Proceedings of Machine Learning Research vol ???:1–13, 2025

Stable convergence of Stochastic Gradient Descent for non-convex objectives

Soham Bonnerjee SOHAMBONNERJEE@UCHICAGO.EDU
University of Chicago

Yuefeng Han YUEFENG.HAN@ND.EDU
University of Notre Dame

Wei Biao Wu WBWU@UCHICAGO.EDU
University of Chicago

Abstract

Stochastic gradient descent (SGD) is a popular algorithm for large-scale data estimation due to its efficiency in computation and memory usage. While most research has focused on convergence rates and asymptotic distributions in convex optimization, this work addresses the stable convergence of true model parameters for SGD variants in non-convex optimization settings. Our contributions are twofold. First, we derive stable convergence results for a broad class of SGD variant iterates under a general non-convex framework. Second, we introduce a Gaussian mixture model-based algorithm to analyze the endpoints of SGD chains, enabling the identification of distinct local optima. These methods have broad practical applications. For instance, by leveraging local optima, the global minimum can be identified through empirical risk, making this approach highly relevant for tackling complex learning problems in modern data science.

Keywords: Stochastic gradient descent (SGD); non-convex optimization; stable convergence

1. Introduction

Estimating model parameters through objective function minimization is central to modern data science. Let $x^\star \in \mathbb{R}^d$ represent the true d-dimensional model parameters. In many models, $x^\star$ minimizes an objective function $F : \mathbb{R}^d \to \mathbb{R}$, formally expressed as:

$$x^\star = \underset{x\in\mathbb{R}^d}{arg\,min}\, F(x), \quad F(x) := \mathbb{E}_{\xi\sim\Pi}[f(x,\xi)], \tag{1.1}$$

where $\xi \in \mathbb{R}^m$ is a random variable drawn from the probability distribution Π, and $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a loss function tailored to the problem at hand. Here, both f and F are assumed to be continuously differentiable with respect to x.

Stochastic optimization methods, pioneered by Kiefer and Wolfowitz (1952) and Lai (2003), are widely applied in scenarios with large datasets or sequential data, such as search queries and transactions. Among these methods, the Robbins-Monro algorithm (Robbins and Monro, 1951; Lai, 2003), commonly known as stochastic gradient descent (SGD), is the most widely used method, particularly in machine learning and statistics. Starting with an initial point x_0, SGD updates iteratively as follows:

$$x_n = x_{n-1} - \eta_n \nabla f(x_{n-1}, \xi_n),\ n = 1, 2, \ldots. \tag{1.2}$$

Here, η_n is a step size that typically decreases over iterations, ξ_n is the n-th random sample from distribution Π, and $\nabla f(x_{n-1}, \xi_n)$ is the gradient of $f(x, \xi_n)$ evaluated at x_{n-1}. The solution can be

either the final iterate or the average of all iterates. SGD offers significant computational and storage benefits compared to traditional deterministic optimization methods. Each iteration requires only one data sample, yielding a per-iteration time complexity of $O(d)$, independent of the dataset size. Furthermore, SGD avoids storing the entire dataset, making it naturally suited for online settings where data arrives in sequence Zhu et al. (2023). It is now the dominant optimization approach for machine learning tasks (Rumelhart et al., 1986; Spall, 2000; Nesterov and Vial, 2008; Nemirovski et al., 2008; Zhou et al., 2019), such as training deep neural networks.

Building on the vanilla SGD framework in Eq. (1.2), numerous variants have been proposed in optimization and statistical learning. Most prior studies emphasize computational convergence for the objective function, or analyze the asymptotic behavior of the solution relative to the true minimizer $x^\star$ in Eq. (1.1), focusing primarily on convex functions. However, assumptions such as strong convexity or the Polyak-Lojasiewicz (PL) conditions are often too restrictive for practical deep learning problems. As noted by Bruna et al. (2017): "While SGD has been rigorously analyzed only for convex loss functions (Schmidt et al., 2017), in deep learning the loss is a non-convex function of the network parameters, hence there are no guarantees that SGD finds the global minimizer." Identifying global minima in complex non-convex optimization remains an unresolved and challenging problem in practice.

Recently, progress has been made in analyzing local convergence properties of SGD variants under non-convex settings, as shown in works like Fehrman et al. (2020); Lei et al. (2020); Ko and Li (2023); An and Lu (2023). These studies often focus on metrics such as the convergence of $\mathbb{E}[|\nabla F(x_n)|]$ or almost sure local convergence of x_n to a distribution over critical points. Despite these advances, stable convergence for local optima in non-convex problems remains largely underexplored, particularly for accelerated variants of SGD. Stable convergence (Hall and Heyde, 1980), a central concept in statistics and probability, is critical for quantifying uncertainty in parameter estimation, especially when the estimator converges to a random variable over a set of parameters rather than to a specific parameter. In this paper, we establish stable convergence results for a range of SGD variants under broad non-convex conditions, demonstrating asymptotic normality around each local optimum. These advancements surpass the capabilities of traditional methods relying on deviation inequalities or generalization error bounds.

In addition to vanilla SGD, this paper examines the momentum-assisted version of SGD (m-SGD), one of the most widely used variants inspired by Polyak's Stochastic Heavy Ball (SHB) method (Poljak, 1964; Gadat and Panloup, 2023). The most common implementation of m-SGD uses a constant step size for momentum, represented as follows:

$$\begin{aligned} v_n &= \beta v_{n-1} + \eta_n \nabla f(x_{n-1}, \xi_n), \\ x_n &= x_{n-1} - v_n. \end{aligned} \tag{1.3}$$

Compared to vanilla SGD, the convergence analysis of m-SGD for non-convex optimization problems remains extremely limited. For a fixed step size ($\eta_n = \eta$), Yu et al. (2019) derived an upper bound for the convergence rate $O((n\eta)^{-1} + \eta(1-\beta)^{-1})$. However, when β approaches 1, the error rate becomes significantly large, failing to fully account for m-SGD's practical competitiveness. In this paper, we address this gap by providing a stable convergence analysis for m-SGD. Our results not only refine the error convergence rate but also highlight a crucial trade-off: the balance between convergence error and m-SGD's capacity to escape saddle points. To the best of our knowledge, this trade-off has not been explicitly explored in existing literature.

Building on these results, multiple SGD runs with moderate iterations can produce outputs that can be effectively represented as a novel Gaussian mixture model, where each cluster mean corresponds to a local minimum. This approach enables the estimation of means of the Gaussian mixture models, which not only helps in distinguishing between different local minima, but also plays a crucial role in identifying the global minimum. The formal procedure is outlined in Algorithm 1.

Algorithm 1 Identifying the Local Optimum Using SGD

Require: Step-size parameters η, α, β; noisy gradient $\nabla f(x, \xi)$; i.i.d. samples $\xi_1, \ldots, \xi_T$; compact set K containing all critical points of $\mathbb{E}[f]$; number of SGD chains B; number of iterations T.

Ensure: Cluster means as estimates of the local optima.

1: Initialize $x_{0,b}$ independently and uniformly within V, for all $b \in [B]$.
2: **for** $b = 1$ *to* B **do**
    for $i = 1$ *to* T **do**
        Update $x_{i,b}$ using either vanilla SGD (1.2) or m-SGD variants (1.3), (3.3).
    end
    end
3: Apply a Gaussian mixture model estimation method to the dataset $(x_{T,b})_{b\in[B]}$ to obtain cluster means and cluster variances.
4: **Output:** Cluster means serve as estimates of the local minima, defined as: $J := \{x : \mathbb{E}[\nabla f(x, \xi)] = 0, \nabla^2 \mathbb{E}[f(x, \xi)] \succ 0\}, \quad J \subseteq V$. Cluster variances estimate the variance of the endpoint iterates corresponding to each identified local optimum.

1.1. Our contribution

We summarize our main results and contributions as follows:

- First, for a fixed initialization x_0, we establish stable convergence results for SGD and m-SGD iterates under a general non-convex setting. Specifically, we show:

$$\eta_n^{-1/2}(x_n(x_0) - a)|\{x_n(x_0) \to a\} \xrightarrow{w} N(0, \Sigma(a)), \ n \to \infty, \tag{1.4}$$

  where a represents a local minimum and $\Sigma(a)$ is a covariance matrix. To the best of our knowledge, this represents the first stable convergence analysis for accelerated variants of SGD in non-convex optimization.

- Second, by leveraging the stable convergence results, we can identify local minima. For sufficiently large iterations T and B independent initializations $x_{0,b}, b \in [B]$, we construct a dataset of endpoints $\{x_T(x_{0,b})\}_{b\in[B]}$. By applying a Gaussian mixture model-based clustering algorithm, we can effectively distinguish the distinct local minima, as demonstrated in Algorithm 1. The global optimum then may be determined as the estimated local optimum with the smallest empirical risk.

The validity of our algorithm is immediate following (1.4). For practical validity, we provide a numerical example of a non-convex function, where, our theoretical framework enables the identification of local minima with significantly fewer SGD iterations, greatly reducing the computational burden.

1.2. Some Related Work

The theoretical analysis of SGD and its variants has been extensively studied in the literature. Here, we focus on reviewing the most relevant works. The convergence of vanilla SGD and the asymptotic normality for strictly convex functions have been thoroughly examined in Blum (1954); Wolfowitz (1956); Sacks (1958); Fabian (1968); Ljung (1977, 1984), among others. For SGD with constant step sizes ($\eta_n = \eta$), it is well known that such step sizes can introduce bias (Dieuleveut et al., 2017), so we focus on step sizes η_n that decrease to zero. To quantify the variability of SGD around the minimum, Polyak and Juditsky (1992); Ruppert (1988) introduced averaged SGD (ASGD), which involves averaging the iterates. The asymptotic normality of ASGD iterates was formally established in Polyak and Juditsky (1992) for convex optimization problems. Another notable asymptotic result in the literature is the end-term central limit theorem (CLT), which exhibits an error rate of $O(\sqrt{\eta_n})$ (Fabian, 1968; Sacks, 1958; Chung, 1954).

While there is a rich body of work on vanilla SGD, the analysis of momentum-based SGD (m-SGD) is less extensive, though still substantial. In convex settings, various convergence properties and asymptotic normality results for m-SGD have been investigated in Gitman et al. (2019); Loizou and Richtárik (2020); Tang et al. (2023); Li et al. (2024), among others. However, asymptotic normality in non-convex settings has been largely unexplored, leaving a significant gap in the literature.

1.3. Organization of the paper

The paper is organized as follows: the first part focuses on the theoretical foundations of Algorithm 1. In particular, Section 2 explores the theoretical results for Vanilla SGD, while Section 3 delves into two widely used variants of momentum-SGD. Finally, Section 4 concludes the paper with a summary of key insights and implications. In the appendix, Appendix A presents numerical studies that validate and support the theoretical findings. Detailed mathematical derivations and proofs are provided in Appendix B and C.

1.4. Notation

We interchangeably use $|\cdot|$ and $\|\cdot\|$ to denote the Euclidean norm on $\mathbb{R}^d$; for $a \in \mathbb{R}^d$, $|a| = \|a\| = \sqrt{\sum_{j=1}^d a_j^2}$. The notation $\|\cdot\|_{\mathcal{L}_q}$ refers the $\mathcal{L}_q$ norm of a random variable, defined as $\|X\|_{\mathcal{L}_q} := \mathbb{E}[\|X\|^q]^{1/q}$. For a positive definite matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote, respectively, the smallest and the largest eigenvalues of A. The symbols ∇ and ∇^2 represent the gradient vector and Hessian matrix of a function $f : \mathbb{R}^d \to \mathbb{R}$, respectively. For positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n \lesssim b_n$ means there exists some constant C such that $a_n \leq Cb_n$ for all sufficiently large n. The constant C is a generic placeholder that may vary from line to line. Finally, $\xrightarrow{w}$ denotes convergence in distribution.

2. Convergence of vanilla SGD

In this section, we lay the theoretical foundation for the stable convergence results of the vanilla SGD algorithm (1.2). To effectively address the non-convexity of the problem, we introduce a set of general assumptions. Let $J = \{x \in \mathbb{R}^d : |\nabla F(x)| = 0\}$ represent the set of critical points, and

$J_0 \subset J$ denote the set of local minima. To proceed, we impose the following standard regularity conditions on F and f, which are essential for establishing our results.

Assumption 2.1 *The set J_0 is non-empty and finite. Moreover, the random sample $\xi_i \in \mathcal{L}_p$ for some $p \geq 2$.*

The assumption that J_0 is non-empty is standard in the non-convex optimization literature, and appears in works such as Gitman et al. (2019); Jin et al. (2022) in the context of almost sure convergence of m-SGD and Adagrad. Additionally, we assume that J_0 is finite. While this may seem restrictive, many well-known non-convex problems, such as matrix sensing, phase retrieval, matrix completion, and regression with non-convex constraints, satisfy this condition; see Ge et al. (2015, 2016); Chi et al. (2019); Tong et al. (2021); Cai et al. (2022); Tan and Vershynin (2023). For neural networks, although saddle-point regions of positive measure may exist, it has been demonstrated that algorithms like SGD and m-SGD can escape these regions with arbitrarily high probability Wang et al. (2021); Kleinberg et al. (2018). Therefore, Assumption 2.1 can be reasonably justified for most practical applications.

