Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization

Vien V. Mai [1] **Mikael Johansson** [1]

Abstract

Stochastic gradient methods with momentum are widely used in applications and at the core of optimization subroutines in many popular machine learning libraries. However, their sample complexities have not been obtained for problems beyond those that are convex or smooth. This paper establishes the convergence rate of a stochastic subgradient method with a momentum term of Polyak type for a broad class of non-smooth, non-convex, and constrained optimization problems. Our key innovation is the construction of a special Lyapunov function for which the proven complexity can be achieved without any tuning of the momentum parameter. For smooth problems, we extend the known complexity bound to the constrained case and demonstrate how the unconstrained case can be analyzed under weaker assumptions than the state-of-the-art. Numerical results confirm our theoretical developments.

1. Introduction

We study the stochastic optimization problem

$$\underset{x\in\mathcal{X}}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s), \quad (1)$$

where $S \sim P$ is a random variable; $f(x;s)$ is the instantaneous loss parameterized by x on a sample $s \in \mathcal{S}$; and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set. In this paper, we move beyond convex and/or smooth optimization and consider f that belongs to a broad class of *non-smooth* and *non-convex* functions called ρ-weakly convex, meaning that

$$x \mapsto f(x) + \rho\,\|x\|_2^2 \;\text{ is convex.}$$

This function class is very rich and important in optimization (Rockafellar, 1982; Vial, 1983). It trivially includes all convex functions, all smooth functions with Lipschitz continuous gradient, and all additive composite functions of the two former classes. More broadly, it includes all compositions of the form

$$f(x) = h(c(x)), \quad (2)$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is convex and L_h-Lipschitz and $c : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth map with L_c-Lipschitz Jacobian. Indeed, the composite function $f = h \circ c$ is then weakly convex with $\rho = L_h L_c$ (Drusvyatskiy & Paquette, 2019). Some representative applications in this problem class include nonlinear least squares (Drusvyatskiy, 2017), robust phase retrieval (Duchi & Ruan, 2018a), Robust PCA (Candès et al., 2011), robust low rank matrix recovery (Charisopoulos et al., 2019), optimization of the Conditional Value-at-Risk (Rockafellar & Uryasev, 2000), graph synchronization (Singer, 2011), and many others.

Stochastic optimization algorithms for solving (1), based on random samples S_k drawn from P, are of fundamental importance in many applied sciences (Bottou, 2010; Shapiro et al., 2014). Since the introduction of the classical stochastic (sub)gradient descent method (SGD) in (Robbins & Monro, 1951), several modifications of SGD have been proposed to improve its practical and theoretical performance. A notable example is the use of a *momentum* term to construct an update direction (Gupal & Bazhenov, 1972; Polyak, 1987; Ruszczynski & Syski, 1983; Tseng, 1998; Sutskever et al., 2013; Ghadimi & Lan, 2016). The basis form of such a method (when $\mathcal{X} = \mathbb{R}^n$) reads:

$$x_{k+1} = x_k - \alpha_k z_k \quad (3a)$$

$$z_{k+1} = \beta_k g_{k+1} + (1-\beta_k) z_k, \quad (3b)$$

where z_k is the search direction, g_k is a stochastic subgradient; α_k is the stepsize, and $\beta_k \in (0,1]$ is the momentum parameter. For instance, the scheme (3) reduces to the stochastic heavy ball method (SHB) (Polyak, 1987):

$$x_{k+1} = x_k - \eta_k g_k + \lambda_k(x_k - x_{k-1}),$$

with $\eta_k = \alpha_k\beta_{k-1}$ and $\lambda_k = (1-\beta_{k-1})\alpha_k/\alpha_{k-1}$. Methods of this type enjoy widespread empirical success in large-scale convex and non-convex optimization, especially in

[1]Division of Decision and Control Systems, EECS, KTH Royal Institute of Technology, Stockholm, Sweden. Correspondence to: V. V. Mai <maivv@kth.se>, M. Johansson <mikaelj@kth.se>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

training neural networks, where they have been used to produce several state-of-the-art results on important learning tasks, e.g., (Krizhevsky et al., 2012; Sutskever et al., 2013; He et al., 2016; Huang et al., 2017).

Sample complexity, namely, the number of observations $S_0, \ldots, S_K$ required to reach a desired accuracy ϵ, has been the most widely adapted metric for evaluating the performance of stochastic optimization algorithms. Although sample complexity results for the standard SGD on problems of the form (1) have been obtained for convex and/or smooth problems (Nemirovski et al., 2009; Ghadimi & Lan, 2013), much less is known in the non-smooth non-convex case (Davis & Drusvyatskiy, 2019). The problem is even more prevalent in momentum-based methods as there is virtually no known complexity results for problems beyond those that are convex or smooth.

1.1. Related work

As many applications in modern machine learning and signal processing cannot be captured by convex models, (stochastic) algorithms for solving non-convex problems have been studied extensively. Below we review some of the topics most closely related to our work.

Stochastic weakly convex minimization Earlier works on this topic date back to Nurminskii who showed subsequential convergence to stationary points for the subgradient method applied to deterministic problems (Nurminskii, 1973). The work (Ruszczyński, 1987) proposes a stochastic gradient averaging-based method and shows the first almost sure convergence for this problem class. Basic sufficient conditions for convergence of stochastic projected subgradient methods is established in (Ermol'ev & Norkin, 1998). Thanks to the recent advances in statistical learning and signal processing, the problem class has been reinvigorated with several new theoretical results and practical applications (see, e.g., (Duchi & Ruan, 2018a; 2016; Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019) and references therein). In particular, based on the theory of non-convex differential inclusions, almost sure convergence is derived in (Duchi & Ruan, 2018b) for a collection of model-based minimization strategies, albeit no rates of convergence are given. An important step toward *non-asymptotic* convergence of stochastic methods is made in (Davis & Grimmer, 2019). There, the authors employ a proximal point technique for which they can show the sample complexity $O(1/\epsilon^2)$ with a certain stationarity measure. Later, the work (Davis & Drusvyatskiy, 2019) shows that the (approximate) proximal point step in (Davis & Grimmer, 2019) is not necessary and establishes the similar complexity for a class of model-based methods including the standard SGD. We also note that there has been a large volume of works in smooth non-convex optimization, e.g., (Ghadimi & Lan, 2013; 2016).

