On Almost Sure Convergence Rates 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 & School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China

Editors: Po-Ling Loh and Maxim Raginsky

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.

Keywords: Stochastic gradient descent, stochastic heavy-ball, stochastic Nesterov's accelerated gradient, almost sure convergence rate

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 (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 descents 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.

1.1. Related work

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 not only proved almost sure convergence for iterations of SGD and SHB, they also established convergence rates that are close to optimal (subject to an ε-factor in the rate) for general convex functions. 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.

1.2. Contributions

We summarize the main contributions of the paper in Table 1 relative to exiting 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.
- 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. In addition to the results listed in Table 1, we also obtained another set of results (Theorem 4) 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 1	Sebbouh et al. (2021) Theorem 1	Sebbouh et al. (2021) Theorem 5
SHB	Theorem 2	Theorem 2	Sebbouh et al. (2021) Theorem 5
SNAG	Theorem 3	Theorem 3	Theorem 5

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

1.3. 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., $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ 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

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

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

Assumption 2 (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 2 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 (see, e.g., Nesterov (2003, Theorem 2.1.5))

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

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

$$\mathbb{E}[\left\|\nabla f(x;\xi)\right\|^2] \leq A(f(x) - f^*) + B\left\|\nabla f(x)\right\|^2 + C, \quad \forall x \in \mathbb{R}^d. \tag{5}$$

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 $\mathbb{E}[\left\|\nabla f(x;\xi)\right\|^2] \leq \sigma^2$ and bounded variance condition $\mathbb{E}[\left\|\nabla f(x;\xi) - \nabla f(x)\right\|^2] \leq \sigma^2$ as special cases. The latter is because, by unbiasedness of $\nabla f(x;\xi)$, bounded variance is equivalent to $\mathbb{E}[\left\|\nabla f(x;\xi)\right\|^2] \leq \left\|\nabla f(x)\right\|^2 + \sigma^2$. 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): $\mathbb{E}[\left\|\nabla f(x;\xi)\right\|^2] \leq 4L(f(x) - f^*) + \sigma^2$, 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 3. For these reasons, we shall use the seemingly most general condition in Assumption 3 throughout this paper.*

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 1 *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. *$\mathbb{E}[Y_{t+1} \,|\, \mathcal{F}_t] \leq (1+\gamma_t)Y_t - X_t + Z_t$ 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 1, provides concrete estimates of almost sure convergence rates for sequences of random variables satisfying a supermartingale property.

Lemma 1 *If $\{Y_t\}$ is a sequence of nonnegative random variables satisfying*

$$\mathbb{E}[Y_{t+1} \,|\, \mathcal{F}_t] \leq (1 - c_1\alpha_t)Y_t + c_2\alpha_t^2, \tag{6}$$

for all $t \geq 1$, where $\alpha_t = \Theta\left(\frac{1}{t^{1-\theta}}\right)$ 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 1 can be found in Appendix A. The following lemma, when used together with Proposition 1, is useful for almost sure convergence rate analysis in a slightly different setting than Lemma 1.

Lemma 2 *Let $\{X_t\}$ a sequence of nonnegative real numbers and $\{\alpha_t\}$ be a decreasing sequence of positive real numbers such that the following holds:*

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

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

$$Y_{t+1} = (1 - w_t)Y_t + w_t X_t, \quad t \geq 1. \tag{7}$$

Then

$$Y_t = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right) \quad \textit{and} \quad \min_{1 \leq i \leq t} X_i = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right). \tag{8}$$

Remark 2 *A concrete convergence rate $o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right)$ results from (8) 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 $\sum_t \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$.*

The proof of Lemma 2 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

$$x_{t+1} = x_t - \alpha_t g_t, \quad t \geq 1, \tag{9}$$

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 1 and 2. These rates match the lower bounds in Agarwal et al. (2012) (see also Nemirovskij and Yudin (1983)) to an ε-factor.

