Journal of Applied Mathematics and Physics, 2024, 12, 1321-1336
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Almost Sure Convergence of Proximal Stochastic Accelerated Gradient Methods

Xin Xiang, Haoming Xia*

Key Laboratory of Optimization Theory and Applications, School of Mathematics and Information, China West Normal University, Nanchong China
Email: xxiang_1028@163.com, *haomingxia1999@163.com

How to cite this paper: Xiang, X. and Xia, H.M. (2024) Almost Sure Convergence of Proximal Stochastic Accelerated Gradient Methods. *Journal of Applied Mathematics and Physics*, **12**, 1321-1336.
https://doi.org/10.4236/jamp.2024.124081

Received: March 15, 2024
Accepted: April 27, 2024
Published: April 30, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Proximal gradient descent and its accelerated version are resultful methods for solving the sum of smooth and non-smooth problems. When the smooth function can be represented as a sum of multiple functions, the stochastic proximal gradient method performs well. However, research on its accelerated version remains unclear. This paper proposes a proximal stochastic accelerated gradient (PSAG) method to address problems involving a combination of smooth and non-smooth components, where the smooth part corresponds to the average of multiple block sums. Simultaneously, most of convergence analyses hold in expectation. To this end, under some mind conditions, we present an almost sure convergence of unbiased gradient estimation in the non-smooth setting. Moreover, we establish that the minimum of the squared gradient mapping norm arbitrarily converges to zero with probability one.

Keywords

Proximal Stochastic Accelerated Method, Almost Sure Convergence, Composite Optimization, Non-Smooth Optimization, Stochastic Optimization, Accelerated Gradient Method

1. Introduction

We consider the following composite optimization problem:

$$\min_{x\in\mathbb{R}^d} F(x) \overset{\text{def}}{=} f(x)+g(x), \tag{1}$$

where $g:\mathbb{R}^d \to \mathbb{R}$ is the average of the smooth functions $g_1,\cdots,g_n$, *i.e.* $g(x)=\frac{1}{n}\sum_{i=1}^{n} g_i(x)$ and $f:\mathbb{R}^d \to \mathbb{R}$ is a closed proper function that can be non-differentiable. One of the most well-studied instances of this type of problem is ℓ_1-regularized least squares [1]:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(a_i^{\mathrm{T}}x-b_i\right)^2+\lambda\|x\|_1$$

where $\|\cdot\|$ denotes the standard ℓ_1-norm.

We frequently encounter optimization problems of this nature in various fields such as machine learning, statistics, signal processing, and imaging [2] [3] [4] [5]. Specifically, we address the task of minimizing the aggregate of two functions: one represents the average of numerous smooth component functions, while the other characterizes a general function amenable to a straightforward proximal mapping. It is imperative for us to ensure that the problem is well-defined, denoted by $\arg\min F\neq\varnothing$, and that each g_i remains bounded from below. Compared with the classical gradient descent (GD) method and stochastic gradient descent (SGD) method [6], the proximal gradient descent (PGD) method has a relatively limited application scope, primarily employed for addressing objective functions that include non-differentiable components. To tackle the non-smooth optimization problem (1), mentioned earlier, we introduce the PGD. It can be delineated by the following update rule for $k=1,2,\cdots$:

$$x_{k+1}=\operatorname{prox}_{t_k f}\left(x_k-t_k\nabla g\left(x_k\right)\right), \tag{2}$$

where t_k is the step size at the *k*th iteration and the proximity operator of *f* $\operatorname{prox}_f(\cdot)$ is defined by:

$$\operatorname{prox}_{tf}(y)=\arg\min_{x\in\mathbb{R}^d}\left\{\frac{1}{2}\|x-y\|^2+tf(x)\right\}.$$

We note that $\operatorname{prox}_{tf}(y)$ maps to a singleton since $f(y)$ is proper and closed, see for example (Beck, Theorem 6.3, 2017 [7]). PGD has adopted attributes to proximal operators no longer rely on $g(x)$, only $f(x)$. That is, $F(x)=g(x)+f(x)$ could be a combination of a very complex differentiable function and a less complex non-differentiable function originally, but with this method, we don't need to consider $g(x)$ (because it is differentiable and the gradient is easy to calculate), we just have to focus on the non-differentiable function $f(x)$, which greatly simplifies our problem. PGD can achieve an error level on the objective function of $O(1/k)$ after *k* iterations [8].

1.1. Related Work

1.1.1. Accelerated Proximal Gradient (APG) Method

Another effective method for solving problem (1) is the accelerated proximal gradient (APG) method, initially proposed by Nesterov [8] for minimizing smooth convex functions with constraints. It was later extended by Beck and Teboulle [9] to composite convex objective functions. The Fast Iterative Shrinkage-Threholding Algorithm (FISTA), an accelerated version of the proximal gradient method, has found applications in various fields, including image and signal processing. APG represents an accelerated variant of the deterministic gradient descent method, incorporating an extrapolation step in the algorithm. One simple version involves selecting an initial point $x_0 = x_{-1} \in \mathbb{R}^n$, and repeating for $k = 1, 2, 3, \cdots$.

$$y_{k+1} := x_k + \beta_k \left(x_k - x_{k-1} \right)$$

$$x_{k+1} := \operatorname{prox}_{t_k f} \left(y_{k+1} - t_k \nabla g \left(y_{k+1} \right) \right)$$

where $\beta_k \in [0,1)$ is an extrapolation parameter and t_k is the usual step size.

These parameters must be chosen in specific ways to achieve the convergence accelerated. One simple choice takes $\beta_k = \dfrac{k}{k+3}$ [10]. It remains to choose the step sizes t_k. When ∇g is Lipschitz continuous with constant *L*, this method can be shown to converge in objective value with $O\left(1/k^2\right)$ when a fixed step size $t_k = t \in \left(0, 1/L\right]$ is used [8] [9]. Following Nesterov, this method is called an accelerated or optimal first-order method and there are several versions of such methods, such as Nesterov [8]; the software package TFOCS [11] is based on and contains several implementations of such methods. Li *et al.* [12] were the first to provide APG-type algorithms for general non-convex and non-smooth problems ensuring that every accumulation point is a critical point.

