AIMS Mathematics, 10(10): 24804–24835.
DOI: 10.3934/math.20251099
Received: 01 August 2025
Revised: 09 September 2025
Accepted: 09 October 2025
Published: 29 October 2025

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

Research article

Novel inertial stochastic Bregman inexact ADMMs for solving large-scale nonconvex and nonsmooth optimization without relying on the Kurdyka–Łojasiewicz property

Yi-xin Yang[1], **Heng-you Lan**[1,2,*] **and Lin-cheng Jiang**[1]

[1] College of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China

[2] Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing, Zigong 643000, China

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

Abstract: As the scale of optimization problems expands, the performance of the alternating direction method of multipliers (ADMM) exhibits a significant downward trend. In this paper, aiming at solving nonconvex, nonsmooth optimization problems under large-scale linear constraints, we proposed a unified framework of novel stochastic inexact ADMMs incorporating inertial terms and Bregman distances. By fusing the Bregman distance with inertial acceleration techniques, the framework not only covers stochastic gradient descent and existing variance-reduced gradient estimation techniques such as the stochastic variance-reduced gradient and stochastic recursive gradient, but also allows for a more flexible double-step strategy in convergence analysis. Without depending on the Kurdyka–Łojasiewicz property and under some suitable mild conditions, we demonstrated global convergence of this unified framework, and showed that it achieves a sublinear convergence rate of $O(1/\mathbb{K})$, where $\mathbb{K}$ is the number of iterations. Further, under error bound conditions, the linear convergence rate of the stochastic inexact ADMMs was established. Finally, the effectiveness of stochastic inexact ADMMs for solving some nonsmooth and nonconvex problems was verified by numerical experiments on the graphically guided fusion LASSO problems.

Keywords: convergence rate; nonconvex, nonsmooth optimization; variance-reduced gradient; stochastic inexact ADMMs; inertial techniques; Bregman distance
Mathematics Subject Classification: 41A25, 62L20, 90C26, 49J52

1. Introduction

In the 1970s, Clark's subgradient theory [1] promoted the development of nonsmooth optimization. However, the nonsmoothness of the objective function in practical problems such as image denoising and reconstruction [2], machine learning [3] and so on, limits the effectiveness of traditional gradient methods. For high-dimensional nonsmooth problems, stochastic optimization is a powerful tool for complex problems due to its fast convergence [4].

It is well known that in practical applications, traditional statistical LASSO models are difficult to directly portray complex dependencies between variables, while graph-guided fusion LASSO models aim to incorporate the variable graph structure into the fusion penalty to reflect the variable correlation, and to improve the model performance and interpretability in a high-dimensional data environment. For the total number $n \in \mathbb{N}$ of samples, whose values may be very large, let for $i \in \{1, 2, \cdots, n\}$, $a_i \in \mathbb{R}^{n_a}$ denote the input data, $b_i \in \{-1, +1\}$ be the corresponding label of i, $\{a_i, b_i\}_{i=1}^n$ be a set of training samples, $f_i(\mathbf{x}) = \dfrac{1}{1+\exp(b_i a_i^T \mathbf{x})} + \dfrac{\gamma_2}{2}\|\mathbf{x}\|_2^2$ be a nonconvex differentiable loss function from $\mathbb{R}^{n_a}$ to $\mathbb{R}$, γ_1 and γ_2 be the regularization parameters, and the matrix $A \in \mathbb{R}^{m\times n_a}$ be the sparse pattern of the graph, which comes from the sparse inverse covariance matrix estimation [5]. The graph-guided fusion LASSO model can be expressed as follows [6]:

$$\min_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) + \gamma_1\|A\mathbf{x}\|_1. \tag{1.1}$$

In general, via introducing an auxiliary variable $\mathbf{z} \in \mathbb{R}^m$ with $\mathbf{z} = A\mathbf{x}$ in (1.1), we would consider reformulating model (1.1) as

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) + \gamma_1\|\mathbf{z}\|_1, \\ &\text{s. t. } A\mathbf{x} - \mathbf{z} = 0. \end{aligned} \tag{1.2}$$

It is noteworthy that many scholars have explored constrained nonconvex optimization problems in the shape of (1.2). See, for example, [7, 8] and the references therein. In particular, Liu et al. [9] effectively dealt with such problems using an iterative soft-thresholding algorithm, with the core objective of achieving sparse feature selection and model regularization. For the specific form of the soft-thresholding operator, the reader is referred to (5.2).

In order to solve (1.1) or (1.2), the general model of (1.2) will be investigated in this paper. In fact, taking the vector $\mathbf{b} \in \mathbb{R}^m$ and two functions $f : \mathbb{R}^{n_a} \to \mathbb{R}$ and $g : \mathbb{R}^{n_b} \to \mathbb{R}$, and setting $\mathbb{R}^m \ni B\mathbf{y} = -\mathbf{z}$ for $B \in \mathbb{R}^{m\times n_b}$, then (1.2) can be extended to the following nonconvex, nonsmooth optimization problem:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x}) + g(\mathbf{y}), \\ \text{s.t.} \quad & A\mathbf{x} + B\mathbf{y} = \mathbf{b}. \end{aligned} \tag{1.3}$$

If $n_b = m$, $B = I$, the identity matrix, $g(\mathbf{y}) = \gamma_1\|\mathbf{y}\|_1$, $\mathbf{y} = -\mathbf{z}$, and $f(\mathbf{x}) = \dfrac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, then (1.3) degenerates to (1.2). Furthermore, f is a nonconvex and smooth function, and g is a locally Lipschitz-continuous function, which may be nonconvex and nonsmooth, and is often used as a regularization term to prevent overfitting. In practice, a large number of problems can also be presented in the forms

of (1.1)–(1.3), where specific applications can be found in [10, 11] and the references therein. To solve the graph-guided fusion LASSO model (1.1), we first consider the solution of (1.3). Further, it is worth exploring a class of novel algorithms to efficiently and stably solve the large-scale optimization problem (1.3).

With the advent of the big data era, the deterministic alternating direction method of multipliers (ADMM) has difficulty handling large-scale optimization problems due to the need to compute the full gradient or solve the subproblems exactly in iterations. Stochastic optimization methods become a powerful tool to solve such problems, and researchers introduced stochastic gradient descent (SGD) into the ADMM, but its convergence speed is limited by the variance of the stochastic gradient [12, 13]. For this reason, researchers reduced the variance of stochastic gradient estimation by the variance reduction (VR) technique to make the algorithm iterations more stable [14, 15]. In particular, gradient estimators such as the stochastic variance-reduced gradient (SVRG) [16] and stochastic recursive gradient (SARAH) [17] have gained increasing popularity, which has greatly promoted the research on the stochastic ADMM. Recent work found that the ADMM and its variants can handle these complex, nonconvex problems well [18, 19]. In the nonconvex setting, Huang et al. [20] first provided a general framework for analyzing the iterative complexity reduction of nonconvex stochastic ADMM variance, including the SVRG-ADMM, stochastic average gradient ADMM (SAG-ADMM), and SAG-ADMM optimization forms. Subsequently, Bian et al. [21] proposed a stochastic ADMM framework for large-scale, nonconvex, nonsmooth optimization problems, which combines the variance-reducing stochastic gradient technique to improve the performance of the algorithms. On the other hand, Zeng et al. [22, 23] considered the convergence of the SVRG-ADMM, which combines Katyusha momentum and the variance-reducing stochastic ADMM, in a nonconvex setting. For more related works, readers may refer to [24, 25] and the references therein.

However, when the problem is extended to nonconvex domains, it becomes challenging to solve the subproblems of the traditional ADMM exactly [26]. To overcome this issue, Wang et al. [26] improved the ADMM by introducing the Bregman distance, which indeed simplifies the original subproblem and improves the algorithm performance. Inspired by this, Chao et al. [27] adopted a linearization method and a suitable Bregman distance, which can better exploit the structure of subproblems to solve nonconvex optimization problems. Here, the Bregman distance plays an important role in the iterative algorithms. Applying the Bregman distance for certain subproblems of the Peaceman–Rachford splitting method, Liu et al. [28] first constructed a Bregman Peaceman–Rachford splitting framework for nonconvex, nonseparable optimization, which overcomes the constraints of traditional algorithms via Bregman distance and double relaxation factors. Further, Liu et al. [28] observed that there is an interesting phenomenon that the descent of the merit function (the augmented Lagrangian function) is guaranteed due to the Bregman distance, and completed a comprehensive theoretical analysis covering global convergence, strong convergence, and the convergence rate. Recently, Guo and Tan [29] proposed another novel Bregman ADMM that can return to the ADMM while avoiding the need for global Lipschitz continuity of the gradient and extended the algorithm to a wider range of nonconvex problems. Thus, to address the challenges of nonconvex optimization, Bregman-type nonconvex splitting algorithms are worthy of further exploration. Whereupon, Liu et al. [9] designed a stochastic inexact ADMMs incorporating Bregman distances and inertial terms to establish global convergence via the Kurdyka–Łojasiewicz

(KL) property. We note that the KL property imposes a stringent restriction on the structure of the function, requiring that the function satisfies the gradient decay condition in its domain of definition, but nonconvex and nonsmooth functions in practice may not satisfy this property. Thus, instead of relying on the powerful KL property utilized in [9, 28, 30], we shall construct a class of novel stochastic ADMMs and achieve global convergence of the novel stochastic ADMMs solving the nonconvex, nonsmooth optimization problem (1.3) through more moderate assumptions such as gradient continuity, bounded variance, a range condition, etc., in combination with the analysis of the decreasing potential function and the boundedness of the iteration stepsize.

In addition, inertial techniques have been increasingly used in various algorithms to improve their numerical efficiency. Polyak [31] proposed the heavy ball method, which introduced the concept of momentum into gradient descent for the first time. Further, inertial techniques have also been applied to the ADMM to solve nonconvex problems. Hien and Papadimitriou [32] introduced inertial techniques to the nonconvex ADMM with nonlinear coupling constraints for the first time to improve the efficiency of the algorithm through noninertial extrapolation. Recently, some accelerated stochastic ADMM methods have also been proposed to efficiently solve large-scale learning problems [33, 34]. In particular, Liu et al. [9] introduced inertia by directly adding a term related to the difference of the variables in the previous step in the iterative update formula, which helped the algorithm to perform global convergence. Different from [9], we will use the extrapolation proposed by Chao et al. [35] to obtain the intermediate variables as inertial terms. This allows the solutions to subsequent subproblems to start from more favorable points, accelerating the convergence speed and thus enhancing the practicality of the algorithm.

Although convergence of the stochastic ADMM under nonconvex conditions has been investigated in [20], a unified theoretical and analytical framework for reduced-variance stochastic ADMM methods for problem (1.3) has not been reported. Recently, Bai et al. [36] proposed an inexact ADMM to solve the subproblems inexactly via an adaptive relative error criterion, which allows the pairwise stepsize s to be in a wide range to enhance flexibility. Zeng et al. [37] employed an inexact approach to solve the subproblems within the stochastic ADMMs framework to address the non-differentiability of nonsmooth terms. In [38], the dual stepsize s is larger than the general stepsize $s = 1$ in the range $\left(0, \frac{1+\sqrt{5}}{2}\right)$. While we further extend the range of the dual stepsize to $s \in (0, 2)$, can we still achieve global convergence? Moreover, in large-scale problems, inexact methods leverage stochastic gradients $\nabla f(\mathbf{x}^k, \xi_M)$ by computing only M samples, where ξ_M denotes a random variable following a certain probability distribution with respect to the sample size M, thus significantly reducing the computational load. So, what kind of results do we get when we add the method of solving the subproblem inexactly?

Inspired by [9, 36, 37], in this paper, we propose a unified framework for a kind of inertial stochastic Bregman inexact ADMMs (ISBI-ADMMs) for solving the nonconvex, nonsmooth optimization problems (1.1) and (1.3). The ISBI-ADMMs proposed in this paper is a unified framework for the stochastic ADMM inexact analysis and is capable of handling a class of nonsmooth, nonconvex problems such as the one described in the form of (1.3). This unified framework not only incorporates Bregman distance and inertial acceleration techniques, but can be also used in conjunction with a variety of stochastic gradient estimators, including both basic SGD [39] and covering sophisticated variance-reduction techniques such as SVRG [16] and SARAH [17], the related algorithms in which can be viewed as special cases of the proposed

ISBI-ADMMs. Our main contributions can be summarized as follows:

(i) We choose suitable Bregman distances to make the subproblems of $\mathbf{x}$ and $\mathbf{y}$ easier to solve. In addition, an inertia technique is incorporated to speed up the convergence.

(ii) We further extend the range of the dual stepsize s to $(0, 2)$ to ensure the stability of the ISBI-ADMMs even under the solution conditions of the inexact subproblem. Unlike existing methods, this analytical framework does not require the assumption that nonsmooth regular terms are convex.

(iii) For the number $\mathbb{K}$ of iterations, we theoretically establish the $O(1/\mathbb{K})$ optimal sublinear convergence rate of the ISBI-ADMMs under some mild conditions by constructing a special potential energy function. In addition, the linear convergence rate of the ISBI-ADMMs is again established under stronger error bound conditions.

(iv) We apply the ISBI-ADMMs to the graph-guided fusion LASSO task, combining various stochastic gradient estimators such as SGD, SVRG, and SARAH for performance evaluation. Our findings indicate that the ISBI-ADMMs achieves the optimal performance when using the SARAH gradient estimator.

The rest of the paper is organized as follows. In Section 2, we provide basic definitions related to the proofs of the main results. In Section 3, we establish the analysis of the ISBI-ADMMs. The convergence rate analysis of the ISBI-ADMMs is established in Section 4. In Section 5, numerical experiments verify the performance benefits of the algorithm in the graphically guided fusion LASSO task. Finally, the whole paper is summarized.

2. Preliminaries

In this section, we introduce some basic notations, lemmas, and definitions, which are closely related to the content of this paper and will be used for further analysis.

We employ the notation $\|\cdot\|$ to respectively signify the Euclidean norm of a vector. P_A and σ_A denote the maximum positive eigenvalue and the minimum positive eigenvalue of matrix $A^\top A$, respectively. Let $\mathcal{F}_k$ be the σ-field produced by the random variables in the initial k iterations, and $\mathbb{E}_k$ be the expectation conditional on $\mathcal{F}_k$. Evidently, $(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ is $\mathcal{F}_k$-measurable because $\mathbf{x}^k$, $\mathbf{y}^k$, and the Lagrangian multiplier $\boldsymbol{\lambda}^k$ rely on the random gradient data from the first k iterations. For a set-valued mapping $\varrho : \mathbb{R}^n \to \mathbb{R}^m$, it is deemed outer semicontinuous at a point $\widetilde{\mathbf{x}}$ if there are $\mathbf{x}^k \to \widetilde{\mathbf{x}}$ and $v^k \to v$ with $v^k \in \varrho(\mathbf{x}^k)$ such that for any $v \in \mathbb{R}^m$, $v \in \varrho(\widetilde{\mathbf{x}})$ is valid. We use $\operatorname{con} C$ to indicate the convex hull of a given closed set C, and the distance from $\mathbf{x}$ to the set C is represented as $\operatorname{dist}(\mathbf{x}, C) := \inf_{\mathbf{y}\in C}\{\|\mathbf{x} - \mathbf{y}\|\}$.