Theorem 1 *Consider the iterates of SGD (9).*

1. *If Assumptions 1, 2, and 3 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 3 hold and* $\{\alpha_t\}$ *is a decreasing sequence of positive real numbers satisfying* $\sum_{t=1}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$*, then almost surely*

$$\min_{1 \le i \le t-1} \|\nabla f(x_t)\|^2 = o\left(\frac{1}{\sum_{i=1}^{t-1} \alpha_i}\right). \tag{10}$$

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*

$$\min_{1 \le i \le t-1} \|\nabla f(x_t)\|^2 = o\left(\frac{1}{t^{\frac{1}{2}-\varepsilon}}\right). \tag{11}$$

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

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\le f(x_t) - f^* - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&= (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - (\alpha_t - \frac{LB\alpha_t^2}{2})\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - 2\mu(\alpha_t - \frac{LB\alpha_t^2}{2})(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2} \\
&= (1 - 2\mu\alpha_t + (LA/2 + LB\mu)\alpha_t^2)(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2} \\
&\le (1 - \mu\alpha_t)(f(x_t) - f^*) + \frac{LC\alpha_t^2}{2},
\end{aligned} \tag{12}$$

provided that $(LA/2 + LB\mu)\alpha_t \le \mu$. The conclusion follows from Lemma 1.

2. For the non-convex case, by L-smoothness and as in (12), we obtain

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\le f(x_t) - f^* - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \left(\alpha_t - \frac{LB\alpha_t^2}{2}\right)\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\le (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \frac{1}{2}\alpha_t \|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2},
\end{aligned} \tag{13}$$

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

Remark 3 *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 1. 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 1). 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 1. 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

$$x_{t+1} = x_t - \alpha_t g_t + \beta(x_t - x_{t-1}), \tag{14}$$

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.

Define

$$z_t = x_t - \frac{\beta}{1-\beta} v_t, \quad v_t = x_t - x_{t-1}. \tag{15}$$

The iteration of SHB can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \alpha_t g_t, \\ z_{t+1} &= z_t - \frac{\alpha_t}{1-\beta} g_t. \end{aligned} \tag{16}$$

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 2 *Consider the iterates of SHB (14).*

1. *If Assumptions 1, 2, and 3 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 3 hold and $\{\alpha_t\}$ is a decreasing sequence of positive real numbers satisfying $\sum_{t=1}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, then almost surely*

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

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

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

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

$$\begin{aligned}\mathbb{E}_t \|v_{t+1}\|^2 &= \beta^2 \|v_t\|^2 - 2\beta\alpha_t\langle \nabla f(x_t), v_t\rangle + \alpha_t^2 \left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\ &\leq \beta^2 \|v_t\|^2 + \varepsilon_1 \|v_t\|^2 + \frac{\alpha_t^2}{\varepsilon_1}\|\nabla f(x_t)\|^2 + \alpha_t^2 \left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right]. \quad (17)\end{aligned}$$

By L-smoothness of f and (2), we have

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

Taking conditional expectation w.r.t. x_t gives

$$\begin{aligned}&\mathbb{E}_t f(z_{t+1}) \\ &\leq f(z_t) - \frac{\alpha_t}{1-\beta}\langle \nabla f(z_t), \nabla f(x_t)\rangle + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\ &= f(z_t) - \frac{\alpha_t}{1-\beta}\|\nabla f(z_t)\|^2 - \frac{\alpha_t}{1-\beta}\langle \nabla f(z_t), \nabla f(x_t) - \nabla f(z_t)\rangle \\ &\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\ &\leq f(z_t) - \frac{\alpha_t}{1-\beta}\|\nabla f(z_t)\|^2 + \frac{\alpha_t}{1-\beta}\|\nabla f(z_t)\|\frac{L\beta}{1-\beta}\|v_t\| \\ &\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\ &\leq f(z_t) - \frac{\alpha_t}{1-\beta}\|\nabla f(z_t)\|^2 + \varepsilon_2\|v_t\|^2 + \frac{\alpha_t^2 L^2\beta^2}{\varepsilon_2(1-\beta)^4}\|\nabla f(z_t)\|^2 \\ &\quad + \frac{L\alpha_t^2}{2(1-\beta)^2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right]. \quad (18)\end{aligned}$$

