J. Nonlinear Var. Anal. 9 (2025), No. 4, pp. 569-598
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.9.2025.4.07

CONVERGENCE ANALYSIS OF A PROXIMAL STOCHASTIC GRADIENT ALGORITHM WITH ADAPTIVE SAMPLING FOR NON-CONVEX AND NON-SMOOTH COMPOSITE OPTIMIZATION PROBLEMS

MENGXIANG ZHANG, SHENGJIE LI*

College of Mathematics and Statistics, Chongqing University, Chongqing, China

Abstract. This paper examines the convergence and computational complexity of a proximal stochastic gradient algorithm that adaptively incorporates sampling techniques for solving large-scale, non-convex, and non-smooth problems, with a particular emphasis on problems that involve the combination of two non-convex functions. This an area that has been scarcely explored by current methods. By adjusting adaptively the sampling size (or mini-batch size) throughout the algorithm's iterations, this method aims to balance the trade-off between stochastic gradient noise and convergence stability. It maintains a convergence rate similar to that of the proximal gradient method. Moreover, when the objective function is a Kurdyka-Łojasiewicz (KL) function, we demonstrate the convergence rate of the expected function value on a case-by-case basis, achieving linear convergence under optimal conditions. Finally, some preliminary numerical results validate the effectiveness and robustness of the proposed method.

Keywords. Adaptive sampling; Kurdyka-Łojasiewicz property; Large-scale; Non-convex and non-smooth; Proximal stochastic gradient.

1. INTRODUCTION

In this paper, our goal is to minimize the composite objective functions presented in the finite-sum form:

$$\min_{x\in\mathbb{R}^d} F(x) := f(x) + h(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) + h(x). \tag{1.1}$$

Assumption 1.1. The functions in equation (1.1) satisfy the following assumptions:

(A.1) $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is continuously differentiable, with an L-Lipschitz continuous gradient, and may be non-convex.

(A.2) $h : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper, lower semicontinuous function, bounded from below, and may be non-convex and non-smooth.

(A.3) F is bounded from below.

Problem (1.1), which is extensively utilized in the fields of statistics [1] and machine learning [2], is also known as regularized empirical risk minimization. In general, $f_i(x) := f_i(x, \xi_i)$, where it is assumed that the training set $\{\xi_i = (a_i, b_i)\}_{i=1}^N$ is a collection of independent and

*Corresponding author.
E-mail address: lisj@cqu.edu.cn (S. Li).
Received 9 January 2025; Accepted 15 February 2025; Published online 15 May 2025.

identically distributed (i.i.d.) samples. Here, a_i and b_i represent the feature vector and class label of the i-th example, respectively, and N denotes the size of the dataset, which is usually quite large. Function h serves as an additional non-smooth regularization term, strategically introduced to mitigate the risk of overfitting or to encourage specific structures, such as sparsity or non-negativity, in the variable x. Over the past few years, there has been a surge in interest towards non-convex regularization techniques, including the l_p norm ($|x|_p$, with $0 \leq p < 1$), smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP). For recent improvements over convex regularizers, we refer to [3–7] and the references therein.

When N is substantially large, the computation of the exact (or full) gradient becomes prohibitively expensive and may even be infeasible. Consequently, there is an increasing focus on approximating the exact gradient of f through sampling methods, drawing on the principles of Stochastic Approximation (SA) method [8]. This methodology entails the random selection of a sample set of indices $\mathcal{N}^{(k)}$, with a sampling size $N^{(k)} \in \{1,2,\cdots,N\}$, to form the stochastic gradient

$$\nabla f^{(k)}_{\mathcal{N}^{(k)}} := \frac{1}{N^{(k)}} \sum_{i \in \mathcal{N}^{(k)}} \nabla f_i \left(x^{(k)} \right).$$

In the field of high-dimensional optimization, the computational cost is primarily determined by the total number of gradient evaluations, and the convergence speed of stochastic algorithms is significantly affected by sampling error. As demonstrated by [9, 10], smaller sampling sizes are sufficient for optimal progress in the early stages of training, while larger sampling sizes are required to maintain performance in the later stages. Therefore, determining the appropriate sampling size has become a critical challenge. To address this challenge, we propose a new sampling strategy that dynamically adjusts the balance between sampling error and sampling size $N^{(k)}$ during the optimization process. Specifically, at each iteration k, we generate a sampling set $\mathcal{N}^{(k)}$ based on the gradient information of the current point $x^{(k)}$, and use this sampling set to calculate the next update point $x^{(k+1)}$. This method provides a robust and practical approach that effectively addresses a wide range of stochastic optimization challenges.

Literature review and and motivation. Over the past decade or so, the field of finite sum optimization with $h = 0$ (1.1) has seen under the research spotlight; see, e.g., [11–16]. Notably, the norm test proposed by Byrd et al. [11] and the inner product test proposed by Bollapragada et al. [12] are two important developments in this area. For composite functions with convex f and convex h, Xie et al. [17] studied the convergence of the inner product test strategy. Recently, many scholars investigated the case where f is non-convex and h is convex. Beiser et al. [18], building upon the norm test, conducted a more in-depth study of the projection stochastic gradient method and the sequential quadratic programming method. Franchini et al. [19] delved into the use of stochastic line search and adaptive sampling within the context of proximal stochastic gradient method. Building on this, Zhang and Li [20] integrated stochastic line search and adaptive sampling into the proximal stochastic quasi-Newton method, and eliminating the need for repeated evaluations of the proximal operator during the stochastic line search process.

However, the existing literature lacks the analysis on the convergence and complexity of adaptive sampling strategies for non-convex f and non-convex h. Current methods [21–26] for solving fully non-convex composite optimization problems (1.1) are primarily based on variance reduction techniques [27], such as SAG [28, 29], SAGA [30, 31], SVRG [31–33], and SARAH [34, 35]. Although variance reduction techniques can effectively dampen the effects

of gradient noise and achieve convergence rates s imilar to those of deterministic methods, they generally either perform full gradient computations periodically, as in SVRG and SARAH, or require the storage of partial gradients, as with SAG and SAGA. Additionally, during the iteration process, these methods also need to consider the sampling size, which significantly affects their performance. Therefore, this motivates us to further explore how adaptive sampling strategies perform in fully non-convex problems.

TABLE 1. Comparison of Prox-ASSG (proposed method) with state-of-the-art algorithms. "Complexity (non-convex)" denotes the expected complexity to achieve an ε-approximate critical point (1.3), "Complexity (KL exponent)" represents the expected complexity to attain an ε-stationary point, and "-" indicates that the result is not provided. κ is a positive constant that can be close to 0.

Algorithm	Full gradient	Historical gradients	Complexity (non-convex)	Complexity (KL exponent)
MBSGA (SG) [37]	No	No	$\mathscr{O}(1/\varepsilon^5)$	-
VRSGA (SVRG) [37]	Yes	No	$\mathscr{O}(n^{2/3}/\varepsilon^3)$	-
SPRING (SAGA) [22]	No	Yes	$\mathscr{O}(n^{2/3}/\varepsilon^2)$	-
SPRING (SARAH) [22]	Yes	Yes	$\mathscr{O}(n^{1/2}/\varepsilon^2)$	-
Prox-ASSG (This paper)	No	No	$\mathscr{O}(1/\varepsilon^{4+\kappa})$	$\mathscr{O}(1/\varepsilon)$, $\theta \in \left[\frac{1}{2},1\right)$

Contributions. In this work, we present Prox-ASSG, a proximal stochastic gradient algorithm that utilizes adaptive sampling to tackle composite finite-sum functions with non-convex and non-smooth regularization terms. The key contributions of our work are as follows:

- We introduce a progressive batching strategy for proximal stochastic gradient algorithms, designed to adaptively increase the sampling size. This strategy not only prevents the sampling size from increasing too rapidly but also effectively manages the variance of stochastic gradients in the direction of the smooth component f, reducing the impact of gradient noise, and thereby restoring the convergence rate to that of the non-stochastic version.
- In the case where f is non-convex and h is non-convex and non-smooth, the algorithm generates iterative points that almost surely (a.s.) converge to a critical point $x^\star$. The rate of convergence is given by

$$\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(R)}\right)\right)\right] = \mathscr{O}\left(\frac{1}{\sqrt{K}}\right), \tag{1.2}$$

  where R is uniformly chosen from the set $\{0,1,\cdots,K\}$ and $\operatorname{dist}(x,\Omega) := \min_{y\in\Omega}\|y-x\|$. To obtain an ε-approximate critical point in expectation, i.e.,

$$\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(R)}\right)\right)\right] \leq \varepsilon, \tag{1.3}$$

  the total number of gradient evaluations (computational complexity) is $\mathscr{O}\left(\frac{1}{\varepsilon^{4+\kappa}}\right)$, where κ is a positive constant that can be close to 0.
- Additionally, if problem (1.1) is a KL function with exponent $\theta \in [0,1)$, we demonstrate that the sequence $\left\{F\left(x^{(k)}\right)\right\}_{k\geq 0}$ generated by Prox-ASSG converges in expectation to $F(x^\star)$ of objective function at the following rates:
  - If $\theta = 1$, then $\mathbb{E}\left[F\left(x^{(k)}\right) - F\left(x^\star\right)\right] = 0$ in a finite number of steps;
  - If $\theta \in \left[\frac{1}{2},1\right)$, then $\mathbb{E}\left[F\left(x^{(k)}\right) - F\left(x^\star\right)\right] \leq \mathscr{O}\left(\bar{\rho}^k\right)$, $\bar{\rho} \in (0,1)$;

· If $\theta \in \left(0, \frac{1}{2}\right)$, then $\mathbb{E}\left[F\left(x^{(k)}\right) - F\left(x^{\star}\right)\right] \leq \mathscr{O}\left(k^{-\frac{1}{1-2\theta}}\right)$.

2. Notations and preliminaries

2.1. **Notations.** The subsequent sections of this paper employ the following notations. Without any specification, we use $\|\cdot\|$ to denote the Euclidean norm ($\|\cdot\|_2$ norm). We consistently use the notation $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product. The ceiling function is represented by $\lceil\cdot\rceil$, and the floor function by $\lfloor\cdot\rfloor$.

To simplify our notation, we adopt the following abbreviations: $f\left(x^{(k)}\right)$ as $f^{(k)}$, $h\left(x^{(k)}\right)$ as $h^{(k)}$, $F\left(x^{(k)}\right)$ as $F^{(k)}$, exact gradient $\nabla f\left(x^{(k)}\right)$ as $\nabla f^{(k)}$, individual $\nabla f_i\left(x^{(k)}\right)$ as $\nabla f_i^{(k)}$, stochastic gradient $\nabla f_{\mathscr{N}^{(k)}}\left(x^{(k)}\right)$ as $\nabla f^{(k)}_{\mathscr{N}^{(k)}}$, subdifferential $\partial h\left(x^{(k)}\right)$ as $\partial h^{(k)}$, subdifferential $\partial F\left(x^{(k)}\right)$ as $\partial F^{(k)}$, $F\left(x^{\star}\right) = f\left(x^{\star}\right) + h\left(x^{\star}\right)$ as $F^{\star} = f^{\star} + h^{\star}$ and $\underline{F} := \inf_x F(x)$ and conditional expectation $\mathbb{E}\left[\cdot|\mathscr{F}^{(k)}\right]$ as $\mathbb{E}_k[\cdot]$.

2.2. **Preliminaries.** We first outline some common assumptions in stochastic methods.

Assumption 2.1. Let $\mathscr{F}^{(k)}$ denote the σ-algebra generated by the sequence $\left\{x^{(0)}, x^{(1)}, \ldots, x^{(k)}\right\}$. For each iteration k, the following assumptions hold

(A.4) $\mathbb{E}_k\left[\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right] = \nabla f^{(k)}$;

(A.5) $\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right] \leq \delta^2 + N^{(k)}\beta^2\mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]$, where δ and β are positive constants.

We would like to underscore two key points for emphasis:

- It is straightforward to confirm that, in problem (1.1), the conditional expectation meets the conditions of Assumption 2.1 (A.4).
- In the literature on stochastic algorithms that utilize mini-batch or adaptive sampling, such as [11, 12, 17, 19, 36], it is often necessary to assume

$$\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right] \leq \delta^2$$

or

$$\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right] \leq \beta^2\mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right].$$

Relative to these algorithms, our assumptions are more lenient. This leniency stems from the design of our algorithm, where the sampling size $N^{(k)}$ escalates with each iteration, and the Assumption 2.1 (A.5) becomes progressively weaker as the iterations advance.

Definition 2.1. (Proximal mapping [38]) The proximal map of a point $x \in \mathbb{R}^d$ under a proper and closed function h with parameter $\alpha > 0$ is defined as:

$$\operatorname{prox}_{\alpha h}(x) := \operatorname{argmin}_z h(z) + \frac{1}{2\alpha}\|z - x\|^2, \tag{2.1}$$

where $\alpha > 0$.

In the specific iteration process, the proximal mapping is used as follows:

$$\begin{aligned} x^{(k+1)} &\in \operatorname{prox}_{\alpha h}\left(x^{(k)}-\alpha\nabla f^{(k)}_{\mathcal{N}(k)}\right) \\ &= \operatorname{argmin}_z\left\langle \nabla f^{(k)}_{\mathcal{N}(k)}, z-x^{(k)}\right\rangle + \frac{1}{2\alpha}\left\|z-x^{(k)}\right\|^2 + h(z), \end{aligned} \tag{2.2}$$

where step size $\alpha > 0$. It is crucial to recognize that if h is a convex function, its associated proximal mapping corresponds to the minimizer of a strongly convex function, resulting in a unique solution. In contrast, when h is non-convex, the proximal mapping may not yield a single solution but rather a set of solutions. In these instances, $\operatorname{prox}_{\alpha h}(x)$ represents an arbitrary element from this set of solutions, which complicates the analytical process. Consequently, it becomes necessary to establish the following definition to navigate these complexities.