Stochastic momentum for non-convex functions Optimization algorithms based on momentum averaging techniques go back to Polyak (1964) who proposed the heavy ball method. In (1983), Nesterov introduced the accelerated gradient method and showed its optimal iteration complexity for the minimization of smooth convex functions. In the last few decades, research on accelerated first-order methods has exploded both in theory and in practice (Beck, 2017; Nesterov, 2018; Bubeck, 2015). The effectiveness of such techniques in the deterministic context has inspired researchers to incorporate momentum terms into stochastic optimization algorithms (Polyak, 1987; Ruszczyński, 1987; Tseng, 1998; Sutskever et al., 2013; Ghadimi & Lan, 2016). Despite evident success, especially, in training neural networks (Krizhevsky et al., 2012; Sutskever et al., 2013; He et al., 2016; Zagoruyko & Komodakis, 2016; Huang et al., 2017), the theory for stochastic momentum methods is not as clear as its deterministic counterpart (cf. Gitman et al. (2019)). As a result, there has been a growing interest in obtaining convergence guarantees for those methods under noisy gradients (Hu et al., 2009; Gitman et al., 2019; Yan et al., 2018; Gadat et al., 2018; Ghadimi & Lan, 2016). In non-convex optimization, almost certain convergence of Algorithm (3) for *smooth and unconstrained* problems is derived in (Ruszczynski & Syski, 1983). Under the *bounded gradient* hypothesis, the convergence rate of the same algorithm has been established in (Yan et al., 2018). The work (Ghadimi et al., 2020) obtains the complexity of a gradient averaging-based method for *constrained* problems. In (Ghadimi & Lan, 2016), the authors study a variant of Nesterov acceleration and establish a similar complexity for smooth and unconstrained problems, while for the constrained case, a mini-batch of samples at each iteration is required to guarantee convergence.

1.2. Contributions

Minimization of weakly convex functions has been a challenging task, especially for stochastic problems, as the objective is neither smooth nor convex. With the recent breakthrough in (Davis & Drusvyatskiy, 2019), this problem class has been the widest one for which provable sample complexity of the standard SGD is known. It is thus intriguing to ask whether such a result can also be obtained for momentum-based methods. The work in this paper aims to address this question. To that end, we make the following contributions:

- We establish the sample complexity of a stochastic subgradient method with momentum of Polyak type for a broad class of non-smooth, non-convex, and constrained optimization problems. Concretely, using a special Lyapunov function, we show the complexity $O(1/\epsilon^2)$ for the minimization of weakly convex functions. The proven complexity is attained in a parameter-free and single

time-scale fashion, namely, the stepsize and the momentum constant are independent of any problem parameters and they have the same scale with respect to the iteration count. To the best of our knowledge, this is the first complexity guarantee for a stochastic method with momentum on non-smooth and non-convex problems.

- We also study the sample complexity of the considered algorithm for smooth and constrained optimization problems. Note that even in this setting, no complexity guarantee of SGD with Polyak momentum has been established before. Under a bounded gradient assumption, we obtain a similar $O(1/\epsilon^2)$ complexity without the need of forming a batch of samples at each iteration, which is commonly required for *constrained* non-convex stochastic optimization (Ghadimi et al., 2016). We then demonstrate how the unconstrained case can be analyzed without the above assumption.

Interestingly, the stated result is achieved in the regime where β can be as small as $O(1/\sqrt{K})$, i.e., one can put much more weight to the momentum term than the fresh subgradient in a search direction. This complements the complexity of SGD attained as $\beta \to 1$. Note that the worst-case complexity $O(1/\epsilon^2)$ is unimprovable in the smooth and unconstrained case (Arjevani et al., 2019).

2. Background

In this section, we first introduce the notation and then provide the necessary preliminaries for the paper.

For any $x, y \in \mathbb{R}^n$, we denote by $\langle x, y\rangle$ the Euclidean inner product of x and y. We use $\|\cdot\|_2$ to denote the Euclidean norm. For a closed and convex set $\mathcal{X}$, $\Pi_{\mathcal{X}}$ denotes the orthogonal projection onto $\mathcal{Z}$, i.e., $y = \Pi_{\mathcal{X}}(x)$ if $y \in \mathcal{X}$ and $\|y - x\|_2 = \min_{z\in\mathcal{X}} \|z - x\|_2$; $\mathsf{I}_{\mathcal{X}}(\cdot)$ denotes the indicator function of $\mathcal{X}$, i.e., $\mathsf{I}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $+\infty$ otherwise. Finally, we denote by $\mathcal{F}_k := \sigma(S_0, \ldots, S_k)$ the σ-field generated by the first $k+1$ random variables $S_0, \ldots, S_k$.

For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the Fréchet subdifferential of f at x, denoted by $\partial f(x)$, consists of all vectors $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + \langle g, y - x\rangle + o(\|y - x\|) \text{ as } y \to x.$$

The Fréchet and conventional subdifferentials coincide for convex functions, while for smooth functions f, $\partial f(x)$ reduces to the gradient $\{\nabla f(x)\}$. A point $x \in \mathbb{R}^n$ is said to be *stationary* for problem (1) if $0 \in \partial f(x) + \partial \mathsf{I}_{\mathcal{X}}(x)$.

The following lemma collects standard properties of weakly convex functions (Vial, 1983).

Lemma 2.1 (Weak convexity)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a ρ-weakly convex function. Then the following hold:*

1. *For any $x, y \in \mathbb{R}^n$ with $g \in \partial f(x)$, we have*
$$f(y) \geq f(x) + \langle g, y - x\rangle - \frac{\rho}{2}\|y - x\|_2^2.$$
2. *For all $x, y \in \mathbb{R}^n$, $\alpha \in [0, 1]$, and $z = \alpha x + (1-\alpha)y$:*
$$f(z) \leq \alpha f(x) + (1-\alpha)f(y) + \frac{\rho\alpha(1-\alpha)}{2}\|x - y\|_2^2.$$

Weakly convex functions admit an implicit smooth approximation through the Moreau envelope:

$$f_\lambda(x) = \inf_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2\lambda}\|x - y\|_2^2\right\}. \tag{4}$$

For small enough λ, the point achieving $f_\lambda(x)$ in (4), denoted by $\operatorname{prox}_{\lambda f}(x)$, is unique and given by:

$$\operatorname{prox}_{\lambda f}(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2\lambda}\|x - y\|_2^2\right\}. \tag{5}$$

The lemma below summarizes two well-known properties of the Moreau envelope and its associated proximal map (Hiriart-Urruty & Lemaréchal, 1993).

