



Almost sure convergence of stochastic composite objective mirror descent for non-convex non-smooth optimization

Yuqing Liang¹ · Dongpo Xu¹ · Naimin Zhang² · Danilo P. Mandic³

Received: 22 August 2022 / Accepted: 5 January 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Stochastic composite objective mirror descent (SCOMID) is an effective method for solving large-scale stochastic composite problems in machine learning. This method can efficiently use the geometric properties of a problem through a general distance function. However, most existing analyses rely on the convexity of the problem and the unbiased assumption of the stochastic gradient. In addition, the convergence results are obtained in expectation. To this end, we present an almost sure convergence analysis of SCOMID with biased gradient estimation in the non-convex non-smooth setting. For this general case, the analysis shows that the minimum of the squared generalized projected gradient norm arbitrarily converges to zero with probability one. We also obtain the almost sure convergence of function values for SCOMID with time-varying stepsizes in the non-convex and non-smooth setting. Numerical experiments support our theoretical findings.

Keywords SCOMID · Almost sure convergence · Non-convex and non-smooth optimization · Biased gradient estimation · Stochastic optimization · Machine learning

✉ Dongpo Xu
xudp100@nenu.edu.cn

Danilo P. Mandic
d.mandic@imperial.ac.uk

¹ Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, People's Republic of China

² College of Mathematics and Physics, Wenzhou University, Wenzhou 325035, People's Republic of China

³ Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK

1 Introduction

Consider the stochastic composite optimization problem of the form

$$\min_{x \in X} \left\{ \Phi(x) = F(x) + R(x) \right\}, \quad (1)$$

where $X \subset \mathbb{R}^d$ is a closed convex set, function $F(x) := \mathbb{E}_{\xi \sim D}[f(x, \xi)]$ is a smooth (possibly non-convex) mapping, where ξ is a random variable which follows an unknown distribution D , and $R(x)$ is a non-convex and non-smooth regularizer used to prevent over-fitting. The problem in (1) is very common in machine learning, statistics, signal processing and imaging, and other related areas [3, 24, 31, 32]. Stochastic composite objective mirror descent (SCOMID) with a general distance function is one of the most popular methods for solving the problem in (1), and takes the form

$$x_{t+1} \in \arg \min_{x \in X} \left\{ \eta \langle g_t, x \rangle + \eta R(x) + B_{\psi_t}(x, x_t) \right\}, \quad (2)$$

where η is the stepsize, g_t is the stochastic gradient, and B_{ψ_t} denotes the Bregman distance with respect to the μ -strongly convex function ψ_t . i.e., $B_{\psi_t}(x, x_t) = \psi_t(x) - \psi_t(x_t) - \langle \nabla \psi_t(x_t), x - x_t \rangle$. It can be observed that by choosing different distance generating functions, ψ_t in (2), we can obtain different stochastic optimization algorithms, some of which are listed in Table 1.

Although SCOMID has been demonstrated to work well for solving the composite optimization problem (1), the classical analysis is performed in the restricted convex setting [9]. However, in practical applications, $F(x)$ and $R(x)$ may be non-convex, and also $R(x)$ is usually non-smooth. Examples include principal component analysis (PCA) with non-convex loss $F(x)$ [19], speckle noise removal model with non-convex regularizer $R(x)$ [15, 28], and ℓ_0 minimization with non-smooth regularizer $R(x)$ [37]. On the other hand, the convergence analyses of stochastic optimization algorithms depend on another restrictive unbiased assumption of

Table 1 Some special cases of SCOMID with different parameter values, where $\|x\|_{G_t^{1/2}} = \langle x, G_t^{1/2} x \rangle$ and $\|x\|^2 = \langle x, x \rangle$

Algorithm	C	R	$\psi_t(x)$	Iterate
AdaGrad [8]	X	$R(x)$	$\ x\ _{G_t^{1/2}}/2$	$x_{t+1} = \arg \min_{x \in X} \left\{ \eta_t \langle g_t, x \rangle + \eta_t R(x) + \frac{1}{2} \ x - x_t\ _{G_t^{1/2}} \right\}$
Prox-SGD [10]	\mathbb{R}^d	$R(x)$	$\ x\ ^2/2$	$x_{t+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ \eta_t \langle g_t, x \rangle + \eta_t R(x) + \frac{1}{2} \ x - x_t\ ^2 \right\}$
SPGD [36]	X	—	$\ x\ ^2/2$	$x_{t+1} = \arg \min_{x \in X} \left\{ \eta_t \langle g_t, x \rangle + \frac{1}{2} \ x - x_t\ ^2 \right\}$
SGD [29]	\mathbb{R}^d	—	$\ x\ ^2/2$	$x_{t+1} = x_t - \eta_t g_t$

C = Constraint, R = Regularizer, AdaGrad = Adaptive stochastic gradient descent, Prox-SGD = Proximal stochastic gradient descent, SPGD = Stochastic projected gradient descent, SGD = Stochastic gradient descent

Table 2 Comparison of problem setting (see $F(x)$, $R(x)$ and Constraint), stochastic gradient estimation g_t , algorithm, and convergence results in this paper with some relevant literature

Citation	$F(x)$	$R(x)$	Constraint	g_t	Algorithm	a.s.
[12]	L -smooth	Convex	X	Unbiased	SCOMID	×
[4, 13]	L -smooth	Convex	\mathbb{R}^d	Unbiased	Prox-SGD	×
[14, 17]	L -smooth	—	\mathbb{R}^d	Unbiased	SGD	×
[35, 38]	L -smooth	—	\mathbb{R}^d	Unbiased	AdaGrad	×
[25]	L -smooth, G -Lipschitz	—	X	Unbiased	SGD	✓
[1]	—	ρ -weakly convex	X	Unbiased	AdaGrad	×
[5]	—	ρ -weakly convex	\mathbb{R}^d	Unbiased	Prox-SGD	×
[6]	L -smooth	Non-convex	\mathbb{R}^d	Biased	Prox-SGD	×
ours	L -smooth	Non-convex	X	Biased	SCOMID	✓

stochastic gradient estimation (i.e., $\mathbb{E}[g_t | \mathcal{F}_t] = \nabla F(x_t)$ ¹) [1, 4, 13]. However, biased gradients commonly exist due to the unknown distribution of the actual data samples [2]. In addition, the gradient estimations generated by effective variance reduction (VR) techniques adopted in machine learning are often biased [6]. Therefore, extending the analysis of SCOMID to a biased stochastic setting is of considerable practical significance. In particular, recent works [3, 22] focus on the convergence of SGD with biased gradient estimation, requiring the stochastic gradient to satisfy $\nabla F(x_t)^T \mathbb{E}[g_t | \mathcal{F}_t] \geq \mu \|\nabla F(x_t)\|^2$. In contrast, we do not impose any assumptions on the first moment of g_t (i.e., $\mathbb{E}[g_t | \mathcal{F}_t]$) in this paper. Also, most existing work focuses on convergence in expectation [4, 9, 12, 33], while the more practically relevant almost sure convergence analysis is still underexplored.

