)})\|^4|\mathcal{F}_k]$$
$$\leq 8\eta^4(8\bar{L}^4\Delta_k^4 + 8\bar{L}^4 + L_\xi'\Delta_k^4 + L_\xi').$$

The first inequality follows from the fact that $(x + y)^4 \leq 8(x^4 + y^4), \forall x, y > 0$. The last inequality follows from Assumptions 2.1 and Lemma A.5. Using the same trick and invoking Cauchy-Schwarz inequality gives

$$\mathbb{E}[\text{II}|\mathcal{F}_k] = 4\eta^2\mathbb{E}[\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^*\rangle^2|\mathcal{F}_k]$$
$$\leq 4\eta^2\Delta_k^2\mathbb{E}[\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2|\mathcal{F}_k]$$
$$\leq 8\eta^2\Delta_k^2(2\bar{L}^2\Delta_k^2 + 2\bar{L}^2 + L_\xi'^{1/2}\Delta_k^2 + L_\xi'^{1/2}).$$

27

Similarly, we have

$$\mathbb{E}[\mathrm{III}|\mathcal{F}_k] =2\eta^2\Delta_k^2\mathbb{E}[\,\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2|\mathcal{F}_k]$$
$$\leq 4\eta^2\Delta_k^2(2\bar{L}^2\Delta_k^2 + 2\bar{L}^2 + L_\xi'^{1/2}\Delta_k^2 + L_\xi'^{1/2})\,.$$

Using Cauchy-Schwarz inequality again, we obtain

$$\mathbb{E}[\mathrm{V}|\mathcal{F}_k] =\mathbb{E}[-4\eta^3\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2\langle\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}),\ \theta_k^{(\eta)} - \theta^*\rangle|\mathcal{F}_k]$$
$$\leq 4\eta^3\mathbb{E}[\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^3\|\theta_k^{(\eta)} - \theta^*\||\mathcal{F}_k]$$
$$=4\eta^3\Delta_k\mathbb{E}\big[\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^3|\mathcal{F}_k\big]$$
$$\leq 4\eta^3\Delta_k\mathbb{E}\big[4\|\nabla f(\theta_k^{(\eta)})\|^3 + 4\|\xi_{k+1}(\theta_k^{(\eta)})\|^3|\mathcal{F}_k\big]\,.$$

Note that by Lemma A.5, it holds for any $k \geq 1$ and $\theta \in \mathbb{R}^d$ that

$$\mathbb{E}[\|\xi_k(\theta)\|^3] \leq L_\xi'^{3/4}(1 + \|\theta - \theta^*\|^3)\,.$$

Combining this with the previous display yields

$$\mathbb{E}[\mathrm{V}|\mathcal{F}_k] \leq 16\eta^3\Delta_k(4\bar{L}^3\Delta_k^3 + 4\bar{L}^3 + L_\xi'^{3/4} + L_\xi'^{3/4}\Delta_k^3)$$
$$=64\bar{L}^3\eta^3\Delta_k^4 + 16\eta^3 L_\xi'^{3/4}\Delta_k^4 + 16\eta^2(\Delta_k\eta 4\bar{L}^3 + \Delta_k\eta L_\xi'^{3/4})\,.$$

Collecting pieces gives

$$\mathbb{E}[\Delta_{k+1}^4|\mathcal{F}_k] \leq \Delta_k^4(1 + 64\eta^4\bar{L}^4 + 64\eta^3\bar{L}^3 + 24\eta^2\bar{L}^2 + 8\eta^2 L_\xi' + 12\eta^2 L_\xi'^{1/2} + 16\eta^2 L_\xi'^{3/4})$$
$$- 4\eta\Delta_k^2\langle\nabla f(\theta_k^{(\eta)}),\ \theta_k^{(\eta)} - \theta^*\rangle$$
$$+ \eta^2\big(64\eta^2\bar{L}^4 + 8\eta^2 L_\xi' + 24\bar{L}^2\Delta_k^2 + 12 L_\xi'\Delta_k^2 + 64\Delta_k^2 + 32(\eta 4\bar{L}^3)^2 + 32(\eta L_\xi'^{3/4})^2\big)$$
$$\leq \Delta_k^4[1 + 32\eta^2(\bar{L}^2 + L_\xi' + L_\xi'^{1/2} + L_\xi'^{3/4})] - 4\eta\Delta_k^2\langle\nabla f(\theta_k^{(\eta)}),\ \theta_k^{(\eta)} - \theta^*\rangle$$
$$+ \eta\big(64\bar{L}^4\eta + 8 L_\xi'\eta + 24\bar{L}^2\Delta_k^2 + 12 L_\xi'^{1/2}\Delta_k^2 + 64\Delta_k^2 + 32(4\bar{L}^3)^2\eta + 32(L_\xi')^{3/2}\eta\big)\,.$$