By L-smoothness of f again, we have

$$\begin{aligned}f(x_t) - f^* &\leq f(z_t) - f^* + \frac{\beta}{1-\beta}\langle \nabla f(z_t), v_t\rangle + \frac{L\beta^2}{2(1-\beta)^2}\|v_t\|^2 \\ &\leq f(z_t) - f^* + \frac{1}{2}\|\nabla f(z_t)\|^2 + \frac{\beta^2}{2(1-\beta)^2}\|v_t\|^2 + \frac{L\beta^2}{2(1-\beta)^2}\|v_t\|^2 , \quad (19)\end{aligned}$$

and

$$\begin{aligned}\|\nabla f(x_t)\|^2 &= \|\nabla f(z_t) + \nabla f(x_t) - \nabla f(z_t)\|^2 \leq 2\|\nabla f(z_t)\|^2 + 2\|\nabla f(x_t) - \nabla f(z_t)\|^2 \\ &\leq 2\|\nabla f(z_t)\|^2 + 2\frac{L^2\beta^2}{(1-\beta)^2}\|v_t\|^2 . \quad (20)\end{aligned}$$

Combining (17)–(20) yields

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

where the constants c_1–c_4 can be straightforwardly determined from (17)–(20). For any $\lambda \in (\beta, 1)$, we can choose $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\beta^2 + \varepsilon_1 + \varepsilon_2 \leq \lambda$. 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 $\frac{\alpha_t}{1-\beta} - c_3\alpha_t^2 \geq c\alpha_t$ for all $t \geq 1$. The above inequality becomes

$$\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \leq (1 + c_1\alpha_t^2)[f(z_t) - f^*] + (\lambda + c_2\alpha_t^2)\|v_t\|^2 - c\alpha_t\|\nabla f(z_t)\|^2 + c_4\alpha_t^2. \tag{21}$$

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 1 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 (19) and (4).

2. If f is non-convex, by (20), inequality (21) 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 1, we have $\sum_{t=1}^{\infty}\alpha_t\|\nabla f(x_t)\|^2 < \infty$ almost surely. The conclusions follow from Lemma 2 and Remark 2. ■

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

$$\begin{aligned} y_{t+1} &= x_t - \alpha_t g_t, \\ x_{t+1} &= y_{t+1} + \beta(x_t - x_{t-1}), \end{aligned} \tag{22}$$

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.

Define z_t and v_t as in (15). The iteration of SNAG can be rewritten as

$$\begin{aligned} v_{t+1} &= \beta v_t - \beta \alpha_t g_t, \\ z_{t+1} &= z_t - \frac{\alpha_t}{1-\beta} g_t. \end{aligned} \tag{23}$$

Indeed, (23) is almost identical to (16) except for the extra β in the first equation for v_{t+1}.

Theorem 3 *Consider the iterates of SNAG (22).*

1. *If Assumptions 1, 2, and 3 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 3 hold and $\{\alpha_t\}$ is a decreasing sequence of positive real numbers satisfying $\sum_{t=1}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, then almost surely*

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

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

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

Proof The proof is similar to that for Theorem 2. Instead of (17), we obtain

$$\mathbb{E}_t \|v_{t+1}\|^2 \le \beta^2 \left(\|v_t\|^2 + \varepsilon_1 \|v_t\|^2 + \frac{\alpha_t^2}{\varepsilon_1} \|\nabla f(x_t)\|^2 + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C \right] \right). \tag{24}$$

The rest of the proof proceeds in the same way (with slightly different constants). We conclude the same convergence rates by Lemmas 1 and 2. ■

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[1] $f(x_t) - f^*$. 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.

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

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 2). 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).