Lemma 2.2 (Moreau envelope)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a ρ-weakly convex function. Then, for a fixed parameter $\lambda^{-1} > \rho$, the following hold:*

1. *f_λ is $\mathcal{C}^1$-smooth with the gradient given by*
$$\nabla f_\lambda(x) = \lambda^{-1}\left(x - \operatorname{prox}_{\lambda f}(x)\right).$$
2. *f_λ is λ^{-1}-smooth, i.e., for all $x, y \in \mathbb{R}^n$:*
$$\left|f_\lambda(y) - f_\lambda(x) - \langle \nabla f_\lambda(x), y - x\rangle\right| \leq \frac{1}{2\lambda}\|x - y\|_2^2.$$

Failure of stationarity test A major source of difficulty in convergence analysis of non-smooth optimization methods is the lack of a controllable stationarity measure. For smooth functions, it is natural to use the norm of the gradient as a surrogate for near stationarity. However, this rule does not make sense in the non-smooth case, even if the function is convex and its gradient existed at all iterates. For example, the *convex* function $f(x) = |x|$ has $\|\nabla f(x)\|_2 = 1$ at each $x \neq 0$, no mater how close x is to the stationary point $x = 0$.

To circumvent this difficulty, we adopt the techniques pioneered in (Davis & Drusvyatskiy, 2019) for convergence of stochastic methods on weakly convex problems. More concretely, we rely on the connection of the Moreau envelope to (near) stationarity: For any $x \in \mathbb{R}^n$, the point $\hat{x} = \operatorname{prox}_{\lambda F}(x)$, where $F(x) = f(x) + \mathsf{I}_{\mathcal{X}}(x)$, satisfies:

$$\begin{cases}\|x - \hat{x}\|_2 = \lambda\|\nabla F_\lambda(x)\|_2, \\ \operatorname{dist}(0, \partial F(\hat{x})) \leq \|\nabla F_\lambda(x)\|_2.\end{cases} \tag{6}$$

Therefore, a small gradient $\|\nabla F_\lambda(x)\|_2$ implies that x is close to a point $\hat{x} \in \mathcal{X}$ that is near-stationary for F. Note that $\hat{x}$ is just a *virtual* point, there is no need to compute it.

3. Algorithm and convergence analysis

We assume that the only access to f is through a stochastic subgradient oracle. In particular, we study algorithms that attempt to solve problem (1) using i.i.d. samples $S_0, S_1, \ldots, S_K \overset{\text{iid}}{\sim} P$.

Assumption A (Stochastic oracle)**.** *Fix a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let S be a sample drawn from P and $f'(x, S) \in \partial f(x, S)$. We make the following assumptions:*

(A1) *For each $x \in \mathrm{dom}(f)$, we have*

$$\mathbb{E}_P[f'(x, S)] \in \partial f(x).$$

(A2) *There exists a real $L > 0$ such that for all $x \in \mathcal{X}$:*

$$\mathbb{E}_P\left[\|f'(x, S)\|_2^2\right] \leq L^2.$$

The above assumptions are standard in stochastic optimization of *non-smooth* functions (see, e.g., (Nemirovski et al., 2009; Davis & Drusvyatskiy, 2019)).

Algorithm To solve problem (1), we use an iterative procedure that starts from $x_0 \in \mathcal{X}$, $z_0 \in \partial f(x_0, S_0)$ and generates sequences of points $x_k \in \mathcal{X}$ and $z_k \in \mathbb{R}^n$ by repeating the following steps for $k = 0, 1, 2, \ldots$:

$$x_{k+1} = \underset{x \in \mathcal{X}}{\mathrm{argmin}} \left\{ \langle z_k, x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \quad (7a)$$

$$z_{k+1} = \beta g_{k+1} + (1 - \beta) \frac{x_k - x_{k+1}}{\alpha}, \quad (7b)$$

where $g_{k+1} \in \partial f(x_{k+1}, S_{k+1})$. When $\mathcal{X} = \mathbb{R}^n$, this algorithm reduces to the procedure (3). For a general convex set $\mathcal{X}$, this scheme is known as the iPiano method in the smooth and deterministic setting (Ochs et al., 2014). For simplicity, we refer to Algorithm 7 as stochastic heavy ball (SHB).

Throughout the paper, we will frequently use the following two quantities:

$$p_k = \frac{1 - \beta}{\beta} (x_k - x_{k-1}) \text{ and } d_k = \frac{1}{\alpha} (x_{k-1} - x_k).$$

Before detailing our convergence analysis, we note that most proofs of $O(1/\epsilon^2)$ sample complexity for subgradient-based methods rely on establishing an iterate relationship on the form (see, e.g., (Nemirovski et al., 2009; Davis & Drusvyatskiy, 2019; Ghadimi & Lan, 2013)):

$$\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] - \alpha \, \mathbb{E}[e_k] + \alpha^2 C^2, \quad (8)$$

where e_k denotes some stationarity measure such as $f(\cdot) - f^\star$ for convex and $\|\nabla f(\cdot)\|_2^2$ for smooth (possibly non-convex) problems, V_k are certain Lyapunov functions, α is the stepsize, and C is some constant. Once (8) is given, a simple manipulation results in the desired complexity provided that α is chosen appropriately. The case with decaying stepsize can be analyzed in the same way with minor adjustments. We follow the same route and identify a Lyapunov function that allows to establish relation (8) for the quantity $\|\nabla F_\lambda(\cdot)\|_2$ in (6).

Since our Lyapunov function is nontrivial, we shall build it up through a series of key results. We start by presenting the following lemma, which quantifies the averaged progress made by one step of the algorithm.

