On the Convergence of Stochastic Splitting Methods for Nonsmooth Nonconvex Optimization

Jia Hu\textsuperscript{a}, Congying Han\textsuperscript{a,b}, Tiande Guo\textsuperscript{a,b}, and Tong Zhao\textsuperscript{a,b}

\textsuperscript{a}School of Mathematical Sciences, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, P.R.China
\textsuperscript{b}Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, No.80 Zhongguancun East Road, Beijing 100190, P.R.China

Abstract

With the growth of data scale, stochastic optimization algorithms have been studied for convex and nonconvex objective functions in recent years. In this paper, we focus on the global convergence for two types of stochastic splitting algorithms, i.e., alternating direction method of multipliers (ADMM) and proximal gradient descent (PGD), for a class of nonsmooth nonconvex composite optimization problems. More specifically, we propose two stochastic algorithms (i.e., SARAH-ADMM and SARAH-PGD) with single loop structure and variance reduction technique. Based on the Kurdyka-Lojasiewicz (KL) assumption, which is satisfied for a wide range of functions including convex and nonconvex functions, we analyze the asymptotic convergence and corresponding rates of the SARAH-ADMM and SARAH-PGD. Finally, we report some numerical experiments to demonstrate the promising performance of the proposed algorithms.

Keywords: Nonsmooth nonconvex optimization; Stochastic approximation; Alternating direction method of multipliers; Proximal gradient descent; Variance reduction

1 Introduction

Mathematical optimization methodology plays a vital role in the development of many applications such as engineering and machine learning. However, the scale of sample data is very large in modern environment, which makes traditional optimization methods no longer applicable. To tackle this problem, a surge of stochastic variants of optimization algorithms including first order and

*Corresponding author: hancy@ucas.ac.cn
second order, such as gradient descent (GD), proximal gradient descent (PGD), alternating direction method of multipliers (ADMM), quasi Newton method, Newton’s method, etc, have been developed. In this paper, we focus on stochastic approaches on two types of splitting methods, i.e., PGD and ADMM (for an appetizer, see for instance [29] and references therein) for solving following nonsmooth nonconvex optimization problem:

\[
\min_{x} \frac{1}{N} \sum_{i=1}^{N} f_i(x) + g(Ax),
\]

where \( N \) is the number of components, \( f_i : \mathbb{R}^m \to \mathbb{R}, i = 1, 2, \ldots, N \) are continuously differentiable (not necessarily convex), \( g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is a possibly nonsmooth and nonconvex function, and \( A \in \mathbb{R}^{n \times m} \) is a matrix. Such a model problem encompasses many important applications in image processing [39], machine learning and artificial intelligence [44], and engineering. In particular, many learning problems are in the form of (1) with each \( f_i \) being a loss function on single data and \( g \) being a regularizer. For instance, when \( A \) is the identity matrix, \( \ell_1 \) and \( \ell_2 \) norm are commonly used convex regularizers while \( \ell_0 \) and \( \ell_{1/2} \) norm are popular nonconvex regularizers, which all promote some sparse properties. In addition, since \( g \) is permitted to take \(+\infty\), some complicated constraint sets are allowed in our model problem (1). There are also some cases where \( A \) is not identity matrix; for example, total variation (TV) regularization [49], generalized lasso regularization [47], etc.

When \( A \) is the identity matrix, the traditional PGD algorithm [3, 7, 8, 15] (also known as the forward-backward splitting algorithm) can be applied without considering the particular finite-sum structure of the problem. But now the number of components \( N \) may be large, so that traditional PGD, in which \( N \) gradients need to be computed at each iteration, is not applicable. Therefore, by using the idea of stochastic approximation and variance reduced techniques [10, 13, 26, 37, 41, 43], a number of stochastic PGD algorithms have been proposed in recent years; see for instance [17,18,24,30,40,50,51]. Note that there are some restrictions for these stochastic variants of PGD: both the smooth and the nonsmooth parts must be convex or the smooth part is nonconvex whereas the nonsmooth part should be convex in the objective function. In contrast, both parts in our model problem (1) can be nonconvex and we can derive the global convergence of the whole sequence generated by the algorithm. Such a significant development is attributed to the well-known Kurdyka-Lojasiewicz (KL) property [5,27,33]. For an in-depth study of the class of KL functions, one can refer to [6], as well as references therein. Additionally, the convergence analysis of deterministic proximal-type algorithm under the KL property is explored in [1–3,7,8,29]. On the other hand, when \( A \) is a general matrix, the proximal step of PGD is not necessarily simple. In this case, by introducing variable \( y \) and letting \( y = Ax \) be constraints, the ADMM [16,19], originally introduced in the early 1970s, is an appropriate approach and it has been extensively researched for convex and nonconvex problems; see, for example, [20–22,25,29,31,34,48]. Similar to the previous discussion, traditional ADMM is costly due to the large
number of components of problem (1). Thus, a series of stochastic ADMM algorithms are developed for the last few years; see, for instance, [23,32,45,46,53,54]. However, at least one part of the composite optimization problem of the form (1) in these literature on stochastic ADMM is assumed to be convex. In this paper, by taking advantage of KL property, we explore the convergence of a stochastic recursive ADMM algorithm (called SARAH-ADMM) when both two parts are possibly nonconvex.

To end this section, we summarize our contributions as follows. First, we propose two stochastic algorithms, named SARAH-ADMM and SARAH-PGD, both with single loop structure and variance reduction technique. Such two single loop algorithmic frameworks provide great convenience for convergence analysis, in contrast, the analysis of some existing stochastic splitting algorithms with two loop for nonconvex problem is tedious. Moreover, the numerical experiments show the effectiveness of our SARAH-ADMM and SARAH-PGD. More specifically, compare to existing stochastic algorithms whose outputs are chosen uniformly from all history iterates, our computational results are more stable. Second, two parts of the model problem (1) we consider are both possibly nonconvex, which is a large difference from the existing related works and hence covers more practical problems arising in a wide range of applications such as statistics and machine learning. Finally, in theory, we show the proposed algorithms (i.e., SARAH-ADMM and SARAH-PGD) both converge globally to some stationary point almost surely under the assumption that the associated function satisfies the KL property. Moreover, we obtain corresponding asymptotic convergence rates.

The remaining parts of this paper are organized as follows. In Sect. 2, we present some notations and preliminaries. The Sect. 3 is divided into four parts. We show our algorithms, i.e., SARAH-ADMM in Subsect. 3.1 and SARAH-PGD in Subsect. 3.2, and analyze the convergence of these two algorithms in Subsect 3.3 and 3.4 respectively. In Sect. 4, we present some numerical experiments to demonstrate the promising efficiency of the proposed algorithms. Finally, some conclusions are made in Sect. 5.

2 Notations and Preliminaries

In this section, we first introduce some notations, which would be used throughout the paper, and then recall some definitions and properties, following those in [7].

**Notation 1.** The gradient of $f$ is denoted as $\nabla f$. $\partial g$ is the subdifferential of $g$. Given a matrix $G$, $\lambda_{\text{max}}(G)$ and $\lambda_{\text{min}}(G)$ denote its maximal and minimal eigenvalue, respectively. We use $\langle \cdot, \cdot \rangle$ to denote the inner product in a finite dimensional Euclidean space and $\|\cdot\|$ to denote the norm induced by the inner product. Given a vector $x$ in a finite dimensional Euclidean space, $\|x\|_G$ is its matrix norm calculated by $\langle Gx, x \rangle^{\frac{1}{2}}$, where $G$ is a positive semidefinite matrix. The closure of a set $C$ is denoted as $\overline{C}$. For a matrix $A$, $A^\ast$ is the transposed matrix of
A. Almost surely is abbreviated as a.s., $\mathbb{E}$ denotes the mathematical expectation, and $\mathbb{E}_k$ denotes conditional expectation on the first $k$ iterations.

Let $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be a point-to-set mapping, its graph is defined by

$$\text{Graph } F := \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n : y \in F(x)\}.$$ 

For any subset $S \subset \mathbb{R}^m$ and any point $x \in \mathbb{R}^m$, the distance from $x$ to $S$ is defined and denoted by

$$\text{dist } (x, S) := \inf \{\|y - x\| : y \in S\}.$$ 

When $S = \emptyset$, we have that $\text{dist } (x, S) = +\infty$ for all $x$.

Let us recall a few definitions concerning subdifferential calculus [35, 42].

Recall that for $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a proper lower semicontinuous function, the domain of $g$ is defined through

$$\text{dom } g := \{x \in \mathbb{R}^n : g(x) < +\infty\}.$$ 

**Definition 1.** Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

(i) For a given $x \in \text{dom } g$, the Fréchet subdifferential of $g$ at $x$, written $\hat{\partial}g(x)$, is the set of all vectors $x^* \in \mathbb{R}^n$ which satisfy

$$\liminf_{y \to x, y \neq x} \frac{g(y) - g(x) - (x^*, y - x)}{\|y - x\|} \geq 0.$$ 

When $x \notin \text{dom } g$, we set $\hat{\partial}g(x) = \emptyset$.

(ii) The limiting-subdifferential, or simply the subdifferential, of $g$ at $x \in \mathbb{R}^n$, written $\partial g (x)$, is defined as follows:

$$\partial g (x) := \{x^* \in \mathbb{R}^n : \exists x_k \to x, g(x_k) \to g(x), x^*_k \in \hat{\partial}g(x_k) \to x^* \}.$$ 

From this definition, we can obtain: (i) $\hat{\partial}g(x) \subset \partial g(x)$ for each $x \in \mathbb{R}^n$, where the first set is closed and convex while the second one is closed; (ii) Let $\{(x_k, x^*_k)\}_{k \in \mathbb{N}}$ be a sequence in Graph $\partial g$ that converges to $(x, x^*)$. By the definition of $\partial g (x)$, if $g(x_k)$ converges to $g(x)$, then $(x, x^*) \in \text{Graph } \partial g$; (iii) If $x \in \mathbb{R}^n$ is a local minimizer of $g$, then $0 \in \partial g(x)$.