Definition 2.2. (Subdifferential [38]) The Frech*é*t subdifferential $\hat{\partial}F(x)$ of $x \in \operatorname{dom} F$ is the set of $u \in \mathbb{R}^d$ such that

$$\liminf_{z \neq x, z \to x} \frac{F(z)-F(x)-u^\top(z-x)}{\|z-x\|} \geq 0,$$

while the (limiting) subdifferential $\partial F(x)$ at $x \in \operatorname{dom} F$ is the graphical closure of $\hat{\partial}F(x)$:

$$\partial F(x) = \left\{u : \exists \left(x^{(k)}, F\left(x^{(k)}\right)\right) \to (x, F(x)), \hat{\partial}F\left(x^{(k)}\right) \ni u^{(k)} \to u\right\}.$$

Specifically, this generalized derivative simplifies to the gradient ∇F when F is continuously differentiable, and it corresponds to the standard subdifferential when F is convex. From the optimality condition of equation (2.2), it follows that

$$0 \in \nabla f^{(k)}_{\mathcal{N}(k)} + \frac{1}{\alpha}(x^{(k+1)} - x^{(k)}) + \partial h(x^{(k+1)}).$$

Definition 2.3. (Critical point) A point $x \in \mathbb{R}^d$ is a critical point of F if and only if (iff) $0 \in \partial F(x)$.

Definition 2.4. (Distance) The distance of a point $x \in \mathbb{R}^d$ to a closed set $\Omega \subseteq \mathbb{R}^d$ is defined as:

$$\operatorname{dist}(x, \Omega) := \min_{y \in \Omega} \|y - x\|. \tag{2.3}$$

Definition 2.5. (KL property [39]) Let $F\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper and lower semicontinuous function. A function $F : \mathbb{R}^d \to (-\infty, +\infty]$ is said to have the KL property at $\bar{x} \in \operatorname{dom}(\partial F)$ if there exist $\varepsilon \in (0, +\infty]$ a neighborhood O of $\bar{x}$ and a continuous concave function $\varphi : [0, \varepsilon) \to \mathbb{R}^+$ such that, for all

$$x \in O \cap \left\{x \in \mathbb{R}^d : F(\bar{x}) < F(x) < F(\bar{x}) + \varepsilon\right\},$$

the following inequality holds

$$\varphi'(F(x) - F(\bar{x}))\operatorname{dist}(0, \partial F(x)) \geq 1, \tag{2.4}$$

where the function $\varphi : [0, \varepsilon) \to [0, +\infty)$ satisfying: i. φ is concave and continuous on $(0, \lambda)$; ii. φ is continuous at 0, $\varphi(0) = 0$; and iii. $\varphi(x) > 0$, $\forall x \in (0, \varepsilon)$.

Definition 2.6. (KL function [39]) A function F is said to be a KL function if it satisfies the KL property at each point of $\operatorname{dom}(\partial F)$.

To put it simply, KL functions become sharp up to reparameterization via φ, a desingularizing function for F. All semi-algebraic and subanalytic functions are the KL function. Notably, for semi-algebraic functions, the desingularizing function $\varphi(x)$ can be selected in the form $\varphi(x) = \frac{c}{\theta}x^{\theta}$, where c is a positive constant and θ is the KL exponent within the interval $(0,1]$. Common examples of semi-algebraic functions encompass real polynomial functions, l_p norms for $p \geq 0$, rational functions, real polynomials, rank functions, and so on, as detailed in [39–41].

3. ALGORITHM

We describe Prox-ASSG in Algorithm 1 below.

Algorithm 1 Prox-ASSG

1: **Input:** Give $x^{(0)} \in \mathbb{R}^d$, $2 \leq N^{(0)} < N$, $N^{(0)} \in \mathbb{N}^+$, α and β are positive constants, and a non-negative sequence $\{\varepsilon^{(k)}\}_{k\geq 0}$ with $\sum_{k=0}^{+\infty}\varepsilon^{(k)} < +\infty$.

2: **for** $k = 0, 1, 2, \ldots K$ **do**

3: Select a sample set $\mathcal{N}^{(k)}$ of size $N^{(k)}$, compute stochastic gradient

$$\nabla f_{\mathcal{N}^{(k)}}^{(k)} := \frac{1}{N^{(k)}} \sum_{i\in\mathcal{N}^{(k)}} \nabla f_i\left(x^{(k)}\right),$$

$x^+ \in \operatorname{prox}_{\alpha h}\left(x^{(k)} - \alpha\nabla f_{\mathcal{N}^{(k)}}^{(k)}\right)$, and sampling gradient variance

$$S^{(k)} := \frac{\sum_{i\in\mathcal{N}^{(k)}} \left\|\nabla f_i^{(k)} - \nabla f_{\mathcal{N}^{(k)}}^{(k)}\right\|^2}{N^{(k)} - 1}.$$

4: **if** $S^{(k)} \cdot \frac{N - N^{(k)}}{N^{(k)}(N-1)} \leq \Delta^{(k)} := \varepsilon^{(k)} + \beta^2\left\|x^+ - x^{(k)}\right\|^2$, **then** the sample set $N^{(k)}$ is considered reasonably valid.

5: **else** $N^{(k)} = \min\left\{N, \max\left\{\lceil\frac{S^{(k)}}{\Delta^{(k)}} \cdot \frac{N}{N - 1 + \frac{S^{(k)}}{\Delta^{(k)}}}\rceil, N^{(k)} + 1\right\}\right\}$. Select a new sample set $\mathcal{N}^{(k)}$ of size $N^{(k)}$, and compute $\nabla f_{\mathcal{N}^{(k)}}^{(k)}$ and x^+.

6: **end if**

7: $x^{(k+1)} = x^+$.

8: **end for**

3.1. **Choice of the sampling size $N^{(k)}$.** Following [42, p.240] (Appendix 7.1), for an arbitrary $i \in \mathcal{N}^{(k)}$, we have that

$$\mathbb{E}_k\left[\left\|\nabla f_{\mathcal{N}^{(k)}}^{(k)} - \nabla f^{(k)}\right\|^2\right] = \frac{\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{N^{(k)}} \cdot \frac{N - N^{(k)}}{N - 1}. \tag{3.1}$$

Consequently, it is evident that, by regulating the sampling size $N^{(k)}$, we can manage the variance bound of the stochastic gradient. Therefore, the inequality

$$\mathbb{E}_k\left[\left\|\nabla f_{\mathcal{N}^{(k)}}^{(k)} - \nabla f^{(k)}\right\|^2\right] \leq \varepsilon^{(k)} + \beta^2\mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] \tag{3.2}$$

holds if $N^{(k)}$ satisfies

$$N^{(k)} \geq \frac{\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{\varepsilon^{(k)} + \beta^2 \mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]} \cdot \frac{N}{N - 1 + \frac{\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{\varepsilon^{(k)} + \beta^2 \mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]}} \tag{3.3}$$

with $\varepsilon^{(k)}$ being a positive value.

In practice, we need to estimate both quantities on the right-hand side of (3.3). As was done, e.g., in [2, 11, 12], the population variance term in the numerator can be approximated by a sampling gradient variance

$$S^{(k)} := \frac{\sum_{i \in \mathcal{N}^{(k)}} \left\|\nabla f_i^{(k)} - \nabla f_{\mathcal{N}^{(k)}}^{(k)}\right\|^2}{N^{(k)} - 1}$$

to approximate $\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]$ in inequality (3.1). And it is desired that the obtained

$$x^+ \in \operatorname{prox}_{\alpha h}\left(x^{(k)} - \alpha \nabla f_{\mathcal{N}^{(k)}}^{(k)}\right)$$

under the current sampling condition satisfies the following inequality:

$$S^{(k)} \frac{N - N^{(k)}}{N^{(k)}(N-1)} \leq \Delta^{(k)} := \varepsilon^{(k)} + \beta^2 \left\|x^+ - x^{(k)}\right\|^2. \tag{3.4}$$

When condition (3.4) is not satisfied, the variance can be reduced by appropriately increasing the sampling size $N^{(k)}$. Considering equation (3.4), we can derive the following after simple manipulation:

$$N^{(k)} \geq \frac{S^{(k)}}{\Delta^{(k)}} \cdot \frac{N}{N - 1 + \frac{S^{(k)}}{\Delta^{(k)}}}.$$

Additionally, we can establish that $N^{(k)}$ must be less than or equal to N. Thus $N^{(k)}$ can be updated by the following rule:

$$N^{(k)} = \min\left\{N, \max\left\{\left\lceil \frac{S^{(k)}}{\Delta^{(k)}} \cdot \frac{N}{N - 1 + \frac{S^{(k)}}{\Delta^{(k)}}} \right\rceil, N^{(k)} + 1\right\}\right\}.$$

Remark 3.1. We would like to emphasize the following points:

i. From inequality (3.1) and Assumption 2.1 (A.5), we can derive that

$$\begin{aligned} \mathbb{E}_k\left[\left\|\nabla f_{\mathcal{N}^{(k)}}^{(k)} - \nabla f^{(k)}\right\|^2\right] &\leq \frac{\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{N^{(k)}} \\ &\leq \frac{\delta^2}{N^{(k)}} + \beta^2 \mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]. \end{aligned} \tag{3.5}$$

Therefore, when $N^{(k)} \geq \frac{\delta^2}{\varepsilon^{(k)}}$, inequality (3.2) holds. Thus the choice of $\varepsilon^{(k)}$ is closely related to the computational complexity. In this paper, we provide two selection rules

for $\varepsilon^{(k)}$: $\frac{C}{(k+1)^{1+\kappa}}$ and $\frac{C}{(1+\kappa)^k}$, where C and κ are two positive constants. The rationale behind these choices is detailed in Corollary 4.1 and Corollary 4.2.

ii. We slightly reformulate inequality (3.3) as follows:

$$N^{(k)} \geq \frac{N}{\frac{N-1}{\Pi}+1}, \tag{3.6}$$

where $\Pi = \frac{\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{\varepsilon^{(k)} + \beta^2 \mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]}$. It can be observed that a smaller value of Π results in a reduction of the right-hand side of inequality (3.6). This implies that the required $N^{(k)}$ decreases, effectively suppressing its growth. Furthermore, in conjunction with Assumption 2.1 (A.5), increasing the upper bound of $\mathbb{E}_k\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]$ is equivalent to increasing the denominator of Π, thereby reducing Π.

iii. We know that if a large number of samples are used, the sample mean is a random variable that approximately follows a normal distribution. In our case, this means

$$\nabla f^{(k)}_{\mathcal{N}^{(k)}} - \nabla f^{(k)} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N^{(k)}}\right),$$

where $\sigma^2 = \mathbb{E}\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]$. Under Assumption 2.1 (A.5), we know that

$$\frac{\sigma^2}{N^{(k)}} = \frac{\mathbb{E}\left[\left\|\nabla f_i^{(k)} - \nabla f^{(k)}\right\|^2\right]}{N^{(k)}} \leq \frac{\delta^2}{N^{(k)}} + \beta^2 \mathbb{E}\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right].$$

In the subsequent proof, it can also be observed that $\mathbb{E}\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] \to 0$ (Lemma 4.3). Additionally, we note that $N^{(k)}$ increases progressively. Consequently, during the iterative process, the variance of the sample mean tends to diminish, and the sample mean converges to the population mean. This indicates that the estimation precision of the sample mean is progressively enhanced, leading to more stable results.

4. Main Results

In Subsection 4.1, we begin by analyzing the convergence rate and computational complexity of Algorithm 1 when applied to non-convex and non-smooth composite functions. Subsequently, we show that KL functions exhibit distinct convergence rates under different conditions, as explained in Subsection 4.2.

4.1. **Non-convex and non-smooth problems.**

Lemma 4.1. *(Robbins-Siegmund) [8] Consider a measurable space $(\Omega, \mathscr{F}, P)$ and a sequence of sub-σ-algebras $\{\mathscr{F}^{(k)}\}_{k\geq 0}$ of $\mathscr{F}$ such that $\mathscr{F}^{(0)} \subset \mathscr{F}^{(1)} \subset \cdots$. Let $\{U^{(k)}\}_{k\geq 0}$, $\{\chi^{(k)}\}_{k\geq 0}$, $\{\xi^{(k)}\}_{k\geq 0}$, and $\{\zeta^{(k)}\}_{k\geq 0}$ be sequences of non-negative, integrable, and $\mathscr{F}^{(k)}$-measurable random variables satisfying the following conditions*

i. $\sum_{k=0}^{\infty} \chi^{(k)} < +\infty$ and $\sum_{k=0}^{\infty} \xi^{(k)} < +\infty$;

ii. For all $k \in \mathbb{N}$, $\mathbb{E}_k\left[U^{(k+1)}\right] \leq \left(1+\chi^{(k)}\right) U^{(k)} + \xi^{(k)} - \zeta^{(k)}$.

Then,

i. $\left\{U^{(k)}\right\}_{k\geq 0}$ *converges almost surely to a finite random variable* U^{∞}.

ii. $\sum_{k=0}^{\infty}\zeta^{(k)}$ *converges almost surely.*

Lemma 4.1 represents a well-established outcome in stochastic analysis, which aids in our examination of the algorithm's convergence.