1.1.2. Proximal Stochastic Gradient Descent (PSGD) Method

With the emergence of big data, the efficiency of deterministic optimization algorithm has gradually become a bottleneck. In the PGD (2), we need to calculate the full gradient of $g(\cdot)$. When the size of the datasets *n* is very large, where the calculation costs will be high, first-order stochastic gradient methods have proven to be very effective thanks to their low iteration complexity. Therefore, one way to reduce calculation is to use stochastic algorithms that take advantage of the finite sum structure of the problem (1) to use cheaper iterations while preserving fast convergence. When $f(x) = 0$, problem (1) reduces to general minimization optimization problem: $\min_{x \in \mathbb{R}^d} F(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} g_i(x)$, where $F(x)$ arise as averages of a very large number of smooth functions. This problem often arises by approximation of the stochastic optimization loss function: $F(x) = \mathbb{E}_{\xi \in \mathcal{D}} \left[g_\xi (x) \right]$, where ξ is a random variable, $g_\xi : \mathbb{R}^d \to \mathbb{R}$ is smooth for all ξ. First-order stochastic methods for the case of a non-smooth regularizer $f(x)$ are an active research area. Non-asymptotic convergence results were first achieved in [13]. For finite-sum problems, Reddi *et al.* [14] were the first to develop a proximal stochastic variance reduced gradient algorithm with improved convergence complexity. Metel *et al.* [15] first presented the non-asymptotic convergence results for the non-smooth non-covex constrained sparse optimization problem and they presented two simple stochastic proximal gradient algorithms, for stochastic and finite-sum optimization problems. Kawashima *et al.* [16] considered $f(x)$ as a non-smooth quasi-convex function and achieved the same convergence complexity as in Ghadimi *et al.* (2016) [13]. A stochastic variant of PGD is proximal stochastic gradient descent (PSGD). At each iteration $k = 1, 2, \cdots$, it picks i_k with probability $\frac{1}{n}$ from $[n]=\{1,2,\cdots,n\}$ via independent identically distribution. The PSGD takes the following update:

$$x_{k+1}=\operatorname{prox}_{t_k f}\left(x_k-t_k\nabla g_{i_k}\left(x_k\right)\right). \tag{3}$$

where $t_k>0$ is a sequence of step size (also known as learning rate). Sampling the index i_k over all indices $[n]=\{1,2,\cdots,n\}$ with i.i.d., the gradient $\nabla g_{i_k}\left(x_k\right)$ satisfies the unbiased estimation:

$$\mathbb{E}\left[\nabla g_{i_k}\left(x_k\right)\right]=\sum_{i=1}^{n}\frac{1}{n}\nabla g_{i_k}\left(x_k\right)=\nabla g\left(x_k\right). \tag{4}$$

The advantage of PSGD over PGD lies in the fact that, at each iteration, PSGD only necessitates the computation of a single gradient $\nabla g_{i_k}\left(x_k\right)$. In contrast, each iteration of PGD evaluates n g gradients. As a result, the computational cost of PSGD per iteration is $\frac{1}{n}$ of that of PGD. Consequently, the computation of $\nabla g_{i_k}\left(x_k\right)$ is approximately n times less expensive than that of $\nabla g\left(x\right)$. Numerous algorithms have been devised to tackle the composite optimization problem (1). Gorbunov *et al.* [17] provided a unified analysis covering a broad range of variants of PSGD. In a study by Cevher *et al.* [18], it was demonstrated that PSGD, assuming strong convexity, displays linear convergence towards a region dominated by noise.

1.1.3. Convergence Criteria

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, stochastic heavy ball (SHB) [19], stochastic Nesterov's accelerated gradient (SNAG) and so forth). However, almost sure convergence (a.s. for short, also known as "convergence with probability 1") [20] properties are important, because they represent what happens to individual trajectories of the stochastic iterations, which are instantiations of the stochastic algorithms actually used in practice. Therefore, almost sure convergence of methods based on stochastic gradient is of practical relevance. For SGD, in the convex and smooth setting, Sebbouh *et al.* [21] provided the almost sure asymptotic convergence rates for a weighted average of the iterates. The almost sure convergence of the last iteration of SGD on non-convex functions is generalized by Orabona [22]. Liu *et al.* [23] provide a unified almost sure convergence rates analysis for SGD, SHB and SNAG. Recently, Liang *et al.* [24] presented an almost sure convergence analysis of stochastic composite objective mirror descent (SCOMID).

Remark 1.1. *It is worth noting that our proof does not rely on the convexity of the function f, so convexity of f is not assumed in the model. However, to ensure the continuity of the generalized gradient operator, we assume that f is convex.*

1.2. Proximal Stochastic Accelerated Gradient (PSAG) Method

Based on the above, we now consider an accelerated version of proximal stochastic

Table 1. Comparison of problem setting, momentum, algorithm and convergence results with some relevant literature.

Citation	$g(x)$	$f(x)$	Constraint	Momentum	Algorithm	a.s.
Khaled *et al.* [25]	L-smooth	--	$\mathbb{R}^d$	×	SGD	×
Gower *et al.* [26]	--		--	--	--	--
Mertikopoulos *et al.* [27]	L-smooth,	--	X	×	SGD	√
	G-Lipschitz		--	--	--	
Sebbouh *et al.* [21]	L-smooth	--	$\mathbb{R}^d$	×	SGD	√
Liu *et al.* [23]	--		--	--	--	--
Sebbouh *et al.* [21]	L-smooth	--	$\mathbb{R}^d$	√	SHB	√
Liu *et al.* [23]	--	--	--	--	--	--
Liu *et al.* [23]	L-smooth	--	$\mathbb{R}^d$	√	SNAG	√
Ward *et al.* [28]	L-smooth	--	$\mathbb{R}^d$	×	AdaGrad	×
Alacaoglu *et al.* [29]	--	ρ-weakly convex	X	×	AdaGrad	×
Davis *et al.* [30]	--	ρ-weakly convex	$\mathbb{R}^d$	×	PSGD	×
Cevher *et al.* [18]	L-smooth	Convex	$\mathbb{R}^d$	×	PSGD	×
Gorbunov *et al.* [17]	--	--	--	--	--	--
Ghadimi *et al.* [13]	L-smooth	Convex	X	×	SCOMID	×
Ours	L-smooth	Convex	$\mathbb{R}^d$	√	PSAG	√

SCOMID = Stochastic Composite Objective Mirror Descent; PSGD = Proximal Stochastic Gradient Descent; SGD = Stochastic Gradient Descent; AdaGrad = Adaptive Stochastic Gradient Descent; SHB = Stochastic Heavy Ball; SNAG = Stochastic Nesterov's Accelerated graDient; PSAG = Proximal Stochastic Accelerated Gradient (in this paper).