For these reasons, this paper aims to establish the convergence analysis of SCOMID in a general setting, by solving the following three issues: (i) non-convex regularizer $R(x)$, (ii) biased gradient estimation, and (iii) almost sure convergence. It should be noted that the constraint set, X , the non-convex regularizer, $R(x)$, and the general distance, B_{ψ_t} , bring challenges to our analysis do not exist in other related work, as elaborated in Table 2. The following are our main contributions:

- We *first* establish the almost sure convergence analysis of SCOMID in the non-convex non-smooth setting (see the last column of Table 2). More specifically, we prove that the generalized projected gradient, $\min_{1 \leq i \leq t} \|G_{\eta_i/2}(x_i, x_i^+)\|^2$, arbitrarily converges to zero with probability one, and the function values, $\Phi(x_t)$, almost surely converge to a finite limit.
- Our analysis is established based on the biased gradient estimation, as shown in the fifth column of Table 2, where the unbiasedness of stochastic gradient g_t [1, 5, 25] is removed. In addition, we develop a new variance assumption on the sto-

¹ \mathcal{F}_t denotes the σ -algebra generated by random variables x_1, x_2, \dots, x_t , i.e., $\mathcal{F}_t = \sigma(x_1, x_2, \dots, x_t)$.

chastic gradient, g_t (see Assumption 1(c)), which is somewhat weaker than the bounded variance assumption (i.e., $\alpha_t = 0$ in (27)) [12, 18, 35].

- We focus on the general composite optimization problem with a non-convex non-smooth regularizer $R(x)$. As shown in the third column of Table 2, this allows us to remove the standard convexity assumption on the regularizer $R(x)$ [4, 12, 13]. Moreover, we develop a non-convex non-smooth regularizer in Sect. 4, and then experimentally verify our theoretical findings.

2 Preliminaries

2.1 Notations

- The projected operator onto X :

$$P_x(y, g, \eta) = \arg \min_{x \in X} \{ \eta \langle g, x \rangle + \eta R(x) + B_{\psi_t}(x, y) \}. \quad (3)$$

- The projection gradient descent associated with $\nabla F(x_t)$:

$$x_t^+ \in P_x(x_t, \nabla F(x_t), \eta). \quad (4)$$

- The generalized projected gradient [12]:

$$G_\eta(x_t, x_t^+) = (x_t - x_t^+)/\eta. \quad (5)$$

Remark 1 The generalized projected gradient function, $G_\eta(x_t, x_t^+)$, is often used to measure convergence for non-convex optimization problems [12]. To have a better understanding of $G_\eta(x_t, x_t^+)$, we explore its role for convergence analysis in the following section.

2.2 Convergence criterion

The convergence criterion is important in studying non-smooth non-convex problems, whereby the function values, $\Phi(x) - \Phi^*$, are typically employed as a criterion in the convex setting. For non-convex but smooth problems, i.e., $R(x) = 0$ in (1), the gradient norm $\|\nabla F(x)\|$ is used to measure stationarity. In this section, we provide some special cases of $G_\eta(x_t, x_t^+)$ to clarify the reason for adopting it as a criterion for non-smooth non-convex objectives.

Example 1 (Gradient) If $X = \mathbb{R}^d$, $R(x) = 0$, and $\psi_t(x) = \|x\|^2/2$, we can obtain $B_{\psi_t}(x, y) = \|x - y\|^2/2$. In addition, due to the smoothness of $F(x)$, the problem in (1) transforms into a smooth one, i.e., $\Phi(x)$ is smooth, to give

$$\min_{x \in \mathbb{R}^d} \{ \Phi(x) = F(x) \}. \quad (6)$$

By the definition of x_t^+ , we can then obtain

$$x_t^+ \stackrel{(4),(3)}{=} \arg \min_{x \in \mathbb{R}^d} \left\{ \eta \langle \nabla F(x_t), x \rangle + \frac{1}{2} \|x - x_t\|^2 \right\} = x_t - \eta \nabla F(x_t), \quad (7)$$

where the last inequality follows from the optimality of x_t^+ , i.e., $\eta \nabla F(x_t) + (x_t^+ - x_t) = 0$. In this case, $G_\eta(x_t, x_t^+)$ directly reduces to the gradient of function F at x_t , i.e.,

$$G_\eta(x_t, x_t^+) \stackrel{(5)}{=} \frac{1}{\eta} (x_t - x_t^+) \stackrel{(7)}{=} \frac{1}{\eta} (x_t - (x_t - \eta \nabla F(x_t))) = \nabla F(x_t), \quad (8)$$

which is used to measure the convergence for non-convex but smooth objective functions [17, 20].

Example 2 (Gradient of the Moreau envelope) If $X = \mathbb{R}^d$, $F(x) = 0$, and $\psi_t(x) = \|x\|^2/2$, then the composite optimization in (1) becomes

$$\min_{x \in \mathbb{R}^d} \{ \Phi(x) = R(x) \}. \quad (9)$$

Further, we can obtain

$$x_t^+ \stackrel{(4),(3)}{\in} \arg \min_{x \in \mathbb{R}^d} \left\{ \eta R(x) + \frac{1}{2} \|x - x_t\|^2 \right\} = \text{prox}_{\eta R}(x_t), \quad (10)$$

Then, $G_\eta(x_t, x_t^+)$ becomes the gradient of the Moreau envelope. i.e.,

$$G_\eta(x_t, x_t^+) \stackrel{(5)}{=} \frac{1}{\eta} (x_t - x_t^+) \stackrel{(10)}{=} \frac{1}{\eta} (x_t - \text{prox}_{\eta R}(x_t)), \quad (11)$$

which measures the convergence when optimizing weakly convex problems [5, 7, 23].

Example 3 (Gradient mapping) If $X = \mathbb{R}^d$, and $\psi_t(x) = \|x\|^2/2$, the problem in (1) assumes the following unconstrained form

$$\min_{x \in \mathbb{R}^d} \{ \Phi(x) = F(x) + R(x) \}. \quad (12)$$

Then, we have

$$\begin{aligned} x_t^+ &\stackrel{(4),(3)}{\in} \arg \min_{x \in \mathbb{R}^d} \left\{ \eta \langle \nabla F(x_t), x \rangle + \eta R(x) + \frac{1}{2} \|x - x_t\|^2 \right\} \\ &= \text{prox}_{\eta R}(x_t - \eta \nabla F(x_t)), \end{aligned} \quad (13)$$

where the last inequality follows from the definition of the proximal operator, i.e., $\text{prox}_h(x) = \arg \min_{u \in \mathbb{R}^d} \{ h(u) + \frac{1}{2} \|u - x\|^2 \}$. Thus, $G_\eta(x_t, x_t^+)$ will reduce to the gradient mapping [27]. i.e.,

$$G_\eta(x_t, x_t^+) \stackrel{(5)}{=} \frac{1}{\eta} (x_t - x_t^+) \stackrel{(13)}{=} \frac{1}{\eta} (x_t - \text{prox}_{\eta R}(x_t - \eta \nabla F(x_t))), \quad (14)$$

which is a standard measure for the convergence of proximal algorithms in the non-convex setting [6, 16, 26].

