Computational Optimization and Applications
https://doi.org/10.1007/s10589-024-00643-y

An inexact ADMM for separable nonconvex and nonsmooth optimization

Jianchao Bai[1,2] · Miao Zhang[3] · Hongchao Zhang[3]

Received: 27 February 2024 / Accepted: 23 December 2024

Abstract

An inexact alternating direction method of multiplies (I-ADMM) with an expansion linesearch step was developed for solving a family of separable minimization problems subject to linear constraints, where the objective function is the sum of a smooth but possibly nonconvex function and a possibly nonsmooth nonconvex function. Global convergence and linear convergence rate of the I-ADMM were established under proper conditions while inexact relative error criterion was used for solving the subproblems. In addition, a unified proximal gradient (UPG) method with momentum acceleration was proposed for solving the smooth but possibly nonconvex subproblem. This UPG method guarantees global convergence and will automatically reduce to an optimal accelerated gradient method when the smooth function in the objective is convex. Our numerical experiments on solving nonconvex quadratic programming problems and sparse optimization problems from statistical learning show that the proposed I-ADMM is very effective compared with other state-of-the-art algorithms in the literature.

This research was partially supported by the National Natural Science Foundation of China (12471298), the Shaanxi Fundamental Science Research Project for Mathematics and Physics (23JSQ031), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012405), and the USA National Science Foundation (DMS-2110722, DMS-2309549).

✉ Hongchao Zhang
hozhang@math.lsu.edu
https://math.lsu.edu/~hozhang

Jianchao Bai
jianchaobai@nwpu.edu.cn

Miao Zhang
mzhan33@lsu.edu

[1] Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China

[2] School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an 710072, China

[3] Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA

Published online: 09 January 2025

Keywords Nonconvex optimization · Nonsmooth optimization · Separable structure · Lipschitz continuous · Inexact ADMM · Accelerated gradient method · Global convergence · Linear convergence rate

Mathematics Subject Classification 65K10 · 65Y20 · 90C26

1 Introduction

We consider the following separable nonconvex and nonsmooth linearly constrained optimization problem

$$\min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{n_\mathbf{x}}\times\mathbb{R}^{n_\mathbf{y}}} F(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + g(\mathbf{y}) \quad \text{subject to} \quad A\mathbf{x} + B\mathbf{y} = \mathbf{b}, \tag{1.1}$$

where $f : \mathbb{R}^{n_\mathbf{x}} \to \mathbb{R}$ is Lipschitz continuously differentiable, but possibly nonconvex, $g : \mathbb{R}^{n_\mathbf{y}} \to \mathbb{R}$ is a proper, lower semi-continuous, possibly nonconvex and nonsmooth function and $A \in \mathbb{R}^{m\times n_\mathbf{x}}$, $B \in \mathbb{R}^{m\times n_\mathbf{y}}$ and $\mathbf{b} \in \mathbb{R}^m$ are given data. Note that constraints of the form $\mathbf{y} \in \mathcal{Y}$ for a closed set $\mathcal{Y} \subset \mathbb{R}^{n_\mathbf{y}}$ can be incorporated in the objective using g as an indicator function of $\mathcal{Y}$. In recent years, problems in the form of (1.1) have attracted sufficient attention both theoretically and numerically, simply due to its special structure and many concrete important applications including statistical learning [10, 20, 46], compressive sensing [64, 65, 67], machine learning [3, 42], phase retrieval [63], image restoration and extraction [13, 69], etc.

It is well-known that the Alternating Direction Method of Multiplies (ADMM) has obtained great success in both theory and numerical efficiency for solving linearly constrained separable convex optimization. Hence, the original ADMM [22, 26] and its variants for solving convex problems have been extended recently to solve the nonconvex structured optimization problem (1.1). Unlike the well-studied Augmented Lagrangian Method (ALM) [41], ADMM can exploit the problem's separable structure and use the special properties of each component function in the objective. Directly extending the original ADMM for solving the problem (1.1) performs the optimization in the following alternative order:

$$\begin{cases} \mathbf{y}^{k+1} \in \arg\min_{\mathbf{y}} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k), \\ \mathbf{x}^{k+1} \in \arg\min_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k), \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right), \end{cases} \tag{1.2}$$

where $s \in (0, \frac{1+\sqrt{5}}{2})$ denotes the stepsize of dual variable $\boldsymbol{\lambda}$ and $\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ is the augmented Lagrangian with penalty parameter $\beta > 0$ defined as