Lemma 3 (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 $p \geq 1$. Furthermore, assume that there exists some $L > 0$ such that $|b_{t+\tau} - b_t| \leq L\left(\sum_{i=t}^{t+\tau-1}\alpha_i b_i + \left\|\sum_{i=t}^{t+\tau-1}\alpha_i w_i\right\|\right)$, where w_t is such that $\sum_{t=1}^{\infty}\alpha_t w_t$ converges. Then b_t converges to 0.*

Theorem 4 *Consider the iterates of SHB (14) and SNAG (22), respectively. Let Assumptions 1 and 3 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 (21) and Proposition 1, we have $\sum_{t=1}^{\infty}\alpha_t\|\nabla f(z_t)\|^2 < \infty$ almost surely. Furthermore, by L-smoothness of f, we have

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

where $w_i = g_i - \nabla f(z_i)$. To show that $\sum_{t\geq 0}\alpha_t w_t$ 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 3 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 (19) 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 (24). We omitted the details here. ■

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. We are not aware of any similar last-iterate almost sure convergence rates in the literature.

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

Theorem 5 *Consider the iterates of SGD (9), SHB (14), and SNAG (22), respectively. 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* $x \to x_*$ *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 4 *While this appears to the first result on last-iterate almost sure convergence rates for SGD 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.*

5. 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.

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.

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.

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.

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.

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

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.

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.

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

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.

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

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

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.

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.

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.

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 1

Proof By the choice of α_t, 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

$$(t+1)^{1-\varepsilon} \leq t^{1-\varepsilon} + (1-\varepsilon)t^{-\varepsilon}, \tag{25}$$

which can be proved, for instance, as follows. Let $g(x) = x^{1-\varepsilon}$. Then $g'(x) = (1-\varepsilon)x^{-\varepsilon}$ is decreasing. 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 (25). Multiplying (6) with $(t+1)^{1-\varepsilon}$ and applying inequality (25) lead to

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

Clearly, as $t \to \infty$, the dominating term in $\frac{1-\varepsilon}{t} - \frac{\eta}{t^{1-\theta}} - \frac{\eta(1-\varepsilon)}{t^{2-\theta}}$ is $-\frac{\eta}{t^{1-\theta}}$. Hence, there exists some $T > 1$ sufficiently large such that, for all $t \geq T$,

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

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 1 are met for all $t \geq$T with $\hat{Y}_t$ in place of Y_t. By Proposition 1, 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 $\sum_{t=T}^{\infty} \frac{\eta}{2t^{1-\theta}} = \infty$. The conclusion follows. ■

Appendix B. Proof of Lemma 2

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 (7) we have

$$Y_{t+1}\sum_{i=1}^{t}\alpha_i = Y_t\sum_{i=1}^{t-1}\alpha_i - \alpha_t Y_t + 2\alpha_t X_t, \quad t \geq 1. \tag{26}$$

Let $\hat{Y}_t = Y_t\sum_{i=1}^{t-1}\alpha_i$. Then conditions of Proposition 1 are met with $\hat{Y}_t$ in place of Y_t, $-\alpha_t Y_t$ in place of Y_t, and $2\alpha_2 X_t$ in place of Z_t. It follows from Proposition 1 that $Y_{t+1}\sum_{i=1}^{t}\alpha_i$ converges[2]

2. While no random sequences are involved here, Proposition 1 is still applicable with almost sure convergence replaced by convergence. A direct proof is possible using the monotone convergence theorem for real numbers.

and $\sum_{t=1}^{\infty} \alpha_t Y_t < \infty$. Since $\sum_{t=1}^{\infty} \frac{\alpha_t}{\sum_{i=1}^{t-1} \alpha_i} = \infty$, $\sum_{t=1}^{\infty} \alpha_t Y_t < \infty$, 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 Theorem 4

