Journal of Machine Learning Research 25 (2024) 1-40 Submitted 11/23; Revised 6/24; Published 9/24

Almost Sure Convergence Rates Analysis and Saddle Avoidance of Stochastic Gradient Methods

Jun Liu[*] J.LIU@UWATERLOO.CA
Department of Applied Mathematics, University of Waterloo, Waterloo, Canada

Ye Yuan[*] YYE@HUST.EDU.CN
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

Editor: Lam Nguyen

Abstract

The vast majority of convergence *rates* analysis for stochastic gradient methods in the literature focus on convergence in expectation, whereas trajectory-wise almost sure convergence is clearly important to ensure that *any instantiation* of the stochastic algorithms would converge with probability one. Here we provide a unified almost sure convergence rates analysis for stochastic gradient descent (SGD), stochastic heavy-ball (SHB), and stochastic Nesterov's accelerated gradient (SNAG) methods. We show, for the first time, that the almost sure convergence rates obtained for these stochastic gradient methods on strongly convex functions, are arbitrarily close to their optimal convergence rates possible. For non-convex objective functions, we not only show that a weighted average of the squared gradient norms converges to zero almost surely, but also the *last iterates* of the algorithms. We further provide *last-iterate almost sure* convergence rates analysis for stochastic gradient methods on general convex smooth functions, in contrast with most existing results in the literature that only provide convergence in expectation for a weighted average of the iterates. The last-iterate almost sure convergence results also enable us to obtain almost sure avoidance of any strict saddle manifold by stochastic gradient methods with or without momentum. To the best of our knowledge, this is the first time such results are obtained for SHB and SNAG methods.

Keywords: stochastic gradient descent; stochastic heavy-ball method; stochastic Nesterov's accelerated gradient; almost sure convergence rate analysis; almost sure saddle avoidance

1. Introduction

Stochastic gradient methods (Robbins and Monro, 1951) have become the *de facto* standard methods for solving large-scale optimization problems in machine learning (Bottou et al., 2018). For this reason, investigating the fundamental theoretical properties of stochastic gradient methods is not only of theoretical interest, but also of practical relevance.

Stochastic gradient descent (SGD) (Robbins and Monro, 1951) and stochastic heavy-ball (SHB) (Polyak, 1964) are among the most popular stochastic gradient methods. SHB adds a momentum term to the iterations of SGD. This was known to accelerate the convergence of deterministic gradient descent methods (Polyak, 1964). Nesterov's accelerated gradient

[*] Co-corresponding authors.

(NAG) methods (Nesterov, 1983) have similar but slightly different iterations from that of the heavy-ball (HB) method. They have been shown to accelerate gradient descent and achieve optimal convergence rates with appropriately chosen parameters in the deterministic settings (Nesterov, 2003, Chapter 2.2). In the stochastic settings, while practical gains of adding a momentum term have been observed (Leen and Orr, 1994; Sutskever et al., 2013), the convergence rates cannot be further improved due to the proven lower bounds in terms of oracle complexity (Agarwal et al., 2012). Nonetheless, understanding the convergence properties of stochastic gradient methods with or without momentum remains a topic of both theoretical and practical interest.

In this paper, we investigate almost sure convergence properties of stochastic gradient methods, including SGD, SHB, and stochastic Nesterov's accelerated gradient (SNAG) methods, and present a unified analysis of these stochastic gradient methods on smooth objective functions. In addition to almost sure convergence rates analysis, we also demonstrate that SGD, SHB, and SNAG almost surely avoid strict saddle manifolds. To the best of our knowledge, this is the first time such results have been obtained for SHB and SNAG methods.

1.1 Related work

1.1.1 Convergence rates analysis

The vast majority of the convergence rates analysis results for stochastic gradient methods in the literature are obtained in terms of the expectation (see, e.g., SGD (Nemirovski et al., 2009; Moulines and Bach, 2011; Ghadimi and Lan, 2013), SHB (Yang et al., 2016; Orvieto et al., 2020; Yan et al., 2018; Mai and Johansson, 2020; Zhou et al., 2020), SNAG (Yan et al., 2018; Assran and Rabbat, 2020; Laborde and Oberman, 2020)). Nonetheless, almost sure convergence properties are important, because they represent what happen to individual trajectories of the stochastic iterations, which are instantiations of the stochastic algorithms actually used in practice.

For this reason, almost sure convergence of stochastic gradient methods is of practical relevance. In fact, the early analysis of SGD (Robbins and Siegmund, 1971) did provide almost sure convergence guarantees. More recent work includes Bertsekas and Tsitsiklis (2000); Bottou (2003); Zhou et al. (2017); Nguyen et al. (2018, 2019); Orabona (2020a); Mertikopoulos et al. (2020). While deterministic HB and NAG methods are well analyzed (Ghadimi et al., 2015; Nesterov, 2003; Wilson et al., 2021), almost sure convergence results for SHB and SNAG are scarce. Gadat et al. (2018) proved almost sure convergence of SHB to a minimizer for non-convex functions, under a uniformly elliptic condition on the noise which helps the algorithm to get out of any unstable point. In Sebbouh et al. (2021), SHB (and SGD) was analyzed for convex (but not strongly convex or non-convex) objective functions. The authors proved almost sure convergence rates for function values at *average* iterates of SGD and *last* iterates of SHB using the iterate moving-average (IMA) viewpoint. The established convergence rates are close to optimal (subject to an ε-factor) for general convex functions (Agarwal et al., 2012). Almost sure convergence rates were analyzed for SGD under locally strongly convex objectives in Pelletier (1998); Godichon-Baggioni (2019). To the best knowledge of the authors, the results in Sebbouh et al. (2021) are the only ones that established almost sure convergence *rates* for SHB on general convex functions. We are not aware of any almost sure convergence *rates* analysis for SHB and SNAG on strongly convex or non-convex functions. The results of this paper aim to fill this theoretical gap and provide a streamlined treatment of almost sure convergence rates analysis for stochastic gradient methods.

1.1.2 Almost sure saddle avoidance

For deterministic gradient descent methods, Lee et al. (2016, 2019) proved that with a step size smaller that $1/L$, where L is the Lipschitz constant of the gradient, gradient descent always avoids strict saddles unless initialized on a set of measure zero (i.e., the stable manifold of the saddles). Various extensions of this result were made, with different assumptions on the gradient oracle, choice of step sizes, and structure of the saddle manifold. Readers are referred to Du et al. (2017); Vlatakis-Gkaragkounis et al. (2019); Jin et al. (2017); Lee et al. (2016, 2019); Panageas and Piliouras (2017); Panageas et al. (2019) and references therein.

For saddle point avoidance by stochastic gradient methods, early work by Pemantle (1990) and Brandière and Duflo (1996) in the context of stochastic approximations showed that standard SGD almost surely avoids hyperbolic saddle points, i.e., points x_* such that $\lambda_{\min}(\nabla^2 f(x_*)) < 0$ and $\det(\nabla^2 f(x_*)) \neq 0$. The work by Benaïm and Hirsch (1995) proved almost sure avoidance of hyperbolic linearly unstable cycle by SGD. Later work by Brandière (1998); Benaïm (1999) extended such results to show that SGD-type algorithms almost surely avoid more general repelling sets. More recently, using different techniques and under different assumptions, Ge et al. (2015) showed that SGD avoids strict saddles points satisfying $\lambda_{\min}(\nabla^2 f(x_*)) < 0$ with high probability. More specifically, they showed that with a constant step size η, SGD produces iterates close to a local minimizer and hence avoids saddle points, with probability at least $1 - \zeta$, after $\Theta(\log(1/\zeta)/\eta^2)$ iterations. The work of Daneshmand et al. (2018); Fang et al. (2019) further obtained results on high-probability avoidance of saddle points and convergence to second-order stationary points, while the more recent work by Vlaski and Sayed (2022) proved efficient escape from saddle points under expectation.

The work closest to ours is that of Mertikopoulos et al. (2020), in which the authors proved that SGD almost surely avoids any strict saddle manifold for a wide spectrum of vanishing step size choices, following earlier work by Pemantle (1990); Benaïm and Hirsch (1995); Benaïm (1999). However, in these works, it is always assumed that the noise on the stochastic gradient is bounded. Moreover, while making an effort to circumvent the bounded trajectory assumption in prior work, Mertikopoulos et al. (2020) also assumed that the objective function is G-Lipschitz, which means the gradient is always bounded. We shall relax these boundedness assumptions in our analysis.

1.2 Contributions

1.2.1 Convergence rates analysis

We summarize the main contributions of the paper in Table 1 relative to existing results in the literature. We only list results that provided *almost sure convergence rates* analysis for SGD, SHB, and SNAG. We emphasize the following results as the main contributions:

- For smooth and strongly convex functions, we establish almost sure convergence rates for SGD, SHB and SNAG that are arbitrarily close to the optimal rates possible implied by information-theoretical lower bounds on oracle complexity of stochastic convex optimization (Agarwal et al., 2012).
- For smooth but non-convex functions, we establish almost sure convergence rates of SHB and SNAG for a weighted average (or the minimum) of the squared gradient norm. We also show almost sure convergence of the last iterates of SHB and SNAG.
- For smooth and general convex functions, we provide almost sure convergence rates of the last iterates of SGD, SHB, and SNAG.

In view of existing results Pelletier (1998); Godichon-Baggioni (2019), our analysis for almost sure convergence rates analysis of SGD on strongly convex functions appears to be more streamlined and unified for SGD, SHB, and SNAG. For analysis of SHB in the general convex case, our result is complementary to that in Sebbouh et al. (2021) because we allow β to be an arbitrarily fixed parameter in $(0, 1)$ (cf. the analysis of deterministic HB in Ghadimi et al. (2015)). This leads to a more unified analysis of SGD, SHB, and SNAG. For general convex functions, the results in Sebbouh et al. (2021) established almost sure convergence rate of SGD for the *average* iterate, whereas Lei et al. (2017) established almost sure convergence rates of the last iterate for SGD type algorithms in a different context (see Remark 14). Our analysis (Theorem 12) provides almost sure convergence rates of the last iterates for SGD, SHB, and SNAG. In addition to the results listed in Table 1, we also obtained another set of results (Theorem 11) on almost sure convergence of the last iterates of SHB and SNAG on non-convex functions, which generalize Orabona (2020a) for SGD.

Algorithm	strongly convex	non-convex	general convex
SGD	Pelletier (1998) Godichon-Baggioni (2019) Theorem 6	Sebbouh et al. (2021) Theorem 6	Sebbouh et al. (2021) Lei et al. (2017) Theorem 13
SHB	Theorem 8	Theorem 8	Sebbouh et al. (2021) Theorem 13
SNAG	Theorem 9	Theorem 9	Theorem 13

Table 1: Summary of the main results relative to existing results on *almost sure* convergence *rates* of stochastic gradient methods.

1.2.2 Almost sure saddle avoidance

To the best of the authors' knowledge, our paper is the first to show that the SHB and SNAG methods almost surely avoid saddle points. Our work also sharpens the analysis for SGD by removing the bounded gradient assumption and relaxing the bounded noise assumption to a local boundedness assumption, which is always satisfied in empirical risk minimization problems such as in neural network training. The key ingredient required to achieve our results was the last-iterate almost sure convergence analysis we achieved in this paper, which showed that both SHB and SNAG almost surely produce iterates with gradients converging to zero, even in the non-convex setting under very weak assumptions (Khaled and Richtárik, 2020) on the stochastic gradient. This almost sure convergence of the gradient, combined with the same asymptotic non-flatness assumption on the objective function as in Mertikopoulos et al. (2020), allowed us to circumvent the bounded gradient and bounded noise assumptions.

A preliminary version of this paper was published in Liu and Yuan (2022). Compared to the conference paper, we provide full proofs of the results, streamline the proofs, and significantly expand the theoretical analysis to include results on almost sure saddle avoidance. Since the publication of Liu and Yuan (2022), other researchers have adopted our almost sure convergence rates, including Liang et al. (2023); Reddy and Vidyasagar (2023). In Liang et al. (2023), the authors extended the almost sure rates analysis to non-smooth objective functions. In Reddy and Vidyasagar (2023), the authors investigated almost sure convergence rate analysis of heavy-ball methods with batch updating and/or approximate gradients.

1.3 Notation Summary

We provide a summary of symbols and notation used in the paper for the convenience of readers.

Symbol	**Description**
$f_t = o(g_t)$	The little-o notation, which indicates that $f_t/g_t \to 0$ as $t \to \infty$, where $\{f_t\}$ and $\{g_t\}$ are positive real-valued sequences indexed by t
$f_t = O(g_t)$	The big-O notation, which indicates that there exists some $C > 0$ such that $f_t \leq C g_t$ for all t sufficiently large
$f_t = \Theta(g_t)$	The big-Θ notation, which indicates that $f_t = O(g_t)$ and $g_t = O(f_t)$
$\mathbb{R}^d$	The d-dimensional Euclidean space
$\mathbb{R}$	The set of real numbers
$\|\cdot\|$	The Euclidean norm
$\mathbb{E}[\cdot]$	The mathematical expectation (expected value) of a random variable
$\mathbb{P}$	The probability measure of a given probability space
$\mathbb{E}[X \mid \mathcal{H}]$	The conditional expectation of a random variable X with respect to a sub-σ-algebra $\mathcal{H}$
$\mathbb{E}[X \mid Y]$	The conditional expectation of a random variable X with respect to (the σ-algebra generated by) the random variable Y
$\mathbb{E}_t$	A shorthand notation for $\mathbb{E}[\cdot \mid x_t]$, where x_t is a random variable indexed by t
∇f	The gradient of a multivariate real-valued function f
$\nabla^2 f$	The Hessian matrix of a multivariate real-valued function f
$\mathcal{C}(f)$	The critical set $\mathcal{C}(f) := \left\{x \in \mathbb{R}^d : \nabla f(x) = 0\right\}$ of f
$\mathcal{S}$	A strict saddle manifold (Section 5)
q^+	The positive part of a quantity q, given by $q^+ = \max(q, 0)$

Table 2: Summary of symbols and notation

2. Problem Formulation and Preliminaries

2.1 Problem statement and assumptions

We are interested in solving the unconstrained minimization problem

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

where $f : \mathbb{R}^d \to \mathbb{R}$, using stochastic gradient methods. For example, with a slight abuse of notation, f may arise from optimizing an expected risk of the form $f(x) = \mathbb{E}[f(x;\xi)]$, where ξ is a source of randomness indicating a sample (or a set of samples), or an empirical risk of the form $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x;\xi_i)$, where $\{\xi_i\}_{i=1}^n$ are realizations of ξ (Bottou et al., 2018). We make the following assumptions.

