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

Anonymous Author(s)
Affiliation
Address
email

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 [1], 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 [2, 3, 4, 5, 6], deep neural networks [7, 8, 9], and robust empirical risk minimization [10, 11, 12]. 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 [13, 14] or local minimizers [15, 2, 16, 17] of the objective function. From a statistical perspective, *oftentimes* these critical points or local minimizers have nice statistical properties [18, 3, 10, 19, 20, 5]; see also [21] for a counterexample. For the purpose of uncertainty quantification in such non-convex settings, 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.

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

$$\min_{\theta \in \mathbb{R}^d} f(\theta)\,. \tag{1}$$

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

$$\theta_{k+1}^{(\eta)} = \theta_k^{(\eta)} - \eta\big(\nabla f(\theta_k^{(\eta)}) + \xi_{k+1}(\theta_k^{(\eta)})\big), \quad k \geq 0\,, \tag{2}$$

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

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

as the optimization problem in (1), where the function $f(\theta)$ is given by

$$f(\theta) := \int F(\theta, Z)\, dP(Z)\,, \tag{3}$$

for an unknown distribution over the random variable $Z \in \mathbb{R}^p$. The function $F(\theta, Z)$ is typically the loss function composed with functions from the hypothesis class parametrized by $\theta \in \mathbb{R}^d$. In online SGD with batch size b, at each iteration k, b independent samples $Z_j \sim P(Z)$ are used to estimate the true gradient with $\frac{1}{b}\sum_{j=1}^{b} \nabla F(\theta_k^{(\eta)}, Z_j)$. The above iterates are indeed a special case of the iterates in (2), with the noise sequence $\{\xi_{k+1}(\theta_k^{(\eta)})\}_{k\geq 0}$ given by

$$\xi_{k+1}(\theta_k^{(\eta)}) := \frac{1}{b}\sum_{j=1}^{b} \left[\nabla F(\theta_k^{(\eta)}, Z_j) - \nabla f(\theta_k^{(\eta)})\right]. \tag{4}$$

Although proposed in the 1950s by [22], SGD has been the algorithm of choice for training statistical models due to its simplicity, and superior performance in large-scale settings [23, 1, 24, 25]. 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 [26, 27, 28]. On the other hand, it is well-known that the SGD iterates in (2) can be viewed as a Markov chain which allows them to converge to a random vector rather than a single critical point [1]. 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 a flexible 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 [29, 30, 31], but they are not studied in the context of non-convex optimization when the algorithm is SGD. In this work, we study the fluctuations and the bias of the averaged SGD iterates in (2), around the first-order critical points of the minimization problem (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 Theorems 3.1 and 3.2 that, with additional local smoothness assumptions on the non-convex objective function f, we can establish a control over the bias in terms of the 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 statistical learning.

Our results provide algorithm-dependent guarantees for uncertainty quantification, and they could be leveraged to obtain confidence intervals (CIs) 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 [10, 32]. 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 [33, 34, 35, 36, 37], with [26] providing a definitive result for strongly convex objectives. In particular, [26] and [36] 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 [38, 39, 27, 40, 41], leverage the asymptotic normality analysis of [26], and compute CIs for SGD. The benefits of constant step size SGD for faster convergence under overparametrization has also be demonstrated in the works of [42, 43, 44, 45]. The use of Markov chain theory to study constant step size stochastic approximation algorithms has been considered in several works [46, 47, 48, 23, 49, 50]. Recently, [1, 51] investigated the asymptotic variance of the constant step size SGD. We emphasize here that most of the above works assume strongly convex and smooth objective functions.