Points whose subdifferential contains 0 are called (limiting-) critical points and the set of critical points of $g$ is denoted by $\text{crit } g$. Also, the domain of $\partial g$ is defined as

$$\text{dom } \partial g := \{x \in \mathbb{R}^n : \partial g(x) \neq \emptyset\}.$$ 

Next let us recall an important property of subdifferentiability.
Proposition 2.1. Assume that \( H(x, y) = f(x) + g(y) \), where \( f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\} \) and \( g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) are proper lower semicontinuous functions. Then for all \((x, y) \in \text{dom } H = \text{dom } f \times \text{dom } g\), we have
\[
\partial H(x, y) = \partial f(x) \times \partial g(y).
\]

Definition 2. (Kurdyka-Lojasiewicz (KL) property) \([5, 27]\) Let \( g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) be a proper lower semicontinuous function. For given real numbers \( a \) and \( b \), set \([a < g < b] := \{x \in \mathbb{R}^n : a < g(x) < b\}\). The function \( g \) is said to have the Kurdyka-Lojasiewicz property at \( x^* \in \text{dom } \partial g \), if there exist \( \eta \in (0, +\infty] \), a neighborhood \( U \) of \( x^* \) and a continuous concave function \( \varphi : [0, \eta) \to \mathbb{R}_+ \), such that
\[
\begin{align*}
(i) \quad & \varphi(0) = 0, \\
(ii) \quad & \varphi \text{ is } C^1 \text{ on } (0, \eta), \\
(iii) \quad & \forall s \in (0, \eta), \varphi'(s) > 0, \\
(iv) \quad & \text{for all } x \in U \cap [g(x^*) < g < g(x^*) + \eta], \text{ the Kurdyka-Lojasiewicz inequality holds:} \\
& \varphi'(g(x) - g(x^*)) \text{ dist } (0, \partial g(x)) \geq 1.
\end{align*}
\]

Definition 3. If \( g \) satisfies the KL property at each point of \( \text{dom } \partial g \), then \( g \) is called a KL function.

There are many functions are KL function, such as semi-algebraic, subanalytic and log-exp (see \([2, 7]\) and references therein), which implies its importance. Also, we denote \( f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \) for convenience and hence \( \nabla f(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x) \).

To introduce the following definitions and describe the algorithm in Sect. 3, we reformulate (1) as
\[
\min_{x,y} \left\{ \frac{1}{N} \sum_{i=1}^{N} f_i(x) + g(y) \mid \text{s.t. } Ax - y = 0 \right\} \quad (2)
\]

to decouple the linear map and the nonsmooth part.

Definition 4. \((x^*, y^*, \lambda^*)\) is called a stationary point of problem (2), if it satisfies that
\[
\begin{align*}
A^* \lambda^* &= \nabla f(x^*) \\
-\lambda^* &= \partial g(y^*) \\
Ax^* - y^* &= 0.
\end{align*}
\]

For problem (1), \( x^* \) is called a stationary point if it satisfies that
\[
0 \in \nabla f(x^*) + \partial g(Ax^*).
\]
Indeed, the former definition about stationary point is basically a reformulation of the latter. Moreover, if both \( f \) and \( g \) are convex, the above definitions imply the global optimality.

Denote the augmented Lagrangian function for (2) by

\[
L_\beta (x, y, \lambda) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) + g(y) - \langle \lambda, Ax - y \rangle + \frac{\beta}{2} \|Ax - y\|^2,
\]

where \( \lambda \) is the Lagrangian multiplier associated with the linear constraints and \( \beta > 0 \) is the penalty parameter. Hence if \((x^*, y^*, \lambda^*)\) is a stationary point, then it is a critical point of \( L_\beta \), i.e., \((x^*, y^*, \lambda^*)\) ∈ \( \text{crit} \ L_\beta \).

We next proceed with the notion of kernel generating distance, following the notation in [8], which would be used in Sect. 3 later.

**Definition 5.** (kernel generating distance) Let \( C \) be a nonempty, convex, and open subset of \( \mathbb{R}^m \). Associated with \( C \), a function \( h : \mathbb{R}^m \to (-\infty, +\infty] \) is called a kernel generating distance if it satisfies the following:

(i) \( h \) is proper lower semicontinuous, and convex, with \( \text{dom} \ h \subset C \) and \( \text{dom} \ \partial h = C \).

(ii) \( h \) is \( C^1 \) on \( \text{int} \ \text{dom} \ h \equiv C \).

Denote the class of kernel generating distances by \( \mathcal{G}(C) \). Given \( h \in \mathcal{G}(C) \), define the proximity measure \( D_h : \text{dom} \ h \times \text{intdom} \ h \to \mathbb{R}_+ \) by

\[
D_h (x, y) := h(x) - [h(y) + \langle \nabla h(y), x - y \rangle].
\]

The proximity measure \( D_h \) is the so-called Bregman distance [9], which measures the proximity of \( x \) and \( y \) due to the following fact:

\[
D_h (x, y) \geq 0, \forall x \in \text{dom} \ h, \ y \in \text{int dom} \ h.
\]

If, in addition, \( h \) is strictly convex, equality holds if and only if \( x = y \). In general, \( D_h \) is not symmetric, unless \( h \) is the energy kernel \( h = \|\cdot\|^2 \), which corresponds to \( D_h (x, y) = \frac{1}{2} \|x - y\|^2 \). Moreover, the Bregman distance have following simple but remarkable properties:

- **The three-point identity.** For any \( y, z \in \text{int dom} \ h \) and \( x \in \text{dom} \ h \), we have
  \[
  D_h (x, z) - D_h (x, y) - D_h (y, z) = \langle \nabla h(y) - \nabla h(z), x - y \rangle.
  \]

- **Linear additivity.** For any \( \alpha, \beta \in \mathbb{R} \), and any function \( h_1 \) and \( h_2 \), we have
  \[
  D_{\alpha h_1 + \beta h_2} (x, y) = \alpha D_{h_1} (x, y) + \beta D_{h_2} (x, y)
  \]
  for all \( x, y \in \text{dom} \ h_1 \cap \text{dom} \ h_2 \) such that both \( h_1 \) and \( h_2 \) are differentiable at \( y \).
3 Stochastic Splitting Methods and Convergence Analysis

In this section, we first present our algorithms for solving problem (1) or (2), and then conduct the convergence analysis. The pseudocodes are outlined as Algorithm 1 and Algorithm 2, respectively, and both two algorithms are based on a stochastic recursive operator with the property of variance reduction—SARAH [28,37]. It uses biased estimates, and has the desirable property that the variance of stochastic gradients decreases to zero, and has better performance for designing other algorithms [40].

3.1 SARAH-ADMM

Algorithm 1 SARAH-ADMM

Let \((x_0, y_0, \lambda_0) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n, x_0 = x_{-1}, v_0 = v_{-1} = \nabla f(x_0); p \in (0, 1], \eta \) and \(\beta > 0\), and a symmetric positive semidefinite matrices \(G \in \mathbb{R}^{m \times m}\).

Output: \((x_k, y_k, \lambda_k)\)

for \(k = 0, 1, \ldots\) do

\(y_{k+1} = \arg\min_y L_{\beta}(x_k, y, \lambda_k)\)

\(v_k = \begin{cases} \nabla f(x_k), & \text{with probability } p \\ \frac{1}{M} \sum_{i \in M_k} (\nabla f_i(x_k) - \nabla f_i(x_{k-1})) + v_{k-1}, & \text{with probability } 1 - p \end{cases}\)

\(M_k\) is a mini-batch set picked uniformly at random from \(\{1, 2, \ldots, n\}\) (with replacement) such that \(|M_k| = M; x_{k+1} = \arg\min_x L_{\beta}^k(x, y_{k+1}, \lambda_k), \)

where

\[ L_{\beta}^k(x, y, \lambda) = f(x_k) + \langle v_k, x - x_k \rangle + \frac{1}{2\eta} \|x - x_k\|^2_G + g(y) - \langle \lambda, Ax - y \rangle + \frac{\beta}{2} \|Ax - y\|^2; \]

\(\lambda_{k+1} = \lambda_k - \beta \langle Ax_{k+1} - y_{k+1} \rangle;\)

If a termination criterion is not met, continue.

end for

We give some remarks here. First, the framework of SARAH-ADMM is similar to that of proximal ADMM [22], except that, in the subproblem with respect to \(x\), the full gradient \(\nabla f(x_k)\) is replaced by a stochastic gradient \(v_k\), but the analysis process is different. From our later analysis, we also show that \(v_k\) is of variance reduction, that is, the gap between \(v_k\) and \(\nabla f(x_k)\) grows progressively smaller in expectation. Second, compare our algorithm with several stochastic ADMM algorithms proposed in [23], the differences may be: (i) Our algorithm is of single loop structure, while theirs have two loops and the convergence analysis is much simplified due to this change; Of course, other variance reduction techniques [10, 26] can also be incorporated into our framework to
obtain corresponding single loop algorithms and the convergence analysis are similar. (ii) In SARAH-ADMM, the stochastic gradient $v_k$ is computed by a probability distribution, where the full gradient is computed with probability $p$, usually equal to $O(1/N)$ at each iteration. However, the full gradient is computed once fixed iterations in theirs; (iii) More importantly, the output is set as the last iterate while their outputs are chosen uniformly at random from all history iterates. Obviously, our algorithm’s output contains more information and is more stable, which can be demonstrated in later experiments. Third, similar to the approach in [38], it is possible to compute an approximation of $\nabla f(x_k)$ based on a minibatch set in the computation of $v_k$, and this may be our future work. Finally, the termination criterion is usually the maximum number of iterations or the maximum time, similarly hereinafter.

3.2 SARAH-PGD

Algorithm 2 SARAH-PGD

Let $x_0 \in \mathbb{R}^m$, $x_0 = x_{-1}$, $v_0 = v_{-1} = \nabla f(x_0)$; $p \in (0, 1]$, $\lambda > 0$

Output: $(x_k, y_k)$;

for $k = 0, 1, \ldots$ do

$v_k = \begin{cases} 
\nabla f(x_k), & \text{with probability } p \\
\frac{1}{M} \sum_{i \in M_k} (\nabla f_i(x_k) - \nabla f_i(x_{k-1})) + v_{k-1}, & \text{with probability } 1 - p 
\end{cases}

$M_k$ is a minibatch set picked uniformly at random from $\{1, 2, \ldots, n\}$ (with replacement) such that $|M_k| = M$;

$x_{k+1} = \arg\min_x \left\{ \langle v_k, x - x_k \rangle + \frac{1}{\lambda} D_h(x, x_k) + g(x) \right\}.$

If a termination criterion is not met, continue.

end for

Similarly, the main difference between SARAH-PGD and traditional proximal gradient algorithm in [8] is that access to the gradient. More specifically, our gradient is computed using a minibatch set with probability $1 - p$ not all $N$ components at each iteration. Since $N$ may be large, in this case, traditional proximal gradient algorithm is not applicable due to high computational cost. A recent work [11] developed similar ideas for analyzing the convergence of a stochastic proximal gradient algorithm applied to the nonconvex composite model problem (a slightly more general model than (1)). However, in their algorithm, only the classical squared Euclidean distance is considered. Recall that the energy kernel generates classical squared Euclidean distance in Sect. 2. Thus, if we choose another kernel, then we obtain a different distance; for more specific examples, one can see [4, 12] and references therein. As stated in the introduction, some complex constraint sets are permissible in our model problem. Consequently, our choice of different distances is beneficial in capturing the geometry of the constraints; for instance, the unit simplex constraint. For the sake of simplicity, we consider $C \equiv \mathbb{R}^m$ (cf. Definition 5) and $h$ is $\gamma$-strongly
convex on $\mathbb{R}^m$ for some $\gamma > 0$.