Definition 1. *A function $\phi : \mathbb{R}^n \to \mathbb{R}$ is said to be Lipschitz continuous with Lipschitz constant L_ϕ, if there exists a positive constant L_ϕ such that for each $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$,*

$$\|\phi(\mathbf{u}_1) - \phi(\mathbf{u}_2)\| \leq L_\phi \|\mathbf{u}_1 - \mathbf{u}_2\|.$$

Definition 2. ([40], Clarke subgradient) *Consider a function $\phi : \mathbb{R}^n \to \mathbb{R}$ that is locally Lipschitz continuous on an open subset $\mathcal{S}$ of $\mathbb{R}^n$. Let C be a part of $\mathcal{S}$ where ϕ is differentiable. Then, the Clarke subgradient set of the function ϕ at a point $\tilde{x} \in \mathcal{S}$ can be formulated as*

$$\partial\phi(\tilde{x}) := \operatorname{con}\{v | \exists x \to \tilde{x}, x \in C, \nabla\phi(x) \to v\},$$

which is nonempty, convex, and compact for every $\tilde{x} \in \mathcal{S}$. Moreover, $\partial\phi$ is outer semicontinuous and locally bounded on $\mathcal{S}$. We further define the set of critical points of ϕ as

$$\operatorname{crit}\phi := \{x \in \mathbb{R}^n | \operatorname{dist}(0, \partial\phi(x)) = 0\}.$$

Definition 3. *Given an accuracy level ϵ within the interval $(0, 1)$, the point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ is referred to as an ϵ-stationary point of* (1.3) *if the following conditions are satisfied:*

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^*) - A^T\boldsymbol{\lambda}^*\right\|^2 \leq \epsilon, \quad \mathbb{E}\left[\operatorname{dist}\left(B^T\boldsymbol{\lambda}^*, \partial g(\mathbf{y}^*)\right)\right]^2 \leq \epsilon, \quad \mathbb{E}\left\|A\mathbf{x}^* + B\mathbf{y}^* - \mathbf{b}\right\|^2 \leq \epsilon.$$

Definition 4. ([26], Bregman distance) *The Bregman distance associated with a convex and differentiable function $\psi : \mathbb{R}^n \to \mathbb{R}$ is defined by*

$$D_\psi(\mathbf{x}, \mathbf{y}) := \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle\nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Specifically, if we define $\psi(\mathbf{x}) = \|\mathbf{x}\|^2$ in the above mentioned context, then it simplifies to $\|\mathbf{x} - \mathbf{y}\|^2$, which is precisely the well-known square of the Euclidean distance.

Definition 5. *We say that $\Omega^* := (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ denotes the set of all stationary points of problem* (1.3) *if it satisfies*

$$A^\top\boldsymbol{\lambda}^* = \nabla f(\mathbf{x}^*), \quad B^\top\boldsymbol{\lambda}^* \in \partial g(\mathbf{y}^*), \quad A\mathbf{x}^* + B\mathbf{y}^* = b.$$

Next, we give some lemmas relevant to this paper.

Lemma 1. ([38]) *The gradient of a function $\phi : \mathbb{R}^n \to \mathbb{R}$ satisfies Lipschitz continuity with Lipschitz constant $L_\phi > 0$. It is straightforward to infer that for arbitrary $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, the following is valid:*

$$|\,\phi(\mathbf{v}) - \phi(\mathbf{u}) - \langle\nabla\phi(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle\,| \leq \frac{L_\phi}{2}\|\mathbf{v} - \mathbf{u}\|^2.$$

Lemma 2. ([41]) *Suppose that A is a nonzero matrix in $\mathbb{R}^{m\times n_a}$, and let σ_A be the smallest positive eigenvalue of $AA^\top$. Then, for any $\mathbf{u} \in \mathbb{R}^{n_a}$, the following inequality holds:*

$$\|G_{A^\top}\mathbf{u}\| \leq \frac{1}{\sigma_A}\|A\mathbf{u}\|,$$

where $G_{A^\top}$ is the orthogonal projection operator onto the column space of the transposed matrix $A^\top$.

For convenience, let us first fix the following notations:

$$\begin{aligned}
&\mathbf{w} := (\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}),\ \mathbf{w}^k := (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k),\ \mathbf{w}^* := (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*),\\
&\overline{\mathbf{w}} := (\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\boldsymbol{\lambda}}),\ \overline{\mathbf{w}}^k := (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k, \overline{\mathbf{x}}^{k-1}, \mathbf{y}^{k-1}, \boldsymbol{\lambda}^{k-1}),\\
&\mathbf{d}_\mathbf{x}^k := \mathbf{x}^{k+1} - \mathbf{x}^k,\ \mathbf{d}_\mathbf{y}^k := \mathbf{y}^{k+1} - \mathbf{y}^k,\ \mathbf{d}_{\boldsymbol{\lambda}}^k := \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k,\ \mathbf{r}^k := A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}.
\end{aligned}$$

3. Presentation of the ISBI-ADMMs

In this section, we present the ISBI-ADMMs. The classical ADMM, originally introduced by Gabay and Mercier [42] and Glowinski and Marrocco [43], is used to solve a class of structurally separable optimization problems. The classical iterative form of the ADMM to problem (1.3) is given as

$$\begin{cases} \mathbf{y}^{k+1} \in \arg\min \left\{ \mathcal{L}_\beta \left(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k \right) \right\}, \\ \mathbf{x}^{k+1} \in \arg\min \left\{ \mathcal{L}_\beta \left(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k \right) \right\}, \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta \left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - b \right), \end{cases}$$

where $s \in (0, 2)$ represents the stepsize of the dual multiplier $\boldsymbol{\lambda}$, $\beta > 0$ is a penalty parameter, and $\mathcal{L}_\beta$ is the augmented Lagrangian function (ALF) expressed as follows:

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}) + g(\mathbf{y}) - \boldsymbol{\lambda}^\top (A\mathbf{x} + B\mathbf{y} - \mathbf{b}) + \frac{\beta}{2} \|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2.$$

In what follows, we shall separately analyze and explain the key rules for updating $\mathbf{y}$, $\mathbf{x}$, and $\boldsymbol{\lambda}$ of the proposed ISBI-ADMMs, laying the foundation for subsequent convergence analysis.

3.1. The update rule of $\mathbf{y}$

In the traditional ADMM, for the $\mathbf{y}$-subproblem involving the nonconvex and nonsmooth function g, its nonconvexity and nonsmoothness make the subproblem difficult to solve and lead to slow convergence. On one hand, based on the ALF of the traditional ADMM, [26] introduced the Bregman distance and found that an appropriate selection of the Bregman distance can indeed simplify the original subproblem. On the other hand, to improve the convergence speed of nonconvex problems, [35] introduced inertial extrapolation points $\overline{\mathbf{y}}^k$ and $\overline{\mathbf{x}}^k$. Inspired by [26, 35], in order to address the challenges brought by nonconvexity and nonsmoothness, we first update the original variable $\mathbf{y}$ to reduce the difficulty of solving the subproblem and improve the iterative efficiency. Its update rule is as follows:

$$\mathbf{y}^{k+1} = \arg\min_{\mathbf{y} \in \mathbb{R}^{n_b}} \mathcal{L}_\beta \left(\overline{\mathbf{x}}^k, \mathbf{y}, \boldsymbol{\lambda}^k \right) + \mathbf{D}_{\psi_1} \left(\mathbf{y}, \mathbf{y}^k \right), \tag{3.1}$$

where $\mathbf{D}_{\psi_1}$ is the Bregman distance. When updating $\mathbf{y}$ here, the Bregman distance and the inertia iteration point $\overline{\mathbf{x}}^k$ are incorporated. The core function of this inertial term is to provide historical directional information for $\mathbf{x}$, and it is generated via extrapolation with its specific form given by $\overline{\mathbf{x}}^k = \mathbf{x}^k + \theta \left(\mathbf{x}^k - \overline{\mathbf{x}}^{k-1} \right)$, where $\theta \in (0, 1]$ denotes the inertia coefficient. Furthermore, the optimality conditions for the variable $\mathbf{y}$ produced at the k-th iteration show that

$$\begin{aligned} & D_{\psi_1} \left(\mathbf{y}^{k+1}, \mathbf{y}^k \right) + \mathcal{L}_\beta \left(\overline{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k \right) \leq \mathcal{L}_\beta \left(\overline{\mathbf{x}}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k \right), \\ & 0 \in \partial g \left(\mathbf{y}^{k+1} \right) + \beta B^T \left(A\overline{\mathbf{x}}^k + B\mathbf{y}^{k+1} - b - \frac{1}{\beta} \boldsymbol{\lambda}^k \right) + \psi_1 \left(\mathbf{y}^{k+1} \right) - \psi_1 \left(\mathbf{y}^k \right), \end{aligned} \tag{3.2}$$

where ∂g stands for the Clarke subgradient of g, as given in Definition 2. By combining (3.2) with $\xi_{\mathbf{y}}^{k+1} \in \partial_{\mathbf{y}} \mathcal{L}_\beta \left(\overline{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k \right)$, for Lipschitz constant L_{ψ_1} of the function ψ_1, we obtain

$$\left\| \xi_{\mathbf{y}}^{k+1} \right\| \leq L_{\psi_1} \left\| \mathbf{y}^{k+1} - \mathbf{y}^k \right\|. \tag{3.3}$$

3.2. The inexact update of $\mathbf{x}$

As we all know, in practical problems, due to the nonconvexity of the function f in (1.3), there may be multiple local optimal solutions of the objective function, and the use of traditional convex optimization methods may not guarantee finding the global optimal solution. Moreover, in nonconvex, nonsmooth problems, the exact solution of subproblems is very difficult to obtain [37]. To overcome this problem, we use an inexact stochastic ADMM to allow for an approximate solution of the subproblems, thus reducing the computational complexity. In addition, for large-scale optimization problems, if the full gradient $\nabla f(\mathbf{x})$ is used, it is necessary to traverse all n samples, which is extremely time-consuming. Therefore, we employ the stochastic gradient estimator $\nabla f(\mathbf{x}, \xi_M)$. Similar to Subsetion 3.1, we also incorporate the Bregman distance to simplify the $\mathbf{x}$-subproblem. For a mini-batch sample size $M \in [1, n]$ in the stochastic gradient estimator $\nabla f(\mathbf{x}, \xi_M) = \frac{1}{M}\sum_{i=1}^{M} \nabla f(\mathbf{x}, \xi_i)$, where n denotes the total number of samples and $\{\xi_i\}$ represents a collection of independent and identically distributed random variables that satisfy $\mathbb{E}[\nabla f(\mathbf{x}, \xi_i)] = \nabla f(\mathbf{x})$ for $i = 1, 2, \cdots, M$, the inexact update of $\mathbf{x}$ is expressed as follows:

$$\mathbf{x}^{k+1} \approx \underset{\mathbf{x}\in\mathbb{R}^{na}}{\arg\min} \left\langle \nabla f\left(\mathbf{x}^k, \xi_M\right), \mathbf{x} - \mathbf{x}^k \right\rangle + \frac{\beta}{2}\left\| A\mathbf{x} + B\mathbf{y}^{k+1} - \mathbf{b} - \frac{\boldsymbol{\lambda}^k}{\beta}\right\|^2 + \mathbf{D}_{\psi_2}\left(\mathbf{x}, \mathbf{x}^k\right), \tag{3.4}$$

where $\mathbf{D}_{\psi_2}$ is the Bregman distance. Furthermore, to improve the generality of the ISBI-ADMMs, we have given the inexact criteria for updating $\mathbf{x}$ as

$$\begin{cases} \mathbb{E}_k\left[\left\|\xi_{\mathbf{x}}^{k+1}\right\|^2\right] \le (L_{\psi_2}\delta)^2\mathbb{E}_k\left(\dfrac{\sigma^2}{M} + \left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 + \left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2\right), \\ \mathbb{E}_k\left[\mathcal{L}_\beta\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k\right)\right] \le \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k\right) - \dfrac{\delta}{2}\mathbb{E}_k\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \dfrac{\left(\zeta_{\psi_2}\delta\right)^2}{2}\dfrac{\sigma^2}{M}, \end{cases} \tag{3.5}$$

where $\xi_{\mathbf{x}}^{k+1} = \nabla_{\mathbf{x}}\mathcal{L}_\beta\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k\right)$, and $\zeta_{\psi_2} > 0$ and $L_{\psi_2} > 0$ are any constant and the Lipschitz constant associated with the function ψ_2, respectively.

3.3. The update rule of $\boldsymbol{\lambda}$

Yin et al. [38] were the first to associate the dual stepsize with the relaxation factor of the Bregman distance, breaking through the limitation that the dual stepsize of the traditional ADMM is fixed at 1. Building on this work, for the update of the dual multiplier $\boldsymbol{\lambda}$, we consider not only the penalty parameter β but also the stepsize s. Specifically, the final step updates the dual multiplier $\boldsymbol{\lambda}$ in accordance with the following rule for the stepsize s of the dual multiplier $\boldsymbol{\lambda}$ and penalty parameter β:

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right).$$

When $s = 1$, the above equation degenerates to the classical form of the dual multiplier $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - \beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)$ given in [35]. In nonconvex and nonsmooth optimization, the objective function contains a large number of local optimal solutions and saddle points, and the dual stepsize s is a core parameter that connects the constraint residual to the update of the dual variable. A narrow range of dual stepsizes may lead to insufficient update magnitude of the dual variable, trapping the

algorithm in suboptimal local regions. On the other hand, expanding this range helps in discovering the global optimum or better local optima, ensures convergence, and avoids residual oscillations caused by excessively large stepsizes. Therefore, in [38], the value range of parameter s has been extended from the traditional $\{1\}$ to $\left(0, \frac{1+\sqrt{5}}{2}\right)$. To further break through this limitation, we expand its range to $(0, 2)$.

On the basis of the aforementioned update rules for variables $\mathbf{y}$, $\mathbf{x}$, and the dual multiplier $\boldsymbol{\lambda}$, we have now fully derived the ISBI-ADMMs (i.e., Algorithm 1) tailored for solving the nonconvex, nonsmooth optimization problem (1.3). The key characteristics of this algorithm are described as follows.

Algorithm 1 ISBI-ADMMs

Input: Initial values of $\mathbf{w}^1 = (\mathbf{x}^1, \mathbf{y}^1, \boldsymbol{\lambda}^1)$. Given two convex and differentiable functions ψ_1 and ψ_2, constants $\beta > 0$, $\theta \in (0, 1]$, and $s \in (0, 2)$. Select the mini-batch sample size M to be the same as in Subsection 3.2 and a sufficiently small $\mathfrak{q}$, and set $k = 1$.

for $k = 1, 2, \cdots$ **do**

1° Compute $\overline{\mathbf{x}}^k = \mathbf{x}^k + \theta\left(\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right)$.

2° Calculate the subproblem (3.1).

3° Solve the subproblem (3.4).

4° Update dual multiplier $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)$.

5° **If** $\rho \leq \mathfrak{q}$, where the parameter ρ appearing could be related to $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|$ and $\|\mathbf{y}^{k+1} - \mathbf{y}^k\|$, **then break**

else

$k = k + 1$ **and continue**

End for

6° **Output:** Iterates $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\lambda}$ chosen uniformly random from $\left\{\mathbf{w}^k\right\}$.

Remark 1. (i) *In* Algorithm 1*, s in Step 4 is the dual stepsize. Expanding the range of the dual stepsize helps to discover the global optimum or better local optima and prevents the algorithm from getting trapped in suboptimal local regions. Therefore, we have extended the value range of s to* $(0, 2)$.
(ii) *$f(\mathbf{x})$ is a nonconvex smooth function. Nonconvexity slows iterative convergence and makes its corresponding subproblems hard to solve. Therefore, we use inertia to reduce subproblem-solving difficulty and speed up convergence. The inertial term provides historical directional information for* $\mathbf{x}$ *and performs extrapolation to generate the equation in Step 1 of* Algorithm 1*, which brings the iterative starting point of* $\mathbf{x}$ *closer to the optimal region and thus accelerates convergence.*
(iii) *In the update scheme of the* $\mathbf{y}$*-subproblem* (3.1) *in Step 2 of* Algorithm 1*, the inertial iteration point* $\overline{\mathbf{x}}^k$ *from Step 1 of* Algorithm 1 *and the update of* $\boldsymbol{\lambda}$ *from Step 4 are incorporated, which allows us to reduce the difficulty of solving the* $\mathbf{y}$*-subproblem. To address the challenges posed by nonconvexity and nonsmoothness, we adopt a stochastic gradient estimator* $\nabla f(\mathbf{x}, \xi_M)$ *and the Bregman distance* $\mathbf{D}_{\psi_2}$ *to simplify* (3.4).