Furthermore, the optimization landscape of many non-convex statistical problems is often benign, allowing us to introduce additional assumptions regarding the behavior of F near the local minima. Specifically, we make the following two assumptions.

Assumption 2.2 (Local μ-strong convexity) *For each $x \in J_0$, there exists $\gamma(x) > 0$ and $\mu(x) > 0$ such that*

$$\nabla F(y)^\top (y - x) \geq \mu(x) \, |y - x|^2, \text{ for all } y \in \mathbb{R}^d \text{ with } |x - y| < \gamma(x). \tag{2.1}$$

This assumption ensures that $\nabla^2 F(y) \succ 0$ for all $y \in J_0$, meaning all local minimas are *Hurwicz-regular*. Note that, due to Assumption 2.1, we can assume $\gamma(x) \equiv \gamma$ and $\mu(x) \equiv \mu$ for all $x \in J_0$.

Assumption 2.3 (Lipschitz condition) *The function F is assumed to be L-smooth, i.e.,*

$$|\nabla F(x) - \nabla F(y)| \leq L|x - y|, \text{ for all } x, y \in \mathbb{R}^d. \tag{2.2}$$

In the context of statistical inference, similar assumptions are widespread in most theoretical analyses of convex SGD. For non-convex optimization problems, Assumption 2.2 guarantees convergence to a local minimum once the iterates enter a neighborhood (ball) around that specific local minimum. Such localized assumptions are commonly employed in non-convex analysis, such as Yu et al. (2021); Zhong et al. (2023).

Additionally, to ensure that the estimation error of SGD exhibits a quantifiable asymptotic behavior, it is crucial to control the randomized gradient $\nabla f(x_{n-1}, \xi_n)$. To address this, we introduce a *Leibniz* assumption, allowing us to interchange integration and differentiation in the analysis. Furthermore, we impose a mild smoothness condition on the gradient noise, defined as $g(x, \xi) = \nabla F(x) - \nabla f(x, \xi)$.

Assumption 2.4 (Regularity of gradient noise) *The function $f(x, \xi)$ is assumed to be continuously differentiable with respect to x for any fixed ξ, and $\mathbb{E}[|\nabla f(x, \xi)|^2]$ is also continuously differentiable. Additionally, it is assumed that $\mathbb{E}[g(x_{n-1}, \xi_n)|\mathcal{F}_1^{n-1}] = 0$ for any $\mathcal{F}_1^{n-1}$-measurable random variable x_{n-1}, where $\mathcal{F}_k^j = \sigma(\xi_k, \ldots, \xi_j)(j > k)$ represents the σ-field generated by all past samples ξ_i up to step k.*

Furthermore, the gradient noise $g(x,\xi)$ must satisfy $\|g(x,\xi)\|_{\mathcal{L}_2} < \infty$ for all $x \in J$. Moreover, there exists constants $L' > 0$, and $\gamma' > 0$, such that for some $p > 2$,

$$\|g(y,\xi) - g(x,\xi)\|_{\mathcal{L}_p} \leq L'|x-y|, \text{ for all } y \in \mathbb{R}^d \text{ with } |x-y| < \gamma' . \tag{2.3}$$

An attentive reader may note that L' and γ' can be chosen to be the same as L and γ in Assumptions 2.3 and 2.2, respectively. Assumption 2.4 is widely used in the literature and can be found in works such as Wei et al. (2023); Zhu et al. (2023). Building on these assumptions, the following theorem provides a conditional Gaussian approximation for the iterates of vanilla SGD.

Theorem 1 *Suppose that the functions f and F satisfy Assumptions 2.1-2.4. Let $V \subseteq \mathbb{R}^d$ be a closed set containing J_0, and consider an initial point $x_0 \in V$. For the SGD iterates defined in (1.2) with step sizes $\eta_i = \eta i^{-\alpha}$, where $\eta > 0$, and $\alpha \in (1/2, 1)$, there exists a random variable $X(x_0)$ supported on J_0, and a function $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ such that*

$$\Sigma(a)^{-1/2}\eta_n^{-1/2}(x_n - a)|\{X = a\} \xrightarrow{w} N(0, I_d), \text{ as } n \to \infty, \text{ if } a \in J_0. \tag{2.4}$$

Remark 2 *For SGD applied to non-convex objectives with step-sizes $\eta_t \approx t^{-1}$, Sirignano and Spiliopoulos (2020) established a central limit theorem under the assumption of a single global minimum. Similar results appear in Hu et al. (2024) in the context of Markov chains. More recently, Zhong et al. (2023) proved stable convergence results for averaged vanilla SGD iterates. Apart from the weaker assumptions for Theorem 1, another key distinction between these works lies in the asymptotic covariance matrix. For Polyak-Ruppert averaged SGD, the asymptotic covariance matrix, in our notation, is given by $A^{-1}SA^{-1}$. In contrast, for the end-term CLT, the covariance matrix Σ uniquely determined as the solution to $A\Sigma + \Sigma A = S$. Additionally, Zhong et al. (2023) requires multiple local Lyapunov and Lindeberg conditions on the noise gradients $\nabla f(x,\xi)$ to achieve the central limit theorem. In comparison, we impose only a much weaker local smoothness assumption as described in (2.3).*

2.1. Proof of Theorem 1

The classical technique in proving central limit theory for SGD iterates involve some form of "linearization" of the noisy gradient; see Polyak and Juditsky (1992); Li et al. (2024). The main idea behind such argument notes that the noisy gradient can be written as

$$\nabla f(x,\xi) = \nabla F(x) - g(x,\xi)$$

for $x \in \mathbb{R}$. Note that, for an SGD sequence $\{x_n\}_{n\geq 1}$ and i.i.d. random variables $\{\xi_n\}_{n\geq 1}$, $\{g(x_{n-1},\xi_n)\}$ is a martingale difference sequence by virtue of Assumption 2.4. Therefore, $\{g(x_{n-1},\xi_n)\}$ can be controlled by classical arguments, such as martingale CLT (Hall and Heyde, 1980). On the other hand, for strongly convex F, if x_0 is the global minima, one uses $\nabla F(x) \approx \nabla^2 F(x_0)^\top (x - x_0)$ to carry out the linearization. However, in this case, owing to non-convexity of the function F, such arguments are not very straightforward. In fact, there are two major technical roadblocks.

- The presence of multiple local minima nullifies the uniqueness of the above Taylor expansion.
- The convexity holds only when x_i reaches the ball inside which local strong convexity (Assumption 2.2) holds. This suggests a conditional argument. Is the Gaussianity of $\{x_n\}$ still preserved conditional on the event that $\{x_n \to a\}$ for some $a \in J_0$?

The first question is answered by a novel projection argument. In fact, we carry out the linearization through the following intermediary oracle SGD iterates. For some $a \in J_0$, define

$$y_n(x_0) = \Pi_{B(a,\gamma)}(y_{n-1}(x_0) - \eta_n \nabla f(y_{n-1}(x_0), \xi_n)),\ y_0(x_0) = x_0 \in V, \tag{2.5}$$

$$y_n^{(1)}(x_0) = \Pi_{B(a,\gamma)}(y_{n-1}^{(1)}(x_0) - \eta_n \nabla F(y_{n-1}^{(1)}(x_0)) + \eta_n g(a, \xi_n)), \tag{2.6}$$

$$y_n^{(2)}(x_0) = \Pi_{B(a,\gamma)}(y_{n-1}^{(2)}(x_0) - \eta_n A y_{n-1}^{(2)}(x_0) + \eta_n g(a, \xi_n)),\ \text{and}, \tag{2.7}$$

$$y_n^{(3)}(x_0) = y_{n-1}^{(3)}(x_0) - \eta_n A y_{n-1}^{(3)}(x_0) + \eta_n g(a, \xi_n), \tag{2.8}$$

where $A = \nabla^2 F(a)$, and $B(a, \gamma) = \{z : |z - a| \leq \gamma\}$ is the ball centered on a, and for a compact set $A \subseteq \mathbb{R}^d$, $\Pi_A : x \mapsto arg\min_y\{|x - y| : y \in A\}$ denotes the projection onto A. This localization of the iterates enable us to carry out the Taylor series factorization successfully, thereby deducing a Gaussian limit. This argument is driven by a successive estimation of error from y_n to $y_n^{(3)}$. The complete result is formally encapsulated as follows, with the proof delegated to Appendix B.

Proposition 3 *Suppose the functions f and F satisfy Assumptions 2.1-2.4. Fix $a \in J$. Let V be any compact subset of $\mathbb{R}^d$. For some $a \in J$, consider the SGD iterates (2.5). Then for a Borel Measurable set $\mathcal{A} \subseteq \mathbb{R}^d$, and with $\psi_n(x_0) := \Sigma(a)^{-1/2}\eta_n^{-1/2}(y_n(x_0) - a)$, it holds that*

$$\sup_{x_0 \in V} |\mathbb{P}(\psi_n(x_0) \in \mathcal{A}) - \Phi(\mathcal{A})| \to 0,\ \text{as } n \to \infty, \tag{2.9}$$

where $\Phi(\cdot)$ denotes the measure induced by $Z \overset{d}{=} N(0, I_d)$, and $\Sigma(a)$ is same as in Theorem 1.

The solution to the second point we raised involves a careful conditional argument leveraging three related mathematical tools:

- The almost sure convergence of $\{x_n\}$ to one of the critical points, as dictated by Jin et al. (2022).
- Controlling the probability of SGD iterates escaping the ball inside which local strong convexity kicks in; this uses a result of Mertikopoulos et al. (2020). This result allows us to, in a sense, "forget" the conditioning event and directly apply Proposition 3.
- Bear upon Proposition 3 on the two preceding points to deduce Gaussianity.

The complete proof is formally stated below. Let $\varepsilon \in (0, 1)$ be given. For each $a \in J_0$, Assumption 2.2 dictates that Lemma 1 of Mertikopoulos et al. (2020) holds with a ball $B(a, \tau) \subseteq B(a, \gamma)$. Choose $\delta > 0$ such that $4\delta + 2\sqrt{\delta} < \tau$. This choice of δ is governed by equations (D.16) and (D.17) of Mertikopoulos et al. (2020). Indeed, from Theorem 4 of Mertikopoulos et al. (2020), there exists $M_{\varepsilon,0} \in \mathbb{N}$ sufficiently large such that for all $m \geq M_{\varepsilon,0}$,

$$\mathbb{P}(\max_{i \geq m} |x_i - a| \leq 4\delta + 2\sqrt{\delta} \mid |x_m - a| \leq 2\delta) \geq 1 - \varepsilon. \tag{2.10}$$

Subsequently we will fix $a \in J_0$. Fix $m \geq M_\varepsilon$, whose particular choice will be mentioned later. For some $b \in \mathbb{R}$, define the projected oracle sequence

$$y_n(b) = \Pi_{B(a,\gamma)}(y_{n-1}(b) - \eta_n \nabla f(y_{n-1}(b), \xi_n)), n \geq m,\ y_m(b) = b.$$

It follows easily from (2.10) that,

$$\begin{aligned}&\mathbb{P}(\eta_n^{-1/2}|x_n - y_n(x_m)| > 0 \mid |x_m - a| \le 2\delta)\\ &\le \varepsilon + \mathbb{P}(\eta_n^{-1/2}|x_n - y_n(x_m)| > 0 \mid \max_{i\ge m}|x_i - a| \le 4\delta + 2\sqrt{\delta}, |x_m - a| \le 2\delta),\end{aligned} \tag{2.11}$$

and the probability in the final expression in (2.11) is exactly zero, since given $\max_{i\ge m}|x_i - a| \in B(0,\gamma)$, $x_n = y_n(x_m)$ holds for all $n \ge m$. Hence, one has

$$\sup_{m,n:M_{\varepsilon,0}\le m\le n} \mathbb{P}(\eta_n^{-1/2}|x_n - y_n(x_m)| > 0 \mid |x_m - a| \le 2\delta) < \varepsilon. \tag{2.12}$$

Let g_n be such a function that $g_n(b, \mathcal{F}_{m+1}^n) := \Sigma(a)^{-1/2}\eta_n^{-1/2}(y_n(b) - a)$. Proposition 3 reveals that all convergence statements therein hold uniformly over the set $|b - a| \le 2\delta$; i.e., given a Borel-measurable set $E \subseteq \mathbb{R}^d$, there exists $N_{\varepsilon,m} \ge m$ such that for all $n \ge N_{\varepsilon,m}$ it holds

$$|\mathbb{P}(g_n(x_m, \mathcal{F}_{m+1}^n) \in E \mid |x_m - a| \le 2\delta) - \Phi(E)| \le \sup_{|b-a|\le 2\delta} |\mathbb{P}(g_n(b, \mathcal{F}_{m+1}^n) \in E) - \Phi(E)| < \varepsilon, \tag{2.13}$$

where $\Phi(\cdot)$ denotes the measure induced by $Z \overset{d}{=} N(0, I_d)$ on $\mathbb{R}^d$. From (2.12) and (2.13) one obtains that for all $n \ge N_{\varepsilon,m}$ with $m \ge M_{\varepsilon,0}$,

$$|\mathbb{P}(\Sigma(a)^{-1/2}\eta_n^{-1/2}(x_n - a) \in E \mid |x_m - a| \le 2\delta) - \Phi(E)| < 2\varepsilon. \tag{2.14}$$