Assumption 1 (L-smoothness) *The continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below by $f^* := \inf_{x\in\mathbb{R}^d} f(x) \in \mathbb{R}$ and its gradient ∇f is L-Lipschitz, i.e., $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$ for all $x, y \in \mathbb{R}^d$.*

A useful consequence of Assumption 1 (see, e.g., Nesterov (2003, Lemma 1.2.3)) is the following

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{2}$$

Assumption 2 (Asymptotic non-flatness) *The function $f : \mathbb{R}^d \to \mathbb{R}$ is not asymptotically flat in the sense that* $\liminf_{\|x\|\to\infty} \|\nabla f(x)\| > 0$.

Assumption 2 is used in Section 5 to show that stochastic gradient descent methods can almost surely avoid strict saddle manifolds. Intuitively, Assumption 2 means that the gradient will not vanish (or the objective function will not be flat) near infinity. The assumption is fairly easy to satisfy, as long as one component of the gradient vector $\nabla f(x)$ does not approach zero, as $\|x\| \to \infty$, for every x. For example, it is straightforward to verify that some popular non-convex optimization benchmark functions[1] such as the Griewank function

$$f(x) = \sum_{i=1}^d \frac{x_i^2}{4000} - \Pi_{i=1}^d \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1,$$

Rastrigin function

$$f(x) = 10d + \sum_{i=1}^d \left[x_i^2 - 10\cos(2\pi x_i)\right],$$

and the Levy function

$$f(x) = \sin^2(\pi w_1) + \sum_{i=1}^{d-1}\left((w_i-1)^2\left(1+10\sin^2(\pi w_i+1)\right)\right) + (w_d-1)^2\left(1+\sin^2(2\pi w_d)\right),$$

where $w_i = 1 + \frac{x_i - 1}{4}$ for $i = 1, \ldots, d$, all satisfy Assumption 2, while having many widespread local minima.

In some settings, we also assume that f is strongly convex.

[1] `https://www.sfu.ca/~ssurjano/optimization.html`

Assumption 3 (μ-strongly convex) *There exists a positive constant μ such that*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 3 with $\mu = 0$ will be referred to as general convexity. When f is convex (strongly or generally), we further assume that f has a minimizer, i.e., $x_* \in \mathbb{R}^d$ such that $f^* = f(x_*)$. A consequence of f being μ-strongly convex is that (see, e.g., Nesterov (2003, Theorem 2.1.10))

$$\frac{1}{2\mu} \|\nabla f(x)\|^2 \geq f(x) - f^*, \quad \forall x \in \mathbb{R}^d. \tag{3}$$

In contrast, if f is generally convex and L-smooth, we have

$$\frac{1}{2L} \|\nabla f(x)\|^2 \leq f(x) - f^*, \quad \forall x \in \mathbb{R}^d, \tag{4}$$

which is a special case (by setting $y = x_*$) of an equivalent condition for f to be generally convex and L-smooth (see, e.g., (2.1.7) of Nesterov (2003, Theorem 2.1.5)), stated below:

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq f(x), \quad \forall x, y \in \mathbb{R}^d. \tag{5}$$

Since we are interested in solving (1) using stochastic gradient methods, we assume at each $x \in \mathbb{R}^d$, we have access to an unbiased estimator of the true gradient $\nabla f(x)$, denoted by $\nabla f(x; \xi)$.

Assumption 4 (ABC condition) *There exist nonnegative constants A, B, and C such that*

$$\mathbb{E}[\|\nabla f(x; \xi)\|^2] \leq A(f(x) - f^*) + B \|\nabla f(x)\|^2 + C, \quad \forall x \in \mathbb{R}^d. \tag{6}$$

Remark 1 *The above assumption was proposed in Khaled and Richtárik (2020) as "the weakest assumption" for analysis of SGD in the non-convex setting. This assumption clearly includes the uniform bound*

$$\mathbb{E}[\|\nabla f(x; \xi)\|^2] \leq \sigma^2$$

and bounded variance condition

$$\mathbb{E}[\|\nabla f(x; \xi) - \nabla f(x)\|^2] \leq \sigma^2$$

as special cases. The latter is because, by unbiasedness of $\nabla f(x; \xi)$, bounded variance is equivalent to

$$\mathbb{E}[\|\nabla f(x; \xi)\|^2] \leq \|\nabla f(x)\|^2 + \sigma^2.$$

Furthermore, in the context of solving stochastic or empirical minimization problems using SGD, by assuming that each realization or individual loss function is L-smooth and convex and that the overall objective function f is strongly convex with a unique minimizer x_*, the following bound can be derived (Nguyen et al., 2019):

$$\mathbb{E}[\|\nabla f(x; \xi)\|^2] \leq 4L(f(x) - f^*) + \sigma^2,$$

where $\sigma^2 = \mathbb{E}[\nabla f(x_*;\xi)]$. If the convexity condition on individual realization or loss function was dropped, a similar bound can still be shown (Nguyen et al., 2019) with $4L$ replaced with $\frac{4L^2}{\mu}$. Both of them are again special cases of the condition in Assumption 4. For these reasons, we shall use the seemingly most general condition in Assumption 4 throughout this paper. Note that, if (4) holds [2]*, the ABC condition (6) can be reduced to*

$$\mathbb{E}[\|\nabla f(x;\xi)\|^2] \leq (A + 2BL)(f(x) - f^*) + C. \tag{7}$$

Nonetheless, we maintain the general form of Assumption 4 and refer readers to Khaled and Richtárik (2020) for discussions on the potential benefits of using (6).

2.2 Lemmas on supermartingale convergence rates

Our almost sure convergence rate analysis relies on the following classical supermartingale convergence theorem (Robbins and Siegmund, 1971).

Proposition 2 *Let $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$ be three sequences of random variables that are adapted to a filtration $\{\mathcal{F}_t\}$. Let $\{\gamma_t\}$ be a sequence of nonnegative real numbers such that $\Pi_{t=1}^{\infty}(1+\gamma_t) < \infty$. Suppose that the following conditions hold:*

1. *X_t, Y_t, and Z_t are nonnegative for all $t \geq 1$.*
2. *$\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] \leq (1+\gamma_t)Y_t - X_t + Z_t$ for all $t \geq 1$.*
3. *$\sum_{t=1}^{\infty} Z_t < \infty$ holds almost surely.*

Then $\sum_{t=1}^{\infty} X_t < \infty$ almost surely and Y_t converges almost surely.

The following lemma, as a corollary of Proposition 2, provides concrete estimates of almost sure convergence rates for sequences of random variables satisfying a supermartingale property.

Lemma 3 *If $\{Y_t\}$ is a sequence of nonnegative random variables satisfying*

$$\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] \leq (1 - c_1\alpha_t)Y_t + c_2\alpha_t^2, \tag{8}$$

for all $t \geq 1$, where $\{\alpha_t\}$ is sequence of positive real numbers such that $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ for some $\theta \in (0, \frac{1}{2})$, and c_1 and c_2 are positive constants. Then, for any $\varepsilon \in (2\theta, 1)$,

$$Y_t = o\left(\frac{1}{t^{1-\varepsilon}}\right), \quad \textit{almost surely.}$$

The proof of Lemma 3 can be found in Appendix A. The following lemma, when used together with Proposition 2, is useful for almost sure convergence rate analysis in a slightly different setting than Lemma 3.

[2] It is shown in Khaled and Richtárik (2020, Lemma 1) that (4) holds under the conditions of Assumption 1, even without convexity or a global minimizer.

Lemma 4 *Let $\{X_t\}$ be a sequence of nonnegative real numbers and $\{\alpha_t\}$ be a decreasing sequence of positive real numbers such that the following holds:*

$$\sum_{t=1}^{\infty} \alpha_t X_t < \infty \quad \text{and} \quad \sum_{t=2}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty.$$

Define $w_t = \frac{2\alpha_t}{\sum_{i=1}^{t} \alpha_i}$, $Y_1 = X_1$, and

$$Y_{t+1} = (1 - w_t) Y_t + w_t X_t, \quad t \geq 1. \tag{9}$$

Then

$$Y_t = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right) \quad \text{and} \quad \min_{1 \leq i \leq t-1} X_i = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right). \tag{10}$$

Remark 5 *A concrete convergence rate $o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right)$ results from (10) if we choose $\alpha_t = \frac{\alpha}{t^{\frac{1}{2}+\varepsilon}}$ for some $\alpha > 0$ and $\varepsilon \in (0, \frac{1}{2})$, because then we have $\sum_{i=1}^{t-1} \alpha_i = \Theta(t^{\frac{1}{2}-\varepsilon})$, $\frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \Theta(\frac{1}{t})$, and $\sum_t \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$. Note that (10) is an asymptotic statement (see an explanation of the little-o notation in Table 2 of Section 1.3), and the summation $\sum_{i=1}^{t-1} \alpha_i$ is non-empty for $t \geq 2$.*

Lemma 4 is inspired by part of the analysis in the proof of (Sebbouh et al., 2021, Theorem 8). The proof of Lemma 4 can be found in Appendix B.

3. Almost sure convergence rate analysis for stochastic gradient methods

In this section, we present a unified almost sure convergence rate analysis for SGD, SHB and SNAG. We primarily focus on two scenarios, namely the strongly convex and non-convex cases.

3.1 Stochastic gradient descent

The iteration of the SGD method is given by

$$x_{t+1} = x_t - \alpha_t g_t, \quad t \geq 1, \tag{11}$$

where $g_t := \nabla f(x_t; \xi_t)$ is the stochastic gradient at x_t (with randomness ξ_t) and α_t is the step size.

We shall prove that, for smooth and strongly convex objective functions, SGD can achieve $o\left(\frac{1}{t^{1-\varepsilon}}\right)$ almost sure convergence rates for any $\varepsilon \in (0, 1)$. To the best knowledge of the authors, this is the first result showing the $o\left(\frac{1}{t^{1-\varepsilon}}\right)$ almost sure convergence rate for SGD under the global strong convexity assumption and relaxed assumption on stochastic gradients (Khaled and Richtárik, 2020). For smooth and non-convex objective functions, the best iterates of SGD can achieve $o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right)$ almost sure convergence rates for any $\varepsilon \in (0, \frac{1}{2})$. This result was already reported in Sebbouh et al. (2021). For locally strongly convex functions, similar rates were obtained in Pelletier (1998); Godichon-Baggioni (2019). Here we provide a somewhat more streamlined proof of both the strongly convex and non-convex cases, enabled by Lemmas 3 and 4. These rates match the lower bounds in Agarwal et al. (2012) (see also Nemirovskij and Yudin (1983)) to an ε-factor.

Theorem 6 *Consider the iterates of SGD (11).*

1. *If Assumptions 1, 3, and 4 hold and $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ for some $\theta \in (0, \frac{1}{2})$, then almost surely*

$$f(x_t) - f^* = o\left(\frac{1}{t^{1-\varepsilon}}\right), \quad \forall \varepsilon \in (2\theta, 1).$$

2. *If Assumptions 1 and 4 hold and $\{\alpha_t\}$ is a decreasing sequence of positive real numbers satisfying $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ and $\sum_{t=2}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, then almost surely*

$$\min_{1 \le i \le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right). \tag{12}$$

In particular, if we choose $\alpha_t = \frac{\alpha}{t^{\frac{1}{2}+\varepsilon}}$ with $\alpha > 0$ and $\varepsilon \in (0, \frac{1}{2})$, then almost surely

$$\min_{1 \le i \le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right). \tag{13}$$

Proof 1. We first consider the strongly convex case. By smoothness of f and (2), we have

$$f(x_{t+1}) \le f(x_t) - \alpha_t \langle \nabla f(x_t), g_t \rangle + \frac{L\alpha_t^2}{2} \|g_t\|^2 .$$

Taking conditional expectation w.r.t. x_t, denoted by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | x_t]$, and using (3) lead to

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\le f(x_t) - f^* - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&= (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - (\alpha_t - \frac{LB\alpha_t^2}{2})\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - 2\mu(\alpha_t - \frac{LB\alpha_t^2}{2})(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2} \\
&= (1 - 2\mu\alpha_t + (LA/2 + LB\mu)\alpha_t^2)(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2} \\
&\le (1 - \mu\alpha_t)(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2},
\end{aligned} \tag{14}$$

provided that $(LA/2 + LB\mu)\alpha_t \le \mu$. The conclusion follows from Lemma 3.

2. For the non-convex case, by L-smoothness and as in (14), we obtain

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\le f(x_t) - f^* - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \left(\alpha_t - \frac{LB\alpha_t^2}{2}\right)\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \frac{1}{2}\alpha_t \|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2},
\end{aligned} \tag{15}$$

provided that $LB\alpha_t \leq 1$. By Proposition 2, $\sum_{t=1}^{\infty} \alpha_t \left\|\nabla f(x_t)\right\|^2 < \infty$ almost surely. The conclusions follow from Lemma 4 and Remark 5. ■

Remark 7 *We choose $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ for $\theta \to 0$ to approach the optimal almost sure convergence rate achievable under Lemma 3. In fact, any step size choice satisfying the classical condition by Robbins and Siegmund (1971): $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ will lead to almost sure convergence under the supermartingale convergence theorem (Proposition 2). What is new here is the analysis of almost sure convergence rate $o\left(\frac{1}{t^{1-\varepsilon}}\right)$ for strongly convex objective functions using Lemma 3. By choosing $\theta \to 0$, we can make $\varepsilon \to 0$. The conditions $(LA/2 + LB\mu)\alpha_t \leq \mu$ and $LB\alpha_t \leq 1$ in the proof can be easily satisfied for all $t \geq 1$, if we scale all α_t's by a constant, or for all t sufficiently large due to the choice of α_t. This difference is insignificant because in the latter case the analysis in the proof holds asymptotically and the same convergence rate follows.*

3.2 Stochastic heavy-ball method

The iteration of the SHB method is given by

$$x_{t+1} = x_t - \alpha_t g_t + \beta(x_t - x_{t-1}), \quad t \geq 1, \tag{16}$$

where $g_t := \nabla f(x_t; \xi_t)$ is the stochastic gradient at x_t, α_t is the step size, and $\beta \in [0, 1)$. Clearly, if $\beta = 0$, SHB reduces to SGD. We take $x_1 = x_0$.

Define

$$z_t = x_t + \frac{\beta}{1-\beta} v_t, \quad v_t = x_t - x_{t-1}, \quad t \geq 1. \tag{17}$$

The iteration of SHB can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \alpha_t g_t, \\ z_{t+1} &= z_t - \frac{\alpha_t}{1-\beta} g_t. \end{aligned} \tag{18}$$

Indeed, the above update rules are easily derived from (16) and (17) as

$$v_{t+1} = x_{t+1} - x_t = -\alpha_t g_t + \beta(x_t - x_{t-1}) = \beta v_t - \alpha_t g_t,$$

and

$$\begin{aligned} z_{t+1} &= x_{t+1} + \frac{\beta}{1-\beta} v_{t+1} \\ &= x_t - \alpha_t g_t + \beta(x_t - x_{t-1}) + \frac{\beta}{1-\beta}(\beta v_t - \alpha_t g_t) \\ &= x_t + [1 + \frac{\beta}{1-\beta}]\beta v_t - [1 + \frac{\beta}{1-\beta}]\alpha_t g_t \\ &= x_t + \frac{\beta}{1-\beta} v_t - \frac{1}{1-\beta}\alpha_t g_t \\ &= z_t - \frac{\alpha_t}{1-\beta} g_t. \end{aligned}$$

To our best knowledge, the following theorem provides the first almost sure convergence rates for SHB under both strongly convex and non-convex assumptions.