Lemma 4.2. *Let Assumption 1.1 holds and* $x^{(k+1)}$ *be derived from equation (2.2). Then*

$$F^{(k+1)}\leq F^{(k)}+\frac{1}{2\eta L}\left\|\nabla f^{(k)}-\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2-\left(\frac{1}{2\alpha}-\frac{(\eta+1)L}{2}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2,$$

where η *is a positive constant.*

Proof. By utilizing the relationship that $F(x)=f(x)+h(x)$, we obtain

$$\begin{aligned}
F^{(k+1)}&\overset{(a)}{\leq} f^{(k)}+\left\langle\nabla f^{(k)},x^{(k+1)}-x^{(k)}\right\rangle+\frac{L}{2}\left\|x^{(k+1)}-x^{(k)}\right\|^2+h^{(k+1)}\\
&\overset{(b)}{\leq} f^{(k)}+\left\langle\nabla f^{(k)},x^{(k+1)}-x^{(k)}\right\rangle+\frac{L}{2}\left\|x^{(k+1)}-x^{(k)}\right\|^2\\
&\quad+h^{(k)}-\left\langle\nabla f^{(k)}_{\mathscr{N}^{(k)}},x^{(k+1)}-x^{(k)}\right\rangle-\frac{1}{2\alpha}\left\|x^{(k+1)}-x^{(k)}\right\|^2\\
&=F^{(k)}+\left\langle\nabla f^{(k)}-\nabla f^{(k)}_{\mathscr{N}^{(k)}},x^{(k+1)}-x^{(k)}\right\rangle-\left(\frac{1}{2\alpha}-\frac{L}{2}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2\\
&\overset{(c)}{\leq} F^{(k)}+\frac{1}{2\eta L}\left\|\nabla f^{(k)}-\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2-\left(\frac{1}{2\alpha}-\frac{(\eta+1)L}{2}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2.
\end{aligned}\tag{4.1}$$

Inequality (a) holds due to the Lipschitz continuity of ∇f. Inequality (b) can be derived from the update formula for $x^{(k+1)}$ given by equation (2.2):

$$h^{(k+1)}+\frac{1}{2\alpha}\left\|x^{(k+1)}-\left(x^{(k)}-\alpha\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right)\right\|^2\leq h^{(k)}+\frac{1}{2\alpha}\left\|x^{(k)}-\left(x^{(k)}-\alpha\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right)\right\|^2,$$

i.e.,

$$h^{(k+1)}\leq h^{(k)}-\left\langle\nabla f^{(k)}_{\mathscr{N}^{(k)}},x^{(k+1)}-x^{(k)}\right\rangle-\frac{1}{2\alpha}\left\|x^{(k+1)}-x^{(k)}\right\|^2.$$

Inequality (c) is established because $2\langle a,b\rangle\leq\eta L\|a\|^2+\frac{1}{\eta L}\|b\|^2$ for all $a,b\in\mathbb{R}^d$ and $\eta>0$. □

Lemma 4.3. *Let Assumptions 1.1 and 2.1 hold. Let the step size be* $0<\alpha<\frac{\eta L}{(\eta^2+\eta)L^2+\beta^2}$ *with* η *being a positive constant. The gradient estimator* $\nabla f^{(k)}_{\mathscr{N}^{(k)}}$ *satisfies relation (3.1), i.e.,*

$$\mathbb{E}_k\left[\left\|\nabla f^{(k)}_{\mathscr{N}^{(k)}}-\nabla f^{(k)}\right\|^2\right]\leq\varepsilon^{(k)}+\beta^2\mathbb{E}_k\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right],\tag{4.2}$$

where the nonnegative sequence $\{\varepsilon^{(k)}\}_{k\geq 0}$ *satisfies* $\sum_{k=0}^{+\infty}\varepsilon^{(k)}<\infty$. *Additionally, in conjunction with Lemma 4.1, the following assertions hold:*

i $F^{(k)}\longrightarrow F^{\infty}$ *a.s.;*

ii. $\sum_{k=0}^{+\infty}\mathbb{E}_k\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]<+\infty$ *a.s.;*

iii. $\sum_{k=0}^{+\infty}\mathbb{E}\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]<+\infty$;

iv. $\mathbb{E}\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]\to 0$.

Proof. By conditioning on $\mathscr{F}^{(k)}$ and computing the expected value in inequality (4.1), we find that

$$\begin{aligned}\mathbb{E}_k\left[F^{(k+1)}-F_\star|\mathscr{F}^{(k)}\right]\leq& F^{(k)}-F_\star-\mathbb{E}\left[\left(\frac{1}{2\alpha}-\frac{(\eta+1)L}{2}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2|\mathscr{F}^{(k)}\right]\\&+\mathbb{E}\left[\frac{1}{2\eta L}\left\|\nabla f^{(k)}-\nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2|\mathscr{F}^{(k)}\right]\\\overset{(3.1)}{\leq}& F^{(k)}-F_\star+\frac{1}{2\eta L}\varepsilon^{(k)}\\&-\mathbb{E}\left[\left(\frac{1}{2\alpha}-\frac{(\eta+1)L}{2}-\frac{\beta^2}{2\eta L}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2|\mathscr{F}^{(k)}\right].\end{aligned}\tag{4.3}$$

Recalling Lemma 4.1 and considering $U^{(k)}=F^{(k)}-F_\star$, $\xi^{(k)}=\frac{1}{2\eta L}\varepsilon^{(k)}$ and

$$\zeta^{(k)}=\mathbb{E}_k\left[\left(\frac{1}{2\alpha}-\frac{(\eta+1)L}{2}-\frac{\beta^2}{2\eta L}\right)\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]$$

with parameters chosen as $\alpha<\frac{\eta L}{(\eta^2+\eta)L^2+\beta^2}$, $\eta>0$, $\sum_{k=0}^{+\infty}\varepsilon^{(k)}<\infty$ and $\varepsilon^{(k)}>0$, we find that Lemma 4.3 *i* and *ii* hold. Following that, we utilize Lemma 4.3 *ii* and the properties of conditional expectation to establish the following:

$$\begin{aligned}\sum_{k=0}^{+\infty}\mathbb{E}\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]&=\sum_{k=0}^{+\infty}\mathbb{E}\left[\mathbb{E}_k\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]\right]\\&=\mathbb{E}\left[\sum_{k=0}^{+\infty}\mathbb{E}_k\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]\right]<+\infty.\end{aligned}$$

Therefore, we obtain Lemma 4.3 *iii* and *iv*. □

Theorem 4.1. *Let the conditions of Lemma 4.3 hold. Then*

i. any limit point of $\left\{x^{(k)}\right\}_{k\geq 0}$ *is a critical point for problem (1.1) a.s.;*

ii.

$$\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(R)}\right)\right)\right]\leq\mathscr{O}\left(\frac{1}{\sqrt{K}}\right),\tag{4.4}$$

where R is uniformly chosen from the set $\{0,1,\cdots,K\}$.

Proof. We presuppose the presence of a subsequence that converges almost surely to $\bar{x}$. That is, there exists a subset $\mathscr{K}\subseteq\mathbb{N}$ for which the following limit holds:

$$\lim_{k\to\infty,\,k\in\mathscr{K}}x^{(k)}=\bar{x}\text{ a.s..}$$

From optimality condition of (2.2), we have

$$0\in\nabla f^{(k)}_{\mathscr{N}^{(k)}}+\frac{1}{\alpha}\left(x^{(k+1)}-x^{(k)}\right)+\partial h^{(k+1)}.$$

Hence, it follows that

$$\nabla f^{(k+1)} - \nabla f^{(k)}_{\mathscr{N}^{(k)}} - \frac{1}{\alpha}\left(x^{(k+1)} - x^{(k)}\right) \in \nabla f^{(k+1)} + \partial h^{(k+1)} = \partial F^{(k+1)}.$$

Furthermore, we can deduce the conclusion that

$$\begin{aligned}\operatorname{dist}\left(0, \partial F^{(k+1)}\right)^2 &\leq \left\|\nabla f^{(k+1)} - \nabla f^{(k)}_{\mathscr{N}^{(k)}} - \frac{1}{\alpha}\left(x^{(k+1)} - x^{(k)}\right)\right\|^2 \\ &\overset{(a)}{\leq} 4\left\|\nabla f^{(k+1)} - \nabla f^{(k)}\right\|^2 + 4\left\|\nabla f^{(k)} - \nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2 + \frac{2}{\alpha^2}\left\|x^{(k+1)} - x^{(k)}\right\|^2.\end{aligned}$$

Inequality (a) is derived from the fact that, for any vectors $a, b \in \mathbb{R}^d$, the inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ holds. Taking the expectation on both sides of the above inequality, we have

$$\begin{aligned}\mathbb{E}\left[\operatorname{dist}\left(0, \partial F^{(k+1)}\right)^2\right] \leq& \mathbb{E}\left[4\left\|\nabla f^{(k+1)} - \nabla f^{(k)}\right\|^2\right] + \mathbb{E}\left[\frac{2}{\alpha^2}\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] \\ &+ \mathbb{E}\left[4\left\|\nabla f^{(k)} - \nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2\right] \\ \overset{(a)}{=}& \mathbb{E}\left[4\left\|\nabla f^{(k+1)} - \nabla f^{(k)}\right\|^2\right] + \mathbb{E}\left[\frac{2}{\alpha^2}\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] \\ &+ \mathbb{E}\left[\mathbb{E}\left[4\left\|\nabla f^{(k)} - \nabla f^{(k)}_{\mathscr{N}^{(k)}}\right\|^2 | \mathscr{F}^{(k)}\right]\right] \\ \overset{(b)}{\leq}& \left(4L^2 + \frac{2}{\alpha^2} + 4\beta^2\right)\mathbb{E}\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] + 4\varepsilon^{(k)}.\end{aligned} \tag{4.5}$$

Equation (a) holds because of the property of conditional expectation $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot|\mathscr{F}]]$. Inequality (b) is valid as the stochastic gradient $\nabla f^{(k)}_{\mathscr{N}^{(k)}}$ complies with relation (4.2), given the Lipschitz continuity of ∇f and $\alpha > 0$. Therefore, by considering

$$\mathbb{E}\left[\operatorname{dist}\left(0, \partial F\left(x^{(k)}\right)\right)\right] \leq \sqrt{\mathbb{E}\left[\operatorname{dist}(0, \partial F\left(x^{(k)}\right))^2\right]},$$

the definition of sequence $\left\{\varepsilon^{(k)}\right\}_{k\geq 0}$, and $\lim_{k\to+\infty}\mathbb{E}[\|x^{(k+1)} - x^{(k)}\|^2] = 0$ (Lemma 4.3 *iv*), we obtain

$$\lim_{k\to+\infty}\mathbb{E}\left[\operatorname{dist}\left(0, \partial F\left(x^{(k)}\right)\right)\right] = 0.$$

Then there exists $\mathscr{K}' \subseteq \mathscr{K}$ such that

$$\lim_{k\to+\infty, k\in\mathscr{K}'} \operatorname{dist}\left(0, \partial F\left(x^{(k)}\right)\right) = 0 \quad \text{a.s..}$$

According to Definition 2.3, $\bar{x}$ is a critical point a.s..

Let us review inequality (4.5). In conjunction with Jensen's inequality, we obtain

$$\begin{aligned}&\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(R)}\right)\right)\right]\\ \leq&\sqrt{\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(R)}\right)\right)^2\right]}\\ \leq&\sqrt{\frac{1}{K}\left(\left(4L^2+\frac{2}{\alpha^2}+4\beta^2\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]+4\sum_{k=0}^{K-1}\varepsilon^{(k)}\right)},\end{aligned} \tag{4.6}$$

where R is uniformly chosen from the set $\{0,1,\cdots,K\}$. By integrating the preceding inequality along with

$$\sum_{k=0}^{+\infty}\mathbb{E}\left[\left\|x^{(k+1)}-x^{(k)}\right\|^2\right]<+\infty$$

and $\sum_{k=0}^{+\infty}\varepsilon^{(k)}<+\infty$, we derive the result expressed as inequality (4.4). □

Corollary 4.1. *Based on the conditions outlined in Theorem 4.1, if we choose $\varepsilon^{(k)}$ to be of the form $C(k+1)^{-1-\kappa_1}$, which is tailored for the worst-case scenario where $N^{(k)}$ grows as $\mathscr{O}\left((k+1)^{1+\kappa_1}\right)$, then the computational complexity to reach an ε-approximate critical point is $\mathscr{O}\left(\frac{1}{\varepsilon^{4+\kappa}}\right)$, where C, $\kappa=2\kappa_1$ and κ_1 are three positive constants.*

Proof. The proof is detailed in Appendix 7.2. □

4.2. **KL functions.** We initially establish the basis for stochastic KL inequality (4.7).

Lemma 4.4. *Let the conditions of Lemma 4.3 hold and the sequence $\{x^{(k)}\}_{k\geq 0}$ be bounded. Let $\omega:=\{x:\exists$ an increasing sequence of integers $\{k_q\}_{q\in\mathbb{N}}$ such that $x^{(k_q)}\to x$ as $q\to+\infty\}$. Then, the following statements hold:*

i. $\sum_{k=0}^{+\infty}\left\|x^{(k+1)}-x^{(k)}\right\|^2<+\infty$ *a.s., and* $\left\|x^{(k+1)}-x^{(k)}\right\|\to 0$ *a.s.;*

ii. $\mathbb{E}\left[F\left(x^{(k)}\right)\right]\to F_\star$, *where* $F_\star\in[\underline{F},\infty)$ *and* $\underline{F}:=\inf_x F(x)$;

iii. $\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^{(k)}\right)\right)\right]\to 0$;

iv. The set ω is non-empty, and for all $x^\star\in\omega$, $\mathbb{E}\left[\operatorname{dist}\left(0,\partial F\left(x^\star\right)\right)\right]=0$;

v. $\operatorname{dist}\left(x^{(k)},\omega\right)\to 0$ *a.s.;*

vi. ω is a.s. compact and connected;

vii. $\mathbb{E}\left[F\left(x^\star\right)\right]=F_\star$ *for all* $x^\star\in\omega$.