Lemma 3.1. *Let Assumptions (A1)–(A2) hold. Let $\beta = \nu\alpha$ for some constant $\nu > 0$ such that $\beta \in (0, 1]$. Let x_k be generated by procedure (7). It holds for any $k \in \mathbb{N}$ that*

$$\begin{aligned} &(1 - \beta) f(x_k) + \mathbb{E}\left[\frac{\nu}{2} \|p_{k+1}\|_2^2 \,|\mathcal{F}_{k-1}\right] \\ &\leq (1 - \beta) f(x_{k-1}) + \frac{\nu}{2} \|p_k\|_2^2 - \alpha \, \mathbb{E}\left[\|d_{k+1}\|_2^2 \,|\mathcal{F}_{k-1}\right] \\ &\quad + \alpha^2 \left(\frac{\rho(1 - \beta)}{2} + \nu\right) L^2. \end{aligned} \quad (9)$$

Proof. See Appendix A. □

In view of (8), the lemma shows that the quantity $\mathbb{E}[\|d_k\|_2^2]$ can be made arbitrarily small. However, this alone is not sufficient to show convergence to stationary points. Nonetheless, we shall show that a small $\mathbb{E}[\|d_k\|_2^2]$ indeed implies a small (averaged) value of the norm of the Moreau envelope defined at a specific point. Toward this goal, we first need to detail the points x and $\hat{x}$ in (6). It seems that taking the most natural candidate $x = x_k$ is unlikely to produce the desired result. Instead, we rely on the following iterates:

$$\bar{x}_k := x_k + \frac{1 - \beta}{\beta} (x_k - x_{k-1}),$$

and construct corresponding *virtual* reference points:

$$\hat{x}_k = \underset{x \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ F(x) + \frac{1}{2\lambda} \|x - \bar{x}_k\|_2^2 \right\},$$

for $\lambda < 1/\rho$. By Lemma 2.2, $\nabla F_\lambda(\bar{x}_k) = \lambda^{-1}(\bar{x}_k - \hat{x}_k)$, where $F_\lambda(\cdot)$ is the Moreau envelope of $F(\cdot) = f(\cdot) + \mathsf{I}_{\mathcal{X}}(\cdot)$.

With these definitions, we can now state the next lemma.

Lemma 3.2. *Assume the same setting of Lemma 3.1. Let $\lambda > 0$ be such that $\lambda^{-1} \geq 2\rho$. Let $\xi = (1 - \beta)/\nu$ and define the function:*

$$\begin{aligned} V_k = F_\lambda(\bar{x}_k) + \frac{\nu \xi^2}{4\lambda^2} \|p_k\|_2^2 + \frac{\alpha \xi^2}{2\lambda^2} \|d_k\|_2^2 \\ + \left(\frac{(1 - \beta)\xi^2}{2\lambda^2} + \frac{\xi}{\lambda}\right) f(x_{k-1}). \end{aligned} \quad (10)$$

Then, for any $k \in \mathbb{N}$,

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_{k-1}\right] \leq V_k - \frac{\alpha}{2}\left\|\nabla F_\lambda(\bar{x}_k)\right\|_2^2 + \frac{\gamma\alpha^2 L^2}{2\lambda}, \quad (11)$$

where $\gamma = \xi^2(\rho(1-\beta)/2+\nu)/\lambda + \rho\xi/2 + 1$.

The proof of this lemma is rather involved and can be found in Appendix B. Lemma 3.2 has established a relation akin to (8) with the Lyapunov function V_k defined in (10). We can now use standard analysis to obtain our sample complexity.

Theorem 1. *Let Assumptions (A1)-(A2) hold. Let k^* be sampled uniformly at random from $\{0,\dots,K\}$. Let $f^\star = \inf_{x\in\mathcal{X}} f(x)$ and denote $\Delta = f(x_0) - f^\star$. If we set $\alpha = \frac{\alpha_0}{\sqrt{K+1}}$ and $\nu = 1/\alpha_0$ for some real $\alpha_0 > 0$, then under the same setting of Lemma 3.2:*

$$\mathbb{E}\left[\left\|\nabla F_\lambda(\bar{x}_{k^*})\right\|_2^2\right] \leq 2\cdot\frac{\gamma_1\Delta + \frac{\gamma L^2}{2\lambda}}{\alpha_0\sqrt{K+1}}, \quad (12)$$

where $\gamma \leq \rho^2\alpha_0^2 + 3\rho\alpha_0 + 1$ and $\gamma_1 \leq 2\rho^2\alpha_0^2 + 2\rho\alpha_0 + 1$. Furthermore, if α_0 is set to $1/\rho$, we obtain

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(\bar{x}_{k^*})\right\|_2^2\right] \leq 10\cdot\frac{\rho\Delta + L^2}{\sqrt{K+1}}.$$

Proof. Taking the expectation on both sides of (11) and summing the result over $k = 0,\dots,K$ yield

$$\mathbb{E}[V_{K+1}] \leq V_0 - \frac{\alpha_0}{2\sqrt{K+1}}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla F_\lambda(\bar{x}_k)\right\|_2^2\right] + \frac{\gamma L^2\alpha_0^2}{2\lambda}.$$

Let $\gamma_1 = 1 + (1-\beta)\xi^2/(2\lambda^2) + \xi/\lambda$, the left-hand-side of the above inequality can be lower-bounded by $\gamma_1 f^\star$. Using the facts that $F_\lambda(x_0) \leq f(x_0)$ and $x_{-1} = x_0$, we get $V_0 \leq \gamma_1 f(x_0)$. Consequently,

$$\mathbb{E}\left[\left\|\nabla F_\lambda(\bar{x}_{k^*})\right\|_2^2\right] \leq 2\cdot\frac{\gamma_1\Delta + \frac{\gamma_1 L^2\alpha_0^2}{2\lambda}}{\alpha_0\sqrt{K+1}},$$

where the last expectation is taken with respect to all random sequences generated by the method and the uniformly distributed random variable k^*. Note that $\nu = 1/\alpha_0$, $\xi = (1-\beta)/\nu$, and $1-\beta \leq 1$. Thus, letting $\lambda = 1/(2\rho)$, the constants γ and γ_1 can be upper-bounded by $\rho^2\alpha_0^2 + 3\rho\alpha_0 + 1$ and $2\rho^2\alpha_0^2 + 2\rho\alpha_0 + 1$, respectively. Therefore, if we let $\alpha_0 = 1/\rho$, we arrive at