gradient method and use the unbiased stochastic gradient (see the last column of **Table 1**), the iteration of the PSAG method is given by:

$$\begin{aligned} y_{k+1} &:= x_k + \beta_k\left(x_k - x_{k-1}\right) \\ x_{k+1} &:= \mathrm{prox}_{t_k f}\left(y_{k+1} - t_k\mathcal{G}_k\right) \end{aligned} \tag{5}$$

where $\beta_k \in [0,1)$ is an extrapolation parameter and t_k is the usual step size, $\mathcal{G}_k = \nabla g_{i_k}\left(y_{k+1}\right)$ is an unbiased estimator of the gradient, where i_k is randomly picked from $[n] = \{1,2,\cdots,n\}$ with probability $\frac{1}{n}$ via the independent identically distribution. Then, we have:

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^d}\left\{ t\left\langle \mathcal{G}_k, x\right\rangle + tf\left(x\right) + \frac{1}{2}\left\|x - x_k\right\|^2 \right\}. \tag{6}$$

where in (5), $t = t_k$ is the stepsize, $\mathcal{G}_k$ is the stochastic gradient and $x_k = y_{k+1}$ is an extrapolation step. For non-smooth composite problem (1), we establish *almost sure convergence rates* of PSAG that the minimum of the squared gradient mapping norm is arbitrarily close to zero with probability one. It is noted that PSAG method reduces to PSGD method when $\beta_k = 0$ in (5).

The rest of this paper is organized as follows. In Section 2, we recall some definitions and known results for further analysis. Then, we present our convergence analysis of PSAG and its convergence rate in Section 3. Finally, we summarize our findings and draw conclusions in Section 4.

2. Preliminaries

2.1. Notations

The optimal solution of problem (1) is denoted by x^*, and the optimal set of problem (1) is non-empty and denoted by X^*. The optimal value of the problem (1) is denoted by $F^* = F\left(x^*\right)$. In the rest of this work, for notational brevity, we will omit the subscript of norm $\|\cdot\|_2$ for $\|\cdot\|$.

2.2. Definitions

Definition 2.1. (*Convexity*). *A function* $g:\mathbb{R}^d \to \mathbb{R}$ *is said to be convex*; *i.e. for all* $x, y \in \mathbb{R}^d$,

$$g(x) \geq g(y) + \langle \nabla g(y), x-y \rangle. \tag{7}$$

Definition 2.2. (*L-smoothness*). *A function* $g:\mathbb{R}^d \to \mathbb{R}$ *is said to be L-smooth if the gradient* $\nabla g(x)$ *is Lipschitz continuous with constant L, i.e. there exists a constant* $L > 0$ *such that*:

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x-y\|, \text{ for all } x, y \in \mathbb{R}^d. \tag{8}$$

It is well-known that Definition 2.2 implies the following inequality (see, e.g. Nesterov [31] (Lemma 1.2.3)):

$$g(x) \leq g(y) + \langle \nabla g(y), x-y \rangle + \frac{L}{2}\|x-y\|^2, \ \forall\, x, y \in \mathbb{R}^d. \tag{9}$$

Remark 2.1. *Definition* 2.2 *essentially implies that gradient descent with sufficiently small step size is well behaved, and we also have* $g(x) = \sum_{i=1}^n g_i(x)$ *is L-smooth where* $L = \sum_i^n L_i$ ($g_i(x)$ *is* L_i*-smooth*).

Definition 2.3. (*Unbiasedness*). *Given a random iterate* $\{x_k\}_{k\geq 0}$. *We call the stochastic gradient* $\nabla g_{i_k}(x_k)$ (i_k *is sampled randomly from* $\{1, 2, \cdots, n\}$) *is unbiased if*:

$$\mathbb{E}\left[\nabla g_{i_k}(x_k)\right] = \nabla g(x_k) \tag{10}$$

holds.

Definition 2.4. (*Prox-grad operator*). *Consider the composite optimization problem* (1), *we have*:

$$x_k^+ = T_{t_k}^{f,g}(x_k)$$

where $T_t^{f,g} : int\left(dom(g)\right) \to \mathbb{E}(t>0)$ is the prox-grad operator defined by:

$$T_t^{f,g}(x) \equiv \operatorname{prox}_{tf}\left(x - t\nabla g(x)\right) \tag{11}$$

Definition 2.5. (*Gradient mapping*). *Suppose that* f, g *satisfy the problem*

(1) *settings. Then, the gradient mapping is the operator* $G_t^{f,g}: int(dom(g)) \to \mathbb{E}$ *defined by*:

$$G_t^{f,g}(x) \equiv \frac{1}{t}\left(x - T_t^{f,g}(x)\right) \tag{12}$$

for any $x \in int(dom(g))$.

3. Convergence Analysis

3.1. Lemmas on Supermartingale Convergence Rates

The proof about almost sure convergence relies on the classical Robbins-Siegmund supermartingale convergence result (Theorem 1 in [32]).

Lemma 3.1. (*Theorem* 1, [32]) *Assume that* $\{X_k\},\{Y_k\}$ *and* $\{Z_k\}$ *are three non-negative sequences of random variable,* $\{\gamma_k\}$ *is a non-negative real sequence and* $\mathcal{F}_k$ *is a σ-algebra. If* X_k, Y_k, Z_k *are all* $\mathcal{F}_k$ *-measurable and the following conditions holds*:

1) $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] \le (1+\gamma_k)Y_k - X_k + Z_k$.

2) $\sum_{k=1}^{\infty}\gamma_k < \infty, \sum_{k=1}^{\infty} Z_k < \infty$ almost surely.

Then, Y_k converges almost surely and $\sum_{k=1}^{\infty} X_k < \infty$ almost surely (a.s.)

Lemma 3.2. (*Lemma* 3, [23]) *Assume that* $\{X_k\}$ *is a non-negative sequence of random variable and* $t_k \ge 0$ *is non-increasing. If the following conditions hold*:

$$\sum_{k=1}^{\infty}\frac{t_{k+1}}{\sum_{s=1}^{k} t_s} = \infty, \quad \sum_{k=1}^{\infty} t_{k+1}X_k < \infty \quad a.s. \tag{13}$$

then we have:

$$\min_{1\le i\le k} X_i = o\left(\frac{1}{\sum_{j=1}^{k} t_s}\right) \quad a.s. \tag{14}$$

where o denotes the higher-order infinitesimal. *i.e.* for two sequences $\{a_k\} \to 0$ and $\{b_k\} \to 0$, $a_k = o(b_k)$ if and only if $\lim_{k\to\infty} a_k / b_k = 0$.