Theorem 1 of Jin et al. (2022) instructs that there exists a discrete random variable X, taking values on J, such that $x_n \overset{a.s.}{\to} X$. We assume $\mathbb{P}(X = a) > 0$, otherwise the assertion in Theorem 1 is trivial. For each $z \in J$, choose $M_{\varepsilon,z}$ such that $\mathbb{P}(|x_{M_{\varepsilon,z}} - z| \le 2\delta \mid X = z) \ge 1 - \varepsilon$. Choose $M_\varepsilon = \max\{M_{\varepsilon,0}, \max_{z\in J} M_{\varepsilon,z}\}$, and $N_\varepsilon = N_{\varepsilon,M_\varepsilon}$. Assumption 2.1 guarantees the existence of a finite M_ε. Clearly, for all $n \ge N_\varepsilon$ with $\psi_X(n,a) := \{\Sigma(a)^{-1/2}\eta_n^{-1/2}(x_n - a) \in E\}$, the following series of inequalities hold.

$$\begin{aligned}&|\mathbb{P}(\psi_X(n,a) \mid X = a) - \Phi(E)|\\ \le& 2\varepsilon + |\mathbb{P}(\psi_X(n,a) \mid X = a) - \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta)|\\ \le& 4\varepsilon + |\mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta, X = a) - \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta)|\end{aligned} \tag{2.15}$$

where the first inequality follows from (2.14) and the second inequality is due to our choice of M_ε. Now, to tackle the second term in (2.15), write

$$\mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta) = \sum_{b\in J} \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta, X = b)\mathbb{P}(X = b \mid |x_{M_\varepsilon} - a| \le 2\delta).$$

Therefore,

$$\begin{aligned}\mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta) \le& \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \le 2\delta, X = a)\\ &+ \sum_{b\in J, b\ne a} \mathbb{P}(X = b \mid |x_{M_\varepsilon} - a| \le 2\delta).\end{aligned} \tag{2.16}$$

Since $B(a,\gamma)$ and $B(b,\gamma)$ are disjoint for $a \neq b$, by virtue of our choice of M_ε it must be true that $\mathbb{P}(|x_{M_\varepsilon} - a| \leq 2\delta \mid X = b) \leq \mathbb{P}(|x_{M_\varepsilon} - b| > 2\delta \mid X = b) < \varepsilon$ for each $b \neq a$. This implies that $\mathbb{P}(X = b \mid |x_{M_\varepsilon} - a| \leq 2\delta) \leq \frac{\varepsilon}{1-\varepsilon}\frac{\mathbb{P}(X=b)}{\mathbb{P}(X=a)}$ for $b \neq a$. Hence, from (2.16), we deduce

$$\mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \leq 2\delta) \leq \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \leq 2\delta, X = a) + \frac{\varepsilon}{1-\varepsilon}\frac{\mathbb{P}(X \neq a)}{\mathbb{P}(X = a)}. \quad (2.17)$$

On the other hand, equation (2.10) along with our choice of $M_\varepsilon \geq M_{\varepsilon,0}$ imply that $\mathbb{P}(X = a \mid |x_{M_\varepsilon} - a| \leq 2\delta) \geq 1 - \varepsilon$. Therefore,

$$\begin{aligned}\mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \leq 2\delta) \geq& \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \leq 2\delta, X = a)\mathbb{P}(X = a \mid |x_{M_\varepsilon} - a| \leq 2\delta) \\ \geq& \mathbb{P}(\psi_X(n,a) \mid |x_{M_\varepsilon} - a| \leq 2\delta, X = a) - \varepsilon. \quad (2.18)\end{aligned}$$

In view of (2.15), (2.17) and (2.18) jointly imply that

$$|\mathbb{P}(\psi_X(n,a) \mid X = a) - \Phi(E)| \leq 4\epsilon + \max\{\varepsilon, \frac{\varepsilon}{1-\varepsilon}\frac{\mathbb{P}(X \neq a)}{\mathbb{P}(X = a)}\} \quad (2.19)$$

for all $n \geq N_\varepsilon$. Finally, (2.19) shows that $\Sigma(a)^{-1/2}\eta_n^{-1/2}(x_n - a)|\{X = a\} \xrightarrow{w} Z = N(0, I_d)$, which completes the proof of Theorem 1.

3. Analysis of SGD with momentum

In this section, we extend our stable convergence results to two different forms of momentum-SGD (m-SGD). The motivation for analyzing m-SGD variants arises from the observation that vanilla SGD is rarely employed in practical applications. Instead, momentum-based variants are commonly used due to their ability to escape saddle points more effectively and achieve significantly faster convergence rates.

We start by examining the constant-momentum version of SGD, as described in (1.3). This variant incorporates a momentum term that helps smooth out the updates, enhancing its robustness in non-convex landscapes. When the momentum parameter β is not excessively large, we establish a stable convergence result analogous to Theorem 1 for this momentum-enhanced variant.

Theorem 4 *Suppose the functions f and F satisfy Assumptions 2.1-2.4. Let $V \subseteq \mathbb{R}^d$ be a closed set containing J_0, and consider an initial point $t_0 = (v_0, x_0) \in V \times V$. For the m-SGD iterates defined in (1.3), assume step sizes $\eta_i = \eta i^{-\alpha}$ with $\eta > 0$, $\alpha \in (1/2, 1)$, and let the momentum parameter β satisfy the condition*

$$0 \leq \beta < \frac{\mu^2}{2L^2 + \mu^2} \wedge \min_{y \in J}\left\{1 + \frac{1}{2\kappa(\nabla^2 F(y))} - \sqrt{1 + \frac{1}{4\kappa^2(\nabla^2 F(y))}}\right\}, \quad (3.1)$$

where $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ is the condition number of the matrix A. Then, there exists a random variable $X(x_0)$ supported on J_0, such that

$$\sqrt{1-\beta}\Sigma(a)^{-1/2}\eta_n^{-1/2}(x_n - a)|\{X = a\} \xrightarrow{w} N(0, I_d), \text{ as } n \to \infty, \text{ if } a \in J_0. \quad (3.2)$$

where $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is same as that in Theorem 1.

This result extends the stable convergence framework to m-SGD, demonstrating that, under appropriate conditions, the method retains theoretical guarantees while incorporating momentum. The proof of this result is included in Appendix C.

Remark 5 *The technical restriction on β warrants further discussion. The term $\frac{\mu^2}{2L^2+\mu^2}$ reflects an interplay between the local convexity μ and smoothness L of the function F. Specifically, achieving a large β requires a large μ and a comparatively small L. This implies that the corresponding local minimum must be a strong attractor, and the function must exhibit low volatility in its vicinity—both of which are reasonable assumptions in many practical scenarios. On the other hand, the condition involving $\kappa(\nabla^2 F(y))$, the condition number of the Hessian at a local minimum y, ensures the regularity of the local minima, further supporting the stability and convergence of the iterates.*

Remark 6 *More importantly, it is worth noting that as β increases, the asymptotic variance grows proportionally to $(1-\beta)^{-1}$. In other words, the asymptotic variance of m-SGD is higher than that of vanilla SGD when $\beta > 0$. At first glance, this may seem counterintuitive, especially given the numerous practical examples where m-SGD demonstrates faster convergence. However, this discrepancy highlights an important trade-off, as mentioned earlier in the introduction. As pointed out in Darken and Moody (1991); Wiegerinck et al. (1994); Sutskever et al. (2013), the primary impact of momentum occurs during the initial "transient" phase of the SGD process. This is also reflected in the ability of m-SGD to escape saddle points with higher probability than vanilla SGD (Xu et al., 2018; Wang et al., 2021). A larger β allows m-SGD to take, on average, larger step sizes compared to vanilla SGD, which explains its improved efficiency in practice for large β. On the other hand, stable convergence characterizes the "local" behavior of the algorithm, focusing on the variability of the iterates once they enter the vicinity of a local minimum. Given that m-SGD typically takes larger step sizes on average compared to vanilla SGD, it naturally exhibits increased variability within the "ball of convergence". This is captured by the additional factor of $\sqrt{1-\beta}$ on the left-hand side of (2.4), highlighting the trade-off between fast initial convergence and higher asymptotic variance.*

While constant-momentum m-SGD is arguably the most widely used variant, a significant body of theoretical work also explores m-SGD with divergent momentum, as seen in Gitman et al. (2019); Li et al. (2024). In particular, we consider the following version of m-SGD,

$$\begin{aligned} v_n &= (1 - r\eta_n)v_{n-1} + r\eta_n \nabla f(x_{n-1}, \xi_n), \\ x_n &= x_{n-1} - \eta_n v_n, \end{aligned} \tag{3.3}$$

where v_n represents the momentum term, η_n is the step size, and r is a scaling parameter. For the analysis of this algorithm, we impose an additional condition:

Assumption 3.1 *The function F belongs to $\mathcal{C}^2\left(\mathbb{R}^d, \mathbb{R}\right)$ and satisfies the following*

$$\lim_{|x| \longrightarrow +\infty} F(x) = +\infty, \quad \sup_{x \in \mathbb{R}^d} |\nabla^2 F(x)| < +\infty, \quad \textit{and} \quad \|\nabla F\|^2 \leqslant c_f f,$$

where c_f is a constant.

This assumption is critical for proving the almost sure convergence of the algorithm (3.3), as demonstrated in Gadat et al. (2018). However, Lemma 13 of Jin et al. (2022) suggests that this

assumption could be relaxed to Assumption 2.3. The following result can be established using techniques similar to those in Theorems 1 and 4 (see also Theorem 3.5 in Barakat et al. (2021) and Theorem 4.1 in Li et al. (2024)).

Theorem 7 *Suppose the functions f and F satisfy Assumptions 2.1-2.4 and 3.1. Let $V \subseteq \mathbb{R}^d$ be a closed set containing J_0, and consider an initial point $x_0 \in V$. For the m-SGD iterates defined in (3.3) with step sizes $\eta_i = \eta i^{-\alpha}$, where $\eta > 0$, $\alpha \in (1/2, 1)$, and some $r > 0$, there exists a random variable $X(x_0)$ supported on the set J_0, and a function $\Gamma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ such that*

$$\Gamma(a)^{-1/2}\eta_n^{-1/2}(x_n - a)|\{X = a\} \xrightarrow{w} N(0, I_d), \text{ as } n \to \infty, \text{ if } a \in J_0. \tag{3.4}$$

This result highlights the stability and convergence properties of m-SGD with divergent momentum, extending the theoretical foundation of momentum-based optimization methods under these conditions.

3.1. Key results for Theorem 4

To provide intuition for the proof of Theorem 4, it is important to revisit the key steps of the proof of Theorem 1. In particular, the three key points listed immediately after Proposition 3 serve as the road-map for this particular proof. The almost sure convergence result of Jin et al. (2022) carries over to the m-SGD algorithm. Thus, after we have established the m-SGD counterparts of (a) Theorem 4 of Mertikopoulos et al. (2020), and (b) Proposition 3, our proof can be followed in an exactly similar manner as (2.10)-(2.19). The following two results correspond to this approach.

Proposition 8 (m-SGD version of Theorem 4.1 of Mertikopoulos et al. (2020)) *Fix some tolerance level $\delta > 0$, let $a \in J$, and suppose that Assumptions 2.2-2.4 hold. Assume further that m-SGD is run with a step-size schedule of the form $\eta_n = \eta(n + m)^{-\alpha}$ for some $\alpha \in (1/2, 1]$ and large enough $m, \eta > 0$. Then, there exist neighborhoods $\mathcal{U}$ and $\mathcal{U}_1$ of a, $\mathcal{U}_1 \subseteq \mathcal{U} \subseteq B(a, \gamma)$, such that, if $\theta_0 \in \mathcal{U}_1$, the event $\Omega_{\mathcal{U}} = \{\theta_n \in \mathcal{U} \text{ for all } n = 1, 2, \ldots\}$ occurs with probability at least $1-\delta$.*

The proof of Proposition 8 is primarily achieved by controlling the escape probability of m-SGD or SGD sequences, is controlling the probability that the iterate can move outside the ball at the n-th iteration, when it has stayed inside the ball at the previous $n - 1$ iterations. The construction of the balls $\mathcal{U}$ and $\mathcal{U}_1$ also needs care in order to ensure the escape probability can be arbitrarily bounded. The detailed proofs can be found in Appendix C.1.1. In order to conclude the asymptotic normality, we invoke the following result.