2.3 Definitions and lemmas

Definition 1 (*L-smoothness*) The differentiable function h is L -smooth if its gradient ∇h is Lipschitz continuous. i.e., there exists $L > 0$, such that

$$\|\nabla h(x) - \nabla h(y)\| \leq L\|x - y\|, \forall x, y \in \text{dom } h. \quad (15)$$

Definition 2 (*Proposition 4.8, [34]*) The function h is μ -strongly convex if there exists $\mu > 0$, such that

$$h(y) \geq h(x) + \langle g(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2. \quad (16)$$

Lemma 1 (*Lemma 1.2.3, [27]*) Assume that the function h is L -smooth, then

$$|h(y) - h(x) - \langle \nabla h(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2, \forall x, y \in \text{dom } h. \quad (17)$$

Lemma 2 (*Theorem 1, [30]*) Assume that $\{X_t\}$, $\{Y_t\}$ and $\{Z_t\}$ are three non-negative sequences of random variable, $\{\gamma_t\}$ is a non-negative real sequence and \mathcal{F}_t is a σ -algebra. If X_t, Y_t, Z_t are all \mathcal{F}_t -measurable and the following conditions hold

- (1) $\mathbb{E}[Y_{t+1} | \mathcal{F}_t] \leq (1 + \gamma_t)Y_t - X_t + Z_t$.
- (2) $\sum_{t=1}^{\infty} \gamma_t < \infty$, $\sum_{t=1}^{\infty} Z_t < \infty$ a.s.

Then, Y_t converges almost surely and $\sum_{t=1}^{\infty} X_t < \infty$ almost surely (a.s.).

Lemma 3 (*Lemma 3, [21]*) Assume that $\{X_t\}$ is a non-negative sequence of random variable and $\eta_t \geq 0$ is non-increasing. If the following conditions hold

$$\sum_{t=1}^{\infty} \frac{\eta_{t+1}}{\sum_{j=1}^t \eta_j} = \infty, \quad \sum_{t=1}^{\infty} \eta_{t+1} X_t < \infty \text{ a.s.} \quad (18)$$

then we have

$$\min_{1 \leq i \leq t} X_i = o\left(\frac{1}{\sum_{j=1}^t \eta_j}\right) \text{ a.s.} \quad (19)$$

where o denotes the higher-order infinitesimal. i.e., for two sequences $\{a_t\}$ and $\{b_t\}$, $a_t = o(b_t)$ if and only if $\lim_{t \rightarrow \infty} a_t/b_t = 0$.

Lemma 4 Suppose that $z \in \arg \min_{x \in X} \{ \eta \langle g, x \rangle + \eta R(x) + B_\psi(x, y) \}$ for some $g, y \in \mathbb{R}^d$ and the μ -strongly convex function $\psi(x)$. If $F(x)$ is L -smooth, then for any $x \in X$, we have

$$\begin{aligned} \Phi(z) &\leq \Phi(x) + \langle \nabla F(y) - g, z - x \rangle \\ &\quad + \frac{1}{\eta} B_\psi(x, y) - \frac{1}{\eta} B_\psi(z, y) + \frac{L}{2} \|y - x\|^2 + \frac{L}{2} \|z - y\|^2. \end{aligned} \quad (20)$$

Proof By the optimality of z , we can show that for any $x \in X$,

$$\eta \langle g, z \rangle + \eta R(z) + B_\psi(z, y) \leq \eta \langle g, x \rangle + \eta R(x) + B_\psi(x, y). \quad (21)$$

Upon rearranging the above equation, we have

$$R(z) \leq R(x) + \langle g, x - z \rangle + \frac{1}{\eta} B_\psi(x, y) - \frac{1}{\eta} B_\psi(z, y). \quad (22)$$

Since $F(x)$ is L -smooth, we can apply Lemma 1 to have

$$F(z) \leq F(y) + \langle \nabla F(y), z - y \rangle + \frac{L}{2} \|z - y\|^2, \quad (23)$$

$$F(x) \geq F(y) + \langle \nabla F(y), x - y \rangle - \frac{L}{2} \|y - x\|^2. \quad (24)$$

Next, upon combining (23) and (24), we arrive at

$$\begin{aligned} F(z) &\stackrel{(23)}{\leq} F(y) + \langle \nabla F(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \\ &\stackrel{(24)}{\leq} F(x) + \langle \nabla F(y), y - x \rangle + \frac{L}{2} \|y - x\|^2 + \langle \nabla F(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \\ &= F(x) + \langle \nabla F(y), z - x \rangle + \frac{L}{2} \|y - x\|^2 + \frac{L}{2} \|z - y\|^2. \end{aligned} \quad (25)$$

By the fact that $\Phi(z) = F(z) + R(z)$, we finally obtain

$$\begin{aligned} \Phi(z) &= F(z) + R(z) \\ &\stackrel{(22)}{\leq} F(z) + R(x) + \langle g, x - z \rangle + \frac{1}{\eta} B_\psi(x, y) - \frac{1}{\eta} B_\psi(z, y) \\ &\stackrel{(25)}{\leq} F(x) + \langle \nabla F(y), z - x \rangle + \frac{L}{2} \|y - x\|^2 + \frac{L}{2} \|z - y\|^2 \\ &\quad + R(x) + \langle g, x - z \rangle + \frac{1}{\eta} B_\psi(x, y) - \frac{1}{\eta} B_\psi(z, y) \\ &= \Phi(x) + \langle \nabla F(y) - g, z - x \rangle + \frac{1}{\eta} B_\psi(x, y) \\ &\quad - \frac{1}{\eta} B_\psi(z, y) + \frac{L}{2} \|y - x\|^2 + \frac{L}{2} \|z - y\|^2. \end{aligned} \quad (26)$$

This complete the proof. \square

3 Almost sure convergence rate analysis

Assumption 1 The objective function $\Phi(x)$ and SCOMID (see (2)) satisfy the following conditions:

- (a) The objective function $\Phi(x)$ is lower bounded, i.e., there exists a constant $\Phi^* > -\infty$, such that $\Phi(x) \geq \Phi^*$ for any x .
- (b) The function $F(x)$ is L -smooth, and $\psi_t(x)$ is μ -strongly convex.
- (c) There exist $\sigma > 0$ and $\alpha_t \geq 0$ such that

$$\mathbb{E}[\|g_t - \nabla F(x_t)\|^2 | \mathcal{F}_t] \leq \alpha_t (\Phi(x_t) - \Phi^*) + \sigma^2, \quad (27)$$

Remark 2 The convergence guarantees of SCOMID often rely on the unbiased estimation g_t and the convexity of the regularizer $R(x)$ [9, 12]. However, these two assumptions are absent in many practical applications, such as under the non-i.i.d. sampling in computer aided diagnosis [11] and the non-convex regularizer in speckle noise removal [15]. It is worth mentioning that, as shown in Assumption 1, we no longer require the unbiasedness of stochastic gradient g_t and the convexity of regularizer $R(x)$. In addition, compared with the bounded variance assumption (i.e., $\alpha_t = 0$ in Assumption 1(c)) used in [12, 18, 35], the slightly weaker Assumption 1(c) can effectively reduce the dependence of variance on the constant σ .