Theorem 8 *Consider the iterates of SHB (16).*

1. *If Assumptions 1, 3, and 4 hold and $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ for some $\theta \in (0, \frac{1}{2})$, then almost surely*

$$f(x_t) - f^* = o\left(\frac{1}{t^{1-\varepsilon}}\right), \quad \forall \varepsilon \in (2\theta, 1).$$

2. *If Assumptions 1 and 4 hold and $\{\alpha_t\}$ is a decreasing sequence of positive real numbers satisfying $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ and $\sum_{t=2}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, then almost surely*

$$\min_{1 \le i \le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right).$$

In particular, if we choose $\alpha_t = \frac{\alpha}{t^{\frac{1}{2}+\varepsilon}}$ with $\alpha > 0$ and $\varepsilon \in (0, \frac{1}{2})$, then almost surely

$$\min_{1 \le i \le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right).$$

Proof We have

$$\|v_{t+1}\|^2 = \beta^2 \|v_t\|^2 - 2\beta\alpha_t \langle g_t, v_t\rangle + \alpha_t^2 \|g_t\|^2.$$

Taking conditional expectation w.r.t. x_t, denoted by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|x_t]$, gives

$$\begin{aligned} \mathbb{E}_t \|v_{t+1}\|^2 &= \beta^2 \|v_t\|^2 - 2\beta\alpha_t \langle \nabla f(x_t), v_t\rangle + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C\right] \\ &\le \beta^2 \|v_t\|^2 + \varepsilon_1 \beta^2 \|v_t\|^2 + \frac{\alpha_t^2}{\varepsilon_1} \|\nabla f(x_t)\|^2 + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C\right], \end{aligned} \tag{19}$$

where we used the elementary inequality $2\langle a, b\rangle \le \varepsilon_1 \|a\|^2 + \frac{1}{\varepsilon_1} \|b\|^2$ with $a = -\beta v_t$, $b = \alpha_t \nabla f(x_t)$, and an arbitrary $\varepsilon_1 > 0$. By L-smoothness of f and (2), we have

$$f(z_{t+1}) \le f(z_t) - \frac{\alpha_t}{1-\beta} \langle \nabla f(z_t), g_t\rangle + \frac{L\alpha_t^2}{2(1-\beta)^2} \|g_t\|^2.$$

By Assumption 4, taking conditional expectation w.r.t. x_t gives

$$\begin{aligned}
&\mathbb{E}_t\, f(z_{t+1}) \\
&\le f(z_t) - \frac{\alpha_t}{1-\beta}\langle \nabla f(z_t), \nabla f(x_t)\rangle + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t)-f^*) + B\left\|\nabla f(x_t)\right\|^2 + C\right] \\
&= f(z_t) - \frac{\alpha_t}{1-\beta}\left\|\nabla f(z_t)\right\|^2 - \frac{\alpha_t}{1-\beta}\langle \nabla f(z_t), \nabla f(x_t) - \nabla f(z_t)\rangle \\
&\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t)-f^*) + B\left\|\nabla f(x_t)\right\|^2 + C\right] \\
&\le f(z_t) - \frac{\alpha_t}{1-\beta}\left\|\nabla f(z_t)\right\|^2 + \frac{\alpha_t}{1-\beta}\left\|\nabla f(z_t)\right\|\frac{L\beta}{1-\beta}\left\|v_t\right\| \\
&\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t)-f^*) + B\left\|\nabla f(x_t)\right\|^2 + C\right] \\
&\le f(z_t) - \frac{\alpha_t}{1-\beta}\left\|\nabla f(z_t)\right\|^2 + \varepsilon_2\left\|v_t\right\|^2 + \frac{\alpha_t^2 L^2\beta^2}{4\varepsilon_2(1-\beta)^4}\left\|\nabla f(z_t)\right\|^2 \\
&\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t)-f^*) + B\left\|\nabla f(x_t)\right\|^2 + C\right], \qquad (20)
\end{aligned}$$

where the second last inequality is by L-smoothness of f and the last inequality is by the use of the elementary inequality $2ab \le \varepsilon_2 a^2 + \frac{1}{\varepsilon_2}b^2$ with $a = \|v_t\|$, $b = \frac{\alpha_t L\beta}{2(1-\beta)^2}\left\|\nabla f(x_t)\right\|$, and an arbitrary $\varepsilon_2 > 0$. By L-smoothness of f again, we have

$$\begin{aligned}
f(x_t) - f^* &\le f(z_t) - f^* + \frac{\beta}{1-\beta}\langle \nabla f(z_t), v_t\rangle + \frac{L\beta^2}{2(1-\beta)^2}\left\|v_t\right\|^2 \\
&\le f(z_t) - f^* + \frac{1}{2}\left\|\nabla f(z_t)\right\|^2 + \frac{\beta^2}{2(1-\beta)^2}\left\|v_t\right\|^2 + \frac{L\beta^2}{2(1-\beta)^2}\left\|v_t\right\|^2, \qquad (21)
\end{aligned}$$

and

$$\begin{aligned}
\left\|\nabla f(x_t)\right\|^2 = \left\|\nabla f(z_t) + \nabla f(x_t) - \nabla f(z_t)\right\|^2 &\le 2\left\|\nabla f(z_t)\right\|^2 + 2\left\|\nabla f(x_t) - \nabla f(z_t)\right\|^2 \\
&\le 2\left\|\nabla f(z_t)\right\|^2 + 2\frac{L^2\beta^2}{(1-\beta)^2}\left\|v_t\right\|^2. \qquad (22)
\end{aligned}$$

Combining (19)–(22) yields

$$\begin{aligned}
\mathbb{E}_t\left[f(z_{t+1}) - f^* + \left\|v_{t+1}\right\|^2\right] \le (1+c_1\alpha_t^2)[f(z_t)-f^*] &+ (\beta^2 + \varepsilon_1\beta^2 + \varepsilon_2 + c_2\alpha_t^2)\left\|v_t\right\|^2 \\
&- \left(\frac{\alpha_t}{1-\beta} - c_3\alpha_t^2\right)\left\|\nabla f(z_t)\right\|^2 + c_4\alpha_t^2,
\end{aligned}$$

where the constants c_1–c_4 can be straightforwardly determined from (19)–(22). For any $\lambda \in (\beta, 1)$, we can choose $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\beta^2 + \varepsilon_1\beta^2 + \varepsilon_2 \le \lambda$. For any $c \in (0, \frac{1}{1-\beta})$, we can choose $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$, for some $\theta \in (0, \frac{1}{2})$, sufficiently small (by changing the constant) such that $\frac{\alpha_t}{1-\beta} - c_3\alpha_t^2 \ge c\alpha_t$ for all $t \ge 1$. The above inequality becomes

$$\begin{aligned}
&\mathbb{E}_t\left[f(z_{t+1}) - f^* + \left\|v_{t+1}\right\|^2\right] \\
&\qquad \le (1+c_1\alpha_t^2)[f(z_t)-f^*] + (\lambda + c_2\alpha_t^2)\left\|v_t\right\|^2 - c\alpha_t\left\|\nabla f(z_t)\right\|^2 + c_4\alpha_t^2. \qquad (23)
\end{aligned}$$

We now consider two different cases:

1. If f is μ-strongly convex, we can use $\|\nabla f(z_t)\|^2 \geq 2\mu(f(z_t) - f^*)$ to further obtain

$$\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \leq (1 - 2c\mu\alpha_t + c_1\alpha_t^2)[f(z_t) - f^*] + (\lambda + c_2\alpha_t^2)\|v_t\|^2 + c_4\alpha_t^2.$$

By choosing $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ sufficiently small, the inequality leads to

$$\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \leq (1 - c_5\alpha_t)[f(z_t) - f^* + \|v_t\|^2] + c_4\alpha_t^2,$$

for some constant $c_5 > 0$. It follows from Lemma 3 that

$$f(z_{t+1}) - f^* + \|v_{t+1}\|^2 = o\left(\frac{1}{t^{1-\varepsilon}}\right)$$

for any $\varepsilon \in (2\theta, 1)$. The conclusion follows from (21) and (4).

2. If f is non-convex, by (22), inequality (23) leads to

$$\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \leq (1 + c_6\alpha_t^2)[f(z_t) - f^* + \|v_t\|^2] - \frac{1}{2}c\alpha_t\|\nabla f(x_t)\|^2 + c_4\alpha_t^2,$$

where $c_6 = \max(c_1, c_2)$, provided that α_t is chosen sufficiently small. By Proposition 2, we have $\sum_{t=1}^{\infty}\alpha_t\|\nabla f(x_t)\|^2 < \infty$ almost surely. The conclusions follow from Lemma 4 and Remark 5. ■

The almost sure convergence rates achieved by SHB are consistent with the best convergence rates possible for strongly convex and non-convex objective functions using stochastic gradient methods (Agarwal et al., 2012) (see also Nemirovskij and Yudin (1983)) subject to an ε-factor.

3.3 Stochastic Nesterov's accelerated gradient

The iteration of the SNAG method is given by

$$\begin{aligned} y_{t+1} &= x_t - \alpha_t g_t, \\ x_{t+1} &= y_{t+1} + \beta(y_{t+1} - y_t), \quad t \geq 1, \end{aligned} \tag{24}$$

where $g_t := \nabla f(x_t; \xi_t)$ is the stochastic gradient at x_t, α_t is the step size, and $\beta \in [0, 1)$. Clearly, if $\beta = 0$, SNAG also reduces to SGD. We take $x_1 = y_1$.

Similar to (17), define

$$z_t = x_t + \frac{\beta}{1-\beta}v_t, \quad v_t = \beta(y_t - y_{t-1}), \quad t \geq 1, \tag{25}$$

where $y_1 = y_0$[3].

The iteration of SNAG can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \beta\alpha_t g_t, \\ z_{t+1} &= z_t - \frac{\alpha_t}{1-\beta}g_t. \end{aligned} \tag{26}$$

[3] This would be consistent with $x_1 = y_1$ and the second equation in (24) with $t = 0$, which is used in (27).

Indeed, similar to how we obtained (18), the above update rules are easily derived from (24) and (25) as

$$\begin{aligned} v_{t+1} &= \beta(y_{t+1} - y_t) \\ &= \beta(x_t - \alpha_t g_t) - \beta y_t \\ &= \beta(y_t + \beta(y_t - y_{t-1}) - \alpha_t g_t) - \beta y_t \\ &= \beta v_t - \beta\alpha_t g_t, \end{aligned} \tag{27}$$

and

$$\begin{aligned} z_{t+1} &= x_{t+1} + \frac{\beta}{1-\beta} v_{t+1} \\ &= y_{t+1} + \beta(y_{t+1} - y_t) + \frac{\beta}{1-\beta} v_{t+1} \\ &= x_t - \alpha_t g_t + v_{t+1} + \frac{\beta}{1-\beta} v_{t+1} \\ &= x_t - \alpha_t g_t + \frac{1}{1-\beta} v_{t+1} \\ &= x_t - \alpha_t g_t + \frac{1}{1-\beta}(\beta v_t - \beta\alpha_t g_t) \\ &= x_t + \frac{\beta}{1-\beta} v_t - \frac{1}{1-\beta}\alpha_t g_t \\ &= z_t - \frac{\alpha_t}{1-\beta} g_t. \end{aligned}$$

Note that (26) is almost identical to (18) except for the extra β in the first equation for v_{t+1}.

Theorem 9 *Consider the iterates of SNAG (24).*

1. *If Assumptions 1, 3, and 4 hold and $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ for some $\theta \in (0, \frac{1}{2})$, then almost surely*
$$f(x_t) - f^* = o\left(\frac{1}{t^{1-\varepsilon}}\right), \quad \forall \varepsilon \in (2\theta, 1).$$

2. *If Assumptions 1 and 4 hold and $\{\alpha_t\}$ is a decreasing sequence of positive real numbers satisfying $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ and $\sum_{t=2}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1}\alpha_i} = \infty$, then almost surely*
$$\min_{1\le i\le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1}\alpha_i}\right).$$
In particular, if we choose $\alpha_t = \frac{\alpha}{t^{\frac{1}{2}+\varepsilon}}$ with $\alpha > 0$ and $\varepsilon \in (0, \frac{1}{2})$, then almost surely
$$\min_{1\le i\le t-1} \|\nabla f(x_i)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right).$$

Proof The proof is similar to that for Theorem 8. Instead of (19), we obtain

$$\mathbb{E}_t \|v_{t+1}\|^2 \leq \beta^2 \left(\|v_t\|^2 + \varepsilon_1 \|v_t\|^2 + \frac{\alpha_t^2}{\varepsilon_1} \|\nabla f(x_t)\|^2 + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C \right] \right). \tag{28}$$

The rest of the proof proceeds in the same way (with slightly different constants). We conclude the same convergence rates by Lemmas 3 and 4. ■

To our best knowledge, the above theorem provides the first result on almost sure convergence rates for SNAG under both strongly convex and non-convex assumptions. It is also evident from the above proofs that we provide a unified treatment the convergence analysis for SHB and SNAG.

4. Last-iterate convergence analysis of stochastic gradient methods

In the previous sections, we have established close-to-optimal almost sure convergence rates for popular stochastic gradient methods. These rates are proved for the last iterate[4] $f(x_t) - f^*$. When strong convexity is absent, convergence (rates) analysis for stochastic gradient methods in terms of the last iterates seems more challenging, even for general convex objective functions. We shall address these issues in this section. Such results are practically relevant, because it is the last iterates of gradient descent methods that are being used in most practical situations.

4.1 Last-iterate convergence analysis of SHB and SNAG for non-convex functions

In the non-convex setting, the convergence analysis in the previous sections shows that a weighted average of the squared gradient norm $\|\nabla f(x_i)\|^2$ converges to zero almost surely, which also implies that the "best" iterate $\min_{1\leq i\leq t} \|\nabla f(x_i)\|^2$ converges to zero almost surely (cf. Lemma 4). It is both theoretically intriguing and practically relevant to know whether the last-iterate gradient $\nabla f(x_t)$ converges almost surely. However, it is usually more challenging to analyze the convergence of the last iterate of SGD. An interesting discussion was made in Orabona (2020a), where the author simplified the long analysis in earlier work by Bertsekas and Tsitsiklis (2000) that proved the last-iterate $\|\nabla f(x_t)\|^2$ converges almost surely to zero for SGD. In this section, we extend this analysis and prove that the last-iterate gradients of SHB and SNAG both converge to zero almost surely.