Proposition 9 *Assume the conditions 2.1-2.4 for the functions F and f. For $a \in J_0$ and $t_0 = (v_0, x_0) \in V \times V$, consider the projected momentum SGD (m-SGD) iterates given as below*

$$\begin{aligned} v_n(t_0) &= \Pi_{B(0,\gamma)}(\beta v_{n-1}(t_0) + \eta_n \nabla f(\theta_{n-1}(t_0), \xi_n)),\ v_0(t_0) = v_0, \\ \theta_n(t_0) &= \Pi_{B(a,\gamma)}(\theta_{n-1}(t_0) - \beta v_{n-1}(t_0) - \eta_n \nabla f(\theta_{n-1}(t_0), \xi_n)),\ \theta_0(t_0) = x_0, \end{aligned} \tag{3.5}$$

where the momentum coefficient β satisfies (3.1); $\eta_n = \eta n^{-\alpha}$, $\alpha \in (1/2, 1)$, is the learning rate. Then, there exists a function $\Sigma(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, such that

$$\Sigma(a)^{-1/2}\eta_n^{-1/2}(\theta_n(t_0) - a) \xrightarrow{w} Z \overset{d}{=} N(0, I_d), \text{ as } n \to \infty. \tag{3.6}$$

Furthermore, this convergence is uniform over starting point t_0 in the sense of (2.9).

The proof of this result echoes the corresponding version for vanilla SGD, Proposition 3, notwithstanding the crucial technical differences necessitated by the presence of the momentum coefficient. However, the key insight can be summarized by the following correspondence between m-SGD and SGD. For iterations (1.3), let $y_n = x_n - \frac{\beta}{1-\beta} v_n$. Then, assuming we are inside $B_{(a,\gamma)}$ for some $a \in J_0$ such that Assumptions 2.2-2.4 can be enforced, one may obtain:

$$\begin{aligned} y_n = x_{n-1} - \frac{1}{1-\beta}(\beta v_{n-1} + \eta_n \nabla f(x_{n-1}, \xi_n)) &= y_{n-1} - \frac{\eta_n}{1-\beta} \nabla f(x_{n-1}, \xi_n) \\ &\approx y_{n-1} - \frac{\eta_n}{1-\beta} A x_{n-1} + C\eta_n g(a, \xi_n) \\ &\approx (I - \frac{\eta_n}{1-\beta} A) y_{n-1} + C\eta_n g(a, \xi_n) - c\frac{\eta_n \beta}{1-\beta} v_{n-1}, \end{aligned}$$

which looks quite similar to (2.8), with an additional term containing v_{n-1}. The Gaussianity for the linearized SGD iterates follows in Lemma 15 from verifying the Lindeberg conditions; similar arguments should hold here for the m-SGD case, provided the momentum term v_n becomes asymptotically negligible. We undertake a step here in that direction.

Lemma 10 *Under the assumptions of Proposition 9,*

$$\sup_{t_0 \in V \times V} \mathbb{E}[|v_n(t_0)|^2] = O(n^{-2\alpha}).$$

Proof We provide an argument for a fixed t_0, which automatically carries over to the $\sup_{t_0 \in V \times V}$ in light of V being closed. As before, let $g(x, \xi) = \nabla F(x) - \nabla f(x, \xi)$ denote the gradient noise. Choose $C_0 > \beta^2(1-\beta^2)^{-1}$. Invoking Assumptions 2.3 and 2.4,

$$\begin{aligned} \mathbb{E}[|v_n|^2] &\leq \mathbb{E}[|\beta v_{n-1} + \eta_n \nabla F(\theta_{n-1})|^2] + \eta_n^2 \mathbb{E}[|g(0, \xi_n)|^2 + L|\theta_{n-1}|^2] \\ &\leq (1 + C_0^{-1})\beta^2 \mathbb{E}[|v_{n-1}|^2] + C\eta_n^2 \\ &\leq C_1 \mathbb{E}[|v_{n-1}|^2] + C\eta_n^2, \end{aligned} \tag{3.7}$$

for some $C_1 \in [0, 1)$, where the second inequality employs $(x+y)^2 \leq (1+c^{-1})x^2 + (1+c)y^2$, as well as $|\nabla F(\theta_n)|^2 \lesssim |\theta_n|^2 \leq \gamma$; (3.7) follows in light of $|\theta_n| \leq \gamma$, $|v_n| \leq \gamma$ and Assumption 2.3. Clearly, (3.7) immediately shows that

$$\mathbb{E}[|v_n|^2] \leq C_1^n |v_0|^2 + C \sum_{i=1}^{n} i^{-2\alpha} C_1^{n-i} = O(n^{-2\alpha}), \tag{3.8}$$

where the final inequality follows in light of $\int_1^a \tau^{-x} x^{-2\alpha} \mathrm{d}x \lesssim C_{\tau,\alpha} \beta^{-a} a^{-2\alpha}$ for $a > 1$, $\tau \in [0, 1)$. We point to the reader that (3.8) is a stronger result than Lemma 7 of Jin et al. (2022), albeit with a stronger assumption of local convexity. ■

The complete proof of Proposition 9 can be found in Appendix C.1.2.

4. Conclusion

This work introduces a unified framework for analyzing the stable convergence of endpoint iterates for vanilla SGD and m-SGD variants in non-convex optimization. Our asymptotic findings support the use of Gaussian mixture models to estimate local minima and subsequently determine the global minimum. To the best of our knowledge, this is the first effort to achieve global minima in complex non-convex optimization, opening new possibilities for advancements in large-scale data science.

References

Jing An and Jianfeng Lu. Convergence of stochastic gradient descent under a local Łojasiewicz condition for deep neural networks. *arXiv preprint arXiv:2304.09221*, 2023. URL https://arxiv.org/abs/2304.09221.

Anas Barakat, Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. *Electron. J. Stat.*, 15(2):3892–3947, 2021. ISSN 1935-7524. doi: 10.1214/21-ejs1880. URL https://doi.org/10.1214/21-ejs1880.

Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, anniversary edition, 2012. ISBN 978-1-118-12237-2. With a foreword by Steve Lalley and a brief biography of Billingsley by Steve Koppes.

Julius R. Blum. Approximation methods which converge with probability one. *Ann. Math. Statistics*, 25:382–386, 1954. ISSN 0003-4851. doi: 10.1214/aoms/1177728794. URL https://doi.org/10.1214/aoms/1177728794.

Joan Bruna, Raja Giryes, Stefano Soatto, and René Vidal. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017. URL https://arxiv.org/abs/1712.04741.

Changxiao Cai, Gen Li, H. Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Operations Research*, 70(2):1219–1237, 2022. doi: 10.1287/opre.2021.2106.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 2020. ISSN 0090-5364,2168-8966. doi: 10.1214/18-AOS1801. URL https://doi.org/10.1214/18-AOS1801.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *Trans. Sig. Proc.*, 67(20):5239–5269, October 2019. ISSN 1053-587X. doi: 10.1109/TSP.2019.2937282. URL https://doi.org/10.1109/TSP.2019.2937282.

K. L. Chung. On a stochastic approximation method. *Ann. Math. Statistics*, 25:463–483, 1954. ISSN 0003-4851. doi: 10.1214/aoms/1177728716. URL https://doi.org/10.1214/aoms/1177728716.

Christian Darken and John Moody. Towards faster stochastic gradient search. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 1009–1016, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *J. Mach. Learn. Res.*, 18:Paper No. 101, 51, 2017. ISSN 1532-4435,1533-7928.

Václav Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.*, 39:1327–1332, 1968. ISSN 0003-4851. doi: 10.1214/aoms/1177698258. URL https://doi.org/10.1214/aoms/1177698258.

Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.*, 21:Paper No. 136, 48, 2020. ISSN 1532-4435,1533-7928.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Process. Appl.*, 156:312–348, 2023. ISSN 0304-4149,1879-209X. doi: 10.1016/j.spa.2022.11.012. URL https://doi.org/10.1016/j.spa.2022.11.012.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12 (1):461–529, 2018. ISSN 1935-7524. doi: 10.1214/18-EJS1395. URL https://doi.org/10.1214/18-EJS1395.

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. PMLR, 2015.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2981–2989, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4eff0720836a198b6174eecf02cbfdbf-Paper.pdf.

P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980. ISBN 0-12-319350-8.

David M Himmelblau et al. *Applied nonlinear programming*. McGraw-Hill, 2018.

Jie Hu, Vishwaraj Doshi, and Do Young Eun. Efficiency ordering of stochastic gradient descent. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Ruinan Jin, Yu Xing, and Xingkang He. On the convergence of mSGD and adagrad for stochastic optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=g5tANwND04i.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*,

pages 2698–2707. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kleinberg18a.html.

Taehee Ko and Xiantao Li. A local convergence theory for the stochastic gradient descent method in non-convex optimization with non-isolated local minima. *J. Mach. Learn.*, 2(2):138–160, 2023. ISSN 2790-203X,2790-2048.

Tze Leung Lai. Stochastic approximation. volume 31, pages 391–406. 2003. doi: 10.1214/aos/1051027873. URL https://doi.org/10.1214/aos/1051027873. Dedicated to the memory of Herbert E. Robbins.

Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(10):4394–4400, 2020. ISSN 2162-237X,2162-2388. doi: 10.1109/tnnls.2019.2952219. URL https://doi.org/10.1109/tnnls.2019.2952219.

Tiejun Li, Tiannan Xiao, and Guoguo Yang. Revisiting the central limit theorems for the SGD-type methods. *Commun. Math. Sci.*, 22(5):1427–1454, 2024. ISSN 1539-6746,1945-0796.

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22 (4):551–575, 1977. doi: 10.1109/tac.1977.1101561. URL https://doi.org/10.1109/tac.1977.1101561.

Lennart Ljung. Analysis of stochastic gradient algorithms for linear regression problems. *IEEE Trans. Inform. Theory*, 30(2):151–160, 1984. ISSN 0018-9448,1557-9654. doi: 10.1109/TIT.1984.1056895. URL https://doi.org/10.1109/TIT.1984.1056895.

Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Comput. Optim. Appl.*, 77(3):653–710, 2020. ISSN 0926-6003,1573-2894. doi: 10.1007/s10589-020-00220-z. URL https://doi.org/10.1007/s10589-020-00220-z.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008. ISSN 1052-6234,1095-7189. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Yu. Nesterov and J.-Ph. Vial. Confidence level solutions for stochastic programming. *Automatica J. IFAC*, 44(6):1559–1568, 2008. ISSN 0005-1098,1873-2836. doi: 10.1016/j.automatica.2008.01.017. URL https://doi.org/10.1016/j.automatica.2008.01.017.

B. T. Poljak. Some methods of speeding up the convergence of iterative methods. *Ž. Vyčisl. Mat i Mat. Fiz.*, 4:791–803, 1964. ISSN 0044-4669.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

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

Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29:373–405, 1958. ISSN 0003-4851. doi: 10.1214/aoms/1177706619. URL https://doi.org/10.1214/aoms/1177706619.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems*, 10(2):124–151, 2020.

James C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Automat. Control*, 45(10):1839–1853, 2000. ISSN 0018-9286,1558-2523. doi: 10.1109/TAC.2000.880982. URL https://doi.org/10.1109/TAC.2000.880982.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1139–III–1147. JMLR.org, 2013.

Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *J. Mach. Learn. Res.*, 24:Paper No. [58], 47, 2023. ISSN 1532-4435,1533-7928.

Kejie Tang, Weidong Liu, Yichen Zhang, and Xi Chen. Acceleration of stochastic gradient descent with momentum by averaging: Finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023. URL https://arxiv.org/abs/2305.17665.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22:Paper No. 150, 63, 2021. ISSN 1532-4435,1533-7928.

Jun-Kun Wang, Chi-Heng Lin, and Jacob Abernethy. Escaping saddle points faster with stochastic momentum. *arXiv preprint arXiv:2106.02985*, 2021.

Ziyang Wei, Wanrong Zhu, and Wei Biao Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *Arxiv Preprint.*, 2023. URL https://arxiv.org/abs/2307.06915.

Wim Wiegerinck, Andrzej Komoda, and Tom Heskes. Stochastic dynamics of learning with momentum in neural networks. *J. Phys. A*, 27(13):4425–4437, 1994. ISSN 0305-4470,1751-8121. URL http://stacks.iop.org/0305-4470/27/4425.

J. Wolfowitz. On stochastic approximation methods. *Ann. Math. Statist.*, 27:1151–1156, 1956. ISSN 0003-4851. doi: 10.1214/aoms/1177728082. URL https://doi.org/10.1214/aoms/1177728082.

Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5535–5545, Red Hook, NY, USA, 2018. Curran Associates Inc.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7184–7193, 2019. URL https://arxiv.org/abs/1905.03817.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 4234–4248, 2021. URL https://proceedings.neurips.cc/paper/2021/file/21ce689121e39821d07d04faab328370-Paper.pdf.

Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Online bootstrap inference with nonconvex stochastic gradient descent estimator. 2023. URL https://arxiv.org/abs/2306.02205.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BylIciRcYQ.