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(\bar{x}_{k^*})\right\|_2^2\right] \leq 10\cdot\frac{\rho\Delta + L^2}{\sqrt{K+1}},$$

as desired. □

Some remarks regarding Theorem 1 are in order:

i) The choice $\nu = 1/\alpha_0$ is just for simplicity; one can pick any constant such that $\beta = \nu\alpha \in (0,1]$. Note that the stepsize used to achieve the rate in (12) does not depend on any problem parameters. Once α is set, the momentum parameter selection is completely parameter-free. Since both α and β scale like $O(1/\sqrt{K})$, Algorithm 7 can be seen as a single time-scale method (Ghadimi et al., 2020; Ruszczynski & Syski, 1983). Such methods contrast those that require at least two time-scales to ensure convergence. For example, stochastic dual averaging for *convex* optimization (Xiao, 2010) requires one fast scale $O(1/K)$ for averaging the subgradients, and one slower scale $O(1/\sqrt{K})$ for the stepsize. To show almost sure convergence of SHB for *smooth and unconstrained* problems, the work (Gitman et al., 2019) requires that both the stepsize and the momentum parameter tend to zero but the former one must do so at a faster speed.[1]

ii) To some extent, Theorem 1 supports the use of a small momentum parameter such as $\beta = 0.1$ or $\beta = 0.01$, which corresponds to the default value $1-\beta = 0.9$ in PyTorch[2] or the smaller $1-\beta = 0.99$ suggested in (Goh, 2017). Indeed, the theorem allows to have β as small as $O(1/\sqrt{K})$, i.e., one can put much more weight to the momentum term than the fresh subgradient and still preserve the complexity. Recall also that SHB reduces to SGD as $\beta \to 1$, which also admits a similar complexity. It is thus quite flexible to set β, without sacrificing the worst-case complexity. We refer to (Gitman et al., 2019, Theorem 2) for a similar discussion in the context of almost sure convergence on smooth problems.

iii) In view of (6), the theorem indicates that $\bar{x}_k$ is nearby a near-stationary point $\hat{x}_k$. Since $\bar{x}_k$ may not belong to $\mathcal{X}$, it is thus more preferable to have the similar guarantee for the iterate x_k. Indeed, we have

$$\begin{aligned}\lambda^{-2}\left\|x_k - \hat{x}_k\right\|_2^2 &\leq 2\lambda^{-2}\left\|\bar{x}_k - \hat{x}_k\right\|_2^2 + 2\lambda^{-2}\left\|x_k - \bar{x}_k\right\|_2^2 \\ &= 2\left\|\nabla F_\lambda(\bar{x}_k)\right\|_2^2 + 2\lambda^{-2}\left\|x_k - \bar{x}_k\right\|_2^2 \\ &= 2\left\|\nabla F_\lambda(\bar{x}_k)\right\|_2^2 + 2\lambda^{-2}\xi^2\left\|d_k\right\|_2^2.\end{aligned}$$

Since both terms on the right converge at the rate $O(1/\sqrt{K})$, it immediately translates into the same guarantee for the term on the left, as desired.

In summary, we have established the convergence rate $O(1/\sqrt{K})$ or, equivalently, the sample complexity $O(1/\epsilon^2)$ of SHB for the minimization of weakly convex functions.

4. Extension to smooth non-convex functions

In this section, we study the convergence property of Algorithm (7) for the minimization of ρ-smooth functions:

$$\left\|\nabla f(x) - \nabla f(x)\right\|_2 \leq \rho\left\|x - y\right\|_2, \quad \forall x, y \in \operatorname{dom} f.$$

[1] Note that our β corresponds to $1-\beta$ in (Gitman et al., 2019).

[2] https://pytorch.org/

Note that ρ-smooth functions are automatically ρ-weakly convex. In this setting, it is more common to replace Assumption (A2) by the following.

Assumption (A3). There exists a real $\sigma > 0$ such that for all $x \in \mathcal{X}$:

$$\mathbb{E}\left[\|f'(x,S) - \nabla f(x)\|_2^2\right] \leq \sigma^2.$$

Deriving convergence rates of stochastic schemes with momentum for non-convex functions under Assumption (A3) can be quite challenging. Indeed, even in *unconstrained* optimization, previous studies often need to make the assumption that the true gradient is bounded, i.e., $\|\nabla f(x)\|_2 \leq G$ for all $x \in \mathbb{R}^n$ (see, e.g., (Yan et al., 2018; Gitman et al., 2019)). This assumption is strong and does not hold even for *quadratic convex* functions. It is more realistic in constrained problems, for example when $\mathcal{X}$ is compact, albeit the constant G could then be large.

Our objective in this section is twofold: First, we aim to extend the convergence results of SHB in the previous section to smooth optimization problems under Assumption (A3). Note that even in this setting, the sample complexity of SHB has not been established before. The rate is obtained without the need of forming a batch of samples at each iteration, which is commonly required for constrained non-convex stochastic optimization (Ghadimi & Lan, 2016; Ghadimi et al., 2016). Second, for unconstrained problems, we demonstrate how to achieve the same complexity without the bounded gradient assumption above.

Let $h(x) = \frac{1}{2}\|x\|_2^2 + \mathsf{I}_{\mathcal{X}}(x)$ and let $h^*(z)$ be its convex conjugate. Our convergence analysis relies on the function:

$$\varphi_k = h^*(x_k - \alpha z_k) - \frac{1}{2}\|x_k\|_2^2 + \alpha \langle x_k, z_k\rangle . \qquad (13)$$

The use of this function is inspired by (Ruszczyński, 1987). Roughly speaking, φ_k is the negative of the optimal value of the function on the RHS of (7a), and hence, $\varphi_k \geq 0$ for all k. This function also underpins the analysis of the dual averaging scheme in (Nesterov, 2009).

The following result plays a similar role as Lemma 3.1.