The non-linear autoregressive (NLAR) process [52, 53, 54] is a specification of our general framework (2) with the noise sequence $\{\xi_k\}_{k\geq 1}$ being a collection of i.i.d mean-zero random vectors with continuous density supported on $\mathbb{R}^d$. However, the methodology for establishing the geometric ergodicity of NLAR [52, 53, 54, 55] is by no means straightforward to carry over to the optimization setting, and does not generalize immediately to the state-dependent noise setup considered in our paper (see Assumption 2.3). In contrast, we establish the geometric ergodicity under easily verifiable assumptions on the objective function using tools from Markov chain theory. Moreover, additional steps are needed to go from geometric ergodicity to CLT results (especially if the chain starts with an arbitrary initial distribution), while we directly obtain a CLT by leveraging the Markov chain structure. Finally, there exists a vast literature on analyzing Langevin diffusion-based sampling algorithms which relies on the much simpler i.i.d. Gaussian noise sequence. We refer the interested reader to [56, 57, 30, 58, 59, 60, 61, 62, 63, 64, 65, 66] 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 (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 $\|\nabla f(\theta)\| \leq L\big(1 + \|\theta\|\big)$ for all $\theta \in \mathbb{R}^d$.*

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. [26, 1]. 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 $\langle \theta, \nabla f(\theta) \rangle \geq \alpha \|\theta\|^2 - \beta$ 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 [67, 29, 31, 68]. 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 non-convex problems that arise in statistical learning such as phase retrieval [50] satisfy Assumption 2.2. We provide examples in Section 4.

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 $t \in \mathbb{R}^d$.*

Assumption 2.3 as formulated above is stronger than what is needed 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 (12). We also emphasize that Assumption 2.3 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 [29, 68, 31].

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

Proposition 2.1 (Ergodicity of SGD)**.** *Let the Assumptions 2.1-2.3 hold, and the step size satisfy*

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

(a) SGD (2) admits a unique stationary distribution $\pi_\eta \in \mathcal{P}_2(\mathbb{R}^d)$, depending on the step size η.

(b) For a test function $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfying $|\phi(\theta)| \le L_\phi(1 + \|\theta\|)$, $\forall \theta \in \mathbb{R}^d$ 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

$$\left|\mathbb{E}\big[\phi\big(\theta_k^{(\eta)}\big)\big] - \pi_\eta(\phi)\right| \le \kappa\, \rho^k (1 + \|\theta_0\|^2)\,, \quad \textit{where } \pi_\eta(\phi) := \int \phi(x) d\pi_\eta(x).$$

The uniqueness of the stationary distribution of the constant step size SGD has been established in [1] 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 [69], which is fundamentally different from the ergodicity analysis in [1]. 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. Note that ρ depends on η and will typically be converging to one if $\eta \to 0$. Thus convergence in Proposition 2.1(b) can be expected to be slower when η becomes smaller. However, smaller η leads to a better control of the asymptotic bias (under additional regularity assumptions), see Theorems 3.1-3.3. Both of those statements (slower convergence for smaller η but smaller bias eventually) are confirmed in our numerical experiments, see Figure 1(d,h).

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.,

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

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 $\sigma^2_{\pi_\eta}(\phi) := \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{\pi_\eta}\big[S_n^2(\phi)\big]$. Then,*

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

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

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

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 [55, Thm 21.2.6]. For the case of strongly convex functions with decreasing step size schedule, it is well-known from the works of [26, 36] that the limiting variance of the averaged SGD iterates achieves the Cramer-Rao lower bound for parameter estimation; see also [70, 28] 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 [71, 72, 10].

There are several important implications of the above CLT for constructing CIs 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 CIs. 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

Proposition 2.1(b) shows that the expectation of a test function evaluated at the k-th iterate converges exponentially fast to the expected value of the stationary distribution π_η. Therefore, a complete characterization of the properties of the SGD requires a control over the asymptotic bias $\pi_\eta(\phi)-\phi(\theta^*)$ for a critical point θ^*. It turns out that this bias behavior is intimately related to the local properties of the objective around its critical points. Therefore, under the mild assumptions that yield the CLT, one cannot expect a tight control over the bias. This section contains three types of bias analyses under different local growth conditions on the objective function f, characterizing the bias behavior in various non-convex and convex settings. We further note that without local regularity conditions, it is still possible to show that the SGD iterates (2) move towards a compact ball containing all critical points exponentially fast; a formal statement of this result along with a corresponding discussion is provided in Proposition B.1, which is deferred to Appendix B. Throughout this section, we make a slightly stronger assumption on the noise sequence.