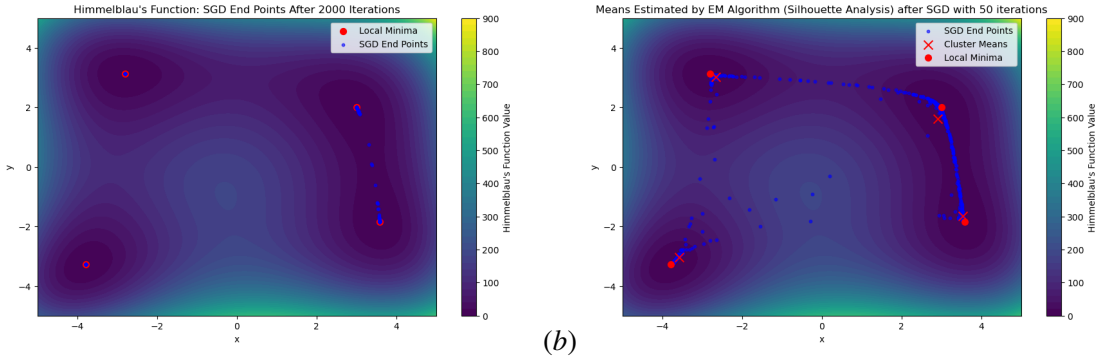
Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.*, 118(541):393–404, 2023. ISSN 0162-1459,1537-274X. doi: 10.1080/01621459.2021.1933498. URL https://doi.org/10.1080/01621459.2021.1933498.

Appendix A. Numerical example: Himmelblau's function

In this section, we present a numerical experiment to validate the theoretical analysis of the central limit theorem for SGD-type methods applied to non-convex optimization problems. More importantly, we demonstrate the use of a Gaussian mixture model applied to the endpoint iterations to identify the global minimum. To that end, consider a randomized version of the *Himmelblau's function* Himmelblau et al. (2018), a widely-used test function in mathematical optimization:

$$f(x, y, \xi) = ((x^2 + y - 11)^2 + (x + y^2 - 7)^2) + \xi(x^2 + y^2),\ \xi \sim N(0, 1). \tag{A.1}$$

It is well-known that $F(x, y) = \mathbb{E}[f(x, y, \xi)]$ has four local minima and one local maximum. Using a vanilla SGD algorithm with $\eta_t = 0.01t^{-0.65}$, and initializing x_0 randomly from Unif$[-4, 4] \times [-4, 4]$, the algorithm requires approximately 2000 iterations to converge to one of the local minima (Fig. 1, left). To identify all the local minima, we run 5000 independently initialized SGD chains in parallel. Following Theorems 1-4, we can apply a Gaussian mixture model algorithm to effectively identify the distinct local minima. As shown in Fig. 1, right, an EM algorithm combined with Silhouette analysis Rousseeuw (1987) successfully identifies the distinct local minima after running each SGD chain for just 50 iterations. Although many chains have not yet reached the local minima, the Gaussian properties of the stable convergence allow the EM algorithm to accurately discern the local minima, overcoming errors from premature chain termination. We further



(*a*) (*b*)

Figure 1: (left) SGD iterates after 2000 iterations near the local minima. (right) After only 50 iterations, a lot of the chains haven't converged, but our Algorithm 1 captures all local minima.

analyze the asymptotic variance of both SGD and m-SGD. To estimate the asymptotic variance of $\eta_n^{-1/2} x_n$ for the algorithms defined in (1.2) or (1.3), we run 1000 independently initialized chains, each for an extended $2,000,000$ iterations. As shown in Zhu et al. (2023); Chen et al. (2020), for averaged-SGD, the rate of convergence of the covariance estimator to the corresponding asymptotic covariance matrix is relatively slow, approximately $n^{-1/8}$ to $n^{-1/6}$, when the step-size parameter α is close to $1/2$. This rate is significantly slower than the asymptotic convergence rate of both the SGD iterates x_n and their averaged counterpart $\bar{x}_n = \sum_{i=1}^n x_i$. Let the estimated covariance matrix for (1.2) and (1.3) based on the 1000 independently initialized SGD chains, be denoted as $\hat{\Sigma}_1$ and $\hat{\Sigma}_2(\beta)$, respectively. Define the element-wise division operator as $\hat{R}_{ij}(\beta) = (\Sigma_2(\beta))_{ij}/(\Sigma_1)_{ij}$. Table 1 provides the value of $|\hat{R}(\beta)|_\infty$ for five different values of $\beta = 0.1, 0.3, 0.5, 0.7, 0.9$, corresponding to the four distinct local minima of $\mathbb{E}[f(x, y, \xi)]$, where $f : \mathbb{R}^3 \to \mathbb{R}$ is defined as in (A.1).

All numerical values have been rounded to 3 decimal places. On average, the $\hat{R}$ values correspond to $(1-\beta)^{-1}$ across all four local minima, justifying the inflated variance in Theorem 4.

β	Local Minima	$\lvert\hat{R}(\beta)\rvert_\infty$	$1/(1-\beta)$
0.1	(-3.779,-3.283)	1.046	1.111
	(3,2)	1.183	
	(-2.805,3.131)	1.200	
	(3.584,-1.848)	1.061	
0.3	(-3.779,-3.283)	1.586	1.429
	(3,2)	1.445	
	(-2.805,3.131)	1.478	
	(3.584,-1.848)	1.181	
0.5	(-3.779,-3.283)	2.258	2.000
	(3,2)	2.019	
	(-2.805,3.131)	2.010	
	(3.584,-1.848)	2.087	
0.7	(-3.779,-3.283)	3.544	3.333
	(3,2)	2.919	
	(-2.805,3.131)	2.894	
	(3.584,-1.848)	3.787	
0.9	(-3.779,-3.283)	9.626	10.000
	(3,2)	12.390	
	(-2.805,3.131)	10.733	
	(3.584,-1.848)	10.467	

Table 1: Empirical ratio of asymptotic covariances of m-SGD (1.3) and vanilla SGD (1.2) iterates.

Appendix B. Auxiliary results for vanilla SGD

In this section we collect all the subsidiary results leading up to the proof of Theorem 1. First we will look at proving Proposition 3. Before we detail the proof of Proposition 3, we establish the convention $A(x) = \nabla^2 F(x)$, and $S(x) = \mathbb{E}[\nabla f(x,\xi)(\nabla f(x,\xi))^\top]$. Moreover, without loss of generality, let $a = 0$; otherwise we can consider $y_n - a$ instead of y_n. We will also be required to define the three following oracle approximations. Let $g(0,\xi_i) = \nabla F(0) - \nabla f(0,\xi_i)$. With the notation $A := A(0)$ and $S := S(0)$, consider the oracle iterates (2.6)-(2.8). Our proof depends on a series of careful approximations facilitating a large sample convergence to a Gaussian random variable. Such approximations will often require the use of constants, accompanying an optimal rate. In particular, we will use the notations C, c_1 and c_2 for all such constants, which may depend on α, η, L, L' and $\min_{y\in J_0}\mu(y)$. The value of these constants may change from line-to line without being explicitly stated.

The proof of Proposition 3 follows directly from the following sequence of results, all of which are stated and proved with the notion that $a = 0$. The main motivation behind the proof is linearizing

the projected SGD sequences $\{y_n\}$, similar to Polyak and Juditsky (1992), which enables global convexity to kick in, ensuring convergence to Gaussianity.

Lemma 11 *Under the assumptions of Proposition 3, it holds that*

$$\sup_{x_0 \in V} \mathbb{E}[|y_n(x_0)|^2] = O(n^{-\alpha}).$$

Lemma 12 *Grant the assumptions of Proposition 3. Then for a given $\lambda > 0$, it holds that,*

$$\sup_{x_0 \in V} \mathbb{P}(\eta_n^{-1/2} |y_n(x_0) - y_n^{(1)}(x_0)| > \lambda) \to 0, \quad \text{as } n \to \infty.$$

Lemma 13 *Under Proposition 3, for a given $\lambda > 0$, one obtains*

$$\sup_{x_0 \in V} \mathbb{P}(\eta_n^{-1/2} |y_n^{(1)}(x_0) - y_n^{(2)}(x_0)| > \lambda) \to 0, \quad \text{as } n \to \infty.$$

Lemma 14 *For $\lambda > 0$, it holds that*

$$\sup_{x_0 \in V} \mathbb{P}(\eta_n^{-1/2} |y_n^{(2)}(x_0) - y_n^{(3)}(x_0))| > \lambda) \to 0,$$

as $n \to \infty$, if we grant the assumptions of Proposition 3.

Lemma 15 *Define $g_n(x_0) := \Sigma(a)^{-1/2} \eta_n^{-1/2} (y_n^{(3)}(x_0) - a)$ for the same $\Sigma(\cdot)$ as in Proposition 3. Then, under the Assumptions of Proposition 3, for a Borel Measurable set $\mathcal{A} \subseteq \mathbb{R}^d$, it holds that*

$$\sup_{x_0 \in V} |\mathbb{P}(g_n(x_0) \in \mathcal{A}) - \Phi(\mathcal{A})| \to 0, \text{ as } n \to \infty, \tag{B.1}$$

where $\Phi(\cdot)$ denotes the measure induced by $Z \overset{d}{=} N(0, I_d)$.

B.1. Proofs of the Lemmas

B.1.1. PROOF OF LEMMA 11

Write

$$|y_n(x_0)|^2 \leq |y_{n-1}(x_0) - \eta_n \nabla F(y_{n-1}(x_0)) + \eta_n g(y_{n-1}(x_0), \xi_n)|^2$$

where, in light of Assumption 2.4, $g(y_{n-1}(x_0), \xi_n) = \nabla F(y_{n-1}(x_0)) - \nabla f(y_{n-1}(x_0), \xi_n)$ are martingale differences with respect to the upward filtration $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$. Crucial to our argument is the observation $|y_n(x_0) - a| \leq \gamma$, which enables the application of Assumptions 2.2 and 2.3. Subsequently, a little algebra reveals that, for all sufficiently large n it holds

$$\begin{aligned}
\mathbb{E}[|y_n(x_0)|^2] &\leq \mathbb{E}[|y_{n-1}(x_0) - \eta_n \nabla F(y_{n-1}(x_0))|^2] + \eta_n^2 \mathbb{E}[|g(y_{n-1}(x_0), \xi_n)|^2] \\
\text{(Assumption 2.2)} \quad &\leq (1 - \eta_n c_1) \mathbb{E}[|y_{n-1}(x_0)|^2] + \eta_n^2 \mathbb{E}[|g(y_{n-1}(x_0), \xi_n)|^2] + C\eta_n^2 \\
\text{(Assumption 2.3)} \quad &\leq (1 - \eta_n c_1) \mathbb{E}[|y_{n-1}(x_0)|^2] + 2\eta_n^2 \mathbb{E}[|g(0, \xi_n)|^2 + L^2 |y_{n-1}(x_0)|^2] + C\eta_n^2 \\
&\leq (1 - \eta_n c_2) \mathbb{E}[|y_{n-1}(x_0)|^2] + C\eta_n^2 \mathbb{E}[|g(0, \xi_n)|^2] \\
&\vdots \\
&\leq |x_0|^2 \prod_{i=1}^{n} (1 - \eta_i c_2) + C \sum_{i=1}^{n} \eta_i^2 \prod_{k=i+1}^{n} (1 - \eta_k c_2).
\end{aligned} \tag{B.2}$$

We will leverage Lemma A.1 of Zhu et al. (2023) on (B.2). In view of elementary inequality $\int_1^a x^{-2\alpha} e^{x^{1-\alpha}}\, \mathrm{d}x \leq C a^{-\alpha} e^{a^{1-\alpha}}$ for $a > 1$, along with V being closed, one obtains

$$\sup_{x_0 \in V} \mathbb{E}[|y_n|^2] \leq C(\exp(-c_1 n^{1-\alpha}) \sup_{x_0 \in V} |x_0|^2 + n^{-\alpha}) = O(n^{-\alpha}),$$

which completes the proof.

B.1.2. PROOF OF LEMMA 12

Note that,

$$\begin{aligned}
\sup_{x_0 \in V} \mathbb{E}[|y_n(x_0) - y_n^{(1)}(x_0)|^2] &\leq \sup_{x_0 \in V} \mathbb{E}[|y_{n-1}(x_0) - y_{n-1}^{(1)}(x_0) - \eta_n(\nabla F(y_{n-1}(x_0)) - \nabla F(y_{n-1}^{(1)}(x_0))) + \\
&\qquad \eta_n(g(y_{n-1}(x_0), \xi_n) - g(0, \xi_n))|^2] \\
\text{(Assumption 2.3)} \quad &\leq (1 - \eta_n c_1) \sup_{x_0 \in V} \mathbb{E}[|y_{n-1}(x_0) - y_{n-1}^{(1)}(x_0)|^2] + \eta_n^2 \sup_{x_0 \in V} \mathbb{E}[|y_{n-1}(x_0)|^2] \\
&\leq (1 - \eta_n c_1) \sup_{x_0 \in V} \mathbb{E}[|y_{n-1}(x_0) - y_{n-1}^{(1)}(x_0)|^2] + O(\eta_n^3), \qquad \text{(B.3)}
\end{aligned}$$

where (B.3) follows from Lemma 11. Therefore, an application of Lemma A.1 of Zhu et al. (2023) along with noting that $\int_1^a x^{-3\alpha} e^{C x^{1-\alpha}} \lesssim a^{-2\alpha} e^{C a^{1-\alpha}}$ for $a > 1$, yields

$$\sup_{x_0 \in V} \mathbb{E}[|y_n(x_0) - y_n^{(1)}(x_0)|^2] \lesssim e^{-C n^{1-\alpha}} \sup_{x_0 \in V} |x_0|^2 + n^{-2\alpha} = O(\eta_n^2).$$

In light of $\eta_n \to 0$ as $n \to \infty$, we obtain Lemma 12.