The above inequalities are based on the fact that $\eta < \frac{1}{10\bar{L}} \wedge 1$ and $xy \leq 2x^2 + 2y^2, \forall x, y > 0$. By Lemma A.4, we handle the term IV as following

$$\mathbb{E}[\Delta_{k+1}^4|\mathcal{F}_k] \leq \Delta_k^4\big(1 - 4\eta\alpha' + 32\eta^2(\bar{L}^2 + L_\xi' + L_\xi'^{1/2} + L_\xi'^{3/4})\big)$$
$$+ \eta\big(4\beta'\Delta_k^2 + 64\bar{L}^4\eta + 8 L_\xi'\eta + 24\bar{L}^2\Delta_k^2 + 12 L_\xi'^{1/2}\Delta_k^2 + 64\Delta_k^2 + 32(4\bar{L}^3)^2\eta + 32 L_\xi'^{3/2}\eta\big)\,.$$

Define $L_\dagger := \bar{L}^2 + 16\big(L_\xi'^{3/4}(1 + (\beta/\alpha)^3) \vee L_\xi'^{1/2}(1 + (\beta/\alpha)^2) \vee L_\xi(1 + (\beta/\alpha)^4)\big)$. Note that $L_\dagger > \bar{L}^2 + L_\xi' + L_\xi'^{1/2} + L_\xi'^{3/4}$. Combing this with the previous display gives

$$\mathbb{E}[\Delta_{k+1}^4|\mathcal{F}_k]$$
$$\leq\Delta_k^4(1 - 4\eta\alpha' + 32\eta^2 L_\dagger)$$
$$+ \eta\big(4\beta'\Delta_k^2 + 64\bar{L}^4\eta + 8 L_\xi'\eta + 24\bar{L}^2\Delta_k^2 + 12 L_\xi'^{1/2}\Delta_k^2 + 64\Delta_k^2 + 32(4\bar{L}^3)^2\eta + 32 L_\xi'^{3/2}\eta\big)$$
$$\leq(1 - 4\eta\alpha' + 32 L_\dagger\eta^2)\Delta_k^4 + \eta(4\beta' + 24\bar{L}^2 + 12 L_\xi'^{1/2} + 64)\Delta_k^2 + \eta^2(64\bar{L}^4 + 8 L_\xi' + 32(4\bar{L}^3)^2 + 32 L_\xi'^{3/2})\,.$$

$\square$

Lemma A.7. *Assume Assumptions [2.1](#)-[2.3](#) holds. With the step size*

$$\eta \leq \frac{\alpha - \sqrt{(\alpha^2 - (3L^2 + L_\xi)) \vee 0}}{3L^2 + L_\xi} \wedge \frac{\alpha}{64L_\dagger} \wedge 1 \,,$$

the chain ([1.2](#)) has the stationary distribution π_η, and the chain has finite 4-th moment:

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^4] \leq \mu_{4,\eta} \,,$$

where

$$\mu_{4,\eta} := \frac{8}{7\alpha} \left((\beta + 6L^2 + 3L_\xi^{1/2} + 16)\mu_{2,\eta} + 16L^4 + 2L_\xi + 128L^6 + 8L_\xi^{3/2} \right)$$

with $\mu_{2,\eta}$ defined in Corollary [A.1](#), and $L_\dagger$ defined in Lemma [A.6](#).

Proof of Lemma [A.7](#). Similar to display ([A.5](#)), we can derive

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^4|\mathcal{F}_k] \leq (1 - 4\eta\alpha + 32L_0^\dagger \eta^2)\|\theta_k^{(\eta)}\|^4$$
$$+ \eta \big[(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\|\theta_k^{(\eta)}\|^2 + \eta(64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2}) \big] \,,$$

where $L_0^\dagger := L^2 + L_\xi + L_\xi^{1/2} + L_\xi^{3/4}$. Recall the definition of $L_\dagger$ in Lemma [A.6](#), it holds that $L_\dagger \geq L_0^\dagger$, which implies

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^4|\mathcal{F}_k] \leq (1 - 4\eta\alpha + 32L_\dagger \eta^2)\|\theta_k^{(\eta)}\|^4$$
$$+ \eta \big[(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\|\theta_k^{(\eta)}\|^2 + \eta(64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2}) \big] \,,$$