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

$$\begin{aligned}\sum_{t=1}^{\infty} \mathbb{E}[\|M_t - M_{t-1}\|^2] &= \sum_{t=1}^{\infty} \alpha_t^2 (\mathbb{E}\,\|g_t\|^2 - \mathbb{E}\,\|\nabla f(x_t)\|^2) \\ &\leq \sum_{t=1}^{\infty} \alpha_t^2 \left[A(\mathbb{E}[f(x_t)] - f^*) + (B-1)\,\mathbb{E}\,\|\nabla f(x_t)\|^2 + C\right], \end{aligned} \tag{27}$$

where we used Assumption 3. Following the same argument as in the proof of Theorem 2, 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 (27) 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

$$\sum_{i=1}^{t} \|\alpha_i(\nabla f(x_i) - \nabla f(z_i))\| \leq \sum_{i=1}^{t} \alpha_i L\, \|x_i - z_i\| = \frac{L\beta}{1-\beta} \sum_{i=1}^{t} \alpha_i\, \|v_i\| \tag{28}$$

$$\leq \frac{L\beta}{1-\beta} \sqrt{\sum_{i=1}^{t} \alpha_i^2} \sqrt{\sum_{i=1}^{t} \|v_i\|^2}. \tag{29}$$

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 (21) as

$$\begin{aligned}\mathbb{E}_t\left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2\right] \leq (1 + c_6\alpha_t^2)[f(z_t) - f^* + \|v_t\|^2] - (1-\lambda)\,\|v_t\|^2 \\ - c\alpha_t\,\|\nabla f(z_t)\|^2 + c_4\alpha_t^2, \end{aligned} \tag{30}$$

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

Appendix D. Proof of Theorem 5

Lemma 4 *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 $t \geq 1$), where $\varepsilon \in (0, \frac{1}{3})$. If*

$$\mathbb{E}[Y_{t+1} \,|\, \mathcal{F}_t] \leq (1 + c_1\alpha_t^2)Y_t + c_2\alpha_t^2, \tag{31}$$

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

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

with some positive constants η_1 and η_2. Multiplying both sides of (31) by $(1+t)^{\frac{1}{3}-\varepsilon}$ leads to

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

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 $\sum_{t=1}^{\infty}\alpha_t Y_t < \infty$. Applying Proposition 1 with $t^{\frac{1}{3}-\varepsilon}Y_t$ in place of Y_t, $X_t = 0$, and $Z_t = c_3\alpha_t Y_t + \frac{c_4}{t^{1+3\varepsilon}}$, 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 5.

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

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

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

$$\begin{aligned}
\mathbb{E}_t\left[f(x_{t+1}) - f^*\right] &\leq f(x_t) - f^* - \alpha_t\|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\left[A(f(x_t) - f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\leq (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \left(\alpha_t - \frac{LB\alpha_t^2}{2}\right)\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2} \\
&\leq (1 + \frac{LA\alpha_t^2}{2})(f(x_t) - f^*) - \frac{1}{2}\alpha_t\|\nabla f(x_t)\|^2 + \frac{LC\alpha_t^2}{2},
\end{aligned} \tag{32}$$

provided that $LB\alpha_t \leq 1$.

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 of f, we obtain

$$\begin{aligned}
\mathbb{E}_t \left[\|x_{t+1} - x^*\|^2 \right] &= \|x_t - x_*\|^2 - 2\alpha_t \langle \nabla f(x_t), x_t - x_* \rangle \\
&\quad + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C \right] \\
&\leq \|x_t - x_*\|^2 - 2\alpha_t \left(f(x_t) - f^* + \frac{1}{2L} \|\nabla f(x_t)\|^2 \right) \\
&\quad + \alpha_t^2 \left[A(f(x_t) - f^*) + B \|\nabla f(x_t)\|^2 + C \right] \\
&= \|x_t - x_*\|^2 - (2\alpha_t - A\alpha_t^2)(f(x_t) - f^*) - \left(\frac{1}{L}\alpha_t - B\alpha_t^2 \right) \|\nabla f(x_t)\|^2 \\
&\quad + \alpha_t^2 C \\
&\leq \|x_t - x_*\|^2 - \alpha_t (f(x_t) - f^*) + \alpha_t^2 C, \qquad (33)
\end{aligned}$$

provided that $A\alpha_t \leq 1$, in addition to $LB\alpha_t \leq 1$.