B.1.3. PROOF OF LEMMA 13

Observe that

$$\begin{aligned}
\mathbb{E}[|y_n^{(1)}(x_0) - y_n^{(2)}(x_0)|] &\leq \mathbb{E}[|y_{n-1}^{(1)}(x_0) - y_{n-1}^{(2)}(x_0) + \eta_n(\nabla F(y_{n-1}^{(1)}(x_0)) - A y_{n-1}^{(2)}(x_0))|] \\
&\leq \mathbb{E}[|(I - \eta_n A)(y_{n-1}^{(1)}(x_0) - y_{n-1}^{(2)}(x_0))|] + \eta_n \mathbb{E}[|\nabla F(y_{n-1}^{(1)}(x_0)) - A y_{n-1}^{(1)}(x_0)|] \\
&\leq (1 - \eta_n c_1) \mathbb{E}[|y_{n-1}^{(1)}(x_0) - y_{n-1}^{(2)}(x_0)|] + C \eta_n \mathbb{E}[|y_{n-1}^{(1)}(x_0)|^2], \qquad \text{(B.4)}
\end{aligned}$$

where the last inequality is due to Taylor Series expansion. A treatment similar to Lemma 11 yields $\sup_{x_0 \in V} \mathbb{E}[|y_{n-1}^{(1)}(x_0)|^2] = O(\eta_n)$. Thus, in light of (B.4), and using Lemma A.1 of Zhu et al. (2023) along with $\int_1^a x^{-2\alpha} e^{C x^{1-\alpha}}\, \mathrm{d}x \leq C a^{-\alpha} e^{C a^{1-\alpha}}$, we arrive at

$$\sup_{x_0 \in V} \mathbb{E}[|y_n^{(1)}(x_0) - y_n^{(2)}(x_0)|] \lesssim e^{-C n^{1-\alpha}} \sup_{x_0 \in V} |x_0| + \sum_{k=1}^{n} \eta_k^2 \prod_{s=k+1}^{n} (1 - \eta_s c_1) = O(\eta_n). \qquad \text{(B.5)}$$

B.1.4. PROOF OF LEMMA 14

Consider the sub-sequence $i_k = k^2, k \geq 1$. Lemma B.3 in Chen et al. (2020) yields $\sup_{x_0 \in V} \mathbb{E}[|y_n^{(3)}|^2] = O(n^{-\alpha})$. Moreover, note that, for $i \in [i_k, i_{k+1})$,

$$y_i^{(3)}(x_0) - y_{i_k}^{(3)}(x_0) = (Y_0^i - Y_0^{i_k}) x_0 + \sum_{s=1}^{i_k} \eta_s (Y_s^i - Y_s^{i_k}) g(0, \xi_s) + \sum_{t=i_k+1}^{i} \eta_t Y_t^i g(0, \xi_t),$$

where $B_l = I - \eta_l A$, and $Y_i^n = \prod_{j=i+1}^n B_l$, with $Y_i^n = I$ for $i \geq n$. The matrices Y_i^n are ubiquitous in Stochastic approximation literature, appearing in the context of asymptotic analysis of SGD sequences as early as Sacks (1958) and Ruppert (1988), as well as in Polyak and Juditsky (1992). Clearly,

$$|y_i^{(3)}(x_0) - y_{i_k}^{(3)}(x_0)| \lesssim |Y_0^i - Y_0^{i_k}| + \sum_{s=1}^{i_k} |g(0,\xi_s)| s^{-\alpha} e^{Cs^{1-\alpha}} (e^{-Ci_k^{1-\alpha}} - e^{-Ci_{k+1}^{1-\alpha}}) + \sum_{s=i_k+1}^{i} |g(0,\xi_s)| s^{-\alpha} e^{Cs^{1-\alpha} - Ci_k^{1-\alpha}}.$$

Hence, due to Kolmogorov's maximal inequality (e.g., Theorem 22.4 in Billingsley (2012)), one obtains

$$\begin{aligned}
&\sup_{x_0 \in V} \mathbb{E}\left[\max_{i_k \leq i < i_{k+1}} |y_i^{(3)}(x_0) - y_{i_k}^{(3)}(x_0)|^2\right] \\
\lesssim & |e^{-Ci_k^{1-\alpha}} - e^{-Ci_{k+1}^{1-\alpha}}| + \mathbb{E}[|g(0,\xi)|^2]\left[(e^{-Ci_k^{1-\alpha}} - e^{-Ci_{k+1}^{1-\alpha}}) \sum_{s=1}^{i_k} s^{-2\alpha} e^{Cs^{1-\alpha}} + e^{-Ci_k^{1-\alpha}} \sum_{s=i_k+1}^{i_{k+1}} s^{-2\alpha} e^{Cs^{1-\alpha}}\right] \\
\lesssim & |e^{-Ci_k^{1-\alpha}} - e^{-Ci_{k+1}^{1-\alpha}}| + \mathbb{E}[|g(0,\xi)|^2](k+1)^{-2\alpha} e^{C(k+1)^{2-2\alpha} - Ck^{2-2\alpha}} = O(k^{-2\alpha}), \qquad \text{(B.6)}
\end{aligned}$$

where we have used $e^{Ck^{1-2\alpha}} = O(1)$, since $\alpha > 1/2$. Consider the set $\mathcal{B}_k(x_0) := \{\max_{n \geq k} |y_n^{(3)}(x_0)| \leq \gamma\}$. From (B.6), one obtains

$$\begin{aligned}
&\sup_{x_0 \in V} \mathbb{P}(\mathcal{B}_k^c(x_0)) \\
\leq & \sum_{s=\lfloor\sqrt{k}\rfloor}^{\infty} \left[\sup_{x_0} \mathbb{P}(|y_{i_s}^{(3)}(x_0)| > \gamma/2) + \sup_{x_0} \mathbb{P}(\max_{i_s \leq i \leq i_{s+1}} |y_i^{(3)}(x_0) - y_{i_s}^{(3)}(x_0)| > \gamma/2)\right] \\
\lesssim & \sum_{s=\lfloor\sqrt{k}\rfloor}^{\infty} s^{-2\alpha} \to 0, \text{ as } k \to \infty.
\end{aligned}$$

Therefore, $\inf_{x_0 \in V} \mathbb{P}(\mathcal{B}_k(x_0)) \to 1$ as $k \to \infty$. On the other hand, for $n \geq k$, under $\mathcal{B}_k(x_0)$ it holds that

$$\begin{aligned}
&\eta_n^{-1/2} |y_n^{(2)}(x_0) - y_n^{(3)}(x_0)| \\
= & \eta_n^{-1/2} |\Pi_{B(0,\gamma)}(y_{n-1}^{(2)}(x_0) - \eta_n A y_{n-1}^{(2)}(x_0) + \eta_n g(0,\xi_n)) - \Pi_{B(0,\gamma)}(y_{n-1}^{(3)}(x_0) - \eta_n A y_{n-1}^{(3)}(x_0) + \eta_n g(0,\xi_n))| \\
\leq & \eta_n^{-1/2}(1 - \eta_n c_1) |y_{n-1}^{(2)}(x_0) - y_{n-1}^{(3)}(x_0)| \\
\leq & |y_k^{(2)}(x_0) - y_k^{(3)}(x_0)| \prod_{s=k+1}^{n} (1 - \eta_s c_1) \lesssim 2\gamma n^{\alpha/2} e^{-Cn^{1-\alpha} + Ck^{1-\alpha}} \to 0, \text{ as } n \to \infty,
\end{aligned}$$

which completes the proof.

B.1.5. PROOF OF LEMMA 15

Recall Y_i^n from the proof of Lemma 14. Note that $g(0,\xi_i)$ are independent random vectors, with $\mathbb{E}[g(0,\xi_i)g(0,\xi_i)^\top] = S$. Let $B_n(x_0) = \mathrm{Cov}(y_n^{(3)}(x_0))$. Clearly, from (2.8),

$$B_n(x_0) = (I - \eta_n A)B_{n-1}(x_0)(I - \eta_n A) + \eta_n^2 S. \tag{B.7}$$

Let Σ be the unique matrix satisfying Lyapunov Equation

$$A\Sigma + \Sigma A = S. \tag{B.8}$$

We pause for a moment to relate Σ to the previous results on asymptotic analysis of stochastic approximations. Indeed, if $d = 1$, $\Sigma = 2^{-1}SA^{-1}$, matching the expression by Chung (1954). In general, letting $A = P\Lambda P^\top$ be the eigen-value decomposition of A, some elementary algebra yields that $\Sigma = PMP^\top$ with

$$M_{ij} = (P^\top SP)_{ij}(\Lambda_{ii} + \Lambda_{jj})^{-1},$$

which again echoes the expression (2.2.7) in Fabian (1968). The crux of our argument relies on proving

$$\eta_n^{-1}\|U_n(x_0)\|_2 = o(1),\ U_n(x_0) = B_n(x_0) - \eta_n\Sigma. \tag{B.9}$$

From (B.7), one obtains

$$U_n(x_0) = (I - \eta_n A)U_{n-1}(x_0)(I - \eta_n A) + R_n, \tag{B.10}$$

where

$$\begin{aligned} R_n &= \eta_{n-1}(I - \eta_n A)\Sigma(I - \eta_n A) - \eta_n\Sigma + \eta_n^2 S \\ &= (\eta_{n-1} - \eta_n)\Sigma + \eta_{n-1}\eta_n^2 A\Sigma A - \eta_n(\eta_{n-1} - \eta_n)S, \end{aligned} \tag{B.11}$$

where the last line utilizes (B.8). Hence, in light of (B.11) and $\alpha \in (1/2, 1)$, (B.10) implies that

$$\begin{aligned} \|U_n(x_0)\|_2 &\le (1 - \eta_n c)\|U_{n-1}(x_0)\|_2 + O(n^{-(\alpha+1)}) \\ &\lesssim e^{-cn^{1-\alpha}}\|U_0(x_0)\|_2 + \sum_{i=1}^{n} i^{-(\alpha+1)}e^{-cn^{1-\alpha}+ci^{1-\alpha}} \\ &\lesssim e^{-cn^{1-\alpha}}\|U_0(x_0)\|_2 + e^{-cn^{1-\alpha}}\int_1^n x^{-\alpha-1}e^{cx^{1-\alpha}}\mathrm{d}x = O(n^{-1}), \end{aligned}$$

which directly shows (B.9). Therefore, $\mathrm{Cov}(\eta_n^{-1/2}y_n(x_0)) \to \Sigma$ as $n \to \infty$. From (2.8), write

$$\eta_n^{-1/2}y_n^{(3)}(x_0) = \eta_n^{-1/2}Y_1^n x_0 + \sum_{i=1}^{n}\eta_n^{-1/2}\eta_i Y_i^n g(0,\xi_i) := D_n(x_0) + \sum_{i=1}^{n} A_{i,n}g(0,\xi_i), \tag{B.12}$$

where we denote $A_{i,n} = \eta_n^{-1/2}\eta_i Y_i^n$. From the treatment above, one obtains that $\Sigma_n := \mathrm{Cov}(\sum_{i=1}^n A_{i,n}g(0,\xi_i)) \to \Sigma$ as $n \to \infty$. Moreover, $\sup_{x_0\in V} D_n(x_0) \overset{a.s.}{\to} 0$ via a direct application of Lemma A.1 of Zhu et al.