3.2. Almost Sure Convergence Rate Analysis for Stochastic Proximal Accelerated Gradient Method

Reviewing the iteration of PSAG (5), using the Definition 2.4, 2.5, we rewrite the PSAG as follows:

$$\begin{aligned} y_{k+1} &:= x_k + \beta(x_k - x_{k-1}), \\ x_{k+1} &:= y_{k+1} - t_k G_{t_k}(y_{k+1}). \end{aligned} \tag{15}$$

where $G_{t_k}(y_{k+1}) = \frac{1}{t_k}\left(y_{k+1} - T_{t_k}^{f,g_{i_k}}(y_{k+1})\right)$,

$T_{t_k}^{f,g_{i_k}}(y_{k+1}) = \text{prox}_{t_k f}\left(y_{k+1} - t_k \nabla g_{i_k}(y_{k+1})\right)$, t_k is the step size and $\beta = \beta_k \in [0,1)$ and i_k is randomly picked from $[n] = \{1,2,\cdots,n\}$ with probability $\frac{1}{n}$ via the independent identically distribution.

Lemma 3.3. *Suppose that* $u \in \arg\min_{x\in\mathbb{R}^d} \left\{ t\langle z,x\rangle + tf(x) + \frac{1}{2}\|x-y\|^2 \right\}$ *for some* $z,y\in\mathbb{R}^d$. *If* $g(x)$ *is L-smooth, the for any* $x\in\mathbb{R}^d$, *we have:*

$$F(u) \le F(x) + \langle \nabla g(y) - z, u-x\rangle + \frac{1}{2t}\|x-y\|^2 - \frac{1}{2t}\|u-y\|^2 + \frac{L}{2}\|u-y\|^2. \quad (16)$$

Proof By the optimality of *u*, we can show that for any $x\in\mathbb{R}^d$,

$$t\langle z,u\rangle + tf(u) + \frac{1}{2}\|u-y\|^2 \le t\langle z,x\rangle + tf(x) + \frac{1}{2}\|x-y\|^2 \quad (17)$$

Upon rearranging the above equation, we have:

$$f(u) \le f(x) + \langle z, x-u\rangle + \frac{1}{2t}\|x-y\|^2 - \frac{1}{2t}\|u-y\|^2. \quad (18)$$

Since $g(x)$ is *L*-smooth, we can apply Definition 2.2 to have:

$$g(u) \le g(y) + \langle \nabla g(y), u-y\rangle + \frac{L}{2}\|u-y\|^2. \quad (19)$$

$$g(x) \ge g(y) + \langle \nabla g(y), x-y\rangle + \frac{1}{2L}\|\nabla g(x) - \nabla g(y)\|^2. \quad (20)$$

Next, upon combining (19) and (20), we arrive at:

$$\begin{aligned} g(u) &\le g(y) + \langle \nabla g(y), u-y\rangle + \frac{L}{2}\|u-y\|^2 \\ &\le g(x) + \langle \nabla g(y), y-x\rangle - \frac{1}{2L}\|\nabla g(x) - \nabla g(y)\|^2 \\ &\quad + \langle \nabla g(y), u-y\rangle + \frac{L}{2}\|u-y\|^2 \\ &\le g(x) + \langle \nabla g(y), u-x\rangle + \frac{L}{2}\|u-y\|^2. \end{aligned} \quad (21)$$

By the fact that $F(u) = g(u) + f(u)$, we finally obtain:

$$\begin{aligned} F(u) &= g(u) + f(u) \\ &\le g(u) + f(x) + \langle z, x-u\rangle + \frac{1}{2t}\|x-y\|^2 - \frac{1}{2t}\|u-y\|^2 \\ &\le g(x) + \langle \nabla g(y), u-x\rangle + \frac{L}{2}\|u-y\|^2 \\ &\quad + f(x) + \langle z, x-u\rangle + \frac{1}{2t}\|x-y\|^2 - \frac{1}{2t}\|u-y\|^2 \\ &= F(x) + \langle \nabla g(y) - z, u-x\rangle + \frac{1}{2t}\|x-y\|^2 - \frac{1}{2t}\|u-y\|^2 + \frac{L}{2}\|u-y\|^2. \end{aligned} \quad (22)$$

Theorem 3.1. *Suppose that* $F(x)$ *is lower bounded and there exists* $\sigma > 0$, $A_k \ge 0$ *such that:*

$$\mathbb{E}\left[\left\|\nabla g(y_{k+1}) - \mathcal{G}_k\right\|^2 \mid \mathcal{F}_k\right] \le A_k\left(F(x_k) - F^*\right) + \sigma^2, \quad (23)$$

where $\mathcal{G}_k = \nabla g_{i_k}\left(y_{k+1}\right)$, i_k is randomly picked from $[n] = \{1, 2, \cdots, n\}$ with probability $\frac{1}{n}$ via the independent identically distribution, the iteration of PSAG (15) is employed with a non-increasing stepsize t_k such that:

$$\sum_{k=1}^{\infty} t_k^2 < \infty,\ \sum_{k=1}^{\infty} A_k t_k < \infty,\ \sum_{k=1}^{\infty} \frac{t_{k+1}}{\sum_{s=1}^{k} t_s} = \infty \ \text{ and } \ t_k \le \frac{1}{4L},$$

where $A_k \ge 0$ is defined in (23), then we have:

$$\left\{F\left(x_k\right) - F^*\right\} \quad a.s.$$

If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, we obtain:

$$\min_{1 \le i \le k} \left\| G_{t_k}\left(y_{k+1}\right) \right\|^2 = o\left(\frac{1}{\sum_{s=1}^{k} t_s}\right) \quad a.s. \tag{24}$$

Proof We begin with the conclusion in (16) of Lemma 3.3. Firstly, by the definition of x_k^+ in Definition 2.4, let $u = x_k^+$, then set $t = \frac{t_k}{2}$, $z = \nabla g\left(x_k\right)$ and $y = x_k$ to meet the condition $u \in \arg\min_{x \in \mathbb{R}^d} \left\{ t\left\langle z, x\right\rangle + tf\left(x\right) + \frac{1}{2}\|x - y\|^2 \right\}$, *i.e.*

$$x_k^+ \in \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{t_k}{2}\left\langle \nabla g\left(x_k\right), x\right\rangle + \frac{t_k}{2} f\left(x\right) + \frac{1}{2}\left\|x - x_k\right\|^2 \right\} \tag{25}$$