Proof. The proof is detailed in Appendix 7.3. □

Lemma 4.5. *([22, Lemma 4.5]) Suppose that the conditions of Lemma 4.4 are satisfied, and $x^{(k)}$ has not reached a critical point of F within a finite number of iterations. Let F be a semialgebraic function that satisfies the KL property, with the desingularizing function $\varphi(x)=\frac{c}{\theta}x^\theta$, where $c>0$ and $\theta\in(0,1]$ are constants. Under these conditions, there exist an index k_0 and a desingularizing function φ such that, for all $k\geq k_0$,*

$$\varphi'\left(\mathbb{E}\left[F^{(k)}-F_k^\star\right]\right)\mathbb{E}\left[\operatorname{dist}\left(0,\partial F^{(k)}\right)\right]\geq 1, \tag{4.7}$$

where $F_k^\star$ is a non-decreasing sequence converging to $\mathbb{E}\left[F\left(x^\star\right)\right]$ for some $x^\star\in\omega$.

Remark 4.1. In [22, Lemma 4.5], the constructed $F_k^\star$ satisfies

$$\mathbb{E}\left[F^{(k)} - F_k^\star\right] > 0, \tag{4.8}$$

which in turn ensures that $\varphi\left(\left[F^{(k)} - F_k^\star\right]\right) > 0$ and $\varphi'\left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right) > 0$. Additionally, since $F_k^\star$ is a non-decreasing sequence converging to $\mathbb{E}[F(x^\star)]$, it follows from Lemma 4.4 *ii* and *vii* that

$$0 < \mathbb{E}\left[F^{(k)} - F_\star\right] = \mathbb{E}\left[F^{(k)} - F^\star\right] \leq \mathbb{E}\left[F^{(k)} - F_k^\star\right],\ k \geq k_0. \tag{4.9}$$

Theorem 4.2. *Let the conditions of Lemma 4.5 hold and $\varepsilon^{(k)} = C\rho^k$, where $\rho \in (0,1)$ and C be two positive constants. Then*

i. If $\theta = 1$, then there exists an k_1 and such that $\mathbb{E}\left[F^{(k)} - F^\star\right] = 0$ for all $k \geq k_1$;

ii. If $\theta = \left[\frac{1}{2}, 1\right)$, then exist positive constants k_2, C_1 and $\bar{\rho} \in (0,1)$ such that

$$\mathbb{E}\left[F^{(k)} - F^\star\right] \leq C_1\bar{\rho}^k,\ k \geq k_2; \tag{4.10}$$

iii. If $\theta = \left(0, \frac{1}{2}\right)$ and $\rho \leq \left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}}$, then exist constants k_3 and C_2 such that

$$\mathbb{E}\left[F^{(k)} - F^\star\right] \leq C_2\left(\frac{1}{k}\right)^{\frac{1}{1-2\theta}},\ k \geq k_3. \tag{4.11}$$

Proof. Applying the expected version of the KL inequality (4.7) with $\varphi(x) = \frac{c}{\theta}x^\theta$ and some constants $c > 0, \theta \in (0,1]$, one sees that there exists k_0 such

$$c\left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right)^{\theta-1}\mathbb{E}\left[\operatorname{dist}\left(0, \partial F^{(k)}\right)\right] \geq 1,\ k \geq k_0. \tag{4.12}$$

Furthermore, inequalities (4.12), (4.5), and (4.3) ensure that

$$\begin{aligned}
1 \leq & c^2\left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right)^{2\theta-2}\mathbb{E}\left[\operatorname{dist}\left(0, \partial F\left(x^{(k)}\right)\right)\right]^2 \\
\overset{(a)}{\leq} & c^2\left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right)^{2\theta-2}\mathbb{E}\left[\operatorname{dist}\left(0, \partial F\left(x^{(k)}\right)\right)^2\right] \\
\overset{(4.5)}{\leq} & c^2\left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right)^{2\theta-2}\left(\left(4L^2 + \frac{2}{\alpha^2} + 4\beta^2\right)\mathbb{E}\left[\left\|x^{(k)} - x^{(k-1)}\right\|^2\right] + 4\varepsilon^{(k)}\right) \\
\overset{(4.3)}{\leq} & \left(\mathbb{E}\left[F^{(k)} - F_k^\star\right]\right)^{2\theta-2}\left(c_1\left(\mathbb{E}\left[F^{(k-1)}\right] - \mathbb{E}\left[F^{(k)}\right]\right) + c_2\varepsilon^{(k-1)}\right),
\end{aligned} \tag{4.13}$$

where

$$c_1 = c^2\frac{4L^2 + \frac{2}{\alpha}^2 + 4\beta^2}{\frac{1}{2\alpha} - \frac{(\eta+1)L}{2} - \frac{\beta^2}{2\eta L}}$$

and $c_2 = c^2\left(\frac{c_1}{2\eta L} + 4\right)$ are two positive constants. Inequality (a) holds due to the Jensen's inequality. Denote

$$r^{(k)} := \mathbb{E}\left[F^{(k)} - F_k^\star\right] \overset{(4.8)}{>} 0. \tag{4.14}$$

Combining (4.13) and the fact that $F_k^\star$ is a non-decreasing sequence, we obtain

$$1 \leq \left(r^{(k)}\right)^{2\theta-2}\left(c_1\left(r^{(k-1)}-r^{(k)}\right)+c_2\varepsilon^{(k-1)}\right),\ k \geq k_0. \tag{4.15}$$

Below, we discuss each case separately for $\theta = 1$, $\theta = \left[\frac{1}{2}, 1\right)$ and $\theta \in \left(0, \frac{1}{2}\right)$.

i. Case $\theta = 1$. In this case, (4.13) becomes

$$1 \leq c_1\left(\mathbb{E}\left[F^{(k-1)}-F^\star\right]-\mathbb{E}\left[F^{(k)}-F^\star\right]\right)+c_2\varepsilon^{(k-1)},\ k \geq k_0.$$

Because $\mathbb{E}\left[F^{(k)}-F^\star\right] \to 0$, $c_1 > 0$, $c_2 > 0$,and $\varepsilon^{(k-1)} \to 0$, this is a contradiction. Thus there exists k_1 such that $\mathbb{E}\left[F^{(k)}-F^\star\right] = 0$ for all $k > k_1$. The algorithm terminates in finite steps.

ii. Case $\theta \in \left[\frac{1}{2}, 1\right)$. In this case, $2-2\theta \in [0,1)$. As $r^{(k)} \to 0$, there exists $\hat{k}_2$ such that $r^{(k)} \leq \left(r^{(k)}\right)^{2-2\theta}$ with $k \geq \hat{k}_2$. Therefore, we obtain

$$r^{(k)} \leq \left(r^{(k)}\right)^{2-2\theta} \overset{(4.15)}{\leq} c_1\left(r^{(k-1)}-r^{(k)}\right)+c_2\varepsilon^{(k-1)},\ k \geq k_2$$

with $k_2 = \max\left\{k_0, \hat{k}_2\right\}$. By performing some simple calculations on the previous expression, we derive the subsequent relationship:

$$r^{(k)} \leq \frac{c_1}{1+c_1}r^{(k-1)}+\frac{c_2}{1+c_1}\varepsilon^{(k-1)},\ k \geq k_2. \tag{4.16}$$

Next, we discuss the following two cases (4.16):

If

$$\frac{c_2}{1+c_1}\varepsilon^{(k-1)} \leq \frac{1}{\nu\left(1+c_1\right)}r^{(k-1)},$$

where $\nu > 1$ is a constant, then

$$r^{(k)} \leq \left(\frac{1}{\nu}+c_1\right)\frac{1}{1+c_1}r^{(k-1)}.$$

If $\varepsilon^{(k)} = C\rho^k$ and

$$\frac{c_2}{1+c_1}\varepsilon^{(k-1)} \geq \frac{1}{\nu\left(1+c_1\right)}r^{(k-1)}$$

then

$$r^{(k)} \leq \frac{\nu c_1 c_2}{1+c_1}\varepsilon^{(k-1)}+\frac{c_2}{1+c_1}\varepsilon^{(k-1)} \leq \frac{c_2(1+\nu c_1)}{1+c_1}C\rho^{k-1},$$

where $\rho \in (0,1)$ is a constant. Combining these two cases, we obtain

$$\mathbb{E}\left[F^{(k)}-F^\star\right] \overset{(4.9)}{\leq} r^{(k)} \leq \max\left\{r^{(k_2)}, \frac{c_2(1+\nu c_1)}{1+c_1}C\rho^{k_2}\right\}\bar{\rho}^{k-k_2-1},\ k \geq k_2,$$

where $\bar{\rho} = \max\left\{\left(\frac{1}{\nu}+c_1\right)\frac{1}{1+c_1}, \rho\right\}$ and $v \geq 1$. Therefore, there exist positive constants k_2, C_1, and $\bar{\rho} \in (0,1)$ such that (4.10) holds.

iii. Case $\theta \in \left(0, \frac{1}{2}\right)$. In case, $2\theta - 2 \in (-2,-1)$ and $2\theta - 1 \in (-1,0)$. By performing a simple transformation and scaling on (4.15), we obtain

$$\begin{aligned}
1 &\overset{(4.15)}{\leq} \left(r^{(k)}\right)^{2\theta-2}\left(c_1\left(r^{(k-1)} - r^{(k)}\right) + c_2\varepsilon^{(k-1)}\right) \\
&\overset{(a)}{=} \left(r^{(k)}\right)^{2\theta-2}\left(c_1\left(r^{(k-1)} - r^{(k)}\right) + c_2 C\rho^{k-1}\right) \\
&= c_1\left(r^{(k)}\right)^{2\theta-2}\left(\left(r^{(k-1)} + c_3 C\rho^{k-1}\right) - \left(r^{(k)} + c_3 C\rho^{k}\right)\right. \\
&\quad \left. - \left(c_3 C - \frac{c_2}{c_1}C - c_3 C\rho\right)\rho^{k-1}\right) \\
&\overset{(b)}{\leq} c_1\left(r^{(k)}\right)^{2\theta-2}\left(\left(r^{(k-1)} + c_3\varepsilon^{(k-1)}\right) - \left(r^{(k)} + c_3\varepsilon^{(k)}\right)\right),\ k \geq k_0.
\end{aligned}$$

Equation (a) holds because $\varepsilon^{(k)} = C\rho^k$, and inequality (b) holds by setting positive constant $c_3 \geq \frac{c_2}{c_1(1-\rho)}$. Set

$$t^{(k)} := r^{(k)} + c_3\varepsilon^{(k)}. \tag{4.17}$$

Then

$$1 \leq c_1\left(r^{(k)}\right)^{2\theta-2}\left(t^{(k-1)} - t^{(k)}\right),\ k \geq k_0. \tag{4.18}$$

As $t^{(k)} \leq t^{(k-1)}$ and $t^{(k)} \to 0$, we see that there exists $\hat{k}_3$ such that

$$\left(t^{(k)}\right)^{2\theta-2} \geq \left(t^{k-1}\right)^{2\theta-2},\ \text{and}\ \left(t^{(\hat{k}_3)}\right)^{2\theta-1} \leq \cdots \leq \left(t^{(k)}\right)^{2\theta-1},\ k \geq \hat{k}_3. \tag{4.19}$$

Define auxiliary function

$$\phi(x) := \frac{c_4}{1-2\theta}x^{2\theta-1}, \tag{4.20}$$

then

$$\phi'(x) = -c_4 x^{2\theta-2}. \tag{4.21}$$

Next, we will discuss in two cases. If

$$\left(t^{(k)}\right)^{2\theta-2} \leq 2\left(t^{(k-1)}\right)^{2\theta-2},$$

then

$$t^{(k)} \geq \left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}} t^{(k-1)}.$$

Substituting $t^{(k)} := r^{(k)} + c_3\varepsilon^{(k)}$ (4.17) into the previous expression, we obtain

$$r^{(k)} \geq \left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}}\left(r^{(k-1)} + c_3\varepsilon^{(k-1)}\right) - c_3\varepsilon^{(k)} \overset{(a)}{\geq} c_3 C\left(\left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}} - \rho\right)\rho^{k-1}. \tag{4.22}$$

Inequality (a) holds because $r^{(k-1)} > 0$ and set $\rho \leq \left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}}$. Therefore,

$$\begin{aligned}
\phi\left(t^{(k)}\right) - \phi\left(t^{(k-1)}\right) &= \int_{t^{(k-1)}}^{t^{(k)}} \phi'(t)dt \overset{(4.21)}{=} c_4 \int_{t^{(k)}}^{t^{(k-1)}} t^{2\theta-2}dt \\
&\geq c_4\left(t^{(k-1)} - t^{(k)}\right)\left(t^{(k-1)}\right)^{2\theta-2} \\
&\geq \frac{c_4}{2}\left(t^{(k-1)} - t^{(k)}\right)\left(t^{(k)}\right)^{2\theta-2} \\
&\overset{(4.18)}{\geq} \frac{c_4}{2c_1}\frac{\left(t^{(k)}\right)^{2\theta-2}}{\left(r^{(k)}\right)^{2\theta-2}} \\
&\overset{(4.17)}{=} \frac{c_4}{2c_1}\left(\frac{r^{(k)} + c_3\varepsilon^{(k)}}{r^{(k)}}\right)^{2\theta-2} \\
&\overset{(4.22)}{\geq} \frac{c_4}{2c_1}\left(1 + \frac{c_3 C\rho^k}{c_3 C\left(\left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}} - \rho\right)\rho^{k-1}}\right)^{2\theta-2} \\
&= \frac{c_4}{2c_1}\left(1 + \frac{\rho}{\left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}} - \rho}\right)^{2\theta-2}, \ k \geq k_3
\end{aligned} \tag{4.23}$$

with $k_3 = \max\left\{k_0, \hat{k}_3\right\}$. If $\left(t^{(k)}\right)^{2\theta-2} \geq 2\left(t^{(k-1)}\right)^{2\theta-2}$, then

$$\left(t^{(k)}\right)^{2\theta-1} \geq 2^{\frac{2\theta-1}{2\theta-2}}\left(t^{(k-1)}\right)^{2\theta-1}. \tag{4.24}$$