Assumption 3.1 (Fourth moment of the noise)**.** *Gradient noise sequence $\{\xi_k\}_{k\geq 1}$ satisfies Assumption 2.3, and $\mathbb{E}\big[\|\xi_1(\theta)\|^4\big] \leq L_\xi(1+\|\theta\|^4)$, for any $\theta\in\mathbb{R}^d$, where L_ξ is as in Assumption 2.3.*

Localized Dissipativity Condition: We now introduce the generalized dissipativity condition which, in addition to the quadratic tail growth property enforced in Assumption 2.2, imposes a local growth within some compact region, 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 $0<\eta<c_{L,\alpha}$ for $c_{L,\alpha}$ in (16) admit the stationary distribution $\theta^{(\eta)}\sim\pi_\eta$ which satisfies*

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

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

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

where

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

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 the 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 [1, Lemma 10]. We highlight that [73] proves 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 Lipschits gradients imply $g(x)\leq x^2$ for small x.

Generalized Łojasiewicz Condition: In this section we work with a generalization of the commonly used Łojasiewicz condition in optimization.

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

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

where $\gamma, R>0$, 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 [74], and for $\kappa = 1$, it reduces to the well-known Polyak-Łojasiewicz (PL) inequality [75]. 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 next 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* $\|\nabla^2 f(\theta)\| \le \tilde{L}(1 + \|\theta\|)$, $\forall \theta \in \mathbb{R}^d$ *and some* $\tilde{L} > 0$. *Then, the SGD iterates with a step size satisfying* $0 < \eta < \frac{2}{\tilde{L}} \wedge c_{L,\alpha} \wedge c^\dagger_{L,\alpha} \wedge 1$ *for* $c_{L,\alpha}, c^\dagger_{L,\alpha}$ *in (16) have the stationary distribution* π_η,

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

$$\text{where} \quad M := 12\tilde{L}\big(L + L_\xi^{1/2} + L_\xi^{1/4}\big)^2\Big(1 + m + m^{3/4} + \textstyle\int \|\theta\|^2 \pi_\eta(d\theta)\Big) \ \text{with}$$
$$m := \tfrac{8}{7\alpha}\Big[\big(\beta + 6L^2 + 3L_\xi^{1/2} + 16\big)\textstyle\int \|\theta\|^2 \pi_\eta(d\theta) + 16L^4 + 2L_\xi + 128L^6 + 8L_\xi^{3/2}\Big]\,.$$

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

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

For smooth objectives with Lipschitz gradient, [75] provides a linear rate under the PL-inequality (see also [76, Lemma 2]), which yields the asymptotic bias $|\pi_\eta(\phi) - \phi(\theta^*)| \le \mathcal{O}(\eta)$. The above result recovers their findings as a special case, and provides a considerable generalization.

Convexity: To make the analysis of constant step size SGD complete, we digress from the main theme of this paper and consider this algorithm 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* $0 < \eta < c_{L,\alpha}$ *for* $c_{L,\alpha}$ *in (16) admit the stationary distribution* π_η*, and for a minimizer* θ^* *it satisfies*

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

for C *in (5). Further, if the test function is given as* $\phi = \tilde{\phi} \circ f$ *for a* $L_{\tilde{\phi}}$*-Lipschitz function* $\tilde{\phi}$ *, then,*

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

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 $\langle \nabla f(\theta), \theta - \theta^* \rangle \ge f(\theta) - f(\theta^*)$, and in the case of PL-inequality $\gamma^{-1}\|f(\theta)\|^2 \ge f(\theta) - f(\theta^*)$.

