AIMS Mathematics, 11(1): 2188–2226.
DOI: 10.3934/math.2026089
Received: 02 December 2025
Revised: 07 January 2026
Accepted: 16 January 2026
Published: 23 January 2026

https://www.aimspress.com/journal/Math

Research article

Preconditioned primal–dual gradient methods for nonconvex composite and finite-sum optimization

Jiahong Guo*

School of Economics and Management, Qilu Normal University, Jinan 250200, China

* **Correspondence:** Email: jhguo0722@163.com.

Abstract: In this paper, we first introduce a preconditioned primal–dual gradient algorithm based on conjugate duality theory. This algorithm is designed to solve a composite optimization problem whose objective function consists of two summands: a continuously differentiable nonconvex function and the composition of a nonsmooth nonconvex function with a linear operator. Under mild conditions, we prove that any cluster point of the generated sequence is a critical point of the composite optimization problem. Under the Kurdyka–Łojasiewicz property, we establish the global convergence and convergence rates for the iterates. Second, for nonconvex finite-sum optimization, we propose a stochastic algorithm that combines the preconditioned primal–dual gradient algorithm with a class of variance-reduced stochastic gradient estimators. Almost sure global convergence and expected convergence rates are derived by relying on the Kurdyka–Łojasiewicz inequality. Finally, preliminary numerical results are presented to demonstrate the effectiveness of the proposed algorithms.

Keywords: nonconvex first-order primal–dual algorithms; Kurdyka–Łojasiewicz inequality; global convergence; convergence rates; stochastic approximation
Mathematics Subject Classification: 90C26, 90C15, 90C06

1. Introduction

In this paper, we first consider the following composite optimization problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) + h(Ax), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable and possibly nonconvex function, $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, and $h : \mathbb{R}^m \to (-\infty, +\infty]$ is a simple and possibly nonsmooth, nonconvex function. Problem (1.1) arises in a variety of practical applications including machine learning, statistics, image processing, and so on. In many applications, the function h is usually referred to the *regularizer*, which is used to guarantee certain regularity properties of the solution. Recently, nonconvex regularizers,

such as ℓ_0, ℓ_p $(0 < p < 1)$, smoothly clipped absolute deviation (SCAD), and the minimax concave penalty (MCP), have drawn a lot of attention and can achieve significant improvement over convex regularizers (see [1] and references therein).

For Problem (1.1) in the fully nonconvex setting (both f and h are nonconvex), there has been an intensive renewed interest in the convergence analysis of various algorithms based on the Kurdyka–Łojasiewicz (KŁ) property in recent years. Attouch et al. [2] established the global convergence of a forward–backward splitting algorithm for (1.1), with A being the identity operator and $(f + h)$ being a KŁ function. Li and Pong [3] demonstrated the convergence of an alternating direction method of multipliers (ADMM) under the assumptions that both f and h are semialgebraic and A is surjective. A nonmonotone linesearch algorithm based on the forward–backward envelope was proposed by [4] and shown to have superlinear convergence rates. In [5], the authors employed a Lyapunov method to establish the global convergence of a bounded sequence to a critical point for several Lagrangian-based methods, including the proximal multipliers method and proximal ADMM, within the semialgebraic setting. By assuming that the associated augmented Lagrangian possesses the KŁ property, it was proved in [6] that the iterates of proximal ADMM converge to a Karush–Kuhn–Tucker point. They also derived convergence rates for both the augmented Lagrangian and the iterates. Algorithms for Problem (1.1) with h being ℓ_0 norm were reviewed in a survey paper [7]. For Problem (1.1) with convex h, there exists a vast literature on various nonconvex composite optimization algorithms [8–10].

Motivated by a class of well-studied primal–dual hybrid gradient (PDHG) algorithms for convex optimization [11–13], and drawing upon the conjugate duality theory for nonconvex optimization, we propose a preconditioned first-order primal–dual algorithm for solving the nonconvex composite optimization Problem (1.1). In most of the aforementioned related algorithms, it is necessary to compute the elements of the generalized proximal (set-valued) mapping for the nonconvex function h and/or nonconvex function f at each iteration. In contrast, at each iteration of our proposed algorithm, we only need to calculate the proximal mapping of the conjugate function h^*, which is always convex and lower semicontinuous. Although certain popular nonconvex regularizers admit closed-form proximal mappings, this property is not guaranteed for general nonconvex functions h.

In the second part of this paper, we extend the proposed algorithm to the following finite-sum optimization problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} f_i(x) + h(Ax), \tag{1.2}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, N$, is continuously differentiable and possibly nonconvex, and $h(Ax)$ is the same as in (1.1). Problem (1.2) arises frequently in the fields of statistics [14], image processing [15], and machine learning [16]. Problem (1.2) is also called *regularized empirical risk minimization*, and the component functions $f_i, i = 1, \ldots, N$ correspond to certain loss models. Moreover, in various interesting problems such as deep learning, dictionary learning, and classification with nonconvex activation functions, the loss functions f_i exhibit nonconvexity. Since the number of components N (which usually represents the size of a dataset) can be extremely large, the exact computation of the full gradient $\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x)$ becomes prohibitively expensive in practice. Consequently, stochastic approximation techniques have gained increasing importance in designing efficient numerical algorithms for finite-sum optimization problems (see, for example, [17–21]). In particular, the success of many popular variance-reduced stochastic algorithms for convex finite-sum optimization has been witnessed in recent years, such as SAG [22], SAGA [23], SVRG [24], and

SARAH [25].

For Problem (1.2) with nonconvex components f_i, $i = 1, \ldots, N$ and a convex h, a large number of algorithms have been developed over the past few years. We only name a few here. The authors of [26] introduced a stochastic proximal gradient algorithm based on variance reduction and established a global linear convergence rate for the nonconvex f_i satisfying the Polyak–Łojasiewicz condition. Nhan et al. [27] presented a stochastic first-order algorithm by combining a proximal gradient step with the SARAH estimator and analyzed the complexity bounds in terms of stochastic first-order oracle calls. Fort and Moulines [28] introduced a stochastic variable metric proximal gradient algorithm by using a mini-batch strategy with variance reduction called SPIDER [29]. Milzarek et al. [30] proposed a stochastic semismooth Newton method for nonsmooth and nonconvex finite-sum optimization and established the almost sure global convergence and local convergence rates with high probability. Wang and Chen [31] studied proximal inexact gradient methods for nonconvex and nonsmooth finite-sum optimization with non-Lipschitz regularization.

We now review the stochastic approximation algorithms for Problem (1.2) in the fully nonconvex setting, where $f_i, i = 1, \ldots, N$ and h are nonconvex. Xu et al. [32] showed that the stochastic proximal gradient methods for Problem (1.2) with a nonconvex h have the same complexities as their counterparts for convex regularized problems to find an approximate stationary point. A stochastic algorithm that combines ADMM with a class of variance-reduced stochastic gradient estimators, including SAGA, SVRG, and SARAH, was proposed by [15]. The global convergence in expectation was established under the condition that $f_i, i = 1, \ldots, N$ and h are semialgebraic, and the convergence rates in the expectation sense were derived by depending on the Łojasiewicz exponent. Using the forward–backward envelope as a Lyapunov function, Latafat et al. [33] proved that the cluster points of the iterates generated by the popular proximal Finito/MISO algorithm are the stationary points almost surely in the fully nonconvex case. They also established the global and linear convergence under the assumption that $f_i, i = 1, \ldots, N$ and h are KŁ functions. In [34], the authors designed a normal map-based proximal random reshuffling method for (1.2) with a weak convex h and proved its convergence under the KŁ property. Bai et al. [35] proposed a single-loop stochastic ADMM with a hybrid gradient estimator for both expectation and finite-sum optimization problems with linear constraints, achieving sublinear convergence. By combining the proposed algorithm for Problem (1.1) with the variance-reduced stochastic gradient estimators (uniformly defined in [15, 36]), we study a stochastic preconditioned first-order primal–dual algorithm for solving the fully nonconvex finite-sum optimization Problem (1.2).

The main contributions of this paper can be summarized as follows.

- We propose a preconditioned primal–dual gradient (PPDG) method for the composite optimization Problem (1.1). This problem presents significant challenges because of its fully nonconvex structure including the smooth nonconvex function f and the nonsmooth nonconvex regularizer h coupled through the linear operator A. Under the mild assumptions that the gradient of f is Lipschitz continuous, the linear operator A is surjective, and the convex hull of h is proper, we prove that any convergent subsequence of the iterates converges to a critical point of the Lagrange function associated with Problem (1.1). This is realized by establishing the nonincreasing property of a properly selected Lyapunov function. With the additional KŁ property of the Lyapunov function, we demonstrate the global convergence of the generated sequence of iterates. We further derive convergence rates for the sequence, provided that the

Lyapunov function has the Łojasiewicz property.

- To address the challenge of solving Problem (1.2) in the fully nonconvex setting, we introduce a stochastic preconditioned primal–dual gradient (SPPDG) method, which can be viewed as a stochastic variant of PPDG. To analyze the convergence of SPPDG, we first establish a crucial descent property related to the expectation of a Lyapunov function based on the Lagrange function of Problem (1.2). Moreover, the upper bound for the conditional expectation of the subgradient of the Lyapunov function is derived. Leveraging these important auxiliary results and assuming that the generated iterates of SPPDG are bounded almost surely, we establish the subsequence convergence in the almost sure sense. Moreover, if the Lyapunov function is a KŁ function, we prove that the whole iteration sequence possesses the finite length property and converges almost surely to a critical point. To the best of our knowledge, such almost sure global convergence result for stochastic algorithms applied to (1.2) in the fully nonconvex setting is new.
- We report the numerical performance of the proposed SPPDG applied to image classification using deep neural network, and to a nonconvex graph-guided fused lasso problem. Compared with the existing popular algorithms, the numerical results verify the advantages of the proposed methods.

The rest of this paper is organized as follows. In Section 2, we explore the convergence of a preconditioned primal–dual gradient method for the composite optimization Problem (1.1). In Section 3, we propose a stochastic preconditioned primal–dual gradient method for the finite-sum Problem (1.2) and provide a convergence analysis. Numerical experiments are presented in Section 4 to show the effectiveness of the proposed algorithms.

Notations and basic definitions. Let $\mathbb{R}^n$ and $\mathbb{R}^m$ be two Euclidean spaces equipped with the standard inner products $\langle\cdot,\cdot\rangle$ and norms $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. The operator norm of a linear operator $A : \mathbb{R}^n \to \mathbb{R}^m$ is

$$\|A\| := \max\{\|Ax\| : x \in \mathbb{R}^n \text{ with } \|x\| \le 1\}.$$

Given a closed set $C \subset \mathbb{R}^n$ and a vector $x \in \mathbb{R}^n$, the *distance* of x to C is given by $\mathrm{dist}(x, C) := \min_{y\in C} \|x - y\|$. Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper lower semicontinuous convex function. The extended *proximal mapping* of f associated with a positive definite linear operator M is defined as

$$\mathrm{prox}_f^M(y) := \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{1}{2}\|x - y\|_M^2 \right\}.$$

Here, $\|x\|_M^2 := \langle Mx, x\rangle$. For an extended real-valued function $f : \mathbb{R}^n \to (-\infty, +\infty]$, let $\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$ be its domain and

$$f^*(y) := \sup_{x\in\mathbb{R}^n}\{\langle y, x\rangle - f(x)\},\ y \in \mathbb{R}^n$$

be its *conjugate function*. The conjugate function f^* is always convex and lower semicontinuous [37, Theorem 4.3]. When f is convex, let ∂f denote its *subdifferential*. A set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be *outer semicontinuous* at $\bar{x}$ if, for any sequence $x^k \subset \mathbb{R}^n$ with $x^k \to \bar{x}$ and any sequence $u^k \subset \mathbb{R}^m$ with $u^k \in F(x^k)$ and $u^k \to u$, it holds that $u \in F(\bar{x})$. If f is proper, lower semicontinuous, and convex, the set-valued mapping ∂f is outer semicontinuous or, equivalently, its graph is closed [38, Theorem 24.4].

2. PPDG for nonconvex composite optimization

In this section, we propose PPDG, a preconditioned primal–dual first-order method based on conjugate duality, for solving the nonconvex composite optimization Problem (1.1) and establish its convergence. We begin by reviewing the preliminary conjugate duality results in Subsection 2.1. The algorithmic framework of PPDG and the main assumptions are described in Subsection 2.2. Subsection 2.3 is devoted to derive the descent property of a Lyapunov function. The subsequence convergence is investigated in Subsection 2.4. Finally, under the KŁ property, the main theoretical results regarding the global convergence and convergence rates are established in Subsection 2.5.

2.1. Conjugate duality and necessary optimality

Going back to Problem (1.1) and drawing upon the conjugate duality theory [39, Section 2.5.3], we find that the dual problem of (1.1) is

$$\max_{y\in\mathbb{R}^m}\left\{\inf_{x\in\mathbb{R}^n}\mathcal{L}(x,y)\right\},\ \text{where}\ \mathcal{L}(x,y):=f(x)+\langle y,Ax\rangle-h^*(y). \tag{2.1}$$

If $\inf_{x\in\mathbb{R}^n}\mathcal{L}(x,y) > -\infty$ for any $y\in\mathbb{R}^m$, then $\bar{x}$ and $\bar{y}$ are the optimal solutions of (1.1) and (2.1), respectively, if and only if the following relations hold true:

$$\begin{cases} \bar{x}\in\arg\min_{x\in\mathbb{R}^n}\mathcal{L}(x,\bar{y}), \\ 0=h(A\bar{x})+h^*(\bar{y})-\langle\bar{y},A\bar{x}\rangle. \end{cases} \tag{2.2}$$

From the definition of the conjugate function, it follows that if (2.2) is satisfied, we have $0\in\partial\mathcal{L}(\bar{x},\bar{y})$, i.e.

$$\begin{cases} 0=\nabla f(\bar{x})+A^T\bar{y}, \\ 0\in-\partial h^*(\bar{y})+A\bar{x}, \end{cases} \tag{2.3}$$

where A^T is the adjoint operator of A. These relations (2.3) provide a set of first-order necessary optimality conditions for the nonconvex Problem (1.1). Let us denote the set of critical points of $\mathcal{L}$ as

$$\mathrm{crit}\mathcal{L}:=\{(\bar{x},\bar{y})\in\mathbb{R}^n\times\mathbb{R}^m:0\in\partial\mathcal{L}(\bar{x},\bar{y})\}.$$

Therefore, the primary aim of this paper is to find a pair $(\bar{x},\bar{y})$ that satisfies (2.3), that is, $(\bar{x},\bar{y})\in\mathrm{crit}\mathcal{L}$. Similarly, for the nonconvex finite-sum Problem (1.2), our goal is to obtain a critical point of

$$\mathcal{L}_s(x,y):=\frac{1}{N}\sum_{i=1}^{N}f_i(x)+\langle y,Ax\rangle-h^*(y).$$

2.2. The PPDG algorithm

The details of PPDG are described in Algorithm 1. This algorithm can be viewed as a first-order primal–dual algorithm by observing the necessary optimality conditions (2.3). In particular, (2.5a) is a standard gradient step associated with the first relation in (2.3), and (2.5b) can be regarded as a

proximal gradient step coupled with the preconditioning technique introduced by [12] for the second relation $0 \in -\partial h^*(\bar{y}) + A\bar{x}$. More precisely, (2.5b) corresponds to the extended proximal mapping of h^* given by

$$y^{k+1} = \text{prox}_{h^*}^M(y^k + M^{-1}A(2x^{k+1} - x^k)).$$

In view of (2.5b), a vector $g^{k+1} \in \partial h^*(y^{k+1})$ exists such that

$$g^{k+1} = -M(y^{k+1} - y^k) + A(2x^{k+1} - x^k). \tag{2.4}$$

If the sequence $\{(x^k, y^k)\}$ converges to $(\bar{x}, \bar{y})$, then (2.4) immediately implies the second relation $A\bar{x} \in \partial h^*(\bar{y})$ due to the outer semicontinuity of ∂h^*.

Algorithm 1: PPDG

1 Initialization: Choose an initial point $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$, a constant $\alpha > 0$, and a positive definite matrix M.

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

3 Update x^k, y^k as follows:

$$\begin{cases} x^{k+1} = x^k - \alpha(\nabla f(x^k) + A^T y^k), & \text{(2.5a)} \\ y^{k+1} = \arg\min\limits_{y \in \mathbb{R}^m} \left\{ h^*(y) - \langle y, A(2x^{k+1} - x^k)\rangle + \frac{1}{2}\|y - y^k\|_M^2 \right\}. & \text{(2.5b)} \end{cases}$$

4 Set $k \leftarrow k + 1$.

Compared with the existing first-order algorithms for nonsmooth and nonconvex optimization problems, one of the main features of Algorithm 1 is that we compute the proximal mapping of the conjugate function h^* rather than dealing with h directly. This is partially motivated by the popular PDHG algorithm [11] for convex optimization problems. Upon observing (2.3), in both the definition of crit$\mathcal{L}$ and the later subsequence convergence analysis of Algorithm 1, we do not need to introduce complex generalized subdifferentials of nonconvex functions, as is often required in many well-studied first-order algorithms for nonsmooth and nonconvex optimization problems (see, e.g., [40–43]).

In order to establish the convergence of Algorithm 1, we impose some standard assumptions throughout this section.

Assumption 1. *Suppose the following:*

(i) The function f is L-smooth over $\mathbb{R}^n$, i.e., f is continuously differentiable and a constant $L > 0$ exists such that for any $x, z \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(z)\| \le L\|x - z\|.$$

(ii) $\inf_{x \in \mathbb{R}^n} \mathcal{L}(x, y) > -\infty$ for any $y \in \mathbb{R}^m$.

(iii) The linear operator A is surjective.

(iv) The convex hull of h is proper.

Remark 1. *Some comments on Assumption 1 are in order.*

(i) Assumption 1(i) holds for many loss functions, including the cross-entropy loss function and the sigmoid loss function in our experiments. A well-known gradient descent lemma under Assumption 1(i) is

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2. \tag{2.6}$$

Moreover, by applying (2.5a) *and Assumption 1(i), we have*

$$\|A^T(y^{k+1} - y^k)\| \leq \left(\frac{1}{\alpha} + L\right)\|x^{k+1} - x^k\| + \frac{1}{\alpha}\|x^{k+2} - x^{k+1}\|. \tag{2.7}$$

(ii) Assumption 1(ii) ensures that the sequence generated by Algorithm 1 is well-defined. It is also indispensable in the subsequence convergence analysis (see Proposition 3).