(2023). Thus, in order to deduce asymptotic normality, all we require is to verify the Lindeberg condition for $\sum_{i=1}^{n} A_{i,n}u_i$, with $u_i = g(0, \xi_i)$ being i.i.d. For a $r > 0$, we aim to show

$$\begin{aligned}&\sum_{i=1}^{n} \mathbb{E}[\|\Sigma_n^{-1/2} A_{i,n}u_i\|_2^2 I\{\|\Sigma_n^{-1/2} A_{i,n}u_i\|_2 > r\}] \\ \leq &\mathbb{E}[u_1^\top (\sum_{i=1}^{n} A_{i,n}\Sigma_n^{-1}A_{i,n}^\top)u_1 I\{\max_{1\leq i\leq n} u_1^\top (A_{i,n}\Sigma_n^{-1}A_{i,n}^\top)u_1 > r^2\}] \to 0 \text{ as } n \to \infty. \end{aligned} \tag{B.13}$$

Clearly, from the convergence of Σ_n, $\|\Sigma_n^{-1}\|_2 = O(1)$. Moreover,

$$\sum_{i=1}^{n} \|A_{i,n}\|^2 \lesssim n^\alpha \sum_{i=1}^{n} i^{-2\alpha} \|Y_i^n\|_2 = O(1)$$

from yet another application of Lemma A.1 of Zhu et al. (2023). On the other hand, $\|\Sigma_n\|_2 = O(1)$, and

$$\max_{1\leq i\leq n} \|A_{i,n}\|_2^2 \lesssim n^\alpha e^{-cn^{1-\alpha}} \max_{1\leq i\leq n} i^{-2\alpha} e^{ci^{1-\alpha}} \leq \max\{n^\alpha e^{-cn^{1-\alpha}}, n^{-\alpha}\} = o(1).$$

Thus

$$\sum_{i=1}^{n} \mathbb{E}[\|\Sigma_n^{-1/2} A_{i,n}u_i\|_2^2 I\{\|\Sigma_n^{-1/2} A_{i,n}u_i\|_2 > r\}] \lesssim \mathbb{E}[\|u_1\|_2^2 I\{\|u_1\|_2 o(1) > r\}] \to 0 \text{ as } n \to \infty,$$

which shows (B.13). Therefore, from the definition of weak convergence, we have

$$\sup_{x_0 \in V} |\mathbb{P}(g_n(x_0) - \Sigma^{-1/2} D_n(x_0) \in \mathcal{A}) - \Phi(\mathcal{A})| \to 0,$$

for any Borel set $\mathcal{A}$. In light of almost sure convergence of $\sup_{x_0\in V} D_n(x_0)$, the proof is completed.

Appendix C. Proof of Theorem 4

The two key ingredients of the proof of Theorem 4 are Propositions 8 and 9. Subsequently, we prove these results.

C.1. Proofs of the Propositions

We will prove the results one-by-one. The proof of Proposition 8 requires the following modified versions of Lemma D.3 and Proposition D.2 of Mertikopoulos et al. (2020).

Lemma 16 (m-SGD version of Lemma D.3 of Mertikopoulos et al. (2020)) *Under the assumptions of Proposition 8, it holds that*

C.1. $\Omega_{n+1} \subseteq \Omega_n$ *and* $E_{n+1} \subseteq E_n$.

C.2. $E_{n-1} \subseteq \Omega_n$.

C.3. Consider the "large noise" event

$$\begin{aligned}\tilde{E}_n &\equiv E_{n-1} \backslash E_n = E_{n-1} \cap \{R_n > \varepsilon\} \\ &= \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \ldots, n-1 \text{ and } R_n > \varepsilon\}\end{aligned}$$

and let $\tilde{R}_n = R_n \mathbf{1}_{E_{n-1}}$ denote the cumulative error subject to the noise being "small" until time n. Then, for a constant $C > 0$, it holds

$$\mathbb{E}\left[\tilde{R}_n\right] \leq \mathbb{E}\left[\tilde{R}_{n-1}\right] + C\eta_n^2 - \varepsilon \mathbb{P}\left(\tilde{E}_{n-1}\right)$$

where, by convention, we write $\tilde{E}_0 = \varnothing$ and $\tilde{R}_0 = 0$.

Lemma 17 *Fix $\delta > 0$. Under the assumptions of Proposition 8, it holds that*

$$\mathbb{P}(E_n) \geq 1 - \delta \quad \text{for all } n = 1, 2, \ldots$$

C.1.1. PROOF OF PROPOSITION 8

The proof follows along the lines of Theorem 4.1 of Mertikopoulos et al. (2020), however, with important changes necessitated by the m-SGD structure. For the sake of completeness, we provide a streamlined proof. Choose $\varepsilon > 0$ such that

$$\mathcal{U} = \left\{x \in \mathbb{R}^d : \|x - a\|^2 \leq 2\varepsilon + \sqrt{\varepsilon}\right\} \subseteq B(a, \gamma),$$

and we will assume that X_1 is initialized in a neighborhood

$$\mathcal{U}_1 = \left\{x \in \mathbb{R}^d : \|x - a\|^2 \leq \varepsilon\right\}.$$

Subsequently, without loss of generality, we will take $a = 0$. Denote $D_n := |\theta_n|^2$. If $\theta_{n-1} \in B(0, \gamma)$, then from (1.3) and Assumption 2.2, it follows

$$\begin{aligned}D_n &= |\theta_{n-1} - \eta_n \nabla F(\theta_{n-1}) - \beta v_{n-1} + \eta_n g(\theta_{n-1}, \xi_n)|^2 \\ &\leq (1 - c\eta_n) D_{n-1} + (\eta_n^2(|\nabla F(\theta_{n-1})|^2 + |g(\theta_{n-1}, \xi_n)|^2) + \beta^2 |v_{n-1}|^2) + 2\beta(\eta_n \nabla F(\theta_{n-1}) - \theta_{n-1})^\top v_{n-1} \\ &\qquad + 2\eta_n(\theta_{n-1} - \eta_n \nabla F(\theta_{n-1}))^\top g(\theta_{n-1}, \xi_n) + 2\beta\eta_n v_{n-1}^\top g(\theta_{n-1}, \xi_n) \\ &\leq D_{n-1} + T_{n-1} + M_{n-1}, \qquad \text{(C.1)}\end{aligned}$$

where

$$\begin{aligned}T_n &:= \eta_{n+1}^2(|\nabla F(\theta_n)|^2 + |g(\theta_n, \xi_{n+1})|^2) + \beta^2 |v_n|^2, \\ M_n &:= 2\beta(\eta_{n+1} \nabla F(\theta_n) - \theta_n)^\top v_n + 2\eta_{n+1}(\theta_n - \eta_{n+1} \nabla F(\theta_n))^\top g(\theta_n, \xi_{n+1}) + 2\beta\eta_{n+1} v_n^\top g(\theta_n, \xi_{n+1}).\end{aligned}$$

Denote $S_n := \sum_{i=1}^n T_i$, and $G_n := \sum_{i=1}^n M_i$. Consider the squared errors

$$R_n := S_n + |G_n|^2,$$

and define the following two events:

$$\Omega_n \equiv \Omega_n(\mathcal{U}) = \{\theta_n \in \mathcal{U} \text{ for all } k = 1, 2, \ldots, n\}$$

and

$$E_n \equiv E_n(\varepsilon) = \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \ldots, n\}.$$

Note that $\Omega_{\mathcal{U}} = \cap_{n=1}^{\infty} \Omega_n$. Therefore, 5.2 in Lemma 16, and Lemma 17 yields

$$\mathbb{P}(\Omega_{\mathcal{U}}) = \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(E_{n-1}) \geq 1 - \delta,$$

which completes the proof.

Now we turn to Proposition 9.

C.1.2. PROOF OF PROPOSITION 9

Without loss of generality, we assume $a = 0$, otherwise we will work with $\theta_n - a$. Moreover, for the sake of cleaner presentation we will ignore the $\sup_{t_0 \in V \times V}$ term from the statements and the proofs of the subsequent assertions, keeping it implicit that all the arguments hold uniformly over $t_0 \in V \times V$, just as in the proof of Proposition 3. In fact, modulo some modifications necessitated by the introduction of momentum, we will majorly mirror the proof of Proposition 3. The following result, corresponding to Lemma 11, yields a control on θ_n. The proof is provided in Section C.2.3.

Lemma 18 *Under the assumptions of Proposition 9,*

$$\mathbb{E}[|\theta_n|^2] = O(n^{-\alpha}).$$

Now, akin to (2.6)-(2.8), we resort to defining a series of intermediate oracle m-SGD sequences. Consider

$$\begin{aligned}
v_n^{(1)} &= \Pi_{B(0,\gamma)}(\beta v_{n-1}^{(1)} + \eta_n \nabla F(\theta_{n-1}^{(1)}) - \eta_n g(0, \xi_n)), \\
\theta_n^{(1)} &= \Pi_{B(0,\gamma)}(\theta_{n-1}^{(1)} - \beta v_{n-1}^{(1)} - \eta_n \nabla F(\theta_{n-1}^{(1)}) + \eta_n g(0, \xi_n)); \qquad \text{(C.2)} \\
v_n^{(2)} &= \Pi_{B(0,\gamma)}(\beta v_{n-1}^{(2)} + \eta_n A\theta_{n-1}^{(2)} - \eta_n g(0, \xi_n)), \\
\theta_n^{(2)} &= \Pi_{B(0,\gamma)}(\theta_{n-1}^{(2)} - \beta v_{n-1}^{(2)} - \eta_n A\theta_{n-1}^{(2)} + \eta_n g(0, \xi_n)); \qquad \text{(C.3)} \\
v_n^{(3)} &= \beta v_{n-1}^{(3)} + \eta_n A\theta_{n-1}^{(3)} - \eta_n g(0, \xi_n), \\
\theta_n^{(3)} &= \theta_{n-1}^{(3)} - v_n^{(3)}. \qquad \text{(C.4)}
\end{aligned}$$

A proof similar to Lemmas 12 and 18 shows that

$$\|\theta_n - \theta_n^{(1)}\|^2 = O(\eta_n^2), \text{which implies } \eta_n^{-1/2}|\theta_n - \theta_n^{(1)}| \xrightarrow{\mathbb{P}} 0. \qquad \text{(C.5)}$$

Moreover, techniques from Lemma 13 can be employed to obtain

$$\|\theta_n^{(1)} - \theta_n^{(2)}\| = O(\eta_n), \text{which implies } \eta_n^{-1/2}|\theta_n^{(1)} - \theta_n^{(2)}| \xrightarrow{\mathbb{P}} 0. \qquad \text{(C.6)}$$

Next, we will argue that $\theta_n^{(3)} \xrightarrow{a.s.} 0$. In fact, a proof along the lines of Lemma 18 shows that $\mathbb{E}[|\theta_n^{(3)}|^2] = O(\eta_n)$, and $\mathbb{E}[|v_n^{(3)}|^2] = O(\eta_n^2)$. Since $\sum_i \eta_i^2 < \infty$, Borel-Cantelli Lemma indicates

that $|v_n^{(3)}| \overset{a.s.}{\to} 0$. Denote $Z_n := \theta_n^{(3)} - \frac{\beta}{1-\beta} v_n^{(3)}$. Note that,

$$\begin{aligned} Z_n &= (I - \frac{\eta_n}{1-\beta}) Z_{n-1} + \frac{\beta}{(1-\beta)^2} \eta_n A v_{n-1}^{(3)} + \frac{\eta_n}{1-\beta} g(0, \xi_n) \\ &\vdots \\ &= Q_0^n Z_0 + \frac{\beta}{(1-\beta)^2} A \sum_{i=1}^n \eta_i Q_i^n v_{i-1}^{(3)} - \frac{1}{1-\beta} \sum_{i=1}^n \eta_i Q_i^n g(0, \xi_i), \end{aligned}$$

where $Q_i^n = (I - \frac{\eta_n}{1-\beta}) \cdots (I - \frac{\eta_{i+1}}{1-\beta})$, $Q_n^n = I$. Therefore, with $i_k := k^2$, for $n \in (i_k, i_{k+1}]$,

$$\begin{aligned} |Z_n - Z_{i_k}| \leq |Q_0^{i_k} - Q_0^n||Z_0| + \frac{\beta}{(1-\beta)^2} \Bigg(\sum_{i=1}^{i_k} \eta_i |Q_i^{i_k} - Q_i^n||v_{i-1}^{(3)}| + \sum_{i=i_k+1}^n \eta_i Q_i^n |v_{i-1}^{(3)}| \Bigg) + \\ \frac{1}{1-\beta} \Bigg(\sum_{i=1}^{i_k} \eta_i |Q_i^{i_k} - Q_i^n||g(0, \xi_i)| + \sum_{i=i_k+1}^n \eta_i Q_i^n |g(0, \xi_i)| \Bigg). \end{aligned} \tag{C.7}$$

For the second term in (C.7), using $\mathbb{E}[|v_i^{(3)}|^2] = O(\eta_i^2)$,

$$\begin{aligned} &\mathbb{E}\Bigg[\max_{i_k < n \leq i_{k+1}} \Bigg(\sum_{i=1}^{i_k} \eta_i |Q_i^{i_k} - Q_i^n||v_{i-1}^{(3)}| + \sum_{i=i_k+1}^n \eta_i Q_i^n |v_{i-1}| \Bigg)^2 \Bigg] \\ \lesssim &(k+1)^2 \Bigg[\sum_{i=1}^{i_k} \eta_i^4 |Q_i^{i_k} - Q_i^{i_{k+1}}|^2 + \sum_{i=i_k+1}^{i_{k+1}} \eta_i^4 |Q_i^n|^2 \Bigg] \\ \lesssim &k^2 \Bigg[(e^{-C i_k^{1-\alpha}} - e^{-C i_{k+1}^{1-\alpha}}) \sum_{i=1}^{k^2} i^{-4\alpha} e^{C i^{1-\alpha}} + e^{-C i_{k+1}} \sum_{i=k^2+1}^{(k+1)^2} i^{-4\alpha} e^{C i^{1-\alpha}} \Bigg] \\ \lesssim &k^{2-6\alpha} e^{C(k+1)^{2-2\alpha} - k^{2-2\alpha}} = O(k^{2-6\alpha}). \end{aligned} \tag{C.8}$$