4. Convergence analysis

In this section, we establish global convergence and the convergence rate of the proposed ISBI-ADMMs under appropriate conditions. Before giving the convergence analysis, we make the following basic assumptions, which are essential to the convergence analysis.

Assumption 1. (a) *The gradient mappings* ∇f *relating to the function* f *in* (1.3) *and* $\nabla\psi_i$ $(i = 1, 2)$ *associated with* ψ_i *in* Algorithm 1 *are separately Lipschitz continuous with Lipschitz constants* L_f *and* L_{ψ_i}.
(b) *There exist constant* $\sigma > 0$ *and a batch size* M *such that*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}, \xi_M) - \nabla f(\mathbf{x})\|^2\right] \leq \frac{\sigma^2}{M},$$

where $\nabla f(\mathbf{x}, \xi_M)$ *is the same as in* (3.4).
(c) *Consider a convex and differentiable function* ψ *and the corresponding Bregman distance* D_ψ. *If* ψ *is* δ*-strongly convex, then for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, *we have*

$$D_\psi(\mathbf{x}, \mathbf{y}) \geq \frac{\delta}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

(d) *The union of the image space of matrix* B *and the vector* $\mathbf{b}$ *is the subset of the image space of matrix* A, *i.e.,* $\mathrm{Im}(B) \cup \mathbf{b} \subseteq \mathrm{Im}(A)$. *Here,* $\mathrm{Im}(\Xi) := \{\Xi x |\, x \in \mathbb{R}^n\}$ *gives the image of any specified matrix* $\Xi \in \mathbb{R}^{m\times n}$. *It is straightforward to deduce that* $\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k = -s\beta\mathbf{r}^{k+1} \in \mathrm{Im}(A)$, *which implies that*

$$\left\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\right\| \leq \sigma_A^{-\frac{1}{2}} \left\|A^\top\left(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\right)\right\|,$$

where σ_A *represents the smallest positive eigenvalue of* $A^\top A$, (*or equivalently, the smallest positive eigenvalue of* $AA^\top$). *In particular,* Assumption 1(d) *can be guaranteed when* A *is nonsingular or has full column or full row rank.*

4.1. Characteristics of the dual multipliers

Before analyzing convergence and the rate of convergence of Algorithm 1, an important lemma on the dual multipliers is first given, which is crucial for the subsequent analytical process.

Lemma 3. *Suppose that* Assumptions 1(a)–(c) *hold, and* $\left\{\mathbf{w}^k := \left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right)\right\}$ *is a sequence of iterations satisfying* (3.5). *Then the following inequality holds:*

$$\begin{aligned}\mathbb{E}\left[\left\|A^\top\mathbf{d}_\lambda^k\right\|^2\right] \leq\ & H_2(s)\mathbb{E}\left(\left\|A^\top\mathbf{d}_\lambda^{k-1}\right\|^2 - \left\|A^\top\mathbf{d}_\lambda^k\right\|^2\right) + H_1(s)\left(2L_f^2 + 4L_{\psi_2}^2\delta^2\right)\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{x}^k\right\|^2\right] \\ & + 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{x}^{k-1}\right\|^2\right] + 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^k\right\|^2\right] \\ & + 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^{k-1}\right\|^2\right] + 8H_1(s)\frac{L_{\psi_2}^2\delta^2\sigma^2}{M},\end{aligned}$$

where

$$H_1(s) = \max\left\{1, \frac{s^2}{(2-s)^2}\right\}, \quad H_2(s) = \max\left\{\frac{1-s}{s}, \frac{s-1}{2-s}\right\}. \tag{4.1}$$

Proof. From the definition of $\xi_{\mathbf{x}}^{k+1} = \nabla_{\mathbf{x}}\mathcal{L}_{\beta}\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k}\right)$, it follows that

$$\xi_{\mathbf{x}}^{k+1} = \nabla f\left(\mathbf{x}^{k+1}\right) + A^{\top}\left[-\boldsymbol{\lambda}^{k} + \beta\mathbf{r}^{k+1}\right],$$

where $\mathbf{r}^{k+1} = A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}$. Subsequently, we obtain

$$A^{\top}\boldsymbol{\lambda}^{k} = \nabla f\left(\mathbf{x}^{k+1}\right) - \xi_{\mathbf{x}}^{k+1} + \beta A^{\top}\mathbf{r}^{k+1}.$$

By $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^{k} - s\beta\mathbf{r}^{k+1}$, one gets

$$sA^{\top}\boldsymbol{\lambda}^{k} = s\left(\nabla f\left(\mathbf{x}^{k+1}\right) - \xi_{\mathbf{x}}^{k+1}\right) + A^{\top}\left(\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{k+1}\right).$$

This results in

$$\begin{aligned}
A^{\top}\boldsymbol{\lambda}^{k+1} &= s\left(\nabla f\left(\mathbf{x}^{k+1}\right) - \xi_{\mathbf{x}}^{k+1}\right) + (1-s)A^{\top}\boldsymbol{\lambda}^{k},\\
A^{\top}\boldsymbol{\lambda}^{k} &= s\left(\nabla f\left(\mathbf{x}^{k}\right) - \xi_{\mathbf{x}}^{k}\right) + (1-s)A^{\top}\boldsymbol{\lambda}^{k-1},\\
A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k} &= s\left(\nabla f\left(\mathbf{x}^{k+1}\right) - \nabla f\left(\mathbf{x}^{k}\right) + \xi_{\mathbf{x}}^{k} - \xi_{\mathbf{x}}^{k+1}\right) + (1-s)A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}.
\end{aligned}$$

Letting $\alpha^{k} = \nabla f\left(\mathbf{x}^{k+1}\right) - \nabla f\left(\mathbf{x}^{k}\right) + \xi_{\mathbf{x}}^{k} - \xi_{\mathbf{x}}^{k+1}$, then we have

$$A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k} = s\alpha^{k} + (1-s)A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}. \tag{4.2}$$

Now, we examine two distinct cases for $s \in (0, 1]$ and $s \in (1, 2)$, respectively.

Case 1: $s \in (0, 1]$.

By combining (4.2) and the convexity of $\|\cdot\|^2$, we have

$$\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2 \leq s\left\|\alpha^{k}\right\|^2 + (1-s)\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2.$$

Subtracting $(1-s)\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2$ and dividing both sides by s from the above mentioned inequality, we get

$$\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2 \leq \left\|\boldsymbol{\alpha}^{k}\right\|^2 + \frac{1-s}{s}\left(\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2 - \left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2\right). \tag{4.3}$$

Case 2: $s \in (1, 2)$.

Based on (4.2), one has

$$\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2 = (1-s)^2\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2 + s^2\left\|\alpha^{k}\right\|^2 + 2s(1-s)\left\langle A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}, \alpha^{k}\right\rangle.$$

Combining the above derived result with the Cauchy–Schwarz inequality, we get for $\nu > 0$,

$$\begin{aligned}
\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k}\right\|^2 &\leq (1-s)^2\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2 + s^2\left\|\alpha^{k}\right\|^2 + s(s-1)\left(\nu\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2 + \frac{1}{\nu}\left\|\alpha^{k}\right\|^2\right)\\
&= \left((1-s)^2 + s(s-1)\nu\right)\left\|A^{\top}\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\|^2 + \left(s^2 + \frac{s(s-1)}{\nu}\right)\left\|\alpha^{k}\right\|^2.
\end{aligned} \tag{4.4}$$

When $\nu = \dfrac{2-s}{s}$ is chosen and (4.4) is reused, we have

$$(1-s)^2 + s(s-1)\nu = s - 1, \quad s^2 + \frac{s(s-1)}{\nu} = \frac{s^2}{2-s},$$

and

$$\left\|A^\top \mathbf{d}_\lambda^k\right\|^2 \leq (s-1)\left\|A^\top \mathbf{d}_\lambda^{k-1}\right\|^2 + \frac{s^2}{2-s}\left\|\alpha^k\right\|^2.$$

By subtracting $(s-1)\left\|A^\top \mathbf{d}_\lambda^k\right\|^2$ and dividing both sides by $(2-s)$ from the above inequality, we obtain

$$\left\|A^\top \mathbf{d}_\lambda^k\right\|^2 \leq \frac{s^2}{(2-s)^2}\left\|\alpha^k\right\|^2 + \frac{s-1}{2-s}\left(\left\|A^\top \mathbf{d}_\lambda^{k-1}\right\|^2 - \left\|A^\top \mathbf{d}_\lambda^k\right\|^2\right). \tag{4.5}$$

Combining (4.3) and (4.5), and taking into account the definitions of H_1 and H_2 in (4.1), one can further get

$$\left\|A^\top \mathbf{d}_\lambda^k\right\|^2 \leq H_1(s)\left\|\alpha^k\right\|^2 + H_2(s)\left(\left\|A^\top \mathbf{d}_\lambda^{k-1}\right\|^2 - \left\|A^\top \mathbf{d}_\lambda^k\right\|^2\right). \tag{4.6}$$

By (3.5), the property $\mathbb{E}[\mathbb{E}[\cdot \mid \mathcal{F}_k]] = \mathbb{E}[\cdot]$, Assumption 1(a), and the Cauchy–Schwartz inequality, we can derive

$$\begin{aligned}
\mathbb{E}\left[\left\|\alpha^k\right\|^2\right] &= \mathbb{E}\left[\left\|\nabla f\left(\mathbf{x}^{k+1}\right) - \nabla f\left(\mathbf{x}^k\right) + \xi_{\mathbf{x}}^k - \xi_{\mathbf{x}}^{k+1}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left(\left\|\nabla f\left(\mathbf{x}^{k+1}\right) - \nabla f\left(\mathbf{x}^k\right)\right\| + \left\|\xi_{\mathbf{x}}^k\right\| + \left\|\xi_{\mathbf{x}}^{k+1}\right\|\right)^2\right] \\
&\leq 2L_f^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + 4\mathbb{E}\left[\left\|\xi_{\mathbf{x}}^k\right\|^2\right] + 4\mathbb{E}\left[\left\|\xi_{\mathbf{x}}^{k+1}\right\|^2\right] \\
&\leq 2L_f^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + 4(L_{\psi_2}\delta)^2\left(\frac{2\sigma^2}{M} + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2\right]\right) \\
&\leq \left(2L_f^2 + 4L_{\psi_2}\delta^2\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + 4(L_{\psi_2}\delta)^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2\right] + 4(L_{\psi_2}\delta)^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] \\
&\quad + 4(L_{\psi_2}\delta)^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2\right] + \frac{8L_{\psi_2}^2\delta^2\sigma^2}{M}.
\end{aligned} \tag{4.7}$$

By combining (4.6) and (4.7), it is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\left\|A^\top \mathbf{d}_\lambda^k\right\|^2\right] \leq\ & H_2(s)\mathbb{E}\left(\left\|A^\top \mathbf{d}_\lambda^{k-1}\right\|^2 - \left\|A^\top \mathbf{d}_\lambda^k\right\|^2\right) + H_1(s)\left(2L_f^2 + 4L_{\psi_2}^2\delta^2\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] \\
&+ 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2\right] + 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] \\
&+ 4H_1(s)L_{\psi_2}^2\delta^2\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2\right] + 8H_1(s)\frac{L_{\psi_2}^2\delta^2\sigma^2}{M}.
\end{aligned}$$

This completes the proof. □

4.2. Global convergence and sublinear convergence rates

We now give important conclusions about global convergence and sublinear convergence rates.

Lemma 4. *Suppose that* Assumptions 1(a)–(d) *hold and the ALF has a lower bound. Let* $\left\{\mathbf{w}^k := \left(\mathbf{x}^k, \mathbf{y}^k, \lambda^k\right)\right\}$ *be a sequence of iterations satisfying* (3.3) and (3.5)*, and choose the parameters in*

this way:

$$\begin{cases} \dfrac{1+\tau}{s\beta\sigma_A}H_1(s)\left(2L_f^2+4L_{\psi_2}^2\delta^2\right)+\widehat{A}\le -w, \\ 2\widehat{A}-\dfrac{\delta}{2}\le -w, \quad \dfrac{L^2}{s^2\beta\sigma_A}\le -w, \\ \left(\dfrac{L^2}{s^2\beta\sigma_A}+\beta P_A\right)\theta^2-\dfrac{\delta}{2}\le -w, \end{cases} \tag{4.8}$$

where $\omega > 0$ *and* $\widehat{A} = 4\dfrac{1+\tau}{s\beta\sigma_A}H_1(s)L_{\psi_2}^2\delta^2$. *We further state that*

$$\begin{aligned} \widehat{\mathcal{L}}^k = \widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right) := \mathcal{L}_\beta\left(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right)+\widehat{A}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2+\widehat{A}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 \\ +\left(\frac{L^2}{s^2\beta\sigma_A}+\beta P_A\right)\theta^2\left\|\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right\|^2+\frac{1+\tau}{s\beta\sigma_A}H_2(s)\left\|A\mathbf{d}_{\lambda}^{k-1}\right\|^2. \end{aligned} \tag{4.9}$$

Here, the constants $\tau > 0$, $H_1(s)$, *and* $H_2(s)$ *are defined as found in* Lemma 3. *Then the following inequality holds:*

$$\min_{k\in\{0,\ldots,\mathbb{K}\}}\left\{\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right]\right\}\le\frac{\mathbb{E}\widehat{\mathcal{L}}^0-\mathbb{E}\widehat{\mathcal{L}}^{k+1}}{\mu(\mathbb{K}+1)}+\frac{\eta}{\mu}\frac{\sigma^2}{M},$$

where $\mu = \min\left\{w,\dfrac{\tau}{s\beta}\right\}$, $\eta = \dfrac{\zeta_{\psi_2}^2\delta^2}{2}+\dfrac{8(1+\tau)H_1(s)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}$, M *denotes the batch size for stochastic gradient estimation, and* $\mathbb{K}$ *denotes the total number of iterations from the initial iteration to the* $\mathbb{K}$*-th iteration.*

Proof. By (3.3) and (3.5), we obtain

$$\mathcal{L}_\beta\left(\overline{\mathbf{x}}^k,\mathbf{y}^{k+1},\boldsymbol{\lambda}^k\right)-\mathcal{L}_\beta\left(\overline{\mathbf{x}}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right)\le-\frac{\delta}{2}\left\|\mathbf{y}^{k+1}-\mathbf{y}^k\right\|^2 \tag{4.10}$$

and

$$\mathbb{E}_k\mathcal{L}_\beta\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1},\boldsymbol{\lambda}^k\right)-\mathcal{L}_\beta\left(\overline{\mathbf{x}}^k,\mathbf{y}^{k+1},\boldsymbol{\lambda}^k\right)\le\frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M}-\frac{\delta}{2}\mathbb{E}_k\left\|\mathbf{x}^{k+1}-\overline{\mathbf{x}}^k\right\|^2. \tag{4.11}$$

Then using the update rule of $\boldsymbol{\lambda}$ in subsection 3.3, one has

$$\mathcal{L}_\beta\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1},\boldsymbol{\lambda}^{k+1}\right)-\mathcal{L}_\beta\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1},\boldsymbol{\lambda}^k\right)=\frac{1}{s\beta}\left\|\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k\right\|^2. \tag{4.12}$$