(iii) The linear operator A is surjective if and only if the matrix associated with AA^T is positive definite. Thus, under Assumption 1(iii), a natural choice of M in Algorithm 1 is the associated matrix of αAA^T. As a special case, if A is the identity operator, then we can choose $M = \alpha I$. Moreover, under Assumption 1(iii), for any $y \in \mathbb{R}^m$, we have

$$\hat{\lambda}\|y\| \leq \|A^T y\|, \tag{2.8}$$

where $\hat{\lambda} := \sqrt{\lambda_{\min}(AA^T)}$ and $\lambda_{\min}(AA^T)$ denotes the smallest eigenvalue of AA^T. Assumption 1(iii) is a standard condition in convergence analysis (see [6, 15, 43]). Even if the linear operator A is not surjective, the proposed algorithm can often be applied effectively in practice.

(iv) The outer semicontinuity of ∂h^*, which is essential in convergence analysis (see Proposition 3), is guaranteed when h^* is proper, lower semicontinuous, and convex. It is known that, without any assumption on h, the conjugate function h^* is lower semicontinuous and convex [37, Theorem 4.3]. However, in order to guarantee that h^* is proper, an additional assumption is required. It is shown in [44, Theorem 11.1] that h^* is proper if Assumption 1(iv) holds. Common nonconvex regularizers, including ℓ_0, ℓ_p, SCAD, and MCP, satisfy Assumption 1(iv).

2.3. A Lyapunov function

As discussed previously, the primary aim of this section is to establish the convergence result that the sequence (x^k, y^k) generated by Algorithm 1 converges to a critical point of the Lagrange function $\mathcal{L}(x, y)$. However, this is difficult to fulfill for the nonconvex composite optimization Problem (1.1) through the usual approach owing to the lack of the descent property of $\mathcal{L}$. Instead, we work with an auxiliary function to alleviate this difficulty.

Let us define the following Lyapunov function:

$$\mathscr{L}(x, y, u, v) := \mathcal{L}(x, y) - a\|x - u\|^2 + b\|x - v\|^2, \ \forall x, u, v \in \mathbb{R}^n, \ y \in \mathbb{R}^m.$$

Here, with the step size α and the Lipschitz constant L of ∇f, let

$$a := \frac{\delta}{\alpha}, \quad b := \frac{1}{2\alpha} - \frac{L}{4} - \frac{\delta}{\alpha} - \frac{\alpha\delta L^2}{2} - \delta L + \frac{\alpha L^2}{4\delta}, \tag{2.9}$$

and let δ be a properly selected constant such that $a > 0$ and $b > 0$. Let c be a constant given by

$$c := b - \frac{\alpha L^2}{2\delta}. \tag{2.10}$$

By an elementary calculation, if we choose $\delta = 1/5$, then the choice of step size $\alpha \in (0, 1/3L)$ is sufficient to guarantee $c > 0$ and, consequently, $a > 0$ and $b > 0$. Therefore, we can safely assume that a, b, and c are positive in the rest of this section.

The convergence analysis of Algorithm 1 will significantly rely on the properties of $\mathscr{L}$, which will be investigated in this subsection. For a start, we show in the following lemma that the critical point set $\mathrm{crit}\mathscr{L}$ is closely related to $\mathrm{crit}\mathcal{L}$.

Lemma 1. *Let $x, u, v \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$. Then, $(x, y, u, v) \in \mathrm{crit}\mathscr{L}$ is equivalent to $(x, y) \in \mathrm{crit}\mathcal{L}$ and $u = v = x$.*

Proof. In view of the definition of $\mathscr{L}$, the condition $(x, y, u, v) \in \mathrm{crit}\mathscr{L}$ reads as

$$\begin{aligned}
0 &= \nabla_x \mathscr{L}(x, y, u, v) = \nabla_x \mathcal{L}(x, y) - 2a(x - u) + 2b(x - v),\\
0 &\in \partial_y \mathscr{L}(x, y, u, v) = \partial_y \mathcal{L}(x, y),\\
0 &= \nabla_u \mathscr{L}(x, y, u, v) = 2a(x - u),\\
0 &= \nabla_v \mathscr{L}(x, y, u, v) = 2b(v - x).
\end{aligned}$$

The latter two relations imply that $u = v = x$. This, together with the first two relations, leads to $(0, 0) \in \partial\mathcal{L}(x, y)$, which implies $(x, y) \in \mathrm{crit}\mathcal{L}$. The converse is obvious. □

Throughout the remainder of this section, we let M be the matrix associated with αAA^T, that is, for all $y, \hat{y} \in \mathbb{R}^m$, $\langle \hat{y}, My\rangle = \langle \hat{y}, \alpha AA^T y\rangle$. Let

$$z^k := (x^k, y^k, x^{k+1}, x^{k-1})$$

and

$$d^k := (\nabla_x \mathscr{L}(z^k), Ax^k - g^k, \nabla_u \mathscr{L}(z^k), \nabla_v \mathscr{L}(z^k)),$$

where $g^k = -M(y^k - y^{k-1}) + A(2x^k - x^{k-1})$. From (2.4), we know that $g^k \in \partial h^*(y^k)$ and hence $d^k \in \partial\mathscr{L}(z^k)$ by the definition of $\mathscr{L}$. We now establish the following descent property of $\mathscr{L}$, which will play a pivotal role in the discussion of global convergence.

Lemma 2. *Let Assumption 1 hold. Then, for all $k \geq 1$, we have*

$$\mathscr{L}(z^{k+1}) + c(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2) \leq \mathscr{L}(z^k), \tag{2.11}$$

where c is defined in (2.10)*. Moreover, it holds that*

$$\|d^k\| \leq \gamma_1 \|x^k - x^{k-1}\| + \gamma_2 \|x^{k+1} - x^k\|, \tag{2.12}$$

where $\gamma_1 := 2L + 4b + \frac{2}{\alpha} + (2 + \alpha L)\|A\|$ and $\gamma_2 := 4a + \frac{1}{\alpha} + \|A\|$, with a, b given in (2.9)*.*

Proof. See Appendix A.1. □

2.4. Subsequence convergence

Let C denote the set of cluster points of the sequence $\{(x^k, y^k)\}$ generated by Algorithm 1. We now establish the subsequence convergence based on the previous lemmas concerning $\mathscr{L}$. These convergence results will be proved under the assumption that the sequence $\{(x^k, y^k)\}$ is bounded, which is a standard assumption in the global convergence analysis of nonconvex optimization algorithms (see [5, 41, 43], for instance).

Proposition 3. *Let the sequence $\{(x^k, y^k)\}$ be bounded, and suppose that Assumption 1 holds. We then have*

(i) $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 < \infty$ *and* $\sum_{k=1}^{\infty} \|y^{k+1} - y^k\|^2 < \infty$;
(ii) C *is a nonempty compact set and* $\lim_{k\to\infty} \mathrm{dist}((x^k, y^k), C) = 0$;
(iii) $C \subseteq \mathrm{crit}\mathcal{L}$;
(iv) $\mathcal{L}$ *is finite and constant on* C.

Proof. Assumption 1 (ii) implies $\inf_k \mathcal{L}(x^k, y^k) > -\infty$, which, together with the boundedness of $\{x^k\}$, leads to $\inf_k \mathscr{L}(z^k) > -\infty$. Since the sequence $\{\mathscr{L}(z^k)\}$ is nonincreasing (cf. Lemma 2) and bounded from below, $\mathscr{L}(z^k)$ converges to a finite value denoted by $\bar{\mathscr{L}}$. Summing (2.11) over $k = 1, \ldots, n$ yields

$$c \sum_{k=1}^{n} (\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2) \le \mathscr{L}(z^1) - \mathscr{L}(z^{n+1}).$$

Let $n \to \infty$, by the convergence of $\{\mathscr{L}(z^k)\}$, we have

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 < \infty.$$

This, together with (2.7) and (2.8), also gives

$$\sum_{k=1}^{\infty} \|y^{k+1} - y^k\|^2 < \infty.$$

Item (i) is derived. Moreover, it implies

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0 \text{ and } \lim_{k\to\infty} \|y^{k+1} - y^k\| = 0. \tag{2.13}$$

The compactness of C follows from the proof of [41, Lemma 5 (iii)]. Since the sequence $\{(x^k, y^k)\}$ is bounded, C is nonempty, and for any $(\bar{x}, \bar{y}) \in C$, a subsequence $\{(x^{k_q}, y^{k_q})\}$ of $\{(x^k, y^k)\}$ exists such that

$$\lim_{q\to\infty} \|x^{k_q} - \bar{x}\| = 0, \ \lim_{q\to\infty} \|y^{k_q} - \bar{y}\| = 0. \tag{2.14}$$

By the definition of the distance function, we have

$$\mathrm{dist}((x^k, y^k), C) \le \|x^k - \bar{x}\| + \|y^k - \bar{y}\| \le \|x^k - x^{k_q}\| + \|x^{k_q} - \bar{x}\| + \|y^k - y^{k_q}\| + \|y^{k_q} - \bar{y}\|.$$

Combining this inequality with (2.13) and (2.14), we obtain the result that $\mathrm{dist}((x^k, y^k), C)$ converges to 0 and hence Item (ii) holds.

For Item (iii), it is sufficient to prove that $(\bar{x}, \bar{y}) \in \text{crit}\mathcal{L}$ for any $(\bar{x}, \bar{y}) \in C$. Let $\bar{z} := (\bar{x}, \bar{y}, \bar{x}, \bar{x})$. Noting that $z^{k_q} \to \bar{z}$, $d^{k_q} \in \partial\mathscr{L}(z^{k_q})$ and $d^{k_q} \to 0$ by (2.12), we derived from the outer semicontinuity of $\partial\mathscr{L}$ that $0 \in \partial\mathscr{L}(\bar{z})$, i.e., $(\bar{x}, \bar{y}, \bar{x}, \bar{x}) \in \text{crit}\mathscr{L}$. Therefore, from Lemma 1, it follows that $(\bar{x}, \bar{y}) \in \text{crit}\mathcal{L}$.

Recall from Remark 1 (iv) that the conjugate function h^* is proper, lower semicontinuous, and convex. Thus, h^* is continuous over its domain $\text{dom}h^*$ [37, Theorem 2.22]. Therefore, $\mathcal{L}$ is continuous over $\mathbb{R}^n \times \text{dom}h^*$ and

$$\lim_{q\to\infty} \mathcal{L}(x^{k_q}, y^{k_q}) = \mathcal{L}(\bar{x}, \bar{y}),$$

which also implies

$$\lim_{q\to\infty} \mathscr{L}(z^{k_q}) = \lim_{q\to\infty} (\mathcal{L}(x^{k_q}, y^{k_q}) - a\|x^{k_q} - x^{k_q+1}\|^2 + b\|x^{k_q} - x^{k_q-1}\|^2) = \mathcal{L}(\bar{x}, \bar{y}) = \mathscr{L}(\bar{z}). \tag{2.15}$$

In the proof of (i), we have shown that $\lim_{k\to\infty} \mathscr{L}(z^k) = \bar{\mathscr{L}}$, which, together with (2.15), implies that $\mathcal{L}(\bar{x}, \bar{y}) = \bar{\mathscr{L}}$. Since $(\bar{x}, \bar{y})$ is arbitrarily chosen in C, Item (iv) is obtained. □

2.5. Global convergence and rates under the KŁ assumption

In this subsection, we will establish the global convergence and convergence rates of Algorithm 1 under the KŁ property, which has been extensively studied in recent years for the convergence of algorithms for nonconvex optimization (see, e.g., [5, 40, 45]).

Given a proper lower semicontinuous function f and the real numbers a, b, let us denote $[a < f < b] := \{x \in \mathbb{R}^n : a < f(x) < b\}$.

Definition 1. *A proper lower semicontinuous function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is said to have the* Kurdyka–Łojasiewicz (KŁ) property *at $\bar{x} \in \text{dom}\partial f := \{x \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$, and a continuous concave function $\varphi : [0, \eta) \to [0, +\infty)$ such that*

(i) $\varphi(0) = 0$;
(ii) φ is continuously differentiable and $\varphi' > 0$ on $(0, \eta)$;
(iii) For all $x \in U \cap [0 < f - f(\bar{x}) < \eta]$, the following KŁ inequality holds:

$$\varphi'(f(x) - f(\bar{x})) \cdot \text{dist}(0, \partial f(x)) \geq 1. \tag{2.16}$$

A proper lower semicontinuous function f, which has the KŁ property at every point of $\text{dom}\partial f$, is called a *KŁ function*. When $\varphi(s) = \sigma s^{1-\theta}$, σ is a positive constant and $\theta \in [0, 1)$, f is said to satisfy the *Łojasiewicz property* with the *exponent* θ.

Remark 2. *It is known that the KŁ property is automatically satisfied at any noncritical point $x \in \mathbb{R}^n$ with a concave function $\varphi(s) = \sigma s$ [40, Section 3.2]. A very wide class of functions, such as nonsmooth semialgebraic functions, real subanalytic functions, and functions definable in an o-minimal structure, satisfy the KŁ property. In particular, for Problem* (1.1)*, $\mathscr{L}$ is considered to be a KŁ function if f and h are semialgebraic (or f is semialgebraic and h^* satisfies a growth condition [41, Section 5]). We refer the readers to [40–42, 46] for more properties and examples of KŁ functions.*

We now establish the global convergence of Algorithm 1.

Theorem 4. *Suppose that $\mathscr{L}$ is a KŁ function. Let Assumption 1 hold and let the sequence $\{(x^k, y^k)\}$ generated by Algorithm 1 be bounded. Then (x^k, y^k) converges to a critical point of $\mathcal{L}$ and*

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty, \quad \sum_{k=1}^{\infty} \|y^{k+1} - y^k\| < \infty.$$

Proof. In the proof of Proposition 3, it has been shown that

$$\lim_{k\to\infty} \mathscr{L}(z^k) = \bar{\mathscr{L}}, \tag{2.17}$$

where $\bar{\mathscr{L}}$ is the constant value of $\mathcal{L}$ over C.

If there exists a number $l_0 > 0$ such that $\mathscr{L}(z^{l_0}) = \bar{\mathscr{L}}$, which, together with Lemma 2, implies that $\mathscr{L}(z^k) = \bar{\mathscr{L}}$ and $x^k = x^{k+1}$ for any $k \geq l_0$. By (2.7), we have $y^k = y^{k+1}$ for any $k \geq l_0$. Thus, $(x^k, y^k) = (x^{k+1}, y^{k+1})$ for any $k \geq l_0$, which proves the claim.

Otherwise, since the sequence $\{\mathscr{L}(z^k)\}$ is nonincreasing by Lemma 2, it follows that $\mathscr{L}(z^k) > \bar{\mathscr{L}}$ for any $k > 0$. Relation (2.17) ensures that for any $\eta > 0$, there exists an integer $l_1 > 0$ such that $\mathscr{L}(z^k) < \bar{\mathscr{L}} + \eta$ for any $k \geq l_1$. Let Ω be the set of cluster points of $\{z^k\}$. Along the same lines as the proof of Proposition 3 (ii) and (iv), we find that the function $\mathscr{L}$ is constant on the nonempty compact set Ω and $\operatorname{dist}(z^k, \Omega) \to 0$ as $k \to \infty$. Thus, for any $\varepsilon > 0$, there exists $l_2 > 0$ such that for $k \geq l_2$, $\operatorname{dist}(z^k, \Omega) < \varepsilon$. Let $K_0 := \max\{l_1, l_2\}$. From the discussion above, one has that $z^k \in \{z : \operatorname{dist}(z, \Omega) < \varepsilon\} \cap [\bar{\mathscr{L}} < \mathscr{L} < \bar{\mathscr{L}} + \eta]$ for all $k \geq K_0$. Since $\mathscr{L}$ is a KŁ function, from the uniformized KŁ property [41, Lemma 6], there exists a continuous concave function φ such that for all $k \geq K_0$,

$$\varphi'(\mathscr{L}(z^k) - \bar{\mathscr{L}}) \cdot \operatorname{dist}(0, \partial\mathscr{L}(z^k)) \geq 1. \tag{2.18}$$

Using the concavity of φ yields

$$\varphi(\mathscr{L}(z^{k+1}) - \bar{\mathscr{L}}) \leq \varphi(\mathscr{L}(z^k) - \bar{\mathscr{L}}) + \varphi'(\mathscr{L}(z^k) - \bar{\mathscr{L}}) \cdot (\mathscr{L}(z^{k+1}) - \mathscr{L}(z^k)). \tag{2.19}$$

The bound (2.12) implies that

$$\operatorname{dist}(0, \partial\mathscr{L}(z^k)) \leq \gamma(\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|), \tag{2.20}$$

where $\gamma := \max\{\gamma_1, \gamma_2\}$. Combining (2.18)–(2.20) with Lemma 2, we obtain that $\mathcal{M}_{m,n} := \varphi(\mathscr{L}(z^m) - \bar{\mathscr{L}}) - \varphi(\mathscr{L}(z^n) - \bar{\mathscr{L}})$ satisfies

$$\begin{aligned} \mathcal{M}_{k,k+1} &\geq \varphi'(\mathscr{L}(z^k) - \bar{\mathscr{L}}) \cdot (\mathscr{L}(z^k) - \mathscr{L}(z^{k+1})) \geq \frac{\mathscr{L}(z^k) - \mathscr{L}(z^{k+1})}{\operatorname{dist}(0, \partial\mathscr{L}(z^k))} \\ &\geq \frac{c(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2)}{\gamma(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|)}, \end{aligned}$$

where c is the constant defined in (2.10). The inequality above is rewritten as

$$\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2 \leq \frac{\gamma}{c}\mathcal{M}_{k,k+1}(\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|). \tag{2.21}$$

This also indicates that

$$\begin{aligned}\|x^{k+1}-x^k\| &\leq \sqrt{\frac{\gamma}{c}\mathcal{M}_{k,k+1}(\|x^k-x^{k-1}\|+\|x^{k+1}-x^k\|)}\\ &\leq \frac{\gamma}{c}\mathcal{M}_{k,k+1}+\frac{1}{4}(\|x^k-x^{k-1}\|+\|x^{k+1}-x^k\|),\end{aligned}$$

which is equivalent to

$$\|x^{k+1}-x^k\| \leq \frac{2\gamma}{c}\mathcal{M}_{k,k+1}+\frac{1}{2}(\|x^k-x^{k-1}\|-\|x^{k+1}-x^k\|).$$

Summing up from $k=K_0$ to n with $n>K_0$, it follows that

$$\sum_{k=K_0}^{n}\|x^{k+1}-x^k\| \leq \frac{2\gamma}{c}\mathcal{M}_{K_0,n+1}+\frac{1}{2}\|x^{K_0}-x^{K_0-1}\|. \tag{2.22}$$

Similarly, from (2.21), we also have

$$\sum_{k=K_0}^{n}\|x^{k}-x^{k-1}\| \leq \frac{2\gamma}{c}\mathcal{M}_{K_0,n+1}+\frac{1}{2}\|x^{n+1}-x^{n}\|. \tag{2.23}$$

Summing (2.22) and (2.23), we obtain

$$\begin{aligned}\sum_{k=K_0}^{n}(\|x^{k+1}-x^k\|+\|x^k-x^{k-1}\|) &\leq \frac{4\gamma}{c}\mathcal{M}_{K_0,n+1}+\frac{1}{2}\|x^{K_0}-x^{K_0-1}\|+\frac{1}{2}\|x^{n+1}-x^n\|\\ &\leq \frac{4\gamma}{c}\varphi(\mathscr{L}(z^{K_0})-\bar{\mathscr{L}})+\frac{1}{2}\|x^{K_0}-x^{K_0-1}\|+\frac{1}{2}\|x^{n+1}-x^n\|,\end{aligned} \tag{2.24}$$

where the second inequality follows from the fact that $\varphi>0$ over $(0,\eta)$. Let $n\to\infty$ in (2.24), then by the first term of (2.13), we have

$$\sum_{k=K_0}^{\infty}\|x^{k+1}-x^k\|<+\infty,$$

which, together with (2.7) and (2.8), implies

$$\sum_{k=K_0}^{\infty}\|y^{k+1}-y^k\|<+\infty.$$

These two inequalities imply that (x^k,y^k) is a Cauchy sequence along the same line of analysis as [41, Theorem 1 (ii)]. Thus, the sequence (x^k,y^k) converges to a limit $(\bar{x},\bar{y})$ that is a critical point of $\mathcal{L}$ by Proposition 3 (iii). □

The convergence rates of the sequence $\{(x^k,y^k)\}$ under the Łojasiewicz exponent are provided in the following theorem which is proved in an analogous way to [42].