We rely on the following lemma from Orabona (2020a), which can be seen as an extension of Alber et al. (1998, Proposition 2) and Mairal (2013, Lemma A.5). Here we also extend the result from $p \geq 1$ to $p > 0$. A proof is included in the Appendix for completeness.

Lemma 10 (Orabona (2020a)) *Let $\{b_t\}$ and $\{\alpha_t\}$ be two nonnegative sequences and $\{w_t\}$ be a sequence of vectors. Assume $\sum_{t=1}^{\infty} \alpha_t b_t^p < \infty$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$, where $p > 0$. Further-*

[4] Similar rates can be easily obtained for $\|x_t - x_*\|^2$ and $\|\nabla f(x_t)\|^2$ using strong convexity.

more, assume that there exists some $L > 0$ such that

$$|b_{t+\tau} - b_t| \leq L \left(\sum_{i=t}^{t+\tau-1} \alpha_i b_i + \left\| \sum_{i=t}^{t+\tau-1} \alpha_i w_i \right\| \right), \quad \forall \tau \geq 1,$$

where w_t is such that $\sum_{t=1}^{\infty} \alpha_t w_t$ converges. Then b_t converges to 0.

Theorem 11 *Consider the iterates of SHB (16) and SNAG (24), respectively. Let Assumptions 1 and 4 hold and the step size $\{\alpha_t\}$ be a sequence of positive real numbers satisfying*

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty.$$

Then we have $\nabla f(x_t) \to 0$ almost surely, as $t \to \infty$, for both the iterates of SHB and SNAG.

Proof We first prove that the last-iterate gradient of SHB converges. By (23) and Proposition 2, we have $\sum_{t=1}^{\infty} \alpha_t \left\| \nabla f(z_t) \right\|^2 < \infty$ almost surely. Furthermore, by L-smoothness of f, we have

$$\begin{aligned} \left| \|\nabla f(z_{t+\tau})\| - \|\nabla f(z_t)\| \right| \leq \|\nabla f(z_{t+\tau}) - \nabla f(z_t)\| \leq L \|z_{t+\tau} - z_t\| &= \frac{L}{1-\beta} \left\| \sum_{i=t}^{t+\tau-1} \alpha_i g_i \right\| \\ &= \frac{L}{1-\beta} \left\| \sum_{i=t}^{t+\tau-1} \alpha_i \nabla f(z_i) + \alpha_i (g_i - \nabla f(z_i)) \right\| \\ &\leq \frac{L}{1-\beta} \left(\sum_{i=t}^{t+\tau-1} \alpha_i \|\nabla f(z_i)\| + \left\| \sum_{i=t}^{t+\tau-1} \alpha_i w_i \right\| \right), \end{aligned}$$

where $w_i = g_i - \nabla f(z_i)$. To show that $\sum_{t\geq 1} \alpha_t w_t$ converges almost surely, we write

$$\alpha_t w_t = \alpha_t (g_t - \nabla f(x_t)) + \alpha_t (\nabla f(x_t) - \nabla f(z_t)).$$

We make the following claims that are proved in Appendix C.

Claim 1: $M_t = \sum_{i=1}^{t} \alpha_i (g_i - \nabla f(x_i))$ is a martingale bounded in $\mathcal{L}^2$ and hence converges almost surely (Williams, 1991, Theorem 12.1)).

Claim 2: $N_t = \sum_{i=1}^{t} \alpha_i (\nabla f(x_i) - \nabla f(z_i))$ converges almost surely.

By Claims 1 and 2, $\sum_{t=1}^{\infty} \alpha_t w_t$ converges almost surely. Applying Lemma 10 with $b_t = \|\nabla f(z_t)\|$ and $p = 2$ shows that $\nabla f(z_t) \to 0$ almost surely. We conclude that $\nabla f(x_t)$ converges to 0 almost surely in view of (22) and that $v_t \to 0$ almost surely (since $\sum_{t=1}^{\infty} \|v_t\|^2 < \infty$ almost surely).

The proof of convergence for SNAG is similar, following (28). We omitted the details here. ■

Remark 12 *Last-iterate convergence analysis in expectation for SHB and SNAG is investigated in Liu et al. (2023) by expressing SHB and SNAG in a unified form with an additional* *parameter that interpolates between SHB and SNAG (termed stochastic unified momentum (SUM); see also Yan et al. (2018)). We expect that the almost sure convergence analysis presented in the current paper can be easily applied to analyze SUM to obtain both almost sure convergence rates and last-iterate convergence. Moreover, our analysis of SHB and SNAG is unified, in the sense that they are written in nearly identical forms (cf. (26) and (18)), allowing the proof of one to be easily extended to the other.*

4.2 Last-iterate convergence *rates* of SGD, SHB, SNAG for general convex functions

We primarily focused on strongly convex and non-convex objective functions in the previous section. For functions that are generally convex, Sebbouh et al. (2021) proved almost sure convergence rates of SGD for a weighted average of the iterates. A natural question to ask is whether one can obtain some last-iterate almost sure convergence rates. Indeed, the vast majority of convergence analysis for stochastic gradient methods under general convexity assumption yields results in terms of a weighted average of the iterates. There is an interesting discussion in Orabona (2020b), where the author derived some last-iterate convergence rates in the context of non-asymptotic analysis for convergence in expectation (see also earlier work Zhang (2004); Shamir and Zhang (2013) with more restricted domains or learning rates). In this section, we provide results on almost sure last-iterate convergence rates for SGD, SHB, and SNAG. Compared with the results in Sebbouh et al. (2021) for SHB, we show that even without the iterate moving-average (IMA) parameter choices, the last iterates of SHB still converge to a minimizer, only assuming smoothness and convexity.

The proof of the following result can be found in Appendix D.

Theorem 13 *Consider the iterates of SGD (11), SHB (16), and SNAG (24), respectively. Let Assumptions 1 and 4 hold and Assumption 3 hold with $\mu = 0$. Suppose that we choose the step size $\alpha_t = \Theta\left(\frac{1}{t^{\frac{2}{3}+\varepsilon}}\right)$ for any $\varepsilon \in (0, \frac{1}{3})$. Then we have $x_t \to x_*$ for some x_* such that $f(x_*) = f^*$ almost surely and $f(x_t) - f^* = O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right)$.*

Remark 14 *While this appears to the first result on last-iterate almost sure convergence rates for SHB and SNAG, the rate $O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right)$ is not close to the lower bound obtained for convergence in expectation (Agarwal et al., 2012). Note that most convergence rates for SGD on general convex function are derived for a weighted average of the iterates. An interesting observation was made in Orabona (2020b) and the author derived a non-asymptotic last-iterate convergence rate of $O\left(\frac{\log(T)}{\sqrt{T}}\right)$ in expectation. It is unclear at this point whether the idea in Orabona (2020b) can be extended to yield a close-to-optimal asymptotic almost sure convergence rate. It would be interesting to investigate whether the law of the iterated logarithm for martingales (Stout, 1970; de la Pena et al., 2004; Balsubramani, 2014) can help determine the sharpest convergence rates in this setting. It is interesting to note that a similar almost sure convergence rate of $O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right)$ was derived in Lei et al. (2017) using totally different techniques for online gradient descent algorithms in reproducing kernel Hilbert spaces (RKHSs) without regularization. Their proof technique, based on convergence* *rates analysis with high probability and the Borel-Cantelli lemma, is different from ours. It remains an intriguing question whether sharper rates can be obtained.*

5. Almost Sure Avoidance of Strict Saddle Manifold

In this section, we analyze almost sure avoidance of saddle manifold by stochastic gradient methods. We further make the following local boundedness assumption on the stochastic gradient.

Assumption 5 (Local boundedness) *For each compact set $K \in \mathbb{R}^n$, there exists a constant C such that $\|\nabla f(x;\xi)\| \le C$ for all $x \in K$ almost surely.*

Remark 15 *The local boundedness assumption is clearly weaker than the assumption of almost surely bounded noise (Mertikopoulos et al., 2020), i.e., there exists a constant C such that*

$$\|\nabla f(x;\xi) - \nabla f(x)\| \le C$$

for all $x \in \mathbb{R}^n$ almost surely. Indeed, since $\nabla f(x)$ is assumed to be (Lipschitz) continuous and hence locally bounded, bounded noise and local boundedness of $\nabla f(x)$ implies local boundedness of $\nabla f(x;\xi)$. Assumption 5 is readily satisfied for stochastic gradient computed using a sample drawn from a finite number of samples where each sample gives a gradient function $\nabla f(x;\xi)$ that is locally bounded. For instance, let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, and suppose that each $\nabla f(x;\xi)$ corresponds to uniformly randomly choosing $i \in \{1,\dots,n\}$ and computing $\nabla f(x;\xi) = \nabla f_i(x)$. If each $\nabla f_i(x)$ is locally bounded, then Assumption 5 holds.

Finally, we need one more assumption on the stochastic gradient.

Assumption 6 *The error between the true gradient and any stochastic gradient is uniformly exciting in the sense that there exists some constant $b > 0$ such that*

$$\mathbb{E}[\langle \nabla f(x;\xi) - \nabla f(x), v\rangle^+] \ge b,$$

where $(\cdot)^+ = \max(\cdot, 0)$ denotes the positive part of a quantity, for all $x \in \mathbb{R}^d$ and all unit vector $v \in \mathbb{R}^n$.

The same assumption was made in prior work (Pemantle, 1990; Benaïm and Hirsch, 1996; Benaïm, 1999; Mertikopoulos et al., 2020). This assumption is naturally satisfied by noisy gradient dynamics (e.g., as in Ge et al. (2015); see also remarks after Assumption 5 in Mertikopoulos et al. (2020)).

Define the critical set as

$$\mathcal{C}(f) = \left\{ x \in \mathbb{R}^d : \nabla f(x) = 0 \right\}. \tag{29}$$

Definition 16 (Mertikopoulos et al. (2020)) *A strict saddle manifold $\mathcal{S}$ of f is a smooth connected component of $\mathcal{C}(f)$ satisfying*

1. *Every $x_* \in S$ is a strict saddle point (Lee et al., 2016, 2019) of f, i.e., $\lambda_{\min}(\nabla^2 f(x_*)) < 0$.*

2. *For all $x_* \in S$, all negative eigenvalues of $\nabla^2 f(x_*)$ are uniformly bounded from above by a negative constant and all positive eigenvalues of $\nabla^2 f(x_*)$ are uniformly bounded from below by a positive constant.*

The above definition is from Mertikopoulos et al. (2020). As stated in Mertikopoulos et al. (2020), the requirement that all positive eigenvalues of $\nabla^2 f(x_*)$ are uniformly bounded from below by a positive number is added for convenience of proof, while the requirement that all negative eigenvalues of $\nabla^2 f(x_*)$ are uniformly bounded from above by a negative number is a more stringent requirement for the technical proof (see proof of (Mertikopoulos et al., 2020, Lemma C.3) and the prerequisite of (Benaïm, 1999, Proposition 9.5) on (Benaïm, 1999, p.48, (iii))). Our proof of Theorem 17 below relies on conclusions of (Benaïm, 1999, Proposition 9.5).

The main result of this section is stated below.

Theorem 17 *Let $\mathcal{S}$ be any strict saddle manifold of f, where f is three times continuously differentiable and satisfies Assumptions 1, 2, 4, 5, 6. Let the step size $\{\alpha_n\}$ be decreasing and satisfy $\alpha_n = \Theta\left(\frac{1}{n^p}\right)$, where $\frac{1}{2} < p \leq 1$. Then $\mathbb{P}(x_n \to \mathcal{S} \text{ as } n \to \infty) = 0$ for both SHB (16) and SNAG (24).*

We provide a brief outline for the proof. The detailed proof can be found in Appendix E.

1. In Section E.1, we summarize preliminary results on the convergence of the sequences generated by SHB and SNAG (Liu and Yuan, 2022). Combined with Benaïm (1996), we show that the limit set of these sequences enjoy the same properties of the limit sets of trajectories of the corresponding gradient flow.

2. In Section E.2, we state previous results by Benaïm and Hirsch (1995); Benaïm (1999) on the construction of a Lyapunov function around the saddle manifold. This result will be used later in the proof.

3. The main proof is presented in Section E.3, where the probabilistic estimates by Pemantle (1990, 1992) are combined with the Lyapunov analysis due to Benaïm and Hirsch (1995); Benaïm (1999) to show both SHB and SNAG almost surely avoid strict saddle manifolds, without the bounded gradient and noise assumptions, compared with results on SGD by Mertikopoulos et al. (2020).

Remark 18 *We briefly highlight the main challenge in establishing Theorem 17. The primary technical challenge is to demonstrate that the last iterates of SGD, SHB, and SNAG converge under relaxed assumptions, a proof presented in Theorem 11 and summarized as Lemma 21. The key ingredients for this proof include the supermartingale convergence theorem, inequality estimates we established for bounding the iterates of SHB and SNAG, and the technical lemma due to Orabona (2020a). Lemma 21 further ensures that the iterates of SHB and SNAG will satisfy the technical assumptions necessary for applying previous results by Benaïm (1996), which characterize the limit sets of stochastic approximations.*

6. Conclusions

In this paper, we have provided a streamlined analysis of almost sure convergence rates for stochastic gradient methods, including SGD, SHB, and SNAG. The rates obtained for strongly convex functions are arbitrarily close to their corresponding optimal rates. For non-convex functions, the rates obtained for the *best* iterates are close to the optimal convergence rates in expectation for general convex functions (Agarwal et al., 2012). For general convex functions, we identified a gap between the last-iterate almost sure convergence rates obtained and the possible optimal rates. Whether it is possible and how to close this gap can be an interesting topic for future work.

Furthermore, our study provides evidence for the effectiveness of various stochastic gradient descent methods, including SGD, SHB, and SNAG, in avoiding strict saddle points. Our analysis expands upon previous work on SGD by removing the requirement for bounded gradients and noise in the objective function, and instead relying on a more practical local boundedness assumption on the noisy gradient. The results of our study demonstrate that even with non-bounded gradients and noise, these methods can still converge to local minimizers. This research contributes to the understanding of the behavior of gradient descent methods in non-convex optimization and has potential implications for their use in solving a wide range of machine learning and optimization problems.

Acknowledgments

This work is partially supported by the NSERC Canada Research Chairs (CRC) program, an NSERC Discovery Grant, an Ontario Early Researcher Award (ERA), and the Jiangsu Industrial Technology Research Institute (JITRI) through a JITRI-Waterloo project. The authors are grateful to the anonymous reviewers for their constructive and detailed comments on the paper, which helped improve its presentation.