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: $\theta^{(\eta)}_{k+1} = \theta^{(\eta)}_k - \frac{\eta}{b_k}\sum_{j=1}^{b_k} \nabla F(\theta^{(\eta)}_k, Z_j)$, for $k \ge 0$, with independent samples $Z_j \sim P(Z)$ used to estimate the true gradient in each iteration k ; and also the semi-stochastic setting, where the noise sequence $\{\xi_k(\theta)\}_{k\ge1}$ is independent of θ and is simply a sequence of i.i.d. random vectors – such a setting helps verifying our assumptions more 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 [77]. Indeed in fields like finance, oftentimes the Student's t-distribution is used to model the heavy-tailed error [78]. In this case, defining the random vector $Z := (X, Y)$, the stochastic optimization problem in (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 $\lambda > 0$. 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

$$f(\theta) := \tfrac{1}{2m}\textstyle\sum_{i=1}^{m} \log\big(1 + (y_i - \langle \mathbf{x}_i,\, \theta\rangle)^2\big) + \tfrac{\lambda}{2}\|\theta\|^2. \tag{6}$$

We consider the finite-sum setup and we verify our assumptions and empirically demonstrate our results on CLT as well as the bias in a clean manner.

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 distributed (df $= 10$) noise. We choose $m = 5000$, $d = 10$, and the Lipschitz 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 $\theta_0 = (1, \ldots, 1)^\top$ and $\theta_0' = (1.5, \ldots, 1.5)^\top$ 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 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 as one can verify that the assumptions of Theorem 2.1 are satisfied.

Lemma 4.1. *The objective function* (6) *satisfies Assumptions 2.1 and 2.2. Further, Assumption 2.3 is also satisfied with the Student-t distributed (df* $= 10$*) noise.*

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 (6) to another non-convex problem

$$f(\theta) := \tfrac{1}{2}\log\big(1 + \|\theta\|^2\big) + \tfrac{\lambda}{2}\|\theta\|^2\,. \tag{7}$$

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 – indeed, a smaller asymptotic bias for a smaller step size η is predicted by Theorem 3.2 while slower convergence can be expected given the discussion after Proposition 2.1. The following lemma proves that our assumptions are satisfied for this objective.

Lemma 4.2. *The objective function* (7) *is non-convex when* λ *is sufficiently small, and it satisfies Assumptions 2.1,2.2, and 3.3. Further, Assumption 3.1 is also satisfied for this example.*

4.1.2 Online Stochastic Gradient Descent

For our online SGD experiments, we use $b_k = 2$, for all k to obtain the stochastic gradient. We also experimented with $m_k = 1, 10, 50$ 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 $\text{Unif}(0,1)$, 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 $\theta_0 = (1, \ldots, 1)^\top$ and $\theta_0^{'} = (2.5, \ldots, 2.5)^\top$ 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 dependence 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. Indeed, all these observations are as predicted by our theory.

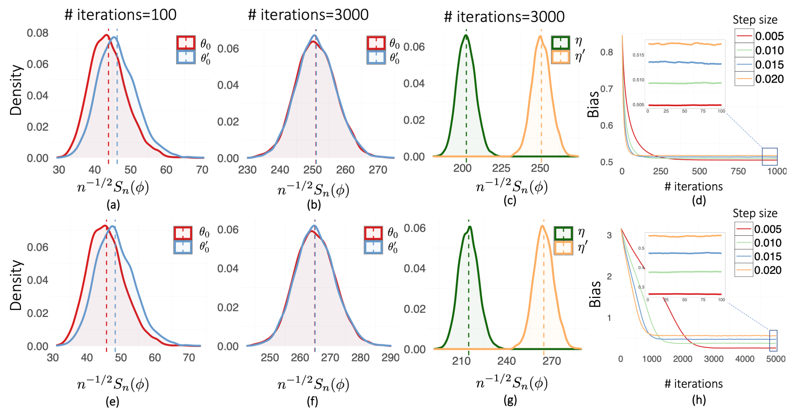

Figure 1: First and second rows correspond to non-convex examples in Sections 4.1.1 and 4.2.1, respectively. Figures (a,b), (e,f) show the density of $n^{-1/2}S_n(\phi) = n^{-1/2}\sum_{k=1}^{n}\phi(\theta_k^{(\eta)})$ with different initializations (red, blue) for different number of iterations. Figures (c,g) show the same density with different step sizes. Figures (d,h) show the evolution of bias against iterations.