Theorem 5. *Assume that the sequence $\{(x^k,y^k)\}$ is bounded and $\mathscr{L}$ is a KŁ function with the Łojasiewicz exponent θ. Let $(\bar{x},\bar{y})$ be the limit of (x^k,y^k). Then, under Assumption 1, the following estimations hold:*

(i) If $\theta = 0$, the sequence $\{(x^k, y^k)\}$ converges in finite steps;

(ii) If $\theta \in (0, \frac{1}{2}]$, then there exist constants $\nu > 0$, $0 < \tau < 1$ and a positive integer K such that for $k \geq K$,

$$\|x^k - \bar{x}\| \leq \nu\tau^{k-K}, \quad \|y^k - \bar{y}\| \leq \nu'\tau^{k-K},$$

where $\nu' := \nu(1/\alpha + L)/\hat{\lambda}$ and $\hat{\lambda}$ is given in (2.8)*;*

(iii) If $\theta \in (\frac{1}{2}, 1)$, then there exist a constant $\mu > 0$ and a positive integer $\bar{K}$ such that for $k \geq \bar{K}$,

$$\|x^k - \bar{x}\| \leq \mu k^{-\frac{1-\theta}{2\theta-1}}, \quad \|y^k - \bar{y}\| \leq \mu' k^{-\frac{1-\theta}{2\theta-1}},$$

where $\mu' := \mu(1/\alpha + L)/\hat{\lambda}$.

Proof. See Appendix A.2. □

3. SPPDG for nonconvex finite-sum optimization

In this section, we consider to solve the nonconvex finite-sum optimization Problem (1.2). By combining Algorithm 1 with certain stochastic gradient estimators, we present a stochastic variant of PPDG, named SPPDG, and establish its almost sure convergence as well as its convergence rates.

3.1. The SPPDG algorithm

In Algorithm 2, we summarize the details of SPPDG for the nonconvex finite-sum optimization problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) + h(Ax), \quad \text{where } f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x).$$

In many applications, the number of components N can be very large, which makes the computation of the full gradient $\nabla f(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x)$ challenging. To circumvent this difficulty, we apply the stochastic gradient estimator $\widetilde{\nabla} f_k$ to approximate $\nabla f(x^k)$ in (3.1a). Hence, Algorithm 2 can be viewed as a stochastic approximate variant of Algorithm 1.

Algorithm 2: SPPDG

1 Initialization: Choose an initial point $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$, a constant $\alpha > 0$, and a positive definite matrix M.

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

3 Update x^k, y^k as follows:

$$\begin{cases} x^{k+1} = x^k - \alpha(\widetilde{\nabla} f_k + A^T y^k), & (3.1a) \\ y^{k+1} = \arg\min\limits_{y\in\mathbb{R}^m} \left\{ h^*(y) - \langle y, A(2x^{k+1} - x^k)\rangle + \frac{1}{2}\|y - y^k\|_M^2 \right\}, & (3.1b) \end{cases}$$

where $\widetilde{\nabla} f_k$ is a stochastic gradient estimator of $\nabla f(x^k)$.

4 Set $k \leftarrow k + 1$.

Let $\mathcal{F}_k$ be the σ-field generated by the random variables of the first k iterations of Algorithm 2 and $\mathbb{E}_k$ be the expectation conditioned on $\mathcal{F}_k$. Since x^k and y^k are both dependent on the random information $\{\widetilde{\nabla} f_0, \widetilde{\nabla} f_1, \ldots, \widetilde{\nabla} f_{k-1}\}$ of the first k iterations, the iterate (x^k, y^k) is $\mathcal{F}_k$-measurable .

In this paper, we will mainly focus on the variance-reduced stochastic gradient estimator $\widetilde{\nabla} f_k$ which is formally defined in [15, 36].

Definition 2. *The stochastic gradient estimator $\widetilde{\nabla} f_k$ is said to be* variance-reduced *if there exist constants $\sigma_1, \sigma_2, \sigma_\Lambda > 0$, $\rho \in (0, 1]$ and the $\mathcal{F}_k$-measurable non-negative random variables Λ_1^k, Λ_2^k of the form $\Lambda_1^k = \sum_{i=1}^t (v_k^i)^2, \Lambda_2^k = \sum_{i=1}^t v_k^i$ for some nonnegative random variables $v_k^i \in \mathbb{R}$ such that for any $k \geq 1$, the following hold:*

(i) The estimator $\widetilde{\nabla} f_k$ satisfies

$$\mathbb{E}_k[\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2] \leq \Lambda_1^k + \sigma_1(\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \|x^k - x^{k-1}\|^2) \tag{3.2}$$

and

$$\mathbb{E}_k[\|\widetilde{\nabla} f_k - \nabla f(x^k)\|] \leq \Lambda_2^k + \sigma_2(\mathbb{E}_k[\|x^{k+1} - x^k\|] + \|x^k - x^{k-1}\|). \tag{3.3}$$

(ii) The sequence $\{\Lambda_1^k\}$ decays geometrically

$$\mathbb{E}_k[\Lambda_1^{k+1}] \leq (1 - \rho)\Lambda_1^k + \sigma_\Lambda(\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \|x^k - x^{k-1}\|^2). \tag{3.4}$$

(iii) If $\{x^k\}$ satisfies $\lim_{k\to\infty} \mathbb{E}[\|x^k - x^{k-1}\|^2] = 0$, then $\mathbb{E}[\Lambda_1^k] \to 0$ and $\mathbb{E}[\Lambda_2^k] \to 0$ as $k \to \infty$.

Remark 3. *A variety of popular stochastic gradient estimators satisfy the conditions in Definition 2, such as, SAGA, SARAH, SAG, and SVRG. Combining* (3.2) *and* (3.4)*, for any $k \geq 1$, we have the following bound:*

$$\mathbb{E}_k[\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2] \leq \frac{1}{\rho}(\Lambda_1^k - \mathbb{E}_k[\Lambda_1^{k+1}]) + \kappa(\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \|x^k - x^{k-1}\|^2), \tag{3.5}$$

where $\kappa := \sigma_1 + \frac{\sigma_\Lambda}{\rho}$. The readers are referred to [15, 36] for a detailed description of the examples and properties of the variance-reduced stochastic gradient estimator.

In the rest of this section, we assume that $\widetilde{\nabla} f_k$ in Algorithm 2 is a variance-reduced gradient estimator satisfying the conditions of Definition 2, and let M be the matrix associated with αAA^T. We shall analyze the convergence of (x^k, y^k) generated by Algorithm 2 under the following assumption.

Assumption 2. *The assumption is the same as that in Assumption 1 except that Assumption 1(i) is replaced by that assumptions that the functions f_i, $i = 1, \ldots, N$ are L-smooth, and $\mathcal{L}$ in Assumption 1(ii) is replaced by $\mathcal{L}_s$ with*

$$\mathcal{L}_s(x, y) = \frac{1}{N}\sum_{i=1}^N f_i(x) + \langle y, Ax\rangle - h^*(y).$$

Unsurprisingly, we will observe that part of the convergence analysis of Algorithm 2 will be performed in a similar way to Algorithm 1 in Section 2. Without any confusion, some notations in Section 2 will be used again in this section.

3.2. Auxiliary lemmas

Let us first define the following Lyapunov function:

$$\mathscr{L}_s(x, y, u, v, w) := \mathcal{L}_s(x, y) - a\|x - u\|^2 + b\|x - v\|^2 + c\|v - w\|^2, \tag{3.6}$$

for any $x, u, v, w \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$. Here, with the step size α and the Lipschitz constant L of ∇f_i, the constants a, b, c are given by

$$a := e_0 + \frac{2\delta_2}{\alpha} + 2\delta_2\alpha\kappa,\ b := e_0 + \frac{9\alpha\kappa}{2\delta_2} + 2\delta_2\alpha\kappa + \frac{\kappa}{2\delta_1} + \frac{3\alpha L^2}{2\delta_2},\ c := \frac{3\alpha\kappa}{2\delta_2},$$

where

$$e_0 := \frac{1}{3\alpha} - \frac{\delta_1 + L}{6} - \frac{\kappa}{3\delta_1} - \frac{4\delta_2 L}{3} - \frac{4\delta_2}{3\alpha} - \frac{2\delta_2\alpha L^2}{3} - \frac{\alpha L^2}{2\delta_2} - \frac{2\alpha\kappa}{\delta_2} - \frac{8\delta_2\alpha\kappa}{3} \tag{3.7}$$

and $\kappa > 0$ is defined in (3.5). In addition, $\delta_1, \delta_2 > 0$ are properly selected constants such that $e_0 > 0$.

In this subsection, we mainly aim to develop the descent property corresponding to the Lyapunov function $\mathscr{L}_s$ in expectation. Following the same line as the proof of Lemma 1, we first derive the relation between $\mathrm{crit}\mathcal{L}_s$ and $\mathrm{crit}\mathscr{L}_s$.

Lemma 6. *For any $x, u, v, w \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $(x, y, u, v, w) \in \mathrm{crit}\mathscr{L}_s$ if and only if $u = v = w = x$, and $(x, y) \in \mathrm{crit}\mathcal{L}_s$.*

The following lemma gives a connection between the two sequences $\{y^k\}$ and $\{x^k\}$.

Lemma 7. *Suppose that Assumption 2 holds. Then, for $k \geq 1$, we have*

$$\begin{aligned}\mathbb{E}_k[\|A^T(y^{k+1} - y^k)\|^2] \leq \frac{4}{\rho}\mathbb{E}_k[\Lambda_1^{k+1} - \Lambda_1^{k+2}] + \frac{4}{\rho}(\Lambda_1^k - \mathbb{E}_k[\Lambda_1^{k+1}]) + 4\kappa\|x^k - x^{k-1}\|^2 \\ + 4\left(\frac{1}{\alpha^2} + \kappa\right)\mathbb{E}_k[\|x^{k+2} - x^{k+1}\|^2] + 4\left(\left(\frac{1}{\alpha} + L\right)^2 + 2\kappa\right)\mathbb{E}_k[\|x^{k+1} - x^k\|^2],\end{aligned}$$

where ρ and κ are defined in Definition 2 and (3.5)*, respectively.*

Proof. See Appendix A.3. □

For the sake of simplicity, define $z^k := (x^k, y^k, x^{k+1}, x^{k-1}, x^{k-2})$. Similar to the process of the convergence analysis in Section 2, we establish the following critical lemma on the descent property of the Lyapunov function $\mathscr{L}_s$.

Lemma 8. *Let Assumption 2 hold. Then, for any $k \geq 1$ and $\delta_1, \delta_2 > 0$, we have*

$$\mathbb{E}[\mathscr{L}^{\Lambda}_{s,k+1}] + \mathbb{E}[e_0(\|x^{k+2} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2)] \leq \mathbb{E}[\mathscr{L}^{\Lambda}_{s,k}], \tag{3.8}$$

where e_0 is defined in (3.7)*,*

$$\mathscr{L}^{\Lambda}_{s,k} := \mathscr{L}_s(z^k) + e_1\Lambda_1^{k+1} + e_2\Lambda_1^k + e_3\Lambda_1^{k-1},$$

and

$$e_1 := \frac{2\delta_2\alpha}{\rho},\ e_2 := \frac{2\delta_2\alpha}{\rho} + \frac{1}{2\delta_1\rho} + \frac{3\alpha}{2\delta_2\rho},\ e_3 := \frac{3\alpha}{2\delta_2\rho}.$$

Proof. See Appendix A.4. □

Remark 4. *As stated previously, the constant e_0 is guaranteed to be positive through a careful selection of δ_1, δ_2, and the step size α. For example, let $\delta_1 = 1$, $\delta_2 = \frac{1}{6}$, and $\alpha \in (0, 1/2(3 + 7L + 6\kappa))$, we have $e_0 > 0$ by a straightforward calculation. Thus, we can assume that e_0 is positive throughout the remainder of this section. Under this condition, Lemma 8 indicates that the sequence $\{\mathbb{E}[\mathscr{L}^{\Lambda}_{s,k}]\}$ is nonincreasing.*

Define

$$d^k := (d^k_1, d^k_2, d^k_3, d^k_4, d^k_5) \tag{3.9}$$

with

$$\begin{aligned}
&d^k_1 := \tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla f_i(x^k) + A^T y^k - 2a(x^k - x^{k+1}) + 2b(x^k - x^{k-1}),\\
&d^k_2 := Ax^k + M(y^k - y^{k-1}) - A(2x^k - x^{k-1}),\\
&d^k_3 := -2a(x^{k+1} - x^k),\ d^k_4 := 2b(x^{k-1} - x^k) + 2c(x^{k-1} - x^{k-2}),\ d^k_5 := 2c(x^{k-2} - x^{k-1}).
\end{aligned}$$

Noting that $g^k = -M(y^k - y^{k-1}) + A(2x^k - x^{k-1}) \in \partial h^*(y^k)$ from (2.4), we can easily check that $d^k \in \partial \mathscr{L}_s(z^k)$. In the following lemma, we derive a bound of d^k.

Lemma 9. *Let Assumption 2 be satisfied. It holds that*

$$\mathbb{E}[\|d^k\|^2] \le r(\mathbb{E}[\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2 + \|x^{k-1} - x^{k-2}\|^2 + \|y^k - y^{k-1}\|^2 + \Lambda^k_1]), \tag{3.10}$$

where $r := \max\{\gamma_3, \gamma_4, \gamma_5, 3\}$ with

$$\gamma_3 := \frac{3}{\alpha^2} + \frac{12a}{\alpha} + 16a^2 + 3\sigma_1,\ \gamma_4 := 20b^2 + 3\sigma_1 + 2\|A\|^2 + 2\|M\|^2,\ \gamma_5 := 12c^2,$$

where σ_1 is defined in Definition 2, and a, b, and c are given in (3.6).

Proof. See Appendix A.5. □

3.3. Convergence analysis

Now, with the help of the auxiliary lemmas established in the previous subsection, we demonstrate that the iterates $\{(x^k, y^k)\}$ of Algorithm 2 exhibit the following elementary convergence property under the assumption that $\{(x^k, y^k)\}$ is bounded almost surely (for short, a.s.). This assumption is also used by [45, 47] for studying stochastic optimization algorithms.