Before proceeding convergence analysis, we make the following assumptions for both algorithms, which are common in the available literature about nonsmooth nonconvex optimization.

**Assumption 1.** We assume:

(i) Each $f_i$ is continuously differentiable and its gradient is Lipschitz continuous with the modulus $L > 0$, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|$, $\forall x, y \in \mathbb{R}^m$;

(ii) $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper lower semicontinuous;

(iii) $AA^* \geq \sigma I$ for some $\sigma > 0$ ($\sigma = 1$ when $A$ is the identity matrix).

We give some remarks on Assumption 1. First, the condition in (i) can be relaxed to

$$\frac{1}{N} \sum_{i=1}^{N} \|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|^2, \forall x, y \in \mathbb{R}^m.$$ 

Second, $g$ is not necessarily convex and hence our model can encompass a wider range of functions. Finally, (iii) says that $A$ is of full row rank.

We begin the convergence analysis with a very useful lemma, which is crucial to derive an estimate for the decreasing amount of the function. Throughout the analysis, we assume that the sets of optimal solutions of minimization subproblems with respect to the function $g$ in Algorithm 1 and 2 are nonempty.

**Lemma 3.1.** [36] Let $f : \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable function with gradient $\nabla f$ assumed Lipschitz continuous with the modulus $L > 0$. Then

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

**Proof.** The proof of this lemma can be found on page 22 of [36].

### 3.3 Convergence Analysis of SARAH-ADMM

In this part, we assume the augmented Lagrangian function is bounded from below. We begin with some properties of the SARAH-ADMM and then obtain a key estimate for the sequence $\{\Phi(x_k, y_k, \lambda_k)\}$ (to be defined later), which will play an essential role in deriving our main convergence results. By the optimality conditions, we have:

$$\begin{align*}
0 &\in \partial g(y_{k+1}) + \lambda_k - \beta (Ax_k - y_{k+1}) \\
0 &\equiv v_k + \frac{1}{\eta} (x_{k+1} - x_k) - A^* \lambda_k + \beta A^* (Ax_{k+1} - y_{k+1}) \\
\lambda_{k+1} &\equiv \lambda_k - \beta (Ax_{k+1} - y_{k+1})
\end{align*}$$

and rearranging terms, we obtain:

$$- \lambda_k - \beta A (x_k - x_{k-1}) \in \partial g(y_k),$$

$$A^* \lambda_{k+1} - \frac{1}{\eta} G(x_{k+1} - x_k) = v_k,$$  \hspace{1cm} (3)

$$\lambda_{k+1} = \lambda_k - \beta (Ax_{k+1} - y_{k+1}).$$
Therefore by the proposition 1, the element
\[
\left( \begin{array}{c}
\nabla f (x_k) - v_k) - \frac{1}{n} G (x_{k+1} - x_k) + A^* (\lambda_{k+1} - \lambda_k) + A^* (\lambda_{k-1} - \lambda_k) \\
\lambda_k - \lambda_{k-1} - \beta A (x_k - x_{k-1}) / \beta
\end{array} \right)
\]
belongs to \( \partial L (x_k, y_k, \lambda_k) \).

The next few lemmas are in preparation for deriving a bound for the difference of \( \mathbb{E}_k [L (x_{k+1}, y_{k+1}, \lambda_{k+1})] \) and \( L (x_k, y_k, \lambda_k) \).

**Lemma 3.2.** Let \( \{(x_k, y_k, \lambda_k)\} \) be the sequence generated by the SARAH-ADMM, then we have
\[
\mathbb{E}_k \|v_k - \nabla f (x_k)\|^2 \leq (1 - p) \|v_{k-1} - \nabla f (x_{k-1})\|^2 + \frac{(1 - p) L^2}{M} \|x_k - x_{k-1}\|^2.
\]

**Proof.** We denote conditional expectation on the first \( k \) iterations and the second event, i.e., the one with probability \( 1 - p \) by \( \mathbb{E}_{k,s} \).

\[
\begin{align*}
\mathbb{E}_k \|v_k - \nabla f (x_k)\|^2 &= (1 - p) \mathbb{E}_{k,s} \|v_{k-1} - \nabla f (x_{k-1}) + \nabla f (x_{k-1}) - \nabla f (x_k) + v_k - v_{k-1}\|^2 \\
&= (1 - p) \left( \|v_{k-1} - \nabla f (x_{k-1})\|^2 + \|\nabla f (x_{k-1}) - \nabla f (x_k)\|^2 + \mathbb{E}_{k,s} \|v_k - v_{k-1}\|^2 \right) \\
&\quad + 2 (1 - p) \langle v_{k-1} - \nabla f (x_{k-1}), \nabla f (x_{k-1}) - \nabla f (x_k) \rangle \\
&\quad + 2 (1 - p) \langle v_{k-1} - \nabla f (x_{k-1}), \mathbb{E}_{k,s} \|v_k - v_{k-1}\| \rangle \\
&\quad + 2 (1 - p) \langle \nabla f (x_{k-1}) - \nabla f (x_k), \mathbb{E}_{k,s} \|v_k - v_{k-1}\| \rangle \\
&\leq (1 - p) \|v_{k-1} - \nabla f (x_{k-1})\|^2 + (1 - p) \mathbb{E}_{k,s} \|v_k - v_{k-1}\|^2.
\end{align*}
\]

\[
\begin{align*}
&= (1 - p) \|v_{k-1} - \nabla f (x_{k-1})\|^2 + (1 - p) \frac{1}{M} \frac{1}{N} \sum_{i=1}^{N} \|\nabla f_i (x_k) - \nabla f_i (x_{k-1})\|^2 \\
&\leq (1 - p) \|v_{k-1} - \nabla f (x_{k-1})\|^2 + \frac{(1 - p) L^2}{M} \|x_k - x_{k-1}\|^2,
\end{align*}
\]
where the last inequality follows from that (i) of assumption 1.

From this inequality, the following result is obvious:
\[
\|v_k - \nabla f (x_k)\|^2 \leq \frac{1}{p} \left( \|v_{k-1} - \nabla f (x_{k-1})\|^2 - \mathbb{E}_k \|v_k - \nabla f (x_k)\|^2 \right) \\
+ \frac{(1 - p) L^2}{pM} \|x_k - x_{k-1}\|^2.
\] (4)
Lemma 3.3. Let \(\{(x_k, y_k, \lambda_k)\}\) be the sequence generated by the SARAH-ADMM, then we have

\[
\mathbb{E}_k \|\lambda_{k+1} - \lambda_k\|^2 \leq \frac{5(2-p)}{\sigma} \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \frac{5\lambda_G^2}{\sigma \eta^2} \mathbb{E}_k \|x_{k+1} - x_k\|^2 \\
+ \left(\frac{5(M+1-p)L^2}{\sigma M} + \frac{5\lambda_G^2}{\sigma \eta^2}\right) \|x_k - x_{k-1}\|^2.
\]

Proof. Since

\[ x_{k+1} = \arg \min_x L^k_{\beta}(x, y_{k+1}, \lambda_k), \]

we have

\[ v_k + \frac{1}{\eta} G(x_{k+1} - x_k) - A^* \lambda_k + \beta A^* (Ax_{k+1} - y_{k+1}) = 0. \]

Then

\[ A^* \lambda_{k+1} = v_k + \frac{1}{\eta} G(x_{k+1} - x_k), \]

which is equivalent to

\[
\|\lambda_{k+1} - \lambda_k\|^2 \leq \frac{1}{\sigma} \|v_k + \frac{1}{\eta} G(x_{k+1} - x_k) - v_{k-1} - \frac{1}{\eta} G(x_k - x_{k-1})\|^2.
\]

Therefore,

\[
\mathbb{E}_k \|\lambda_{k+1} - \lambda_k\|^2 \\
\leq \frac{5}{\sigma} \mathbb{E}_k \|v_k - \nabla f(x_k)\|^2 + \frac{5}{\sigma} \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \frac{5}{\sigma} \|v_{k-1} - \nabla f(x_{k-1})\|^2 \\
+ \frac{5}{\sigma} \mathbb{E}_k \left\|\frac{1}{\eta} G(x_{k+1} - x_k)\right\|^2 + \frac{5}{\sigma} \left\|\frac{1}{\eta} G(x_k - x_{k-1})\right\|^2 \\
\leq \frac{5(2-p)}{\sigma} \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \left(\frac{5(M+1-p)L^2}{\sigma M} + \frac{5\lambda_G^2}{\sigma \eta^2}\right) \|x_k - x_{k-1}\|^2 \\
+ \frac{5\lambda_G^2}{\sigma \eta^2} \mathbb{E}_k \|x_{k+1} - x_k\|^2.
\]

Lemma 3.4. Let \(\{(x_k, y_k, \lambda_k)\}\) be the sequence generated by the SARAH-ADMM, then we have

\[
\mathbb{E}_k \left[L_{\beta}(x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_{\beta}(x_k, y_{k+1}, \lambda_k)\right] \\
\leq \frac{1-p}{2} \|v_{k-1} - \nabla f(x_{k-1})\|^2 - \left(\frac{\beta \sigma}{2} + \frac{\lambda_G}{\eta} - \frac{L+1}{2}\right) \mathbb{E}_k \|x_{k+1} - x_k\|^2 \\
+ \frac{(1-p)L^2}{2M} \|x_k - x_{k-1}\|^2 + \frac{1}{\beta} \mathbb{E}_k \|\lambda_{k+1} - \lambda_k\|^2.
\]
Proof. Firstly, we have
\[ L_\beta (x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_\beta (x_k, y_{k+1}, \lambda_k) \]
\[ = L_\beta (x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_\beta (x_{k+1}, y_{k+1}, \lambda_k) + L_\beta (x_{k+1}, y_{k+1}, \lambda_k) - L_\beta (x_k, y_{k+1}, \lambda_k) \]
\[ = \frac{1}{\beta} \| \lambda_{k+1} - \lambda_k \|^2 + L_\beta (x_{k+1}, y_{k+1}, \lambda_k) - L_\beta (x_k, y_{k+1}, \lambda_k) . \]

and
\[ L_\beta (x_{k+1}, y_{k+1}, \lambda_k) - L_\beta (x_k, y_{k+1}, \lambda_k) \]
\[ = f(x_{k+1}) - f(x_k) - \langle \lambda_k, A x_{k+1} - A x_k \rangle + \frac{\beta}{2} \| A x_{k+1} - y_{k+1} \|^2 - \frac{\beta}{2} \| A x_k - y_{k+1} \|^2 \]
\[ \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 - \langle \lambda_k, A x_{k+1} - A x_k \rangle + \frac{\beta}{2} \| A x_{k+1} - y_{k+1} \|^2 \]
\[ - \frac{\beta}{2} \| A x_k - y_{k+1} \|^2 . \]