Theorem 1 Suppose that Assumption 1 holds and SCOMID (see (2)) is employed with a non-increasing stepsize, η_t , such that

$$\sum_{t=1}^{\infty} \eta_t^2 < \infty, \quad \sum_{t=1}^{\infty} \eta_t \alpha_t < \infty, \quad \sum_{t=1}^{\infty} \frac{\eta_{t+1}}{\sum_{j=1}^t \eta_j} = \infty, \quad \text{and } \eta_t \leq \frac{\mu}{8L}, \quad (28)$$

where $\alpha_t \geq 0$ is defined in (27). Then, we have

$$\min_{1 \leq i \leq t} \eta_i \|G_{\eta_i/2}(x_i, x_i^+)\|^2 = o\left(\frac{1}{\sum_{j=1}^t \eta_j}\right) \text{ a.s.} \quad (29)$$

Proof Since from Assumption 1 (b) the function $F(x)$ is L -smooth, we begin with the conclusion in (20) of Lemma 4. First, by the definition of x_t^+ in Sect. 2.1, let $z = x_t^+$, then set $\eta = \eta_t/2$, $\psi = \psi_t$, $g = \nabla F(x_t)$ and $y = x_t$ to meet the condition $z \in \arg \min_{x \in X} \{\eta \langle g, x \rangle + \eta R(x) + B_{\psi}(x, y)\}$, i.e.,

$$x_t^+ \in \arg \min_{x \in X} \left\{ \frac{\eta_t}{2} \langle \nabla F(x_t), x \rangle + \frac{\eta_t}{2} R(x) + B_{\psi_t}(x, x_t) \right\}. \quad (30)$$

From the update in (2) of SCOMID, we can show that the iterate $x_t \in X$ for any $t \geq 1$. Upon applying Lemma 4 with $z = x_t^+$, $\eta = \eta_t/2$, $\psi = \psi_t$, $g = \nabla F(x_t)$ and $y = x_t$, for $x = x_t \in X$, we have

$$\begin{aligned}
 \Phi(x_t^+) &\leq \Phi(x_t) + \langle \nabla F(x_t) - \nabla F(x_t), x_t^+ - x_t \rangle + \frac{L}{2} \|x_t - x_t\|^2 \\
 &\quad + \frac{L}{2} \|x_t - x_t^+\|^2 + \frac{2}{\eta_t} B_{\psi_t}(x_t, x_t) - \frac{2}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
 &= \Phi(x_t) + \frac{L}{2} \|x_t - x_t^+\|^2 - \frac{2}{\eta_t} B_{\psi_t}(x_t^+, x_t),
 \end{aligned} \tag{31}$$

Next, by the update rule in (2) of the SCOMID algorithm, we choose $z = x_{t+1}$, then set $\eta = \eta_t$, $\psi = \psi_t$, $g = g_t$ and $y = x_t$ to meet the condition $z \in \arg \min_{x \in X} \{ \eta \langle g, x \rangle + \eta R(x) + B_{\psi}(x, y) \}$, i.e.,

$$x_{t+1} \in \arg \min_{x \in X} \left\{ \eta_t \langle g_t, x \rangle + \eta_t R(x) + B_{\psi_t}(x, x_t) \right\}, \tag{32}$$

Similarly, from the definition of x_t^+ in (30), we can show that for any $t \geq 1$, the sequence $x_t^+ \in X$. Now, upon applying Lemma 4 with $z = x_{t+1}$, $\eta = \eta_t$, $\psi = \psi_t$, $g = g_t$ and $y = x_t$, for $x = x_t^+ \in X$, we then obtain

$$\begin{aligned}
 \Phi(x_{t+1}) &\leq \Phi(x_t^+) + \langle \nabla F(x_t) - g_t, x_{t+1} - x_t^+ \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
 &\quad + \frac{L}{2} \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) - \frac{1}{\eta_t} B_{\psi_t}(x_{t+1}, x_t) \\
 &\stackrel{(34)}{\leq} \Phi(x_t^+) + \langle \nabla F(x_t) - g_t, x_{t+1} - x_t^+ \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
 &\quad + \frac{L}{2} \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) - \frac{\mu}{2\eta_t} \|x_{t+1} - x_t\|^2 \\
 &= \Phi(x_t^+) + \left(\frac{L}{2} - \frac{\mu}{2\eta_t} \right) \|x_{t+1} - x_t\|^2 + \frac{L}{2} \|x_t - x_t^+\|^2 \\
 &\quad + \underbrace{\langle \nabla F(x_t) - g_t, x_{t+1} - x_t^+ \rangle + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t)}_{\clubsuit},
 \end{aligned} \tag{33}$$

where the second inequality holds by the μ -strongly convexity of $\psi_t(x)$ in Assumption 1 (b), i.e.,

$$B_{\psi_t}(x_{t+1}, x_t) = \psi_t(x_{t+1}) - \psi_t(x_t) - \langle \nabla \psi_t(x_t), x_{t+1} - x_t \rangle \geq \frac{\mu}{2} \|x_{t+1} - x_t\|^2. \tag{34}$$

The inner product \clubsuit in (33) can be estimated as follows

$$\begin{aligned}
\clubsuit &\leq \|\nabla F(x_t) - g_t\| \cdot \|x_{t+1} - x_t^+\| \\
&\leq \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \frac{\mu}{8\eta_t} \|x_{t+1} - x_t^+\|^2 \\
&= \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \frac{\mu}{8\eta_t} \|x_{t+1} - x_t + x_t - x_t^+\|^2 \\
&\leq \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \frac{\mu}{4\eta_t} \|x_{t+1} - x_t\|^2 + \frac{\mu}{4\eta_t} \|x_t - x_t^+\|^2,
\end{aligned} \tag{35}$$

where we use the Cauchy-Schwarz $|\langle a, b \rangle| \leq \|a\| \|b\|$ in the first inequality, and the second inequality follows from $ab \leq (1/2)(a^2 + b^2)$ with $a = (2\sqrt{\eta_t}/\sqrt{\mu}) \|\nabla F(x_t) - g_t\|$ and $b = (\sqrt{\mu}/(2\sqrt{\eta_t})) \|x_{t+1} - x_t^+\|$. The last inequality holds by $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ with $a = x_{t+1} - x_t$ and $b = x_t - x_t^+$. Upon substituting (35) back into (33), we further have

$$\begin{aligned}
\Phi(x_{t+1}) &\stackrel{(33)}{\leq} \Phi(x_t^+) + \clubsuit + \left(\frac{L}{2} - \frac{\mu}{2\eta_t}\right) \|x_{t+1} - x_t\|^2 + \frac{L}{2} \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
&\stackrel{(35)}{\leq} \Phi(x_t^+) + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \frac{\mu}{4\eta_t} \|x_{t+1} - x_t\|^2 + \frac{\mu}{4\eta_t} \|x_t - x_t^+\|^2 \\
&\quad + \left(\frac{L}{2} - \frac{\mu}{2\eta_t}\right) \|x_{t+1} - x_t\|^2 + \frac{L}{2} \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
&= \Phi(x_t^+) + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \left(\frac{L}{2} - \frac{\mu}{4\eta_t}\right) \|x_{t+1} - x_t\|^2 \\
&\quad + \left(\frac{L}{2} + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t).
\end{aligned} \tag{36}$$