Proposition 10. *Suppose that $\{(x^k, y^k)\}$ is bounded almost surely. Then, under Assumption 2, we have*

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < \infty \text{ a.s.} \quad \text{and} \quad \sum_{k=0}^{\infty} \|y^{k+1} - y^k\|^2 < \infty \text{ a.s.}$$

Proof. Summing (3.8) over $k = 1, \ldots, n$ yields

$$e_0 \sum_{k=1}^{n} \mathbb{E}[\|x^{k+2} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2] \le \mathbb{E}[\mathscr{L}^{\Lambda}_{s,1}] - \mathbb{E}[\mathscr{L}^{\Lambda}_{s,n+1}]. \tag{3.11}$$

From Assumption 2, $\mathcal{L}_s$ is bounded from below, which, together with the almost sure boundedness of $\{x^k\}$, ensures that $\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}]$ is bounded from below. Since $\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}]$ is nonincreasing (cf. Remark 4), $\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}]$ converges to a finite value. From (3.11), it follows that

$$\sum_{k=1}^{\infty} \mathbb{E}[\|x^k - x^{k-1}\|^2] = \sum_{k=0}^{\infty} \mathbb{E}[\|x^{k+1} - x^k\|^2] < \infty. \tag{3.12}$$

This also implies that

$$\lim_{k\to\infty} \mathbb{E}[\|x^{k+1} - x^k\|^2] = 0 \tag{3.13}$$

and

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < \infty \quad \text{a.s.} \tag{3.14}$$

Furthermore, from Item (iii) in Definition 2, it follows that

$$\lim_{k\to\infty} \mathbb{E}[\Lambda_1^k] = 0 \text{ and } \lim_{k\to\infty} \mathbb{E}[\Lambda_2^k] = 0. \tag{3.15}$$

By Lemma 7 and (2.8), we have

$$\begin{aligned} \mathbb{E}_k[\|y^{k+1} - y^k\|^2] \leq{}& \frac{4}{\hat{\lambda}^2\rho}\left(\mathbb{E}_k[\Lambda_1^{k+1} - \Lambda_1^{k+2}] + (\Lambda_1^k - \mathbb{E}_k[\Lambda_1^{k+1}])\right) \\ &+ \frac{4}{\hat{\lambda}^2}\left(\left(\frac{1}{\alpha} + L\right)^2 + 2\kappa\right)\left(\mathbb{E}_k[\|x^{k+2} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2] + \|x^k - x^{k-1}\|^2\right). \end{aligned} \tag{3.16}$$

Taking expectation on both sides and summing it over $k = 1, \ldots, n$, we have

$$\begin{aligned} \sum_{k=1}^{n} \mathbb{E}[\|y^{k+1} - y^k\|^2] \leq{}& \frac{4}{\hat{\lambda}^2\rho}\mathbb{E}[\Lambda_1^1 + \Lambda_1^2 - \Lambda_1^{n+1} - \Lambda_1^{n+2}] \\ &+ \frac{4}{\hat{\lambda}^2}\left(\left(\frac{1}{\alpha} + L\right)^2 + 2\kappa\right)\sum_{k=1}^{n} \mathbb{E}[\|x^{k+2} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2]. \end{aligned} \tag{3.17}$$

Let $n \to \infty$. Then, by using (3.12) and (3.15), one has

$$\sum_{k=0}^{\infty} \mathbb{E}[\|y^{k+1} - y^k\|^2] < \infty, \tag{3.18}$$

which implies that

$$\lim_{k\to\infty} \mathbb{E}[\|y^{k+1} - y^k\|^2] = 0 \tag{3.19}$$

and

$$\sum_{k=0}^{\infty} \|y^{k+1} - y^k\|^2 < \infty \quad \text{a.s.} \tag{3.20}$$

The proof is completed. □

Remark 5. *Because of the random nature of $\widetilde{\nabla} f_k$, we can define a suitable sample space Ω based on the structure of Algorithm 2. The sequence $\{(x^k(\omega), y^k(\omega))\}$ with each sample $\omega \in \Omega$ then corresponds to the iterates generated by a single run of Algorithm 2. The sample space Ω can be equipped with a σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$ to form a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consequently, the assumption that $\{(x^k, y^k)\}$ is bounded almost surely implies that there is an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) = 1$ such that the sequence $\{(x^k(\omega), y^k(\omega))\}$ is bounded for every $\omega \in \mathcal{A}$.*

The following proposition establishes the subsequence convergence by showing that any cluster point of the sequence $\{(x^k, y^k)\}$ is a critical point of $\mathcal{L}_s$ with probability 1.

Proposition 11. *Let Assumption 2 be satisfied and let $\{(x^k, y^k)\}$ be bounded almost surely. Then, there exists an event $\mathcal{A}$ with a measure 1 such that, for all $\omega \in \mathcal{A}$, the following statements hold:*

(i) The set C_ω containing all cluster points of $\{(x^k(\omega), y^k(\omega))\}$ is nonempty and compact, and $\operatorname{dist}((x^k(\omega), y^k(\omega)), C_\omega) \to 0$;
(ii) $C_\omega \subseteq \operatorname{crit}\mathcal{L}_s$;
(iii) $\mathcal{L}_s$ is finite and constant on C_ω.

Proof. See Appendix A.6. □

Remark 6. *Under the assumptions in Proposition 11, from Item (i) and Item (iii), there exists an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) = 1$ such that for all $\omega \in \mathcal{A}$, $\operatorname{dist}((x^k(\omega), y^k(\omega)), C_\omega) \to 0$, and $\mathcal{L}_s$ equals a constant value $\bar{\mathcal{L}}_{s,\omega}$ over C_ω. Hence, it follows that $\mathbb{E}[\mathcal{L}_s(x^k, y^k)] \to \bar{\mathcal{L}}_s$ with $\bar{\mathcal{L}}_s := \mathbb{E}[\bar{\mathcal{L}}_{s,\omega}]$. It also follows from* (3.6) *and $z^k = (x^k, y^k, x^{k+1}, x^{k-1}, x^{k-2})$ that*

$$\mathscr{L}_s(z^k) = \mathcal{L}_s(x^k, y^k) - a\|x^k - x^{k+1}\|^2 + b\|x^k - x^{k-1}\|^2 + c\|x^{k-1} - x^{k-2}\|^2,$$

which, together with (3.13)*, implies that $\mathbb{E}[\mathscr{L}_s(z^k)] \to \bar{\mathcal{L}}_s$ as $k \to \infty$.*

We now present the main theorem of this section about the finite length property and the almost sure convergence of the whole sequence $\{(x^k, y^k)\}$ generated by Algorithm 2 depending on the KŁ property of the Lyapunov function $\mathscr{L}_s$.

Theorem 12. *Suppose that Assumption 2 holds and that $\mathscr{L}_s$ is a KŁ function with the Łojasiewicz exponent $\theta \in [0, 1)$. Let the sequence $\{(x^k, y^k)\}$ be bounded almost surely. The following then hold:*

(i) It holds that

$$\sum_{k=0}^{\infty} \mathbb{E}[\|x^{k+1} - x^k\|] < \infty, \quad \sum_{k=0}^{\infty} \mathbb{E}[\|y^{k+1} - y^k\|] < \infty;$$

(ii) The sequence $\{(x^k, y^k)\}$ converges almost surely to a random vector $(\bar{x}, \bar{y})$, and $(\bar{x}, \bar{y}) \in \operatorname{crit}\mathcal{L}_s$ a.s.

Proof. Let us begin with the proof of the simple fact that $\sum_{k=0}^{\infty} \mathbb{E}[\|x^{k+1} - x^k\|] < \infty$ and $\sum_{k=0}^{\infty} \mathbb{E}[\|y^{k+1} - y^k\|] < \infty$ if $\sum_{k=0}^{\infty} \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty$. In other words, if this fact is true, in order to derive Item (i), it is sufficient to prove $\sum_{k=0}^{\infty} \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty$. Indeed, by Jensen's inequality, the fact that $\sum_{k=0}^{\infty} \mathbb{E}[\|x^{k+1} - x^k\|] < \infty$ is obvious. By (3.16) (with $k = k - 1$) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, there exists a constant $\gamma_6 > 0$ such that

$$\begin{aligned} &\sqrt{\mathbb{E}[\|y^k - y^{k-1}\|^2]} \\ &\le \gamma_6\big(\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} + \sqrt{\mathbb{E}[\Lambda_1^{k-1}]}\big). \end{aligned} \tag{3.21}$$

Using (3.4), $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and $\sqrt{1-\rho} \leq 1 - \frac{\rho}{2}$, it follows that

$$\begin{aligned}\sqrt{\mathbb{E}[\Lambda_1^k]} &\leq \sqrt{(1-\rho)\mathbb{E}[\Lambda_1^{k-1}] + \sigma_\Lambda(\mathbb{E}[\|x^k - x^{k-1}\|^2] + \mathbb{E}[\|x^{k-1} - x^{k-2}\|^2])} \\ &\leq (1 - \frac{\rho}{2})\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} + \sqrt{\sigma_\Lambda \mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\sigma_\Lambda \mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]}.\end{aligned} \tag{3.22}$$

Rearranging this inequality, we obtain

$$\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} \leq \frac{2}{\rho}(\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} - \sqrt{\mathbb{E}[\Lambda_1^k]}) + \frac{2}{\rho}\sqrt{\sigma_\Lambda \mathbb{E}[\|x^k - x^{k-1}\|^2]} + \frac{2}{\rho}\sqrt{\sigma_\Lambda \mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]}. \tag{3.23}$$

Therefore, by substituting (3.23) into (3.21), we have

$$\sum_{k=0}^{\infty} \mathbb{E}[\|y^{k+1} - y^k\|] \leq \sum_{k=0}^{\infty} \sqrt{\mathbb{E}[\|y^{k+1} - y^k\|^2]} < \infty.$$

Hence, the simple fact is proved.

We next prove that $\sum_{k=0}^{\infty} \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty$. If $\mathscr{L}_s$ is a KŁ function with the exponent θ, an integer K_0 and a function $\varphi_0(s) = \sigma_0 s^{1-\theta}$ exist such that the following holds

$$\varphi_0'(\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k})\mathbb{E}[\text{dist}(0, \partial\mathscr{L}_s(z^k))] \geq 1, \ \forall k \geq K_0, \tag{3.24}$$

where $\{\bar{\mathscr{L}}_{s,k}\}$ is a nondecreasing sequence satisfying $\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k} > 0$ [36, Lemma 4.5] and converging to a finite value $\bar{\mathcal{L}}_s$ that is given in Remark 6.

When $\theta = 0$, we show that $\mathbb{E}[\mathscr{L}_{s,k}^\Lambda] = \bar{\mathcal{L}}_s$ holds after a finite number of iterations by contradiction. Otherwise, Inequality (3.24) implies that

$$\mathbb{E}[\text{dist}(0, \partial\mathscr{L}_s(z^k))] \geq \frac{1}{\sigma_0}, \ \forall k \geq K_0. \tag{3.25}$$

From (3.25), (3.10), and Jensen's inequality, we have

$$\tfrac{1}{\sigma_0^2} \leq (\mathbb{E}[\text{dist}(0, \partial\mathscr{L}_s(z^k))])^2 \leq r(\mathbb{E}[\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2 + \|x^{k-1} - x^{k-2}\|^2 + \|y^k - y^{k-1}\|^2 + \Lambda_1^k]).$$

Applying this inequality to (3.8), we have

$$\begin{aligned}\mathbb{E}[\mathscr{L}_{s,k}^\Lambda] &\leq \mathbb{E}[\mathscr{L}_{s,k-1}^\Lambda] - e_0\mathbb{E}[\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2 + \|x^{k-1} - x^{k-2}\|^2] \\ &\leq \mathbb{E}[\mathscr{L}_{s,k-1}^\Lambda] - \tfrac{e_0}{r\sigma_0^2} + e_0\mathbb{E}[\|y^k - y^{k-1}\|^2] + e_0\mathbb{E}[\Lambda_1^k],\end{aligned}$$

which is impossible after a large enough number of iterations by noticing that $\mathbb{E}[\|y^k - y^{k-1}\|^2] \to 0$ (cf. (3.19)), $\mathbb{E}[\Lambda_1^k] \to 0$ (cf. (3.15)), and $\mathbb{E}[\mathscr{L}_{s,k}^\Lambda] \to \bar{\mathcal{L}}_s$ (cf. Remark 6). Therefore, there exists an integer $\bar{K} \geq 0$ such that $\mathbb{E}[\mathscr{L}_{s,k}^\Lambda] = \bar{\mathcal{L}}_s$ holds for $k \geq \bar{K}$. In view of (3.8), we have $\mathbb{E}[\|x^k - x^{k-1}\|^2] = 0$ for $k \geq \bar{K}$, and hence $\sum_{k=0}^{\infty} \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty$.

We now consider $\theta \in [\frac{1}{2}, 1)$. By (3.10), Jensen's inequality, and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, it holds that

$$\begin{aligned}\mathbb{E}[\text{dist}(0, \partial\mathscr{L}_s(z^k))] \leq \sqrt{r}(&\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|y^k - y^{k-1}\|^2]} \\ &+ \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} + \sqrt{\mathbb{E}[\Lambda_1^k]}).\end{aligned} \tag{3.26}$$

Substituting (3.21) into (3.26), we then have

$$\begin{aligned}\mathbb{E}[\mathrm{dist}(0,\partial\mathscr{L}_s(z^k))] \le\ & \gamma_6\sqrt{r}\sqrt{\mathbb{E}[\Lambda_1^{k-1}]}+\sqrt{r\mathbb{E}[\Lambda_1^k]}\\ &+(\sqrt{r}+\gamma_6\sqrt{r})\Big(\sqrt{\mathbb{E}[\|x^{k+1}-x^k\|^2]}+\sqrt{\mathbb{E}[\|x^k-x^{k-1}\|^2]}+\sqrt{\mathbb{E}[\|x^{k-1}-x^{k-2}\|^2]}\Big).\end{aligned}\tag{3.27}$$

Applying (3.23) to the last two terms in (3.27) and letting $\gamma := \sqrt{r}+\gamma_6\sqrt{r}+\frac{2\sqrt{r\sigma_\Lambda}}{\rho}+\frac{2\gamma_6\sqrt{r\sigma_\Lambda}}{\rho}$, one has

$$\begin{aligned}\mathbb{E}[\mathrm{dist}(0,\partial\mathscr{L}_s(z^k))] \le\ & \gamma\Big(\sqrt{\mathbb{E}[\|x^{k+1}-x^k\|^2]}+\sqrt{\mathbb{E}[\|x^k-x^{k-1}\|^2]}+\sqrt{\mathbb{E}[\|x^{k-1}-x^{k-2}\|^2]}\Big)\\ &+\frac{2\gamma_6\sqrt{r}}{\rho}\Big(\sqrt{\mathbb{E}[\Lambda_1^{k-1}]}-\sqrt{\mathbb{E}[\Lambda_1^k]}\Big)+\frac{2\sqrt{r}}{\rho}\Big(\sqrt{\mathbb{E}[\Lambda_1^k]}-\sqrt{\mathbb{E}[\Lambda_1^{k+1}]}\Big).\end{aligned}$$

Let Σ_k denote the right-hand side of the inequality above. Obviously, $\Sigma_k > 0$. Combining the inequality $\mathbb{E}[\mathrm{dist}(0,\partial\mathscr{L}_s(z^k))] \le \Sigma_k$ with (3.24) and $\varphi_0(s) = \sigma_0 s^{1-\theta}$ gives

$$\frac{\sigma_0(1-\theta)\Sigma_k}{(\mathbb{E}[\mathscr{L}_s(z^k)]-\bar{\mathscr{L}}_{s,k})^\theta}\ge 1,\ \forall k\ge K_0.\tag{3.28}$$

Note that for $\theta\in[\frac{1}{2},1)$, there are positive constants β_0, κ_2,κ_3, and a sufficiently large integer $K_1 > 0$ such that for $k \ge K_1$,

$$\begin{aligned}&(\mathbb{E}[e_1\Lambda_1^{k+1}+e_2\Lambda_1^k+e_3\Lambda_1^{k-1}])^\theta \le \kappa_2(\mathbb{E}[\Lambda_1^{k+1}+\Lambda_1^k+\Lambda_1^{k-1}])^\theta \le \kappa_2\sqrt{\mathbb{E}[\Lambda_1^{k+1}+\Lambda_1^k+\Lambda_1^{k-1}]}\\ &\le \kappa_3(\sqrt{\mathbb{E}[\Lambda_1^k]}+\sqrt{\mathbb{E}[\Lambda_1^{k-1}]}+\sqrt{\mathbb{E}[\|x^{k+1}-x^k\|^2]}+\sqrt{\mathbb{E}[\|x^k-x^{k-1}\|^2]})\le\beta_0\Sigma_k,\end{aligned}$$

where the second inequality is deduced from $\mathbb{E}[\Lambda_1^k]\to 0$ for $k\to\infty$ (cf. (3.15)), the third inequality is obtained by $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, and (3.22), and the last inequality is from (3.23) and the definition of Σ_k. Take a constant $\beta > 0$ such that $\beta\sigma_0(1-\theta)\ge\sigma_0(1-\theta)+\beta_0$. From (3.28) and the fact that $(a+b)^\theta\le a^\theta+b^\theta$ for $\theta\in[\frac{1}{2},1]$, it then holds for $k\ge K := \max\{K_0,K_1\}$,

$$\begin{aligned}\frac{\beta\sigma_0(1-\theta)\Sigma_k}{(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})^\theta} &\ge \frac{\beta\sigma_0(1-\theta)\Sigma_k}{(\mathbb{E}[\mathscr{L}_s(z^k)]-\bar{\mathscr{L}}_{s,k})^\theta+(\mathbb{E}[e_1\Lambda_1^{k+1}+e_2\Lambda_1^k+e_3\Lambda_1^{k-1}])^\theta}\\ &\ge \frac{\beta\sigma_0(1-\theta)\Sigma_k}{\sigma_0(1-\theta)\Sigma_k+\beta_0\Sigma_k}\ge 1.\end{aligned}\tag{3.29}$$

Let $\varphi_1(s) := \beta\sigma_0 s^{1-\theta}$, and for any $k\ge K$, (3.29) is rewritten as

$$\varphi_1'(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})\Sigma_k\ge 1.\tag{3.30}$$

Since φ_1 is concave, we have

$$\begin{aligned}&\varphi_1(\mathbb{E}[\mathscr{L}_{s,k+1}^\Lambda]-\bar{\mathscr{L}}_{s,k+1})\\ &\le\varphi_1(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})+\varphi_1'(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})\mathbb{E}[\mathscr{L}_{s,k+1}^\Lambda-\bar{\mathscr{L}}_{s,k+1}-\mathscr{L}_{s,k}^\Lambda+\bar{\mathscr{L}}_{s,k}]\\ &\le\varphi_1(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})+\varphi_1'(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})\mathbb{E}[\mathscr{L}_{s,k+1}^\Lambda-\mathscr{L}_{s,k}^\Lambda]\\ &\le\varphi_1(\mathbb{E}[\mathscr{L}_{s,k}^\Lambda]-\bar{\mathscr{L}}_{s,k})-\frac{e_0}{\Sigma_k}\mathbb{E}[\|x^{k+2}-x^{k+1}\|^2+\|x^{k+1}-x^k\|^2+\|x^k-x^{k-1}\|^2],\end{aligned}\tag{3.31}$$

where the second inequality is obtained by $\bar{\mathscr{L}}_{s,k} \leq \bar{\mathscr{L}}_{s,k+1}$, the third inequality is from Lemma 8 and (3.30). Let $\mathcal{M}_{m,n} := \varphi_1(\mathbb{E}[\mathscr{L}^{\Lambda}_{s,m}] - \bar{\mathscr{L}}_{s,m}) - \varphi_1(\mathbb{E}[\mathscr{L}^{\Lambda}_{s,n}] - \bar{\mathscr{L}}_{s,n})$. Then, (3.31) implies

$$\mathcal{M}_{k,k+1} \geq \frac{e_0}{\Sigma_k}\mathbb{E}[\|x^{k+2} - x^{k+1}\|^2].$$