By the optimality condition
\[ \langle v_k + \frac{1}{\eta} G (x_{k+1} - x_k) - A^* \lambda_k + \beta A^* (A x_{k+1} - y_{k+1}), x_{k+1} - x_k \rangle \leq 0, \]
we obtain
\[ - \langle \lambda_k, A x_{k+1} - A x_k \rangle + \frac{\beta}{2} \| A x_{k+1} - y_{k+1} \|^2 - \frac{\beta}{2} \| A x_k - y_{k+1} \|^2 \]
\[ \leq - (v_k, x_{k+1} - x_k) - \frac{\beta}{2} \| A x_{k+1} - A x_k \|^2 - \frac{1}{\eta} \| x_{k+1} - x_k \|^2 . \]

And hence
\[ E_k [L_\beta (x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_\beta (x_k, y_{k+1}, \lambda_k)] \]
\[ \leq \frac{1}{\beta} E_k \| \lambda_{k+1} - \lambda_k \|^2 + E_k \langle \nabla f(x_k), v_k, x_{k+1} - x_k \rangle + \frac{L}{2} E_k \| x_{k+1} - x_k \|^2 \]
\[ - \frac{\beta}{2} E_k \| A x_{k+1} - A x_k \|^2 - \frac{1}{\eta} E_k \| x_{k+1} - x_k \|_G^2 \]
\[ \leq \frac{1}{\beta} E_k \| \lambda_{k+1} - \lambda_k \|^2 + \frac{1}{2} E_k \| v_k - \nabla f(x_k) \|^2 + \frac{1}{2} E_k \| x_{k+1} - x_k \|^2 \]
\[ + \frac{L}{2} E_k \| x_{k+1} - x_k \|^2 - \frac{\beta}{2} E_k \| A x_{k+1} - A x_k \|^2 - \frac{1}{\eta} E_k \| x_{k+1} - x_k \|_G^2 \]
\[ \leq - \left( \frac{\beta \sigma}{2} + \frac{\lambda G}{\eta} - \frac{L + 1}{2} \right) E_k \| x_{k+1} - x_k \|^2 + \frac{(1 - p) L^2}{2M} \| x_k - x_{k-1} \|^2 \]
\[ + \frac{1}{\beta} E_k \| \lambda_{k+1} - \lambda_k \|^2 + \frac{1}{2} \| v_{k-1} - \nabla f(x_{k-1}) \|^2 . \]
By the fact that \( y_{k+1} \) is the global minimizer of \( L_{\beta}(x_k, y, \lambda_k) \), we have \( L_{\beta}(x_k, y_{k+1}, \lambda_k) \leq L_{\beta}(x_k, y_k, \lambda_k) \). Then we obtain the following lemma.

**Lemma 3.5.** Let \( \{(x_k, y_k, \lambda_k)\} \) be the sequence generated by the SARAH-ADMM, then we have

\[
E_k \left[ L_{\beta}(x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_{\beta}(x_k, y_k, \lambda_k) \right] \\
\leq \left[ \frac{5(2-p)}{\sigma \beta} + \frac{1-p}{2} \right] \| v_k - \nabla f(x_k) \|^2 - \left( \frac{\beta \sigma}{2} + \frac{\lambda \eta}{\eta} - \frac{L+1}{2} - \frac{5\lambda G}{\sigma \beta \eta^2} \right) E_k \| x_{k+1} - x_k \|^2 \\
+ \left[ \frac{5 \left( (M+1-p) \eta^2 L^2 + M \lambda_G^2 \right)}{\sigma \beta M \eta^2} + \frac{(1-p) L^2}{2M} \right] \| x_k - x_{k-1} \|^2.
\]

**Proof.** Since \( L_{\beta}(x_k, y_{k+1}, \lambda_k) \leq L_{\beta}(x_k, y_k, \lambda_k) \), lemma 4 also holds with \( L_{\beta}(x_k, y_{k+1}, \lambda_k) \) replaced by \( L_{\beta}(x_k, y_k, \lambda_k) \). By the lemma 3, we have

\[
E_k \left[ L_{\beta}(x_{k+1}, y_{k+1}, \lambda_{k+1}) - L_{\beta}(x_k, y_k, \lambda_k) \right] \\
\leq \left[ \frac{5(2-p)}{\sigma \beta} + \frac{1-p}{2} \right] \| v_k - \nabla f(x_k) \|^2 - \left( \frac{\beta \sigma}{2} + \frac{\lambda \eta}{\eta} - \frac{L+1}{2} - \frac{5\lambda G}{\sigma \beta \eta^2} \right) E_k \| x_{k+1} - x_k \|^2 \\
+ \left[ \frac{5 \left( (M+1-p) \eta^2 L^2 + M \lambda_G^2 \right)}{\sigma \beta M \eta^2} + \frac{(1-p) L^2}{2M} \right] \| x_k - x_{k-1} \|^2.
\]

\[\square\]

For ease of presentation, denote

\[
\alpha_1 = \frac{5(2-p)}{\sigma p \beta} + \frac{1-p}{2p}; \\
\alpha_2 = \frac{5 \left( (M+2p+2) \eta^2 L^2 + M p \lambda_G^2 \right)}{\sigma \beta M p \eta^2} + \frac{(1-p) L^2}{2Mp}; \\
\delta = \frac{\beta \sigma}{2} - \frac{5 \left( (M+2p+2) \eta^2 L^2 + 2M \lambda_G^2 \right)}{\sigma \beta M p \eta^2} + \frac{\lambda \eta}{\eta} - \frac{L+1}{2} - \frac{(1-p) L^2}{2Mp}; (5)
\]

and

\[
\Phi(x_k, y_k, \lambda_k) = L_{\beta}(x_k, y_k, \lambda_k) + \alpha_1 \| v_{k-1} - \nabla f(x_{k-1}) \|^2 + \alpha_2 \| x_k - x_{k-1} \|^2.
\]

By means of (4) and lemma 5, the consequence is described in the following lemma.
Lemma 3.6. Let \( \{(x_k, y_k, \lambda_k)\} \) be the sequence generated by the SARAH-ADMM, then we have

\[
E_k \left[ \phi(x_{k+1}, y_{k+1}, \lambda_{k+1}) \right] \leq \Phi(x_k, y_k, \lambda_k) - \delta E_k \|x_{k+1} - x_k\|^2.
\]

**Proof.** From lemma 2, we get

\[
E_k \|v_k - \nabla f(x_k)\|^2 \leq (1-p) \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \frac{(1-p) L^2}{M} \|x_k - x_{k-1}\|^2,
\]
i.e.,

\[
\|v_{k-1} - \nabla f(x_{k-1})\|^2 \leq \frac{1}{p} \left( \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \frac{(1-p) L^2}{M} \|x_k - x_{k-1}\|^2 - E_k \|v_k - \nabla f(x_k)\|^2 \right).
\]

Hence

\[
E_k \left[ L_\beta (x_{k+1}, y_{k+1}, \lambda_k) - L_\beta (x_k, y_k, \lambda_k) \right]
\leq \left[ 5 \frac{(2-p)}{\sigma p \beta} + \frac{1-p}{2p} \right] \|v_{k-1} - \nabla f(x_{k-1})\|^2 - \left[ 5 \frac{(2-p)}{\sigma p \beta} + \frac{1-p}{2p} \right] E_k \|v_k - \nabla f(x_k)\|^2
\]

\[+ \left[ \frac{5 \left( (Mp - 2p + 2) \eta^2 L^2 + Mp \lambda_G^2 \right)}{\sigma \beta M \eta^2} + \frac{(1-p) L^2}{2M} \right] \|x_k - x_{k-1}\|^2
\]

\[- \left( \frac{\beta \sigma}{2} - \frac{\lambda_G}{\eta} - \frac{L+1}{2} - \frac{5 \lambda_G^2}{\sigma \beta \eta^2} \right) E_k \|x_{k+1} - x_k\|^2.
\]

Rearranging terms, the following holds:

\[
E_k \left[ L_\beta (x_{k+1}, y_{k+1}, \lambda_k) + \alpha_1 \|v_k - \nabla f(x_k)\|^2 + \alpha_2 \|x_{k+1} - x_k\|^2 \right]
\leq L_\beta (x_k, y_k, \lambda_k) + \alpha_1 \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \alpha_2 \|x_k - x_{k-1}\|^2 - \delta E_k \|x_{k+1} - x_k\|^2,
\]

where

\[
\alpha_1 = \frac{5 \frac{(2-p)}{\sigma p \beta} + \frac{1-p}{2p}}{\sigma p \beta}, \quad \alpha_2 = \frac{5 \left( (Mp - 2p + 2) \eta^2 L^2 + Mp \lambda_G^2 \right)}{\sigma \beta M \eta^2} + \frac{(1-p) L^2}{2M},
\]

and

\[
\delta = \frac{\beta \sigma}{2} - \frac{5 \left( (Mp - 2p + 2) \eta^2 L^2 + 2Mp \lambda_G^2 \right)}{\sigma \beta M \eta^2} + \frac{\lambda_G}{\eta} - \frac{L+1}{2} - \frac{(1-p) L^2}{2M}.
\]

Note that

\[
\phi(x_k, y_k, \lambda_k) = L_\beta (x_k, y_k, \lambda_k) + \alpha_1 \|v_{k-1} - \nabla f(x_{k-1})\|^2 + \alpha_2 \|x_k - x_{k-1}\|^2,
\]

we can come to the conclusion. \( \square \)
The lemma 6 is very crucial, from which several conclusions can be drawn. And we summarize them in the corollary 1, whose proof is given in the appendix.

**Corollary 3.1.** Let \( \left\{ (x_k, y_k, \lambda_k) \right\} \) be the sequence generated by the SARAH-ADMM which is assumed to be bounded and assume that \( \delta > 0 \) (or \( \beta \) is larger than some constant). Let \( \omega(z_0) \) denote the set of limit points of \( z_k \), where \( z_k := (x_k, y_k, \lambda_k) \). Then the following assertions hold.