On the other hand, since

$$\begin{aligned} &\mathcal{L}_\beta\left(\overline{\mathbf{x}}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right)-\mathcal{L}_\beta\left(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right) \\ &=f\left(\overline{\mathbf{x}}^k\right)-f\left(\mathbf{x}^k\right)-\left\langle\boldsymbol{\lambda}^kA^\top,\overline{\mathbf{x}}^k-\mathbf{x}^k\right\rangle+\frac{\beta}{2}\left\|A\overline{\mathbf{x}}^k+B\mathbf{y}^k-b\right\|^2-\frac{\beta}{2}\left\|A\mathbf{x}^k+B\mathbf{y}^k-b\right\|^2, \end{aligned} \tag{4.13}$$

from the optimal value condition for the $\mathbf{x}$-subproblem and Lemma 1, it follows that

$$f\left(\overline{\mathbf{x}}^k\right)-f\left(\mathbf{x}^k\right)\le\left\langle\nabla f\left(\mathbf{x}^k\right),\overline{\mathbf{x}}^k-\mathbf{x}^k\right\rangle+\frac{L}{2}\left\|\overline{\mathbf{x}}^k-\mathbf{x}^k\right\|^2=\left\langle\boldsymbol{\lambda}^kA^\top,\overline{\mathbf{x}}^k-\mathbf{x}^k\right\rangle+\frac{L}{2}\left\|\overline{\mathbf{x}}^k-\mathbf{x}^k\right\|^2. \tag{4.14}$$

One can easily obtain

$$\frac{\beta}{2}\left\|A\mathbf{x}^k + B\mathbf{y}^k - b\right\|^2 = \frac{1}{2s^2\beta}\left\|\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right\|^2. \tag{4.15}$$

Furthermore, we use the inequality $(a+b)^2 \le 2(a^2+b^2)$ to know that

$$\begin{aligned}\frac{\beta}{2}\left\|A\overline{\mathbf{x}}^k + B\mathbf{y}^k - b\right\|^2 &= \frac{\beta}{2}\left\|\frac{1}{s\beta}\left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right) + A\left(\overline{\mathbf{x}}^k - \mathbf{x}^k\right)\right\|^2 \\ &\le \frac{1}{s^2\beta}\left\|\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right\|^2 + \beta P_A\left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2. \end{aligned}\tag{4.16}$$

By Lemma 2 and the optimal value condition for the **x**-subproblem, we have

$$\begin{aligned}\left\|\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right\|^2 &\le \frac{1}{\sigma_A}\left\|A^\top\left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right)\right\|^2 = \frac{1}{\sigma_A}\left\|\nabla f\left(\mathbf{x}^{k-1}\right) - \nabla f\left(\mathbf{x}^k\right)\right\|^2 \\ &\le \frac{L^2}{\sigma_A}\left\|\mathbf{x}^{k-1} - \mathbf{x}^k\right\|^2 \le \frac{2L^2}{\sigma_A}\left(\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2 + \left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2\right), \end{aligned}\tag{4.17}$$

where $\left\|\mathbf{x}^{k-1} - \mathbf{x}^k\right\|^2 \le 2\left(\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2 + \left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2\right)$.

Substituting (4.14)–(4.17) into (4.13), one has

$$\mathcal{L}_\beta\left(\overline{\mathbf{x}}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) \le \frac{L^2}{s^2\beta\sigma_A}\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2 + \left(\beta P_A + \frac{L^2}{s^2\beta\sigma_A}\right)\left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2. \tag{4.18}$$

From equation

$$\frac{1}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_\lambda^k\right\|^2\right] \le \frac{1+\tau}{s\beta\sigma_A}\mathbb{E}\left[\left\|A\mathbf{d}_\lambda^k\right\|^2\right] - \frac{\tau}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_\lambda^k\right\|^2\right]$$

with $\tau \in (0,1)$, we take the expected values of (4.10)–(4.12) and (4.18) and add them to get

$$\begin{aligned}\mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}\right) &\le \mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) + \frac{1}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_\lambda^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \\ &\quad + \frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M} + \frac{L^2}{s^2\beta\sigma_A}\mathbb{E}\left[\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2\right] + \left(\beta P_A + \frac{L^2}{s^2\beta\sigma_A}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2\right] \\ &\le \mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) + \frac{1+\tau}{s\beta\sigma_A}\mathbb{E}\left[\left\|A\mathbf{d}_\lambda^k\right\|^2\right] - \frac{\tau}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_\lambda^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \\ &\quad + \frac{L^2}{s^2\beta\sigma_A}\mathbb{E}\left[\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2\right] + \left(\beta P_A + \frac{L^2}{s^2\beta\sigma_A}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2\right] + \frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M}. \end{aligned}\tag{4.19}$$

By applying Lemma 3 to (4.19), we get

$$\begin{aligned}\mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}\right) &\le \mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - \frac{\tau}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_\lambda^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^k\right\|^2\right] - \frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \\ &\quad + \frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M} + \frac{L^2}{s^2\beta\sigma_A}\mathbb{E}\left[\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2\right] + \left(\beta P_A + \frac{L^2}{s^2\beta\sigma_A}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}^k - \mathbf{x}^k\right\|^2\right] \\ &\quad + 4\frac{1+\tau}{s\beta\sigma_A}H_1(s)L_{\psi_2}^2\delta^2\left(\mathbb{E}\left[\left\|\mathbf{d}_\mathbf{x}^{k-1}\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_\mathbf{y}^{k-1}\right\|^2\right]\right) \end{aligned}$$

$$+\frac{1+\tau}{s\beta\sigma_A}H_1(s)\left(2L_f^2+4L_{\psi_2}^2\delta^2\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right]$$
$$+\frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left[\left\|A^\top\mathbf{d}_{\lambda}^{k-1}\right\|^2-\left\|A^\top\mathbf{d}_{\lambda}^{k}\right\|^2\right]+8\frac{1+\tau}{s\beta\sigma_A}H_1(s)\frac{L_{\psi_2}^2\delta^2\sigma^2}{M}.$$

Now, we define

$$\mathcal{L}_\beta(k):=\mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right),\quad \widehat{B}:=\frac{1+\tau}{s\beta\sigma_A}H_1(s)\left(2L_f^2+4L_{\psi_2}^2\delta^2\right),\quad \widehat{A}:=4\frac{1+\tau}{s\beta\sigma_A}H_1(s)L_{\psi_2}^2\delta^2.$$

Thus, we obtain

$$\begin{aligned}
&\mathcal{L}_\beta(k+1)\\
&\le \mathcal{L}_\beta(k)+\widehat{A}\left(\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2\right]-\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right]\right)+\left(\widehat{B}+\widehat{A}\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right]+\widehat{A}\left(\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2\right]-\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right]\right)\\
&\quad+\left(2\widehat{A}-\frac{\delta}{2}\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right]-\frac{\tau}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^{k}\right\|^2\right]+\frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left[\left\|A^\top\mathbf{d}_{\lambda}^{k-1}\right\|^2-\left\|A^\top\mathbf{d}_{\lambda}^{k}\right\|^2\right]\\
&\quad-\frac{\delta}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\overline{\mathbf{x}}^k\right\|^2\right]+\frac{L^2}{s^2\beta\sigma_A}\mathbb{E}\left[\left\|\mathbf{x}^{k-1}-\overline{\mathbf{x}}^k\right\|^2\right]+\left(\beta P_A+\frac{L^2}{s^2\beta\sigma_A}\right)\theta^2\mathbb{E}\left[\left\|\mathbf{x}^{k}-\overline{\mathbf{x}}^{k-1}\right\|^2\right]\\
&\quad+\frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M}+8\frac{1+\tau}{s\beta\sigma_A}H_1(s)\frac{L_{\psi_2}^2\delta^2\sigma^2}{M}.
\end{aligned}\tag{4.20}$$

Recalling the definition of the potential function $\widehat{\mathcal{L}}^k$ in (4.9) and substituting it into (4.20), one has

$$\begin{aligned}
\mathbb{E}\widehat{\mathcal{L}}^{k+1}&\le\mathbb{E}\widehat{\mathcal{L}}^k+\left(\widehat{B}+\widehat{A}\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right]+\left(2\widehat{A}-\frac{\delta}{2}\right)\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right]\\
&\quad+\left[\left(\beta P_A+\frac{L^2}{s^2\beta\sigma_A}\right)\theta^2-\frac{\delta}{2}\right]\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\overline{\mathbf{x}}^k\right\|^2\right]+\frac{L^2}{s^2\beta\sigma_A}\mathbb{E}\left[\left\|\mathbf{x}^{k-1}-\overline{\mathbf{x}}^k\right\|^2\right]\\
&\quad-\frac{\tau}{s\beta}\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^{k}\right\|^2\right]+\frac{(\zeta_{\psi_2}\delta)^2}{2}\frac{\sigma^2}{M}+8\frac{1+\tau}{s\beta\sigma_A}H_1(s)\frac{L_{\psi_2}^2\delta^2\sigma^2}{M}.
\end{aligned}\tag{4.21}$$

Letting $\mu=\min\left\{w,\frac{\tau}{s\beta}\right\}$ and $\eta=\frac{\zeta_{\psi_2}^2\delta^2}{2}+\frac{8(1+\tau)H_1(s)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}$, then by (4.8), one further finds that

$$\begin{aligned}
&\mu\left\{\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\overline{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{x}^{k-1}-\overline{\mathbf{x}}^k\right\|^2\right]\right\}\\
&\le\mathbb{E}\widehat{\mathcal{L}}^k-\mathbb{E}\widehat{\mathcal{L}}^{k+1}+\left(\frac{(\zeta_{\psi_2}\delta)^2}{2}+\frac{8(1+\tau)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}H_1(s)\right)\frac{\sigma^2}{M},
\end{aligned}\tag{4.22}$$

and

$$\begin{aligned}
&\mu\sum_{k=0}^{\mathbb{K}}\left\{\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\overline{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{x}^{k-1}-\overline{\mathbf{x}}^k\right\|^2\right]\right\}\\
&\le\mathbb{E}\widehat{\mathcal{L}}^0-\mathbb{E}\widehat{\mathcal{L}}^{k+1}+(\mathbb{K}+1)\frac{\sigma^2}{M}\eta.
\end{aligned}\tag{4.23}$$

We assume that the batch size M is linearly related to the number of iterations $\mathbb{K}$, i.e., $M = O(\mathbb{K}+1)$. It follows from (4.23) that:

$$\begin{aligned}
&\mu \sum_{k=0}^{\mathbb{K}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \right\} \\
&\leq \mu \sum_{k=0}^{\mathbb{K}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{x}^{k-1} - \overline{\mathbf{x}}^k\right\|^2\right] \right\} \\
&\leq \mathbb{E}\widehat{\mathcal{L}}^0 - \mathbb{E}\widehat{\mathcal{L}}^{k+1} + O\left(\sigma^2\eta\right) \\
&\leq \mathbb{E}\widehat{\mathcal{L}}^0 - \widehat{\mathcal{L}}^\star + O\left(\sigma^2\eta\right),
\end{aligned}$$

where $\widehat{\mathcal{L}}^\star$ is the lower bound of $\widehat{\mathcal{L}}^k$, and

$$\begin{aligned}
&\sum_{k=0}^{\mathbb{K}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \right\} < +\infty, \\
&\min_{k\in\{0,\ldots,\mathbb{K}\}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] \right\} \\
&\leq \min_{k\in\{0,\ldots,\mathbb{K}\}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \right\} \\
&\leq \frac{\mathbb{E}\widehat{\mathcal{L}}^0 - \widehat{\mathcal{L}}^{k+1}}{\mu(\mathbb{K}+1)} + \frac{\eta}{\mu} O\left(\frac{\sigma^2}{M}\right).
\end{aligned}$$

In summary, when $\mathbb{K}$ is large enough, we have

$$\min_{k\in\{0,\ldots,\mathbb{K}\}} \left\{ \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^k\right\|^2\right] \right\} = O\left(\frac{1}{\mathbb{K}}\right),$$

and

$$\lim_{k\to\infty} \mathbb{E}\left\|\mathbf{d}_{\mathbf{x}}^k\right\| = 0, \quad \lim_{k\to\infty} \mathbb{E}\left\|\mathbf{d}_{\mathbf{y}}^k\right\| = 0, \quad \lim_{k\to\infty} \mathbb{E}\left\|\mathbf{d}_{\lambda}^k\right\| = 0, \quad \lim_{k\to\infty} \mathbb{E}\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\| = 0. \tag{4.24}$$

Additionally, according to the update rule of $\boldsymbol{\lambda}^k$ in subsection 3.3, which is $\mathbf{r}^{k+1} = -\dfrac{\mathbf{d}_{\lambda}^k}{s\beta}$, one gets

$$\lim_{k\to\infty} \mathbb{E}\left\|\mathbf{r}^k\right\| = 0. \tag{4.25}$$

□

Theorem 1. *Assuming that the conditions and assumptions in* Lemma 4 *hold and the total number of iterations is denoted by* $\mathbb{K}$*, we have*

$$\min_{1\leq k\leq\mathbb{K}} \mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)\right)^2\right] = O\left(1/\mathbb{K}\right).$$

Proof. First of all, we define

$$\begin{cases} \varepsilon_1^k := A^\top\left(\lambda^{k-1}-\lambda^k\right)+\left(2\widehat{A}-L_{\psi_2}\right)\left(\mathbf{x}^k-\mathbf{x}^{k-1}\right)+2\widehat{C}\left(\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right),\\ \varepsilon_2^k := B^\top\left(\lambda^{k-1}-\lambda^k\right)-\beta B^\top A\theta\left(\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right)+\left(2\widehat{A}-L_{\psi_2}\right)\left(\mathbf{y}^k-\mathbf{y}^{k-1}\right),\\ \varepsilon_3^k := \left(\frac{1}{s\beta}+\frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\right)\left(\lambda^{k-1}-\lambda^k\right),\\ \varepsilon_4^k := 2\widehat{C}\left(\overline{\mathbf{x}}^{k-1}-\mathbf{x}^k\right),\\ \varepsilon_5^k := 2\widehat{A}\left(\mathbf{y}^{k-1}-\mathbf{y}^k\right),\\ \varepsilon_6^k := \frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\left(\lambda^{k-1}-\lambda^k\right). \end{cases}$$

By the definition of the potential energy function $\widehat{\mathcal{L}}^k$, and noting that $\widehat{C}=\left(\frac{L^2}{s^2\beta\sigma_A}+\beta P_A\right)\theta^2$, we have

$$\begin{cases} \partial_{\mathbf{x}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=\nabla f\left(\mathbf{x}^k\right)-A^\top\lambda^k+\beta A^\top\left(A\mathbf{x}^k+B\mathbf{y}^k-b\right)\\ \qquad\qquad +2\widehat{A}\left(\mathbf{x}^k-\mathbf{x}^{k-1}\right)+2\widehat{C}\left(\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right),\\ \partial_{\mathbf{y}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=\partial g\left(\mathbf{y}^k\right)-B^\top\lambda^k+\beta B^\top\left(A\mathbf{x}^k+B\mathbf{y}^k-b\right)\\ \qquad\qquad +2\widehat{A}\left(\mathbf{y}^k-\mathbf{y}^{k-1}\right),\\ \partial_{\lambda}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=\left(\frac{1}{s\beta}+\frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\right)\left(\lambda^{k-1}-\lambda^k\right),\\ \partial_{\bar{\mathbf{x}}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=2\widehat{C}\left(\overline{\mathbf{x}}^{k-1}-\mathbf{x}^k\right),\\ \partial_{\bar{\mathbf{y}}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=2\widehat{A}\left(\mathbf{y}^{k-1}-\mathbf{y}^k\right),\\ \partial_{\bar{\lambda}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)=\frac{1+\tau}{s\beta\sigma_A}2H_2(s)\left\|A^\top\left(\lambda^{k-1}-\lambda^k\right)\right\|\le\frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\left(\lambda^{k-1}-\lambda^k\right). \end{cases}$$