Lemma 4.1. *Let Assumptions (A1) and (A3) hold. Let $\alpha \in (0, 1/\rho)$ and $\beta = \nu\alpha$ for some constant $\nu > 0$ such that $\beta \in (0,1]$. Let $\alpha \in (0, 1/(4\rho)]$ and $\xi = (1-\beta)/\nu$, and define the function:*

$$W_k = 2f(x_k) + \frac{\varphi_k}{\nu\alpha^2} + \frac{\xi}{2}\|d_k\|_2^2 .$$

Then, it holds for any $k \in \mathbb{N}$ that

$$\mathbb{E}\left[W_{k+1}|\mathcal{F}_k\right] \leq W_k - \alpha\|d_{k+1}\|_2^2 + 4\nu\alpha^2\sigma^2. \qquad (14)$$

Proof. Since the proof is rather technical, we defer details to Appendix C and sketch only the main arguments here.

By smoothness of h^*, weak convexity of f and the optimality condition for the update formula (7a) we get

$$\mathbb{E}\left[f(x_{k+1}) + \frac{\varphi_{k+1}}{\nu\alpha^2}\middle|\mathcal{F}_k\right] \leq f(x_k) + \frac{\varphi_k}{\nu\alpha^2} - (\alpha - \frac{\rho\alpha^2}{2})\|d_{k+1}\|_2^2 + \frac{1}{2\nu}\mathbb{E}\left[\|z_k - z_{k+1}\|_2^2|\mathcal{F}_k\right].$$

The proof of this relation can be found in Lemma C.1. The preceding inequality admits very useful properties as we have terms that form a telescoping sum, and the constant associated with $\|d_{k+1}\|_2^2$ has the right order-dependence on the stepsize. However, we still have a remaining term $\|z_k - z_{k+1}\|_2^2$. Thus, in view of relation (8), our next strategy is to bound this term in a way that still keeps all the favourable features described above, and at the most introduces an additional term of order $O(\alpha^2\sigma^2)$. As shown in Lemma C.2, we can establish the following inequality

$$\mathbb{E}\left[\frac{1}{2\nu}\|z_{k+1} - z_k\|_2^2|\mathcal{F}_k\right] \leq f(x_k) - f(x_{k+1}) + \frac{\xi}{2}\|d_k\|_2^2 - \frac{\xi}{2}\|d_{k+1}\|_2^2 - \left(\alpha - \frac{\alpha^3\rho^2 + 3\rho\alpha^2}{2}\right)\|d_{k+1}\|_2^2 + 4\nu\alpha^2\sigma^2.$$

Now, (14) follows immediately from combining the two previous inequalities and the fact that $\alpha \in (0, 1/(4\rho)]$. □

We remark that Lemma 4.1 does not require the bounded gradient assumption and readily indicates the convergence rate $O(1/\sqrt{K})$ for $\mathbb{E}[\|d_k\|_2^2]$. However, to establish the rate for $\mathbb{E}[\|\nabla F_\lambda(\bar{x}_k)\|_2^2]$, we need to impose such an assumption in the theorem below. Nonetheless, the assumption is much more realistic in this setting than the unconstrained case.

Theorem 2. *Let Assumptions (A1) and (A3) hold. Assume further that $\|\nabla f(x)\|_2 \leq G$ for all $x \in \mathcal{X}$. Let k^*, $\bar{x}_{k^*}$, λ, Δ, γ, and γ_1 be defined as in Theorem 1. If we set $\alpha = \frac{\alpha_0}{\sqrt{K+1}}$ and $\nu = 1/\alpha_0$ for some real $\alpha_0 > 0$, then*

$$\mathbb{E}\left[\|\nabla F_\lambda(\bar{x}_{k^*})\|_2^2\right] \leq 2 \cdot \frac{\gamma_1\Delta + \gamma(\sigma^2 + G^2)/(2\lambda)}{\alpha_0\sqrt{K+1}}.$$

Furthermore, if α_0 is set to $1/\rho$, we obtain

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(\bar{x}_{k^*})\right\|_2^2\right] \leq 10 \cdot \frac{\rho\Delta + \sigma^2 + G^2}{\sqrt{K+1}}.$$

Proof. The proof is a verbatim copy of that of Theorem 1 with L^2 replaced by $\sigma^2 + G^2$; see Appendix D. □

Some remarks are in order:

i) To the best of our knowledge, this is the first convergence rate result of a stochastic (or even deterministic) method

with Polyak momentum for smooth, non-convex, and constrained problems.

ii) The algorithm enjoys the same single time-scale and parameter-free properties as in the non-smooth case.

iii) The rate in the theorem readily translates into an analogous estimate for the norm of the so-called *gradient mapping* $\mathcal{G}_{1/\rho}$, which is commonly adapted in the literature, e.g., (Ghadimi et al., 2016). This is because for ρ-smooth functions (Drusvyatskiy & Paquette, 2019):

$$\left\|\mathcal{G}_{1/\rho}(x)\right\|_2 \leq \frac{3}{2}(1+\frac{1}{\sqrt{2}})\left\|\nabla F_{1/(2\rho)}(x)\right\|_2, \quad \forall x \in \mathbb{R}^n.$$

Since the bounded gradient assumption is rather restrictive in the unconstrained case, our final result shows how the desired complexity can be attained without this assumption.

Theorem 3. *Let Assumptions (A1) and (A3) hold. Let* $\lambda^{-1} \in (3\rho/2, 2\rho]$. *Let* k^* *be sampled uniformly at random from* $\{-1, \ldots, K-1\}$. *If we set* $\alpha = \frac{\alpha_0}{\sqrt{K+1}}$ *and* $\nu = 1/\alpha_0$, *where* $\alpha_0 \in (0, 1/(4\rho)]$, *then under the same setting of Lemma 4.1:*

$$\mathbb{E}\left[\left\|\nabla F_\lambda(\bar{x}_{k^*})\right\|_2^2\right] \leq c \cdot \frac{(1+2\alpha_0^2/\lambda^2)\Delta + \frac{(1+8\alpha_0/\lambda)\sigma^2\alpha_0^2}{2\lambda}}{\alpha_0\sqrt{K+1}},$$

where $c = 2\lambda^{-1}/(2\lambda^{-1}-3\rho)$. *Furthermore, let* $\lambda = 1/(2\rho)$, *we obtain*

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(\bar{x}_{k^*})\right\|_2^2\right] \leq 4 \cdot \frac{\left(1+8\rho^2\alpha_0^2\right)\Delta + (\rho+16\alpha_0\rho^2)\sigma^2\alpha_0^3}{\alpha_0\sqrt{K+1}}.$$

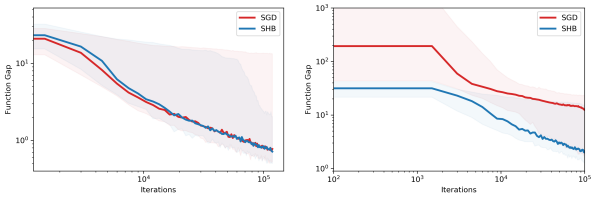
Proof. See Appendix E. □

It should be mentioned that a similar result has been attained very recently in (Ghadimi et al., 2020) using a different analysis, albeit no sample complexity is given for the non-smooth case. It is still an open question if one can preserve the complexity in Theorem 2 without the bounded gradient hypothesis.