References

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Ya I Alber, Alfredo N. Iusem, and Mikhail V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, 81 (1):23–35, 1998.

Mahmoud Assran and Mike Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. In *International Conference on Machine Learning*, pages 410–420. PMLR, 2020.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Michel Benaïm. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

Michel Benaïm and Morris W Hirsch. Dynamics of morse-smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, 1995.

Michel Benaïm and Morris W Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8:141–176, 1996.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Léon Bottou. Stochastic learning. In *Summer School on Machine Learning*, pages 146–168. Springer, 2003.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Odile Brandière. Some pathological traps for stochastic approximation. *SIAM Journal on Control and Optimization*, 36(4):1293–1314, 1998.

Odile Brandière and Marie Duflo. Les algorithmes stochastiques contournent-ils les pièges? *Annales de l'IHP Probabilités et statistiques*, 32(3):395–427, 1996.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164. PMLR, 2018.

Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, pages 1902–1933, 2004.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing systems*, 30, 2017.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234. PMLR, 2019.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

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.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315. IEEE, 2015.

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

Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Maxime Laborde and Adam Oberman. A Lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case. In *International Conference on Artificial Intelligence and Statistics*, pages 602–612. PMLR, 2020.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176:311–337, 2019.

John Lee. *Introduction to Smooth Manifolds.* Springer Science & Business Media, 2012.

Todd K Leen and Genevieve B Orr. Optimal stochastic search and adaptive momentum. *Advances in Neural Information Processing Systems*, pages 477–477, 1994.

Yunwen Lei, Lei Shi, and Zheng-Chu Guo. Convergence of unregularized online learning algorithms. *J. Mach. Learn. Res.*, 18(1):6269–6301, 2017.

Yuqing Liang, Dongpo Xu, Naimin Zhang, and Danilo P Mandic. Almost sure convergence of stochastic composite objective mirror descent for non-convex non-smooth optimization. *Optimization Letters*, pages 1–19, 2023.

Jinlan Liu, Dongpo Xu, Yinghua Lu, Jun Kong, and Danilo P Mandic. Last-iterate convergence analysis of stochastic momentum methods for neural networks. *Neurocomputing*, 527:27–35, 2023.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pages 2963–2983. PMLR, 2022.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, pages 6630–6639. PMLR, 2020.

Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. *arXiv preprint arXiv:1306.4650*, 2013.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *arXiv preprint arXiv:2006.11144*, 2020.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24: 451–459, 2011.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons, 1983.

Yurii Nesterov. *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, volume 269. Doklady Akademii Nauk Sssr, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Science & Business Media, 2003.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. SGD and Hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

Lam M Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takác, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20:176–1, 2019.

Francesco Orabona. Almost sure convergence of SGD on smooth nonconvex functions. Blogpost on `http://parameterfree.com`, available at `https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/`, 2020a.

Francesco Orabona. Last iterate of SGD converges (even in unbounded domains). Blogpost on `http://parameterfree.com`, available at `https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/`, 2020b.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Uncertainty in Artificial Intelligence*, pages 356–366. PMLR, 2020.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 2. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

Mariane Pelletier. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Processes and Their Applications*, 78(2):217–244, 1998.

Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

Robin Pemantle. Vertex-reinforced random walk. *Probability Theory and Related Fields*, 92 (1):117–136, 1992.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Tadipatri Uday Kiran Reddy and Mathukumalli Vidyasagar. Convergence of momentum-based heavy ball method with batch updating and/or approximate gradients. *arXiv preprint arXiv:2303.16241*, 2023.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

Rex Clark Robinson. *An Introduction to Dynamical Systems: Continuous and Discrete*, volume 19. American Mathematical Society, 2012.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79. PMLR, 2013.

Michael Shub. *Global Stability of Dynamical Systems.* Springer, 1987.

William F Stout. A martingale analogue of Kolmogorov's law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 15(4):279–290, 1970.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147. PMLR, 2013.

Stefan Vlaski and Ali H. Sayed. Second-order guarantees of stochastic gradient descent in nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(12):6489–6504, 2022. doi: 10.1109/TAC.2021.3131963.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing systems*, 32, 2019.

David Williams. *Probability with Martingales.* Cambridge University Press, 1991.

Ashia C Wilson, Ben Recht, and Michael I Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2955–2961, 2018.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

Beitong Zhou, Jun Liu, Weigao Sun, Ruijuan Chen, Claire J Tomlin, and Ye Yuan. pbsgd: Powered stochastic gradient descent methods for accelerated non-convex optimization. In *International Joint Conferences on Artificial Intelligence*, pages 3258–3266, 2020.

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. *Advances in Neural Information Processing Systems*, 30:7040–7049, 2017.

Appendix A. Proof of Lemma 3

Proof By the choice of $\alpha_t = \Theta(\frac{1}{t^{1-\theta}})$, where $\theta \in (0, \frac{1}{2})$, there exists some $\eta > 0$ such that $c_1\alpha_t \geq \frac{\eta}{t^{1-\theta}}$ for all $t \geq 1$. We shall make use of the elementary inequality

$$(t+1)^{1-\varepsilon} \leq t^{1-\varepsilon} + (1-\varepsilon)t^{-\varepsilon}, \quad t > 0, \quad \varepsilon \in (0,1), \tag{30}$$

which can be proved, for instance, as follows. Let $g(x) = x^{1-\varepsilon}$ for $x > 0$. Then $g'(x) = (1-\varepsilon)x^{-\varepsilon}$ is decreasing on $(0, \infty)$. By the mean value theorem,

$$(t+1)^{1-\varepsilon} - t^{1-\varepsilon} = g'(\xi) \leq g'(t) = (1-\varepsilon)t^{-\varepsilon},$$

where $\xi \in (t, t+1)$, which implies inequality (30). For the convenience of the readers, recall (8) as

$$\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] \leq (1 - c_1\alpha_t)Y_t + c_2\alpha_t^2,$$

Multiplying both sides with $(t+1)^{1-\varepsilon}$ and applying inequality (30) and $c_1\alpha_t \geq \frac{\eta}{t^{1-\theta}}$ lead to

$$\begin{aligned}
\mathbb{E}[(t+1)^{1-\varepsilon}Y_{t+1} \mid \mathcal{F}_t] &\leq (t+1)^{1-\varepsilon}(1 - c_1\alpha_t)Y_t + c_2(t+1)^{1-\varepsilon}\alpha_t^2 \\
&\leq \left[t^{1-\varepsilon} + (1-\varepsilon)t^{-\varepsilon}\right]\left(1 - \frac{\eta}{t^{1-\theta}}\right)Y_t + c_2(t+1)^{1-\varepsilon}\alpha_t^2 \\
&= \left(1 + \frac{1-\varepsilon}{t}\right)\left(1 - \frac{\eta}{t^{1-\theta}}\right)t^{1-\varepsilon}Y_t + c_2(t+1)^{1-\varepsilon}\alpha_t^2 \\
&= \left[1 + \frac{1-\varepsilon}{t} - \frac{\eta}{t^{1-\theta}} - \frac{\eta(1-\varepsilon)}{t^{2-\theta}}\right]t^{1-\varepsilon}Y_t + c_2(t+1)^{1-\varepsilon}\alpha_t^2,
\end{aligned}$$

where the last two equations are straightforward algebraic rearrangements. Clearly, as $t \to \infty$, the term $\frac{1-\varepsilon}{t}$ is dominated by $\frac{\eta}{2t^{1-\theta}}$. Hence, there exists some $T > 1$ sufficiently large such that, for all $t \geq T$, $\frac{1-\varepsilon}{t} - \frac{\eta}{t^{1-\theta}} - \frac{\eta(1-\varepsilon)}{t^{2-\theta}} \leq \frac{\eta}{2t^{1-\theta}} - \frac{\eta}{t^{1-\theta}} = -\frac{\eta}{2t^{1-\theta}}$. It follows that

$$\mathbb{E}[(t+1)^{1-\varepsilon} Y_{t+1} \,|\, \mathcal{F}_t] \leq t^{1-\varepsilon} Y_t - \frac{\eta}{2t^{1-\theta}} t^{1-\varepsilon} Y_t + c_2 (t+1)^{1-\varepsilon} \alpha_t^2, \quad t \geq T.$$

With $\hat{Y}_t = t^{1-\varepsilon} Y_t$, $X_t = \frac{\eta}{2t^{1-\theta}} t^{1-\varepsilon} Y_t$, $Z_t = c_2 (t+1)^{1-\varepsilon} \alpha_t^2 = \Theta\left(\frac{1}{t^{1+\varepsilon-2\theta}}\right)$, and $\gamma_t = 0$, the conditions of Proposition 2 are met for all $t \geq$T with $\hat{Y}_t$ in place of Y_t. By Proposition 2, we have $t^{1-\varepsilon} Y_t$ converges and $\sum_{t=T}^{\infty} X_t < \infty$ almost surely. We must have $t^{1-\varepsilon} Y_t \to 0$ almost surely, since $\sum_{t=T}^{\infty} \frac{\eta}{2t^{1-\theta}} = \infty$ and $\sum_{t=T}^{\infty} X_t = \sum_{t=T}^{\infty} \frac{\eta}{2t^{1-\theta}} t^{1-\varepsilon} Y_t < \infty$ almost surely. Otherwise, a contradiction would arise by the limit comparison test for convergence of infinite series. The conclusion follows. ■

Appendix B. Proof of Lemma 4

Proof Note that $w_1 = 2$ and $Y_2 = Y_1$. Since α_t is monotonically decreasing, $w_t \in [0, 1]$ for $t \geq 2$. It follows that, for each $t \geq 2$, Y_t is a weighted average of all numbers in $\{X_1, \cdots, X_{t-1}\}$. Furthermore, by (9) we have

$$Y_{t+1} \sum_{i=1}^{t} \alpha_i = Y_t \sum_{i=1}^{t-1} \alpha_i - \alpha_t Y_t + 2\alpha_t X_t, \quad t \geq 1. \tag{31}$$

Let $\hat{Y}_t = Y_t \sum_{i=1}^{t-1} \alpha_i$. Then conditions of Proposition 2 are met with $\hat{Y}_t$ in place of Y_t, $\alpha_t Y_t$ in place of X_t, and $2\alpha_t X_t$ in place of Z_t. It follows from Proposition 2 that $Y_{t+1} \sum_{i=1}^{t} \alpha_i$ converges[5] and $\sum_{t=1}^{\infty} \alpha_t Y_t < \infty$. Since $\sum_{t=2}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, $\sum_{t=1}^{\infty} \alpha_t Y_t < \infty$, and $\lim_{t\to\infty} \frac{\alpha_t Y_t}{\frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i}} = \lim_{t\to\infty} Y_t \sum_{i=1}^{t-1} \alpha_i$ exists, we must this limit equal 0 by the limit comparison test for series. Hence $Y_t = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right)$. The other part of the conclusion follows by noting $\min_{1 \leq i \leq t-1} X_i \leq Y_t$, because Y_t is a weighted average of $\{X_1, \cdots, X_{t-1}\}$. ■

Appendix C. Proof of Lemma 10 and Claims in the Proof of Theorem 11

Proof of Lemma 10: Since $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t b_t^p < \infty$, we must have $\liminf_{t\to\infty} b_t = 0$. For the sake of deriving a contradiction, suppose that $\limsup_{t\to\infty} b_t > 0$ or $\limsup_{t\to\infty} b_t = \infty$. Then there exists some $\varepsilon > 0$ such that we have a subsequence $\{b_{n_k}\}$ satisfying $b_{n_k} \leq \frac{\varepsilon}{2}$ for all $k \geq 1$ and another subsequence $\{b_{m_k}\}$ satisfying $b_{m_k} \geq \varepsilon$ for all $k \geq 1$.

Given $\varepsilon > 0$, define a constant $C_\varepsilon = \max\left(\varepsilon^{1-p}, (\varepsilon/2)^{1-p}\right)$. Since both $\sum_{t=1}^{\infty} \alpha_t b_t^p$ and $\sum_{t=1}^{\infty} \alpha_t w_t$ converge, the partial sums of these series are Cauchy sequences. There exists

[5] While no random sequences are involved here, Proposition 2 is still applicable with almost sure convergence replaced by convergence. A direct proof is possible using the monotone convergence theorem for real numbers.

some $N > 0$ sufficiently large such that, for all $t \geq N$ and $\tau \geq 0$, we have

$$\sum_{i=t}^{t+\tau} \alpha_i b_i^p < \frac{\varepsilon}{4LC_\varepsilon}, \quad \left\| \sum_{i=t}^{t+\tau} \alpha_i w_i \right\| < \frac{\varepsilon}{4L}. \tag{32}$$

We consider the case $0 < p < 1$ and $p \geq 1$ separately.

1) Assume $0 < p < 1$. Pick some $n_k \geq N$ such that $b_{n_k} \leq \frac{\varepsilon}{2}$. Then pick the first $m_{\hat{k}} > n_k$ such that $b_{m_{\hat{k}}} \geq \varepsilon$, which implies that, for all i with $n_k \leq i < m_{\hat{k}}$, $b_i < \varepsilon$. By the assumption of the Lemma and (32), we have

$$\begin{aligned}\left| b_{m_{\hat{k}}} - b_{n_k} \right| &\leq L \left(\sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i b_i + \left\| \sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i w_i \right\| \right) \\ &= L \left(\sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i b_i^p b_i^{1-p} + \left\| \sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i w_i \right\| \right) \\ &\leq L \left(\sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i b_i^p \varepsilon^{1-p} + \left\| \sum_{i=n_k}^{m_{\hat{k}}-1} \alpha_i w_i \right\| \right) < L(\frac{\varepsilon}{4LC_\varepsilon} C_\varepsilon + \frac{\varepsilon}{4L}) = \frac{\varepsilon}{2},\end{aligned}$$

which contradicts that $b_{n_k} \leq \frac{\varepsilon}{2}$ and $b_{m_{\hat{k}}} \geq \varepsilon$.

2) Similarly, for $p \geq 1$, pick some $m_k \geq N$ such that $b_{m_k} \geq \varepsilon$. Then pick the first $n_{\hat{k}} > m_k$ such that $b_{n_{\hat{k}}} \leq \frac{\varepsilon}{2}$. We have, for all i with $m_k \leq i < n_{\hat{k}}$, $b_i \geq \frac{\varepsilon}{2}$. It follows that

$$\begin{aligned}\left| b_{n_{\hat{k}}} - b_{m_k} \right| &\leq L \left(\sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i b_i + \left\| \sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i w_i \right\| \right) \\ &= L \left(\sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i b_i^p b_i^{1-p} + \left\| \sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i w_i \right\| \right) \\ &\leq L \left(\sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i b_i^p (\varepsilon/2)^{1-p} + \left\| \sum_{i=m_k}^{n_{\hat{k}}-1} \alpha_i w_i \right\| \right) < L(\frac{\varepsilon}{4LC_\varepsilon} C_\varepsilon + \frac{\varepsilon}{4L}) = \frac{\varepsilon}{2},\end{aligned}$$

which contradicts that $b_{m_k} \geq \varepsilon$ and $b_{n_{\hat{k}}} \leq \frac{\varepsilon}{2}$.