Invoking the optimality condition, one gets

$$\begin{cases} A^\top\left(\lambda^{k-1}-\lambda^k\right)+\left(2\widehat{A}-L_{\psi_2}\right)\left(\mathbf{x}^k-\mathbf{x}^{k-1}\right)+2\widehat{C}\left(\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right)\in\partial_{\mathbf{x}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right),\\ B^\top\left(\lambda^{k-1}-\lambda^k\right)-\beta B^\top A\theta\left(\mathbf{x}^k-\overline{\mathbf{x}}^{k-1}\right)+\left(2\widehat{A}-L_{\psi_2}\right)\left(\mathbf{y}^k-\mathbf{y}^{k-1}\right)\in\partial_{\mathbf{y}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right),\\ \left(\frac{1}{s\beta}+\frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\right)\left(\lambda^{k-1}-\lambda^k\right)=\partial_{\lambda}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right),\\ 2\widehat{C}\left(\overline{\mathbf{x}}^{k-1}-\mathbf{x}^k\right)=\partial_{\bar{\mathbf{x}}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right),\\ 2\widehat{A}\left(\mathbf{y}^{k-1}-\mathbf{y}^k\right)=\partial_{\bar{\mathbf{y}}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right),\\ \frac{1+\tau}{s\beta\sigma_A}2H_2(s)P_A\left(\lambda^{k-1}-\lambda^k\right)\in\partial_{\bar{\lambda}}\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right). \end{cases}$$

Thus, we are certain that $\left(\varepsilon_1^{k+1},\varepsilon_2^{k+1},\varepsilon_3^{k+1},\varepsilon_4^{k+1},\varepsilon_5^{k+1},\varepsilon_6^{k+1}\right)^\top\in\partial\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)$ holds. Moreover, according to (4.24), there is a real number ζ_0 such that

$$d\left(0,\partial\widehat{\mathcal{L}}_\beta\left(\overline{\mathbf{w}}^k\right)\right)\le\left\|\left(\varepsilon_1^{k+1},\varepsilon_2^{k+1},\varepsilon_3^{k+1},\varepsilon_4^{k+1},\varepsilon_5^{k+1},\varepsilon_6^{k+1}\right)\right\|$$

$$\leq \zeta_0 \left(\left\| \mathbf{x}^k - \mathbf{x}^{k-1} \right\| + \left\| \mathbf{x}^k - \overline{\mathbf{x}}^{k-1} \right\| + \left\| \mathbf{y}^k - \mathbf{y}^{k-1} \right\| + \left\| \boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k \right\| \right). \tag{4.26}$$

Taking expectations on both sides of equation (4.26), we have

$$\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial \widehat{\mathcal{L}}_\beta \left(\overline{\mathbf{w}}^k \right) \right) \right] \leq \zeta_0 \left(\mathbb{E}\left[\left\| \mathbf{d}_{\mathbf{y}}^{k-1} \right\| \right] + \mathbb{E}\left[\left\| \mathbf{d}_{\mathbf{x}}^{k-1} \right\| \right] + \mathbb{E}\left[\left\| \mathbf{d}_{\lambda}^{k-1} \right\| \right] + \mathbb{E}\left[\left\| \mathbf{x}^k - \overline{\mathbf{x}}^{k-1} \right\| \right] \right).$$

From (4.24), it follows that:

$$\lim_{k \to \infty} \mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial \widehat{\mathcal{L}}_\beta \left(\overline{\mathbf{w}}^k \right) \right)^2 \right] = 0, \tag{4.27}$$

and using $M = O(\mathbb{K} + 1)$, a sufficiently large $\mathbb{K}$, one gets

$$\min_{1 \leq k \leq \mathbb{K}} \mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial \widehat{\mathcal{L}}_\beta \left(\overline{\mathbf{w}}^k \right) \right)^2 \right] = O(1/\mathbb{K}).$$

□

4.3. Linear convergence rate

In this section, the local linear convergence of the iterative sequence $\left\{ \mathbf{w}^k \right\}$ and the potential energy function $\left\{ \mathbb{E}\left\{ \widehat{\mathcal{L}}^k \right\} \right\}$ will be established under specific assumptions.

Assumption 2. (a) (Error-bound condition) *For any $\xi \geq \inf_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w})$, there are $\varsigma > 0$ and $\tau > 0$ such that the following holds:*

$$\operatorname{dist}(\mathbf{w}, \Omega^*) \leq \tau \operatorname{dist}(\mathbf{0}, \partial \mathcal{L}_\beta(\mathbf{w})),$$

whenever $\operatorname{dist}(\mathbf{0}, \partial \mathcal{L}_\beta(\mathbf{w})) \leq \varsigma$ *and* $\mathcal{L}_\beta(\mathbf{w}) \leq \xi$.
(b) *For any $\xi \geq \inf_{\overline{\mathbf{w}}} \widehat{\mathcal{L}}_\beta(\overline{\mathbf{w}})$, there are $\varsigma, \tau > 0$ such that* $\operatorname{dist}\left(\overline{\mathbf{w}}, \widehat{\Omega} \right) \leq \tau \operatorname{dist}\left(\mathbf{0}, \partial \widehat{\mathcal{L}}_\beta(\overline{\mathbf{w}}) \right)$, *whenever* $\operatorname{dist}\left(\mathbf{0}, \partial \widehat{\mathcal{L}}_\beta(\overline{\mathbf{w}}) \right) \leq \varsigma$ *and* $\widehat{\mathcal{L}}_\beta(\overline{\mathbf{w}}) \leq \xi$.
(c) *The set Ω^* is nonempty, and there exists a positive constant ω^* such that $\|\mathbf{w}_1 - \mathbf{w}_2\| \geq \omega^*$ whenever $\mathbf{w}_1, \mathbf{w}_2 \in \Omega^*$ and $f(\mathbf{x}_1) + g(\mathbf{y}_1) \neq f(\mathbf{x}_2) + g(\mathbf{y}_2)$ for every $\mathbf{x}_i \in \mathbb{R}^{n_a}$ and each $\mathbf{y}_i \in \mathbb{R}^{n_b}$ $(i = 1, 2)$.*
(d) *The function g shows local weak convexity near*

$$\Omega_{\mathbf{y}}^* := \{ \mathbf{y} \in \mathbb{R}^{n_b} | \text{ there exist } \mathbf{x} \text{ and } \boldsymbol{\lambda} \text{ such that } (\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \in \Omega^* \},$$

which implies that there are $\varsigma, \delta, \sigma > 0$. For all $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{n_b}$ with $\operatorname{dist}(\mathbf{y}_1, \Omega_{\mathbf{y}}^*) \leq \varsigma$, $\operatorname{dist}(\mathbf{y}_2, \Omega_{\mathbf{y}}^*) \leq \varsigma$, *and $\|\mathbf{y}_1 - \mathbf{y}_2\| \leq \delta$, and for all $\varpi \in \partial g(\mathbf{y}_2)$, the following holds:*

$$g(\mathbf{y}_1) \geq g(\mathbf{y}_2) + \langle \varpi, \mathbf{y}_1 - \mathbf{y}_2 \rangle - \sigma \|\mathbf{y}_1 - \mathbf{y}_2\|^2.$$

Lemma 5. *Let $S(\mathbf{w}_0)$ be the set of limit points of the iterates $\left\{ \mathbf{w}^k := \left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k \right) \right\}$. Suppose that* Assumption 1 *and the conditions in* Theorem 1 *are satisfied. Then there is an $\mathcal{F}^*$ such that* (i) *when M is very large or $\sigma^2 \to 0$* (using a VR-gradient estimator)*, the sequence $\left\{ \mathbb{E}\left[\widehat{\mathcal{L}}^k \right] \right\}$ does not increase and*

$$\lim_{k \to \infty} \mathbb{E}\left[\widehat{\mathcal{L}}^k \right] = \lim_{k \to \infty} \mathbb{E}\left[\mathcal{L}_\beta(\mathbf{w}^k) \right] = \mathcal{F}^*; \tag{4.28}$$

(ii) $S(\mathbf{w}_0) \subseteq \operatorname{crit} \mathcal{L}_\beta$ *almost surely* (a.s.).

Proof. (i) By combining a sufficiently large M or $\sigma^2 \to 0$ with (4.8), (4.21), and $\mu = \min\left\{w, \frac{\tau}{s\beta}\right\}$, we obtain

$$\mathbb{E}\widehat{\mathcal{L}}^{k+1} \leq \mathbb{E}\widehat{\mathcal{L}}^{k} - \mu\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{x}}^{k}\right\|^2\right] - \mu\mathbb{E}\left[\left\|\mathbf{d}_{\mathbf{y}}^{k}\right\|^2\right] - \mu\mathbb{E}\left[\left\|\mathbf{d}_{\lambda}^{k}\right\|^2\right] - \mu\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^{k}\right\|^2\right],$$

and so, the sequence $\mathbb{E}\widehat{\mathcal{L}}^{k}$ is monotonically nonincreasing. Since $\left\{\widehat{\mathcal{L}}^{k}\right\}$ is bounded from below, one has $\lim_{k\to\infty}\mathbb{E}\left[\widehat{\mathcal{L}}_{\beta}(\overline{\mathbf{w}}^{k})\right] = \mathcal{F}^{*}$ for some $\mathcal{F}^{*}$. Then, from the definition of $\widehat{\mathcal{L}}^{k}$ in (4.9), it follows that (4.24) and (4.25) hold. Thus, we have

$$\lim_{k\to\infty}\mathbb{E}\left[\widehat{\mathcal{L}}^{k}\right] = \lim_{k\to\infty}\mathbb{E}\left[\mathcal{L}_{\beta}(\mathbf{w}^{k})\right] = \mathcal{F}^{*}.$$

(ii) For each $\widetilde{\mathbf{w}} \in S(\mathbf{w}_0)$, it can be proved that $\widetilde{\mathbf{w}} \in \operatorname{crit}\mathcal{L}_{\beta}$. Given $\widetilde{\mathbf{w}} \in S(\mathbf{w}_0)$ by Definition 2, we have $\mathbf{w}^{k_q} \to \widetilde{\mathbf{w}}$, $d^{k_q} \in \partial\mathcal{L}_{\beta}\left(\mathbf{w}^{k_q}\right)$, and $d^{k_q} \to 0$ a.s. Noting the outer semicontinuity of the Clarke subgradient $\partial\mathcal{L}_{\beta}\left(\mathbf{w}^{k_q}\right)$, it implies that $0 \in \partial\mathcal{L}_{\beta}(\widetilde{\mathbf{w}})$. Therefore, we show that $\widetilde{\mathbf{w}} \in \operatorname{crit}\mathcal{L}_{\beta}$ for any $\widetilde{\mathbf{w}} \in S(\mathbf{w}_0)$, which is equivalent to $S(\mathbf{w}_0) \subseteq \operatorname{crit}\mathcal{L}_{\beta}$ a.s. □

Theorem 2. *Assume that the conditions in* Lemma 4 *and* Assumption 2 *are satisfied. Let* $\left\{\mathbf{w}^{k} := \left(\mathbf{x}^{k}, \mathbf{y}^{k}, \lambda^{k}\right)\right\}$ *be the iterations produced by* Algorithm 1, Ω^{*} *be the cluster point set of the sequence* $\left\{\mathbf{w}^{k}\right\}$, *and* $\widehat{\Omega}$ *denote the cluster point set of the sequence* $\left\{\overline{\mathbf{w}}^{k}\right\}$. *Then the following assertions hold:*

(i) $\lim_{k\to\infty}\operatorname{dist}\left(\overline{\mathbf{w}}^{k}, \widehat{\Omega}\right) = 0$, $\lim_{k\to\infty}\operatorname{dist}\left(\mathbf{w}^{k}, \Omega^{*}\right) = 0$ *a.s.*

(ii) *There exist constants* $\widetilde{C} \in (0,1)$ *and* $\check{C} > 0$ *such that*

$$\mathbb{E}\widehat{\mathcal{L}}^{k} - F^{*} \leq \left(\widetilde{C}\right)^{k}\left(\mathbb{E}\widehat{\mathcal{L}}^{0} - F^{*}\right) + \check{C}\frac{\sigma^2}{M} \quad \text{a. s.} \tag{4.29}$$

If $\dfrac{\sigma^2}{M} = 0$ (the noiseless case), *then* (4.29) *simplifies to*

$$\mathbb{E}\widehat{\mathcal{L}}^{k} - F^{*} \leq \left(\widetilde{C}\right)^{k}\left(\mathbb{E}\widehat{\mathcal{L}}^{0} - F^{*}\right) \quad \text{a. s.},$$

which shows that $\mathbb{E}\widehat{\mathcal{L}}^{k}$ *a.s. converges to* F^{*} *at a linear rate as the iteration number* k *approaches infinity.*

Proof. (i) From (4.27) and (4.28), we know that there is a constant $\zeta \geq \inf_{\overline{\mathbf{w}}}\widehat{\mathcal{L}}_{\beta}(\overline{\mathbf{w}})$ such that $\mathbb{E}\widehat{\mathcal{L}}_{\beta}\left(\overline{\mathbf{w}}^{k}\right) \leq \zeta$ for every k. By Assumption 2(b), when $\xi = \zeta$, we obtain

$$\operatorname{dist}\left(\overline{\mathbf{w}}^{k}, \widehat{\Omega}\right) \leq \tau\operatorname{dist}\left(\mathbf{0}, \partial\widehat{\mathcal{L}}_{\beta}\left(\overline{\mathbf{w}}^{k}\right)\right)$$

and

$$\lim_{k\to\infty}\operatorname{dist}\left(\mathbf{0}, \partial\overline{\mathbf{w}}_{\beta}\left(\overline{\mathbf{w}}^{k}\right)\right) = 0.$$

Thus, we have $\lim_{k\to\infty}\operatorname{dist}\left(\overline{\mathbf{w}}^{k}, \widehat{\Omega}\right) = 0$ a.s. Similarly, it follows from Assumption 2(a) that $\lim_{k\to\infty}\operatorname{dist}\left(\mathbf{w}^{k}, \Omega^{*}\right) = 0$ is a.s. clearly established.