Note that the chain starts from the stationary distribution π_η, taking the expectation on both sides gives

$$(4\eta\alpha - 32L_\dagger \eta^2)\mathbb{E}[\|\theta_k^{(\eta)}\|^4]$$
$$\leq \eta(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\mathbb{E}[\|\theta_k^{(\eta)}\|^2] + \eta^2(64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2}) \,.$$

We also note that $\mathbb{E}[\|\theta_k^{(\eta)}\|^2] = \mu_{2,\eta}$ for $\mu_{2,\eta}$ from Corollary [A.1](#). Plugging this into the previous display and rearranging the inequality gives

$$\mathbb{E}[\|\theta_{k+1}^{(\eta)}\|^4]$$
$$\leq \frac{\eta}{4\eta\alpha - 32L_\dagger \eta^2}(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\mu_{2,\eta} + \frac{\eta^2}{4\eta\alpha - 32L_\dagger \eta^2}(64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2})$$
$$\leq \frac{\eta}{4\eta\alpha - 32L_\dagger \eta^2}(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\mu_{2,\eta} + \frac{\eta}{4\eta\alpha - 32L_\dagger \eta^2}(64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2})$$
$$\leq \frac{2}{7\alpha} \Big[(4\beta + 24L^2 + 12L_\xi^{1/2} + 64)\mu_{2,\eta} + (64L^4 + 8L_\xi + 32(4L^3)^2 + 32L_\xi^{3/2}) \Big]$$

as desired. □

We are now ready to prove Proposition [3.1](#).

Proof of Proposition 3.1. Define $\Delta_k := \|\theta_k^{(\eta)} - \theta^*\|$. By Lemma A.6, we have

$$\mathbb{E}[\Delta_{k+1}^4 | \mathcal{F}_k]$$
$$\leq (1 - 4\eta\alpha' + 32L_\dagger \eta^2)\Delta_k^4 + \eta(4\beta' + 24\bar{L}^2 + 12L_\xi'^{1/2} + 64)\Delta_k^2$$
$$+ \eta^2(64\bar{L}^4 + 8L_\xi' + 32(4\bar{L}^3)^2 + 32L_\xi'^{3/2}).$$

Taking expectation on both sides then gives

$$\mathbb{E}[\Delta_{k+1}^4]$$
$$\leq (1 - 4\eta\alpha' + 32L_\dagger \eta^2)\mathbb{E}[\Delta_k^4] + \eta(4\beta' + 24\bar{L}^2 + 12L_\xi'^{1/2} + 64)\mathbb{E}[\Delta_k^2]$$
$$+ \eta^2(64\bar{L}^4 + 8L_\xi' + 32(4\bar{L}^3)^2 + 32L_\xi'^{3/2}).$$

Set

$$\varrho := 1 - 4\eta\alpha' + 32L_\dagger \eta^2$$
$$A_1 := 64\bar{L}^4 + 8L_\xi' + 32(4\bar{L}^3)^2 + 32L_\xi'^{3/2}$$
$$A_2 := 4\beta' + 24\bar{L}^2 + 12L_\xi'^{1/2} + 64.$$

By Cauchy-Schwatz inequality, we then have

$$\mathbb{E}[\Delta_{k+1}^4] \leq \varrho \mathbb{E}[\Delta_k^4] + A_1 \eta^2 + A_2 \mathbb{E}^{1/2}[\Delta_k^4]\eta.$$

Note that when $0 < \eta < \frac{\alpha' - \sqrt{(\alpha'^2 - 4L_\dagger)}}{16L_\dagger}\mathbb{1}(\alpha'^2 > 8L_\dagger) + \frac{\alpha'}{32L_\dagger}\mathbb{1}(\alpha'^2 \leq 8L_\dagger)$, it follows that

$$\varrho > \frac{1}{2}\mathbb{1}(\alpha'^2 \geq 8L_\dagger) + (1 - \frac{3\alpha'^2}{32L_\dagger^2})\mathbb{1}(\alpha'^2 < 8L_\dagger) \geq \frac{1}{4}.$$

Set $D := \sqrt{A_1} \vee A_2$. We then find

$$\mathbb{E}^{1/2}[\Delta_{k+1}^4] \leq \sqrt{\varrho}\,\mathbb{E}^{1/2}[\Delta_k^4] + D\eta.$$