Proof of Claim 1: It is straightforward to verify by definition that it is a martingale. It is well known (see, e.g., (Williams, 1991, Theorem 12.1)) that M_t is bounded in $\mathcal{L}^2$ if and only if

$$\sum_{t=1}^{\infty} \mathbb{E}[\|M_t - M_{t-1}\|^2] < \infty.$$

The latter is verified by

$$\begin{aligned}\sum_{t=1}^{\infty} \mathbb{E}[\|M_t - M_{t-1}\|^2] &= \sum_{t=1}^{\infty} \alpha_t^2 (\mathbb{E}\, \|g_t\|^2 - \mathbb{E}\, \|\nabla f(x_t)\|^2) \\ &\leq \sum_{t=1}^{\infty} \alpha_t^2 \left[A(\mathbb{E}[f(x_t)] - f^*) + (B-1)\, \mathbb{E}\, \|\nabla f(x_t)\|^2 + C \right],\end{aligned} \tag{33}$$

where we used Assumption 4. Following the same argument as in the proof of Theorem 8, except that we take expectation on all the inequalities involved, we can show that $\mathbb{E}[f(x_t)] - f^*$ converges as $t \to \infty$ and $\sum_{t=1}^{\infty} \alpha_t \mathbb{E} \|\nabla f(x_t)\|^2 < \infty$. Since $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$, we have $\alpha_t \to 0$ as $t \to 0$. By comparing the series on the right-hand side of (33) with convergent series $\sum_{t=1}^{\infty} \alpha_t^2$ and $\sum_{t=1}^{\infty} \alpha_t \mathbb{E} \|\nabla f(x_t)\|^2$, we conclude that $\sum_{t=1}^{\infty} \mathbb{E}[\|M_t - M_{t-1}\|^2] < \infty$.

Proof of Claim 2: By L-smoothness of f, we have

$$\sum_{i=1}^{t} \|\alpha_i(\nabla f(x_i) - \nabla f(z_i))\| \le \sum_{i=1}^{t} \alpha_i L \|x_i - z_i\| = \frac{L\beta}{1-\beta} \sum_{i=1}^{t} \alpha_i \|v_i\| \tag{34}$$

$$\le \frac{L\beta}{1-\beta} \sqrt{\sum_{i=1}^{t} \alpha_i^2} \sqrt{\sum_{i=1}^{t} \|v_i\|^2}. \tag{35}$$

It follows that N_t converges almost surely, provided that $\sum_{t=1}^{\infty} \|v_t\|^2 < \infty$ almost surely. To show the latter, recall (23) as

$$\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \le (1 + c_6\alpha_t^2)[f(z_t) - f^* + \|v_t\|^2] - (1-\lambda)\|v_t\|^2$$
$$- c\alpha_t \|\nabla f(z_t)\|^2 + c_4\alpha_t^2, \tag{36}$$

where $c_6 = \max(c_1, c_2)$. Proposition 2 implies that $\sum_{t=1}^{\infty} \|v_t\|^2 < \infty$ almost surely.

Appendix D. Proof of Theorem 13

Lemma 19 *Suppose that Y_t is a sequence of nonnegative random variables that are adapted to a filtration $\{\mathcal{F}_t\}$. Let $\{\alpha_t\}$ be a sequence chosen as $\alpha_t = \Theta\left(\frac{1}{t^{\frac{2}{3}+\varepsilon}}\right)$ (for $t \ge 1$), where $\varepsilon \in (0, \frac{1}{3})$. If*

$$\mathbb{E}[Y_{t+1} \,|\, \mathcal{F}_t] \le (1 + c_1\alpha_t^2)Y_t + c_2\alpha_t^2, \tag{37}$$

for some constants $c_1, c_2 > 0$ and $\sum_{t=1}^{\infty} \alpha_t Y_t < \infty$ almost surely, then $Y_t = O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right)$ almost surely.

Proof Suppose that

$$\frac{\eta_1}{t^{\frac{2}{3}+\varepsilon}} \le \alpha_t \le \frac{\eta_2}{t^{\frac{2}{3}+\varepsilon}}, \quad \forall t \ge 1,$$

with some positive constants η_1 and η_2. Multiplying both sides of (37) by $(1+t)^{\frac{1}{3}-\varepsilon}$ leads to

$$\begin{aligned}
\mathbb{E}_t[(1+t)^{\frac{1}{3}-\varepsilon}Y_{t+1}\,|\,\mathcal{F}_t] &\le (1+t)^{\frac{1}{3}-\varepsilon}(1+c_1\alpha_t^2)Y_t + c_2(1+t)^{\frac{1}{3}-\varepsilon}\alpha_t^2 \\
&\le \left[t^{\frac{1}{3}-\varepsilon} + \left(\frac{1}{3}-\varepsilon\right)t^{-\frac{2}{3}-\varepsilon}\right](1+c_1\alpha_t^2)Y_t + c_2(1+t)^{\frac{1}{3}-\varepsilon}\alpha_t^2 \\
&= (1+c_1\alpha_t^2)t^{\frac{1}{3}-\varepsilon}Y_t + \left(\frac{1}{3}-\varepsilon\right)t^{-\frac{2}{3}-\varepsilon}(1+c_1\alpha_t^2)Y_t + c_2(1+t)^{\frac{1}{3}-\varepsilon}\alpha_t^2 \\
&\le (1+c_1\alpha_t^2)t^{\frac{1}{3}-\varepsilon}Y_t + (c_1\eta_2^2+1)\left(\frac{1}{3}-\varepsilon\right)t^{-\frac{2}{3}-\varepsilon}Y_t + \frac{c_2\eta_2^2}{t^{1+3\varepsilon}}\frac{(1+t)^{\frac{1}{3}-\varepsilon}}{t^{\frac{1}{3}-\varepsilon}} \\
&\le (1+c_1\alpha_t^2)t^{\frac{1}{3}-\varepsilon}Y_t + c_3\alpha_tY_t + \frac{c_4}{t^{1+3\varepsilon}},
\end{aligned}$$

where we can take $c_3 = (c_1\eta_2^2+1)\left(\frac{1}{3}-\varepsilon\right)/\eta_1$ and $c_4 = c_2\eta_2^2\sqrt[3]{2}$. Recall that $\sum_{t=1}^{\infty}\alpha_tY_t < \infty$. Applying Proposition 2 with $t^{\frac{1}{3}-\varepsilon}Y_t$ in place of Y_t, $X_t = 0$, and $Z_t = c_3\alpha_tY_t + \frac{c_4}{t^{1+3\varepsilon}}$, we have $\sum_{t=1}^{\infty}Z_t < \infty$ and $t^{\frac{1}{3}-\varepsilon}Y_t$ converges almost surely. The conclusion follows. ■

With this lemma, we are ready to present the proof of Theorem 13.

Proof 1) We show the proof for SGD first. By smoothness of f and (2), we have

$$f(x_{t+1}) \le f(x_t) - \alpha_t\langle\nabla f(x_t), g_t\rangle + \frac{L\alpha_t^2}{2}\|g_t\|^2.$$

Taking conditional expectation w.r.t. x_t, denoted by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|x_t]$, leads to

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\le f(x_t) - f^* - \alpha_t\|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\le (1+\frac{LA\alpha_t^2}{2})(f(x_t)-f^*) - \left(\alpha_t - \frac{LB\alpha_t^2}{2}\right)\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\le (1+\frac{LA\alpha_t^2}{2})(f(x_t)-f^*) - \frac{1}{2}\alpha_t\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2}, \qquad (38)
\end{aligned}$$

provided that $LB\alpha_t \le 1$.

Let x_* be a minimizer, i.e., $f(x_*) = f^*$. We have

$$\|x_{t+1} - x_*\|^2 = \|x_t - x_*\|^2 - 2\alpha_t\langle g_t, x_t - x_*\rangle + \alpha_t^2\|g_t\|^2.$$

Take conditional expectation w.r.t. x_t from both side. By convexity and L-smoothness of f, we obtain

$$\begin{aligned}
\mathbb{E}_t\left[\|x_{t+1}-x_*\|^2\right] &\leq \|x_t-x_*\|^2 - 2\alpha_t\langle\nabla f(x_t), x_t-x_*\rangle \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\leq \|x_t-x_*\|^2 - 2\alpha_t\left(f(x_t)-f^* + \frac{1}{2L}\|\nabla f(x_t)\|^2\right) \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&= \|x_t-x_*\|^2 - (2\alpha_t - A\alpha_t^2)(f(x_t)-f^*) - \left(\frac{1}{L}\alpha_t - B\alpha_t^2\right)\|\nabla f(x_t)\|^2 \\
&\quad + \alpha_t^2 C \\
&\leq \|x_t-x_*\|^2 - \alpha_t(f(x_t)-f^*) + \alpha_t^2 C, \qquad (39)
\end{aligned}$$

where the second inequality follows from (5) with $y = x_t$ and $x = x_*$, provided that $A\alpha_t \leq 1$, in addition to $LB\alpha_t \leq 1$.

By (38) and Proposition 2, $\sum_{t=1}^\infty \alpha_t \|\nabla f(x_t)\|^2 < \infty$ almost surely and $f(x_t)$ converges almost surely. By (39) and Proposition 2, $\sum_{t=1}^\infty \alpha_t(f(x_t)-f^*) < \infty$ almost surely and $\|x_{t+1}-x_*\|$ converges almost surely. Since $\sum_{t=1}^\infty \alpha_t = \infty$, the almost sure limit of $f(x_t)$ must be f^*. By almost sure convergence of $\|x_{t+1}-x_*\|$, $\{x_t\}$ almost surely has a convergent subsequence. The limit of this subsequence, denoted by $x(\omega)$ must satisfy $f(x(\omega)) = f^*$. Hence $x(\omega)$ is also a minimizer. Since the choice of minimizer in (39) is arbitrary, we must have x_t converges almost surely to some random variable. It follows that $\nabla f(x_t)$ exists almost surely and the limit must be 0 (either by using the fact that the limit of x_t is a minimizer almost surely or that $\sum_{t=1}^\infty \alpha_t \|\nabla f(x_t)\|^2 < \infty$ almost surely and $\sum_{t=1}^\infty \alpha_t = \infty$).

We now derive a concrete convergence rate for $f(x_t)-f^*$. Let $Y_t = f(x_t)-f^*$. By (38) (and dropping the term $-\frac{1}{2}\alpha_t\|\nabla f(x_t)\|^2$), (37) of Lemma 19 holds with $c_1 = \frac{LA}{2}$ and $c_2 = \frac{LC}{2}$. The conclusion follows from that of Lemma 19.

2) We now prove the case for SHB. Recall (23) as

$$\begin{aligned}
\mathbb{E}_t\left[f(z_{t+1})-f^* + \|v_{t+1}\|^2\right] &\leq (1+c_6\alpha_t^2)[f(z_t)-f^* + \|v_t\|^2] - (1-\lambda)\|v_t\|^2 \\
&\quad - c\alpha_t\|\nabla f(z_t)\|^2 + c_4\alpha_t^2, \qquad (40)
\end{aligned}$$

where $c_6 = \max(c_1, c_2)$ defined in (23). Proposition 2 implies that $\sum_{t=1}^\infty \|v_t\|^2 < \infty$, $f(z_t) - f^*$ converges, and $\sum_{t=1}^\infty \alpha_t \|\nabla f(z_t)\|^2 < \infty$, almost surely.

Similar to (39), by convexity of f and iterates of SHB in (18), we obtain

$$\begin{aligned}
\mathbb{E}_t\left[\|z_{t+1}-x_*\|^2\right] &\le \|z_t-x_*\|^2 - \frac{2\alpha_t}{1-\beta}\langle \nabla f(x_t), z_t-x_*\rangle \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*)+B\left\|\nabla f(x_t)\right\|^2+C\right] \\
&= \|z_t-x_*\|^2 - \frac{2\alpha_t}{1-\beta}\langle \nabla f(z_t), z_t-x_*\rangle + \frac{2\alpha_t}{1-\beta}\langle \nabla f(z_t)-f(x_t), z_t-x_*\rangle \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*)+B\left\|\nabla f(x_t)\right\|^2+C\right] \\
&\le \|z_t-x_*\|^2 - \frac{2\alpha_t}{1-\beta}\left(f(z_t)-f^*+\frac{1}{2L}\left\|\nabla f(z_t)\right\|^2\right) + \frac{\beta^2L^2}{(1-\beta)^4}\left\|v_t\right\|^2 \\
&\quad + \alpha_t^2\left\|z_t-x_*\right\|^2 + \alpha_t^2\left[A(f(x_t)-f^*)+B\left\|\nabla f(x_t)\right\|^2+C\right] \\
&\le (1+\alpha_t^2)\left\|z_t-x_*\right\|^2 - (\frac{2\alpha_t}{1-\beta}-c_7\alpha_t^2)(f(z_t)-f^*) \\
&\quad - \left(\frac{1}{L(1-\beta)}\alpha_t - c_8\alpha_t^2\right)\left\|\nabla f(z_t)\right\|^2 + c_9\left\|v_t\right\|^2+\alpha_t^2C,
\end{aligned} \tag{41}$$

where c_7, c_8, and c_9 are some positive constants. The first inequality above follows from convexity of f, L-Lipschitzness of ∇f, and the elementary inequality $2\langle a,b\rangle \le \|a\|^2+\|b\|^2$. The third inequality follows from (21) and (22). By (41), choosing α_t sufficiently small leads to

$$\mathbb{E}_t\left[\|z_{t+1}-x_*\|^2\right] \le (1+\alpha_t^2)\left\|z_t-x_*\right\|^2 - \frac{\alpha_t}{1-\beta}(f(z_t)-f^*+\|v_t\|^2) + c_{10}\left\|v_t\right\|^2+\alpha_t^2C, \tag{42}$$

where $\frac{\alpha_t}{1-\beta}+c_9 \le c_{10}$. Since $\sum_{t=1}^{\infty}\|v_t\|^2 < \infty$ almost surely, Proposition 2 implies that $\sum_{t=1}^{\infty}\frac{\alpha_t}{1-\beta}(f(z_t)-f^*+\|v_t\|^2)<\infty$ and $\|z_t-x_*\|^2$ converges almost surely. By a similar argument as in the proof for SGD, we have z_t converges to a minimizer almost surely. To obtain a concrete convergence rate, let $Y_t = f(z_t)-f^*+\|v_t\|^2$. By the choice of α_t and Lemma 19, we have

$$Y_t = f(z_t)-f^*+\|v_t\|^2 = O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right), \quad \text{almost surely.}$$

From (21) and the fact that

$$\|\nabla f(z_t)\|^2 \le 2L(f(z_t)-f^*),$$

which is from (4), we obtain

$$f(x_t)-f^* = O\left(\frac{1}{t^{\frac{1}{3}-\varepsilon}}\right), \quad \text{almost surely.}$$

3) The case for SNAG is very similar in view of (28) and omitted. ■

Appendix E. Proof of Theorem 17

This section is dedicated to the proof of Theorem 17.