On the other hand, for the third term in (C.7), a treatment same as (B.6), we obtain

$$\mathbb{E}\Bigg[\max_{i_k < n \leq i_{k+1}} \Bigg(\sum_{i=1}^{i_k} \eta_i |Q_i^{i_k} - Q_i^n||g(0, \xi_i)| + \sum_{i=i_k+1}^n \eta_i Q_i^n |g(0, \xi_i)| \Bigg)^2 \Bigg] = O(k^{-2\alpha}). \tag{C.9}$$

Combining (C.7)-(C.9), we arrive at

$$\mathbb{E}[\max_{i_k < n \leq i_{k+1}} |Z_n - Z_{i_k}|^2] = O(k^{-2\alpha} + k^{2-6\alpha}).$$

Since $\sum_i (i^{-2\alpha} + i^{2-6\alpha}) < \infty$ in view of $\alpha > 1/2$, therefore via Borel-Cantelli Lemma and $|Z_{i_k}| \overset{a.s.}{\to} 0$ (recall that $\mathbb{E}[|Z_i|^2] = O(\eta_i)$), we have $Z_n \overset{a.s.}{\to} 0$. Finally we have $\theta_n^{(3)} \overset{a.s.}{\to} 0$ by virtue of $|v_n^{(3)}| \overset{a.s.}{\to} 0$. For some $k \in \mathbb{N}$, consider the set $\mathcal{A}_k := \{\max_{n \geq k} |\theta_n| \vee |v_n| \leq \gamma\}$. Moreover, for $n \geq k$, it holds conditional on $\mathcal{A}_k$,

$$\begin{aligned} \theta_n^{(3)} &= \Pi_{B(0,\gamma)}((I - \eta_n A)\theta_{n-1}^{(3)} - \beta v_{n-1}^{(3)} + \eta_n g(0, \xi_n)) \\ v_n^{(3)} &= \Pi_{B(0,\gamma)}(\beta v_{n-1}^{(3)} + \eta_n A \theta_{n-1}^{(3)} - \eta_n g(0, \xi_n)). \end{aligned}$$

Therefore, letting $S_n = \theta_n^{(3)} - \theta_n^{(2)}$, $T_n = v_n^{(3)} - v_n^{(2)}$, and $W_n = |S_n| + \frac{\beta}{1-\beta}|T_{n-1}|$, it is immediate that

$$|S_n| \leq |(I - \eta_n A)S_{n-1} - \beta T_{n-1}| \tag{C.10}$$
$$|T_n| \leq |\beta T_{n-1} + \eta_n A S_n|. \tag{C.11}$$

Clearly,

$$\begin{aligned} |T_n| &\leq \beta |T_{n-1}| + O(\eta_n), \\ &\vdots \\ &\leq \beta^{n-k}|T_k| + C \sum_{i=k+1}^{n} \beta^{n-i}\eta_i \\ &\leq O(\eta_n) \end{aligned} \tag{C.12}$$

holds almost surely for all $n \geq k$ conditional on $\mathcal{A}_k$. On the other hand, the following implication can be deduced from (C.11):

$$W_n \leq (1 - \eta_n \lambda_{\min}(A) + \eta_n \lambda_{\max}(A)\frac{2\beta - \beta^2}{1-\beta})W_{n-1} + \frac{\beta}{1-\beta}\eta_n(\lambda_{\min}(A) - \lambda_{\max}(A)\frac{\beta}{1-\beta})|T_{n-2}| + O(\eta_n^2). \tag{C.13}$$

From the choice of β in (3.1), β satisfies

$$\kappa(A)^{-1} = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)} \geq \frac{2\beta - \beta^2}{1-\beta} > \frac{\beta}{1-\beta}.$$

Therefore, (C.13) and (C.12) can be condensed to yield

$$W_n \leq (1 - \eta_n c)W_{n-1} + O(\eta_n^2)$$

for all $n \geq k$ conditional on $\mathcal{A}_k$. From Lemma 4 of Chung (1954), $|W_n| = O(\eta_n)$ conditional on $\mathcal{A}_n$, which readily implies, in light of $\mathbb{P}(\mathcal{A}_k) \to 1$ as $k \to \infty$, that $\eta_n^{-1/2}|\theta_n^{(2)} - \theta_n^{(3)}| \to 0$ almost surely as $n \to \infty$. Finally we deal with the convergence of $\theta_n^{(3)}$. Define

$$L_n = \theta_n^{(3)} - \frac{\beta}{1-\beta}v_n^{(3)}.$$

Note that, from (C.4) and $v_n^{(3)}$, it follows that

$$L_n = (I - \frac{\eta_n}{1-\beta}A)L_{n-1} + \frac{\eta_n}{1-\beta}g(0, \xi_n) + O_{\mathbb{P}}(\eta_n^2). \tag{C.14}$$

Therefore, following the analysis of Lemma 15, we have the Gaussian approximation result, which we then combine with Proposition 8 in the manner of (2.10) to (2.19) to obtain (3.6). Note that the extra factor of $(1 - \beta)$ can be seen to occur here (C.14), when contrasted with corresponding equation of vanilla SGD, (2.8).

C.2. Auxiliary results

In this section we document the proofs of all the Lemmas used in the proofs of Theorem 4.

C.2.1. PROOF OF LEMMA 16

7.1. is trivial by definition of Ω_n and E_n. Following Mertikopoulos et al. (2020), we employ mathematical induction for 7.2. Note that $E_0 \subseteq \Omega_1$ trivially, since $\Omega_1 = \Omega$. For the inductive step, suppose $E_{n-1} \subseteq \Omega_n$ holds. Consider a sample point in E_n, i.e $R_k \leq \varepsilon$ for all $k = 1(1)n$. Since $E_n \subseteq E_{n-1} \subseteq \Omega_n$, it follows that $\theta_k \in \mathcal{U} \subseteq B(0,\gamma)$ for all $k = 1(1)n$. Therefore, applying (C.1) in a telescopic manner for $k = 1, \ldots, n$, we arrive at

$$D_{n+1} \leq D_1 + S_n + G_n \leq D_1 + R_n + \sqrt{R_n} \leq 2\varepsilon + \sqrt{\varepsilon},$$

which shows Ω_{n+1} occurs, and thus $E_n \subseteq \Omega_{n+1}$. For 7.3., decompose $\tilde{R}_n$ as

$$\begin{aligned}\tilde{R}_n = R_n \mathbf{1}_{E_{n-1}} &= R_{n-1}\mathbf{1}_{E_{n-1}} + (R_n - R_{n-1})\,\mathbf{1}_{E_{n-1}} \\ &= R_{n-1}\mathbf{1}_{E_{n-2}} - R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}} + (R_n - R_{n-1})\,\mathbf{1}_{E_{n-1}} \\ &= \tilde{R}_{n-1} + (R_n - R_{n-1})\,\mathbf{1}_{E_{n-1}} - R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}} \end{aligned} \tag{C.15}$$

where we used the fact that $E_{n-1} = E_{n-2}\backslash\tilde{E}_{n-1}$ so $\mathbf{1}_{E_{n-1}} = \mathbf{1}_{E_{n-2}} - \mathbf{1}_{\tilde{E}_{n-1}}$ that $E_{n-1} \subseteq E_{n-2}$). From the definition of R_n, one finds,

$$R_n - R_{n-1} = T_n + M_n^2 + 2G_{n-1}M_n.$$

Now, due to Assumption 2.4,

$$\begin{aligned}\mathbb{E}[M_n\mathbf{1}_{E_{n-1}}|\mathcal{F}_n] =& 2\beta\mathbb{E}[\mathbf{1}_{E_{n-1}}(\eta_{n+1}\nabla F(\theta_n) - \theta_n)^\top v_n|\mathcal{F}_n] \\ \leq& 2\beta\mathbb{E}[\mathbf{1}_{\Omega_n}(\eta_{n+1}\nabla F(\theta_n) - \theta_n)^\top v_n|\mathcal{F}_n].\end{aligned}$$

Therefore, from $E_{n-1} \subseteq \Omega_n$,

$$\begin{aligned}\mathbb{E}[G_{n-1}M_n\mathbf{1}_{E_{n-1}}] &\leq \sqrt{\varepsilon}\mathbb{E}[M_n\mathbf{1}_{E_{n-1}}] \\ &\leq 2\sqrt{\gamma}\beta\mathbb{E}[\mathbf{1}_{\Omega_n}(|\eta_{n+1}\nabla F(\theta_n)^\top v_n| + |\theta_n^\top v_n|)] \\ &\leq C\eta_n^2,\end{aligned}$$

where the final assertion follows from an argument same as Lemma 18. Note that multiplication with $\mathbf{1}_{\Omega_n}$ allows all the assumptions 2.2-2.4 to be applicable, since $\mathcal{U} \subseteq B(0,\gamma)$. Similarly,

$$\mathbb{E}[M_n^2\mathbf{1}_{E_{n-1}}] \leq C\eta_n^3, \text{ and } \mathbb{E}[T_n\mathbf{1}_{E_{n-1}}] \leq C\eta_n^2.$$

Putting it all together, clearly

$$\mathbb{E}[(R_n - R_{n-1})\,\mathbf{1}_{E_{n-1}}] \leq C\eta_n^2, \tag{C.16}$$

for a constant $C > 0$. Moreover, we have $R_{n-1} > \varepsilon$ if $\tilde{E}_{n-1}$ occurs, so the last term becomes

$$\mathbb{E}\left[R_{n-1}\mathbf{1}_{\tilde{E}_{n-1}}\right] \geq \varepsilon\mathbb{E}\left[\mathbf{1}_{\tilde{E}_{n-1}}\right] = \varepsilon\mathbb{P}\left(\tilde{E}_{n-1}\right). \tag{C.17}$$

The proof of 5.3 is completed by combining (C.15), (C.16) and (C.17).

C.2.2. PROOF OF LEMMA 17

Follows directly from Lemma 16 and the proof of Proposition D.4 in Mertikopoulos et al. (2020).

C.2.3. PROOF OF LEMMA 18

Recall Lemma 10. Note that, via Assumptions 2.2-2.4,

$$\begin{aligned}\mathbb{E}[\|\theta_n\|^2] &\leq \mathbb{E}[\|\theta_{n-1} - \eta_n \nabla F(\theta_{n-1}) - \beta v_{n-1}\|^2] + \eta_n^2 \mathbb{E}[\|g(\theta_{n-1}, \xi_n)\|^2] \\ &\leq (1 - \eta_n \mu)\|\theta_{n-1}\|^2 - 2\beta \mathbb{E}[\theta_{n-1}^\top v_{n-1}] + O(\eta_n^2), \end{aligned} \tag{C.18}$$

where the final assertion follows from (3.8). Moreover, letting $Z_n := \mathbb{E}[\theta_n^\top v_n]$, one has

$$\begin{aligned}|Z_n| &\leq |\mathbb{E}[(\theta_{n-1} - \beta\eta_n \nabla F(\theta_{n-1}))^\top (\beta v_{n-1} + \eta_n \nabla F(\theta_{n-1}))]| + O(\eta_n^2) \\ &\leq \beta(1 - \eta_n \mu)|Z_{n-1}| + \eta_n (L^2 \mu^{-1} - \beta\mu\eta_n)\mathbb{E}[\|\theta_{n-1}\|^2] + O(\eta_n^2).\end{aligned}$$

Choose $\zeta \in (2\beta(1-\beta)^{-1}, \mu^2 L^{-2} \wedge 1)$. By our choice of ζ, $(2+\zeta)\beta < \zeta$, and $\mu > \zeta L^2 \mu^{-1}$. Therefore, with $c := \mu - \zeta L^2 \mu^{-1} > 0$,

$$\begin{aligned}&\mathbb{E}[\|\theta_n\|^2] + \zeta |Z_n| \\ \leq &(1 - \eta_n \mu + \eta_n (L^2 \mu^{-1} - \beta\mu\eta_n)\zeta)\mathbb{E}[\|\theta_{n-1}\|^2] + (2\beta + \zeta\beta(1 - \eta_n\mu))|Z_{n-1}| + O(\eta_n^2) \\ \leq &(1 - \eta_n c)(\mathbb{E}[\|\theta_{n-1}\|^2] + \zeta |Z_{n-1}|) + O(\eta_n^2). \end{aligned} \tag{C.19}$$

A treatment similar to that following (B.2) in Lemma 11 yields that $\|\theta_n\|^2 \leq \|\theta_n\|^2 + \zeta|Z_{n-1}| = O(\eta_n)$. Note that this automatically implies that $|Z_n| = O(\eta_n^2)$.