By a straightforward induction, we have

$$\mathbb{E}^{1/2}[\Delta_k^4] \leq \varrho^{k/2}\mathbb{E}^{1/2}[\Delta_0^4] + \frac{D\eta}{1 - \sqrt{\varrho}}.$$

Notice that $\eta \leq \frac{\alpha'}{16L_\dagger}$, it then follows that

$$\varrho = 1 - 4\eta\alpha' + 32L_\dagger \eta^2 \leq 1 - 2\eta\alpha',$$

which implies

$$\frac{1}{1 - \sqrt{\varrho}} \leq \frac{1}{1 - \sqrt{1 - 2\eta\alpha'}} \leq \frac{1}{\eta\alpha'}.$$

Combining this with previous display gives

$$\mathbb{E}^{1/2}[\Delta_k^4] \leq \varrho^{k/2}\mathbb{E}^{1/2}[\Delta_0^4] + \frac{D}{\alpha'}.$$

By Proposition 2.1, there exists a unique stationary distribution π_η.

Consider the chain starting from the stationary distribution π_η. Note that $\mathbb{E}[\Delta_0^4] \leq 8(\mathbb{E}[\|\theta_0^{(\eta)}\|^4] + \|\theta^*\|^4)$. By Lemma A.7, it follows that

$$\mathbb{E}[\Delta_0^4] \leq 8\mu_{4,\eta} + 8\|\theta^*\|^4 ,$$

where the constant $\mu_{4,\eta}$ is defined in Lemma A.7. Plugging this into previous display provides us with

$$\left(\int \|\theta - \theta^*\|^4 \pi_\eta(d\theta) \right)^{1/4} = \mathcal{O}(1) .$$

Note that it holds for the L_ϕ-Lipschitz continuous test function ϕ that

$$|\pi_\eta(\phi) - \phi(\theta^*)| \leq L_\phi \int \|\theta - \theta^*\| \pi_\eta(d\theta)$$

$$\leq L_\phi \left[\int \|\theta - \theta^*\|^4 \pi_\eta(d\theta) \right]^{1/4} ,$$

Thus, we obtain

$$|\pi_\eta(\phi) - \phi(\theta^*)| = \mathcal{O}(1)$$

as desired.

$\qquad\square$

Lemma A.8. *For any $a, b, \delta > 0$, it holds for any $x \geq \frac{\delta}{a} + \sqrt{\frac{b}{a}}$ that*

$$ax^2 - b \geq \delta x .$$

Proof of Lemma A.8. Define the function $h(x) := ax^2 - b - \delta x$. When $x \geq \frac{\delta + \sqrt{\delta^2 + 4ab}}{2a}$, it holds that $h(x) \geq 0$. Note that $\sqrt{\delta^2 + 4ab} \leq \delta + \sqrt{4ab}$, it follows that when

$$x \geq \frac{\delta + \delta + \sqrt{4ab}}{2a} ,$$

it holds that $h(x) \geq 0$. The desired result then follows readily. $\qquad\square$

Proof of Theorem 3.1. Consider the chain $\{\theta_k^{(\eta)}\}_{k \geq 0}$ starting from the stationary distribution π_η. Define $\Delta_k := \|\theta_k^{(\eta)} - \theta^*\|$. Note that under Assumptions 3.2 and 3.1, Lemma A.5 still holds. By Assumptions 2.1, 3.1, and Lemma A.5 , we have

$$\mathbb{E}\big[\Delta_{k+1}^2 | \mathcal{F}_k\big]$$
$$= \mathbb{E}\big[\Delta_k^2 + \eta^2 \|\nabla f(\theta_k^{(\eta)})\|^2 + \eta^2 \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 - 2\eta \langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle | \mathcal{F}_k\big]$$
$$\leq \Delta_k^2 + \eta^2 \big(3L^2(2\Delta_k^2 + 2\|\theta^*\|^2 + 3) + L_\xi'^{1/2}(1 + \Delta_k^2)\big) - 2\eta \langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle$$
$$= \Delta_k^2 + 6L^2\eta^2\Delta_k^2 + L_\xi'^{1/2}\eta^2\Delta_k^2 + \eta^2 C_1 - 2\eta \langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle$$

where $C_1 := 6\|\theta^*\|^2 L^2 + 9L^2 + L_\xi'^{1/2}$. Note that the chain starts from the stationary distribution π_η, which implies $\mathbb{E}[\Delta_{k+1}^2] = \mathbb{E}[\Delta_k^2]$ for all $k \geq 0$. Taking the expectation on both sides and rearranging the inequality yields

$$\mathbb{E}[\langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle] \leq \eta(3L^2 + L_\xi'^{1/2})\mathbb{E}[\Delta_k^2] + \frac{\eta}{2}C_1 .$$