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 [79] for details. In this case, defining the random vector $Z := (X, Y)$, the stochastic optimization problem in (3) is given by the expectation of the function $F(Z,\theta) := \log\left(\nu + e^{-(Y-\langle X,\theta\rangle)^2}\right) + \frac{\lambda}{2}\|\theta\|^2$, for $\nu > 0$. 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

$$f(\theta) = -\tfrac{1}{2m}\sum_{i=1}^{m}\log\left(\nu + e^{-(y_i-\langle \mathbf{x}_i,\theta\rangle)^2}\right) + \tfrac{\lambda}{2}\|\theta\|^2, \quad \nu > 0\,. \tag{8}$$

4.2.1 Semi-stochastic Gradient Descent

Asymptotic normality: In the experiments, we use the same setup and parameters as in Section 4.1.1. Fig 1-(e,f,g) demonstrates the asymptotic normality of the SGD with heavy-tailed gradient noise Student-t($\mathrm{df} = 6$). 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.

Lemma 4.3. *The objective function* (8) *satisfies Assumptions 2.1, 2.2. Further, Assumption 2.3 is also satisfied with the Student-t (df* $= 10$*) noise.*

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

$$f(\theta) := -\tfrac{1}{2}\log\left(\nu + e^{-\|\theta\|^2}\right) + \tfrac{\lambda}{2}\|\theta\|^2, \quad \nu > 0\,. \tag{9}$$

We observe in Fig.1-h that smaller step sizes lead to smaller asymptotic bias. One can verify that this can be predicted from Theorem 3.1.

Lemma 4.4. *The objective function* (9) *is non-convex when* λ *is sufficiently small, and it satisfies Assumptions 2.1 and 3.2. Further, Assumption 3.1 is also satisfied for this example.*

4.2.2 Online Stochastic Gradient Descent

Asymptotic normality: In the experiments, we use the same setup as in Section 4.1.2. 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 $\theta_0'' = (1.5, \ldots, 1.5)$

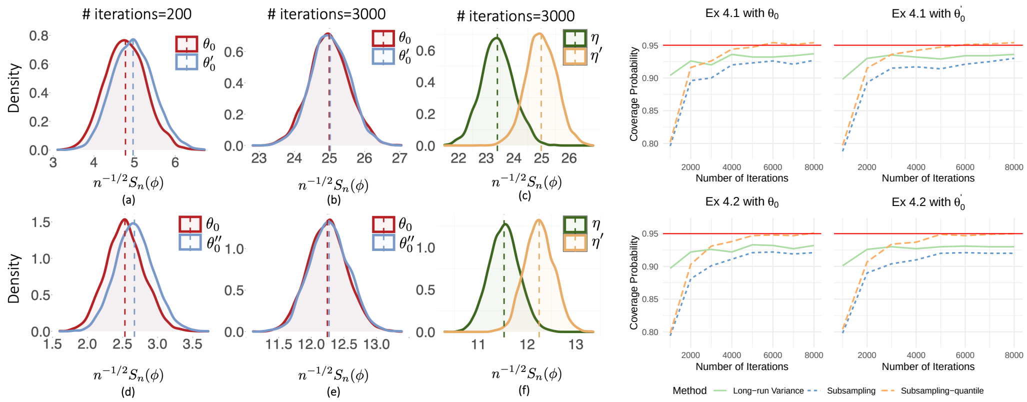

Figure 2: **Left:** First and second rows correspond to non-convex examples in Sections 4.1.2 and 4.2.2, respectively. Figures (a,b), (d,e) show the density of $n^{-1/2}S_n(\phi) = n^{-1/2}\sum_{k=1}^{n}\phi(\theta_k^{(\eta)})$ with different initializations (red, blue) for different number of iterations. Figures (c,f) show the same density with different step sizes. **Right:** Coverage probabilities for Subsampling quantile, Subsampling var, and Long-run var as functions of the number of iterations. Subsampling quantile method outmatches the others in terms of coverage probability and achieves the nominal level with larger iterations.

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 [1], we provided theoretical results characterizing the fluctuations and bias of SGD for non-convex and non-smooth optimization which arises frequently in statistical learning.