(ii) Let us define $\widetilde{\mathbf{w}}^k \in \Omega^*$ such that for any iterate $\mathbf{w}^k$, $\text{dist}\left(\mathbf{w}^k, \Omega^*\right) = \left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|$. Due to the closedness of the set Ω^*, such a $\widetilde{\mathbf{w}}^k$ must exist. Then, based on the conclusion of (i), we get

$$\lim_{k\to\infty} \left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\| = 0 \quad \text{a. s.} \tag{4.30}$$

In addition, by (4.24) and $\left\|\mathbf{w}^k - \mathbf{w}^{k-1}\right\| \leq \left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\| + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\| + \left\|\mathbf{d}_{\lambda}^{k-1}\right\|$, one has

$$\lim_{k\to\infty} \left\|\mathbf{w}^k - \mathbf{w}^{k-1}\right\| = 0 \quad \text{a. s.} \tag{4.31}$$

Hence, from $\left\|\widetilde{\mathbf{w}}^k - \widetilde{\mathbf{w}}^{k-1}\right\| \leq \left\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\right\| + \left\|\mathbf{w}^k - \mathbf{w}^{k-1}\right\| + \left\|\mathbf{w}^{k-1} - \widetilde{\mathbf{w}}^{k-1}\right\|$, (4.30), and (4.31), it further implies that

$$\lim_{k\to\infty} \left\|\widetilde{\mathbf{w}}^k - \widetilde{\mathbf{w}}^{k-1}\right\| = 0 \quad \text{a. s.}$$

Combining with Assumption 2(c) and $\widetilde{\mathbf{w}}^k \in \Omega$, there exists a constant $\widetilde{F}^*$ such that for all k large enough,

$$\mathcal{L}_\beta\left(\widetilde{\mathbf{w}}^k\right) = \mathcal{L}_\beta\left(\widetilde{\mathbf{x}}^k, \widetilde{\mathbf{y}}^k, \widetilde{\lambda}^k\right) = F\left(\widetilde{\mathbf{x}}^k, \widetilde{\mathbf{y}}^k\right) = \widetilde{F}^* \quad \text{a. s.} \tag{4.32}$$

By Assumption 2, the sequence $\left\{\mathbf{w}^k\right\}$ has a cluster point denoted as $\mathbf{w}^*$. That is, there is a subsequence $\left\{\mathbf{w}^{k_i}\right\}$ converging to $\mathbf{w}^*$. Thus, from Lemma 5, we infer that $\mathbf{w}^* \in \Omega^*$. Moreover, according to (4.30), one gets

$$\lim_{i\to\infty} \left\|\widetilde{\mathbf{w}}^{k_i} - \mathbf{w}^*\right\| \leq \lim_{i\to\infty} \left(\left\|\widetilde{\mathbf{w}}^{k_i} - \mathbf{w}^{k_i}\right\| + \left\|\mathbf{w}^{k_i} - \mathbf{w}^*\right\|\right) = 0 \quad \text{a. s.}$$

Again, from (4.32), $\mathbf{w}^* \in \Omega^*$, and Assumption 2(c), it follows that:

$$\mathcal{L}_\beta\left(\mathbf{w}^*\right) = \widetilde{F}^* \quad \text{a. s.}$$

Then, due to the lower semicontinuity of the function $\mathcal{L}_\beta(\cdot)$, we have

$$\widetilde{F}^* = \mathcal{L}_\beta\left(\mathbf{w}^*\right) \leq \lim_{i\to\infty} \mathcal{L}_\beta\left(\mathbf{w}^{k_i}\right) = F^* \quad \text{a. s.}, \tag{4.33}$$

where $F^* = \lim_{k\to\infty} \widehat{\mathcal{L}}^k = \lim_{k\to\infty} \mathcal{L}_\beta\left(\mathbf{w}^k\right)$ a. s., as given in (4.28).

Based on the definition of $\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$, the update rules of $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\lambda}$, along with some calculations, we get

$$\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}\right) = \frac{1}{s\beta}\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^k\right)^\top \left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right), \tag{4.34}$$

$$\begin{aligned}
&\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}\right) - \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}\right) \\
&= g\left(\mathbf{y}^k\right) - g\left(\mathbf{y}\right) + \boldsymbol{\lambda}^\top B\left(\mathbf{y} - \mathbf{y}^k\right) + \frac{\beta}{2}\left(\left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2 - \left\|A\mathbf{x}^k + B\mathbf{y} - \mathbf{b}\right\|^2\right),
\end{aligned} \tag{4.35}$$

and

$$\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}\right) - \mathcal{L}_\beta\left(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}\right)$$

$$= f\left(\mathbf{x}^k\right) - f\left(\mathbf{x}\right) + \boldsymbol{\lambda}^\top A\left(\mathbf{x} - \mathbf{x}^k\right) + \frac{\beta}{2}\left(\left\|A\mathbf{x}^k + B\mathbf{y} - \mathbf{b}\right\|^2 - \|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2\right). \tag{4.36}$$

Summing up (4.34)–(4.36), and for all k large enough, based on (4.32) and (4.33), one gets

$$\begin{aligned}
\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - F^* &\leq \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - \widetilde{F}^* = \mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - \mathcal{L}_\beta\left(\widetilde{\mathbf{x}}^k, \widetilde{\mathbf{y}}^k, \widetilde{\boldsymbol{\lambda}}^k\right) \\
&= \frac{1}{s\beta}\left(\widetilde{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\right)^\top\left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right) + f\left(\mathbf{x}^k\right) - f\left(\widetilde{\mathbf{x}}^k\right) + \left\langle A^\top\widetilde{\boldsymbol{\lambda}}^k, \widetilde{\mathbf{x}}^k - \mathbf{x}^k\right\rangle \\
&\quad + g\left(\mathbf{y}^k\right) - g\left(\widetilde{\mathbf{y}}^k\right) + \left\langle B^\top\widetilde{\boldsymbol{\lambda}}^k, \widetilde{\mathbf{y}}^k - \mathbf{y}^k\right\rangle + \frac{\beta}{2}\left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2 \\
&= \frac{1}{s\beta}\left(\widetilde{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\right)^\top\left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right) + f\left(\mathbf{x}^k\right) - f\left(\widetilde{\mathbf{x}}^k\right) + \left\langle A^\top\widetilde{\boldsymbol{\lambda}}^k, \widetilde{\mathbf{x}}^k - \mathbf{x}^k\right\rangle \\
&\quad + g\left(\mathbf{y}^k\right) - g\left(\widetilde{\mathbf{y}}^k\right) + \left\langle B^\top\widetilde{\boldsymbol{\lambda}}^k, \widetilde{\mathbf{y}}^k - \mathbf{y}^k\right\rangle + \frac{\beta}{2}\left\|\frac{\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}}{s\beta}\right\|^2 \quad \text{a. s.},
\end{aligned} \tag{4.37}$$

where the first equation is due to $\widetilde{\mathbf{w}}^k \in \Omega^* : A\widetilde{\mathbf{x}}^k + B\widetilde{\mathbf{y}}^k = \mathbf{b}$, and the last one is from $\mathbf{d}_{\boldsymbol{\lambda}}^{k-1} = -s\beta\mathbf{r}^k$.

From Assumption 1(a) and $A^\top\widetilde{\boldsymbol{\lambda}}^k = \nabla f\left(\widetilde{\mathbf{x}}^k\right)$, we can derive

$$f\left(\mathbf{x}^k\right) - f\left(\widetilde{\mathbf{x}}^k\right) + \left\langle A^\top\widetilde{\boldsymbol{\lambda}}^k, \widetilde{\mathbf{x}}^k - \mathbf{x}^k\right\rangle \leq \frac{L}{2}\left\|\widetilde{\mathbf{x}}^k - \mathbf{x}^k\right\|^2. \tag{4.38}$$

Remembering (3.3), there exists a $\xi_{\mathbf{y}}^k \in \partial_{\mathbf{y}}\mathcal{L}_\beta\left(\overline{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1}\right)$, i.e.,

$$\boldsymbol{\varpi}^k := \xi_{\mathbf{y}}^k + B^\top\boldsymbol{\lambda}^{k-1} - \beta B^\top\left(A\overline{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b}\right) \in \partial g\left(\mathbf{y}^k\right),$$

with $\left\|\xi_{\mathbf{y}}^{k+1}\right\| \leq L_{\psi_1}\left\|\mathbf{y}^k - \mathbf{y}^{k-1}\right\| = L_{\psi_1}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|$ such that

$$\begin{aligned}
\left\|\boldsymbol{\varpi}^k - B^\top\widetilde{\boldsymbol{\lambda}}^k\right\| &= \left\|\xi_{\mathbf{y}}^k + B^\top\boldsymbol{\lambda}^{k-1} - \beta B^\top\left(A\overline{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b}\right) - B^\top\widetilde{\boldsymbol{\lambda}}^k\right\| \\
&\leq \left\|\xi_{\mathbf{y}}^k + B^\top\boldsymbol{\lambda}^{k-1} - B^\top\widetilde{\boldsymbol{\lambda}}^k\right\| \\
&\leq \left\|\xi_{\mathbf{y}}^k\right\| + \left\|B^\top\left(\boldsymbol{\lambda}^{k-1} - \widetilde{\boldsymbol{\lambda}}^k\right)\right\| \\
&\leq L_{\psi_1}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\| + \|B\|\left(\left\|\mathbf{d}_{\boldsymbol{\lambda}}^{k-1}\right\| + \left\|\boldsymbol{\lambda}^k - \widetilde{\boldsymbol{\lambda}}^k\right\|\right).
\end{aligned}$$

By Assumption 2(d), we can conclude that

$$g\left(\mathbf{y}^k\right) - g\left(\widetilde{\mathbf{y}}^k\right) + \left\langle\boldsymbol{\varpi}^k, \widetilde{\mathbf{y}}^k - \mathbf{y}^k\right\rangle \leq \sigma\left\|\widetilde{\mathbf{y}}^k - \mathbf{y}^k\right\|^2. \tag{4.39}$$

Thus, substituting (4.38) and (4.39) into (4.37), and using the inequality $ab \leq \dfrac{a^2 + b^2}{2}$ for all $a, b \in \mathbb{R}$, we derive

$$\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - F^* \leq \frac{1}{s\beta}\left(\widetilde{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\right)^\top\left(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k\right) + \frac{L}{2}\left\|\widetilde{\mathbf{x}}^k - \mathbf{x}^k\right\|^2$$

$$
\begin{aligned}
&+\frac{1}{2\beta s^2}\left\|\mathbf{d}_{\lambda}^{k-1}\right\|^2+\sigma\left\|\widetilde{\mathbf{y}}^k-\mathbf{y}^k\right\|^2+\left\|\boldsymbol{\varpi}^k-B^{\top}\widetilde{\boldsymbol{\lambda}}^k\right\|\left\|\widetilde{\mathbf{y}}^k-\mathbf{y}^k\right\| \\
&\leq\left(\frac{1}{2s\beta}+\frac{\|B\|^2}{2}\right)\left\|\widetilde{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\right\|^2+\frac{L}{2}\left\|\widetilde{\mathbf{x}}^k-\mathbf{x}^k\right\|^2+\left(\sigma+\frac{3}{2}\right)\left\|\widetilde{\mathbf{y}}^k-\mathbf{y}^k\right\|^2 \\
&\quad+\left(\frac{1}{2\beta s}+\frac{1}{2\beta s^2}+\frac{\|B\|^2}{2}+\frac{\|B\|^2}{2s^2}\right)\left\|\mathbf{d}_{\lambda}^{k-1}\right\|^2+\frac{L_{\psi_1}^2}{2}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 \\
&\leq\left(\frac{1}{2s\beta}+\frac{\|B\|^2}{2}\right)\left\|\widetilde{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\right\|^2+\frac{L}{2}\left\|\widetilde{\mathbf{x}}^k-\mathbf{x}^k\right\|^2+\left(\sigma+\frac{3}{2}\right)\left\|\widetilde{\mathbf{y}}^k-\mathbf{y}^k\right\|^2 \\
&\quad+\left(\frac{1}{2\beta s}+\frac{1}{2\beta s^2}+\frac{\|B\|^2}{2}+\frac{\|B\|^2}{2s^2}\right)\left\|\mathbf{d}_{\lambda}^{k-1}\right\|^2+\frac{L_{\psi_1}^2}{2}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2+\frac{\beta^2}{2}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 \\
&\leq C_1\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2+\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2+\left\|\mathbf{d}_{\lambda}^{k-1}\right\|^2\right)+C_2\left\|\widetilde{\mathbf{w}}^k-\mathbf{w}^k\right\|^2 \quad \text{a. s.},
\end{aligned}
\tag{4.40}
$$

where

$$
C_1=\max\left\{\frac{1}{2\beta s}+\frac{1}{2\beta s^2}+\frac{\|B\|^2}{2}+\frac{\|B\|^2}{2s^2},\frac{L_{\psi_1}}{2},\frac{\beta^2}{2}\right\},\ C_2=\max\left\{\frac{1}{2s\beta}+\frac{\|B\|^2}{2},\frac{L}{2},\sigma+\frac{3}{2}\right\}.
$$

Using the notation of the iterates produced by Algorithm 1 as $\left\{\mathbf{w}^k:=\left(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k\right)\right\}$ and performing some straightforward calculations, we find that

$$
\begin{cases}
\partial_{\mathbf{x}}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)=\nabla f\left(\mathbf{x}^k\right)-A^{\top}\boldsymbol{\lambda}^k+\beta A^{\top}\mathbf{r}^k \\
\qquad\qquad=\nabla_{\mathbf{x}}\mathcal{L}_{\beta}\left(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^{k-1}\right)-A^{\top}\mathbf{d}_{\lambda}^{k-1}=\xi_{\mathbf{x}}^k-A^{\top}\mathbf{d}_{\lambda}^{k-1}, \\
\partial_{\mathbf{y}}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)=\partial_y g\left(\mathbf{y}^k\right)-B^{\top}\boldsymbol{\lambda}^{k-1}+\beta B^{\top}\left(A\mathbf{x}^k+B\mathbf{y}^k-\mathbf{b}\right)+B^{\top}\boldsymbol{\lambda}^{k-1}-B^{\top}\boldsymbol{\lambda}^k \\
\qquad\qquad=-\beta B^{\top}\left(A\overline{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b}\right)+\nabla\psi_1\left(\mathbf{y}^{k-1}\right)-\nabla\psi_1\left(\mathbf{y}^k\right) \\
\qquad\qquad\quad+\beta B^{\top}\frac{1}{s\beta}\left(\boldsymbol{\lambda}^{k-1}-\boldsymbol{\lambda}^k\right)+B^{\top}\left(\boldsymbol{\lambda}^{k-1}-\boldsymbol{\lambda}^k\right) \\
\qquad\qquad\leq L_{\psi_1}\mathbf{d}_y^{k-1}+\left(B^{\top}+\frac{B^{\top}}{s}\right)\mathbf{d}_{\lambda}^{k-1}, \\
\partial_{\lambda}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)=\partial_{\lambda}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)=-\mathbf{r}^k=-\frac{1}{s\beta}\mathbf{d}_{\lambda}^{k-1}.
\end{cases}
\tag{4.41}
$$

By applying the optimal conditions (3.3) and (3.5), along with (4.24) and (4.25), and considering $M=O(\mathbb{K}+1)$ for a sufficiently large $\mathbb{K}$, we can infer that

$$
\begin{aligned}
\lim_{k\to\infty}\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0},\partial\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)\right)\right]&=\lim_{k\to\infty}\mathbb{E}\begin{pmatrix}\partial_{\mathbf{x}}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)\\ \partial_{\mathbf{y}}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)\\ \partial_{\lambda}\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)\end{pmatrix} \\
&\leq\lim_{k\to\infty}\mathbb{E}\begin{pmatrix}\xi_{\mathbf{x}}^k-A^{\top}\mathbf{d}_{\lambda}^{k-1}\\ L_{\psi_1}\mathbf{d}_{\mathbf{y}}^{k-1}+\left(B^{\top}+\frac{B^{\top}}{s}\right)\mathbf{d}_{\lambda}^{k-1}\\ -\mathbf{r}^k\end{pmatrix}=0,
\end{aligned}
$$

and then we get $\lim_{k\to\infty}\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0},\partial\mathcal{L}_{\beta}\left(\mathbf{w}^k\right)\right)\right]=0$.