(i) \( \sum_{k=0}^{\infty} \| z_{k+1} - z_k \|^2 < \infty \text{ a.s., and hence the limit } \lim_{k \to \infty} \| z_{k+1} - z_k \| = 0 \text{ a.s.} \)

(ii) \( \omega(z_0) \) a.s. a nonempty, compact and connected set and \( \lim_{k \to \infty} \text{dist} (z_k, w(z_0)) = 0 \text{ a.s.} \)

(iii) \( \lim_{k \to \infty} \mathbb{E} L_\beta (z_k) = \mathbb{E} L_\beta (z^*) \equiv L^*, \forall z^* \in \omega(z_0), \text{ where } L^* \text{ is the limit of the sequence of } \{ \mathbb{E} L_\beta (z_k) \} \).

(iv) \( \lim_{k \to \infty} \mathbb{E} \text{dist} (0, \partial L_\beta (z_k)) = 0. \)

Our goal is to prove the global convergence of iterates of SARAH-ADMM, which is not covered in [23]. For that purpose we consider now that the augmented Lagrangian function for (2) is semi-algebraic, which is also a KL function with \( \phi \) of the form \( \phi(\theta) = cs^{1-\theta} \) where \( c \) is positive and \( \theta \in [0, 1) \) is called KL exponent. Next, we present a lemma analogous to the Uniformed KL property in [7], which has been developed in [11].

**Lemma 3.7.** Let \( \left\{ (x_k, y_k, \lambda_k) \right\} \) be the sequence generated by the SARAH-ADMM which is assumed to be bounded and suppose that \( z_k := (x_k, y_k, \lambda_k) \) is not a critical point after a finite number of iterations. Assume that \( L_\beta \) is a semi-algebraic function, then there exist \( K > 0 \) and a continuous concave function \( \phi \) such that for \( k \geq K \), the following inequality holds a.s.:

\[
\phi' (\mathbb{E} L_\beta (z_k) - L^*_k) \mathbb{E} \text{dist} (0, \partial L_\beta (z_k)) \geq 1.
\]

where \( \{ L^*_k \} \) is a non-decreasing sequence converging to \( L^* \).

Now we present the main convergence results.

**Theorem 3.1.** Suppose that \( L_\beta \) is a semi-algebraic function. Let \( \left\{ (x_k, y_k, \lambda_k) \right\} \) be the sequence generated by the SARAH-ADMM which is assumed to be bounded. Then either \( z_k := (x_k, y_k, \lambda_k) \) is a stationary point after a finite number of iterations, or satisfies the finite length property in expectation a.s.:

\[
\sum_{k=0}^{\infty} \mathbb{E} \| z_{k+1} - z_k \| < \infty.
\]

Moreover, the sequence \( \{ z_k \} \) converges globally to the stationary point \( z^* \) of problem (2) a.s.
Proof. We give the proof of the second conclusion here, while the proof of the first conclusion is given in the appendix. Due to
\[
\sum_{k=0}^{\infty} \mathbb{E} \| z_{k+1} - z_k \| < \infty,
\]
we get that the sequence \( \{ z_k \} \) is globally convergent a.s., and the limit is denoted by \( z^* := (x^*, y^*, \lambda^*) \). Invoking the last two equalities of (3), the following two relations hold a.s.:
\[
\begin{align*}
A^* \lambda^* &= \nabla f (x^*) \quad (7) \\
Ax^* - y^* &= 0. \quad (8)
\end{align*}
\]
Moreover, by the Algorithm 1, we have
\[
L_\beta (x_k, y_{k+1}, \lambda_k) \leq L_\beta (x_k, y^*, \lambda_k).
\]
Letting \( k \) goes to \( \infty \) in the above inequality, we obtain
\[
\limsup_{k \to \infty} g (y_{k+1}) \leq g (y^*).
\]
But because
\[
\liminf_{k \to \infty} g (y_{k+1}) \geq g (y^*),
\]
hence we get
\[
\lim_{k \to \infty} g (y_k) = g (y^*).
\]
Also, \( \lim_{k \to \infty} (-\lambda_k - \beta A (x_k - x_{k-1})) = -\lambda^* \) a.s., thus
\[
0 \in \partial g (y^*) + \lambda^* \quad (9)
\]
due to the first inclusion relation and the property of subdifferential. Relations (7), (8) and (9) show that \( (x^*, y^*, \lambda^*) \) is a stationary point of problem (2). \( \square \)

Convergence rate can be also derived for the traditional proximal algorithm [1]. Motivated by this, we obtain a similar result about convergence rate in the stochastic setting.

Theorem 3.2 (convergence rate of SARAH-ADMM). Suppose that \( L_\beta \) is a semi-algebraic function with KL exponent \( \theta \). Let \( \{(x_k, y_k, \lambda_k)\} \) be the sequence generated by the SARAH-ADMM which is assumed to be bounded and \( \beta \) is sufficiently large. Then the following estimation hold a.s.:

(i) If \( \theta = 0 \), then there exists a positive constant \( K \) such that \( \mathbb{E} L_\beta (z_k) = L^* \) for all \( k \geq K \).
(ii) If \( \theta \in (0, \frac{1}{2}] \), then there exists \( c > 0 \) and \( \tau \in [0, 1) \) such that \( \mathbb{E} \| z_k - z^* \| \leq ck^{\frac{\theta}{2\tau^*}} \).

(iii) If \( \theta \in (\frac{1}{2}, 1) \), then there exists \( c > 0 \) such that \( \mathbb{E} \| z_k - z^* \| \leq ck^{\frac{\theta}{2\tau^*}} \).

Proof. From the proof of lemma 3.6, we can actually obtain following inequality by adjusting the coefficients \( \alpha_1, \alpha_2 \) and \( \delta_1, \delta_2, \delta \):

\[
\mathbb{E}_k \Phi (x_{k+1}, y_{k+1}, \lambda_{k+1}) \leq \Phi (x_k, y_k, \lambda_k) - \delta' \mathbb{E}_k \| x_{k+1} - x_k \|^2 - \delta'' \| x_k - x_{k-1} \|^2,
\]

where \( \delta', \delta'' > 0 \) are two constants. Moreover, from the proof of theorem 3.1, we have the estimation for the \( \sum_{i=k_1}^{k} \mathbb{E} \| x_{i+1} - x_i \|^2 \):

\[
\mathbb{E} \| x_{i+1} - x_i \|^2 \leq 2 \mathbb{E} \| x_{k_1} - x_{k_1-1} \|^2 + \mathbb{E} \| x_{k_1-1} - x_{k_1-2} \|^2 + M_1 \Delta_{k_1,k+1}
\]

\[
+ M_2 \sqrt{\mathbb{E} \| v_{k_1-2} - \nabla f (x_{k_1-2}) \|^2},
\]

where \( M_1, M_2 > 0 \) are two constants. Applying the same argument, we have the similar estimation for \( \sum_{i=k_1}^{k} \mathbb{E} \| x_{i} - x_{i-1} \|^2 \):

\[
\mathbb{E} \| x_{i} - x_{i-1} \|^2 \leq \left( \mathbb{E} \| x_{k_1} - x_k \|^2 - \mathbb{E} \| x_{k_1} - x_{k_1-1} \|^2 \right) + \left( \mathbb{E} \| x_{k_1-1} - x_{k_1-2} \|^2 - \mathbb{E} \| x_{k_1-2} - x_{k_1-1} \|^2 \right)
\]

\[
+ M_3 \Delta_{k_1,k+1} + M_4 \sqrt{\mathbb{E} \| v_{k_1-2} - \nabla f (x_{k_1-2}) \|^2},
\]

where \( M_3, M_4 > 0 \) are two constants. Adding the above two estimations and letting \( k \to \infty \), we get

\[
\sum_{i=k_1}^{\infty} \mathbb{E} \| x_{i+1} - x_i \|^2 + \mathbb{E} \| x_i - x_{i-1} \|^2
\]

\[
\leq \sqrt{\mathbb{E} \| x_k - x_{k-1} \|^2} + 2 \mathbb{E} \| x_{k_1-1} - x_{k_1-2} \|^2 - (M_1 + M_3) \left( \mathbb{E} \left[ \Phi (z_{k_1}) - L_k^* \right] \right)^{1-\theta}
\]

\[
+ (M_2 + M_4) \sqrt{\mathbb{E} \| v_{k_1-2} - \nabla f (x_{k_1-2}) \|^2}.
\]

To reach the desired conclusion, the remaining argument is similar to that in [11]. \( \square \)

3.4 Convergence Analysis of SARAH-PGD

We now analyze the convergence properties of SARAH-PGD. For ease of presentation, denote \( \psi (x) := f (x) + g (x), \delta := \frac{\gamma}{2} x - L - \frac{(1-p) L}{M} \left( L + \frac{1-p}{2p} \right), \) and \( \Psi (x_k) := \psi (x_k) + \frac{1-p}{2Lp} \| v_{k-1} - \nabla f (x_{k-1}) \|^2 + \frac{(1-p) L}{M} \left( L + \frac{1-p}{2p} \right) \| x_k - x_{k-1} \|^2. \)
Using lemma 3.1 and 3.2, we can easily obtain the following key estimate for the sequence \( \{ \Psi (x_k) \} \).

**Lemma 3.8.** Let \( \{ x_k \} \) be the sequence generated by the SARAH-PGD, then we have

\[
\mathbb{E}_k \Psi (x_{k+1}) \leq \Psi (x_k) - \delta_k \| x_{k+1} - x_k \|^2.
\]

Moreover, if \( \lambda < \frac{\gamma}{2[L + \frac{1-p}{2Lp}]} \), then the sequence \( \{ \mathbb{E}_k (\Psi (x_k)) \} \) is strictly decreasing.

**Proof.** Firstly, we have the following inequality:

\[
g(x_{k+1}) + \langle v_k, x_{k+1} - x_k \rangle + \frac{1}{\lambda} D_h (x_{k+1}, x_k) \leq g(x_k) \tag{10}
\]

by taking \( x = x_k \). Invoking lemma 3.1 for \( f \), and using the above inequality (10), we get

\[
\psi (x_{k+1}) = f(x_{k+1}) + g(x_{k+1}) \\
\leq f(x_k) + \langle \nabla f (x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 + g(x_{k+1}) \\
\leq f(x_k) + g(x_k) + \langle \nabla f (x_k) - v_k, x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 - \frac{1}{\lambda} D_h (x_{k+1}, x_k) \\
\leq \psi (x_k) + \frac{1}{2L} \| v_k - \nabla f (x_k) \|^2 + \left( L - \frac{\gamma}{2\lambda} \right) \| x_{k+1} - x_k \|^2 \\
\leq \psi (x_k) + \frac{1}{2L} \| v_k - \nabla f (x_k) \|^2 + \left( L - \frac{\gamma}{2\lambda} \right) \| x_{k+1} - x_k \|^2,
\]

where the last inequality holds since \( h \) is \( \gamma \)-strongly convex. Then by taking the conditional expectation on \( x_k \) on both sides of the inequality, using lemma 3.2 and (4), and arranging terms, we obtain

\[
\mathbb{E}_k \left[ \psi (x_{k+1}) + \frac{1-p}{2Lp} \| v_k - \nabla f (x_k) \|^2 \right] \\
\leq \psi (x_k) + \frac{1-p}{2Lp} \| v_{k-1} - \nabla f (x_{k-1}) \|^2 - \left( \frac{\gamma}{2\lambda} - L \right) \mathbb{E}_k \| x_{k+1} - x_k \|^2 \\
+ \frac{(1-p)L}{M} \left( L + \frac{1-p}{2p} \right) \| x_k - x_{k-1} \|^2,
\]

which yields the desired conclusion.