From the update in (2), we can show that the iterate $x_k \in \mathbb{R}^d$ for any $k \ge 1$. Upon applying Lemma 3.3 with $u = x_k^+$, $t = \frac{t_k}{2}$, $z = \nabla g\left(x_k\right)$ and $y = x_k$ for $x = x_k \in \mathbb{R}^d$, we have:

$$\begin{aligned} F\left(x_k^+\right) &\le F\left(x_k\right) + \left\langle \nabla g\left(x_k\right) - \nabla g\left(x_k\right), x_k^+ - x_k \right\rangle + \frac{1}{t_k}\left\|x_k - x_k\right\|^2 \\ &\quad - \frac{1}{t_k}\left\|x_k^+ - x_k\right\|^2 + \frac{L}{2}\left\|x_k^+ - x_k\right\|^2 \\ &= F\left(x_k\right) + \left(\frac{L}{2} - \frac{1}{t_k}\right)\left\|x_k^+ - x_k\right\|^2. \end{aligned} \tag{26}$$

Next, by the update rule in (6) of PSAG algorithm, we choose $u = x_{k+1}$, then set $t = t_k$, $z = \mathcal{G}_k = \nabla g_i\left(y_{k+1}\right)$ and $y = y_{k+1}$ to meet the condition $u \in \arg\min_{x \in \mathbb{R}^n} \left\{ t\left\langle z, x\right\rangle + tf\left(x\right) + \frac{1}{2}\|x - y\|^2 \right\}$, *i.e.*

$$x_{k+1} \in \arg\min_{x \in \mathbb{R}^d} \left\{ t_k \left\langle \mathcal{G}_k, x\right\rangle + t_k f\left(x\right) + \frac{1}{2}\left\|x - y_{k+1}\right\|^2 \right\} \tag{27}$$

Similarly, from the definition of x_k^+ in (25), we can show that for any $k \ge 1$, the sequence $x_k^+ \in \mathbb{R}^d$. Now, upon applying Lemma 3.3 with $u = x_{k+1}$, $t = t_k$, $z = \mathcal{G}_k$, and $y = y_{k+1}$, for $x = x_k^+ \in \mathbb{R}^d$, we then obtain:

$$
\begin{aligned}
F\left(x_{k+1}\right) \leq & F\left(x_k^{+}\right)+\left\langle\nabla g\left(y_{k+1}\right)-\mathcal{G}_k, x_{k+1}-x_k^{+}\right\rangle+\frac{1}{2 t_k}\left\|x_k^{+}-y_{k+1}\right\|^2 \\
& -\frac{1}{2 t_k}\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{L}{2}\left\|x_{k+1}-y_{k+1}\right\|^2 \\
= & F\left(x_k^{+}\right)+\left(\frac{L}{2}-\frac{1}{2 t_k}\right)\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{1}{2 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
& +\left\langle\nabla g\left(y_{k+1}\right)-\mathcal{G}_k, x_{k+1}-x_k^{+}\right\rangle .
\end{aligned} \tag{28}
$$

Now, we consider $\left\langle\nabla g\left(y_{k+1}\right)-\mathcal{G}_k, x_{k+1}-x_k^{+}\right\rangle$,

$$
\begin{aligned}
& \left\langle\nabla g\left(y_{k+1}\right)-\mathcal{G}_k, x_{k+1}-x_k^{+}\right\rangle \\
& \leq\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\| \cdot\left\|x_{k+1}-x_k^{+}\right\| \\
& \leq 2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2+\frac{1}{8 t_k}\left\|x_{k+1}-x_k^{+}\right\|^2 \\
& =2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2+\frac{1}{8 t_k}\left\|x_{k+1}-y_{k+1}+y_{k+1}-x_k^{+}\right\|^2 \\
& \leq 2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2+\frac{1}{4 t_k}\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{1}{4 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2,
\end{aligned} \tag{29}
$$

where we use the Cauchy-Schwarz $|\langle a, b\rangle| \leq\|a\|\|b\|$ in the first inequality, and the second inequality follows from $ab \leq(1/2)\left(a^2+b^2\right)$ with $a=2 \sqrt{t_k}\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|$ and $b=\frac{1}{2 \sqrt{t_k}}\left\|x_{k+1}-x_k^{+}\right\|$. The last inequality holds by $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$ with $a=x_{k+1}-y_{k+1}$ and $b=y_{k+1}-x_k^{+}$.

Upon substituting (29) back into (28), we further have:

$$
\begin{aligned}
F\left(x_{k+1}\right) \leq & F\left(x_k^{+}\right)+\left(\frac{L}{2}-\frac{1}{2 t_k}\right)\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{1}{2 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
& +\left\langle\nabla g\left(y_{k+1}\right)-\mathcal{G}_k, x_{k+1}-x_k^{+}\right\rangle \\
\leq & F\left(x_k^{+}\right)+\left(\frac{L}{2}-\frac{1}{2 t_k}\right)\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{1}{2 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
& +2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2+\frac{1}{4 t_k}\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{1}{4 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
= & F\left(x_k^{+}\right)+\left(\frac{L}{2}-\frac{1}{4 t_k}\right)\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{3}{4 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
& +2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 \\
\leq & F\left(x_k\right)+\left(\frac{L}{2}-\frac{1}{t_k}\right)\left\|x_k^{+}-x_k\right\|^2+2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 \\
& +\left(\frac{L}{2}-\frac{1}{4 t_k}\right)\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{3}{4 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 \\
\leq & F\left(x_k\right)+2 t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2-\frac{1}{8 t_k}\left\|x_{k+1}-y_{k+1}\right\|^2+\frac{3}{4 t_k}\left\|y_{k+1}-x_k^{+}\right\|^2 .
\end{aligned} \tag{30}
$$

where the last inequality holds by the stepsize condition $t_k \leq 1/4L$, *i.e.*

$L/2-1/4t_k \le -1/8t_k$.