Estimating the Asymptotic Variance: As discussed in Section 2, in order to use the established CLT to compute CIs 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 asymptotic variance $\sigma^2_{\pi_\eta}(\phi)$ can be estimated by adopting the online approach of [80] to the constant step size setting. The variance $\sigma^2_{\pi_\eta}(\phi)$ can also be estimated by the Newey-West long-run variance estimation [81, 82] or empirical variance estimation based on sub-sampling [83, Sections 4.2 and 4.6] for a single trajectory.
- 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 [41, 39, 84] for the constant step size SGD setting in order to obtain estimates for $\pi_\eta(\phi)$ and $\sigma^2_{\pi_\eta}(\phi)$.

As a confirmation of the practicability of constructing CIs, we provide preliminary experimental results for constructing CIs with minibatch SGD. We consider the data generation setup described in Sections 4.1 and 4.2 with step size 0.3, and run online SGD with batch size 2. In each run, the first 200 values are discarded. CIs are constructed for each trajectory based on sub-sampling with empirical CDF (Subsampling quantile) and variance estimation (Subsampling var) [83, Sections 4.2 and 4.6], and Newey-West long-run variance estimation (Long-run var) with data-driven bandwidth selection [81, 82]. Empirical coverage results as a function of iteration numbers (nominal level = 95%, 4000 Monte Carlo replications) for the three methods and different initializations ($\theta_0 = (1, \dots, 1)^\top$ and $\theta_0' = (1.5, \dots, 1.5)^\top$) are reported in Figure 2 (right). A non-asymptotic justification of the relative merits of the above variance estimation approaches are left as future work.

References

[1] 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. (Cited on pages 1, 2, 3, 4, 5, 9, and 16.)

[2] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015. (Cited on page 1.)

[3] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016. (Cited on page 1.)

[4] 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. (Cited on page 1.)

[5] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. (Cited on page 1.)

[6] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1861–1872, 2019. (Cited on page 1.)

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. (Cited on page 1.)

[8] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017. (Cited on page 1.)

[9] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13132–13143, 2019. (Cited on page 1.)

[10] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017. (Cited on pages 1, 2, and 4.)

[11] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018. (Cited on page 1.)

[12] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018. (Cited on page 1.)

[13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. (Cited on page 1.)

[14] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018. (Cited on page 1.)

[15] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. (Cited on page 1.)

[16] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017. (Cited on page 1.)

[17] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in neural information processing systems*, pages 2899–2908, 2018. (Cited on page 1.)

[18] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016. (Cited on page 1.)

[19] Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. In *Conference on Learning Theory*, pages 1476–1515, 2017. (Cited on page 1.)

[20] Andreas Elsener and Sara van de Geer. Sharp oracle inequalities for stationary points of nonconvex penalized m-estimators. *IEEE Transactions on Information Theory*, 65(3):1452–1472, 2018. (Cited on page 1.)

[21] DA Freedman and P Diaconis. On inconsistent m-estimators. *The Annals of Statistics*, 10(2):454–461, 1982. (Cited on page 1.)

[22] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. (Cited on page 2.)

[23] Jean-Claude Fort and Gilles Pages. Asymptotic behavior of a markovian stochastic algorithm with constant step. *SIAM journal on control and optimization*, 37(5):1456–1482, 1999. (Cited on page 2.)

[24] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017. (Cited on page 2.)

[25] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413, 2018. (Cited on page 2.)

[26] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. (Cited on pages 2, 3, and 4.)

[27] John Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *Arxiv Preprint*, 2018. (Cited on page 2.)

[28] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, pages 115–137, 2019. (Cited on pages 2 and 4.)

[29] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017. (Cited on pages 2, 3, and 4.)

[30] 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. (Cited on pages 2 and 3.)

[31] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018. (Cited on pages 2, 3, and 4.)

[32] 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. (Cited on page 2.)

[33] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954. (Cited on page 2.)

[34] Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958. (Cited on page 2.)

[35] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968. (Cited on page 2.)

[36] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. (Cited on pages 2 and 4.)

[37] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, 17(2):841–858, 1989. (Cited on page 2.)

[38] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan. Averaging stochastic gradient descent on riemannian manifolds. *arXiv preprint arXiv:1802.09128*, 2018. (Cited on page 2.)