Next, from the stepsize condition $\eta_t \leq \mu/(8L) < \mu/(2L)$, we can obtain $L/2 - \mu/(4\eta_t) < 0$, so we arrive at

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t^+) + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 \\
&\quad + \left(\frac{L}{2} + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t).
\end{aligned} \tag{37}$$

Upon combining (37) and (31), we arrive at

$$\begin{aligned}
 \Phi(x_{t+1}) &\stackrel{(37)}{\leq} \Phi(x_t^+) + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \left(\frac{L}{2} + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
 &\stackrel{(31)}{\leq} \Phi(x_t) + \frac{L}{2} \|x_t - x_t^+\|^2 - \frac{2}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
 &\quad + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 + \left(\frac{L}{2} + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 + \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) \\
 &= \Phi(x_t) + \left(L + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 - \frac{1}{\eta_t} B_{\psi_t}(x_t^+, x_t) + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 \quad (38) \\
 &\stackrel{(39)}{\leq} \Phi(x_t) + \left(L + \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 - \frac{\mu}{2\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 \\
 &= \Phi(x_t) + \left(L - \frac{\mu}{4\eta_t}\right) \|x_t - x_t^+\|^2 + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2 \\
 &\leq \Phi(x_t) - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t}{\mu} \|\nabla F(x_t) - g_t\|^2,
 \end{aligned}$$

where the last inequality holds by the stepsize condition $\eta_t \leq \mu/(8L)$, i.e., $L - \mu/(4\eta_t) \leq -\mu/(8\eta_t)$. We use the μ -strongly convexity of $\psi_t(x)$ in the third inequality above, i.e.,

$$B_{\psi_t}(x_t^+, x_t) = \psi_t(x_t^+) - \psi_t(x_t) - \langle \nabla \psi_t(x_t), x_t^+ - x_t \rangle \geq \frac{\mu}{2} \|x_t - x_t^+\|^2. \quad (39)$$

Now, the conditional expectation on (38) with respect to \mathcal{F}_t gives

$$\begin{aligned}
 \mathbb{E}[\Phi(x_{t+1})|\mathcal{F}_t] &\leq \mathbb{E}[\Phi(x_t)|\mathcal{F}_t] - \frac{\mu}{8\eta_t} \mathbb{E}[\|x_t - x_t^+\|^2|\mathcal{F}_t] + \frac{2\eta_t}{\mu} \mathbb{E}[\|\nabla F(x_t) - g_t\|^2|\mathcal{F}_t] \\
 &= \Phi(x_t) - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t}{\mu} \mathbb{E}[\|\nabla F(x_t) - g_t\|^2|\mathcal{F}_t] \\
 &\leq \Phi(x_t) - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t}{\mu} \left(\alpha_t(\Phi(x_t) - \Phi^*) + \sigma^2\right) \\
 &= \Phi(x_t) - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t\alpha_t}{\mu} (\Phi(x_t) - \Phi^*) + \frac{2\sigma^2\eta_t}{\mu}. \quad (40)
 \end{aligned}$$

where the first equality holds by the fact that x_t is \mathcal{F}_t -measurable, i.e., $\mathbb{E}[\Phi(x_t)|\mathcal{F}_t] = \Phi(x_t)$ and $\mathbb{E}[\|x_t - x_t^+\|^2|\mathcal{F}_t] = \|x_t - x_t^+\|^2$. The last inequality follows from Assumption 1 (c), i.e., $\mathbb{E}[\|g_t - \nabla F(x_t)\|^2|\mathcal{F}_t] \leq \alpha_t(\Phi(x_t) - \Phi^*) + \sigma^2$. Upon subtracting Φ_* from both sides of (40), we have

$$\begin{aligned}
& \mathbb{E}[\Phi(x_{t+1}) - \Phi^* | \mathcal{F}_t] \\
& \leq \Phi(x_t) - \Phi^* - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\eta_t\alpha_t}{\mu} (\Phi(x_t) - \Phi^*) + \frac{2\sigma^2\eta_t}{\mu} \\
& = \left(1 + \frac{2\eta_t\alpha_t}{\mu}\right) (\Phi(x_t) - \Phi^*) - \frac{\mu}{8\eta_t} \|x_t - x_t^+\|^2 + \frac{2\sigma^2\eta_t}{\mu} \\
& = \left(1 + \frac{2\eta_t\alpha_t}{\mu}\right) (\Phi(x_t) - \Phi^*) - \frac{\mu\eta_t}{32} \|G_{\eta_t/2}(x_t, x_t^+)\|^2 + \frac{2\sigma^2\eta_t}{\mu}.
\end{aligned} \tag{41}$$

where the last equality follows from the definition of $G_{\eta}(x_t, x_t^+)$ and (30), i.e.,

$$G_{\eta}(x_t, x_t^+) \stackrel{(5)}{=} \frac{x_t - x_t^+}{\eta} \stackrel{(30)}{=} \frac{x_t - x_t^+}{\eta_t/2} = \frac{2(x_t - x_t^+)}{\eta_t}, \tag{42}$$

which implies that $\|x_t - x_t^+\|^2 = (\eta_t^2/4) \|G_{\eta_t/2}(x_t, x_t^+)\|^2$. Upon multiplying (41) by η_{t+1} , we can obtain

$$\begin{aligned}
& \mathbb{E}[\eta_{t+1}(\Phi(x_{t+1}) - \Phi^*) | \mathcal{F}_t] \\
& \leq \left(1 + \frac{2\eta_t\alpha_t}{\mu}\right) \eta_{t+1} (\Phi(x_t) - \Phi^*) - \frac{\mu\eta_{t+1}\eta_t}{32} \|G_{\eta_t/2}(x_t, x_t^+)\|^2 + \frac{2\sigma^2\eta_t\eta_{t+1}}{\mu} \\
& \leq \left(1 + \frac{2\eta_t\alpha_t}{\mu}\right) \eta_t (\Phi(x_t) - \Phi^*) - \frac{\mu\eta_{t+1}\eta_t}{32} \|G_{\eta_t/2}(x_t, x_t^+)\|^2 + \frac{2\sigma^2\eta_t^2}{\mu}.
\end{aligned} \tag{43}$$

where the last inequality follows from the the non-increasing behaviour of the step-size η_t , i.e., $\eta_{t+1} \leq \eta_t$.