Next, we consider $\frac{3}{4t_k}\left\|y_{k+1}-x_k^+\right\|^2$ and combine Definition 2.4 with (5) and (15), we have:

$$x_k^+ = T_{t_k}^{f,g_{i_k}}\left(y_{k+1}\right) = \operatorname{prox}_{t_k f}\left(y_{k+1}-t_k\nabla g_{i_k}\left(y_{k+1}\right)\right),$$

and

$$\begin{aligned} G_{t_k}\left(y_{k+1}\right) &= \frac{1}{t_k}\left(y_{k+1}-T_{t_k}^{f,g_{i_k}}\left(y_{k+1}\right)\right) \\ &= \frac{1}{t_k}\left(y_{k+1}-x_k^+\right), \end{aligned}$$

Due to the convexity of *f*, we know that there exists $M>0$ such that:

$$\left\|G_{t_k}\left(y_{k+1}\right)\right\| < M.$$

Therefore,

$$\frac{3}{4t_k}\left\|y_{k+1}-x_k^+\right\|^2 = \frac{3t_k}{4}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 \le \frac{3M^2}{4}t_k. \tag{31}$$

Then, by (15), we also have:

$$-\frac{1}{8t_k}\left\|x_{k+1}-y_{k+1}\right\|^2 = -\frac{t_k}{8}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2. \tag{32}$$

Upon combining (31) and (32) with (30), we have:

$$F\left(x_{k+1}\right) \le F\left(x_k\right) - \frac{t_k}{8}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 + 2t_k\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 + \frac{3M^2}{4}t_k. \tag{33}$$

Now, the conditional expectation on (33) with respect to $\mathcal{F}_k$ gives:

$$\begin{aligned} &\mathbb{E}\left[F\left(x_{k+1}\right) \mid \mathcal{F}_k\right] \\ &\le \mathbb{E}\left[F\left(x_k\right) \mid \mathcal{F}_k\right] - \frac{t_k}{8}\mathbb{E}\left[\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 \mid \mathcal{F}_k\right] \\ &\quad + 2t_k\mathbb{E}\left[\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 \mid \mathcal{F}_k\right] + \frac{3M^2}{4}t_k \\ &= F\left(x_k\right) - \frac{t_k}{8}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 + 2t_k\mathbb{E}\left[\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 \mid \mathcal{F}_k\right] + \frac{3M^2}{4}t_k \\ &\le F\left(x_k\right) - \frac{t_k}{8}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 + 2t_k\left(A_k\left(F\left(x_k\right)-F^*\right)+\sigma^2\right) + \frac{3M^2}{4}t_k \\ &= F\left(x_k\right) - \frac{t_k}{8}\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 + 2A_k t_k\left(F\left(x_k\right)-F^*\right) + \left(2\sigma^2+\frac{3M^2}{4}\right)t_k. \end{aligned} \tag{34}$$

where the first equality holds by the fact that x_k is $\mathcal{F}_k$-measurable, *i.e.* $\mathbb{E}\left[F\left(x_k\right) \mid \mathcal{F}_k\right] = F\left(x_k\right)$ and $\mathbb{E}\left[\left\|G_{t_k}\left(y_{k+1}\right)\right\|^2 \mid \mathcal{F}_k\right] = \left\|G_{t_k}\left(y_{k+1}\right)\right\|^2$. The last inequality follows from $\mathbb{E}\left[\left\|\nabla g\left(y_{k+1}\right)-\mathcal{G}_k\right\|^2 \mid \mathcal{F}_k\right] \le A_k\left(F\left(x_k\right)-F^*\right)+\sigma^2$.

Upon subtracting F^* from both sides of (34), we have:

$$\begin{aligned}&\mathbb{E}\left[F(x_{k+1})-F^* \mid \mathcal{F}_k\right]\\&\le F(x_k)-F^*-\frac{t_k}{8}\left\|G_{t_k}(y_{k+1})\right\|^2+2A_kt_k\left(F(x_k)-F^*\right)+\left(2\sigma^2+\frac{3M^2}{4}\right)t_k \qquad (35)\\&=(1+2A_kt_k)\left(F(x_k)-F^*\right)-\frac{t_k}{8}\left\|G_{t_k}(y_{k+1})\right\|^2+\left(2\sigma^2+\frac{3M^2}{4}\right)t_k.\end{aligned}$$

Upon multiplying (35) by t_{k+1}, we can obtain:

$$\begin{aligned}&\mathbb{E}\left[t_{k+1}\left(F(x_{k+1})-F^* \mid \mathcal{F}_k\right)\right]\\&\le(1+2A_kt_k)t_{k+1}\left(F(x_k)-F^*\right)-\frac{t_kt_{k+1}}{8}\left\|G_{t_k}(y_{k+1})\right\|^2+\left(2\sigma^2+\frac{3M^2}{4}\right)t_kt_{k+1} \qquad (36)\\&\le(1+2A_kt_k)t_k\left(F(x_k)-F^*\right)-\frac{t_kt_{k+1}}{8}\left\|G_{t_k}(y_{k+1})\right\|^2+\left(2\sigma^2+\frac{3M^2}{4}\right)t_k^2.\end{aligned}$$

where the last inequality follows from the non-increasing behaviour of the stepsize t_k, *i.e.* $t_{k+1}<t_k$.

Finally, let $Y_k=t_k\left(F(x_k)-F^*\right)$, $\gamma_k=2A_kt_k$, $X_k=\frac{t_k}{8}\left\|G_{t_k}(y_{k+1})\right\|^2$ and $Z_k=\left(2\sigma^2+\frac{3M^2}{4}\right)t_k^2$, then (36) becomes:

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k]\le(1+\gamma_k)Y_k-t_{k+1}X_k+Z_k.$$

Recalling the stepsize conditions $\sum_{k=1}^{\infty}t_k^2<\infty$ and $\sum_{k=1}^{\infty}A_kt_k<\infty$, we know that $\sum_{k=1}^{\infty}Z_k<\infty$, $\sum_{k=1}^{\infty}\gamma_k<\infty$. Thus, by Lemma 3.1, we have:

$$\sum_{k=1}^{\infty}t_{k+1}X_k<\infty \ a.s. \text{ and } \left\{F(x_k)-F^*\right\} \ a.s.$$

Finally, with the condition $\sum_{k=1}^{\infty}\frac{t_{k+1}}{\sum_{s=1}^{k}t_s}=\infty$ of Lemma 3.2, we have:

$$\min_{1\le r\le k}t_r\left\|G_{t_r}(y_{k+1})\right\|^2=o\left(\frac{1}{\sum_{s=1}^{k}t_s}\right) \ a.s. \qquad (37)$$

This completes the proof.