By (32) and Proposition 1, $\sum_{t=1}^{\infty} \alpha_t \|\nabla f(x_t)\|^2 < \infty$ almost surely and $f(x_t)$ converges almost surely. By (33) and Proposition 1, $\sum_{t=1}^{\infty} \alpha_t (f(x_t) - f^*) < \infty$ almost surely and $\|x_{t+1} - x^*\|$ converges almost surely. Since $\sum_{t=1}^{\infty} \alpha_t = \infty$, we have $f(x_t)$ converges to f^* almost surely. 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 (33) is arbitrary, we must have x_t converges almost surely to some random variable. It follows that $\nabla f(x_t)$ exists almost and the limit must be 0 (either by using the fact that the limit of x_t is a minimizer almost surely or that $\sum_{t=1}^{\infty} \alpha_t \|\nabla f(x_t)\|^2 < \infty$ 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 (32) (and dropping the term $-\frac{1}{2}\alpha_t \|\nabla f(x_t)\|^2$), (31) of Lemma 4 holds with $c_1 = \frac{LA}{2}$ and $c_2 = \frac{LC}{2}$. The conclusion follows from that of Lemma 4.

2) We now prove the case for SHB. Recall (21) as

$$\begin{aligned}
\mathbb{E}_t \left[f(z_{t+1}) - f^* + \|v_{t+1}\|^2 \right] &\leq (1 + c_6\alpha_t^2)[f(z_t) - f^* + \|v_t\|^2] - (1 - \lambda) \|v_t\|^2 \\
&\quad - c\alpha_t \|\nabla f(z_t)\|^2 + c_4\alpha_t^2, \qquad (34)
\end{aligned}$$

where $c_6 = \max(c_1, c_2)$ defined in (21). Proposition 1 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 (33), by convexity of f and iterates of SHB in (16), we obtain

$$\begin{aligned}
\mathbb{E}_t\left[\|z_{t+1}-x^*\|^2\right] &= \|z_t-x_*\|^2 - \frac{2\alpha_t}{1-\beta}\langle\nabla f(x_t), z_t-x_*\rangle \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&= \|z_t-x_*\|^2 - \frac{2\alpha_t}{1-\beta}\langle\nabla f(z_t), z_t-x_*\rangle + \frac{2\alpha_t}{1-\beta}\langle\nabla f(z_t)-f(x_t), z_t-x_*\rangle \\
&\quad + \alpha_t^2\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\le \|z_t-x_*\|^2 - 2\alpha_t\left(f(z_t)-f^* + \frac{1}{2L}\|\nabla f(z_t)\|^2\right) + \frac{\beta^2L^2}{(1-\beta)^4}\|v_t\|^2 \\
&\quad + \alpha_t^2\|z_t-x_*\|^2 + \alpha_t^2\left[A(f(x_t)-f^*) + B\|\nabla f(x_t)\|^2 + C\right] \\
&\le (1+\alpha_t^2)\|z_t-x_*\|^2 - (2\alpha_t - c_7\alpha_t^2)(f(z_t)-f^*) \\
&\quad - \left(\frac{1}{L}\alpha_t - c_8\alpha_t^2\right)\|\nabla f(z_t)\|^2 + c_9\|v_t\|^2 + \alpha_t^2 C,
\end{aligned} \tag{35}$$

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 second inequality follows from (19). By (35), choosing α_t sufficiently small leads to

$$\mathbb{E}_t\left[\|z_{t+1}-x^*\|^2\right] \le (1+\alpha_t^2)\|z_t-x_*\|^2 - \alpha_t(f(z_t)-f^*+\|v_t\|^2) + c_{10}\|v_t\|^2 + \alpha_t^2 C, \tag{36}$$

where $\alpha_t + c_9 \le c_{10}$. Since $\sum_{t=1}^\infty \|v_t\|^2 < \infty$, Proposition 1 implies $\sum_{t=1}^\infty \alpha_t(f(z_t)-f^*+\|v_t\|^2) < \infty$ 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 4, we have

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

From (19) and the fact that

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

we obtain

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

3) The case for SNAG is very similar in view of (24) and omitted. ■