Finally, let $Y_t = \eta_t(\Phi(x_t) - \Phi^*)$, $\gamma_t = (2\eta_t\alpha_t)/\mu$, $X_t = (\mu\eta_t \|G_{\eta_t/2}(x_t, x_t^+)\|^2)/32$ and $Z_t = (2\sigma^2\eta_t^2)/\mu$. Then, (43) becomes

$$\mathbb{E}[Y_{t+1} | \mathcal{F}_t] \leq (1 + \gamma_t)Y_t - \eta_{t+1}X_t + Z_t. \tag{44}$$

Recall the stepsize conditions $\sum_{t=1}^{\infty} \eta_t\alpha_t < \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then we know that $\sum_{t=1}^{\infty} \gamma_t < \infty$ and $\sum_{t=1}^{\infty} Z_t < \infty$. Thus, by Lemma 2, we have

$$\sum_{t=1}^{\infty} \eta_{t+1}X_t < \infty \text{ a.s.} \tag{45}$$

Finally, a combination of Lemma 3 with the condition $\sum_{t=1}^{\infty} \frac{\eta_{t+1}}{\sum_{j=1}^t \eta_j} = \infty$ gives

$$\min_{1 \leq i \leq t} \eta_i \|G_{\eta_i/2}(x_i, x_i^+)\|^2 = o\left(\frac{1}{\sum_{j=1}^t \eta_j}\right) \text{ a.s.} \tag{46}$$

This completes the proof. \square

Corollary 1 *Following the setting of Theorem 1 and choosing the stepsize $\eta_t = \beta/(1 + \gamma\beta t^{\frac{1}{2}+\epsilon})$ for any $\gamma, \beta \geq 0$ and $\epsilon \in (0, 1/2)$ gives*

$$\min_{1 \leq i \leq t} \|G_{\eta_t/2}(x_i, x_i^+)\|^2 = o(t^{2\epsilon}) \text{ a.s.} \quad (47)$$

Proof Since $t \geq 1$ in the stepsize condition $\eta_t = \beta/(1 + \gamma\beta t^{\frac{1}{2}+\epsilon})$, we arrive at

$$1 + \gamma\beta t^{\frac{1}{2}+\epsilon} \leq t^{\frac{1}{2}+\epsilon} + \gamma\beta t^{\frac{1}{2}+\epsilon} = (1 + \gamma\beta)t^{\frac{1}{2}+\epsilon}, \quad (48)$$

which implies that $\eta_t \geq \beta/((1 + \gamma\beta)t^{\frac{1}{2}+\epsilon})$. Upon by the integral test inequality, we have

$$\begin{aligned} \sum_{j=1}^t \eta_j &\stackrel{(48)}{\geq} \sum_{j=1}^t \frac{\beta}{(1 + \gamma\beta)j^{\frac{1}{2}+\epsilon}} \geq \int_1^t \frac{\beta}{(1 + \gamma\beta)x^{\frac{1}{2}+\epsilon}} dx \\ &= \frac{\beta(t^{\frac{1}{2}-\epsilon} - 1)}{(1 + \gamma\beta)(\frac{1}{2} - \epsilon)} \geq \frac{\beta t^{-\frac{1}{2}-\epsilon}(t - 1)}{1 + \gamma\beta}, \end{aligned} \quad (49)$$

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

$$h(y) \leq h(x) + h'(x)(y - x). \quad (50)$$

In other words, by taking $y = 1$ and $x = t$, we can get $t^{\frac{1}{2}-\epsilon} - 1 \geq (\frac{1}{2} - \epsilon)t^{-\frac{1}{2}-\epsilon}(t - 1)$. Next, combining (48) and (49), we can obtain

$$\left(\sum_{j=1}^t \eta_j\right) \eta_t \stackrel{(48),(49)}{\geq} \frac{\beta t^{-\frac{1}{2}-\epsilon}(t - 1)}{1 + \gamma\beta} \frac{\beta}{(1 + \gamma\beta)t^{\frac{1}{2}+\epsilon}} = \frac{\beta^2 t^{-1-2\epsilon}(t - 1)}{(1 + \gamma\beta)^2}. \quad (51)$$

Upon setting $G_i = \|G_{\eta_t/2}(x_i, x_i^+)\|^2$, and applying the inequality $\min_{1 \leq i \leq t} \eta_t G_i \geq \eta_t \min_{1 \leq i \leq t} G_i$, we have

$$\left(\sum_{j=1}^t \eta_j\right) \min_{1 \leq i \leq t} \eta_t G_i \geq \left(\sum_{j=1}^t \eta_j\right) \eta_t \min_{1 \leq i \leq t} G_i \stackrel{(51)}{\geq} \frac{\beta^2 t^{-1-2\epsilon}(t - 1)}{(1 + \gamma\beta)^2} \min_{1 \leq i \leq t} G_i \geq 0. \quad (52)$$

For the left hand side of (52), we apply Theorem 1 to get

$$\lim_{t \rightarrow \infty} \left(\sum_{j=1}^t \eta_j\right) \min_{1 \leq i \leq t} \eta_t G_i = 0 \text{ a.s.} \quad (53)$$

The application of Squeeze theorem in conjunction with (52) gives

$$\lim_{t \rightarrow \infty} t^{-1-2\epsilon}(t - 1) \min_{1 \leq i \leq t} G_i = 0 \text{ a.s.} \quad (54)$$

which implies that $\min_{1 \leq i \leq t} G_i = o(t^{1+2\epsilon}/(t-1))$. Since² $t^{1+2\epsilon}/(t-1) \sim t^{2\epsilon}$, $(t \rightarrow \infty)$, we then have $o(t^{1+2\epsilon}/(t-1)) = o(t^{2\epsilon})$. Therefore, we finally obtain

$$\min_{1 \leq i \leq t} \|G_{\eta_i/2}(x_i, x_i^+)\|^2 = \min_{1 \leq i \leq t} G_i = o(t^{2\epsilon}) \text{ a.s.} \quad (55)$$

This completes the proof. \square

Remark 3 It should be noted that the boundedness of the squared generalized projected gradient norm can be obtained only in expectation, by an additional randomized step [12]. In contrast, we achieve the almost sure convergence of SCOMID without the bounded variance assumption (i.e., $\alpha_t = 0$ in Assumption 1(c)) and the unbiasedness of the stochastic gradient g_t . In addition, by choosing $\epsilon \rightarrow 0$ in the stepsize η_t , we can conclude that $o(t^{2\epsilon}) \rightarrow o(1)$. Then, Corollary 1 indicates that $\min_{1 \leq i \leq t} \|G_{\eta_i/2}(x_i, x_i^+)\|^2$ is arbitrarily close to zero with probability one.

Theorem 2 Suppose that Assumption 1 holds and SCOMID (see (2)) is employed with a stepsize η_t that satisfies $\sum_{t=1}^{\infty} \eta_t < \infty$, $\sum_{t=1}^{\infty} \eta_t \alpha_t < \infty$ and $\eta_t \leq \mu/(8L)$. Then, there exists a constant $\bar{\Phi} < \infty$ such that

$$\lim_{t \rightarrow \infty} \Phi(x_t) = \bar{\Phi} \text{ a.s.} \quad (56)$$

Proof By the L -smoothness of $F(x)$, the inequality (41) in the proof of Theorem 1 still holds. i.e.,

$$\begin{aligned} \mathbb{E}[\Phi(x_{t+1}) - \Phi^* | \mathcal{F}_t] &\leq \left(1 + \frac{2\eta_t \alpha_t}{\mu}\right) \left(\Phi(x_t) - \Phi^*\right) \\ &\quad - \frac{\mu \eta_t}{32} \|G_{\eta_t/2}(x_t, x_t^+)\|^2 + \frac{2\sigma^2 \eta_t}{\mu}. \end{aligned} \quad (57)$$

In view of the conditions $\sum_{t=1}^{\infty} \eta_t < \infty$ and $\sum_{t=1}^{\infty} \eta_t \alpha_t < \infty$, upon applying Lemma 2 we conclude that $\Phi(x_t) - \Phi^*$ converges almost surely. Thus, there exists a constant $\bar{\Phi} < \infty$ such that $\lim_{t \rightarrow \infty} \Phi(x_t) = \bar{\Phi}$ a.s.. \square

Remark 4 The almost sure convergence of function values $\Phi(x_t) - \Phi^*$ for SCOMID is obtained in Theorem 2. Nevertheless, by additionally assuming the G -Lipschitz continuity of $F(x)$ and the unbiasedness of g_t , the almost sure convergence of function values for SGD has been achieved in [25]. It should be noted that G -Lipschitz continuity implies the boundedness of the gradient norm $\|\nabla F(x_t)\|$, which takes the convergence of subsequence $\|\nabla F(x_{t,k})\|$ a priori.