Remark 3.1. (23) *is a new variance assumption on the stochastic gradient, where* $\mathcal{G}_k$ *is weaker than the bounded variance assumption* (*i.e.* $A_k=0$ *in* [33]).

Remark 3.2. *It is noting that we get the boundedness of the gradient mapping* $G_{t_k}(y_{k+1})$ *due to the continuity of the prox-grad operator* $T_{t_k}^{f,g_{i_k}}(y_{k+1})$ *which is deduced by the convexity of f.*

Corollary 3.1. *Following the setting of Theorem* 3.1 *and choosing the stepsize* $t_k=\frac{\beta}{1+\gamma\beta k^{\frac{1}{2}+\varepsilon}}$ *for any* $\gamma,\beta\ge0$ *where* $\varepsilon\in\left(0,\frac{1}{2}\right)$ *gives:*

$$\min_{1\le r\le k}\left\|G_{t_r}(y_{k+1})\right\|^2=o(1) \ a.s. \qquad (38)$$

Proof Stepsize satisfying $t_k = \frac{\beta}{1+\gamma\beta k^{\frac{1}{2}+\varepsilon}}$ has been studied in [23]. It follows that:

$$1+\gamma\beta k^{\frac{1}{2}+\varepsilon} \le k^{\frac{1}{2}+\varepsilon}+\gamma\beta k^{\frac{1}{2}+\varepsilon} = \left(1+\gamma\beta\right)k^{\frac{1}{2}+\varepsilon}, \tag{39}$$

which implies that $t_k \ge \frac{\beta}{\left(1+\gamma\beta\right)k^{\frac{1}{2}+\varepsilon}}$. Upon by the integral test inequality, we have:

$$\begin{aligned}\sum_{s=1}^{k} t_s &\ge \sum_{s=1}^{k}\frac{\beta}{\left(1+\gamma\beta\right)s^{\frac{1}{2}+\varepsilon}} \ge \int_1^k \frac{\beta}{\left(1+\gamma\beta\right)x^{\frac{1}{2}+\varepsilon}}\mathrm{d}x \\ &= \frac{\beta\left(k^{\frac{1}{2}-\varepsilon}-1\right)}{\left(1+\gamma\beta\right)\left(\frac{1}{2}-\varepsilon\right)} \ge \frac{\beta k^{-\frac{1}{2}-\varepsilon}\left(k-1\right)}{1+\gamma\beta},\end{aligned} \tag{40}$$

where the last inequality follows from the concavity of $h\left(x\right)=x^{\frac{1}{2}-\varepsilon}$, so that:

$$h\left(y\right)\le h\left(x\right)+h'\left(x\right)\left(y-x\right).$$

In other words, by taking $y=1$ and $x=k$, we can get $k^{\frac{1}{2}-\varepsilon}-1 \ge \left(\frac{1}{2}-\varepsilon\right)k^{-\frac{1}{2}-\varepsilon}\left(k-1\right)$.

Next, combining (39) and (40), we can obtain:

$$\left(\sum_{s=1}^{k} t_s\right)t_k \ge \frac{\beta k^{-\frac{1}{2}-\varepsilon}\left(k-1\right)}{1+\gamma\beta}\frac{\beta}{\left(1+\gamma\beta\right)k^{\frac{1}{2}+\varepsilon}} = \frac{\beta^2 k^{-1-2\varepsilon}\left(k-1\right)}{\left(1+\gamma\beta\right)^2}. \tag{41}$$

Upon setting $G_r = \left\|G_{t_r}\left(y_{k+1}\right)\right\|^2$, and applying the inequality $\min_{1\le r\le k} t_r G_r \ge t_k \min_{1\le r\le k} G_r$, we have:

$$\left(\sum_{s=1}^{k} t_s\right)\min_{1\le r\le k} t_r G_r \ge \left(\sum_{s=1}^{k} t_s\right)t_k \min_{1\le r\le k} G_r \ge \frac{\beta^2 k^{-1-2\varepsilon}\left(k-1\right)}{\left(1+\gamma\beta\right)^2}\min_{1\le r\le k} G_r \ge 0. \tag{42}$$

For the left hand side of (42), we apply Theorem 3.1 to get:

$$\lim_{k\to\infty}\left(\sum_{s=1}^{k} t_s\right)\min_{1\le r\le k} t_k G_r = 0 \;\; a.s.$$

The application of Squeeze theorem in conjunction with (42) gives:

$$\lim_{k\to\infty} k^{-1-2\varepsilon}\left(k-1\right)\min_{1\le r\le k} G_r = 0 \;\; a.s.$$

When $\varepsilon \to 0$, we have $\frac{k-1}{k^{1+2\varepsilon}} \to 1$, which implies that:

$$\lim_{k\to\infty}\min_{1\le r\le k} G_r = 0 \;\; a.s.$$

Therefore, we finally obtain:

$$\min_{1\le r\le k} G_r = \min_{1\le r\le k} \left\| G_{t_r}\left(y_{k+1}\right)\right\|^2 = o(1) \ a.s.$$

This completes the proof.

Remark 3.3. *Corollary* 3.1 *indicates that* $\min_{1\le r\le k} \left\| G_{t_r}\left(y_{k+1}\right)\right\|^2$ *is arbitrarily close to* 0 *with probability one*.

4. Conclusion

This paper presents PSAG with unbiased gradient estimation and analyzes its almost sure convergence for solving composite optimization problems, wherein the objective function comprises a smooth component and a non-smooth component. By leveraging certain key assumptions, we have established the almost sure convergence of PSAG with unbiased gradient estimation. Furthermore, we have demonstrated that the minimum of the squared gradient mapping norm approaches zero arbitrarily closely with probability one.