Rewriting this inequality and using $4\sqrt{ab} \leq a/\gamma + 4\gamma b$ for any $\gamma > 0$ yields

$$4\sqrt{\mathbb{E}[\|x^{k+2} - x^{k+1}\|^2]} \leq 4\sqrt{\frac{\mathcal{M}_{k,k+1}\Sigma_k}{e_0}} \leq \frac{\Sigma_k}{\gamma} + \frac{4\gamma\mathcal{M}_{k,k+1}}{e_0},$$

which, together with the definition of Σ_k, gives

$$\begin{aligned} 4\sqrt{\mathbb{E}[\|x^{k+2} - x^{k+1}\|^2]} \leq\ & \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} \\ & + \frac{4\gamma\mathcal{M}_{k,k+1}}{e_0} + \frac{2\gamma_6\sqrt{r}}{\rho\gamma}\left(\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} - \sqrt{\mathbb{E}[\Lambda_1^{k}]}\right) + \frac{2\sqrt{r}}{\rho\gamma}\left(\sqrt{\mathbb{E}[\Lambda_1^{k}]} - \sqrt{\mathbb{E}[\Lambda_1^{k+1}]}\right). \end{aligned}$$

Summing up from $k = K$ to n, we have

$$\begin{aligned} \sum_{k=K}^{n}\sqrt{\mathbb{E}[\|x^{k+2} - x^{k+1}\|^2]} \leq\ & 3\sqrt{\mathbb{E}[\|x^{K+1} - x^K\|^2]} + 2\sqrt{\mathbb{E}[\|x^K - x^{K-1}\|^2]} \\ & + \sqrt{\mathbb{E}[\|x^{K-1} - x^{K-2}\|^2]} + \sum_{k=K}^{n}\frac{4\gamma\mathcal{M}_{k,k+1}}{e_0} + \frac{2\gamma_6\sqrt{r}}{\rho\gamma}\sqrt{\mathbb{E}[\Lambda_1^{K-1}]} + \frac{2\sqrt{r}}{\rho\gamma}\sqrt{\mathbb{E}[\Lambda_1^{K}]}. \end{aligned} \tag{3.32}$$

By the definition of $\mathcal{M}_{k,k+1}$, it holds that $\sum_{k=K}^{n}\mathcal{M}_{k,k+1} = \mathcal{M}_{K,n+1} \leq \varphi_1(\mathbb{E}[\mathscr{L}^{\Lambda}_{s,K}] - \bar{\mathscr{L}}_{s,K})$. If we let $n \to \infty$ in (3.32), it then follows that

$$\sum_{k=0}^{\infty}\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty.$$

For $\theta \in (0, \frac{1}{2})$, we show that it can be reduced to the case that $\theta = \frac{1}{2}$. Indeed, from Remark 6, we can let K_0 be large enough such that $\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k} < 1$. Since (3.24) holds with $\theta \in (0, \frac{1}{2})$, we can see that (3.24) also holds with $\theta = \frac{1}{2}$. Thus, the claim follows immediately from the analysis for the case that $\theta \in [\frac{1}{2}, 1)$.

Combining these results, we obtain $\sum_{k=0}^{\infty}\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} < \infty$ for all $\theta \in [0, 1)$, and hence Item (i) is derived by the previously mentioned simple fact.

In the proof of Proposition 11, we have shown that there exists an event $\mathcal{A}$ with a measure 1 such that, for any $\omega \in \mathcal{A}$, every convergent subsequence of $\{(x^k(\omega), y^k(\omega))\}$ converges to a point $(\bar{x}(\omega), \bar{y}(\omega))$ belonging to $\mathrm{crit}\mathcal{L}_s$. If follows from Item (i) that

$$\sum_{k=0}^{\infty}\|x^{k+1} - x^k\| < \infty \text{ a.s.}, \quad \sum_{k=0}^{\infty}\|y^{k+1} - y^k\| < \infty \text{ a.s.},$$

and consequently

$$\sum_{k=0}^{\infty}\|x^{k+1}(\omega) - x^k(\omega)\| < \infty, \quad \sum_{k=0}^{\infty}\|y^{k+1}(\omega) - y^k(\omega)\| < \infty.$$

In other words, $\{(x^k(\omega), y^k(\omega))\}$ is a Cauchy sequence. Thus, the sequence $\{(x^k(\omega), y^k(\omega))\}$ converges to $(\bar{x}(\omega), \bar{y}(\omega))$. Therefore, there is a random vector $(\bar{x}, \bar{y})$ such that $\{(\bar{x}, \bar{y})\} \in \text{crit}\mathcal{L}_s$ a.s. and $\{(x^k, y^k)\}$ converges almost surely to $(\bar{x}, \bar{y})$. Item (ii) is proved. □

Finally, we establish the convergence rates of the sequence $\{(x^k, y^k)\}$ under the Łojasiewicz exponent in the following theorem.

Theorem 13. *Suppose that Assumption 2 is satisfied and that $\mathscr{L}_s$ is a KŁ function with the Łojasiewicz exponent $\theta \in [0, 1)$. Let the sequence $\{(x^k, y^k)\}$ be bounded almost surely and let $\{(x^k, y^k)\}$ converge almost surely to some random vector $(\bar{x}, \bar{y})$. The following statements then hold:*

(i) If $\theta = 0$, the sequence $\{(x^k, y^k)\}$ converges in expectation after finite steps;

(ii) If $\theta \in (0, \frac{1}{2}]$, then there exist constants $\nu, \bar{\nu} > 0$, $\tau, \bar{\tau} \in (0, 1)$ and a sufficiently large integer K such that for $k \geq K$,

$$\mathbb{E}[\|x^k - \bar{x}\|] \leq \nu\tau^{k-K}, \quad \mathbb{E}[\|y^k - \bar{y}\|] \leq \bar{\nu}\bar{\tau}^{k-K};$$

(iii) If $\theta \in (\frac{1}{2}, 1)$, then there exist constants $\mu, \bar{\mu} > 0$ and a sufficiently large integer $\bar{K}$ such that for $k \geq \bar{K}$,

$$\mathbb{E}[\|x^k - \bar{x}\|] \leq \mu k^{-\frac{1-\theta}{2\theta-1}}, \quad \mathbb{E}[\|y^k - \bar{y}\|] \leq \bar{\mu} k^{-\frac{1-\theta}{2\theta-1}}.$$

Proof. See Appendix A.7. □

4. Preliminary numerical experiments

In this section, we show the efficiency of our proposed algorithms, and compare them with several state-of-the-art algorithms on a variety of test problems. All numerical experiments are carried out using MATLAB R2023a on a desktop computer with an Intel Core i5 with 2.5GHz and 32GB memory.

4.1. Deep learning for image classification

In this subsection, we use a one-hidden layer deep neural network for image classification using the dataset `CIFAR-10`*, which consists of 60,000 color images of size 32×32 divided into 10 classes. Within this dataset, 50,000 images have already been designated for training, while the remaining 10,000 are reserved for testing. The one-hidden-layer neural network consists of an input layer, a hidden layer, and an output layer. The hidden layer size is 175. We emphasize that this experiment serves only as a simple proof-of-concept to demonstrate the efficiency of our algorithm.

We adopt the following notation:

- N: the number of input samples, m: the number of neurons in the hidden layer, d: the dimension of each input sample;
- $x_i \in \mathbb{R}^d$: the i-th input sample, $i = 1, \ldots, N$, $z_{ij} \in \mathbb{R}$: the j-th element of the actual output, $y_{ij} \in \mathbb{R}$: the j-th element of the desired output, $j = 1, \ldots, 10$;
- w_{kl}: the weights of the connections between the input nodes and the hidden layer, v_{jk}: the weights of the connections between the hidden layer neurons and the output neuron, b_k, b_j: the bias parameters of the hidden layer and output layer, $k = 1, \ldots, m$, $j = 1, \ldots, 10$.

*The dataset can be found in `https://www.cs.toronto.edu/~kriz/cifar.html`

Training the neural network amounts to obtaining the value of the model parameter (w, b) such that, for each point of input data x, the output z of the model predicts the real value y with satisfying accuracy. To achieve this, it is necessary to solve the following finite-sum optimization problem:

$$\min_{w,b} \frac{1}{N} \sum_{i=1}^{N} f_i(w, b) + \lambda(\|w\|_1 + \|b\|_1), \tag{4.1}$$

where

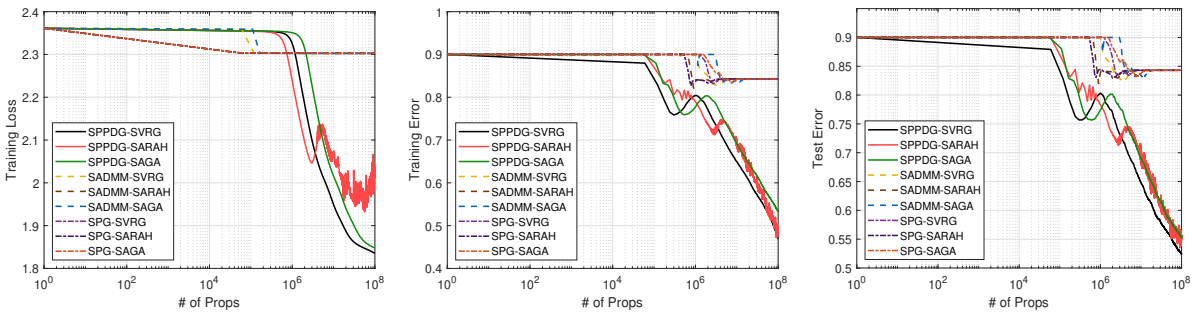
$$f_i(w, b) = -\sum_{j=1}^{10} y_{ij} \log(z_{ij})$$

is the cross-entropy loss function,

$$z_{ij} = g_2\left(\sum_{k=1}^{m} v_{jk} g_1\left(\sum_{l=1}^{d} w_{kl} x_{il} + b_k\right) + b_j\right),$$

and $\lambda > 0$ represents a regularization parameter. Here, g_1 is the sigmoid activation function and g_2 is the softmax activation function.

We choose a normalized vector drawn from the standard normal distribution as the initial point x^0 and set $\lambda = 1e{-}4$. Alongside the SAGA, SVRG, and SARAH estimators, we apply SPPDG (Algorithm 2), the stochastic linearized ADMM (SADMM) proposed recently by [15], and the stochastic proximal gradient method (SPG) to train the neural network on the training set. After training, we evaluate the classification performance of this neural network on the test set. The numerical results are presented in Figure 1, which displays the training loss, training error, and test error as functions of the total number of propagations for all methods. By observing this figure, all three SPPDG methods obviously outperform the methods associated with SADMM and SPG, and the gradient estimator SVRG seems to be more competitive than SAGA and SARAH.



(a) Training loss vs Props. **(b)** Training error vs Props. **(c)** Test error vs Props.

Figure 1. Comparison of SPPDG, SADMM, and SPG for image classification using a one-hidden-layer neural network.

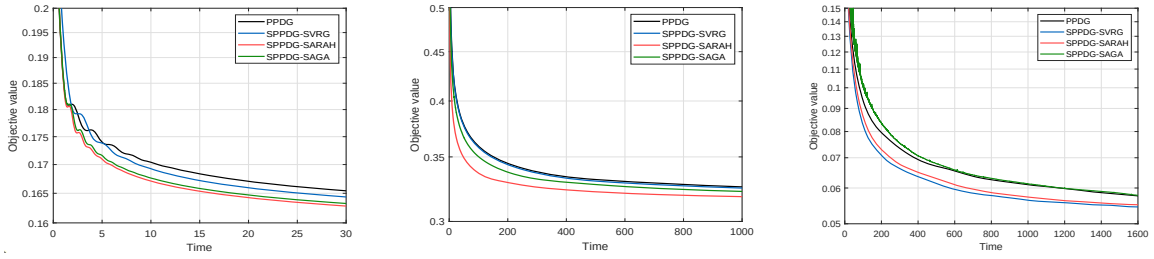
4.2. Nonconvex graph-guided fused lasso

In this subsection, we consider the following problem:

$$\min_{x \in \hat{\mathcal{D}}} \frac{1}{N} \sum_{i=1}^{N} f_i(x) + \lambda \|Ax\|_p^p. \tag{4.2}$$

Here, $A = [V; I] \in \mathbb{R}^{m\times n}$, with $V \in \mathbb{R}^{n\times n}$ being the sparsity pattern of the graph obtained by sparse inverse covariance estimation [48], the set $\hat{\mathcal{D}}$ is defined as $\hat{\mathcal{D}} = \{x \in \mathbb{R}^n : \|Ax\|_\infty \leq r\}$, $f_i(x) = 1 - \tanh(b_i \cdot \langle a_i, x\rangle)$ is the sigmoid loss function which is nonconvex, and $\|u\|_p$, $p \in (0, 1)$ is the ℓ_p-norm. Evidently, Problem (4.2) can be categorized as an instance of the fully nonconvex finite-sum optimization Problem (1.2) with $h(u) = \lambda\|u\|_p^p + \mathcal{I}_{\mathcal{D}}(u)$, $\mathcal{D} = \{u : \|u\|_\infty \leq r\}$. In what follows, we choose $\lambda = 1e-4$ and $r = 1$.

In this experiment, we test Problem (4.2) on the datasets `CINA`[†], `MNIST`[‡], and `gisette` [49]. To apply our algorithms, PPDG (Algorithm 1) and SPPDG (Algorithm 2), we should calculate $\text{prox}_{h^*}^M(y^k + M^{-1}A(2x^{k+1} - x^k))$ with $M = \alpha AA^T$ at each iteration. However, since the extended proximal mapping $\text{prox}_{h^*}^M$ is difficult to obtain directly, in practice, we calculate the term $\text{prox}_{\beta h^*}\left(y^k + \beta A(2x^{k+1} - x^k)\right)$ as an approximation, where $\beta = 1/(\alpha\|A\|^2)$. The numerical result is displayed in Figure 2 with the initial point $x^0 = 0$, $q = 0.5$ and a fixed mini-batch sample size $\lfloor 0.01N \rfloor$. Because of the fully nonconvex structure and the existence of the linear operator A, Problem (4.2) cannot be solved by SADMM and SPG directly as in the previous subsection. However, we can observe from Figure 2 that both PPDG and SPPDG with the gradient estimators SAGA, SVRG, and SARAH are able to solve Problem (4.2) efficiently. Moreover, with the same CPU time, SPPDG outperforms PPDG significantly, and among the stochastic variants, SPPDG-SARAH exhibits better performance than SPPDG-SVRG and SPPDG-SAGA.



(a) Obj. vs Time (CINA) **(b)** Obj. vs Time (MNIST) **(c)** Obj. vs Time (gisette)

Figure 2. PPDG and SPPDG for the nonconvex graph-guided fused lasso problem.

5. Conclusions

In this paper, we delve into an exploration of the first-order primal–dual methods for composite optimization and nonconvex finite-sum optimization in the fully nonconvex setting. Inspired by the existing first-order primal–dual methods for convex optimization, with the help of conjugate duality, we propose a preconditioned primal–dual gradient method and its stochastic approximate variant. The proposed methods are shown to be effective on a variety of nonconvex applications.

Use of Generative-AI tools declaration

The author declares she has not used Artificial Intelligence (AI) tools in the creation of this article.

[†]The dataset is available in `http://www.causality.inf.ethz.ch/data/CINA.html`
[‡]The dataset is available in `http://yann.lecun.com/exdb/mnist`

Acknowledgments

This research was funded by project ZR2025QC1919Z supported by Shandong Provincial Natural Science Foundation.

Conflict of interest

The author declares no conflict of interest in this paper.

References

1. F. Wen, L. Chu, P. Liu, R. C. Qiu, A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning, *IEEE Access*, **6** (2018), 69883–69906. https://doi.org/10.1109/ACCESS.2018.2880454
2. H. Attouch, J. Bolte, B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Math. Program.*, **137** (2013), 91–129. https://doi.org/10.1007/s10107-011-0484-9
3. G. Li, T. K. Pong, Global convergence of splitting methods for nonconvex composite optimization, *SIAM J. Optim.*, **25** (2015), 2434–2460. https://doi.org/10.1137/140998135
4. A. Themelis, L. Stella, P. Patrinos, Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms, *SIAM J. Optim.*, **28** (2018), 2274–2303. https://doi.org/10.1137/16M1080240
5. J. Bolte, S. Sabach, M. Teboulle, Nonconvex Lagrangian-based optimization: Monitoring schemes and global convergence, *Math. Oper. Res.*, **43** (2018), 1210–1232. https://doi.org/10.1287/moor.2017.0900
6. R. I. Boţ, D. K. Nguyen, The proximal alternating direction method of multipliers in the nonconvex setting: Convergence analysis and rates, *Math. Oper. Res.*, **45** (2020), 682–712. https://doi.org/10.1287/moor.2019.1008
7. A. M. Tillmann, D. Bienstock, A. Lodi, A. Schwartz, Cardinality minimization, constraints, and regularization: A survey, *SIAM Rev.*, **66** (2024), 403–477. https://doi.org/10.1137/21M142770X
8. R. H. Byrd, J. Nocedal, F. Oztoprak, An inexact successive quadratic approximation method for ℓ_1 regularized optimization, *Math. Program.*, **157** (2016), 375–396. https://doi.org/10.1007/s10107-015-0941-y
9. S. Bonettini, I. Loris, F. Porta, M. Prato, S. Rebegoldi, On the convergence of a linesearch based proximal-gradient method for nonconvex optimization, *Inverse Probl.*, **33** (2017), 055005. https://doi.org/10.1088/1361-6420/aa5bfd
10. S. Bonettini, M. Prato, S. Rebegoldi, Convergence of inexact forward-backward algorithms using the forward-backward envelope, *SIAM J. Optim.*, **30** (2020), 3069–3097. https://doi.org/10.1137/19M1254155
11. A. Chambolle, T. Pock, A first-order primal–dual algorithm for convex problems with applications to imaging, *J. Math. Imaging Vision*, **40** (2011), 120–145. https://doi.org/10.1007/s10851-010-0251-1

12. T. Pock, A. Chambolle, Diagonal preconditioning for first order primal–dual algorithms in convex optimization, In: *2011 International Conference on Computer Vision*, 2011, 1762–1769.