Therefore, we obtain

$$\begin{aligned}
\phi\left(t^{(k)}\right) - \phi\left(t^{(k-1)}\right) &\overset{(4.20)}{=} \frac{c_4}{1-2\theta}\left(\left(t^{(k)}\right)^{2\theta-1} - \left(t^{(k-1)}\right)^{2\theta-1}\right) \\
&\overset{(4.24)}{\geq} \frac{c_4}{1-2\theta}\left(2^{\frac{2\theta-1}{2\theta-2}} - 1\right)\left(t^{(k-1)}\right)^{2\theta-1} \\
&\overset{(4.19)}{\geq} \frac{c_4}{1-2\theta}\left(2^{\frac{2\theta-1}{2\theta-2}} - 1\right)\left(t^{(k_3)}\right)^{2\theta-1}, \ k \geq k_3.
\end{aligned} \tag{4.25}$$

Letting

$$c_5 = \min\left\{\frac{c_4}{2c_1}\left(1 + \frac{\rho}{\left(\frac{1}{2}\right)^{\frac{1}{2-2\theta}} - \rho}\right)^{2\theta-2}, \frac{c_4}{1-2\theta}\left(2^{\frac{2\theta-1}{2\theta-2}} - 1\right)\left(t^{(k_3)}\right)^{2\theta-1}\right\},$$

we can combine cases (4.23) and (4.25) to obtain

$$\phi\left(t^{(k)}\right) - \phi\left(t^{(k-1)}\right) \geq c_5, \ k \geq k_3,$$

and

$$\phi\left(t^{(k)}\right) \geq \phi\left(t^{(k)}\right) - \phi\left(t^{(k_3)}\right) = \sum_{i=k_3+1}^{k} \phi\left(t^{(i)}\right) - \phi\left(t^{(i-1)}\right) \geq (k-k_3)\, c_5,\ k \geq k_3.$$

Reviewing the definition of $\phi(t) := \frac{c_4}{1-2\theta} t^{2\theta-1}$ (4.20) and performing some simple transformations on the above expression, it can be inferred that

$$\frac{c_4}{1-2\theta}\left(t^{(k)}\right)^{2\theta-1} \leq (k-k_3)\, c_5,\ k \geq k_3.$$

Substituting $t^{(k)} := r^{(k)} + c_3\varepsilon^{(k)}$ (4.17) and $r^{(k)} := \mathbb{E}\left[F^{(k)} - F_k^\star\right]$ (4.14) into the previous expression, we obtain

$$\begin{aligned}\mathbb{E}\left[F^{(k)} - F^\star\right] \overset{(4.9)}{\leq} r^{(k)} &\leq \left(\frac{c_4}{(k-k_3)\, c_5\,(1-2\theta)}\right)^{\frac{1}{1-2\theta}} - c_3\varepsilon^{(k)} \\ &\leq \left(\frac{c_4}{c_5\,(1-2\theta)}\right)^{\frac{1}{1-2\theta}} \left(\frac{1}{k-k_3}\right)^{\frac{1}{1-2\theta}},\ k \geq k_3.\end{aligned}$$

Therefore, there exist constants k_3 and C_2 such that (4.11) holds. □

Corollary 4.2. *Based on the conditions outlined in Theorem 4.2 ii, if we choose $\varepsilon^{(k)}$ to be of the form $C(1+\kappa)^{-k}$, which is tailored for the worst-case scenario where $N^{(k)}$ grows as $\mathcal{O}\left((1+\kappa)^k\right)$, then the computational complexity to reach an ε-solution is $\mathcal{O}(1/\varepsilon)$.*

Proof. The proof is detailed in Appendix 7.4. □

5. Numerical Experiments

To evaluate the performance of our proposed method, we apply it to a fundamental minimization problem in the training of a binary classifier. The problem is formulated as follows:

$$F(x) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \tanh\left(b_i\left\langle a_i, x\right\rangle\right)\right) + h(x). \tag{5.1}$$

We investigate the l_0 regularization term $\lambda\|x\|_0$ and the $l_{1/2}$ regularization term $\lambda\|x\|_{1/2}$, with λ set to $\frac{1}{100N}$. The development environment consists of Python 3.10.9, which was executed on a personal computer with Windows 10 as the operating system. The PC was powered by an AMD Ryzen 7 5700X octa-core Processor with a clock speed of 3.40 GHz and was equipped with 32.0 GB of RAM. We assessed various algorithms using three distinct datasets sourced from the LIBSVM dataset repository:https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Table 2. Summary of datasets

dataset	training size N	testing size	feature
w8a	49,749	14,951	300
rcv1.binary.	20,242	677,399	47,236

We conducted a comparative analysis with the state-of-the-art algorithms reported in current academic literature.

- Prox-SG [43]: fixed sampling size and vanishing step size $\alpha^{(k)} = \frac{\alpha^{(0)}}{1+0.1\lfloor \frac{k}{N} \rfloor}$.
- Prox-SVRG [33] and Prox-SARAH [35]: number of internal cycles $2N$, fixed sampling size and fixed step size.
- Prox-SAGA [30]: fixed sampling size and fixed step size.
- Prox-ASSG: adaptive sampling and adaptive step size $\alpha^{(k)} = \min\left\{\alpha^{(0)}\frac{N^{(k)}}{N^{(0)}}, \alpha_{\max}\right\}$.

It is well-known that there is a subtle relationship between sampling size and step size. Based on the research in references [10, 44], we apply a linear scaling rule in the Prox-ASSG algorithm: as the sampling size increases, the step size is also scaled up by a corresponding factor. We selected the step size from the collection $\{2^{-8}, 2^{-6}, 2^{-4}, 2^{-2}, 1, 2^2, 2^4, 2^6, 2^7\}$ and the sampling size from the collection $\{1, 2^4, 2^6, 2^7, 2^8, 2^9\}$. The optimal parameters were identified based on the convergence rate and the objective function's performance on the training dataset. These optimal values $F(x^\star)$ were obtained through exhaustive iterations employing the Prox-SG method. We subsequently detail the optimal parameters for the algorithms in the subsequent tables.

TABLE 3. Optimal parameters (l_0)

algorithm	w8a	rcv1.binary.
Prox-SG	$\alpha^{(0)} = 2^7, N^{(k)} = 2^4$	$\alpha^{(0)} = 2^6, N^{(k)} = 2^4$
Prox-SVRG	$\alpha = 2^4, N^{(k)} = 2^9$	$\alpha = 2^7, N^{(k)} = 2^8$
Prox-SAGA	$\alpha = 2^2, N^{(k)} = 2^4$	$\alpha = 2^4, N^{(k)} = 2^4$
Prox-SARAH	$\alpha = 2^{-2}, N^{(k)} = 2^4$	$\alpha = 2^7, N^{(k)} = 2^9$
Prox-ASSG	$\alpha^{(0)} = 2^4$, $\alpha_{\max} = 2^7$ $N^{(0)} = 16$	

TABLE 4. Optimal parameters $(l_{1/2})$

algorithm	w8a	rcv1.binary.
Prox-SG	$\alpha^{(0)} = 2^6, N^{(k)} = 2^6$	$\alpha^{(0)} = 2^6, N^{(k)} = 2^4$
Prox-SVRG	$\alpha = 2^{-2}, N^{(k)} = 2^4$	$\alpha = 2^7, N^{(k)} = 2^7$
Prox-SAGA	$\alpha = 1, N^{(k)} = 2^4$	$\alpha = 2^6, N^{(k)} = 2^6$
Prox-SARAH	$\alpha = 2^{-2}, N^{(k)} = 2^6$	$\alpha = 2^7, N^{(k)} = 2^9$
Prox-ASSG	$\alpha^{(0)} = 2^4$, $\alpha_{\max} = 2^7$, $N^{(0)} = 16$	

We have defined $\Delta^{(k)}$ in the adaptive sampling strategy as $\varepsilon^{(k)} = 10 \cdot 0.999^k$ and set $\beta = 10^{-4}$. The ensuing Figures 1-16 depict the performance of several algorithms when applied to non-convex and non-smooth functions, as specified by equation (5.1), across various datasets.

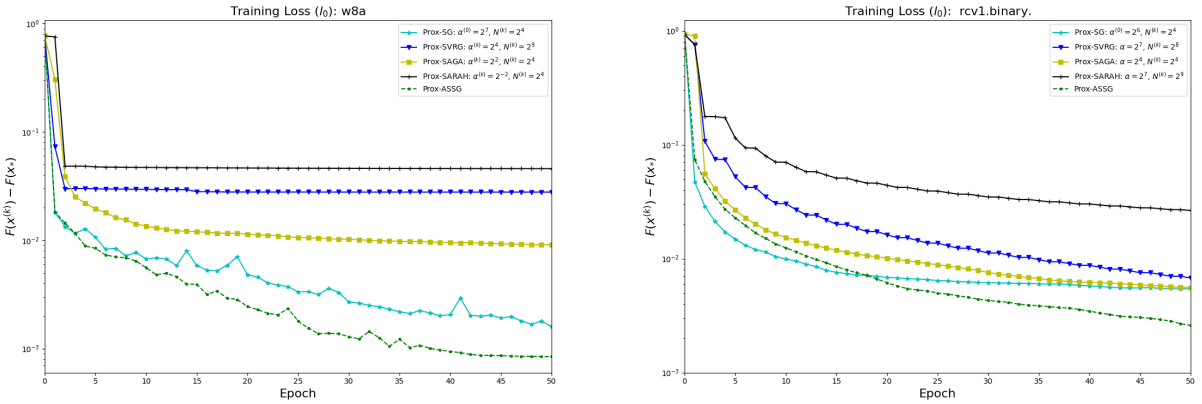

FIGURE 1. The function values of algorithms change with epochs on training set (l_0).

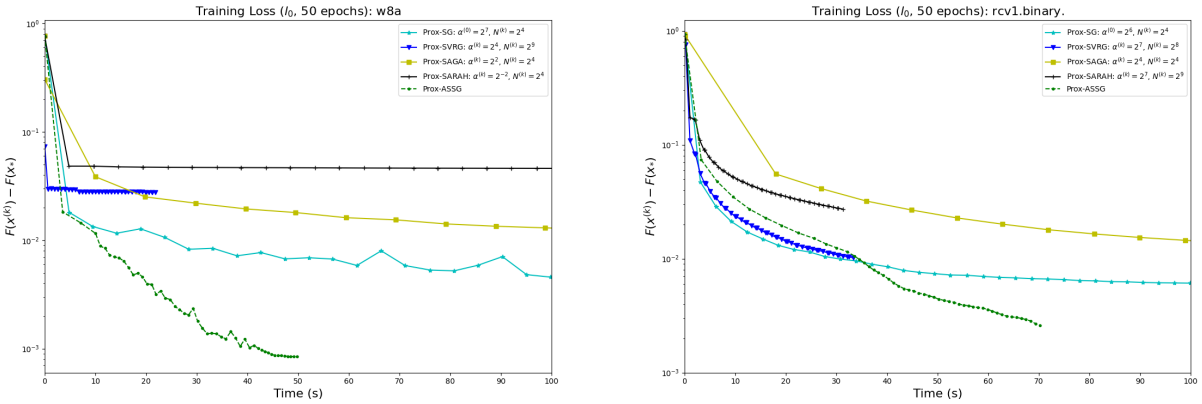

FIGURE 2. The function values of algorithms change with time on training set (l_0, 50 epochs).

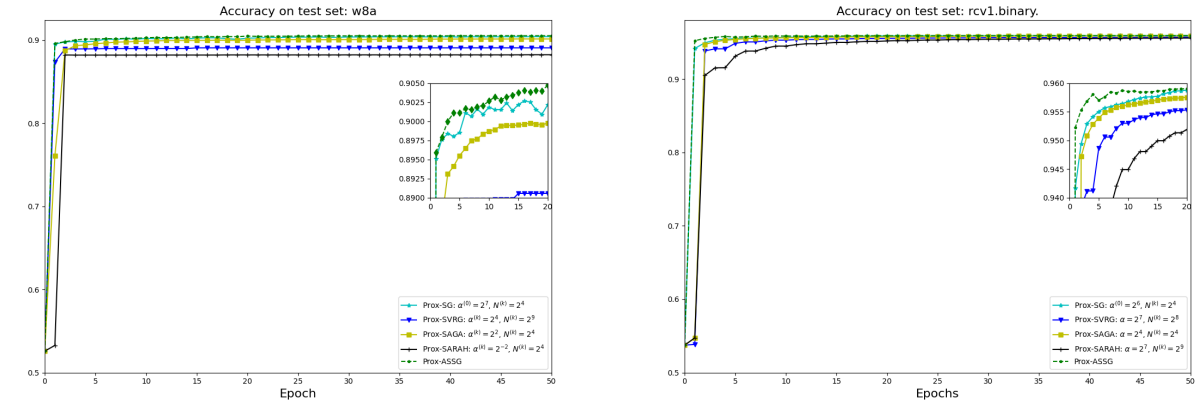

FIGURE 3. The accuracy of algorithms change with epochs on test set (l_0).