By Corollary A.1, it follows that

$$\mathbb{E}[\langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle] \leq C_2 \eta, \tag{A.6}$$

where $C_2 := 2(3L^2 + L_\xi'^{1/2})(\mu_{2,\eta} + \|\theta^*\|^2) + C_1/2$ and $\mu_{2,\eta}$ is defined in Corollary A.1. Moreover, by Assumption 3.2, Lemma A.8, and Jensen's inequality, we have

$$\mathbb{E}[\langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle] \geq \delta \mathbb{E}[\Delta_k \mathbb{1}(\Delta_k \geq R)] + g(\mathbb{E}[\Delta_k \mathbb{1}(\Delta_k < R)]).$$

Combining this with previous display provides us with

$$\mathbb{E}[\Delta_k \mathbb{1}(\Delta_k \geq R)] \leq \frac{C_2}{\delta} \eta,$$

and

$$\mathbb{E}[\Delta_k \mathbb{1}(\Delta_k < R)] \leq g^{-1}(C_2 \eta).$$

Collecting pieces then gives

$$\begin{aligned}
\mathbb{E}[\Delta_k] &= \mathbb{E}[\Delta_k \mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| < R)] + \mathbb{E}[\Delta_k \mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| \geq R)] \\
&\leq \frac{C_2}{\delta} \eta + g^{-1}(C_2 \eta).
\end{aligned}$$

Thus, it holds for the L_ϕ-Lipschitz continuous test function ϕ that

$$|\pi_\eta(\phi) - \phi(\theta^*)| \leq L_\phi \int \|\theta - \theta^*\| \pi_\eta(d\theta) \leq L_\phi \Big(\frac{C_2}{\delta} \eta + g^{-1}(C_2 \eta) \Big).$$

$\square$

Proof of Theorem 3.2. Consider the chain $\{\theta_k^{(\eta)}\}_{k \geq 0}$ starting from the stationary distribution π_η. Note that by the assumption that $\|\nabla^2 f(\theta)\| \leq \tilde{L}(1 + \|\theta\|)$ and Taylor expansion, we have

$$\begin{aligned}
f(\theta_{k+1}^{(\eta)}) &= f(\theta_k^{(\eta)}) + \langle \nabla f(\theta_k^{(\eta)}), \theta_{k+1}^{(\eta)} - \theta_k^{(\eta)} \rangle + \frac{1}{2}(\theta_{k+1}^{(\eta)} - \theta_k^{(\eta)})^\top \nabla^2 f(\tilde{\theta})(\theta_{k+1}^{(\eta)} - \theta_k^{(\eta)}) \\
&\leq f(\theta_k^{(\eta)}) + \langle \nabla f(\theta_k^{(\eta)}), \theta_{k+1}^{(\eta)} - \theta_k^{(\eta)} \rangle + \frac{1}{2}\tilde{L}\|\theta_{k+1}^{(\eta)} - \theta_k^{(\eta)}\|^2(1 + \|\tilde{\theta}\|),
\end{aligned}$$

where $\tilde{\theta} \in \mathbb{R}^d$ is a convex combination between $\theta_{k+1}^{(\eta)}$ and $\theta_k^{(\eta)}$. By definition of SGD iterates in (1.2), it follows that