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = \mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) + \frac{\beta}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2 \tag{1.3}$$

and $\mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ is the Lagrangian function of (1.1) defined as

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

The global convergence and complexity of 2-block ADMM, such as (1.2) for solving convex problems have been well-studied [18, 39]. Multi-block ADMMs have also received intensive research in both deterministic and stochastic setting including [2, 4, 5, 12, 14, 17, 23, 27, 35, 38, 48, 49]. However, all the above mentioned work focuses on convex optimization problems. The studies of ADMM for solving (1.1) with nonconvex objective function are much limited despite its high demands in applications. Indeed, unlike solving convex problems, ADMM for solving nonconvex problems could fail for arbitrary choice of the penalty parameter $\beta > 0$. However, with proper choice of β, the excellent performance of nonconvex ADMM has been observed in recent applications [61]. These mysteries of practical success in fact triggered the recent rigorous study on ADMM for nonconvex optimization. For example, under suitable assumptions, ADMM and its variants have been shown convergent [30, 31, 44, 45] for solving two block and multi-block nonconvex problems. Moreover, ADMMs have been applied to solve some special nonconvex models with particular choices of A and B [43, 47] and certain nonconvex signal/image recovery problems [6, 69]. Note that the dominant computation in each iteration of ADMM is to solve its subproblems. Hence, how to solve these subproblems inexactly while still maintaining nice convergence properties will be critical for the overall success of ADMM, especially when no closed-form solution of the subproblem exists [33, 34, 64]. However, the current work of nonconvex ADMM for (1.1) still lacks sufficient rigorous study on solving its subproblems inexactly in a more practical way. A nice theoretical framework on nonconvex ADMM is discussed in [61], but the theories therein still assume exact subproblem solution of the proposed ADMM and its global convergence under the adaptive inexact criteria in [61] remains incomplete. Moreover, no numerical experiments and only unit dual stepsize are considered in [61], while larger range of dual stepsize $s \in (0, \frac{\sqrt{5}+1}{2})$ is allowed in the original ADMM.

In this paper, motivated by the recent surged interests for studying nonconvex ADMM and the adaptive relative error strategy used in ALM and convex ADMM (Ex. [33]), we propose an Inexact ADMM (I-ADMM) framework with an expansion linesearch step (see Algorithm 3.1) to solve the nonconvex problem (1.1). Our proposed I-ADMM has the following major features.

(a) *The proposed I-ADMM solves the subproblems inexactly to adaptive accuracy while guarantees global convergence and linear convergence rate under proper conditions.* In the literature, unless special structure of f or g exits, almost all efficient ADMMs for convex problems solve the subproblems inexactly [15, 18, 28, 37]. Among these inexact ADMMs, one usual way is to solve the subproblems to the accuracy based on some absolute summable error criteria, but often without guidance on how to adaptively select the error tolerance except requiring it to be summable. Moreover, ADMM is just a splitting version of ALM, for which nice convergence theory and encouraging numerical results are often obtained [19, 55] using adaptive relative subproblem stopping criteria. Hence, ideally we should also solve the subproblems of I-ADMM

to an adaptive accuracy while maintaining desirable convergence properties. In this paper, we establish global convergence and linear convergence rate of I-ADMM under a local error bound condition and a weakly convex property of g.

(b) *The proposed I-ADMM allows more flexible stepsize* $s \in (0, 2)$ *of the dual variable stepsize and applies an expansion linesearch step to accelerate the convergence.* It is well-known that the dual stepsize s of ADMM for solving convex optimization can be arbitrary in the interval $(0, (\sqrt{5}+1)/2)$ [2, 21, 26]. Hence, it is desirable to allow a more flexible dual stepsize of I-ADMM while not losing convergence. But only fixed dual stepsize $s = 1$ was discussed in almost all the current nonconvex ADMMs [6, 43, 47, 61], except the methods in [68, 69] allow $s \in (0, (\sqrt{5}+1)/2)$ for an image recovery problem as original ADMM and $s \in (0, 2)$ for a linearized ADMM. However, both methods assume exact subproblem or linearized subproblem solution. In this paper, applying a much different potential energy function, we show that the dual stepsize interval can be $(0, 2)$ even with inexact subproblem solution. In addition, an expansion linesearch step (see step 6 of Algorithm 3.1) is applied in our I-ADMM, which not only improves the numerical performance but also reduces the sensitivity of algorithm parameters as well.