[39] Weijie Su and Yuancheng Zhu. Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018. (Cited on pages 2 and 9.)

[40] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017. (Cited on page 2.)

[41] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018. (Cited on pages 2 and 9.)

[42] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013. (Cited on page 2.)

[43] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014. (Cited on page 2.)

[44] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334, 2018. (Cited on page 2.)

[45] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019. (Cited on page 2.)

[46] Yuri Kifer. Random perturbations of dynamical systems. *Nonlinear Problems in Future Particle Accelerators. World Scientific*, page 189, 1988. (Cited on page 2.)

[47] Michel Benaim. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996. (Cited on page 2.)

[48] P Priouret and A Yu Veretenikov. A remark on the stability of the lms tracking algorithm. *Stochastic analysis and applications*, 16(1):119–129, 1998. (Cited on page 2.)

[49] 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(3):872–899, 2000. (Cited on page 2.)

[50] 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. (Cited on pages 2 and 3.)

[51] Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics*, pages 1476–1485, 2018. (Cited on page 2.)

[52] Rabi Bhattacharya and Chanho Lee. On geometric ergodicity of nonlinear autoregressive models. *Statistics & Probability Letters*, 22(4):311–315, 1995. (Cited on page 3.)

[53] Eckhard Liebscher. Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689, 2005. (Cited on page 3.)

[54] HZ An and SG Chen. A note on the ergodicity of non-linear autoregressive model. *Statistics & probability letters*, 34(4):365–372, 1997. (Cited on page 3.)

[55] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018. (Cited on pages 3, 4, and 19.)

[56] Arnak S Dalalyan and Avetik G Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017. (Cited on page 3.)

[57] Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *COLT*, 2017. (Cited on page 3.)

[58] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. (Cited on page 3.)

[59] 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(3):651–676, 2017. (Cited on page 3.)

[60] 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. (Cited on page 3.)

[61] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 2018. (Cited on page 3.)

[62] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018. (Cited on page 3.)

[63] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018. (Cited on page 3.)

[64] Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic runge-kutta accelerates langevin monte carlo and beyond. In *Advances in Neural Information Processing Systems*, pages 7748–7760, 2019. (Cited on page 3.)

[65] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019. (Cited on page 3.)

[66] 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. (Cited on page 3.)

[67] 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(2):185–232, 2002. (Cited on pages 3 and 19.)

[68] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018. (Cited on pages 3 and 4.)

[69] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. (Cited on pages 4, 18, 19, and 20.)

[70] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011. (Cited on page 4.)

[71] Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 22(4):1993–2010, 1994. (Cited on page 4.)

[72] Alexander Shapiro. On the asymptotics of constrained local m-estimators. *Annals of statistics*, pages 948–960, 2000. (Cited on page 4.)

[73] 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. (Cited on page 5.)

[74] Bin Gao, Xin Liu, Xiaojun Chen, and Ya-xiang Yuan. On the $\{\backslash L\}$ ojasiewicz exponent of the quadratic sphere constrained optimization problem. *arXiv preprint arXiv:1611.08781*, 2016. (Cited on page 6.)

[75] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016. (Cited on page 6.)

[76] Soham De, Abhay Yadav, David Jacobs, and Tom Goldstein. Automated inference with adaptive batches. In *Artificial Intelligence and Statistics*, pages 1504–1513, 2017. (Cited on page 6.)

[77] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004. (Cited on page 6.)

[78] Jianqing Fan and Qiwei Yao. *The elements of financial econometrics*. Cambridge University Press, 2017. (Cited on page 6.)

[79] Andrew Blake and Andrew Zisserman. *Visual reconstruction*. 1987. (Cited on page 8.)

[80] 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. (Cited on page 9.)

[81] Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. Technical report, National Bureau of Economic Research, 1986. (Cited on page 9.)

[82] Whitney K Newey and Kenneth D West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994. (Cited on page 9.)

[83] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999. (Cited on pages 9 and 30.)

[84] 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(1):251–273, 2020. (Cited on page 9.)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 2 and 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices A, B and C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]