² $f(x) \sim g(x)$: there exist x_0 , such that $\lim_{x \rightarrow x_0} f(x)/g(x) = 1$.

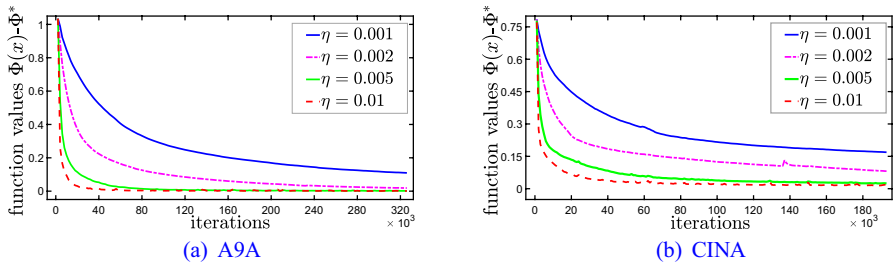


Fig. 1 Performance of the SCOMID algorithm with different fixed stepsizes for solving regularized logistic regression problem on A9A and CINA

4 Numerical experiments

The SCOMID algorithm (2) was tested on solving the regularized logistic regression problem. More specifically, we constructed a new non-convex non-smooth regularizer $R(x) = \lambda \sum_{j=1}^d |x_j|/(1 + |x_j|)$ based on the non-convex smooth regularizer $R(x) = \lambda \sum_{j=1}^d x_j^2/(1 + x_j^2)$ adopted in [39]. Then, the logistic regression problem can be formulated as

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i \cdot a_i^T x)) + \lambda \sum_{j=1}^d \frac{|x_j|}{1 + |x_j|}, \quad (58)$$

where $\{a_i, b_i\}_{i=1}^N$ is the training set with $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$, $\lambda > 0$ denotes the regularization factor, and x_j is the j -th component of x . Similarly to [39], we used two relatively binary classification datasets A9A ($N = 16,281, d = 123$) and CINA ($N = 3,206, d = 133$) from LIBSVM³.

To illustrate the performance of the SCOMID algorithm with different stepsizes, we followed [7, 10, 12] to choose the distance generating function $\psi_t(x) = \|x\|^2/2$ in (2). In addition, owing to the variability in size of different datasets, we set the total number of iterations T of each experiment to depend on the data volume N , i.e., $T = 20 \times N$ for A9A dataset and $T = 60 \times N$ for CINA dataset. For better comparison, we also run SCOMID with the same regularization factor $\lambda = 10^{-2}/N$ and the same random initial point in each experiment. We examined the following three stepsize schemes.

(a) Comparison of different fixed stepsizes: Our first experiment aimed to compare the behaviour of SCOMID using different fixed stepsizes η . We considered four stepsize choices: $\eta \in \{0.001, 0.002, 0.005, 0.01\}$. The results are provided in Fig. 1, and show that the function value sequences of SCOMID with fixed stepsizes mentioned above are all convergent. For the considered datasets and stepsizes, overall $\eta = 0.01$ provides the best performance.

³ LIBSVM website: <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

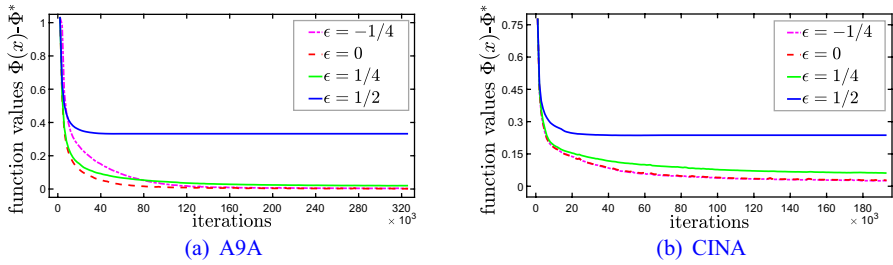


Fig. 2 Performance of the SCOMID algorithm with different diminishing stepsizes for solving regularized logistic regression problem on A9A and CINA

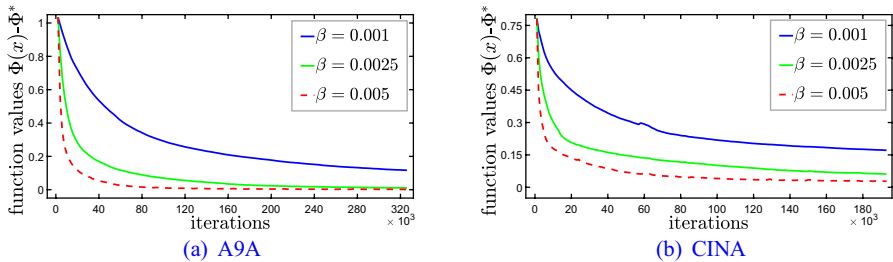


Fig. 3 Performance of the SCOMID algorithm with different parameters to diminishing stepsizes for solving regularized logistic regression problem on A9A and CINA

(b) Comparison of different diminishing stepsizes: Next, we tested the SCOMID algorithm with diminishing stepsizes on two different datasets. According to Corollary 1 in Sect. 3, we considered the diminishing stepsizes of the form $\eta_t = \beta / (1 + \gamma \beta t^{\frac{1}{2} + \epsilon})$ with $\gamma = 0.05$. For a fair comparison, we fixed the parameter $\beta = 0.005$ in the experiment, then select ϵ from the set $\{-1/4, 0, 1/4, 1/2\}$. Figure 2 shows the convergence behavior of SCOMID with different diminishing stepsizes mentioned above. We observe from Fig. 2 that the loss decays with an increasingly fast rate as ϵ gradually decreases to zero (i.e., η_t increases to $\beta / (1 + \gamma \beta t^{\frac{1}{2}})$). However, once ϵ goes down below zero (e.g., $\epsilon = -1/4$), the performance of SCOMID will not improve further, which indicates that $\epsilon = 0$ is optimal among all considered values. This is consistent with our theoretical analysis in Corollary 1 and Remark 3, which shows that the almost sure convergence result improves by choosing ϵ close to zero.

(c) Comparison of the diminishing stepsize with different parameters: Further, we tested the SCOMID algorithm using the optimal diminishing stepsize $\eta_t = \beta / (1 + \gamma \beta t^{\frac{1}{2}})$, suggested by Corollary 1. Then, we fixed $\gamma = 0.05$ as (b) and tuned the parameter β in the set $\{0.001, 0.0025, 0.005\}$. The performance is provided in Fig. 3 and shows that the larger parameter $\beta = 0.005$ performs much better than the other two settings.