(c) *We propose a unified proximal gradient (UPG) method with momentum acceleration to solve the nonconvex smooth* $\mathbf{x}$*-subproblem*. Our UPG method is motivated by the extrapolation techniques for solving both convex and nonconvex optimization [8, 62]. Uniform proximal gradient methods were also proposed in [24, 25]. However, [24] requires all iterates must belong to a bounded set for global convergence and the method in [25] could just reduce to a simple proximal descent method without any momentum acceleration steps for nonconvex optimization. Our UPG method is particularly designed for solving $\mathbf{x}$-subproblem arising in our I-ADMM. This UPG method guarantees global convergence for solving the smooth but possibly nonconvex subproblem problem and will automatically reduce to an optimal gradient method, maintaining optimal complexity, when the function f in the objective is convex.

(d) *The framework of I-ADMM is more general and flexible than most of ADMMs in the literature.* When no expansion step (Step 6 of Algorithm 3.1) is used, this I-ADMM will just reduce to a particular inexact version of nonconvex ADMM without a relaxation step. But our linesearch expansion step often allows much larger stepsize than the fixed relaxation stepsize used in [18, 36, 40]. Convergence of the ADMM-type methods in [11, 47] were established for (1.1) with $B = \mathbf{I}$ and $\mathbf{b} = \mathbf{0}$ under the Kurdyka-Łojasiewicz property, while we have used more general problem settings and different assumptions for establishing global convergence and linear convergence rate. Although the over-relaxation step was adopted in [29], the involved subproblems were also solved exactly. Moreover, our numerical experiments show that the proposed I-ADMM is very effective compared with other state-of-the-art algorithms in the literature and could obtain more accurate solution.

The paper is organized as follows. In Sect. 2, we introduce some notations, definitions and some well-known results in the literature. Section 3 describes the framework of our proposed I-ADMM algorithm. The global convergence and convergence rate of I-ADMM are studied in Sect. 4. In Sect. 5, we propose a Unified Proximal Gradient (UPG) method with momentum acceleration for solving the smooth but possibly

nonconvex subproblem. Numerical experiments on solving some nonconvex quadratic programming problems and sparse optimization problems from statistical learning are given in Sect. 6. Conclusions are drawn in Sect. 7.

2 Notation and preliminaries

Let $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ be the sets of real numbers, n dimensional real column vectors, and $n\times m$ real matrices, respectively. Let $\mathbf{I}$ denote the identity matrix and $\mathbf{0}$ denote zero matrix/vector. For symmetric matrices A and B of the same dimension, $A \succ B$ ($A \succeq B$) means $A - B$ is a positive definite (semidefinite) matrix. For two vectors $\mathbf{v}$ and $\mathbf{u}$ in $\mathbb{R}^n$, $\mathbf{u} > \mathbf{v}$ ($\mathbf{u} \geq \mathbf{v}$) means $\mathbf{u}$ is component-wise larger (not less than) $\mathbf{v}$. We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the standard Euclidean norm in $\mathbb{R}^n$ and the associated inner product. For any positive semidefinite matrix $D \succeq \mathbf{0}$, let $\|\mathbf{x}\|_D^2 = \mathbf{x}^{\mathsf{T}} D\mathbf{x}$. For a matrix A, Range(A) denotes the range of A and for a nonempty closed set $\mathcal{C} \subseteq \mathbb{R}^n$, we use dist$(\mathbf{x}, \mathcal{C})$ to denote the Euclidean distance from $\mathbf{x}$ to $\mathcal{C}$, i.e., dist$(\mathbf{x}, \mathcal{C}) = \inf_{\mathbf{z}\in\mathcal{C}} \|\mathbf{x} - \mathbf{z}\|$. Given an extended real-valued function $h : \mathbb{R}^n \to [-\infty, \infty]$, dom $h := \{\mathbf{x} \in \mathbb{R}^n : h(\mathbf{x}) < \infty\}$ denotes its effective domain. A function h is said to be proper if $h(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathbb{R}^n$ and dom h is nonempty. For a proper lower semi-continuous function h, its *(limiting-) subdifferential* [54, Definition 8.3 (b)] at $\mathbf{x} \in$ dom h, denoted as $\partial h(\mathbf{x})$, is defined as

$$\partial h(\mathbf{x}) := \left\{ \boldsymbol{\nu} \in \mathbb{R}^n : \exists \mathbf{x}^k \to \mathbf{x},\, h(\mathbf{x}^k) \to h(\mathbf{x}),\, \boldsymbol{\nu}^k \to \boldsymbol{\nu} \text{ with } \boldsymbol{\nu}^k \in \widehat{\partial} h(\mathbf{x}^k) \right\}, \quad (2.1)$$