13. Y. Liu, Y. Xu, W. Yin, Acceleration of primal–dual methods by preconditioning and simple subproblem procedures, *J. Sci. Comput.*, **86** (2021), 1–34. https://doi.org/10.1007/s10915-020-01371-1

14. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2 Eds., New York: Springer, 2009, https://doi.org/10.1007/978-0-387-84858-7

15. 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. https://doi.org/10.1088/1361-6420/ac0966

16. L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.*, **60** (2018), 223–311. https://doi.org/10.1137/16M1080173

17. G. Lan, *First-Order and Stochastic Optimization Methods for Machine Learning*, Berlin: Springer, 2020. https://doi.org/10.1007/978-3-030-39568-1

18. A. Chambolle, M. J. Ehrhardt, P. Richtárik, C. B. Schönlieb, Stochastic primal–dual hybrid gradient algorithm with arbitrary sampling and imaging applications, *SIAM J. Optim.*, **28** (2018), 2783–2808. https://doi.org/10.1137/17M1134834

19. J. Bai, W. W. Hager, H. Zhang, An inexact accelerated stochastic ADMM for separable convex optimization, *Comput. Optim. Appl.*, **81** (2022), 479–518. https://doi.org/10.1007/s10589-021-00338-8

20. J. Bai, F. Bian, X. Chang, L. Du, Accelerated stochastic Peaceman-Rachford method for empirical risk minimization, *J. Oper. Res. Soc. China*, **11** (2023), 783–807. https://doi.org/10.1007/s40305-023-00470-8

21. J. Bai, Y. Chen, X. Yu, H. Zhang, Generalized asymmetric forward-backward-adjoint algorithms for convex-concave saddle-point problem, *J. Sci. Comput.*, **102** (2025), 80. https://doi.org/10.1007/s10915-025-02802-7

22. M. Schmidt, N. Le Roux, F. Bach, Minimizing finite sums with the stochastic average gradient, *Math. Program.*, **162** (2017), 83–112. https://doi.org/10.1007/s10107-016-1030-6

23. A. Defazio, F. Bach, S. Lacoste-Julien, SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives, *Adv. Neural Inform. Proc. Syst.*, **2** (2014), 1646–1654.

24. R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*, **1** (2013), 315–323.

25. L. M. Nguyen, J. Liu, K. Scheinberg, M. Taká, SARAH: A novel method for machine learning problems using stochastic recursive gradient, In: *Proceedings of the 34th International Conference on Machine Learning*, **70** (2017), 2613–2621.

26. Z. Li, J. Li, A simple proximal stochastic gradient method for nonsmooth nonconvex optimization, *Adv. Neural Inform. Proc. Syst.*, **31** (2018), 5569–5579.

27. 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.

28. G. Fort, E. Moulines, Stochastic variable metric proximal gradient with variance reduction for non-convex composite optimization, *Stat. Comput.*, **33** (2023), 65. https://doi.org/10.1007/s11222-023-10230-6

29. C. Fang, C. J. Li, Z. Lin, T. Zhang, SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator, In: *Advances in Neural Information Processing Systems*, **31** (2018), 687–697.

30. A. Milzarek, X. Xiao, S. Cen, Z. Wen, M. Ulbrich, A stochastic semismooth Newton method for nonsmooth nonconvex optimization, *SIAM J. Optim.*, **29** (2019), 2916–2948. https://doi.org/10.1137/18M1181249

31. X. Wang, X. Chen, Complexity of finite-sum optimization with nonsmooth composite functions and non-lipschitz regularization, *SIAM J. Optim.*, **34** (2024), 2472–2502. https://doi.org/10.1137/23M1546701

32. Y. Xu, R. Jin, T. Yang, Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems, *Adv. Neural Inform. Proc. Syst.*, **32** (2019), 2630–2640.

33. P. Latafat, A. Themelis, P. Patrinos, Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems, *Math. Program.*, **193** (2022), 195–224. https://doi.org/10.1007/s10107-020-01599-7

34. J. Qiu, X. Li, A. Milzarek, A new random reshuffling method for nonsmooth nonconvex finite-sum optimization, *J. Mach. Learn. Res.*, **26** (2025), 1–46.

35. Y. Zeng, J. Bai, S. Wang, Z. Wang, X. Shen, A hybrid stochastic alternating direction method of multipliers for nonconvex and nonsmooth composite optimization, *Eur. J. Oper. Res.*, **329** (2026), 63–78. https://doi.org/10.1016/j.ejor.2025.10.024

36. D. Driggs, J. Tang, J. Liang, M. Davies, C. B. Schönlieb, A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization, *SIAM J. Imaging Sci.*, **14** (2021), 1932–1970. https://doi.org/10.1137/20M1387213

37. A. Beck, *First-Order Methods in Optimization*, Philadelphia: SIAM, 2017. https://doi.org/10.1137/1.9781611974997

38. R. T. Rockafellar, *Convex Analysis*, Princeton: Princeton University Press, 1970. https://doi.org/10.1515/9781400873173

39. J. F. Bonnans, A. Shapiro, *Perturbation Analysis of Optimization Problems*, New York: Springer-Verlag, 2000. https://doi.org/10.1007/978-1-4612-1394-9

40. H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality, *Math. Program.*, **35** (2010), 5–16. https://doi.org/10.1287/moor.1100.0449

41. J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.*, **146** (2014), 459–494. https://doi.org/10.1007/s10107-013-0701-9

42. H. Attouch, J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Math. Program.*, **116** (2009), 5–16. https://doi.org/10.1007/s10107-007-0133-5

43. R. I. Boţ, E. R. Csetnek, D. K. Nguyen, A proximal minimization algorithm for structured nonconvex and nonsmooth problems, *SIAM J. Optim.*, **29** (2019), 1300–1328. https://doi.org/10.1137/18M1190689

44. R. T. Rockafellar, R. J. B. Wets, *Variational Analysis*, Berlin: Springer-Verlag, 1998. https://doi.org/10.1007/978-3-642-02431-3

45. X. Li, A. Milzarek, J. Qiu, Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality, *SIAM J. Optim.*, **33** (2023), 1092–1120. https://doi.org/10.1137/21M1468048

46. D. Davis, D. Drusvyatskiy, S. Kakade, J. D. Lee, Stochastic subgradient method converges on tame functions, *Found. Comput. Math.*, **20** (2020), 119–154. https://doi.org/10.1007/s10208-018-09409-5

47. D. Davis, The asynchronous PALM algorithm for nonsmooth nonconvex problems, preprint paper, 2016. https://doi.org/10.48550/arXiv.1604.00526

48. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9** (2008), 432–441. https://doi.org/10.1093/biostatistics/kxm045

49. I. Guyon, S. Gunn, A. Ben-Hur, G. Dror, Result analysis of the NIPS 2003 feature selection challenge, *Adv. Neural Inform. Proc. Syst.*, 2004, 545–552.

50. D. P. Bertsekas, *Convex Optimization Algorithms*, Belmont: Athena Scientific, 2015.

A. Appendix

A.1. Proof of Lemma 2

Proof. We first present a recursive relation for $\mathcal{L}$. From (2.5a) and (2.6), it follows that

$$f(x^{k+1}) \leq f(x^k) - \langle y^k, A(x^{k+1} - x^k)\rangle - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2.$$

Taking $k = k - 1$ in (2.4) and using the convexity of h^*, we have

$$-h^*(y^{k+1}) \leq -h^*(y^k) + \langle y^k - y^{k+1}, -M(y^k - y^{k-1}) + A(2x^k - x^{k-1})\rangle.$$

Combining these two inequalities, adding $\langle y^{k+1}, Ax^{k+1}\rangle$ on both sides and recalling that $\mathcal{L}(x, y) = f(x) + \langle y, Ax\rangle - h^*(y)$, we obtain

$$\begin{aligned}\mathcal{L}(x^{k+1}, y^{k+1}) \leq \mathcal{L}(x^k, y^k) &- \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 \\ &+ \langle y^{k+1} - y^k, A(x^{k+1} - x^k + x^{k-1} - x^k) + M(y^k - y^{k-1})\rangle.\end{aligned} \tag{A.1}$$

Applying (2.5a) again, one has

$$\begin{aligned}&\langle y^{k+1} - y^k, A(x^{k+1} - x^k + x^{k-1} - x^k) + M(y^k - y^{k-1})\rangle \\ &= \langle y^{k+1} - y^k, (\alpha AA^T - M)(y^{k-1} - y^k)\rangle + \langle y^{k+1} - y^k, \alpha A(\nabla f(x^{k-1}) - \nabla f(x^k))\rangle \\ &= \langle y^{k+1} - y^k, \alpha A(\nabla f(x^{k-1}) - \nabla f(x^k))\rangle.\end{aligned}$$

Substituting this relation into (A.1), we have

$$\mathcal{L}(x^{k+1}, y^{k+1}) \leq \mathcal{L}(x^k, y^k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 + \langle y^{k+1} - y^k, \alpha A(\nabla f(x^{k-1}) - \nabla f(x^k))\rangle. \tag{A.2}$$

We now prove (2.11). From (A.2), we have

$$\begin{aligned}\mathcal{L}(x^{k+1}, y^{k+1}) &\leq \mathcal{L}(x^k, y^k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 + \langle y^{k+1} - y^k, \alpha A(\nabla f(x^{k-1}) - \nabla f(x^k))\rangle \\ &\leq \mathcal{L}(x^k, y^k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{\alpha\delta}{2}\|A^T(y^{k+1} - y^k)\|^2 + \frac{\alpha L^2}{2\delta}\|x^k - x^{k-1}\|^2,\end{aligned} \tag{A.3}$$

where the second inequality is deduced from the Lipschitz continuity of ∇f and the fact that $\langle x, y\rangle \leq \frac{\delta}{2}\|x\|^2 + \frac{1}{2\delta}\|y\|^2$. From (2.7), it follows that

$$\|A^T(y^{k+1} - y^k)\|^2 \leq 2\left(\frac{1}{\alpha} + L\right)^2\|x^{k+1} - x^k\|^2 + \frac{2}{\alpha^2}\|x^{k+2} - x^{k+1}\|^2.$$

Substituting this inequality into (A.3) and recalling the definitions of a, b, and c, we conclude that

$$\mathcal{L}(x^{k+1}, y^{k+1}) \leq \mathcal{L}(x^k, y^k) - (a + b + c)\|x^{k+1} - x^k\|^2 + (b - c)\|x^k - x^{k-1}\|^2 + a\|x^{k+2} - x^{k+1}\|^2.$$

Rewriting this inequality gives

$$\begin{aligned}&\mathcal{L}(x^{k+1}, y^{k+1}) - a\|x^{k+1} - x^{k+2}\|^2 + b\|x^{k+1} - x^k\|^2 + c(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2) \\ &\leq \mathcal{L}(x^k, y^k) - a\|x^k - x^{k+1}\|^2 + b\|x^k - x^{k-1}\|^2.\end{aligned}$$

The proof of (2.11) is completed by recalling the definition of $\mathscr{L}$.

Finally, we consider the bound of d^k. For the first component of d^k, we have

$$\begin{aligned}\|\nabla_x\mathscr{L}(z^k)\| &= \|\nabla f(x^k) + A^T y^k - 2a(x^k - x^{k+1}) + 2b(x^k - x^{k-1})\| \\ &\leq (L + 2b)\|x^k - x^{k-1}\| + 2a\|x^k - x^{k+1}\| + \|A^T(y^k - y^{k-1})\| + \|\nabla f(x^{k-1}) + A^T y^{k-1}\|,\end{aligned}$$

which, together with (2.5a) and (2.7), gives

$$\|\nabla_x\mathscr{L}(z^k)\| \leq 2\left(L + b + \frac{1}{\alpha}\right)\|x^k - x^{k-1}\| + \left(2a + \frac{1}{\alpha}\right)\|x^{k+1} - x^k\|. \tag{A.4}$$

For the second component of d^k, by (2.4), we obtain

$$\|Ax^k - g^k\| = \|M(y^k - y^{k-1}) - A(x^k - x^{k-1})\| \leq \alpha\|A\|\|A^T(y^k - y^{k-1})\| + \|A\|\|x^k - x^{k-1}\|,$$

which, together with (2.7), yields

$$\|Ax^k - g^k\| \leq (2 + \alpha L)\|A\|\|x^k - x^{k-1}\| + \|A\|\|x^{k+1} - x^k\|. \tag{A.5}$$

For $\nabla_u\mathscr{L}(z^k)$ and $\nabla_v\mathscr{L}(z^k)$, one has

$$\|\nabla_u\mathscr{L}(z^k)\| = 2a\|x^{k+1} - x^k\|, \quad \|\nabla_v\mathscr{L}(z^k)\| = 2b\|x^k - x^{k-1}\|. \tag{A.6}$$

Combining (A.4), (A.5), and (A.6) together, we have

$$\|d^k\| \leq \left(2L + 4b + \frac{2}{\alpha} + (2 + \alpha L)\|A\|\right)\|x^k - x^{k-1}\| + \left(4a + \frac{1}{\alpha} + \|A\|\right)\|x^{k+1} - x^k\|.$$

The proof is completed. □

A.2. Proof of Theorem 5

Proof. Consider $\theta = 0$, let $K_1 := \max\{k \in \mathbb{N} : x^{k+1} \neq x^k\}$. We now show that K_1 is a finite number. On the contrary, we assume that K_1 is sufficiently large that (2.18) holds for all $k \geq K_1$. Note that $\varphi(s) = \sigma s$ when $\theta = 0$, then (2.18) and (2.20) read

$$\gamma(\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|) \geq \operatorname{dist}(0, \partial\mathscr{L}(z^k)) \geq \frac{1}{\sigma},\ k \geq K_1,$$

which, together with Lemma 2 and $a^2 + b^2 \geq (a+b)^2/2$, yields

$$\mathscr{L}(z^{k+1}) \leq \mathscr{L}(z^k) - c(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2) \leq \mathscr{L}(z^k) - \frac{c}{2\gamma^2\sigma^2}.$$

Let $k \to \infty$. In the proof of Proposition 3, it has been shown that $\lim_{k\to\infty} \mathscr{L}(z^k) = \bar{\mathscr{L}} = \mathcal{L}(\bar{x}, \bar{y})$, consequently,

$$\mathcal{L}(\bar{x}, \bar{y}) \leq \mathcal{L}(\bar{x}, \bar{y}) - \frac{c}{2\gamma^2\sigma^2},$$

which is a contradiction. Therefore, K_1 is a finite number and $\{x^k\}$ converges in finite steps. From (2.7) and (2.8), we obtain

$$\begin{aligned}\|y^{k+1} - y^k\| &\leq \frac{1/\alpha + L}{\hat{\lambda}}\|x^{k+1} - x^k\| + \frac{1}{\alpha\hat{\lambda}}\|x^{k+2} - x^{k+1}\| \\ &\leq \frac{1/\alpha + L}{\hat{\lambda}}\left(\|x^{k+1} - x^k\| + \|x^{k+2} - x^{k+1}\|\right).\end{aligned}$$

Hence, $\{y^k\}$ also converges in finite steps and Item (i) holds.

Let $\Delta_k := \sum_{q=k}^{\infty} \|x^{q+1} - x^q\| + \|x^q - x^{q-1}\|$. The results in Theorem 4 state that $\Delta_k < +\infty$ for any $k \geq 1$ and the sequence $\{(x^k, y^k)\}$ converges to $(\bar{x}, \bar{y})$ which is a critical point of $\mathcal{L}$. The triangle inequality implies that $\|x^k - \bar{x}\| \leq \Delta_k$ and

$$\|y^k - \bar{y}\| \leq \sum_{q=k}^{\infty} \|y^{q+1} - y^q\| \leq \frac{1/\alpha + L}{\hat{\lambda}}\Delta_k.$$

Therefore, it is sufficient to establish the estimations in (ii) and (iii) for Δ_k. If $\Delta_k = 0$ for some k, it follows that $\|x^{q+1} - x^q\| = 0$ for $q \geq k$ and $\{(x^k, y^k)\}$ converges in finite steps. Thus, without loss of generality, we assume $\Delta_k > 0$ for any $k \geq 1$.

For $\theta \in (0, 1)$, noting that $\varphi(s) = \sigma s^{1-\theta}$, letting $n \to \infty$ in (2.24) and using (2.18), we have

$$\begin{aligned}\Delta_{k+1} \leq \Delta_k &\leq \frac{4\gamma\sigma}{c}(\mathscr{L}(z^k) - \mathcal{L}(\bar{x}, \bar{y}))^{1-\theta} + \frac{1}{2}\|x^k - x^{k-1}\| \\ &\leq \frac{4\gamma\sigma^{\frac{1}{\theta}}}{c}((1-\theta)\operatorname{dist}(0, \partial\mathscr{L}(z^k)))^{\frac{1-\theta}{\theta}} + \frac{1}{2}\|x^k - x^{k-1}\|\end{aligned}$$

for any $k \geq K_0$. The inequality above, together with the definition of Δ_k and (2.20), yields

$$\begin{aligned}\Delta_{k+1} &\leq \frac{4\gamma\sigma^{\frac{1}{\theta}}}{c}[\gamma(1-\theta)(\Delta_k - \Delta_{k+1})]^{\frac{1-\theta}{\theta}} + \frac{1}{2}(\Delta_k - \Delta_{k+1}) \\ &= \gamma'(\Delta_k - \Delta_{k+1})^{\frac{1-\theta}{\theta}} + \frac{1}{2}(\Delta_k - \Delta_{k+1}),\end{aligned} \tag{A.7}$$

where $\gamma' := \frac{4}{c}(\gamma\sigma)^{\frac{1}{\theta}}(1-\theta)^{\frac{1-\theta}{\theta}}$.

Consider $\theta \in (0, \frac{1}{2}]$. Noting that $0 < \Delta_k - \Delta_{k+1} < 1$ for $k \geq K$ with K large enough ($K \geq K_0$). From (A.7) and $\frac{1-\theta}{\theta} \geq 1$, it follows that

$$\Delta_{k+1} \leq (\gamma' + \frac{1}{2})(\Delta_k - \Delta_{k+1}).$$

By rearranging the inequality above and setting $\tau := (\gamma' + \frac{1}{2})/(\gamma' + \frac{3}{2}) < 1$, one has $\Delta_{k+1} \leq \tau\Delta_k$. Therefore, for any $k \geq K$, it holds that $\Delta_k \leq \nu\tau^{k-K}$, where $\nu := \Delta_K$ is a finite number. Item (ii) is derived.