$$\begin{aligned}
f(\theta_{k+1}^{(\eta)}) \leq & f(\theta_k^{(\eta)}) - \eta \langle \nabla f(\theta_k^{(\eta)}), \nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}) \rangle + \frac{\tilde{L}}{2}\eta^2\|\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\|^2(1 + \|\tilde{\theta}\|) \\
= & f(\theta_k^{(\eta)}) - \eta \langle \nabla f(\theta_k^{(\eta)}), \nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}) \rangle \\
& + \frac{\tilde{L}}{2}\eta^2\big(\|\nabla f(\theta_k^{(\eta)})\|^2 + \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 + 2\langle \nabla f(\theta_k^{(\eta)}), \xi_{k+1}(\theta_k^{(\eta)}) \rangle\big)(1 + \|\tilde{\theta}\|) \\
\leq & f(\theta_k^{(\eta)}) - \eta \langle \nabla f(\theta_k^{(\eta)}), \nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)}) \rangle \\
& + \frac{\tilde{L}}{2}\eta^2\big(\|\nabla f(\theta_k^{(\eta)})\|^2 + \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 + 2\langle \nabla f(\theta_k^{(\eta)}), \xi_{k+1}(\theta_k^{(\eta)}) \rangle\big)(1 + \max\{\|\theta_k^{(\eta)}\|, \|\theta_{k+1}^{(\eta)}\|\}) \\
\leq & f(\theta_k^{(\eta)}) - \eta\|\nabla f(\theta_k^{(\eta)})\|^2 - \eta \langle \nabla f(\theta_k^{(\eta)}), \xi_{k+1}(\theta_k^{(\eta)}) \rangle \\
& + \frac{\tilde{L}}{2}\eta^2\big(\|\nabla f(\theta_k^{(\eta)})\|^2 + \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 + 2\langle \nabla f(\theta_k^{(\eta)}), \xi_{k+1}(\theta_k^{(\eta)}) \rangle\big) \\
& + \frac{\tilde{L}}{2}\eta^2\big(\|\nabla f(\theta_k^{(\eta)})\|^2 + \|\xi_{k+1}(\theta_k^{(\eta)})\|^2 + 2\langle \nabla f(\theta_k^{(\eta)}), \xi_{k+1}(\theta_k^{(\eta)}) \rangle\big)(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|).
\end{aligned}$$

Taking the conditional expectation on both sides, using Cauchy-Schwarz inequality, Assumption 3.1 and the fact that $(1 + x^4)^{1/2} \leq 1 + x^2, \forall x > 0$ gives

$$\mathbb{E}[f(\theta_{k+1}^{(\eta)})|\mathcal{F}_k]$$

$$\leq f(\theta_k^{(\eta)}) + (\frac{\tilde{L}}{2}\eta^2 - \eta)\|\nabla f(\theta_k^{(\eta)})\|^2 + \frac{\tilde{L}}{2}L_\xi \eta^2(1 + \|\theta_k^{(\eta)}\|^2) + 0$$

$$+ \frac{\tilde{L}}{2}\eta^2 \mathbb{E}\left[\|\nabla f(\theta_k^{(\eta)})\|^2(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)|\mathcal{F}_k\right]$$

$$+ \frac{\tilde{L}}{2}\eta^2 \mathbb{E}^{1/2}\left[\|\xi_{k+1}(\theta_k^{(\eta)})\|^4|\mathcal{F}_k\right]\mathbb{E}^{1/2}\left[(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)^2|\mathcal{F}_k\right]$$

$$+ 0 + \tilde{L}\eta^2 \mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|\|\xi_{k+1}(\theta_k^{(\eta)})\|\|\theta_{k+1}^{(\eta)}\||\mathcal{F}_k]$$

$$\leq f(\theta_k^{(\eta)}) + (\frac{\tilde{L}}{2}\eta^2 - \eta)\|\nabla f(\theta_k^{(\eta)})\|^2 + \frac{\tilde{L}}{2}L_\xi \eta^2(1 + \|\theta_k^{(\eta)}\|^2)$$

$$+ \frac{\tilde{L}}{2}\eta^2 \mathbb{E}\left[\|\nabla f(\theta_k^{(\eta)})\|^2(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)|\mathcal{F}_k\right]$$

$$+ \frac{\tilde{L}}{2}\eta^2 L_\xi^{1/2}(1 + \|\theta_k^{(\eta)}\|^2)\mathbb{E}^{1/2}\left[(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)^2|\mathcal{F}_k\right]$$

$$+ \tilde{L}\eta^2 \mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|\|\xi_{k+1}(\theta_k^{(\eta)})\|\|\theta_{k+1}^{(\eta)}\||\mathcal{F}_k].$$

We then take expectation on both sides. For this, we bound the last three terms separately. Note that the chain starts from the initial distribution π_η. By Hölder's inequality, we have

$$\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)]$$

$$\leq \mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2\|\theta_k^{(\eta)}\|] + \mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2\|\theta_{k+1}^{(\eta)}\|]$$

$$\leq \mathbb{E}^{1/2}[\|\nabla f(\theta_k^{(\eta)})\|^4]\mathbb{E}^{1/2}[\|\theta_k^{(\eta)}\|^2] + \mathbb{E}^{1/2}[\|\nabla f(\theta_k^{(\eta)})\|^4]\mathbb{E}^{1/2}[\|\theta_k^{(\eta)}\|^2].$$