5. Numerical evaluations

In this section, we perform experiments to validate our theoretical developments and to demonstrate that despite sharing the same worst-case complexity, SHB can be better in terms of speed and robustness to problem and algorithm parameters than SGD.

We consider the robust phase retrieval problem (Duchi & Ruan, 2018a;b): Given a set of m measurements $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$, the phase retrieval problem seeks for a vector $x^\star$ such that $\langle a_i, x^\star\rangle^2 \approx b_i$ for most $i = 1, \ldots, m$. Whenever



(a) $\kappa = 10$, $\alpha_0 = 0.1$ (b) $\kappa = 10$, $\alpha_0 = 0.25$

(c) $\kappa = 1$, $\alpha_0 = 0.1$ (d) $\kappa = 1$, $\alpha_0 = 0.15$

Figure 1. The function gap $f(x_k) - f(x^\star)$ versus iteration count for phase retrieval with $p_{\text{fail}} = 0.2$, $\beta = 10/\sqrt{K}$.

the problem is corrupted with gross outliers, a natural exact penalty form of this (approximate) system of equations yields the minimization problem:

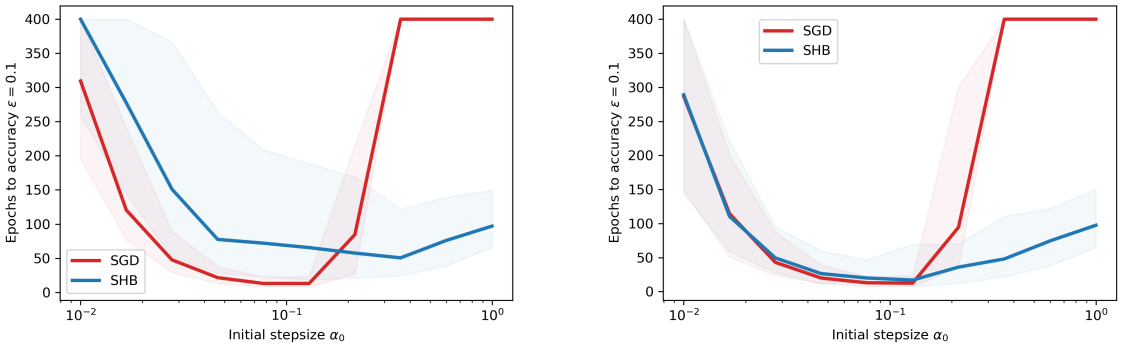
$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \frac{1}{m}\sum_{i=1}^m \left|\langle a_i, x\rangle^2 - b_i\right|.$$

This objective function is non-smooth and non-convex. In view of (2), it is the composition of the Lipschitz-continuous convex function $h(y) = \|y\|_1$ and the smooth map c with $c_i(x) = \langle a_i, x\rangle^2 - b_i$. Hence, it is weakly convex.

In each experiment, we set $m = 300$, $n = 100$ and select $x^\star$ uniformly from the unit sphere. We generate A as $A = QD$, where $Q \in \mathbb{R}^{m\times n}$ is a matrix with standard normal distributed entries, and D is a diagonal matrix with linearly spaced elements between $1/\kappa$ and 1, with $\kappa \geq 1$ playing the role of a condition number. The elements b_i of the vector b are generated as $b_i = \langle a_i, x^\star\rangle^2 + \delta\zeta_i$, $i = 1, \ldots, m$, where $\zeta_i \sim \mathcal{N}(0, 25)$ models the corruptions, and $\delta \in \{0, 1\}$ is a binary random variable taking the value 1 with probability p_{fail}, so that $p_{\text{fail}} \cdot m$ measurements are corrupted. The algorithms are all randomly initialized at $x_0 \sim \mathcal{N}(0, 1)$. The stochastic subgradient is simply given as an element of the subdifferential of $g(x) = \left|\langle a, x\rangle^2 - b\right|$:

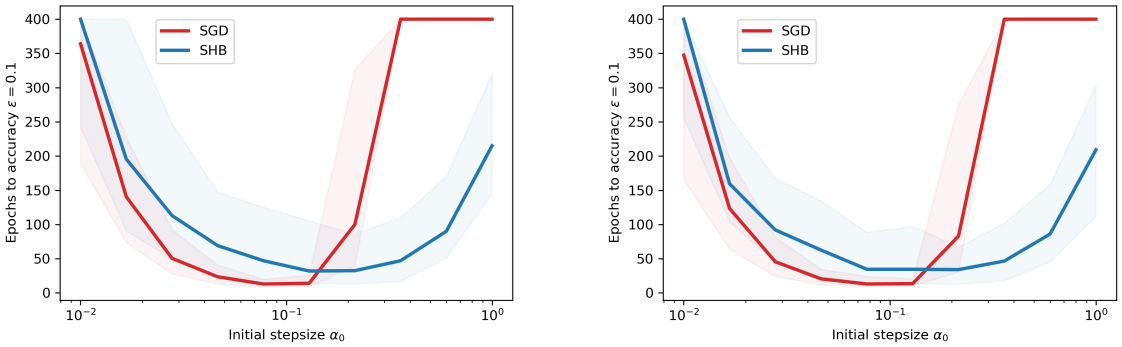
$$\partial g(x) = 2\langle a, x\rangle a \cdot \begin{cases} \text{sign}\left(\langle a, x\rangle - b\right)^2 & \text{if } \langle a, x\rangle^2 \neq b, \\ [-1, 1] & \text{otherwise.} \end{cases}$$

In each of our experiments, we set the stepsize as $\alpha_k = \alpha_0/\sqrt{k+1}$, where α_0 is an initial stepsize. We note that this stepsize can often make a little faster progress (for both SGD and SHB) at the beginning of the optimization process than the constant one $\alpha_k = \alpha_0/\sqrt{K+1}$. However, after a



(a) $p_{\text{fail}} = 0.3$, $\beta = 1/\sqrt{K}$ (b) $p_{\text{fail}} = 0.3$, $\beta = 1/\alpha_0/\sqrt{K}$

Figure 2. The number of epochs to achieve ϵ-accuracy versus initial stepsize α_0 for phase retrieval with $\kappa = 10$.