Consider $\theta \in (\frac{1}{2}, 1)$. Let $\bar{K} \geq K_0$ be large enough such that $0 < \Delta_k - \Delta_{k+1} < 1$ for all $k \geq \bar{K}$. Noting that $0 < \frac{1-\theta}{\theta} < 1$, we see from (A.7) that

$$\Delta_{k+1} \leq (\gamma' + \frac{1}{2})(\Delta_k - \Delta_{k+1})^{\frac{1-\theta}{\theta}}$$

for all $k \geq \bar{K}$. Following the same line of the proof of [42, p.14], there exists a constant $\mu_1 > 0$ such that for all $k \geq \bar{K}$,

$$(\Delta_{k+1})^{\nu_1} - (\Delta_k)^{\nu_1} \geq \mu_1,$$

where $\nu_1 := (1-2\theta)/(1-\theta) < 0$. Summing up from $k = \bar{K}$ to n for any $n \geq \bar{K}$ yields

$$(\Delta_n)^{\nu_1} \geq (n - \bar{K})\mu_1 + (\Delta_{\bar{K}})^{\nu_1},$$

which, together with $\nu_1 < 0$, implies that for any $n \geq \bar{K}$,

$$\Delta_n \leq [(n-\bar{K})\mu_1 + (\Delta_{\bar{K}})^{\nu_1}]^{\frac{1}{\nu_1}} \leq \mu n^{\frac{1}{\nu_1}},$$

where μ is a positive constant. Item (iii) is obtained. □

A.3. Proof of Lemma 7

Proof. Using (3.1a) twice yields

$$\begin{aligned} \|A^T(y^{k+1} - y^k)\| &= \left\| \left(\frac{x^{k+1} - x^{k+2}}{\alpha} - \widetilde{\nabla} f_{k+1} \right) - \left(\frac{x^k - x^{k+1}}{\alpha} - \widetilde{\nabla} f_k \right) \right\| \\ &\leq \frac{1}{\alpha}\|x^{k+2} - x^{k+1}\| + \frac{1}{\alpha}\|x^{k+1} - x^k\| + \|\widetilde{\nabla} f_{k+1} - \widetilde{\nabla} f_k\|. \end{aligned} \tag{A.8}$$

Since f is L-smooth, we have

$$\begin{aligned} \|\widetilde{\nabla} f_{k+1} - \widetilde{\nabla} f_k\| &\leq \|\widetilde{\nabla} f_{k+1} - \nabla f(x^{k+1})\| + \|\nabla f(x^{k+1}) - \nabla f(x^k)\| + \|\widetilde{\nabla} f_k - \nabla f(x^k)\| \\ &\leq \|\widetilde{\nabla} f_{k+1} - \nabla f(x^{k+1})\| + \|\widetilde{\nabla} f_k - \nabla f(x^k)\| + L\|x^{k+1} - x^k\|. \end{aligned} \tag{A.9}$$

Substituting (A.9) into (A.8), one has

$$\begin{aligned} \|A^T(y^{k+1} - y^k)\|^2 \leq \frac{4}{\alpha^2}\|x^{k+2} - x^{k+1}\|^2 + 4\left(\frac{1}{\alpha} + L\right)^2 \|x^{k+1} - x^k\|^2 \\ + 4\|\widetilde{\nabla} f_{k+1} - \nabla f(x^{k+1})\|^2 + 4\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2. \end{aligned}$$

Finally, taking the conditional expectation on both sides and using (3.5), we derive the claim. □

A.4. Proof of Lemma 8

Proof. Since the function f is L-smooth, we have

$$\begin{aligned}
f(x^{k+1}) &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) + \langle \widetilde{\nabla} f_k, x^{k+1} - x^k \rangle + \langle \nabla f(x^k) - \widetilde{\nabla} f_k, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\le f(x^k) + \langle \frac{x^k - x^{k+1}}{\alpha} - A^T y^k, x^{k+1} - x^k \rangle + \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 + \frac{\delta_1 + L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 - \langle y^k, A(x^{k+1} - x^k) \rangle,
\end{aligned}$$

where the second inequality is deduced from (3.1a) and $\langle x, y\rangle \le \frac{\delta_1}{2}\|x\|^2 + \frac{1}{2\delta_1}\|y\|^2$ for any $\delta_1 > 0$. Together with the convexity of h^* and $g^k \in \partial h^*(y^k)$, it also indicates

$$\begin{aligned}
&f(x^{k+1}) - h^*(y^{k+1}) + \langle y^{k+1}, Ax^{k+1} \rangle \\
&\le f(x^k) - h^*(y^k) + \langle y^k, Ax^k \rangle - \langle y^k, Ax^k \rangle + \langle y^{k+1}, Ax^{k+1} \rangle - \langle y^k, A(x^{k+1} - x^k) \rangle \\
&\quad + \langle y^k - y^{k+1}, g^k \rangle - \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 \\
&= f(x^k) - h^*(y^k) + \langle y^k, Ax^k \rangle + \langle y^{k+1} - y^k, Ax^{k+1} - g^k \rangle \\
&\quad - \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2.
\end{aligned}$$

Recalling the definition of $\mathcal{L}_s$ and substituting (2.4) (let $k + 1 = k$) into the inequality above, one has

$$\begin{aligned}
\mathcal{L}_s(x^{k+1}, y^{k+1}) \le \mathcal{L}_s(x^k, y^k) &+ \langle y^{k+1} - y^k, A(x^{k+1} - x^k + x^{k-1} - x^k) + M(y^k - y^{k-1}) \rangle \\
&- \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2.
\end{aligned}$$

Using (3.1a) and the fact that $\langle x, y\rangle \le \frac{\delta_2}{2}\|x\|^2 + \frac{1}{2\delta_2}\|y\|^2$ for the second term of the right-hand side, and letting $M = \alpha AA^T$, we have

$$\begin{aligned}
\mathcal{L}_s(x^{k+1}, y^{k+1}) \le \mathcal{L}_s(x^k, y^k) &- \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{\delta_2 \alpha}{2}\|A^T(y^{k+1} - y^k)\|^2 \\
&+ \frac{1}{2\delta_1}\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 + \frac{\alpha}{2\delta_2}\|\widetilde{\nabla} f_{k-1} - \widetilde{\nabla} f_k\|^2,
\end{aligned}$$

which, together with (A.9) (take $k = k - 1$), yields

$$\begin{aligned}
&\mathcal{L}_s(x^{k+1}, y^{k+1}) \\
&\le \mathcal{L}_s(x^k, y^k) - \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{\delta_2 \alpha}{2}\|A^T(y^{k+1} - y^k)\|^2 + \frac{3\alpha L^2}{2\delta_2}\|x^k - x^{k-1}\|^2 \\
&\quad + \left(\frac{1}{2\delta_1} + \frac{3\alpha}{2\delta_2}\right)\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 + \frac{3\alpha}{2\delta_2}\|\nabla f(x^{k-1}) - \widetilde{\nabla} f_{k-1}\|^2.
\end{aligned} \tag{A.10}$$

Taking the conditional expectation on both sides of (A.10), and applying Lemma 7 as well as (3.5), we have

$$\begin{aligned}
&\mathbb{E}_{k-1}[\mathcal{L}_s(x^{k+1}, y^{k+1})] \\
&\leq \mathbb{E}_{k-1}[\mathcal{L}_s(x^k, y^k)] + \left(2\delta_2\alpha\kappa + \frac{\kappa}{2\delta_1} + \frac{3\alpha(L^2+2\kappa)}{2\delta_2}\right)\mathbb{E}_{k-1}[\|x^k - x^{k-1}\|^2] \\
&\quad - \left(\frac{1}{\alpha} - \frac{\delta_1 + L}{2} - \frac{\kappa}{2\delta_1} - \frac{3\alpha\kappa}{2\delta_2} - 2\delta_2\alpha\left(\left(\frac{1}{\alpha} + L\right)^2 + 2\kappa\right)\right)\mathbb{E}_{k-1}[\|x^{k+1} - x^k\|^2] \\
&\quad + 2\delta_2\alpha\left(\frac{1}{\alpha^2} + \kappa\right)\mathbb{E}_{k-1}[\|x^{k+2} - x^{k+1}\|^2] + \left(\frac{2\delta_2\alpha}{\rho} + \frac{1}{2\delta_1\rho} + \frac{3\alpha}{2\delta_2\rho}\right)\mathbb{E}_{k-1}[\Lambda_1^k - \Lambda_1^{k+1}] \\
&\quad + \frac{3\alpha\kappa}{2\delta_2}\|x^{k-1} - x^{k-2}\|^2 + \frac{2\delta_2\alpha}{\rho}\mathbb{E}_{k-1}[\Lambda_1^{k+1} - \Lambda_1^{k+2}] + \frac{3\alpha}{2\delta_2\rho}(\Lambda_1^{k-1} - \mathbb{E}_{k-1}[\Lambda_1^k]).
\end{aligned}$$

Therefore, taking the expectation on both sides implies that

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_s(x^{k+1}, y^{k+1})] \leq \mathbb{E}[\mathcal{L}_s(x^k, y^k)] &- e_4\mathbb{E}[\|x^{k+1} - x^k\|^2] + e_5\mathbb{E}[\|x^{k+2} - x^{k+1}\|^2] + e_6\mathbb{E}[\|x^k - x^{k-1}\|^2] \\
&+ e_7\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2] + e_1\mathbb{E}[\Lambda_1^{k+1} - \Lambda_1^{k+2}] + e_2\mathbb{E}[\Lambda_1^k - \Lambda_1^{k+1}] + e_3\mathbb{E}[\Lambda_1^{k-1} - \Lambda_1^k],
\end{aligned}$$

where

$$\begin{aligned}
&e_1 = \tfrac{2\delta_2\alpha}{\rho},\ e_2 = \tfrac{2\delta_2\alpha}{\rho} + \tfrac{1}{2\delta_1\rho} + \tfrac{3\alpha}{2\delta_2\rho},\ e_3 = \tfrac{3\alpha}{2\delta_2\rho}, \\
&e_4 = \tfrac{1}{\alpha} - \tfrac{\delta_1+L}{2} - \tfrac{\kappa}{2\delta_1} - \tfrac{3\alpha\kappa}{2\delta_2} - 2\delta_2\alpha((\tfrac{1}{\alpha} + L)^2 + 2\kappa), \\
&e_5 = 2\delta_2\alpha(\tfrac{1}{\alpha^2} + \kappa),\ e_6 = 2\delta_2\alpha\kappa + \tfrac{\kappa}{2\delta_1} + \tfrac{3\alpha(L^2+2\kappa)}{2\delta_2},\ e_7 = \tfrac{3\alpha\kappa}{2\delta_2}.
\end{aligned}$$

Recalling the definitions of a, b, c, and e_0, we have $e_0 = \frac{1}{3}(e_4 - e_5 - e_6 - e_7)$, $a = e_0 + e_5$, $b = e_0 + e_6 + e_7$, and $c = e_7$, and thus

$$\mathbb{E}[\mathscr{L}_{s,k+1}^{\Lambda} + e_0(\|x^{k+2} - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2)] \leq \mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}].$$

This proof is completed. □

A.5. Proof of Lemma 9

Proof. It is sufficient to bound the five components of d^k. First, from (3.1a), we have

$$\begin{aligned}
\|d_1^k\|^2 &= \|\nabla f(x^k) + A^T y^k - 2a(x^k - x^{k+1}) + 2b(x^k - x^{k-1})\|^2 \\
&\leq \left(\|\nabla f(x^k) - \widetilde{\nabla} f_k\| + \|A^T y^k + \widetilde{\nabla} f_k\| + 2a\|x^{k+1} - x^k\| + 2b\|x^k - x^{k-1}\|\right)^2 \\
&= \left(\|\nabla f(x^k) - \widetilde{\nabla} f_k\| + \left(\frac{1}{\alpha} + 2a\right)\|x^{k+1} - x^k\| + 2b\|x^k - x^{k-1}\|\right)^2 \\
&= 3\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 + 3\left(\frac{1}{\alpha} + 2a\right)^2\|x^{k+1} - x^k\|^2 + 12b^2\|x^k - x^{k-1}\|^2.
\end{aligned}$$

Taking the conditional expectation on both sides and using (3.2) yields

$$\mathbb{E}_k[\|d_1^k\|^2] \leq 3\left(\frac{1}{\alpha^2} + \frac{4a}{\alpha} + 4a^2 + \sigma_1\right)\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + 3(4b^2 + \sigma_1)\|x^k - x^{k-1}\|^2 + 3\Lambda_1^k.$$

For the other four components of d^k, it follows that

$$\begin{aligned}
&\mathbb{E}_k[\|d_2^k\|^2] \le 2\|M\|^2 \cdot \|y^k - y^{k-1}\|^2 + 2\|A\|^2 \cdot \|x^k - x^{k-1}\|^2,\\
&\mathbb{E}_k[\|d_3^k\|^2] = 4a^2\mathbb{E}_k[\|x^{k+1} - x^k\|^2],\\
&\mathbb{E}_k[\|d_4^k\|^2] \le 8b^2\|x^k - x^{k-1}\|^2 + 8c^2\|x^{k-1} - x^{k-2}\|^2,\\
&\mathbb{E}_k[\|d_5^k\|^2] = 4c^2\|x^{k-1} - x^{k-2}\|^2.
\end{aligned}$$

Combining these results, we derive the conclusion. □

A.6. Proof of Proposition 11

Proof. Because $\{(x^k, y^k)\}$ is bounded almost surely, from Remark 5 there is an event $\mathcal{A}$ with a measure 1 such that the sequence $\{(x^k(\omega), y^k(\omega))\}$ is bounded for any fixed $\omega \in \mathcal{A}$. Hence, the set C_ω is nonempty. For any $(\bar{x}(\omega), \bar{y}(\omega)) \in C_\omega$, there is a subsequence $\{(x^{k_q}(\omega), y^{k_q}(\omega))\}$ of $\{(x^k(\omega), y^k(\omega))\}$ such that

$$x^{k_q}(\omega) \to \bar{x}(\omega) \text{ and } y^{k_q}(\omega) \to \bar{y}(\omega). \tag{A.11}$$

From (3.14) and (3.20), we have

$$\lim_{k\to\infty} \|x^{k+1}(\omega) - x^k(\omega)\| = 0 \text{ and } \lim_{k\to\infty} \|y^{k+1}(\omega) - y^k(\omega)\| = 0. \tag{A.12}$$

Thus, we obtain that C_ω is compact and $\mathrm{dist}((x^k(\omega), y^k(\omega)), C_\omega) \to 0$ by following the same line of the proof of Proposition 3 (ii). Item (i) is derived.

We next prove that for any $(\bar{x}(\omega), \bar{y}(\omega)) \in C_\omega$, $\bar{z}(\omega) := (\bar{x}(\omega), \bar{y}(\omega), \bar{x}(\omega), \bar{x}(\omega), \bar{x}(\omega)) \in \mathrm{crit}\mathscr{L}_s$, i.e., $0 \in \partial\mathscr{L}_s(\bar{z}(\omega))$, by using the outer semicontinuity of $\partial\mathscr{L}_s$. Let

$$z^{k_q}(\omega) := (x^{k_q}(\omega), y^{k_q}(\omega), x^{k_q+1}(\omega), x^{k_q-1}(\omega), x^{k_q-2}(\omega)).$$

It immediately follows from (A.11) that $z^{k_q}(\omega) \to \bar{z}(\omega)$. Let $d^{k_q}(\omega)$ be defined in a similar way to (3.9) with respect to ω. We then have $d^{k_q}(\omega) \in \partial\mathscr{L}_s(z^{k_q}(\omega))$. Therefore, because of the outer semicontinuity of $\partial\mathscr{L}_s$, in order to obtain $0 \in \partial\mathscr{L}_s(\bar{z}(\omega))$, it is sufficient to show that $d^{k_q}(\omega) \to 0$. By rearranging (3.4), we obtain

$$\mathbb{E}[\Lambda_1^{k_q}] \le \frac{1}{\rho}\mathbb{E}[\Lambda_1^{k_q} - \Lambda_1^{k_q+1}] + \frac{\sigma_\Lambda}{\rho}(\mathbb{E}[\|x^{k_q+1} - x^{k_q}\|^2] + \mathbb{E}[\|x^{k_q} - x^{k_q-1}\|^2]),$$

which, together with (3.10) (take $k = k_q$), yields

$$\begin{aligned}
\mathbb{E}[\|d^{k_q}\|^2] \le (r + \frac{r\sigma_\Lambda}{\rho})\mathbb{E}[\|x^{k_q+1} - x^{k_q}\|^2 + \|x^{k_q} - x^{k_q-1}\|^2 + \|x^{k_q-1} - x^{k_q-2}\|^2]\\
+ r\mathbb{E}[\|y^{k_q} - y^{k_q-1}\|^2] + \frac{r}{\rho}\mathbb{E}[\Lambda_1^{k_q} - \Lambda_1^{k_q+1}].
\end{aligned}$$

Summing up from $k_q = 2$ to ∞ and using (3.12), (3.18), and (3.15), one has

$$\sum_{k_q=2}^{\infty} \mathbb{E}[\|d^{k_q}\|^2] < \infty.$$

Hence $d^{k_q} \to 0$ almost surely, which also implies that $d^{k_q}(\omega) \to 0$. Thus, we finish the proof of $\bar{z}(\omega) \in \mathrm{crit}\mathscr{L}_s$. Furthermore, we derive Item (ii) by Lemma 6.