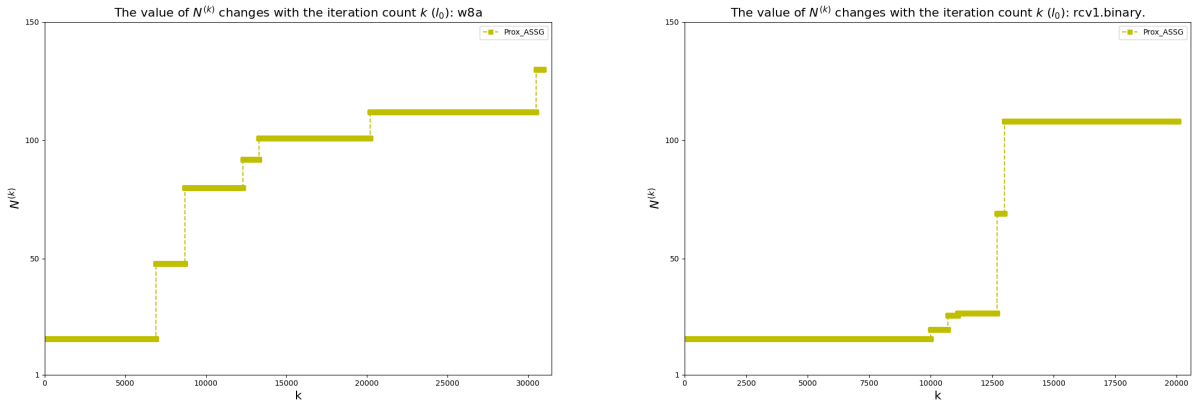

FIGURE 4. The sampling size $N^{(k)}$ change with the number of iterations k (l_0).

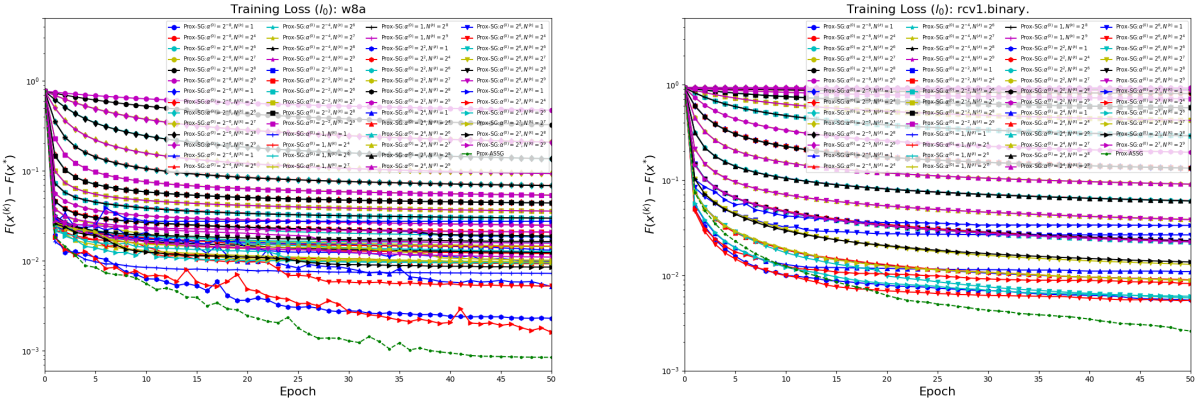

FIGURE 5. The step size and sampling size effects on Prox-SG algorithm (l_0).

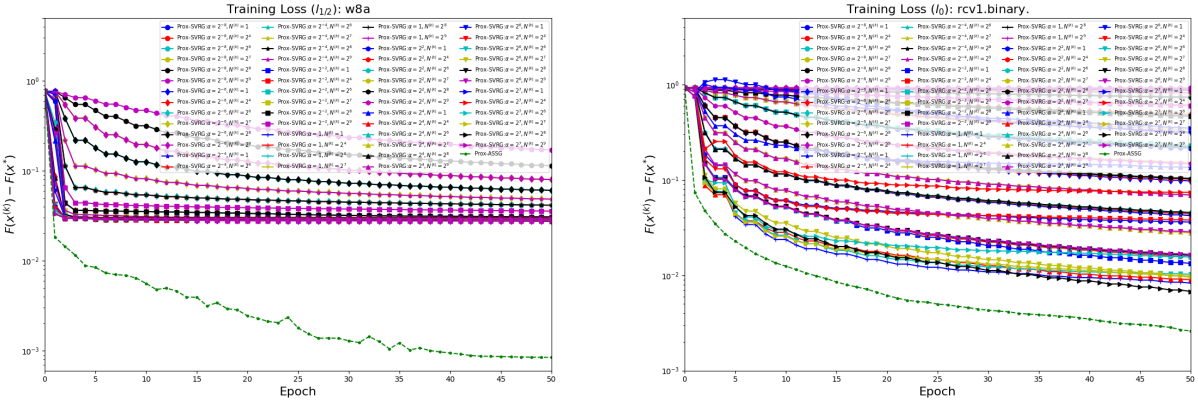

FIGURE 6. The step size and sampling size effects on Prox-SVRG algorithm (l_0).

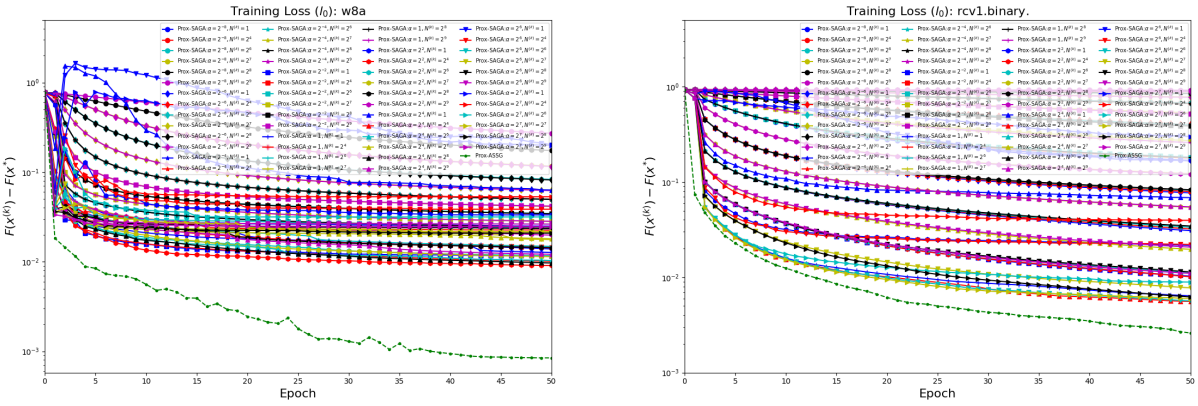

FIGURE 7. The step size and sampling size effects on Prox-SAGA algorithm (l_0).

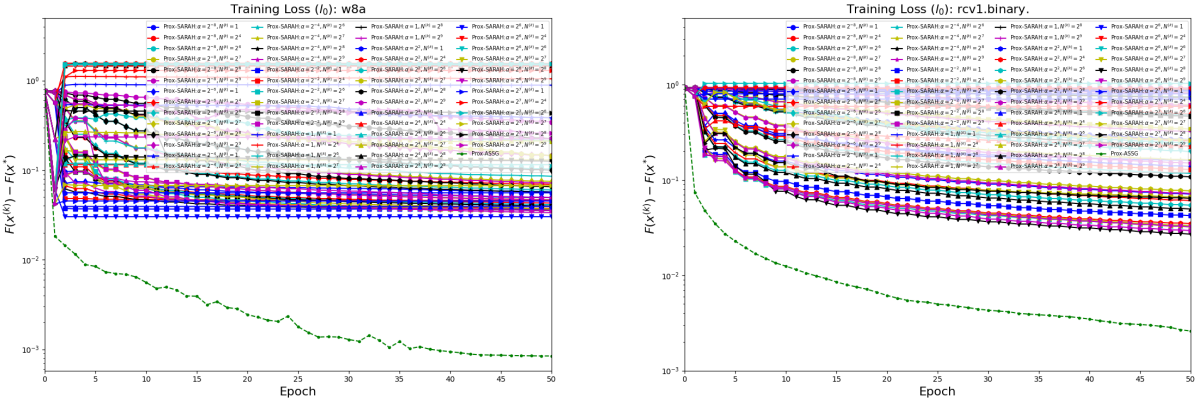

FIGURE 8. The step size and sampling size effects on Prox-SARAH algorithm (l_0).

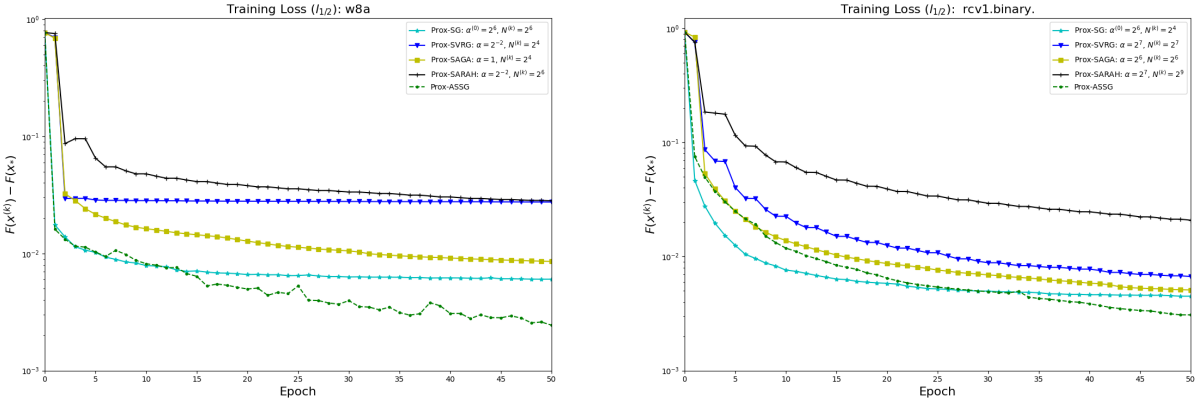

FIGURE 9. The function values of algorithms change with epochs on training set ($l_{1/2}$).

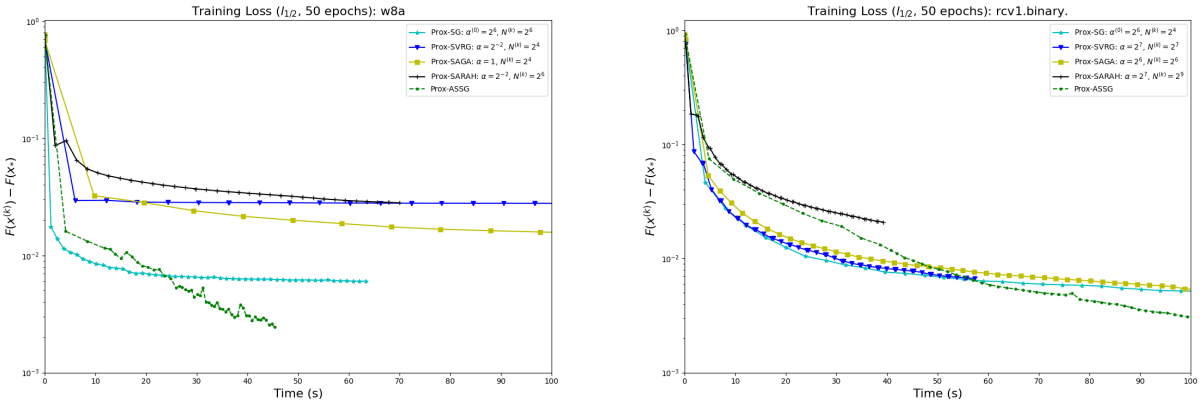

FIGURE 10. The function values of algorithms change with time on training set ($l_{1/2}$, 50 epochs).

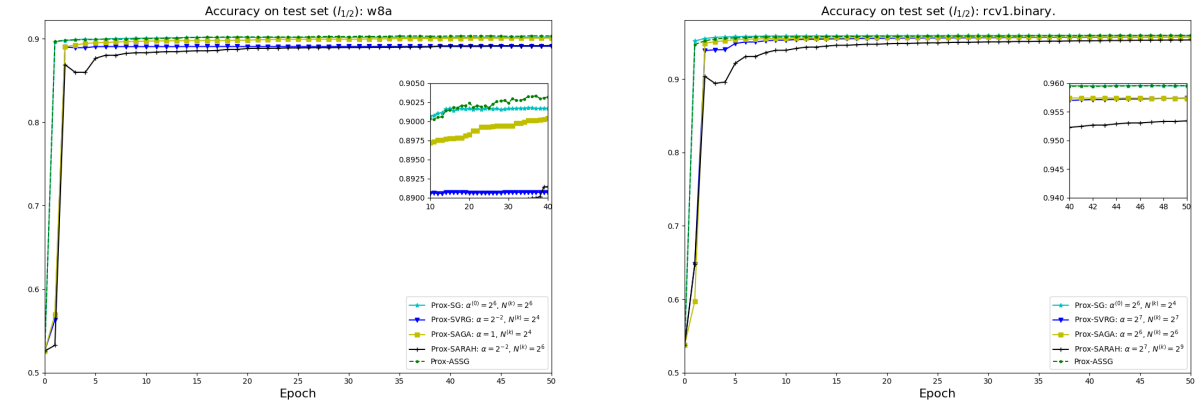

FIGURE 11. The accuracy of algorithms change with epochs on test set ($l_{1/2}$).

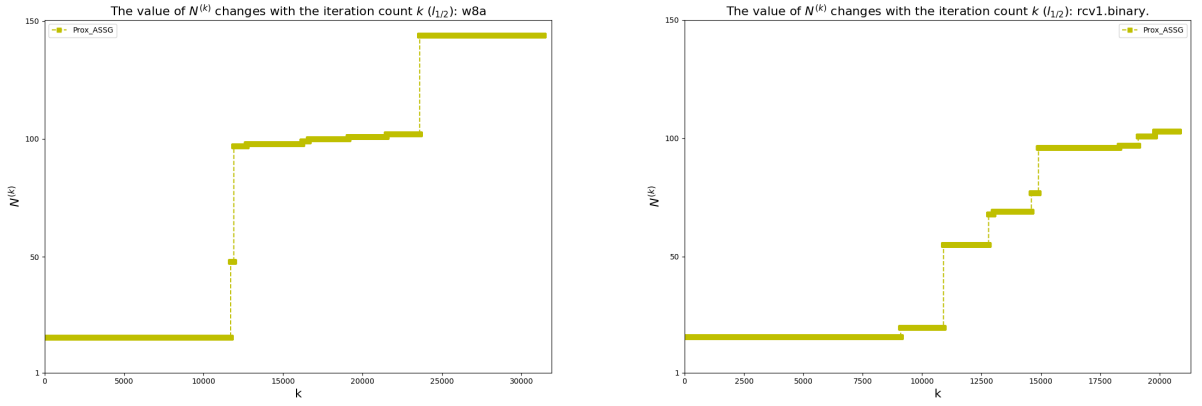

FIGURE 12. The sampling size $N^{(k)}$ change with the number of iterations k ($1_{1/2}$).