By Assumption 2.1 and the fact that $(x + y)^4 \leq 9(x^4 + y^4), \forall x, y \in \mathbb{R}$, we have

$$\mathbb{E}^{1/2}[\|\nabla f(\theta_k^{(\eta)})\|^4] \leq L^2\mathbb{E}^{1/2}[(1 + \|\theta_k^{(\eta)}\|)^4] \leq 3L^2\sqrt{1 + \mathbb{E}[\|\theta_k^{(\eta)}\|^4]}.$$

By Lemma A.7, it holds that $\mathbb{E}[\|\theta_k^{(\eta)}\|^4] < \mu_{4,\eta}$, where the constant $\mu_{4,\eta}$ is defined in Lemma A.7. Moreover, by Corollary A.1, we also have $\mathbb{E}[\|\theta_k^{(\eta)}\|^2] \leq \mu_{2,\eta}$, where the constant $\mu_{2,\eta}$ is defined in Corollary A.1. Combining these with previous display gives

$$\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)] \leq 6L^2\sqrt{1 + \mu_{4,\eta}}\sqrt{\mu_{2,\eta}}.$$

Using the same trick, we obtain

$$\mathbb{E}\left[(1 + \|\theta_k^{(\eta)}\|^2)\mathbb{E}^{1/2}[(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)^2|\mathcal{F}_k]\right]$$

$$\leq \mathbb{E}^{1/2}[(1 + \|\theta_k^{(\eta)}\|^2)^2]\mathbb{E}^{1/2}\left[\mathbb{E}[(\|\theta_k^{(\eta)}\| + \|\theta_{k+1}^{(\eta)}\|)^2|\mathcal{F}_k]\right]$$

$$\leq \mathbb{E}^{1/2}[2 + 2\|\theta_k^{(\eta)}\|^4]\mathbb{E}^{1/2}\left[\mathbb{E}[2\|\theta_k^{(\eta)}\|^2 + 2\|\theta_{k+1}^{(\eta)}\|^2|\mathcal{F}_k]\right]$$

$$\leq 4\mathbb{E}^{1/2}[1 + \|\theta_k^{(\eta)}\|^4]\mathbb{E}^{1/2}[\|\theta_k^{(\eta)}\|^2]$$

$$\leq 4\sqrt{\mu_{2,\eta}}\sqrt{1 + \mu_{4,\eta}}.$$

33

By Assumptions 2.1 and 3.1, we have

$$
\mathbb{E}\Big[\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|\|\xi_{k+1}(\theta_k^{(\eta)})\|\|\theta_{k+1}^{(\eta)}\| | \mathcal{F}_k]\Big]
$$

$$
\leq \mathbb{E}\Big[\|\nabla f(\theta_k^{(\eta)})\|\mathbb{E}^{1/2}[\|\xi_{k+1}(\theta_k^{(\eta)})\|^2 | \mathcal{F}_k]\mathbb{E}^{1/2}[\|\theta_{k+1}^{(\eta)}\|^2 | \mathcal{F}_k]\Big]
$$

$$
\leq L_\xi^{1/2} L \mathbb{E}\Big[(1+\|\theta_k^{(\eta)}\|)(1+\|\theta_k^{(\eta)}\|^2)^{1/2}\mathbb{E}^{1/2}[\|\theta_{k+1}^{(\eta)}\|^2 | \mathcal{F}_k]\Big]
$$

$$
\leq L L_\xi^{1/2} \mathbb{E}\Big[(1+\|\theta_k^{(\eta)}\|)^2 \mathbb{E}^{1/2}[\|\theta_{k+1}^{(\eta)}\|^2 | \mathcal{F}_k]\Big]
$$

$$
\leq L L_\xi^{1/2} \mathbb{E}^{1/2}[(1+\|\theta_k^{(\eta)}\|)^4]\mathbb{E}^{1/4}[\|\theta_k^{(\eta)}\|^4]
$$

$$
\leq L L_\xi^{1/2} \sqrt{8+8\mu_{4,\eta}}(\mu_{4,\eta})^{1/4}
$$

$$
= 3 L L_\xi^{1/2}(\mu_{4,\eta}+\mu_{4,\eta}^{3/4}).
$$

Collecting pieces then gives

$$
\mathbb{E}[f(\theta_{k+1}^{(\eta)})]
$$

$$
\leq \mathbb{E}[f(\theta_k^{(\eta)})] + (\frac{\tilde{L}}{2}\eta^2 - \eta)\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2] + \tilde{L}L_\xi\eta^2(1+\mu_{2,\eta})
$$