From the results in (4.41) and Assumption 2(a) once more, it follows that

$$\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial \mathcal{L}_\beta\left(\mathbf{w}^k\right)\right)\right] \leq \mathbb{E}\left\{\left\|\xi_{\mathbf{x}}^k\right\| + \left(\|A\| + \left\|B + \frac{B^\top}{s}\right\| + \frac{1}{s\beta}\right)\left\|\mathbf{d}_\lambda^{k-1}\right\| + L_{\psi_1}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|\right\},$$

and

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\| &= \mathbb{E}\left[\operatorname{dist}\left(\mathbf{w}^k, \Omega^*\right)\right] \leq \tau\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial \mathcal{L}_\beta\left(\mathbf{w}^k\right)\right)\right] \\ &\leq \tau\mathbb{E}\left\{\left\|\xi_{\mathbf{x}}^k\right\| + \left(\|A\| + \left\|B + \frac{B^\top}{s}\right\| + \frac{1}{s\beta}\right)\left\|\mathbf{d}_\lambda^{k-1}\right\| + L_{\psi_1}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|\right\}. \end{aligned} \tag{4.42}$$

Combining (4.40) with (4.42), one can derive

$$\begin{aligned}&\mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - F^* \leq C_1\mathbb{E}\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left\|\mathbf{d}_\lambda^{k-1}\right\|^2\right) \\ &\quad + 3C_2\tau\mathbb{E}\left[\left\|\xi_{\mathbf{x}}^k\right\|^2 + \left(\|A\| + \left\|B + \frac{B^\top}{s}\right\| + \frac{1}{s\beta}\right)^2\left\|\mathbf{d}_\lambda^{k-1}\right\|^2 + L_{\psi_1}^2\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2\right] \\ &\leq C_3\mathbb{E}\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left\|\mathbf{d}_\lambda^{k-1}\right\|^2\right) + 3C_2\tau(L_{\psi_2}^2\delta^2)\frac{\sigma^2}{M} \quad \text{a. s.}, \end{aligned} \tag{4.43}$$

where the positive constant

$$C_3 = \max\left\{\begin{array}{l} C_1 + 3C_2\tau L_{\psi_2}^2\delta^2, C_1 + 3C_2\tau L_{\psi_1}^2 + 3C_2\tau L_{\psi_1}^2\delta^2, \\ C_1 + 3C_2\tau\left(\|A\| + \left\|B + \frac{B^\top}{s}\right\| + \frac{1}{s\beta}\right)^2 \end{array}\right\}.$$

Adding $\widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left(\frac{L^2}{s^2\beta\sigma_A} + \beta P_A\right)\theta^2\mathbb{E}\left\|\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right\|^2 + \frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left\|A\mathbf{d}_\lambda^{k-1}\right\|^2$ to both sides of (4.43) and recalling the definition of $\widehat{\mathcal{L}}^k$, one gets

$$\begin{aligned}&\mathbb{E}\mathcal{L}_\beta\left(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k\right) - F^* + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 \\ &\quad + \left(\frac{L^2}{s^2\beta\sigma_A} + \beta P_A\right)\theta^2\mathbb{E}\left\|\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right\|^2 + \frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left\|A\mathbf{d}_\lambda^{k-1}\right\|^2 \\ &\leq C_3\mathbb{E}\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left\|\mathbf{d}_\lambda^{k-1}\right\|^2\right) + 3C_2\tau L_{\psi_2}^2\delta^2\frac{\sigma^2}{M} + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 \\ &\quad + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left(\frac{L^2}{s^2\beta\sigma_A} + \beta P_A\right)\theta^2\mathbb{E}\left\|\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right\|^2 + \frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left\|A\mathbf{d}_\lambda^{k-1}\right\|^2 \quad \text{a. s.}, \end{aligned}$$

that is,

$$\begin{aligned}\mathbb{E}\widehat{\mathcal{L}}^k - F^* &\leq C_3\mathbb{E}\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left\|\mathbf{d}_\lambda^{k-1}\right\|^2\right) + 3C_2\tau L_{\psi_2}^2\delta^2\frac{\sigma^2}{M} + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 \\ &\quad + \widehat{A}\mathbb{E}\left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left(\frac{L^2}{s^2\beta\sigma_A} + \beta P_A\right)\theta^2\mathbb{E}\left\|\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right\|^2 + \frac{1+\tau}{s\beta\sigma_A}H_2(s)\mathbb{E}\left\|A\mathbf{d}_\lambda^{k-1}\right\|^2 \\ &\leq \overline{C}\left(\left\|\mathbf{d}_{\mathbf{x}}^{k-1}\right\|^2 + \left\|\mathbf{d}_{\mathbf{y}}^{k-1}\right\|^2 + \left\|\mathbf{d}_\lambda^{k-1}\right\|^2 + \left\|\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right\|^2\right) + 3C_2\tau L_{\psi_2}^2\delta^2\frac{\sigma^2}{M} \quad \text{a. s.}, \end{aligned}$$

where $\overline{C} = \max\left\{C_3 + \widehat{A}, C_3 + \dfrac{1+\tau}{s\beta\sigma_A}H_2(s)\|A\|^2, \left(\dfrac{L^2}{s^2\beta\sigma_A} + \beta P_A\right)\theta^2\right\}$. By reusing (4.22), it follows that

$$\mathbb{E}\widehat{\mathcal{L}}^k - F^* \leq \frac{\overline{C}}{\mu}\left[\mathbb{E}\widehat{\mathcal{L}}^{k-1} - \mathbb{E}\widehat{\mathcal{L}}^k + \left(\frac{(\zeta_{\psi_2}\delta)^2}{2} + 8\frac{(1+\tau)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}H_1(s)\right)\frac{\sigma^2}{M}\right] + 3C_2\tau L_{\psi_2}^2\delta^2\frac{\sigma^2}{M}.$$

The term $\dfrac{\overline{C}}{\mu}\left(\mathbb{E}\widehat{\mathcal{L}}^k - F^*\right)$ is added to both sides of the above inequality. Then one has

$$\left(1 + \frac{\overline{C}}{\mu}\right)\left(\mathbb{E}\widehat{\mathcal{L}}^k - F^*\right) \leq \frac{\overline{C}}{\mu}\left(\mathbb{E}\widehat{\mathcal{L}}^{k-1} - F^*\right) + \left[\frac{\overline{C}}{\mu}\left(\frac{(\zeta_{\psi_2}\delta)^2}{2} + 8\frac{(1+\tau)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}H_1(s)\right) + 3C_2\tau L_{\psi_2}^2\delta^2\right]\frac{\sigma^2}{M},$$

and

$$\mathbb{E}\widehat{\mathcal{L}}^k - F^* \leq \frac{\overline{C}}{\mu + \overline{C}}\left(\mathbb{E}\widehat{\mathcal{L}}^{k-1} - F^*\right) + \widehat{C}\frac{\sigma^2}{M} \quad \text{a. s.},$$

where $\widehat{C} := \dfrac{\left[\frac{\overline{C}}{\mu}\left(\frac{(\zeta_{\psi_2}\delta)^2}{2} + 8\frac{(1+\tau)L_{\psi_2}^2\delta^2}{s\beta\sigma_A}H_1(s)\right) + 3C_2\tau L_{\psi_2}^2\delta^2\right]}{\left(1 + \frac{\overline{C}}{\mu}\right)}$.

Hence, by rearranging the equation, we can further derive

$$\left[\mathbb{E}\widehat{\mathcal{L}}^k - F^* - \widehat{C}\frac{\sigma^2}{M}\left(1 + \frac{\overline{C}}{\mu}\right)\right] \leq \frac{\overline{C}}{\mu + \overline{C}}\left[\mathbb{E}\widehat{\mathcal{L}}^{k-1} - F^* - \widehat{C}\frac{\sigma^2}{M}\left(1 + \frac{\overline{C}}{\mu}\right)\right],$$

$$\mathbb{E}\widehat{\mathcal{L}}^k - F^* \leq \left(\frac{\overline{C}}{\mu + \overline{C}}\right)^k\left(\mathbb{E}\widehat{\mathcal{L}}^0 - F^*\right) + \left(1 - \left(\frac{\overline{C}}{\mu + \overline{C}}\right)^k\right)\widehat{C}\frac{\sigma^2}{M}\left(1 + \frac{\overline{C}}{\mu}\right) \quad \text{a. s.}$$

Defining $\widetilde{C} = \dfrac{\overline{C}}{\mu + \overline{C}}$ and $\breve{C} = \widehat{C}\left(1 + \dfrac{\overline{C}}{\mu}\right)$, we finish the proof. □

5. Numerical experiments

In this section, we present the application of the proposed ISBI-ADMMs (Algorithm 1) to the graphically guided fused LASSO problem (1.2). To highlight the performance advantages of the ISBI-ADMMs and to verify the effectiveness of Algorithm 1, it is compared with several state-of-the-art methods, including the conventional ADMM [42, 43], SGD-ADMM [44] with stochastic gradient descent, and the SVRG-ADMM [45] with variance-reduction techniques. Furthermore, after integrating the aforementioned existing methods with the ISBI-ADMMs, we then compare these integrated methods with the original methods to further demonstrate the advantages of the ISBI-ADMMs. All codes are implemented in MATLAB 2024b and executed on a Windows 10 system equipped with an Intel i7-10510U CPU.

During the experiments, we differentiate and label the different gradient estimation strategies: the SADMM algorithms integrating the SGD and SVRG gradient estimators are named the SADMM and SVRG, respectively. Similarly, the ISBI-ADMMs algorithms equipped with the above mentioned gradient estimators are labeled as the BSADMM and BSVRG in turn. In addition, the integration

of the ISBI-ADMMs algorithm with the SARAH gradient estimator proposed by Bian et al. [21] is denoted as BSARAH.

We apply the proposed Algorithm 1 to the previously mentioned graph-guided fused LASSO model (1.2), which can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) + \gamma_1\|\mathbf{z}\|_1, \\ &\text{s.t. } A\mathbf{x} - \mathbf{z} = 0. \end{aligned} \tag{5.1}$$

We have already mentioned the specifics in detail in (1.1) and (1.2), where $A \in \mathbb{R}^{m\times n_a}$, $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^{n_a}$. Equation (5.1) is an adapted form designed in this paper for the graph-guided fusion LASSO model (1.2), which essentially involves transforming the constraint $A\mathbf{x} - \mathbf{z} = 0$ of model (1.2) into the general constraint form $A\mathbf{x} + B\mathbf{y} = \mathbf{b}$ corresponding to Assumption 1(d). By setting $B = I$, $\mathbf{y} = -\mathbf{z}$, and $\mathbf{b} = 0$, the constraint $A\mathbf{x} - \mathbf{z} = 0$ can be rewritten as $A\mathbf{x} + I\mathbf{y} = 0$. The core of the graph-guided fused LASSO approach lies in characterizing the graph-structured relationships between variables via matrix A. The graph structures corresponding to the datasets used in the experiments are all connected graphs, which implies that the row vectors of A are linearly independent, i.e., A is a full row rank matrix and $\mathrm{Im}(A) = \mathbb{R}^m$. Meanwhile, the image space of the m-dimensional identity matrix I is $\mathrm{Im}(I) = \mathbb{R}^m$. Thus, $\mathrm{Im}(I) = \mathbb{R}^m = \mathrm{Im}(A)$, which clearly satisfies $\mathrm{Im}(I) \subseteq \mathrm{Im}(A)$. Additionally, $\mathbf{b} = 0$ obviously belongs to the image space of any matrix. Therefore, the condition $\mathrm{Im}(I) \cup \{0\} \subseteq \mathrm{Im}(A)$ holds entirely. Consequently, we have verified that (5.1) satisfies Assumption 1(d).

In this experiment, since there exists an explicit solution to the approximation operator for g, we let $\psi_1 = 0$ and $\mathbf{z}^{k+1}$ can be easily obtained in Step 2 of Algorithm 1. Note that this adaptation scenario does not apply to $\mathbf{x}^{k+1}$. To ensure that the $\mathbf{x}$-subproblem has the property of an explicit solution, $\psi_2 = \frac{\tau}{2}\|\mathbf{x}\|_Q^2$ can be parameterized in Step 3 of Algorithm 1. In this setting, the solution of the $\mathbf{x}$-subproblem will take the following specific form:

$$\mathbf{x}^{k+1} = (\beta A^\top A + \tau Q)^{-1}\left(\tau Q\mathbf{x}^k - \left(\nabla f(\mathbf{x}^k, \xi_M) - A^\top \boldsymbol{\lambda}^k - \beta A^\top \mathbf{z}^{k+1}\right)\right).$$

To avoid computing the inverse $(\beta A^\top A + \tau Q)^{-1}$, we define $Q = I - \tau^{-1}\beta A^\top A$. We then construct the following framework to solve (5.1):

$$\begin{cases} \overline{\mathbf{x}}^k = \mathbf{x}^k + \theta\left(\mathbf{x}^k - \overline{\mathbf{x}}^{k-1}\right), \\ \mathbf{z}^{k+1} \in S_{\frac{\lambda_1}{\beta}}\left(A\overline{\mathbf{x}}^k - \dfrac{\boldsymbol{\lambda}^k}{\beta}\right), \\ \mathbf{x}^{k+1} \approx \mathbf{x}^k - \dfrac{1}{\tau}\left(\nabla f(\mathbf{x}, \xi_M) - A^\top \boldsymbol{\lambda}^k + \beta A^\top\left(A\mathbf{x}^k - \mathbf{z}^{k+1}\right)\right), \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\left(A\mathbf{x}^{k+1} - \mathbf{z}^{k+1}\right), \end{cases}$$

where $S_{\frac{\lambda_1}{\beta}}(\cdot)$ is the soft threshold operator. The soft-threshold function is often used in sparse reconstruction to deal with ℓ_1 regularization subproblems, and the soft-threshold operator is specifically defined in [46] as follows for an input signal r and a threshold parameter $c > 0$:

$$\mathrm{Prox}(r, c) = \mathrm{sign}(r) \cdot \max\{\mid r \mid -c,\ 0\}, \tag{5.2}$$

and

$$\text{sign}(r) = \begin{cases} -1, & r < 0, \\ 0, & r = 0, \\ 1, & r > 0. \end{cases}$$

Hence, the soft-thresholding operation is a nonlinear shrinkage function that induces sparsity in the solution by causing smaller coefficients to be set to zero, and larger coefficients to shrink by the threshold value. In the $(k+1)$-th iteration, the primal residual is expressed as $r^{k+1} = A\mathbf{x}^{k+1} - \mathbf{z}^{k+1}$, and the dual residual is presented as $s^{k+1} = \beta A^T(\mathbf{z}^{k+1} - \mathbf{z}^k)$, which satisfy the following inequalities:

$$\begin{cases} \|r^{k+1}\|_2 \leq \sqrt{n_a} \times 10^{-4} + 10^{-2} \cdot \max\left(\|A\mathbf{x}^k\|_2, \|\mathbf{z}^k\|_2\right), \\ \|s^{k+1}\|_2 \leq \sqrt{n_a} \times 10^{-4} + 10^{-2} \cdot \|\beta A^T \boldsymbol{\lambda}^k\|_2, \end{cases}$$

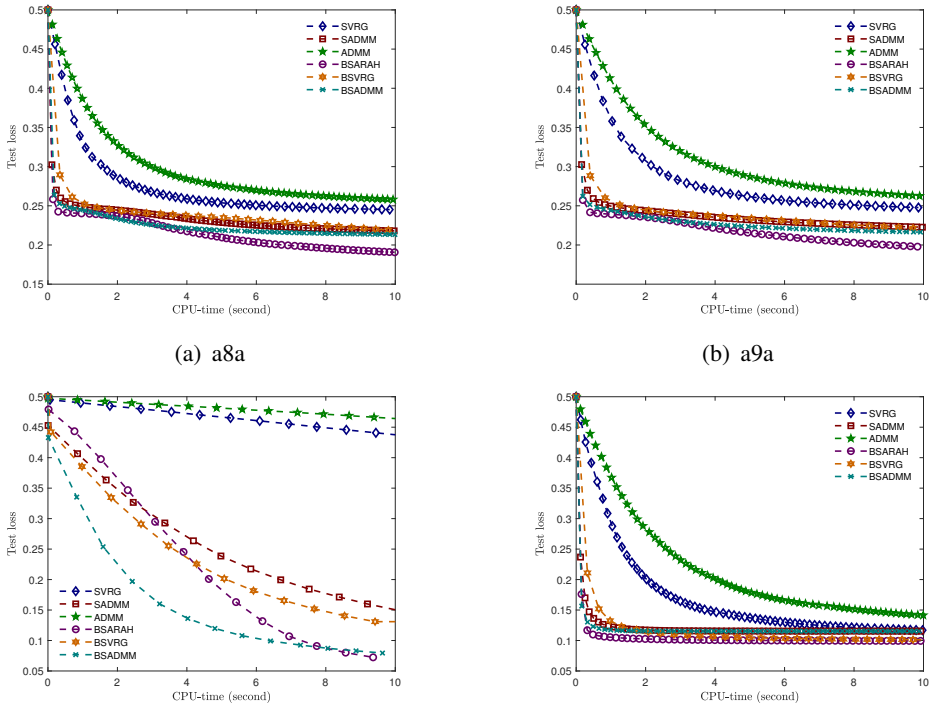
where n_a denotes the column number of matrix A, and then the iteration termination occurs.