E.1 Preliminary convergence results and property of limit sets

Consider the continuous-time gradient flow for the objective function f:

$$\dot{x} = -\nabla f(x). \tag{43}$$

Let $\{\Phi_t\}$ denote the flow associated with (43), i.e., Φ_t maps any initial condition x to the value of the solution to (43) at time t, $\Phi_t(x)$.

The following lemma is purely deterministic, but can be used to show limit points of the sequences produced by SHB (16) and SNAG (24) basically enjoy the same properties as the omega limit sets of the trajectories of the gradient flow (43).

Lemma 20 (Benaïm (1996)) *Let $\{z_n\}$, $\{u_n\}$, and $\{b_n\}$ be sequences in $\mathbb{R}^d$ such that*

$$z_{n+1} = z_n + \alpha_n(-\nabla f(z_n) + u_n + b_n),$$

where $\{\alpha_n\}$ is a positive decreasing sequence satisfying $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\lim_{n\to\infty} \alpha_n = 0$. Assume:

1. *$\{z_n\}$ is bounded;*

2. *$\lim_{n\to\infty} b_n = 0$; and*

3. *for each $T > 0$,*

$$\lim_{n\to\infty} \sup_{\{k\colon 0\le\tau_k-\tau_n\le T\}} \left\| \sum_{i=n}^{k-1} \alpha_i u_i \right\| = 0,$$

where $\tau_n = \sum_{i=1}^{n} \alpha_i$. Then the limit set of $\{z_n\}$ is a nonempty, compact, connected set which is invariant under the flow $\{\Phi_t\}$ of (43). Furthermore, the limit set belongs to the chain recurrent set of (43).

We state another lemma that asserts convergence properties of the sequences produced by SHB (18) and SNAG (26). While its proof is already included in the proof of Theorem 11 in Appendix C, we state the conclusions separately for clarity.

Lemma 21 *Suppose that Assumptions 1 and 4 hold. Furthermore, $\{\alpha_n\}$ satisfies*

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

Then the following results hold:

1. *$\nabla f(x_n) \to 0$, $\nabla f(z_n) \to 0$, and $v_n \to 0$, as $n \to \infty$, almost surely;*

2. *$\sum_{i=1}^{n} \alpha_i(\nabla f(x_n) - g_n)$ is a martingale bounded in L^2 and hence converges almost surely.*

Based on these two lemmas, we can prove the following result.

Proposition 22 *Suppose that Assumptions 1, 2, and 4 and $\{\alpha_n\}$ is a positive decreasing sequence that satisfies*

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty.$$

Then the sequence $\{z_n\}$ obtained from SHB (16) and SNAG (24) almost surely satisfies the assumptions of Lemma 20. Hence its limit set satisfies the conclusion of Lemma 20 almost surely.

Proof For SHB, write

$$z_{n+1} = z_n + \frac{\alpha_n}{1-\beta} \left(-\nabla f(z_n) + (\nabla f(x_n) - g_n) + (\nabla f(z_n) - \nabla f(x_n)) \right).$$

Let $u_n = \nabla f(x_n) - g_n$ and $b_n = \nabla f(z_n) - \nabla f(x_n)$. Then Lemma 21 implies that $b_n \to 0$ as $n \to \infty$ and $\sum_{i=1}^{n} \alpha_i u_i$ converges. It follows that

$$\lim_{n\to\infty} \sup_{k \geq n+1} \left\| \sum_{i=n}^{k-1} \alpha_i u_i \right\| = 0.$$

Boundedness of $\{z_n\}$ follows from the fact that $\nabla f(z_n) \to 0$ as $n \to \infty$ (Lemma 21) and Assumption 2.

Hence the assumptions of Lemma 20 are met and its conclusion follows. The proof for SNAG (24) is almost identical and therefore omitted. ■

Remark 23 *Since $z_n = x_n - \frac{\beta}{1-\beta} v_n$ and $v_n \to 0$, the limit sets of $\{x_n\}$ and $\{z_n\}$ coincide and hence both enjoy the property stated in the conclusion of Lemma 20.*

E.2 Lyapunov analysis around strict saddle manifold

The saddle avoidance analysis relies on the construction of a Lyapunov function around the saddle manifold due to (Benaïm, 1999, Proposition 9.5).

In this section, we assume that f is three times continuously differentiable[6]. Since $\mathcal{S}$ is a strict saddle manifold, the center manifold theorem (Robinson, 2012; Shub, 1987) implies that there exists a submanifold $\mathcal{M}$ of $\mathbb{R}^d$, namely the center stable manifold of $\mathcal{S}$, that is locally invariant under the flow $\{\Phi_t\}$ in the sense that there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$ and a positive time t_0 such that $\Phi_t(\mathcal{U} \cap \mathcal{M}) \subset \mathcal{M}$ for all $|t| \leq t_0$. Furthermore, for each $x_* \in \mathcal{S}$, we have $\mathbb{R}^d = T_{x_*}\mathcal{M} \oplus E^u_{x_*}$, where $E^u_{x_*}$ is the unstable subspace of $\mathbb{R}^n$ for (43) at x_*. Due to the assumption on the strict saddle manifold, the dimension of $E^u_{x_*}$ is at least one

[6] This stringent requirement is to ensure that the vector field of the gradient flow (43) is twice continuously differentiable, which in turn ensures that V defined in Proposition 24 is twice continuously differentiable ((Benaïm, 1999, Proposition 9.5); cf. (Benaïm and Hirsch, 1995, beginning of Section 3 and proof of Proposition 3.1) for more clarity on the smoothness requirement of the vector field.)

and the dimension of $\mathcal{M}$ is at most $d-1$. Relying on center manifold theory and geometric arguments, one can construct a Lyapunov function V based on the following function

$$\rho(y) = \|\Pi(y) - y\|,$$

which maps from a neighborhood $\mathcal{U}_0$ of $\mathcal{S}$ to $\mathbb{R}_{\geq 0}$, where $\Pi(y)$ projects y on $\mathcal{M}$ along the unstable directions of (43). The following result was proved in Benaïm (1999) (see also Mertikopoulos et al. (2020) for more specific discussions to strict saddle manifold as defined by Definition 16).

Proposition 24 (Benaïm (1999)) *There exists a compact neighborhood $\mathcal{U}_\mathcal{S}$ of $\mathcal{S}$ and positive constants τ and c such that the function $V : \mathcal{U}_\mathcal{S} \to \mathbb{R}$ given by*

$$V(x) = \int_0^\tau \rho\left(\Phi_{-t}(x)\right) dt,$$

where $\{\Phi_t\}$ is the flow generated by (43), satisfies the following properties:

1. *V is twice continuously differentiable on $\mathcal{U}_\mathcal{S} \setminus \mathcal{M}$. For all $x \in \mathcal{U}_\mathcal{S} \cap \mathcal{M}$, V admits a right derivative $DV(x) : \mathbb{R}^d \to \mathbb{R}^d$ which is Lipschitz, convex, and positively homogeneous.*

2. *For all $x \in \mathcal{U}_\mathcal{S}$,*
$$DV(x)[-\nabla f(x)] \geq cV(x).$$

3. *There exists a positive constant C such that, for all $x \in \mathcal{U}_\mathcal{S}$,*
$$DV(x)[v] \geq -C\|v\|, \tag{44}$$
for all $v \in \mathbb{R}^d$.

4. *There exists a constant $\gamma > 0$ and a neighborhood V of the origin of $\mathbb{R}^d$ such that for all $x \in \mathcal{U}_\mathcal{S}$ and $v \in V$, we have*
$$V(x+v) \geq V(x) + DV(x)[v] - \frac{\gamma}{2}\|v\|^2. \tag{45}$$

5. *There exists a constant $m > 0$ such that for all $x \in \mathcal{U}_\mathcal{S} \setminus \mathcal{M}$,*
$$\|\nabla V(x)\| \geq m,$$
and for all $x \in \mathcal{U}_\mathcal{S} \cap \mathcal{M}$ and $v \in \mathbb{R}^d$,
$$DV(x)[v] \geq m\|v - D\Pi(x)v\|.$$

For more details on this construction and proof of the above proposition[7], readers are referred to (Benaïm, 1999, Proposition 9.5) (see also (Mertikopoulos et al., 2020, Appendix C)). For more background information on the topic, we refer the readers to Benaïm and Hirsch (1995); Benaïm (1999); Lee (2012); Shub (1987); Robinson (2012).

[7] Proposition 24(3) was not explicitly stated in Benaïm (1999), but can be easily derived from (35) and (36) in the proof of (Benaïm, 1999, Proposition 9.5).

E.3 Almost sure saddle avoidance analysis

In this section, we analyze almost sure avoidance of any strict saddle manifold. The following lemma due to (Pemantle, 1992, Lemma 5.5) plays an important role in the probabilistic argument of the proof.

Lemma 25 (Pemantle (1992)) *Let $\{S_n\}$ be a nonnegative stochastic process defined as $S_n = S_1 + \sum_{i=2}^n Z_i$, where $\{Z_n\}$ is adapted to a filtration $\{\mathcal{F}_n\}$. Suppose that $\{\alpha_n\}$ satisfies $\alpha_n = \Theta\left(\frac{1}{n^p}\right)$, where $\frac{1}{2} < p \leq 1$. Suppose there exist positive constants b_1, b_2, and b_3 such that the following hold almost surely for all n sufficiently large:*

1. $\|Z_{n+1}\| \leq b_1\alpha_n$;
2. $\mathbf{1}_{\{S_n > b_2\alpha_n\}}\,\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq 0$;
3. $\mathbb{E}[S_{n+1}^2 - S_n^2 \mid \mathcal{F}_n] \geq b_3\alpha_n^2$.

Then $\mathbb{P}(\lim_{n\to\infty} S_n = 0) = 0$.

The lemma was proved in (Pemantle, 1992, Lemma 5.5) for $p = 1$, but the proof for $\frac{1}{2} < p \leq 1$ is the same. The same result was proved and used in (Pemantle, 1990, Theorem 1), but not explicitly stated as a standalone lemma. See (Benaïm, 1999, Lemma 9.6) for a more general form of this result, with a proof using the same technique as Pemantle (1992).

We need another technical lemma that states when the sequence $\{x_n\}$ is uniformly bounded, then by Assumption 5, we can obtain a uniform rate of convergence by v_n to zero, at least for $\{\alpha_n\}$ chosen as in Lemma 25.

Lemma 26 *Let $\{x_n\}$ and $\{v_n\}$ be obtained from SHB (18) or SNAG (26) with $\{\alpha_n\}$ satisfying $\alpha_n = \Theta(\frac{1}{n^p})$, where $\frac{1}{2} < p \leq 1$. Suppose that there exists some constant $B_0 > 0$ such that $\|x_n\| \leq B_0$ for all $n \geq 1$. Then $\|v_n\| = O(\frac{1}{n^p})$.*

Proof For SHB, we have

$$\|v_{n+1}\|^2 = \beta^2\|v_n\|^2 - 2\beta\alpha_n\langle g_n, v_n\rangle + \alpha_n^2\|g_n\|^2.$$

For SNAG, we have

$$\|v_{n+1}\|^2 = \beta^2\|v_n\|^2 - 2\beta^2\alpha_n\langle g_n, v_n\rangle + \beta^2\alpha_n^2\|g_n\|^2.$$

In either case, using the elementary inequality that $2\langle a, b\rangle \leq \varepsilon\|a\|^2 + \frac{1}{\varepsilon}\|b\|^2$, we can find two constants $\varepsilon > 0$ and $C_0 > 0$, where C_0 only depends on β and $\varepsilon > 0$ can be made arbitrarily small, such that

$$\|v_{n+1}\|^2 \leq (\beta^2 + \varepsilon)\|v_n\|^2 + C_0\alpha_n^2\|g_n\|^2.$$

We can choose ε such that $\beta^2 + \varepsilon < 1$. Let $\lambda = \beta^2 + \varepsilon$. By Assumption 5, there exists another constant C_1 such that

$$\|v_{n+1}\|^2 \leq \lambda\|v_n\|^2 + C_1\alpha_n^2.$$

First, observe that the above implies

$$\|v_{n+1}\|^2 - \|v_n\|^2 \leq C_1\alpha_n^2.$$

Summing both sides from 1 to m shows $\|v_{m+1}\|^2 \leq \|v_1\|^2 + C_1\sum_{n=1}^{m}\alpha_n^2$, which shows that $\left\{\|v_n\|^2\right\}$ is uniformly bounded, provided the uniform bound on $\{\|x_n\|\}$.

To show a specific rate estimate for $\{\|v_n\|\}$, we claim that for n sufficiently large, $\|v_n\|^2 = O\left(\frac{1}{n^{2p}}\right)$, i.e., there exists some constant C_2 such that $\|v_n\|^2 \leq \frac{C_2}{n^{2p}}$. Since $\alpha_n^2 = \Theta\left(\frac{1}{n^{2p}}\right)$, there exist positive constants A_1 and B_1 such that

$$\frac{A_1}{n^{2p}} \leq \alpha_n^2 \leq \frac{B_1}{n^{2p}},$$

for all n. Fix any $\mu \in (\lambda, 1)$. Choose C_2 such that $C_2 \geq \frac{C_1B_1}{\mu-\lambda}$. By induction, suppose $\|v_n\|^2 \leq \frac{C_2}{n^{2p}}$ holds for some n such that $\frac{n^{2p}}{(n+1)^{2p}} \geq \mu$ (which also holds for all subsequent n). We have

$$\begin{aligned}
\|v_{n+1}\|^2 \leq \lambda\|v_n\|^2 + C_1\alpha_n^2 &\leq \frac{\lambda C_2}{n^{2p}} + \frac{C_1B_1}{n^{2p}} \\
&= \frac{C_2}{(n+1)^{2p}} - \frac{C_2}{(n+1)^{2p}} + \frac{\lambda C_2}{n^{2p}} + \frac{C_1B_1}{n^{2p}} \\
&\leq \frac{C_2}{(n+1)^{2p}} - \frac{\mu C_2 - \lambda C_2 - C_1B_1}{n^{2p}} \\
&\leq \frac{C_2}{(n+1)^{2p}},
\end{aligned}$$

by the choice of C_2. Hence the estimate holds for all n sufficiently large. ■

With these preliminary results, we are ready to prove Theorem 17.