Similar to the analysis of SARAH-ADMM, we easily obtain many nice properties of SARAH-PGD. The proofs are analogue and hence omitted.

**Corollary 3.2.** Let \( \{ x_k \} \) be the sequence generated by the SARAH-PGD which is assumed to be bounded and assume that \( \lambda < \frac{\gamma}{L+1 + \frac{1-p}{2p}} \). Let \( \omega (x_0) \) denote the set of limit points of \( x_k \). Then the following assertions hold.
(i) $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 < \infty$ a.s., and hence the limit $\lim_{k \to \infty} \|x_{k+1} - x_k\| = 0$ a.s.

(ii) $\omega(x_0)$ a.s. a nonempty, compact and connected set and $\lim_{k \to \infty} \text{dist}(x_k, w(x_0)) = 0$ a.s.

(iii) $\lim_{k \to \infty} \mathbb{E} \psi(x_k) = \mathbb{E} \psi(x^*) \equiv \psi^*$, $\forall x^* \in \omega(x_0)$, where $\psi^*$ is the limit of the sequence of $\{\mathbb{E} \psi(x_k)\}$.

(iv) $\lim_{k \to \infty} \mathbb{E} \text{dist}(0, \partial \psi(x_k)) = 0$.

**Lemma 3.9.** Let $\{x_k\}$ be the sequence generated by the SARAH-PGD which is assumed to be bounded and suppose that $x_k$ is not a stationary point after a finite number of iterations. Assume that $\psi$ is a semi-algebraic function, then there exist $K > 0$ and a continuous concave function $\varphi$ such that for $k \geq K$, the following inequality holds a.s.:

$$\varphi'(\mathbb{E} \psi(x_k) - \psi^*) \mathbb{E} \text{dist}(0, \partial \psi(x_k)) \geq 1.$$ 

We now present our main convergence results for the SARAH-PGD algorithm.

**Theorem 3.3.** Suppose that $\psi$ is a semi-algebraic function. Let $\{x_k\}$ be the sequence generated by the SARAH-PGD which is assumed to be bounded. Then either $x_k$ is a stationary point after a finite number of iterations, or satisfies the finite length property in expectation a.s.:

$$\sum_{k=0}^{\infty} \mathbb{E} \|x_{k+1} - x_k\| < \infty.$$

Moreover, the sequence $\{x_k\}$ converges globally to the stationary point $x^*$ of problem (1) a.s.

Similar to the argument in theorem 3.2, we can also obtain the convergence rate of iterates generated by the SARAH-PGD.

**Theorem 3.4** (convergence rate of SARAH-PGD). Suppose that $\psi$ is a semi-algebraic function with KL exponent $\theta$. Let $\{x_k\}$ be the sequence generated by the SARAH-PGD which is assumed to be bounded and $\lambda$ is sufficiently small. Then the following estimation hold a.s.:

(i) If $\theta = 0$, then there exists a positive constant $K$ such that $\mathbb{E} \psi(x_k) = \psi^*$ for all $k \geq K$.

(ii) If $\theta \in (0, \frac{1}{2}]$, then there exists $c > 0$ and $\tau \in [0, 1)$ such that $\mathbb{E} \|x_k - x^*\| \leq c \tau^k$.

(iii) If $\theta \in \left(\frac{1}{2}, 1\right)$, then there exists $c > 0$ such that $\mathbb{E} \|x_k - x^*\| \leq c k^{\frac{\theta - 1}{(\theta - 1)}}$. 

19
4 Experiments

In this section, we report some experimental results of solving binary classification problem and generalized lasso problem. The datasets we used in the experiments are publicly available\(^1\). All numerical experiments are implemented in MATLAB R2014a on a PC with i7-6700K CPU.

4.1 Binary Classification

Given a set of samples \(\{(a_i, b_i)\}_{i=1}^{N}\), where \(a_i \in \mathbb{R}^m, b_i \in \mathbb{R}\). We consider the following classification problem:

\[
\min_x \frac{1}{N} \sum_{i=1}^{N} f_i(x) + \lambda \|Ax\|_{1/2}^{1/2},
\]

where \(f_i(x) = \frac{1}{1+\exp(b_i a_i^T x)}\) and \(\lambda > 0\) is a parameter that needs to be given in advance. For the nonsmooth term, we use the \(\ell_{1/2}\) regularization [52], which can enforce further sparsity than \(\ell_1\) norm, and is nonconvex but semi-algebraic. The matrix \(A\) specifies the desired structured sparsity pattern of \(x\). Like the graph-guided fused lasso model, \(A = [G; I]\) where \(I\) is \(m \times m\) identity matrix and \(G\) encodes the sparsity pattern, and the graph is obtained by sparse inverse covariance estimation [14]. In our experiments, we compare SARAH-ADMM with other three ADMM-type methods, i.e., deterministic ADMM (abbreviated as DETER-ADMM), SVRG-ADMM, and SPIDER-ADMM on four data sets, whose information is summarized in the Table 1. Since SVRG-ADMM and SPIDER-ADMM in [23] are basically two loop algorithms and their outputs are chosen uniformly from all iterates, we output 10 times for each epoch and average these outputs for stability and fairer comparison. As in the previous literature, \(\lambda = 10^{-5}\) and the initial points are generated from a standard normal distribution for all algorithms. Also, we set \(p\) to be \(1/N\). We conducted experiments many times with the same penalty parameter \(\beta\) and step size, and relative trends in these experimental results are roughly the same. The experimental results on different datasets are shown in Figure 1, where “optimality gap” means current objective function value minus that calculated by the deterministic algorithm. As we speculated in Sect. 3, since the outputs of SVRG-ADMM and SPIDER-ADMM are chosen uniformly from all history iterates, their results would not be very stable. However, the output of our algorithm SARAH-ADMM is the last iterate and hence contains more information. In addition, since four algorithms perform the same number of iterations, the run time of DETER-ADMM is relatively long while that of other three algorithms including SARAH-ADMM is short. Moreover, we find that, in some case, SVRG-ADMM and SPIDER-ADMM may be worse than DETER-ADMM due to their uniform outputs in terms of optimality gap in the same amount of time.

\(^1\)https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
Figure 1: Optimality gap versus Time on \(a9a\) (top left), covtype (top right), \(ijcnn1\) (bottom left) and \(w8a\) (bottom right), respectively.
Table 1: Data sets for binary classification

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Sample size</th>
<th>Dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>a9a</td>
<td>32561</td>
<td>123</td>
</tr>
<tr>
<td>covtype</td>
<td>581012</td>
<td>54</td>
</tr>
<tr>
<td>Ijcnn1</td>
<td>49990</td>
<td>22</td>
</tr>
<tr>
<td>w8a</td>
<td>49749</td>
<td>300</td>
</tr>
</tbody>
</table>

Table 2: Data sets for generalized lasso

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Sample size</th>
<th>Dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>abalone</td>
<td>4177</td>
<td>8</td>
</tr>
<tr>
<td>cadata</td>
<td>20640</td>
<td>8</td>
</tr>
<tr>
<td>cpusmall</td>
<td>8192</td>
<td>12</td>
</tr>
<tr>
<td>YearPredictionMSD</td>
<td>51630</td>
<td>90</td>
</tr>
</tbody>
</table>

4.2 Generalized Lasso

In this subsection we consider generalized lasso problem, which is formulated as following:

\[
\frac{1}{2N} \sum_{i=1}^{N} (a_i^T x - b_i)^2 + \lambda \|Ax\|_1^{1/2},
\]

where \( \lambda \) is a regularization parameter and \( \{(a_i, b_i)\}_{i=1}^{N} \) are samples. Such a problem arises frequently in many applications such as statistical machine learning. The matrix \( A \) is obtained in the same way as in the previous subsection. We use four data sets presented in Table 2 to evaluate the performance of algorithms. The initial points are generated from a standard normal distribution for all algorithms and \( p \) is equal to \( 1/N \). Like the previous subsection, we performed experiments many times with the same penalty parameter \( \beta \) and step size, and relative trends in these experimental results are roughly the same. The four algorithms also perform the same number of iterations. In Figure 2, we observe a similar phenomenon as in the binary classification experiment. And these numerical results are consistent with our analysis.

5 Conclusions

In this paper, we propose two stochastic splitting algorithms, which are single loop and whose outputs are the last iterates not chosen uniformly from all iterates as in the existing literature. Based on the Kurdyka-Lojasiewicz (KL) property, which is satisfied by a wide range of functions, we analyze the asymptotic convergence and corresponding rates. In the numerical experiments, we
Figure 2: Optimality gap versus Time on abalone (top left), cadata (top right), cpusmall (bottom left) and YearPredictionMSD (bottom right), respectively.
also observe the instability of existing algorithms due to their uniform output. For our future work, we may consider the multiblock composite optimization problems and discuss the convergence properties for some stochastic splitting algorithms.