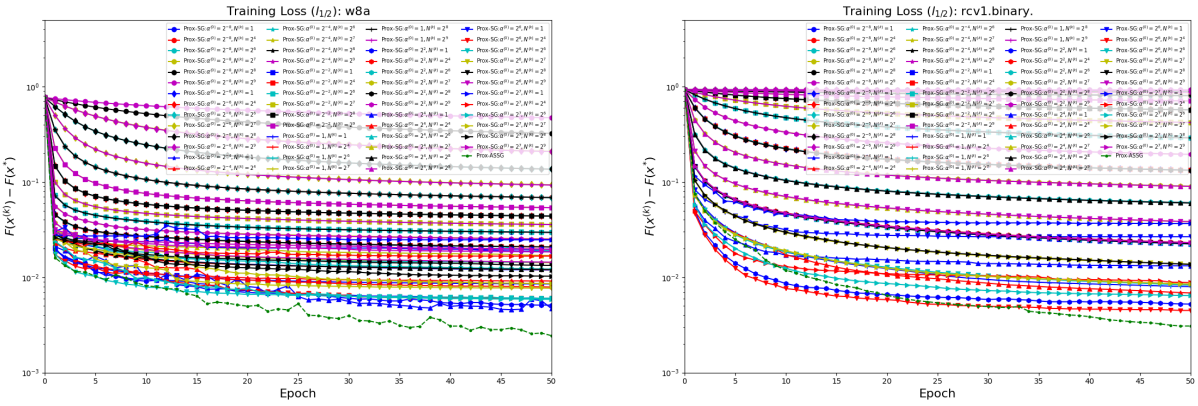

FIGURE 13. The step size and sampling size effects on Prox-SG algorithm ($l_{1/2}$).

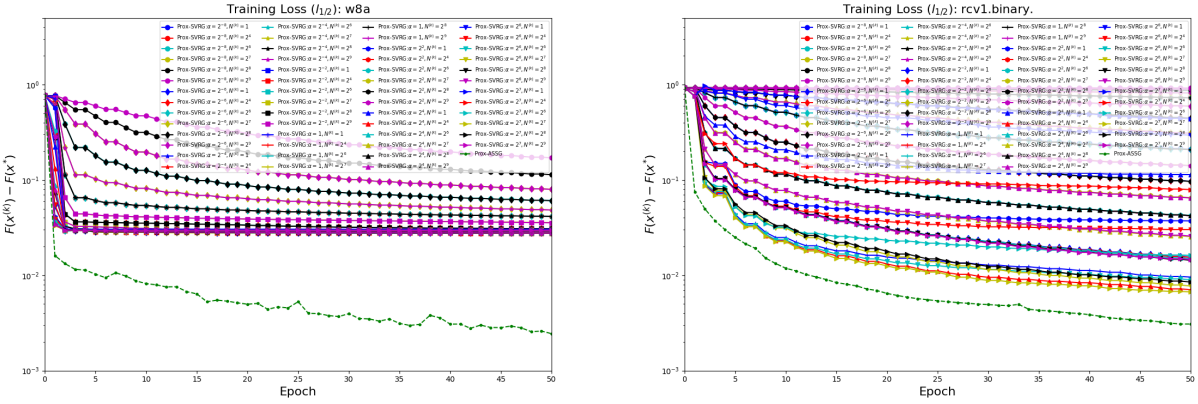

FIGURE 14. The step size and sampling size effects on Prox-SVRG algorithm ($l_{1/2}$).

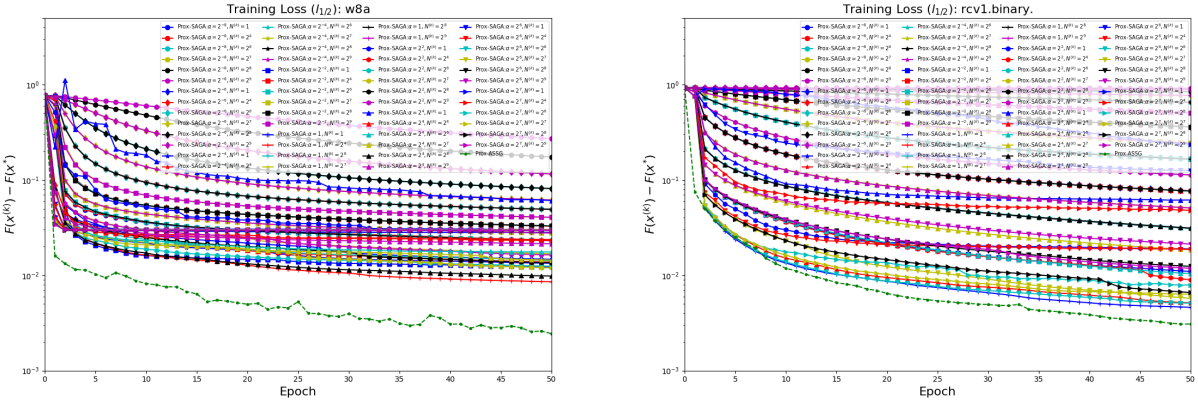

FIGURE 15. The step size and sampling size effects on Prox-SAGA algorithm ($l_{1/2}$).

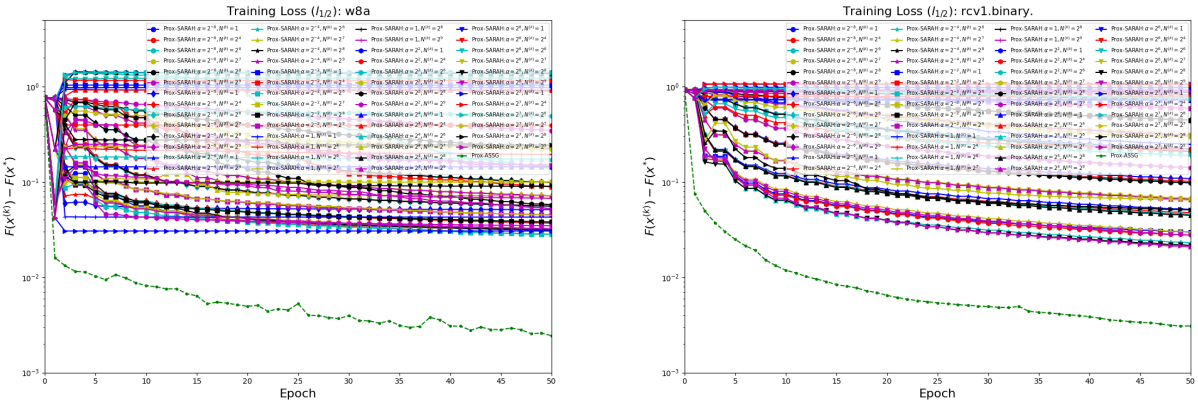

FIGURE 16. The step size and sampling size effects on Prox-SARAH algorithm ($l_{1/2}$).

5.1. **Numerical experiment results.** Figures 1-16 display the results of the various methods applied to each dataset. These figures indicate that the algorithm we proposed demonstrates promising and competitive performance across all test cases. Subsequently, we offer an in-depth analysis based on the findings of our numerical experiments:

- In fact, the Prox-ASSG algorithm employs an adaptive sampling strategy, which ensures a more suitable sampling size. As a result, the Prox-ASSG algorithm not only matches the computational speed of comparative algorithms but also outperforms them in certain scenarios. This is observed even when the comparative algorithms are operating with the optimal step size and sampling size, as illustrated in Figures 1-3 and 9-11. Although the algorithm may initially lag due to an inappropriate current sampling size, it is designed to catch up during later iterations through the adjustments made by the adaptive sampling strategy, as depicted in Figures 2 and 10.
- Secondly, it is important to note that the optimal sampling size and step size for various algorithms differ across different datasets. This variability is observed even for the same algorithm when applied to different datasets, as shown in Tables 3 and 4, and Figures 5-8 and 13-16. In light of this, we have adopted the strategy proposed in references [10,44], which posits that the step size should increase linearly with the sampling size, yet remain bounded.
- Lastly, Figures 4 and 12 demonstrate the efficacy of our proposed adaptive sampling strategy in mitigating the rapid growth of $N^{(k)}$.

6. CONCLUSIONS

In existing literature, most methods aimed at reducing the noise of stochastic gradients and accelerating convergence for non-convex and non-smooth composite problems are based on frameworks such as SVRG, SAGA, and SARAH, which typically use a fixed sampling size. This paper introduced an innovative adaptive sampling strategy and provides a detailed analysis of its application in the proximal stochastic gradient algorithm. We demonstrated that the algorithm can converge to critical points, and in functions that satisfy the KL property, for the exponent $\theta \in [0,1)$, we studied the convergence speed under different values of θ and found

that, under optimal conditions, the convergence speed is linear. This method not only has excellent generalization performance and robustness to parameter changes but also exhibits rapid convergence characteristics. The experimental results clearly demonstrate the superiority of our method.

However, it should also be noted that, at present, we only established the convergence properties of the objective function under the KL exponent. This is primarily because we need to utilize the property of the "exponent of block-separable sum for KL functions" [41, Theorem 3.3], which is currently only applicable to functions of the KL exponent type. Therefore, how to achieve global convergence under the condition that the objective function only satisfies the KL property is a topic worthy of further exploration. Additionally, we believe that the proposed adaptive sampling strategy is also worth further investigation in the context of accelerating stochastic algorithms.

Acknowledgments

The author is grateful to the reviewers for useful suggestions which improved the contents of this paper. The second author was supported by the National Natural Science Foundation of China (No. 11971078), Graduate Research and Innovation Foundation of Chongqing, China (Grant No. CYB23009).

REFERENCES

[1] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction, Springer, New York, 2009.

[2] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Rev. 60 (2018), 223–311.

[3] R. Chartrand, V. Staneva, Restricted isometry properties and nonconvex compressive sensing, Inverse Probl. 24 (2008), 035020.

[4] H. Wang, R. Li, C.L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007), 553–568.

[5] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Stat. 38 (2010), 894–942.

[6] Z. Xu, X. Chang, F. Xu, H. Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, IEEE Trans. Neural Networks Learn. Syst. 23(2012), 1013–1027.

[7] S. Zhou, X. Xiu, Y. Wang, D. Peng, Revisiting l_q $(0 \leq q < 1)$ norm regularized optimization, 2023, arXiv:2306.14394.

[8] H. Robbins, D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, in: Optimizing methods in statistics, pp. 233–257, Elsevier, 1971.

[9] S. McCandlish, J. Kaplan, D. Amodei, O.D. Team, An empirical model of large-batch training, 2018, arXiv:1812.06162.

[10] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch SGD: Training imagenet in 1 hour, 2017, arXiv:1706.02677.

[11] R.H. Byrd, G.M. Chin, J. Nocedal, Y. Wu, Sample size selection in optimization methods for machine learning, Math. Program. 134 (2012), 127–155.

[12] R. Bollapragada, R.H. Byrd, J. Nocedal, Adaptive sampling strategies for stochastic optimization, SIAM J. Optim. 28 (2018), 3312–3343.

[13] X. Qin, A weakly convergent method for splitting problems with nonexnpansive mappings, J. Nonlinear Convex Anal. 24 (2023), 1033-1043.

[14] A. Moudafi, Proximal linear methods for DC composite minimization problems, J. Appl. Numer. Optim. 5 (2023), 391-398.