where $\widehat{\partial} h(\mathbf{x})$ denotes the *regular subdifferential* [54, Definition 8.3 (a)] of h at $\mathbf{x}$ given as

$$\widehat{\partial} h(\mathbf{x}) := \left\{ \boldsymbol{\nu} \in \mathbb{R}^n : \liminf_{\mathbf{z}\to\mathbf{x}, \mathbf{z}\neq\mathbf{x}} \frac{h(\mathbf{z}) - h(\mathbf{x}) - \langle \nu, \mathbf{z} - \mathbf{x}\rangle}{\|\mathbf{z} - \mathbf{x}\|} \geq 0 \right\}.$$

It is well-known that the subdifferential (2.1) coincides with the classical subdifferential of a proper closed convex function h and is the gradient of h, denoted as ∇h, when h is continuously differentiable. However, the limiting subdifferential plays a much wider role in nonsmooth and nonconvex analysis and optimization [54, Exercise 8.8 and Proposition 8.12]. For example, the Fermat's rule remains true, that is, if $\mathbf{x}$ is a local minimizer of h, then $\mathbf{0} \in \partial h(\mathbf{x})$ [54, Theorem 10.1].

3 Algorithm description

We propose an inexact ADMM (I-ADMM, i.e., Algorithm 3.1) with an expansion linesearch step to solve the possibly nonsmooth and nonconvex problem (1.1). At each iteration, both the $\mathbf{y}$-subproblem, i.e.,

$$\min_{\mathbf{y}\in\mathbb{R}^{n_\mathbf{y}}} \mathcal{L}_\mathbf{y}^k(\mathbf{y}) := \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}^k\|_{\mathcal{D}_\mathbf{y}^k}^2, \quad (3.1)$$

and the $\mathbf{x}$-subproblem, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^{n_\mathbf{x}}} \mathcal{L}_\mathbf{x}^k(\mathbf{x}) := \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_\mathbf{x}^k}, \tag{3.2}$$

are allowed to be solved inexactly, where $\mathcal{D}_\mathbf{x}^k \succeq \mathbf{0}$ and $\mathcal{D}_\mathbf{y}^k \succeq \mathbf{0}$ could be two adaptively chosen uniformly upper bounded positive semidefinite matrices. More precisely, in Algorithm 3.1, it requires the $\mathbf{y}^{k+1}$ generated at the k-th iteration satisfies

$$\frac{\beta}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2_{\mathcal{D}_\mathbf{y}} + \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \le \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k) \tag{3.3}$$

for some positive definite matrix $\mathcal{D}_\mathbf{y} \succ \mathbf{0}$, and there exists a positive constant $c_\mathbf{y} > 0$ and some $\xi_\mathbf{y}^{k+1} \in \partial_\mathbf{y}\mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$ such that

$$\|\xi_\mathbf{y}^{k+1}\| \le c_\mathbf{y}\beta\|\mathbf{y}^{k+1} - \mathbf{y}^k\|. \tag{3.4}$$

For inexact solution of $\mathbf{x}$-subproblem, it requires the $\widehat{\mathbf{x}}^k$ generated at the k-th iteration of Algorithm 3.1 satisfies

$$\frac{\beta}{2}\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|^2_{\mathcal{D}_\mathbf{x}} + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \le \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \tag{3.5}$$

for some positive definite matrix $\mathcal{D}_\mathbf{x} \succ \mathbf{0}$, and there exists a positive constant $c_\mathbf{x} > 0$ such that $\xi_\mathbf{x}^{k+1} = \nabla_\mathbf{x}\mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$ satisfies

$$\|\xi_\mathbf{x}^{k+1}\| \le c_\mathbf{x}\beta\Big(\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|\Big). \tag{3.6}$$

The algorithm stops when R^{k+1} is sufficiently small, where

$$R^{k+1} = \|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\| + \|\widehat{\mathbf{r}}^{k+1}\|, \tag{3.7}$$

and $\widehat{\mathbf{r}}^{k+1} = A\widehat{\mathbf{x}}^k + B\mathbf{y}^{k+1} - \mathbf{b}$. Furthermore, we see that an expansion linesearch step for $\mathbf{x}$-iterates is applied in Step 6 of Algorithm 3.1. From this expansion step, we have $\phi(\alpha_k) = \mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$, $\phi(1) = \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$ and the stepsize $\alpha_k \ge 1$ is chosen such that

$$\mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) \le \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2, \tag{3.8}$$

where $\delta \in (0, 1)$ is an algorithm parameter. As standard linesearch techniques in optimization, this Armijo-type linesearch step could significantly improve the algorithm performance as well as reduce the sensitivity of the choice of algorithm parameters.