Appendix

Proof of Corollary 1

(i) Since

\[ \mathbb{E}_k \left[ \Phi (x_{k+1}, y_{k+1}, \lambda_{k+1}) \right] \leq \Phi (x_k, y_k, \lambda_k) - \delta \mathbb{E}_k \| x_{k+1} - x_k \|^2 \]

and \( \Phi \) has a lower bound, we get

\[ \mathbb{E} \left[ \sum_{k=0}^{\infty} \| x_{k+1} - x_k \|^2 \right] < \infty, \]

which implies that

\[ \sum_{k=0}^{\infty} \| x_{k+1} - x_k \|^2 < \infty \]

almost surely and

\[ \lim_{k \to \infty} \| x_{k+1} - x_k \| = 0 \]

almost surely. Due to lemma 3, we have

\[ \mathbb{E}_k \| \lambda_{k+1} - \lambda_k \|^2 \]

\[ \leq \frac{5 (2 - p)}{\sigma} \| v_{k-1} - \nabla f (x_{k-1}) \|^2 + \left( \frac{5 (M + 1 - p) L^2}{\sigma M} + \frac{5 \lambda_G^2}{\sigma \eta^2} \right) \| x_k - x_{k-1} \|^2 \]

\[ + \frac{5 \lambda_G^2}{\sigma \eta^2} \mathbb{E}_k \| x_{k+1} - x_k \|^2, \]

which implies that

\[ \mathbb{E} \left[ \sum_{k=0}^{\infty} \| \lambda_{k+1} - \lambda_k \|^2 \right] < \infty; \]

and

\[ \sum_{k=0}^{\infty} \| \lambda_{k+1} - \lambda_k \|^2 < \infty \]

almost surely and

\[ \lim_{k \to \infty} \| \lambda_{k+1} - \lambda_k \| = 0 \]

almost surely are followed. Invoking that

\[ \begin{cases} \lambda_{k+1} = \lambda_k - \beta (Ax_{k+1} - y_{k+1}) \\ \lambda_k = \lambda_{k-1} - \beta (Ax_k - y_k) \end{cases}, \]

24
we have
\[ \lambda_{k+1} - \lambda_k = \lambda_k - \lambda_{k-1} + \beta (A x_k - A x_{k+1}) + \beta (y_{k+1} - y_k), \]
which implies
\[ \|y_{k+1} - y_k\|^2 \leq \frac{3}{\beta^2} \left( \|\lambda_{k+1} - \lambda_k\|^2 + \|\lambda_k - \lambda_{k-1}\|^2 + \beta^2 \|A^* A\| \|x_{k+1} - x_k\|^2 \right), \]
\[ \mathbb{E} \left[ \sum_{k=0}^{\infty} \|y_{k+1} - y_k\|^2 \right] < \infty, \]
\[ \sum_{k=0}^{\infty} \|y_{k+1} - y_k\|^2 < \infty \]
almost surely and
\[ \lim_{k \to \infty} \|y_{k+1} - y_k\| = 0 \]
almost surely. Therefore, \( \sum_{k=0}^{\infty} \|z_{k+1} - z_k\|^2 < \infty \) almost surely, and \( \lim_{k \to \infty} \|z_{k+1} - z_k\| = 0 \) almost surely.

(ii) This part of the proof is similar to lemma 5 in [7].

(iii) Since \( \mathbb{E} \Phi (x_k, y_k, \lambda_k) \) is bounded from below and nonincreasing, it has a limit. Moreover,
\[ \mathbb{E}\|v_k - \nabla f (x_k)\|^2 \leq (1 - p) \mathbb{E}\|v_{k-1} - \nabla f (x_{k-1})\|^2 + \frac{(1 - p) \sum_{l=0}^{k-1} L^2 M \mathbb{E}\|x_l - x_{l-1}\|^2}{M} \]

hence
\[ \sum_{k=0}^{\infty} \mathbb{E}\|v_k - \nabla f (x_k)\|^2 \leq \sum_{k=0}^{k} \left( \sum_{l=0}^{k-1} \frac{(1 - p)^{k-l+1} L^2 M \mathbb{E}\|x_l - x_{l-1}\|^2}{M} \right) \]
\[ \leq \sum_{k=0}^{\infty} \sum_{l=0}^{k} \frac{(1 - p)^{k-l+1} L^2 M \mathbb{E}\|x_l - x_{l-1}\|^2}{M} \]
\[ \leq \frac{L^2 B \sum_{l=0}^{k} (1 - p)^{k+1}}{M p} \]
\[ < \infty, \]
where the second inequality follows from the sequence \( \left\{ \mathbb{E}\|x_l - x_{l-1}\|^2 \right\} \) is convergent and has an upper bound \( B \). These results show that
\[ \lim_{k \to \infty} \mathbb{E} L \beta (x_k, y_k, \lambda_k) = \lim_{k \to \infty} \mathbb{E} \Phi (x_k, y_k, \lambda_k) = L^*. \]
For any $z^* \in \omega (z_0)$, there exist a subsequence $\{z_{k_j}\}$ converges to $z^* := (x^*, y^*, \lambda^*)$. Also, we have $L_\beta (x_k, y_{k+1}, \lambda_k) \leq L_\beta (x_k, y^*, \lambda_k)$. Letting $k = k_j$ and $j$ goes to $\infty$, we obtain

$$
\limsup_{j \to \infty} g (y_{k_j}) \leq g (y^*) .
$$

And since $g$ is lower semicontinuous,

$$
\liminf_{j \to \infty} g (y_{k_j}) \geq g (y^*) ,
$$

therefore $\lim_{j \to \infty} g (y_{k_j}) = g (y^*)$ and $\lim_{j \to \infty} \mathbb{E} L_\beta (x_{k_j}, y_{k_j}, \lambda_{k_j}) = \mathbb{E} L_\beta (x^*, y^*, \lambda^*)$ because other function is continuous in $L_\beta$. Invoking that the whole sequence $\{\mathbb{E} L_\beta (x_k, y_k, \lambda_k)\}$ is convergent, we have $\forall z^* \in \omega (z_0)$,

$$
\lim_{k \to \infty} \mathbb{E} L_\beta (z_k) = \mathbb{E} L_\beta (z^*) .
$$

(iv) Invoking the optimality condition for Algorithm 1, we have

$$
\begin{align*}
0 & \in \partial g (y_{k+1}) + \lambda_k - \beta (Ax_k - y_{k+1}) \\
0 & = v_k + \frac{1}{\eta} G (x_{k+1} - x_k) - A^* \lambda_k + \beta A^* (Ax_{k+1} - y_{k+1}), \\
\lambda_{k+1} & = \lambda_k - \beta (Ax_{k+1} - y_{k+1})
\end{align*}
$$

rearranging terms, we obtain

$$
\begin{align*}
-\lambda_k - \beta A (x_k - x_{k-1}) & \in \partial g (y_k) \\
A^* \lambda_{k+1} - \frac{1}{\eta} G (x_{k+1} - x_k) & = v_k \\
\lambda_{k+1} & = \lambda_k - \beta (Ax_{k+1} - y_{k+1})
\end{align*}
$$

Using the property of subdifferentiability, we get

$$
\begin{pmatrix}
A^* (\lambda_{k+1} - \lambda_k) + A^* (\lambda_{k-1} - \lambda_k) - \frac{1}{\eta} G (x_{k+1} - x_k) + (\nabla f (x_k) - v_k) \\
\lambda_k - \lambda_{k-1} - \beta A (x_k - x_{k-1})
\end{pmatrix} \in \partial \mathbb{E} L_\beta (x_k, y_k, \lambda_k)
$$

and there exist $\zeta_i > 0, i = 1, 2, \ldots, 5$ such that

$$
\text{dist} (0, \partial \mathbb{E} L_\beta (x_k, y_k, \lambda_k)) \leq \zeta_1 \| \lambda_{k+1} - \lambda_k \| + \zeta_2 \| \lambda_k - \lambda_{k-1} \| + \zeta_3 \| v_k - \nabla f (x_k) \| + \zeta_4 \| x_{k+1} - x_k \| + \zeta_5 \| x_k - x_{k-1} \| .
$$

Next we bound the $\| \lambda_{k+1} - \lambda_k \|$ and $\| v_k - \nabla f (x_k) \|$.

$$
\| \lambda_{k+1} - \lambda_k \| \\
\leq \frac{1}{\sigma} \left\| v_k + \frac{1}{\eta} G (x_{k+1} - x_k) - v_{k-1} - \frac{1}{\eta} G (x_k - x_{k-1}) \right\| \\
= \frac{1}{\sigma} \left\| v_k - \nabla f (x_k) + \nabla f (x_k) - \nabla f (x_{k-1}) + \nabla f (x_{k-1}) - v_{k-1} + \frac{1}{\eta} G (x_{k+1} - x_k) - \frac{1}{\eta} G (x_k - x_{k-1}) \right\| \\
\leq \frac{1}{\sigma} \| v_k - \nabla f (x_k) \| + \frac{1}{\sigma} \| v_{k-1} - \nabla f (x_{k-1}) \| + \frac{\lambda_{G}}{\sigma \eta} \| x_{k+1} - x_k \| + \left( \frac{L}{\sigma} + \frac{\lambda_{G}}{\sigma \eta} \right) \| x_k - x_{k-1} \| .
$$
Due to lemma 2, we have
\[
E \|v_k - \nabla f (x_k)\|^2 \leq (1 - p) E \|v_{k-1} - \nabla f (x_{k-1})\|^2 + \frac{(1 - p) L^2}{M} E \|x_k - x_{k-1}\|^2 \\
\leq \sum_{l=0}^{k-1} \frac{(1 - p)^{k-l+1} L^2}{M} E \|x_l - x_{l-1}\|^2.
\]

By the Jensen's inequality,
\[
E_k \|v_k - \nabla f (x_k)\| \\
\leq \sqrt{E_k \|v_k - \nabla f (x_k)\|^2} \\
\leq \sqrt{(1 - p) \|v_{k-1} - \nabla f (x_{k-1})\|^2 + \frac{(1 - p) L^2}{M} \|x_k - x_{k-1}\|^2} \\
\leq \sqrt{1 - p \|v_{k-1} - \nabla f (x_{k-1})\| + \frac{(1 - p) L^2}{M} \|x_k - x_{k-1}\|}.
\]

Therefore
\[
E \|v_k - \nabla f (x_k)\| \leq \sum_{l=0}^{k-1} \frac{(1 - p)^{k-l+1} L}{\sqrt{M}} E \|x_l - x_{l-1}\|.
\]

Based on these results, we can bound the Edist \((0, \partial L_{\beta} (z_k))\) using sequence \(\{\|x_{k+1} - x_k\|\}\). Thus \(\lim_{k \to \infty} \text{Edist} (0, \partial L_{\beta} (z_k)) = 0\) due to (i).

**Proof of Theorem 1**

Some of the following statements are almost surely to hold, and we omit the notation a.s. for the sake of simplicity.