In our experiments, we used four publicly available datasets [47], the relevant details of which are listed in Table 1. The sum of the training size and testing size in Table 1 yields the total number of our samples, denoted as n. Additionally, the number of columns of matrix A, denoted as n_a, is defined as the number of features in Table 1 plus 1.

Table 1. Datasets for graph-guided fused LASSO.

Datasets	Training size	Testing size	Feature	Classes
a8a	22696	9865	123	2
a9a	32561	16281	123	2
w8a	49749	14951	300	2
ijcnn1	49990	91701	22	2

On the a8a, a9a, w8a, and ijcnn1 datasets, by setting the regularization parameters $\gamma_1 = 1e-3$, $\gamma_2 = 1e-4$, $\beta = 5$, and for the ISBI-ADMMs using SGD, we use $\theta = 0.5$, and select $\tau = 0.2$ and the dual stepsize-related parameter s closer to 2 for testing. We found that $s = 1.5$ yields the optimal performance. Moreover, the test experiments demonstrate that this choice not only avoids the issue of slow convergence when $s < 1$ but also prevents residual oscillations, which occur when $s > 1.6$. In the ISBI-ADMMs with various variance shrinking gradient estimators, we use $\theta = 0.4$. Under these parameter settings and Assumptions 1(a)–(c), we further systematically optimized the batch size M, and finally determined $M = 500$. The low-variance property brought by this selection can reduce the number of iterations required for the algorithm to converge to a stable loss, thereby improving convergence efficiency. Meanwhile, it can also meet the requirements of the four datasets used in our experiments. We initialize $x^1 = 0$. In Figure 1, we give the results of the various methods based on the test losses of the different datasets while keeping the same running time. In addition to this, we found that Bian et al. [21] proposed a gradient estimator for SARAH with variance reduction, and BSARAH, obtained by the ISBI-ADMMs piggybacking on this gradient estimator, shows the fastest performance in most cases.



(a) a8a (b) a9a

(c) w8a (d) ijcnn1

Figure 1. Comparison of convergence efficiency of test loss with CPU time for different algorithms on typical datasets.

By comparing the test loss convergence characteristics of the ISBI-ADMMs family of algorithms equipped with SGD, SARAH, and SVRG variance-reduced gradient estimation with those of the SVRG-ADMM, SADMM and conventional ADMM, the results show that BSARAH exhibits the fastest convergence speed on all datasets, with the test loss decreasing rapidly and stabilizing in a short period of time. The BSADMM and BSVRG also significantly outperform the SADMM with the corresponding stochastic gradient estimator as well as the conventional ADMM, with more efficient loss reduction and lower final convergence values. Moreover, the curve of the conventional ADMM always changes slowly in the range of relatively high loss, which further verifies the dual advantages of the ISBI-ADMMs in both convergence efficiency and final accuracy.

This result validates the adaptability of the ISBI-ADMMs framework for stochastic gradient estimation and the advantage of parameter design in guaranteeing numerical stability while breaking through the convergence performance through variance reduction techniques. On the contrary, the traditional and similar comparison algorithms have slow loss reduction and high convergence values due to underutilizing the gradient optimization strategy, which further highlights the efficient and stable solving capability of the ISBI-ADMMs in large-scale optimization problems of the form (1.3).

6. Conclusions

In this paper, we studied the nonconvex and nonsmooth optimization problem (1.3). By integrating a class of variance-reducing gradient estimation techniques, such as SVRG and SARAH, we proposed

a unified framework for the inexact stochastic ADMMs. In addition, the framework also integrated inertial acceleration techniques with Bregman distance regularization, which significantly improves the convergence efficiency and practical applicability of the ISBI-ADMMs. Theoretical analysis showed that under some mild assumptions, we established the global convergence and sublinear convergence rate $\mathcal{O}(1/\mathbb{K})$ of the algorithm in nonconvex scenarios. Meanwhile, under stronger assumptions, we established the linear convergence rate of the algorithm, which provided theoretical support for the stability of the algorithm. To verify the theoretical contribution and effectiveness of the method, we conducted multiple sets of comparative numerical experiments for the graph-guided fusion LASSO problem (1.1) or (5.1), and the results fully demonstrated the superiority of the ISBI-ADMMs in complex optimization scenarios.

If the inertial term in the ISBI-ADMMs for the variable $\mathbf{y}$ or the multiplier $\boldsymbol{\lambda}$ is considered, the rigor of the theoretical analysis for the ISBI-ADMMs is worth exploring. Hence, in the future research, we will focus on the design, convergence analysis, and application of the stochastic ADMM algorithm under distributed systems. It is also worth investigating whether embodying the stochastic nature of the ISBI-ADMMs in the objective function can make the problem more generalized.

Author contributions

Yi-xin Yang: Conceptualization, Methodology, Software, Validation, Visualization, Writing-original draft, Writing-review & editing; Heng-you Lan: Conceptualization, Validation, Visualization, Writing-original draft, Writing-review & editing, Supervision, Project administration; Lin-cheng Jiang: Software, Visualization. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the Innovation Fund of Postgraduate, Sichuan University of Science & Engineering (Y2025105), the Scientific Research and Innovation Team Program of Sichuan University of Science and Engineering (SUSE652B002), and the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (2024QZJ01).

Conflict of interest

The authors declare no conflicts of interest.

References

1. F. H. Clarke, Generalized gradients and applications, *Trans. Am. Math. Soc.*, **205** (1975), 247–262. https://doi.org/10.2307/1997202

2. T. Chan, S. Esedoglu, F. Park, A. Yip, Total variation image restoration: Overview and recent developments, In: *Handbook of Mathematical Models in Computer Vision*, New York: Springer, 2006, 17–31. https://doi.org/10.1007/0-387-28831-7_2
3. F. Iutzeler, J. Malick, Nonsmoothness in machine learning: Specific structure, proximal identification, and applications, *Set-Valued Var. Anal.*, **28** (2020), 661–678. https://doi.org/10.1007/s11228-020-00561-1
4. D. Davis, B. Edmunds, M. Udell, The sound of apalm clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm, In: *Advances in Neural Information Processing Systems*, **29** (2016), 226–234.
5. C. J. Hsieh, M. A. Sustik, I. S. Dhillon, P. Ravikumar, QUIC: Quadratic approximation for sparse inverse covariance estimation, *J. Mach. Learn. Res.*, **15** (2014), 2911–2947. https://doi.org/10.5555/2627435.2697058
6. S. Kim, K. A. Sohn, E. P. Xing, A multivariate regression approach to association analysis of a quantitative trait network, *Bioinformatics*, **25** (2009), i204–i212. https://doi.org/10.1093/bioinformatics/btp218
7. Y. Liu, F. Shang, H. Liu, L. Kong, L. Jiao, Z. Lin, Accelerated variance reduction stochastic ADMM for large-scale machine learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2020), 4242–4255. https://doi.org/10.1109/TPAMI.2020.3000512
8. Y. Cheung, J. Lou, Proximal average approximated incremental gradient descent for composite penalty regularized empirical risk minimization, *Mach. Learn.*, **106** (2017), 595–622. https://doi.org/10.1007/s10994-016-5609-1
9. L. Liu, C. Han, T. Guo, S. Liao, An inertial stochastic Bregman generalized alternating direction method of multipliers for nonconvex and nonsmooth optimization, *Expert Syst. Appl.*, **276** (2025), 126939. https://doi.org/10.1016/j.eswa.2025.126939
10. P. Gong, C. Zhang, Z. Lu, J. Huang, J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, In: *International Conference on Machine Learning*, 2013, 37–45.
11. M. R. Metel, A. Takeda, Stochastic proximal methods for non-smooth non-convex constrained sparse optimization, *J. Mach. Learn. Res.*, **22** (2021), 1–36.
12. R. G. Crespo, E. Verdú, M. Khari, A. K. Garg, Gesture recognition of RGB and RGB-D static images using convolutional neural networks, *Int. J. Interact. Multimed. Artif. Intell.*, **5** (2019), 22–27. https://doi.org/10.9781/ijimai.2019.09.002
13. Y. Tian, Y. Zhang, H. Zhang, Recent advances in stochastic gradient descent in deep learning, *Mathematics*, **11** (2023), 682. https://doi.org/10.3390/math11030682
14. W. Zhong, J. Kwok, Fast stochastic alternating direction method of multipliers, In: *International Conference on Machine Learning, PMLR*, (2014), 46–54.
15. T. Suzuki, Stochastic dual coordinate ascent with alternating direction method of multipliers, In: *International Conference on Machine Learning, PMLR*, 2014, 736–744.
16. R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, In: *Advances in Neural Information Processing Systems, New York*, **26** (2013), 315–323.

17. L. M. Nguyen, J. Liu, K. Scheinberg, M. Takáč, SARAH: A novel method for machine learning problems using stochastic recursive gradient, In: *International Conference on Machine Learning, PMLR*, 2017, 2613–2621.

18. Y. Zhao, M. Li, X. Pan, J. Tan, Partial symmetric regularized alternating direction method of multipliers for non-convex split feasibility problems, *AIMS Math.*, **10** (2025), 3041–3061. https://doi.org/10.3934/math.2025142

19. M. Yashtini, Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization, *J. Glob. Optim.*, **84** (2022), 913–939. https://doi.org/10.1007/s10898-022-01174-8

20. F. Huang, S. Chen, Z. Lu, Stochastic alternating direction method of multipliers with variance reduction for nonconvex optimization, preprint paper, 2016. https://doi.org/10.48550/arXiv.1610.02758

21. 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

22. Y. Zeng, Z. Wang, J. Bai, X. Shen, An accelerated variance reduction stochastic ADMM for nonconvex nonsmooth optimization, In: *2022 China Automation Congress (CAC), Institute of Electrical and Electronics Engineers*, 2022, 4853–4858. https://doi.org/10.1109/CAC57257.2022.10055828

23. Y. Zeng, Z. Wang, J. Bai, X. Shen, An accelerated stochastic ADMM for nonconvex and nonsmooth finite-sum optimization, *Automatica*, **163** (2024), 111554. https://doi.org/10.1016/j.automatica.2024.111554

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

25. J. Bai, D. Han, H. Sun, H. Zhang, Convergence on a symmetric accelerated stochastic ADMM with larger stepsizes, *CSIAM Trans. Appl. Math.*, **3** (2022), 448–479. https://doi.org/10.1103/physrevd.105.032007

26. F. Wang, Z. Xu, H. K. Xu, Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems, preprint paper, 2014. https://doi.org/10.48550/arXiv.1410.8625

27. M. Chao, Z. Deng, J. Jian, Convergence of linear Bregman ADMM for nonconvex and nonsmooth problems with nonseparable structure, *Complexity*, **2020** (2020), 6237942. https://doi.org/10.1155/2020/6237942

28. P. J. Liu, J. B. Jian, B. He, X. Z. Jiang, Convergence of Bregman Peaceman-Rachford splitting method for nonconvex nonseparable optimization, *J. Oper. Res. Soc. China*, **11** (2023), 707–733. https://doi.org/10.1007/s40305-022-00411-x

29. L. Tan, K. Guo, Bregman ADMM: A new algorithm for nonconvex nonconvex wuth linear constraints, *J. Nonlinear Var. Anal.*, **9** (2025), 179–196. https://doi.org/10.23952/jnva.9.2025.2.02

30. P. J. Liu, J. B. Jian, H. Shao, X. Q. Wang, J. W. Xu, X. Y. Wu, A Bregman-style improved ADMM and its linearized version in the nonconvex setting: convergence and rate analyses, *J. Oper. Res. Soc. China.*, **12** (2024), 298–340. https://doi.org/10.1007/s40305-023-00535-8

31. B. T. Polyak, Some methods of speeding up the convergence of iteration methods, *USSR Comput. Math. Math. Phys.*, **4** (1964), 1–17. https://doi.org/10.1016/0041-5553(64)90137-5

32. L. T. K. Hien, D. Papadimitriou, An inertial ADMM for a class of nonconvex composite optimization with nonlinear coupling constraints, *J. Glob. Optim.*, **89** (2024), 927–948. https://doi.org/10.1007/s10898-024-01382-4

33. F. Huang, S. Chen, H. Huang, Faster stochastic alternating direction method of multipliers for nonconvex optimization, In: *International Conference on Machine Learning, PMLR*, 2019, 2839–2848.

34. Z. Lin, H. Li, C. Fang, Accelerated stochastic algorithms, In: *Accelerated Optimization for Machine Learning*, 2020, 137–207. https://doi.org/10.1007/978-981-15-2910-8_5

35. M. Chao, Y. Geng, Y. Zhao, A method of inertial regularized ADMM for separable nonconvex optimization problems, *Soft Comput.*, **27** (2023), 16741–16757. https://doi.org/10.1007/s00500-023-09017-8

36. J. Bai, M. Zhang, H. Zhang, An inexact ADMM for separable nonconvex and nonsmooth optimization, *Comput. Optim. Appl.*, **90** (2025), 445–479. https://doi.org/10.1007/s10589-024-00643-y

37. Y. Zeng, J. Bai, S. Wang, Z. Wang, A unified inexact stochastic admm for composite nonconvex and nonsmooth optimization, preprint paper, 2024. https://doi.org/10.48550/arXiv.2403.02015

38. J. Yin, C. Tang, J. Jian, Q. Huang, A partial Bregman ADMM with a general relaxation factor for structured nonconvex and nonsmooth optimization, *J. Glob. Optim.*, **89** (2024), 899–926. https://doi.org/10.1007/s10898-024-01384-2

39. H. Robbins, S. Monro, A stochastic approximation method, *Ann. Math. Statist.*, 1951, 400–407. https://doi.org/10.1109/TSMC.1971.4308316

40. R. T. Rockafellar, R. J. B. Wets, *Variational analysis*, Berlin: Springer Science & Business Media, 1998. https://doi.org/10.1007/978-3-642-02431-3

41. M. L. N. Gonçalves, J. G. Melo, R. D. C. Monteiro, Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems, *Pac. J. Optim.*, **15** (2019), 379–398.

42. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. Appl.*, **2** (1976), 17–40. https://doi.org/10.1016/0898-1221(76)90003-1

43. R. Glowinski, A. Marroco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires, *Rev. Fr. Autom. Inform. Rech. Opér., Anal. Numér.*, **9** (1975), 41–76. https://doi.org/10.1051/m2an/197509R200411

44. F. Huang, S. Chen, Mini-batch stochastic ADMMs for nonconvex nonsmooth optimization, preprint paper, 2018. https://doi.org/10.48550/arXiv.1802.03284

45. S. Zheng, J. T. Kwok, Fast-and-Light stochastic ADMM, In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, Palo Alto, California: AAAI Press, 2016, 2407–2613.

46. S. J. Wright, R. D. Nowak, M. A. T. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Trans. Signal Process.*, **57** (2009), 2479–2493. https://doi.org/10.1109/ICASSP.2008.4518374

47. C. C. Chang, C. J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, **2** (2011), 1–27. https://doi.org/10.1145/1961189.1961199

AIMS Press