We now have the following comments regarding the conditions (3.3), (3.4), (3.5) and (3.6) for the subproblem solutions. First, since $\{\mathcal{D}_\mathbf{x}^k\}$ and $\{\mathcal{D}_\mathbf{y}^k\}$ are chosen uniformly upper bounded, supposing functions $\mathcal{L}_\mathbf{x}^k(\cdot)$ and $\mathcal{L}_\mathbf{y}^k(\cdot)$ are bounded from below, we can

Initialization: parameters $\beta > 0$, $s \in (0,2)$, $\delta \in (0,1)$ and $\eta > 1$, starting point $\mathbf{w}^0 = (\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0)$.

For $k = 0, 1, 2, \ldots$

1. Choose uniformly upper bounded matrices $\mathcal{D}_{\mathbf{y}}^k \succeq \mathbf{0}$ and $\mathcal{D}_{\mathbf{x}}^k \succeq \mathbf{0}$.
2. Solve $\mathbf{y}^{k+1} \approx \arg\min_{\mathbf{y}\in\mathbb{R}^{n_\mathbf{y}}} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}^k\|^2_{\mathcal{D}_{\mathbf{y}}^k}$ inexactly such that (3.3) and (3.4) are satisfied.
3. Solve $\widehat{\mathbf{x}}^k \approx \arg\min_{\mathbf{x}\in\mathbb{R}^{n_\mathbf{x}}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_{\mathbf{x}}^k}$ inexactly such that (3.5) and (3.6) are satisfied.
4. If R^{k+1} defined in (3.7) is sufficiently small, stop.
5. Update the Lagrange multiplier:
   $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta(A\widehat{\mathbf{x}}^k + B\mathbf{y}^{k+1} - \mathbf{b})$.
6. Expansion step for the $\mathbf{x}$-iterate:
   $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \widehat{\mathbf{d}}_{\mathbf{x}}^k$, where $\widehat{\mathbf{d}}^k = \widehat{\mathbf{x}}^k - \mathbf{x}^k$ and $\alpha_k = \eta^j$ with $j \geq 0$ being the largest integer such that
   $\phi(\alpha_k) \leq \phi(1) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2$ and $\phi(\alpha) = \mathcal{L}_\beta(\mathbf{x}^k + \alpha\widehat{\mathbf{d}}_{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$.

end

Algorithm 3.1 An inexact ADMM (I-ADMM) for separable nonconvex optimization problem (1.1)

find $\mathbf{y}^{k+1}$ and $\widehat{\mathbf{x}}^k$ such that (3.4) and (3.6) will be satisfied. In addition, if $R^{k+1} = 0$, we can derive that $\mathbf{w}^k := (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ is a stationary point of the problem (1.1) (see definition (4.22)). On the other hand, if $\{\mathcal{D}_{\mathbf{x}}^k\}$ and $\{\mathcal{D}_{\mathbf{y}}^k\}$ are chosen such that

$$\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|^2_{\mathcal{D}_{\mathbf{x}}^k} \geq \eta_\mathbf{x}\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|^2 \quad \text{and} \quad \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2_{\mathcal{D}_{\mathbf{y}}^k} \geq \eta_\mathbf{y}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2$$