(a) $p_{\text{fail}} = 0.3$, $\beta = 0.1$ (b) $p_{\text{fail}} = 0.3$, $\beta = 0.01$

Figure 3. The number of epochs to achieve ϵ-accuracy versus initial stepsize α_0 for phase retrieval with $\kappa = 10$ and popular choices of β.

few iterations, both of them yield very similar results and will not change the qualitative aspects of our plots in any way. We also refer m stochastic iterations as one epoch (pass over the data). Within each individual run, we allow the considered stochastic methods to perform $K = 400m$ iterations. We conduct 50 experiments for each stepsize and report the median of the epochs-to-ϵ-accuracy; the shaded areas in each plot cover the 10th to 90th percentile of convergence times.

Figure 1 shows the function gap versus iteration count for different values of κ and α_0, with $p_{\text{fail}} = 0.2$, $\beta = 10/\sqrt{K}$. It is evident that SHB converges with a theoretically justified parameter β and is much less sensitive to problem and algorithm parameters than the vanilla SGD. Note that the sensitivity issue of SGD is rather well documented; a slight change in its parameters can have a severe effect on the overall performance of the algorithm (Nemirovski et al., 2009; Asi & Duchi, 2019). For example, Fig. 1d shows that SGD exhibits a transient exponential growth before eventual convergence. This behaviour can occur even when minimizing the smooth quadratic function $\frac{1}{2}x^2$ (Asi & Duchi, 2019, Example 2). In contrast, SHB converges in all settings of the figure, suggesting that using a momentum term can help to improve the robustness of the standard SGD. This is expected as the update formula (3b) acts like a lowpass filter, averaging past stochastic subgradients, which may have stabilizing effect on the sequence $\{x_k\}$.

To further clarify this observation, in the next set of experiments, we test the sensitivity of SHB and SGD to the initial stepsize α_0. Figure 2 shows the number of epochs required to reach ϵ-accuracy for phase retrieval with $\kappa = 10$ and $p_{\text{fail}} = 0.3$. We can see that the standard SGD has good performance for a narrow range of stepsizes, while wider convergence range can be achieved with SHB.

Finally, it would be incomplete without reporting experiments for some of the most popular momentum parameters used by practitioners. Figure 3 shows a similar story to Fig. 2 for the parameter $1 - \beta = 0.9$ and $1 - \beta = 0.99$ as discussed in remark ii) after Theorem 1. This together with Fig. 2 demonstrates that SHB is able to find good approximate solutions for diverse values of the momentum constant over a wider (often significantly so) range of algorithm parameters than SGD.

6. Conclusion

Using a carefully constructed Lyapunov function, we established the first sample complexity results for the SHB method on a broad class of non-smooth, non-convex, and constrained optimization problems. The complexity is attained in a parameter-free fashion in a single time-scale. A notable feature of our results is that they justify the use of a large amount of momentum in search directions. We also improved some complexity results for SHB on smooth problems. Numerical results show that SHB exhibits good performance and low sensitivity to problem and algorithm parameters compared to the standard SGD.

Acknowledgements

This work was supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Research Council and the Swedish Foundation for Strategic Research.

References

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Beck, A. *First-order methods in optimization*, volume 25. SIAM, 2017.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4): 231–357, 2015.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Charisopoulos, V., Chen, Y., Davis, D., Díaz, M., Ding, L., and Drusvyatskiy, D. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Drusvyatskiy, D. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

Duchi, J. C. and Ruan, F. Asymptotic optimality in stochastic optimization. *arXiv preprint arXiv:1612.05612*, 2016.

Duchi, J. C. and Ruan, F. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2018a.

Duchi, J. C. and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018b.

Ermol'ev, Y. M. and Norkin, V. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2): 196–215, 1998.

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

Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pp. 310–315. IEEE, 2015.

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

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Ghadimi, S., Ruszczyński, A., and Wang, M. A single time-scale stochastic approximation method for nested stochastic optimization. *SIAM J. on Optimization*, 2020. Accepted for publication (arXiv preprint arXiv:1812.01094).

Gitman, I., Lang, H., Zhang, P., and Xiao, L. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pp. 9630–9640, 2019.

Goh, G. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

Gupal, A. M. and Bazhenov, L. G. Stochastic analog of the conjugant-gradient method. *Cybernetics and Systems Analysis*, 8(1):138–140, 1972.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms*, volume 305. Springer science & business media, 1993.

Hu, C., Pan, W., and Kwok, J. T. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, pp. 781–789, 2009.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

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

Nesterov, Y. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nurminskii, E. A. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9 (1):145–150, 1973.

Ochs, P., Chen, Y., Brox, T., and Pock, T. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

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

Polyak, B. T. *Introduction to optimization*. Optimization Software, 1987.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rockafellar, R. T. Favorable classes of Lipschitz continuous functions in subgradient optimization. In Nurminski, E. A. (ed.), *Progress in Nondifferentiable Optimization*, CP-82-S8, pp. 125–143, 1982.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of risk*, (2):21–42, 2000.

Ruszczyński, A. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

Ruszczynski, A. and Syski, W. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12): 1097–1105, 1983.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Singer, A. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147, 2013.

Tseng, P. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

Vial, J.-P. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Yan, Y., Yang, T., Li, Z., Lin, Q., and Yang, Y. A unified analysis of stochastic momentum methods for deep learning. In *International Joint Conference on Artificial Intelligence*, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zavriev, S. and Kostyuk, F. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.