If \(z_k\) is a critical point, then our Algorithm 1 terminates. If not, for semi-algebraic problems, \(\phi\) can be chosen to be of the form
\[
\varphi (s) = cs^{1-\theta},
\]
where \(c\) is positive real number and \(\theta\) belongs to \([0, 1)\). In this case, the Kurdyka-Lojasiewicz inequality reduces to
\[
\frac{(1 - \theta) c}{\left(\sigma (u) - \sigma (\bar{u})\right)^{\theta}} \text{dist} (0, \partial \sigma (u)) \geq 1.
\]

By the lemma 7, we have
\[
\frac{(1 - \theta) c}{(E [L_{\beta} (z_k)] - L_{Z_k}^{\Gamma})^{\theta}} E [\text{dist} (0, \partial L_{\beta} (z_k))] \geq 1.
\]
for all \( k \geq K \). Invoking the proof of (iv) of corollary 1, we have

\[
E \left[ \text{dist} \left( 0, \partial L_\beta (z_k) \right) \right] \leq \zeta_1 E \left\| \lambda_{k+1} - \lambda_k \right\| + \zeta_2 E \left\| \lambda_k - \lambda_{k-1} \right\| + \zeta_3 E \left\| v_k - \nabla f (x_k) \right\|
\]

\[
+ \zeta_4 E \left\| x_{k+1} - x_k \right\| + \zeta_5 E \left\| x_k - x_{k-1} \right\|
\]

\[
\leq \zeta_1 \sqrt{E \left\| \lambda_{k+1} - \lambda_k \right\|^2} + \zeta_2 \sqrt{E \left\| \lambda_k - \lambda_{k-1} \right\|^2} + \zeta_3 \sqrt{E \| v_k - \nabla f (x_k) \|^2}
\]

\[
+ \zeta_4 \sqrt{E \| x_{k+1} - x_k \|^2} + \zeta_5 \sqrt{E \| x_k - x_{k-1} \|^2}
\]

\[= Q_k, \]

where the second inequality follows from the Jensen’s inequality. Obviously when \( k \) is sufficiently large, there exists a positive constant \( K_1 \),

\[Q_k \geq K_1 (E \left[ \Phi (z_k) - L_\beta (z_k) \right] )^\theta \geq K_1 (E \left[ \Phi (z_k) - L_\beta (z_k) \right] )^\theta, \theta \in \left[ \frac{1}{2}, 1 \right]\]

and

\[(E [L_\beta (z_k)] - L_k^*)^2 \leq (E [L_\beta (z_k)] - L_k^*)^\theta, \theta \in \left[ 0, \frac{1}{2} \right].\]

Due to the inequality

\[ (E [\Phi (z_k) - L_k^*])^\theta \leq (E [\Phi (z_k) - L_\beta (z_k)])^\theta + (E [\Phi (z_k) - L_k^*])^\theta, \theta \in [0, 1) \]

we obtain

\[Q_k \geq K_2 (E [\Phi (z_k) - L_k^*])^\theta, \]

where \( K_2 > 0 \) and \( \theta \in \left[ \frac{1}{2}, 1 \right] \). For \( \theta \in \left[ 0, \frac{1}{2} \right] \), \( Q_k \geq K_3 (E [\Phi (z_k) - L_k^*])^\frac{1}{2} \) for some \( K_3 > 0 \), which implies that we only need to consider the case of \( \theta \in \left[ \frac{1}{2}, 1 \right] \).

Let \( K = \max \{ K_1, K_2, K_3 \} \), the following inequality holds when \( k \) is sufficiently large:

\[Q_k \geq K (E [\Phi (z_k) - L_k^*])^\theta, \theta \in \left[ \frac{1}{2}, 1 \right].\]

Define the following quantity

\[\Delta_{p,q} := (E [\Phi (z_p) - L_p^*])^{1-\theta} - (E [\Phi (z_q) - L_q^*])^{1-\theta}.\]

By using the concavity of function \( s > 0 \to s^{1-\theta} \) and the above inequality, we obtain that

\[Q_k \Delta_{k,k+1} = Q_k \left[ (E [\Phi (z_k) - L_k^*])^{1-\theta} - (E [\Phi (z_{k+1}) - L_{k+1}^*])^{1-\theta} \right]
\]

\[\geq \frac{Q_k (1 - \theta)}{(E [\Phi (z_k) - L_k^*])^{1-\theta}} E (\Phi (z_k) - L_k^* - \Phi (z_{k+1}) + L_{k+1}^*)
\]

\[\geq \frac{Q_k (1 - \theta)}{(E [\Phi (z_k) - L_k^*])^{1-\theta}} E (\Phi (z_k) - \Phi (z_{k+1}))
\]

\[\geq K (1 - \theta) E (\Phi (z_k) - \Phi (z_{k+1}))
\]

\[\geq \delta K (1 - \theta) E \| x_{k+1} - x_k \|^2, \]
Moreover, by the relation

\[
\sqrt{\mathbb{E}[v_{k-1} - \nabla f(x_{k-1})]^2} \leq \sqrt{(1 - p)} \mathbb{E}[v_{k-2} - \nabla f(x_{k-2})]^2 + \frac{(1 - p) L^2}{M} \mathbb{E}|x_{k-1} - x_{k-2}|^2 \\
\leq \left(1 - \frac{p}{2}\right) \mathbb{E}[v_{k-2} - \nabla f(x_{k-2})]^2 + \frac{(1 - p) L^2}{M} \mathbb{E}|x_{k-1} - x_{k-2}|^2
\]

where the last inequality holds due to \(\sqrt{1 - p} \leq 1 - \frac{p}{2}\), there exist \(\rho_i > 0, i = 1, 2\) such that

\[
Q_k \leq \rho_1 \left(\mathbb{E}[x_{k+1} - x_k]^2 + \mathbb{E}[x_k - x_{k-1}]^2 + \mathbb{E}[x_{k-1} - x_{k-2}]^2\right) + \rho_2 \mathbb{E}[v_{k-2} - \nabla f(x_{k-2})]^2 \\
\leq \rho_3 \left(\mathbb{E}[x_{k+1} - x_k]^2 + \mathbb{E}[x_k - x_{k-1}]^2 + \mathbb{E}[x_{k-1} - x_{k-2}]^2\right) + \rho_4 \left(\mathbb{E}[v_{k-2} - \nabla f(x_{k-2})]^2 - \mathbb{E}[v_{k-1} - \nabla f(x_{k-1})]^2\right).
\]

where \(\rho_3 = \rho_1 + \frac{2}{p} \sqrt{\frac{(1 - p) L^2}{M}}\) and \(\rho_4 = \frac{2\rho_3}{p}\). Denote

\[V_{k-1} := \sqrt{\mathbb{E}[v_{k-2} - \nabla f(x_{k-2})]^2} - \mathbb{E}[v_{k-1} - \nabla f(x_{k-1})]^2\]

Using the fact that \(2\sqrt{\alpha \beta} \leq \alpha + \beta\) for all \(\alpha, \beta \geq 0\), we infer

\[
4\sqrt{\mathbb{E}[x_{k+1} - x_k]^2} \\
\leq 2 \left(\sqrt{\mathbb{E}[x_{k+1} - x_k]^2} + \mathbb{E}[x_k - x_{k-1}]^2 + \mathbb{E}[x_{k-1} - x_{k-2}]^2\right) + \frac{\rho_4}{\rho_3} V_{k-1} \times \frac{\rho_3}{\delta K (1 - \theta)} \Delta_{k,k+1} \\
\leq \left(\sqrt{\mathbb{E}[x_{k+1} - x_k]^2} + \mathbb{E}[x_k - x_{k-1}]^2 + \mathbb{E}[x_{k-1} - x_{k-2}]^2\right) + \frac{\rho_4}{\rho_3} V_{k-1} + \frac{4\rho_3}{\delta K (1 - \theta)} \Delta_{k,k+1},
\]

so that

\[
3\sqrt{\mathbb{E}[x_{k+1} - x_k]^2} \leq \mathbb{E}[x_k - x_{k-1}]^2 + \mathbb{E}[x_{k-1} - x_{k-2}]^2 + \frac{\rho_4}{\rho_3} V_{k-1} + \frac{4\rho_3}{\delta K (1 - \theta)} \Delta_{k,k+1}.
\]

Thus for any \(k, k_1 \geq 1\)

\[
3 \sum_{i=k_1}^{k} \sqrt{\mathbb{E}[x_{i+1} - x_i]^2} \leq \sum_{i=k_1}^{k} \mathbb{E}[x_{i} - x_{i-1}]^2 + \sum_{i=k_1}^{k} \mathbb{E}[x_{i-1} - x_{i-2}]^2 \\
+ \frac{\rho_4}{\rho_3} \sum_{i=k_1}^{k} V_{i-1} + \frac{4\rho_3}{\delta K (1 - \theta)} \sum_{i=k_1}^{k} \Delta_{i,i+1},
\]

29
which is equivalent to
\[
\sum_{i=k_1}^{k} \sqrt{\mathbb{E}\|x_{i+1} - x_i\|^2} \\
\leq \sum_{i=k_1}^{k} \left( \sqrt{\mathbb{E}\|x_i - x_{i-1}\|^2} - \sqrt{\mathbb{E}\|x_{i+1} - x_i\|^2} \right) + \sum_{i=k_1}^{k} \left( \sqrt{\mathbb{E}\|x_{i-1} - x_{i-2}\|^2} - \sqrt{\mathbb{E}\|x_{i+1} - x_i\|^2} \right) \\
+ \frac{\rho_3}{\rho_1} \sum_{i=k_1}^{k} V_{i-1} + \frac{4\rho_3}{\delta K (1 - \theta)} \sum_{i=k_1}^{k} \Delta_{i,i+1} \\
\leq 2\sqrt{\mathbb{E}\|x_{k_1} - x_{k_1-1}\|^2} + \sqrt{\mathbb{E}\|x_{k_1-1} - x_{k_1-2}\|^2} + \frac{4\rho_3}{\delta K (1 - \theta)} \Delta_{k_1,k+1} \\
+ \frac{\rho_3}{\rho_1} \left( \sqrt{\mathbb{E}\|v_{k_1-2} - \nabla f(x_{k_1-2})\|^2} - \sqrt{\mathbb{E}\|v_{k-1} - \nabla f(x_{k-1})\|^2} \right) \\
< \infty,
\]
where the last inequality is due to the definition of \( \Delta_{k_1,k+1} \) and the finiteness of \( \sqrt{\mathbb{E}\|v_{k_1-2} - \nabla f(x_{k_1-2})\|^2} \). Therefore,
\[
\sum_{i=k_1}^{k} \mathbb{E}\|x_{i+1} - x_i\| \leq \sum_{i=k_1}^{k} \sqrt{\mathbb{E}\|x_{i+1} - x_i\|^2} < \infty.
\]
Consequently,
\[
\sum_{k=0}^{\infty} \mathbb{E}\|x_{k+1} - x_k\| < \infty.
\]
Invoking the proof of the corollary 1, we also obtain
\[
\sum_{k=0}^{\infty} \mathbb{E}\|y_{k+1} - y_k\| < \infty
\]
and
\[
\sum_{k=0}^{\infty} \mathbb{E}\|\lambda_{k+1} - \lambda_k\| < \infty.
\]
Therefore,
\[
\sum_{k=0}^{\infty} \mathbb{E}\|z_{k+1} - z_k\| < \infty.
\]

References