[15] C. Paquette, K. Scheinberg, A stochastic line search method with expected complexity analysis, SIAM J. Optim. 30 (2020), 349-376.
[16] R. Bollapragada, S.M. Wild, Adaptive sampling quasi-Newton methods for zeroth-order stochastic optimization, Math. Program. Comput. 15 (2023), 327-364.
[17] Y. Xie, R. Bollapragada, R.H. Byrd, J. Nocedal, Constrained and composite optimization via adaptive sampling methods, IMA J. Numer. Anal. 44 (2024), 680-709.
[18] F. Beiser, B. Keith, S. Urbainczyk, B. Wohlmuth, Adaptive sampling strategies for risk-averse stochastic optimization with constraints, IMA J. Numer. Anal. 43 (2023), 3729-3765.
[19] G. Franchini, F. Porta, V. Ruggiero, I. Trombini, A line search based proximal stochastic gradient algorithm with dynamical variance reduction, J. Sci. Comput. 94 (2023), 23.
[20] M. Zhang, S. Li, A proximal stochastic quasi-Newton algorithm with dynamical sampling and stochastic line search, J. Sci. Comput. 102 (2025), 102:23.
[21] Y. Xu, R. Jin, T. Yang, Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems, NeurIPS 32 (2019).
[22] D. Driggs, J. Tang, J. Liang, M. Davies, C.B. Schonlieb, A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization, SIAM J. Imaging Sci. 14 (2021), 1932–1970.
[23] F. Bian, J. Liang, X. Zhang, A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization, Inverse Probl. 37 (2021), 075009.
[24] J. Guo, X. Wang, X. Xiao, Preconditioned primal-dual gradient methods for nonconvex composite and finite-sum optimization, 2023, arXiv:2309.13416.
[25] Q. Wang, D. Han, A Bregman stochastic method for nonconvex nonsmooth problem beyond global Lipschitz gradient continuity, Optim. Method. Softw. 38 (2023), 914–946.
[26] Q. Wang, Z. Liu, C. Cui, D. Han, A Bregman proximal stochastic gradient method with extrapolation for nonconvex nonsmooth problems, AAAI 38 (2024), 15580–15588.
[27] D. Driggs, J. Liang, C.B. Schönlieb, On biased stochastic gradient estimation, JMLR 23 (2022), 1–43.
[28] N. Roux, M. Schmidt, F. Bach, A stochastic gradient method with an exponential convergence rate for finite training sets, NeurIPS 25 (2012).
[29] M. Schmidt, N. Le Roux, F. Bach, Minimizing finite sums with the stochastic average gradient, Math. Program. 162 (2017), 83–112.
[30] A. Defazio, F. Bach, S. Lacoste-Julien, SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives, NeurIPS 27 (2014).
[31] S.J. Reddi, S. Sra, B. Poczos, and A.J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. NeurIPS 29 (2016).
[32] R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, NeurIPS 26 (2013).
[33] L. Xiao, T. Zhang, A proximal stochastic gradient method with progressive variance reduction, SIAM J. Optim. 24 (2014), 2057–2075.
[34] L. M. Nguyen, J. Liu, K. Scheinberg, M. Takáč, SARAH: A novel method for machine learning problems using stochastic recursive gradient, ICML, 2017, 2613–2621.
[35] N.H. Pham, L.M. Nguyen, D.T. Phan, Q. Tran-Dinh, ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization, J. Mach. Learn. Res. 21 (2020), 1–48.
[36] S. Ghadimi, G. Lan, H. Zhang, Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, Math. Program. 155 (2016), 267—305.
[37] M. Metel, A. Takeda, Simple stochastic gradient methods for non-smooth non-convex regularized optimization, In: Proceedings of the 36 th International Conference on Machine Learning, 2019, 4537-4545.
[38] R.T. Rockafellar, R.J.B. Wets, Variational analysis, Vol. 317, Springer Science & Business Media, America, 2009.
[39] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (2014), 459–494.
[40] P. Frankel, G. Garrigos, J. Peypouquet, Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates, J. Optim. Theory Appl. 165 (2015), 874–900.

[41] G. Li, T.K. Pong, Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods, Found. Comput. Math. 18 (2018), 1199–1232.
[42] I. Miller, M. Miller, J.E. Freund, John E. Freund's Mathematical Statistics with Applications. 8th edition, Pearson Education Limited, America, 2014.
[43] S. Ghadimi, G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework, SIAM J. Optim. 22 (2012), 1469–1492.
[44] S. Smith, E. Elsen, S. De, On the generalization benefit of noise in stochastic gradient descent, In: ICML'20: Proceedings of the 37th International Conference on Machine Learning, 2020, 9058–9067.

7. Appendix

7.1. **Variance of the sample mean in finite samples.**

Lemma 7.1. *[42, p.240] If $\overline{X}$ is the mean of a random sample of size n taken without replacement from a finite population of size N with the mean μ and the variance* $\mathrm{var}(X)$*, then*

$$\mathbb{E}(\overline{X}) = \mu \quad \text{and} \quad \mathrm{var}(\overline{X}) = \frac{\mathrm{var}(X)}{n} \cdot \frac{N-n}{N-1}.$$

7.2. **Proof of Corollary 4.1.**

Proof. From (4.6), we obtain

$$\mathbb{E}\left[\mathrm{dist}\left(0, \partial F\left(x^{(R)}\right)\right)\right] \leq \frac{\tau}{\sqrt{K}} \leq \varepsilon, \tag{7.1}$$

where

$$\tau := \sqrt{\left(4(L^2+\beta^2) + \frac{2}{\alpha^2}\right) \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right] + 4\sum_{k=0}^{K-1} \varepsilon^{(k)}}.$$

Therefore, in order to show that inequality (7.1) holds, it is necessary that $K \geq \frac{\tau^2}{\varepsilon^2}$. $\varepsilon^{(k)}$ could be defined as $\frac{C}{(k+1)^{1+\kappa_1}}$, where C and κ_1 are positive constants. This definition ensures that the series $\sum_{k=0}^{+\infty} \varepsilon^{(k)}$ converges to a finite value and that $\varepsilon^{(k)}$ approaches zero as k tends to infinity. Concurrently, based on equation (3.5), it can be deduced that

$$\mathbb{E}_k\left[\left\|\nabla f^{(k)}_{\mathcal{N}^{(k)}} - \nabla f^{(k)}\right\|^2\right] \leq \frac{C}{(k+1)^{1+\kappa_1}} + \beta^2 \mathbb{E}_k\left[\left\|x^{(k+1)} - x^{(k)}\right\|^2\right]$$

is valid under the condition that the sampling size is given by

$$N^{(k)} = \lceil \frac{\delta^2 (k+1)^{1+\kappa_1}}{C} \rceil.$$

Upon completion of iteration k, the total number of gradient computations is approximately

$$\begin{aligned}\sum_{k=0}^{K-1} N^{(k)} = \sum_{k=0}^{\lceil \frac{\tau^2}{\varepsilon^2} \rceil} \lceil \frac{\delta^2 (k+1)^{1+\kappa_1}}{C} \rceil &\leq \sum_{k=1}^{\lceil \frac{\tau^2}{\varepsilon^2} \rceil + 1} \frac{\delta^2 k^{1+\kappa_1}}{C} + 1 \\ &\leq \int_1^{\frac{\tau^2}{\varepsilon^2}+2} \frac{\delta^2}{C} x^{1+\kappa_1} + 1 \, dx \\ &\leq \frac{\delta^2}{C} \frac{\left(\frac{\tau^2 + 2\varepsilon^2}{\varepsilon^2}\right)^{2+\kappa_1}}{2+\kappa_1} + \frac{\tau}{\varepsilon}.\end{aligned}$$

Consequently, the computational complexity to reach an ε-approximate critical point is $\mathscr{O}\left(\frac{1}{\varepsilon^{4+\kappa}}\right)$. □

7.3. **Proof of Lemma 4.4.** The proof for Lemma 4.4 *iv-vii* is primarily based on Lemma 4.3 from reference [22].

Proof. For Lemma 4.4 *i-iii*, the results have been established in Lemma 4.3 and Theorem 4.1.

To prove Lemma 4.4 part *iv*, we start by assuming that $x^\star$ is a limit point of $\{x^{(k)}\}_{k\geq 0}$. The presence of such a limit point is warranted by our presupposition that the sequence $\{x^{(k)}\}_{k\geq 0}$ is bounded. Therefore, there exists a subsequence $\{x^{(k_q)}\}_{q\in\mathbb{N}}$ that converges to $x^\star$, as demonstrated by the limit:

$$\lim_{q\to\infty} x^{(k_q)} \to x^\star. \tag{7.2}$$

Since h is lower semicontinuous, then

$$\liminf_{q\to\infty} h\left(x^{(k_q)}\right) \geq h(x^\star). \tag{7.3}$$

By the updata rule for $x^{(k+1)}$ (2.2). Letting $z = x^\star$, we obtain

$$\begin{aligned}&\left\langle \nabla f^{(k)}_{\mathscr{N}^{(k)}}, x^{(k+1)} - x^{(k)} \right\rangle + \frac{1}{2\alpha} \left\| x^{(k+1)} - x^{(k)} \right\|^2 + h^{(k+1)} \\ &\leq \left\langle \nabla f^{(k)}_{\mathscr{N}^{(k)}}, x^\star - x^{(k)} \right\rangle + \frac{1}{2\alpha} \left\| x^\star - x^{(k)} \right\|^2 + h^\star,\end{aligned}$$

i.e.,

$$h^{(k+1)} \leq \left\langle \nabla f^{(k)}_{\mathscr{N}^{(k)}}, x^\star - x^{(k+1)} \right\rangle + \frac{1}{2\alpha} \left\| x^\star - x^{(k)} \right\|^2 - \frac{1}{2\alpha} \left\| x^{(k+1)} - x^{(k)} \right\|^2 + h^\star.$$

Setting $k = k_q$, and taking the limit $q \to \infty$, one has

$$\begin{aligned}\limsup_{q\to\infty} h^{(k_q+1)} \leq &\limsup_{q\to\infty} \left\langle \nabla f^{(k_q)}, x^\star - x^{(k_q+1)} \right\rangle + \left\langle \nabla f^{(k_q)}_{\mathscr{N}^{(k_q)}} - \nabla f^{(k_q)}, x^\star - x^{(k_q+1)} \right\rangle \\ &+ \frac{1}{2\alpha} \left\| x^\star - x^{(k_q)} \right\|^2 - \frac{1}{2\alpha} \left\| x^{(k_q+1)} - x^{(k_q)} \right\|^2 + h^\star.\end{aligned}$$

As $x^{(k_q)} \to x^\star$ (7.2), we can infer from the gradient's Lipschitz continuity that $\nabla f^{(k_q)}$ is bounded. Moreover, according to Lemma 4.4 *iii*, it follows that $x^{(k_q+1)} - x^{(k_q)} \to 0$ a.s.. Additionally, based on inequalities (3.1), $\varepsilon^{(k)} \to 0$ and $\left\| x^{(k+1)} - x^{(k)} \right\| \to 0$ a.s., we can deduce that $\nabla f^{(k_q)}_{\mathscr{N}^{(k_q)}} -$

$\nabla f^{(k_q)} \to 0$ a.s.. Therefore, we can obtain $\limsup_{q\to\infty} h^{(k_q+1)} \leq h^\star$ a.s.. Together with equation (7.3), this implies that $h^{(k_q+1)} \to h^\star$ a.s.. Recalling that f is continuous, we can immediately conclude that

$$\lim_{q\to\infty} F\left(x^{(k_q)}\right) = \lim_{q\to\infty} f\left(x^{(k_q)}\right) + h\left(x^{(k_q)}\right) = F\left(x^\star\right) \ a.s..$$

Together with Lemma 4.4 *iii* and the fact that the subdifferential of F is closed, we have $\mathbb{E}\left[\operatorname{dist}\left(0, \partial F\left(x^\star\right)\right)\right] = 0$.

Lemma 4.4 *v* and *vi* hold for any sequence satisfying $\left\|x^{(k+1)} - x^{(k)}\right\| \to 0$ a.s. [39, Remark 5] and [22, Lemma 4.3].

Finally, we must show that F has constant expectation over ω. According to Lemma 4.4 *ii*, we deduce that $\mathbb{E}\left[F\left(x^{(k)}\right)\right] \to F_\star$, which further suggests that $\mathbb{E}\left[F\left(x^{(k_q)}\right)\right] \to F_\star$ for every subsequence $\left\{x^{(k_q)}\right\}_{q\in\mathbb{N}}$ converging to some $x^\star \in \omega$. During the proof of Lemma 4.4 *iv*, it is demonstrated that $F\left(x^{(k_q)}\right) \to F\left(x^\star\right)$ a.s., leading to the conclusion that $\mathbb{E}\left[F\left(x^\star\right)\right] = F_\star$ for all $x^\star \in \omega$. □

7.4. **Proof of Corollary 4.2.**

Proof. From equation (4.10), we deduce that the number of iterations k required to ensure

$$\mathbb{E}\left[F^{(k)} - F^\star\right] \leq C_2 \bar{\rho}^k \leq \varepsilon,$$

where $\bar{\rho} = \max\left\{\left(\frac{1}{\nu} + c_{2,1}\right)\frac{1}{1+c_{2,1}}, \frac{1}{1+\kappa_2}\right\}$. Therefore, the condition

$$k \geq \frac{\log(\varepsilon^{-1}) + \log(\kappa_1)}{\log(\bar{\rho}^{-1})}$$

must be satisfied. If $N^{(k)}$ grows in accordance with $N^{(k)} = \lceil \frac{\delta^2}{\kappa_1}(1+\kappa_2)^k \rceil$, then

$$\varepsilon^{(k)} \overset{(3.5)}{\leq} \frac{\kappa_1}{(1+\kappa_2)^k}$$

must hold true. Upon completion of iteration k, the total number of gradient computations is approximately

$$\sum_{j=0}^{k} N^{(j)} = \sum_{j=0}^{k} \frac{\delta^2}{\kappa_1}(1+\kappa_2)^j = \frac{\delta^2}{\kappa_1}\frac{(1+\kappa_2)^{k+1} - 1}{(1+\kappa_2) - 1}.$$

Now, let us set k to $\frac{\log(\varepsilon^{-1})+\log(c_1)}{\log(\bar{\rho}^{-1})}$, and then

$$\begin{aligned}(1+\kappa_2)^k &= (1+\kappa_2)^{\frac{\log(\varepsilon^{-1})+\log(C_2)}{\log(\bar{\rho}^{-1})}} \\ &= \exp\left(\frac{\log(\varepsilon^{-1}) + \log(C_2)}{\log(\bar{\rho}^{-1})}\log(1+\kappa_2)\right) \\ &= \left(\frac{C_2}{\varepsilon}\right)^{\frac{\log((1+\kappa_2)^{-1})}{\log(\bar{\rho})}}.\end{aligned}$$

In particular, if we let $\frac{1}{1+\kappa_2} = \left(\frac{1}{\nu} + c_{2,1}\right)\frac{1}{1+c_{2,1}}$, then

$$\sum_{j=0}^{k} N^{(j)} = \frac{\delta^2}{C}\frac{\frac{C_2}{\varepsilon}(1+\kappa_2)-1}{\kappa_2}.$$

Consequently, the computational complexity to reach an ε-solution is $\mathscr{O}\left(\frac{1}{\varepsilon}\right)$. □