Founding

This paper was supported by the Natural Science Foundation of China (Grant No. 11801455), the Sichuan Science and Technology Program (Grant No. 2023NSFSC1922), the Innovation Team Funds of China West Normal University (Grant No. KCXTD2023-3), and the Fundamental Research Funds of China West Normal University (Grant No. 23kc010).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wright, S.J., Nowak, R.D. and Figueiredo, M.A.T. (2009) Sparse Reconstruction by Separable Approximation. *IEEE Transactions on Signal Processing*, **57**, 2479-2493. https://doi.org/10.1109/TSP.2009.2016892

[2] Bottou, L., Curtis, F.E. and Nocedal, J. (2018) Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, **60**, 223-311. https://doi.org/10.1137/16M1080173

[3] Mandic, D. and Chambers, J. (2001) Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. Wiley, New York. https://doi.org/10.1002/047084535X

[4] Shalev-Shwartz, S. and Ben-David, S. (2014) Understanding Machine Learning: from Theory to Algorithms. Cambridge University Press, New York. https://doi.org/10.1017/CBO9781107298019

[5] Sun, R.Y. (2020) Optimization for Deep Learning: An Overview. *Journal of the Operations Research Society of China*, **8**, 249-294. https://doi.org/10.1007/s40305-020-00309-6

[6] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *Annals of Mathematical Statistics*, **22**, 400-407. https://doi.org/10.1214/aoms/1177729586

[7] Beck, A. (2017) First-Order Methods in Optimization. Society for Industrial and Ap-

plied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611974997

[8] Nesterov, Y. (2013) Gradient Methods for Minimizing Composite Functions. *Mathematical Programming*, **140**, 125-161. https://doi.org/10.1007/s10107-012-0629-5

[9] Beck, A. and Teboulle, M. (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**, 183-202. https://doi.org/10.1137/080716542

[10] Vandenberghe, L. (2010) Fast Proximal Gradient Methods. http://faculty.bicmr.pku.edu.cn/~wenzw/courses/fgrad.pdf

[11] Becker, S., Candès, E. and Grant, M. (2011) Templates for Convex Cone Problems with Applications to Sparse Signal Recovery. *Mathematical Programming Computation*, **3**, 165-218. https://doi.org/10.1007/s12532-011-0029-5

[12] Li, H. and Lin, Z.C. (2015) Accelerated Proximal Gradient Methods for Nonconvex Programming. *The* 28*th International Conference on Neural Information Processing Systems*, Montreal, 7-12 December 2015, 379-387.

[13] Ghadimi, S., Lan, G. and Zhang, H. (2016) Mini-Batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization. *Mathematical Programming*, **155**, 267-305. https://doi.org/10.1007/s10107-014-0846-1

[14] Reddi, S.J., Sra, S., Póczos, B. and Smola, A.J. (2016) Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization. *Advances in Neural Information Processing Systems*, **29**, 1145-1153.

[15] Metel, M.R. and Takeda, A. (2019) Stochastic Proximal Methods for Non-Smooth Non-Covex Constrained Sparse Optimization. *Journal of Machine Learning Research*, **22**, 1-36.

[16] Kawashima, T. and Fujisawa, H. (2018) Stochastic Gradient Descent for Stochastic Doubly-Nonconvex Composite Optimization. arXiv: 1805.07960.

[17] Gorbunov, E., Hanzely, F. and Richtárik, P. (2020) A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent. arXiv: 1905.11261. https://arxiv.org/abs/1905.11261

[18] Cevher, V. and Vũ, B.C. (2019) On the Linear Convergence of the Stochastic Gradient Method with Constant Step-Size. *Optimization Letters*, **13**, 1177-1187. https://doi.org/10.1007/s11590-018-1331-1

[19] Polyak, T.B. (1964) Some Methods of Speeding up the Convergence of Iteration Methods. *USSR Computational Mathematics and Mathematical Physics*, **4**, 1-17. https://doi.org/10.1016/0041-5553(64)90137-5

[20] Biscarat, C.J. (1994) Almost Sure Convergence of a Class of Stochastic Algorithms. *Stochastic Processes and their Applications*, **50**, 83-99. https://doi.org/10.1016/0304-4149(94)90149-X

[21] Sebbouh, O., Gower, M.R. and Defazio, A. (2021) Almost Sure Convergence Rates for Stochastic Gradient Descent and Stochastic Heavy Ball. *Proceedings of Machine Learning Research*, **134**, 1-37.

[22] Orabona, F. (2020) Almost Sure Convergence of SGD on Smooth Nonconvex Functions. http://parameterfree.com https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/

[23] Liu, J. and Yuan, Y. (2022) On Almost Sure Convergence Rates of Stochastic Gradient Methods. arXiv: 2202.04295. https://arxiv.org/abs/2202.04295

[24] Liang, Y.Q., Xu, D.P., Zhang, N.M. and Mandic, D.P. (2023) Almost Sure Convergence of Stochastic Composite Objective Mirror Descent for Non-Convex Non-Smooth Op-

timization. *Optimization Letters*. https://doi.org/10.1007/s11590-023-01972-3

[25] Khaled, A. and Richtárik, P. (2020) Better Theory for SGD in the Nonconvex World. arXiv: 2002.03329.

[26] Gower, R., Sebbouh, O. and Loizou, N. (2021) SGD for Structured Nonconvex Functions: Learning Rates, Mini-Batching and Interpolation. *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics*, Vol. 130, 13-15 April 2021, 1315-1323.

[27] Mertikopoulos, P., Hallak, N., Kavis, A. and Cevher, V. (2020) On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems. *Optimization and Control*, **33**, 1117-1128. https://arxiv.org/abs/2006.11144

[28] Ward, R., Wu, X. and Bottou, L. (2020) AdaGrad Stepsizes: Sharp Convergence over Nonconvex Landscapes. *Journal of Machine Learning Research*, **21**, 9047-9076.

[29] Alacaoglu, A., Malitsky, Y. and Cevher, V. (2020) Convergence of Adaptive Algorithms for Weakly Convex Constrained Optimization. arXiv: 2006.06650.

[30] Davis, D. and Drusvyatskiy, D. (2019) Stochastic Model-Based Minimization of Weakly Convex Functions. *SIAM Journal on Optimization*, **29**, 207-239. https://doi.org/10.1137/18M1178244

[31] Nesterov, Y. (2003) Introductory Lectures on Convex Optimization: A Basic Course. Springer Science & Business Media, New York. https://doi.org/10.1007/978-1-4419-8853-9

[32] Robbins, H. and Siegmund, D. (1971) A Convergence Theorem for Non-Negative Almost Supermartingales and Some Applications. In: Rustagi, J.S., Ed., *Optimizing Methods in Statistics*, Academic Press, Cambridge, 233-257. https://doi.org/10.1016/B978-0-12-604550-5.50015-8

[33] Nesterov, Y. (2018) Smooth Convex Optimization. *Lectures on Convex Optimization*, **137**, 59-137. https://doi.org/10.1007/978-3-319-91578-4_2