5 Conclusions

We have provided the almost sure convergence analysis of SCOMID with biased gradient estimation in the non-convex non-smooth setting. Under a slightly weaker Assumption 1(c) and only L -smoothness of $F(x)$, we have proved that the minimum of the squared generalized projected gradient norm is arbitrarily close to zero with probability one. Furthermore, the almost sure convergence of function values has been established by selecting an appropriate sequence of time-varying stepsizes. Finally, we have verified the theoretical findings through several numerical experiments. In view of the general distance function B_{ψ_i} used in the iterate (2), our theoretical analysis of almost sure convergence is quite general and can be extended to a wide range of stochastic optimization algorithms.

Acknowledgements The authors wish to thank the anonymous reviewers for their insightful and very helpful expert comments and suggestions. This work was funded in part by National Key R & D Program of China (No. 2021YFA1003400), in part by the National Natural Science Foundation of China (No. 62176051), and in part by the Fundamental Research Funds for the Central Universities of China (No. 2412020FZ024).

Data Availability Statement Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Alacaoglu, A., Malitsky, Y., Cevher, V.: Convergence of adaptive algorithms for weakly convex constrained optimization. arXiv preprint [arXiv:2006.06650](https://arxiv.org/abs/2006.06650) (2020)
- Atchadé, Y.F., Fort, G., Moulines, E.: On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.* **18**(1), 310–342 (2017)
- Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. *SIAM Rev.* **60**(2), 223–311 (2018)
- Cevher, V., Vũ, B.C.: On the linear convergence of the stochastic gradient method with constant step-size. *Optim. Lett.* **13**(5), 1177–1187 (2019)
- Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* **29**(1), 207–239 (2019)
- Driggs, D., Liang, J., Schönlieb, C.B.: On biased stochastic gradient estimation. *J. Mach. Learn. Res.* **23**, 24–1 (2022)
- Drusvyatskiy, D., Paquette, C.: Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.* **178**(1), 503–558 (2019)
- Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**(7), 2121–2159 (2011)
- Duchi, J., Shalev-Shwartz, S., Singer, Y., Tewari, A.: Composite objective mirror descent. In: *Conference on Learning Theory*, vol. 10, pp. 14–26 (2010)
- Duchi, J., Singer, Y.: Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.* **10**, 2899–2934 (2009)
- Dundar, M., Krishnapuram, B., Bi, J., Rao, R.B.: Learning classifiers when the training data is not iid. In: *IJCAI*, vol. 2007, pp. 756–61 (2007)

12. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.* **155**(1), 267–305 (2016)
13. Gorbunov, E., Hanzely, F., Richtárik, P.: A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In: *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR (2020)
14. Gower, R., Sebbouh, O., Loizou, N.: SGD for structured nonconvex functions: Learning rates, mini-batching and interpolation. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR (2021)
15. Han, Y., Feng, X., Baciú, G., Wang, W.: Nonconvex sparse regularizer based speckle noise removal. *Pattern Recognit.* **46**(3), 989–1001 (2013)
16. J. Reddi, S., Sra, S., Póczos, B., Smola, A.J.: Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In: *Advances in Neural Information Processing Systems*, vol. 29, pp. 1153–1161 (2016)
17. Khaled, A., Richtárik, P.: Better theory for SGD in the nonconvex world. arXiv preprint [arXiv:2002.03329](https://arxiv.org/abs/2002.03329) (2020)
18. Li, Z., Li, J.: Simple and optimal stochastic gradient methods for nonsmooth nonconvex optimization. *J. Mach. Learn. Res.* **23**(239), 1–61 (2022)
19. Liang, J., Monteiro, R.D.: An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems. *SIAM J. Optim.* **31**(1), 217–243 (2021)
20. Liu, J., Kong, J., Xu, D., Qi, M., Lu, Y.: Convergence analysis of AdaBound with relaxed bound functions for non-convex optimization. *Neural Netw.* **145**, 300–307 (2022)
21. Liu, J., Yuan, Y.: On almost sure convergence rates of stochastic gradient methods. arXiv preprint [arXiv:2202.04295](https://arxiv.org/abs/2202.04295) (2022)
22. Luo, J., Liu, J., Xu, D., Zhang, H.: SGD- α : A real-time α -suffix averaging method for SGD with biased gradient estimates. *Neurocomputing* **487**, 1–8 (2022)
23. Mai, V., Johansson, M.: Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In: *International Conference on Machine Learning*, pp. 6630–6639. PMLR (2020)
24. Mandic, D., Chambers, J.: *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. Wiley, New York (2001)
25. Mertikopoulos, P., Hallak, N., Kavis, A., Cevher, V.: On the almost sure convergence of stochastic gradient descent in non-convex problems. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 1117–1128 (2020)
26. Metel, M.R., Takeda, A.: Stochastic proximal methods for non-smooth non-convex constrained sparse optimization. *J. Mach. Learn. Res.* **22**, 115–1 (2021)
27. Nesterov, Y.: *Lectures on Convex Optimization*, vol. 137. Springer, Cham (2018)
28. Nikolova, M., Ng, M.K., Tam, C.P.: Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Trans. Image Process.* **19**(12), 3073–3088 (2010)
29. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* pp. 400–407 (1951)
30. Robbins, H., Siegmund, D.: A convergence theorem for non-negative almost supermartingales and some applications. In: *Optimizing methods in statistics*, pp. 233–257. Elsevier (1971)
31. Shalev-Shwartz, S., Ben-David, S.: *Understanding Machine Learning: from Theory to Algorithms*. Cambridge University Press, New York (2014)
32. Sun, R.Y.: Optimization for deep learning: An overview. *J. Oper. Res. Soc. China* **8**(2), 249–294 (2020)
33. Tao, W., Pan, Z., Wu, G., Tao, Q.: Primal averaging: A new gradient evaluation step to attain the optimal individual convergence. *IEEE T. Cybern.* **50**(2), 835–845 (2018)
34. Vial, J.P.: Strong and weak convexity of sets and functions. *Math. Oper. Res.* **8**(2), 231–259 (1983)
35. Ward, R., Wu, X., Bottou, L.: AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *J. Mach. Learn. Res.* **21**, 1–30 (2020)
36. Wood, K., Bianchin, G., Dall’Anese, E.: Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Syst. Lett.* **6**, 1646–1651 (2022)
37. Zhang, H., Pan, L., Xiu, N.: Optimality conditions for locally Lipschitz optimization with l_0 -regularization. *Optim. Lett.* **15**(1), 189–203 (2021)
38. Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., Gu, Q.: On the convergence of adaptive gradient methods for nonconvex optimization. arXiv preprint [arXiv:1808.05671](https://arxiv.org/abs/1808.05671) (2018)
39. Zhou, Y., Wang, Z., Ji, K., Liang, Y., Tarokh, V.: Proximal gradient algorithm with momentum and flexible parameter restart for nonconvex optimization. arXiv preprint [arXiv:2002.11582](https://arxiv.org/abs/2002.11582) (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.