Proof of Theorem 17

Let $\mathcal{U}_{\mathcal{S}}$ be the neighborhood defined in Proposition 24. Consider the sequences $\{x_n\}$ and $\{z_n\}$ generated from SHB or SNAG. For each $n \geq 1$, let $\mathcal{F}_n$ be the σ-algebra generated by $\{x_1, \ldots, x_n\}$. Without loss of generality, assume $z_1 = x_1 \in \mathcal{U}_{\mathcal{S}}$ (the proof for $z_N \in \mathcal{U}_{\mathcal{S}}$ for any N is identical). For any $k \geq 1$, define the stopping time

$$T_{\mathcal{S}}^k = \inf\left\{n \geq 1:\ z_n \notin \mathcal{U}_{\mathcal{S}} \text{ or } \|x_n\| > k\right\},$$

which is the first exit time of $\{z_n\}$ from $\mathcal{U}_{\mathcal{S}}$ or $\{x_n\}$ from the k-radius ball. Define two sequences of random variables $\{Z_n\}$ and $\{S_n\}$ as follows[8]:

$$Z_{n+1} = (V(z_{n+1}) - V(z_n))\mathbf{1}_{\{n\leq T_{\mathcal{S}}^k\}} + \alpha_n\mathbf{1}_{\{n>T_{\mathcal{S}}^k\}}, \quad n \geq 1, \tag{46}$$

and

$$S_1 = V(z_1), \quad S_n = S_1 + \sum_{i=2}^{n} Z_i, \quad n \geq 2. \tag{47}$$

[8] Note that we should have a superscript k on $\{Z_n\}$ and $\{S_n\}$ as they depend on k, but we omit it to simplify the notation.

It follows that $\{Z_n\}$ is adapted to $\{\mathcal{F}_n\}$. Clearly, if $T^k_{\mathcal{S}} = \infty$, then $S_n = V(z_n)$ for all $n \geq 1$ by telescoping.

We verify that $\{Z_n\}$ and $\{S_n\}$ defined above satisfy the conditions of Lemma 25.

Condition 1: It is clearly satisfied if $n > T^k_{\mathcal{S}}$. If $n \leq T^k_{\mathcal{S}}$, since V is locally Lipschitz and the stochastic gradient is locally bounded (Assumption 5), we have $\|Z_{n+1}\| \leq b_1\alpha_n$ for some $b_1 > 0$.

Condition 2: If $n > T^k_{\mathcal{S}}$, we have $Z_{n+1} = \alpha_n$ and

$$\mathbf{1}_{\{n>T^k_{\mathcal{S}}\}}\,\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq \mathbf{1}_{\{n>T^k_{\mathcal{S}}\}}\alpha_n \geq 0. \tag{48}$$

If $n \leq T^k_{\mathcal{S}}$, we have $z_n \in \mathcal{U}_{\mathcal{S}}$ and, by Proposition 24,

$$\begin{aligned}
Z_{n+1} = V(z_{n+1}) - V(z_n) &\geq DV(z_n)[-\frac{\alpha_n}{1-\beta}g_n] - \frac{\gamma\alpha_n^2}{2(1-\beta)^2}\left\|g_n\right\|^2 \\
&\geq \frac{\alpha_n}{1-\beta}DV(z_n)[-\nabla f(z_n)] + \frac{\alpha_n}{1-\beta}DV(z_n)[\nabla f(z_n) - g_n] - \frac{\gamma\alpha_n^2}{2(1-\beta)^2}\left\|g_n\right\|^2 \\
&\geq \frac{\alpha_n c}{1-\beta}V(z_n) + \frac{\alpha_n}{1-\beta}DV(z_n)[\nabla f(z_n) - g_n] - C_1\alpha_n^2, \qquad (49)
\end{aligned}$$

where $C_1 > 0$ is a constant that can be derived from the bound k for $\{x_n\}$ and Assumption 4. Taking the conditional expectation w.r.t. $\mathcal{F}_n$ gives

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq \frac{\alpha_n c}{1-\beta}V(z_n) + \frac{\alpha_n}{1-\beta}\,\mathbb{E}[DV(z_n)[\nabla f(z_n) - g_n] \mid \mathcal{F}_n] - C_1\alpha_n^2. \tag{50}$$

Now we can use convexity of the right derivative of V (Proposition 24) and the conditional Jensen's inequality to obtain

$$\begin{aligned}
\mathbb{E}[DV(z_n)[\nabla f(z_n) - g_n] \mid \mathcal{F}_n] &\geq DV(z_n)[\nabla f(z_n) - \mathbb{E}[g_n \mid \mathcal{F}_n]] \\
&\geq DV(z_n)[\nabla f(z_n) - \nabla f(x_n)] \\
&\geq -C_2\left\|v_n\right\|,
\end{aligned}$$

where $C_2 > 0$ is a constant that can be derived from Proposition 24, the Lipschitz continuity of ∇f, and (17). Putting this back to (50) and using Lemma 26, we obtain

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq \frac{\alpha_n c}{1-\beta}V(z_n) - \frac{\alpha_n C_2\left\|v_n\right\|}{1-\beta} - C_1\alpha_n^2 \geq \frac{\alpha_n c}{1-\beta}V(z_n) - C_3\alpha_n^2, \tag{51}$$

for some $C_3 > 0$. In other words, we have shown

$$\mathbf{1}_{\{n\leq T^k_{\mathcal{S}}\}}\,\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq \mathbf{1}_{\{n\leq T^k_{\mathcal{S}}\}}\left(\frac{\alpha_n c}{1-\beta}V(z_n) - C_3\alpha_n^2\right). \tag{52}$$

Clearly, if we choose $b_2 = \frac{C_3(1-\beta)}{c}$, then $S_n = V(z_n) > b_2\alpha_n$ implies $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq 0$. Condition 2 is verified.

Condition 3: We have

$$\begin{aligned}
\mathbb{E}[S_{n+1}^2 - S_n^2 \mid \mathcal{F}_n] &= \mathbb{E}[Z_{n+1}^2 + 2S_nZ_{n+1} \mid \mathcal{F}_n] \\
&= \mathbb{E}[Z_{n+1}^2 \mid \mathcal{F}_n] + 2S_n\,\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n].
\end{aligned}$$

If $S_n > b_2\alpha_n$, condition 2 implies that $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq 0$ and hence the right-hand side of the above equation is non-negative. If $S_n \leq b_2\alpha_n$, it follows from (48) and (52) that

$$2S_n\,\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \geq -2b_2C_3\alpha_n^3.$$

Hence, to verify condition 3, it suffices to show that there exists a constant $b_4 > 0$ such that

$$\mathbb{E}[Z_{n+1}^2 \mid \mathcal{F}_n] \geq b_4\alpha_n^2,$$

for all n sufficiently large. If $n > T_{\mathcal{S}}^k$, this obviously holds. For $n \leq T_{\mathcal{S}}^k$, we investigate $\mathbb{E}[Z_{n+1}^+ \mid \mathcal{F}_n]$. In view of Jensen's inequality

$$\mathbb{E}[Z_{n+1}^2 \mid \mathcal{F}_n] \geq \mathbb{E}[Z_{n+1}^+ \mid \mathcal{F}_n]^2,$$

we only need to show $\mathbb{E}[Z_{n+1}^+ \mid \mathcal{F}_n] = \Omega(\alpha_n)$. Consider two cases: (i) $z_n \in \mathcal{M}$; (ii) $z_n \notin \mathcal{M}$. If $z_n \notin \mathcal{M}$, the right derivative in (49) becomes the gradient and from it we obtain

$$\begin{aligned}
Z_{n+1} &\geq \frac{\alpha_n}{1-\beta}\langle \nabla V(z_n), (\nabla f(z_n) - \nabla f(x_n)) + (\nabla f(x_n) - g_n)\rangle - C_1\alpha_n^2 \\
&\geq -\frac{\alpha_n C_2}{1-\beta}\|v_n\| + \frac{\alpha_n}{1-\beta}\langle \nabla V(z_n), \nabla f(x_n) - g_n\rangle - C_1\alpha_n^2 \\
&\geq \frac{\alpha_n}{1-\beta}\langle \nabla V(z_n), \nabla f(x_n) - g_n\rangle - C_3\alpha_n^2, \qquad (53)
\end{aligned}$$

where C_1, C_2, and C_3 are as defined above in the proof for condition 2. Taking conditional expectation on the positive part, we obtain

$$\begin{aligned}
\mathbb{E}[Z_{n+1}^+ \mid \mathcal{F}_n] &\geq \frac{\alpha_n}{1-\beta}\,\mathbb{E}[\langle \nabla V(z_n), \nabla f(x_n) - g_n\rangle^+ \mid \mathcal{F}_n] - C_3\alpha_n^2 \\
&\geq \frac{\alpha_n}{1-\beta}\|\nabla V(z_n)\|\, b - C_3\alpha_n^2 \\
&\geq \frac{\alpha_n m b}{1-\beta} - C_3\alpha_n^2, \qquad (54)
\end{aligned}$$

where we used Assumption 6 on the unit vector $-\nabla V(z_n)/\|\nabla V(z_n)\|$ and then Proposition 24. Hence we do have $\mathbb{E}[Z_{n+1}^+ \mid \mathcal{F}_n] = \Omega(\alpha_n)$ in this case.

If $z_n \in \mathcal{M}$, since the dimension of $\mathcal{M}$ is at most $d-1$, we can choose a unit vector u_n such that $\langle u_n, y\rangle = 0$ for all $y \in T_{z_n}\mathcal{M}$. Since $D\Pi(z_n)$ takes values in $T_{z_n}\mathcal{M}$ (Benaïm, 1999, p. 51), we have $\langle u_n, D\Pi(z_n)v\rangle = 0$ for any $v \in \mathbb{R}^d$. In view of (49) and by Proposition 24, we estimate

$$\begin{aligned}
DV(z_n)[\nabla f(z_n) - g_n] &\geq m\,\|(\nabla f(z_n) - g_n) - D\Pi(z_n)[\nabla f(z_n) - g_n]\| \\
&\geq \langle u_n, (\nabla f(z_n) - g_n) - D\Pi(z_n)[\nabla f(z_n) - g_n]\rangle \\
&= \langle u_n, \nabla f(z_n) - g_n\rangle, \qquad (55)
\end{aligned}$$

where the first inequality is by Proposition 24, the second one is Cauchy-Schwartz, and the equality is by the choice of u_n above. Continuing from (55), we obtain

$$\begin{aligned}
\langle u_n, \nabla f(z_n) - g_n\rangle &= \langle u_n, \nabla f(z_n) - \nabla f(x_n)\rangle + \langle u_n, \nabla f(x_n) - g_n\rangle \\
&\geq -\frac{L\beta}{1-\beta}\|v_n\| + \langle u_n, \nabla f(x_n) - g_n\rangle, \qquad (56)
\end{aligned}$$

where we used Lipschitz continuity of ∇f. Putting (55) and (56) in (49) and using Lemma 26 and Assumption 6, we obtain

$$\mathbb{E}[Z_{n+1}^{+} \mid \mathcal{F}_n] \geq \frac{\alpha_n}{1-\beta}\mathbb{E}[\langle u_n, \nabla f(x_n) - g_n\rangle^{+} \mid \mathcal{F}_n] - C_4\alpha_n^2 \geq \frac{\alpha_n}{1-\beta}b - C_4\alpha_n^2,$$

for sufficiently large n, where $b > 0$ is from Assumption 6, and C_4 can be derived from Lemma 26. Hence we have $\mathbb{E}[Z_{n+1}^{+} \mid \mathcal{F}_n] = \Omega(\alpha_n)$ in the second case as well.

Since conditions 1–3 of Lemma 25 are verified, we conclude by Lemma 25 that $\mathbb{P}(S_n \to 0 \text{ as } n \to \infty) = 0$. Recall the slight abuse of notation (see footnote on page 37), we have in fact proved $\mathbb{P}(S_n^k \to 0 \text{ as } n \to \infty) = 0$ for each $k \geq 1$. We now complete the proof of $\mathbb{P}(x_n \to \mathcal{S} \text{ as } n \to \infty) = 0$ (conclusion of Theorem 17) in a few steps.

1) Let Ω_0 denote the event on which the conclusion of Lemma 21 holds. Then $\mathbb{P}(\Omega_0) = 1$. Let B_k denote the event $\{\sup_n \|x_n\| \leq k\} \cap \Omega_0$. By Assumption 2, almost every $\{x_n\}$ will be ultimately bounded, because $\nabla f(x_n) \to 0$ as $n \to \infty$ and $\liminf_{\|x\|\to\infty} \|\nabla f(x)\| > 0$. It follows that $\cup_{k=1}^{\infty} B_k = \Omega_0$.

2) We prove that $\mathbb{P}(T_{\mathcal{S}}^k = \infty) = 0$ for any k. Suppose that there exists some k such that $\mathbb{P}(T_{\mathcal{S}}^k = \infty) > 0$. For almost every path in $\left\{T_{\mathcal{S}}^k = \infty\right\}$, by Proposition 22, the limit set of $\{z_n\}$, denoted by $L(\{z_n\})$ forms an invariant subset of $\mathcal{U}_{\mathcal{S}}$ under the flow $\{\Phi_t\}$. Pick any limit point $z \in L(\{z_n\}) \subset \mathcal{U}_{\mathcal{S}}$, we have $\Phi_t(z) \in L(\{z_n\}) \subset \mathcal{U}_{\mathcal{S}}$ for all $t \geq 0$. By Proposition 24(2), $V(\Phi_t(z)) \geq e^{ct}V(z)$ for all $t > 0$, where $c > 0$. By compactness of $\mathcal{U}_{\mathcal{S}}$, continuity of V, and the fact that V is nonnegative, we must have $V(z) = 0$. Recall that, when $T_{\mathcal{S}}^k = \infty$, by a telescoping sum, $S_n^k = V(z_n)$ for all $n \geq 1$. It follows that $S_n^k = V(z_n) \to 0$. Since $\mathbb{P}(S_n^k \to 0 \text{ as } n \to \infty) = 0$ for each k, we must have $\mathbb{P}(T_{\mathcal{S}}^k = \infty) = 0$. It follows that $\mathbb{P}(T_{\mathcal{S}}^k < \infty) = 1$ for all k.

3) On each B_k, since we have $T_{\mathcal{S}}^k < \infty$ almost surely, it follows that $\{z_n\}$ eventually exits $\mathcal{U}_{\mathcal{S}}$ (in fact infinitely often by repeating the argument in this proof) almost surely. As a result, $z_n \not\to \mathcal{S}$ as $n \to \infty$ on B_k for each k almost surely, and hence entirely on Ω_0. Since $x_n = z_n - \frac{\beta}{1-\beta}v_n$ and $v_n \to 0$ almost surely (Lemma 21), we have $x_n \not\to \mathcal{S}$ as $n \to \infty$ on Ω_0. The proof is complete by noticing $\mathbb{P}(\Omega_0) = 1$. ■