$$
+ 3\tilde{L}L^2\eta^2\mu_{2,\eta}^{1/2}\sqrt{1+\mu_{4,\eta}} + 2\tilde{L}L_\xi^{1/2}\eta^2\mu_{2,\eta}^{1/2}\sqrt{1+\mu_{4,\eta}} + 3\tilde{L}LL_\xi^{1/2}\eta^2(\mu_{4,\eta}+\mu_{4,\eta}^{3/4})
$$

$$
\leq \mathbb{E}[f(\theta_k^{(\eta)})] + (\frac{\tilde{L}}{2}\eta^2 - \eta)\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2] + 12\eta^2\tilde{L}(L+L_\xi^{1/2}+L_\xi^{1/4})^2\Big(1+\mu_{2,\eta}+\mu_{4,\eta}+\mu_{4,\eta}^{3/4}\Big).
$$

Recall that the iterates $\{\theta_k^{(\eta)}\}_{k\geq 0}$ starts from the stationary distribution π_η and $\eta < \frac{2}{\tilde{L}}$. Rearranging the above display gives

$$
\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2] \leq \frac{2\tilde{M}\eta}{2-\tilde{L}\eta},
$$

where

$$
\tilde{M} := 12\tilde{L}(L+L_\xi^{1/2}+L_\xi^{1/4})^2\Big(1+\mu_{2,\eta}+\mu_{4,\eta}+\mu_{4,\eta}^{3/4}\Big).
$$

By Assumption 3.3 and Jensen's inequality, it holds that

$$
\mathbb{E}[\|\nabla f(\theta_k^{(\eta)})\|^2] \geq \mathbb{E}[g(f(\theta_k^{(\eta)}) - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| \leq R)] + \gamma\mathbb{E}[(f(\theta_k^{(\eta)})] - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| > R)]
$$

$$
\geq g(\mathbb{E}[(f(\theta_k^{(\eta)})] - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| \leq R)]) + \gamma\mathbb{E}[(f(\theta_k^{(\eta)})] - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| > R)].
$$

Combing this with previous display gives

$$
0 \leq \mathbb{E}[(f(\theta_k^{(\eta)})] - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| \leq R)] \leq g^{-1}\Big(\frac{2\tilde{M}\eta}{2-\tilde{L}\eta}\Big)
$$

$$
0 \leq \mathbb{E}[(f(\theta_k^{(\eta)})] - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| > R)] \leq \frac{2\tilde{M}\eta}{2-\tilde{L}\eta}.
$$

This implies

$$
0 \leq \pi_\eta(f) - f^* = \mathbb{E}[(f(\theta_k^{(\eta)})] - f^*
$$

$$
= \mathbb{E}[(f(\theta_k^{(\eta)}) - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| \leq R)] + \mathbb{E}[(f(\theta_k^{(\eta)}) - f^*)\mathbb{1}(\|\theta_k^{(\eta)} - \theta^*\| > R)]
$$

$$
\leq g^{-1}\Big(\frac{2\tilde{M}\eta}{2-\tilde{L}\eta}\Big) + \frac{2\tilde{M}\eta}{2-\tilde{L}\eta}.
$$

When the test function ϕ satisfies $\phi = \tilde{\phi} \circ f$ with the $L_{\tilde{\phi}}$-Lipschitz function $\tilde{\phi}$, we obtain

$$|\pi_\eta(\phi) - \phi(\theta^*)| \le L_{\tilde{\phi}}(\pi_\eta(f) - f^*) \le L_{\tilde{\phi}}\left(g^{-1}\left(\frac{2\tilde{M}\eta}{2 - \tilde{L}\eta}\right) + \frac{2\tilde{M}\eta}{2 - \tilde{L}\eta}\right)$$

as desired. □

Proof of Theorem 3.3. Consider the chain $\{\theta_k^{(\eta)}\}_{k \ge 0}$ starting from the stationary distirbution π_η. By display (A.6), it holds that

$$\mathbb{E}[\langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle] \le C_2\eta,$$

where C_2 is a positive constant defined in Theorem 3.1. Note that f is convex, this implies

$$0 \le f(\theta_k^{(\eta)}) - f^* \le \langle \nabla f(\theta_k^{(\eta)}), \theta_k^{(\eta)} - \theta^* \rangle.$$

Taking the expectation on both sides and combing this with the previous display gives

$$0 \le \pi_\eta(f) - f^* \le C_2\eta.$$

The desired result readily follows for the test function ϕ satisfies $\phi = \tilde{\phi} \circ f$ with the $L_{\tilde{\phi}}$-Lipschitz function $\tilde{\phi}$. □