To prove Item (iii), let us first show that $\sum_{k=1}^{\infty} W_k < \infty$ almost surely, where

$$W_k := \frac{\delta_1 + L}{2}\|x^{k+1} - x^k\|^2 + \frac{\delta_2\alpha}{2}\|A^T(y^{k+1} - y^k)\|^2 + \frac{3\alpha L^2}{2\delta_2}\|x^k - x^{k-1}\|^2 + \left(\frac{1}{2\delta_1} + \frac{3\alpha}{2\delta_2}\right)\|\nabla f(x^k) - \widetilde{\nabla} f_k\|^2 + \frac{3\alpha}{2\delta_2}\|\nabla f(x^{k-1}) - \widetilde{\nabla} f_{k-1}\|^2.$$

It follows from (3.5) that

$$\mathbb{E}[\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2] \le \frac{1}{\rho}(\mathbb{E}[\Lambda_1^k] - \mathbb{E}[\Lambda_1^{k+1}]) + \kappa(\mathbb{E}[\|x^{k+1} - x^k\|^2] + \mathbb{E}[\|x^k - x^{k-1}\|^2]),$$

which, together with the facts that $\mathbb{E}[\Lambda_1^k] \to 0$ and $\sum_{k=0}^{\infty}\mathbb{E}[\|x^{k+1} - x^k\|^2] < \infty$ from (3.15) and (3.12), indicates that

$$\sum_{k=1}^{\infty}\mathbb{E}[\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2] < \infty,$$

and hence $\sum_{k=1}^{\infty}\|\widetilde{\nabla} f_k - \nabla f(x^k)\|^2 < \infty$ almost surely. Therefore, using Proposition 10, we obtain that $\sum_{k=1}^{\infty} W_k < \infty$ almost surely.

In a completely analogous way to (A.10), we can prove that for any fixed $\omega \in \mathcal{A}$,

$$\begin{aligned}
&\mathcal{L}_s(x^{k+1}(\omega), y^{k+1}(\omega)) \\
&\le \mathcal{L}_s(x^k(\omega), y^k(\omega)) + \frac{\delta_1 + L}{2}\|x^{k+1}(\omega) - x^k(\omega)\|^2 + \frac{\delta_2\alpha}{2}\|A^T(y^{k+1}(\omega) - y^k(\omega))\|^2 \\
&\quad + \frac{3\alpha L^2}{2\delta_2}\|x^k(\omega) - x^{k-1}(\omega)\|^2 + \left(\frac{1}{2\delta_1} + \frac{3\alpha}{2\delta_2}\right)\|\nabla f(x^k(\omega)) - \widetilde{\nabla} f_k(\omega)\|^2 \\
&\quad + \frac{3\alpha}{2\delta_2}\|\nabla f(x^{k-1}(\omega)) - \widetilde{\nabla} f_{k-1}(\omega)\|^2 \\
&= \mathcal{L}_s(x^k(\omega), y^k(\omega)) + W_k(\omega).
\end{aligned}$$

Because $\sum_{k=1}^{\infty} W_k < \infty$ almost surely, we have $\sum_{k=1}^{\infty} W_k(\omega) < \infty$. Therefore, from [50, Proposition A.4.4], it follows that $\{\mathcal{L}_s(x^k(\omega), y^k(\omega))\}$ converges to a finite value. Since $\mathcal{L}_s$ is continuous over $\mathbb{R}^n \times \mathrm{dom}h^*$, from (A.11), we have

$$\lim_{q\to\infty}\mathcal{L}_s(x^{k_q}(\omega), y^{k_q}(\omega)) = \mathcal{L}_s(\bar{x}(\omega), \bar{y}(\omega)).$$

Combining these results with the definition of C_ω, we derive that $\mathcal{L}_s$ is finite and constant on C_ω. The proof is completed. □

A.7. Proof of Theorem 13

Proof. Item (i) has been presented in the proof of Theorem 12.

Let us point out that for the case that $\theta \in (0, 1/2)$, by the same reason as that in the proof of Theorem 12, the analysis in the following can reduce to the case that $\theta = 1/2$. Therefore, it is sufficient to consider the case that $\theta \in [1/2, 1)$. Let

$$\Delta_k := \sum_{q=k}^{\infty} \sqrt{\mathbb{E}[\|x^{q+1} - x^q\|^2]} + \sum_{q=k}^{\infty} \sqrt{\mathbb{E}[\|x^q - x^{q-1}\|^2]} + \sum_{q=k}^{\infty} \sqrt{\mathbb{E}[\|x^{q-1} - x^{q-2}\|^2]}.$$

Noticing that (3.32) holds for all $\theta \in [1/2, 1)$. Similarly, we can also have

$$\begin{aligned} \sum_{k=K}^{n} \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} \leq\ & \sqrt{\mathbb{E}[\|x^{n+1} - x^n\|^2]} + \sqrt{\mathbb{E}[\|x^{K-1} - x^{K-2}\|^2]} + \sum_{k=K}^{n} \frac{4\gamma \mathcal{M}_{k,k+1}}{e_0} \\ & + \frac{2\gamma_6 \sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^{K-1}]} + \frac{2\sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^K]} \end{aligned} \tag{A.13}$$

and

$$\begin{aligned} \sum_{k=K}^{n} \sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} \leq\ & 2\sqrt{\mathbb{E}[\|x^K - x^{K-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{K-1} - x^{K-2}\|^2]} + \sum_{k=K}^{n} \frac{4\gamma \mathcal{M}_{k,k+1}}{e_0} \\ & + \frac{2\gamma_6 \sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^{K-1}]} + \frac{2\sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^K]}. \end{aligned} \tag{A.14}$$

Then, combining (3.32), (A.13), and (A.14) (let $n \to \infty$), for any $k \geq K$, it follows from the definition of $\mathcal{M}_{m,n}$ and $\varphi_1(s) = \beta\sigma_0 s^{1-\theta}$ that

$$\begin{aligned} \Delta_{k+1} &\leq 4\Big(\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} \Big) \\ &\quad + \frac{6\gamma_6 \sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^{k-1}]} + \frac{6\sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^k]} + \frac{12\gamma}{e_0} \varphi_1(\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}] - \bar{\mathscr{L}}_{s,k}) \\ &= 4\Big(\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} \Big) \\ &\quad + \frac{6\gamma_6 \sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^{k-1}]} + \frac{6\sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^k]} + \beta_1 (\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}] - \bar{\mathscr{L}}_{s,k})^{1-\theta}, \end{aligned} \tag{A.15}$$

where $\beta_1 := 12\gamma\beta\sigma_0/e_0$. By the definition of $\mathscr{L}_{s,k}^{\Lambda}$ and $(a+b)^{1-\theta} \leq a^{1-\theta} + b^{1-\theta}$ for $\theta \in [1/2, 1)$, it follows that

$$(\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}] - \bar{\mathscr{L}}_{s,k})^{1-\theta} \leq (\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k})^{1-\theta} + \beta_2 (\mathbb{E}[\Lambda_1^{k+1}] + \mathbb{E}[\Lambda_1^k] + \mathbb{E}[\Lambda_1^{k-1}])^{1-\theta}, \tag{A.16}$$

where $\beta_2 = \max\{e_1, e_2, e_3\}^{1-\theta}$. Let $\bar{\Sigma}_k$ be the right-hand side of (3.27), then there exists a constant $\beta_4 > 0$ such that

$$\begin{aligned} & 4\big(\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]} \big) \\ & + \frac{6\gamma_6 \sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^{k-1}]} + \frac{6\sqrt{r}}{\rho\gamma} \sqrt{\mathbb{E}[\Lambda_1^k]} \leq \beta_4 \bar{\Sigma}_k. \end{aligned} \tag{A.17}$$

Plugging (A.16) and (A.17) into (A.15) yields

$$\Delta_{k+1} \leq \beta_4 \bar{\Sigma}_k + \beta_1 (\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k})^{1-\theta} + \beta_1\beta_2 (\mathbb{E}[\Lambda_1^{k+1}] + \mathbb{E}[\Lambda_1^k] + \mathbb{E}[\Lambda_1^{k-1}])^{1-\theta}. \tag{A.18}$$

From (3.24), it follows that

$$(\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k})^{1-\theta} \leq (\sigma_0(1-\theta)\mathbb{E}[\mathrm{dist}(0, \partial\mathscr{L}_s(z^k))])^{\frac{1-\theta}{\theta}}. \tag{A.19}$$

Since $2\theta \geq 1$, we have

$$\begin{aligned}(\mathbb{E}[\Lambda_1^{k+1}] + \mathbb{E}[\Lambda_1^k] + \mathbb{E}[\Lambda_1^{k-1}])^{1-\theta} &\leq (\mathbb{E}[\Lambda_1^{k+1}] + \mathbb{E}[\Lambda_1^k] + \mathbb{E}[\Lambda_1^{k-1}])^{\frac{1-\theta}{2\theta}} \\ &\leq (\sqrt{\mathbb{E}[\Lambda_1^{k+1}]} + \sqrt{\mathbb{E}[\Lambda_1^k]} + \sqrt{\mathbb{E}[\Lambda_1^{k-1}]})^{\frac{1-\theta}{\theta}},\end{aligned}$$

which further implies that a constant $\beta_3 > 0$ exists such that

$$(\mathbb{E}[\Lambda_1^{k+1}] + \mathbb{E}[\Lambda_1^k] + \mathbb{E}[\Lambda_1^{k-1}])^{1-\theta} \leq \beta_3 \bar{\Sigma}_k^{\frac{1-\theta}{\theta}}. \tag{A.20}$$

It follows from (A.19) and (3.27) that

$$(\mathbb{E}[\mathscr{L}_s(z^k)] - \bar{\mathscr{L}}_{s,k})^{1-\theta} \leq (\sigma_0(1-\theta)\bar{\Sigma}_k)^{\frac{1-\theta}{\theta}}. \tag{A.21}$$

Substituting (A.21) and (A.20) into (A.18) gives

$$\Delta_{k+1} \leq ((\sigma_0(1-\theta))^{\frac{1-\theta}{\theta}}\beta_1 + \beta_1\beta_2\beta_3)\bar{\Sigma}_k^{\frac{1-\theta}{\theta}} + \beta_4\bar{\Sigma}_k \leq \varrho\bar{\Sigma}_k^{\frac{1-\theta}{\theta}}, \tag{A.22}$$

where the second inequality is derived from $\frac{1-\theta}{\theta} \leq 1$, $\bar{\Sigma}_k \to 0$ and $\varrho := (\sigma_0(1-\theta))^{\frac{1-\theta}{\theta}}\beta_1 + \beta_1\beta_2\beta_3 + \beta_4$. To bound $\bar{\Sigma}_k$, from (3.22) and (3.23), we have

$$\begin{aligned}\bar{\Sigma}_k \leq& (\sqrt{r} + \gamma_6\sqrt{r} + \sqrt{r\sigma_\Lambda})\Big(\sqrt{\mathbb{E}[\|x^{k+1}-x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k-x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1}-x^{k-2}\|^2]}\Big) \\ &+ (2\gamma_6\sqrt{r} + (2-\rho)\sqrt{r})\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} - (\gamma_6\sqrt{r} + (1-\frac{\rho}{2})\sqrt{r})\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} \\ \leq& \beta_5\Big(\sqrt{\mathbb{E}[\|x^{k+1}-x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k-x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1}-x^{k-2}\|^2]}\Big) \\ &+ 2\beta_6\Big(\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} - \sqrt{\mathbb{E}[\Lambda_1^k]}\Big) - (\gamma_6\sqrt{r} + (1-\frac{\rho}{2})\sqrt{r})\sqrt{\mathbb{E}[\Lambda_1^{k-1}]},\end{aligned} \tag{A.23}$$

where $\beta_5 := \sqrt{r} + \gamma_6\sqrt{r} + \frac{4+4\gamma_6-\rho}{\rho}\sqrt{r\sigma_\Lambda}$, $\beta_6 := \frac{2+2\gamma_6-\rho}{\rho}\sqrt{r}$. Let $\Delta_k^\Lambda := \Delta_k + \frac{2\beta_6(1-\frac{\rho}{4})}{\beta_5}\sqrt{\mathbb{E}[\Lambda_1^{k-1}]}$. Then,

$$\begin{aligned}(\Delta_{k+1}^\Lambda)^{\frac{\theta}{1-\theta}} &\leq \frac{2^{\frac{\theta}{1-\theta}}}{2}\Delta_{k+1}^{\frac{\theta}{1-\theta}} + \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(\frac{2\beta_6(1-\frac{\rho}{4})}{\beta_5}\sqrt{\mathbb{E}[\Lambda_1^k]}\right)^{\frac{\theta}{1-\theta}} \\ &\leq \frac{(2\varrho)^{\frac{\theta}{1-\theta}}}{2}\bar{\Sigma}_k + \frac{(\frac{4\beta_6(1-\frac{\rho}{4})}{\beta_5})^{\frac{\theta}{1-\theta}}}{2}\sqrt{\mathbb{E}[\Lambda_1^k]},\end{aligned} \tag{A.24}$$

where the first inequality is obtained by using $\frac{\theta}{1-\theta} \geq 1$ and $(a+b)^v \leq 2^{v-1}a^v + 2^{v-1}b^v$ for any $v \geq 1$, and the second inequality is from (A.22). Substituting (A.23) into (A.24) and rearranging the terms, we have

$$(\Delta_{k+1}^\Lambda)^{\frac{\theta}{1-\theta}} \leq \frac{\beta_5(2\varrho)^{\frac{\theta}{1-\theta}}}{2}(\Delta_k^\Lambda - \Delta_{k+1}^\Lambda),$$

which also gives

$$\Delta_{k+1}^{\Lambda} \leq 2\varrho(\frac{\beta_5}{2})^{\frac{1-\theta}{\theta}}(\Delta_k^{\Lambda} - \Delta_{k+1}^{\Lambda})^{\frac{1-\theta}{\theta}}. \tag{A.25}$$

Note that, the result of (A.25) is very similar to that of (A.7). Hence, the rest of the proof can be conducted in a similar way to that of Theorem 5. In specific, if $\theta \in (\frac{1}{2}, 1)$, there exist an integer $\bar{K}$, constants $\mu > 0$ and $\nu_1 = \frac{1-2\theta}{1-\theta} < 0$ such that for $n > \bar{K}$, $\Delta_n^{\Lambda} \leq \mu n^{\frac{1}{\nu_1}}$; if $\theta \in (0, \frac{1}{2}]$, there exists constants $\nu > 0$ and $\tau = \frac{\varrho\beta_5}{1+\varrho\beta_5} < 1$ such that for $k \geq K$, $\Delta_{k+1}^{\Lambda} \leq \nu\tau^{k-K}$. Since $\mathbb{E}[\|x^k - \bar{x}\|] \leq \Delta_{k+1}^{\Lambda}$, the estimations for $\mathbb{E}[\|x^k - \bar{x}\|]$ in (ii) and (iii) are derived.

Finally, we consider the estimations for $\mathbb{E}[\|y^k - \bar{y}\|]$. Combining (3.21) and (3.23) yields

$$\begin{aligned}
&\sqrt{\mathbb{E}[\|y^q - y^{q-1}\|^2]} \\
\leq&\gamma_6(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\Big(\sqrt{\mathbb{E}[\|x^{q+1} - x^q\|^2]} + \sqrt{\mathbb{E}[\|x^q - x^{q-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{q-1} - x^{q-2}\|^2]}\Big) \\
&+ \frac{2\gamma_6}{\rho}\left(\sqrt{\mathbb{E}[\Lambda_1^{q-1}]} - \sqrt{\mathbb{E}[\Lambda_1^{q}]}\right).
\end{aligned}$$

Summing up from $q = k$ to ∞, we have

$$\begin{aligned}
\sum_{q=k}^{\infty}\sqrt{\mathbb{E}[\|y^{q+1} - y^q\|^2]} &\leq \gamma_6(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\Delta_{k+1} + \frac{2\gamma_6}{\rho}\sum_{q=k}^{\infty}\left(\sqrt{\mathbb{E}[\Lambda_1^{q}]} - \sqrt{\mathbb{E}[\Lambda_1^{q+1}]}\right) \\
&\leq \gamma_6(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\Delta_{k+1} + \frac{2\gamma_6}{\rho}\sqrt{\mathbb{E}[\Lambda_1^{k}]}.
\end{aligned} \tag{A.26}$$

Let $V_k := \gamma_6(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\Delta_k + \frac{2\gamma_6}{\rho}\sqrt{\mathbb{E}[\Lambda_1^{k-1}]}$. Then, by (A.15), it holds that

$$\begin{aligned}
V_{k+1} \leq\ & 4\gamma_6(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\Big(\sqrt{\mathbb{E}[\|x^{k+1} - x^k\|^2]} + \sqrt{\mathbb{E}[\|x^k - x^{k-1}\|^2]} + \sqrt{\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2]}\Big) \\
&+ \frac{6\gamma_6^2\sqrt{r}}{\rho\gamma}(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\sqrt{\mathbb{E}[\Lambda_1^{k-1}]} + \frac{6\gamma_6\sqrt{r}}{\rho\gamma}(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})\sqrt{\mathbb{E}[\Lambda_1^{k}]} \\
&+ \gamma_6\beta_1(1 + \frac{2\sqrt{\sigma_\Lambda}}{\rho})(\mathbb{E}[\mathscr{L}_{s,k}^{\Lambda}] - \bar{\mathscr{L}}_{s,k})^{1-\theta} + \frac{2\gamma_6}{\rho}\sqrt{\mathbb{E}[\Lambda_1^{k}]}.
\end{aligned}$$

By the same line used to obtain (A.22), there is a constant $\varrho' > 0$ such that $V_{k+1} \leq \varrho'\bar{\Sigma}_k^{\frac{1-\theta}{\theta}}$. Similar to the method used to obtain the estimations of Δ_{k+1}^{Λ}, for

$$V_k^{\Lambda} := V_k + \frac{2\beta_6(1 - \frac{\rho}{4})}{\beta_5}\sqrt{\mathbb{E}[\Lambda_1^{k-1}]},$$

we can obtain that

$$V_k^{\Lambda} \leq \bar{\mu}k^{\frac{1}{\nu_1}} \text{ for } \theta \in (1/2, 1) \tag{A.27}$$

and

$$V_{k+1}^{\Lambda} \leq \bar{\nu}\tau^{k-K} \text{ for } \theta \in (0, 1/2], \tag{A.28}$$

where $\bar{\mu}$ and $\bar{\nu}$ are some positive constants. The triangle inequality gives

$$\mathbb{E}[\|y^k - \bar{y}\|] \leq V_{k+1} \leq V_{k+1}^{\Lambda},$$

which, together with (A.27) and (A.28), implies the estimations for $\mathbb{E}[\|y^k - \bar{y}\|]$ in (ii) and (iii). The proof is completed. □

AIMS Press