for some constants $\eta_\mathbf{x} > 0$ and $\eta_\mathbf{y} > 0$, then for any $\widehat{\mathbf{x}}^k$ satisfying $\mathcal{L}_\mathbf{x}^k(\widehat{\mathbf{x}}^k) \leq \mathcal{L}_\mathbf{x}^k(\mathbf{x}^k)$ and any $\mathbf{y}^{k+1}$ satisfying $\mathcal{L}_\mathbf{y}^k(\mathbf{y}^{k+1}) \leq \mathcal{L}_\mathbf{y}^k(\mathbf{y}^k)$, the conditions (3.3) and (3.5) will hold with $\mathcal{D}_\mathbf{x} = \eta_\mathbf{x}\mathbf{I}$ and $\mathcal{D}_\mathbf{y} = \eta_\mathbf{y}\mathbf{I}$. Obviously, one simple choice could be letting $\mathcal{D}_\mathbf{x}^k = \eta_\mathbf{x}\mathbf{I}$ and $\mathcal{D}_\mathbf{y}^k = \eta_\mathbf{y}\mathbf{I}$ for all $k \geq 0$. However, under certain circumstances, it is not even necessary to require positive definiteness of $\{\mathcal{D}_\mathbf{x}^k\}$ or $\{\mathcal{D}_\mathbf{y}^k\}$ in order to satisfy the conditions (3.3) and (3.5). For instance, denoting $L > 0$ as the Lipschitz constant of ∇f, if $A^\mathsf{T}A + \mathcal{D}_\mathbf{x}^k \succ \mathbf{0}$ and the parameter β is sufficiently large such that $\beta(A^\mathsf{T}A + \mathcal{D}_\mathbf{x}^k) \succeq (L + 2\eta\beta)\mathbf{I}$ for some $\eta > 0$, the objective function $\mathcal{L}_\mathbf{x}^k(\cdot)$ of the $\mathbf{x}$-subproblem (5.1) will be uniformly strongly convex with modulus greater than $2\eta\beta > 0$. Under this case, all points sufficiently close to the minimizer of the $\mathbf{x}$-subproblem (5.1) will satisfy (3.5) with $\mathcal{D}_\mathbf{x} = \eta\mathbf{I}$. Hence, in the following, we assume that we can solve the subproblems (3.1) and (3.2) inexactly to meet the conditions (3.3), (3.4), (3.5) and (3.6).

4 Convergence analysis

In this section, we would like to study the convergence properties of Algorithm 3.1. For the convergence analysis, we need the following assumptions throughout the paper:

Assumption 4.1 The gradient of f is Lipschitz continuous, i.e., there exits a constant $L > 0$ such that

$$\|\nabla f(\mathbf{z}_1) - \nabla f(\mathbf{z}_2)\| \leq L\|\mathbf{z}_1 - \mathbf{z}_2\| \tag{4.1}$$

for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_\mathbf{x}}$.

Assumption 4.2 $(\text{Range}(B) \cup \mathbf{b}) \subseteq \text{Range}(A)$.

Based on Assumption 4.2, we have $\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k = -s\beta\mathbf{r}^{k+1} \in \text{Range}(A)$, which implies

$$\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\| \leq \sigma_A^{-\frac{1}{2}}\|A^\mathsf{T}(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k)\|, \tag{4.2}$$

where σ_A is the smallest positive eigenvalue of $A^\mathsf{T}A$ (or equivalently the smallest positive eigenvalue of AA^T). Certainly, Assumption 4.2 holds if A is nonsingular or has full column or full row rank.

4.1 Technical preliminaries

In the following, to facilitate the analysis, for all $k \geq 0$ let us denote

$$\widehat{\mathbf{d}}_\mathbf{x}^k := \widehat{\mathbf{x}}^k - \mathbf{x}^k, \quad \widetilde{\mathbf{d}}_\mathbf{x}^k := \mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k, \quad \mathbf{d}_\mathbf{y}^k := \mathbf{y}^{k+1} - \mathbf{y}^k \quad \text{and} \quad \mathbf{d}_\lambda^k := \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k,$$

and define

$$\psi_1(s) = \max\left\{1, \frac{s^2}{(2-s)^2}\right\} \quad \text{and} \quad \psi_2(s) = \max\left\{\frac{1-s}{s}, \frac{s-1}{2-s}\right\}. \tag{4.3}$$

It is easy to see that $\psi_1(s) > 0$ and $\psi_2(s) \geq 0$ for any $s \in (0, 2)$. Then, we have the following lemma.

Lemma 4.1 *Suppose the Assumption 4.1 holds and the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 3.1 satisfy the condition (3.6). Then, for all $k \geq 1$, we have*

$$\begin{aligned}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \leq{}& \psi_2(s)\left(\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2\right) + 2\psi_1(s)(L + c_\mathbf{x}\beta)^2\|\widehat{\mathbf{d}}_\mathbf{x}^k\|^2 \\ &+8\psi_1(s)L^2\|\widetilde{\mathbf{d}}_\mathbf{x}^{k-1}\|^2 + 8\psi_1(s)c_\mathbf{x}^2\beta^2\left(\|\widehat{\mathbf{d}}_\mathbf{x}^{k-1}\|^2 + \|\mathbf{d}_\mathbf{y}^k\|^2 + \|\mathbf{d}_y^{k-1}\|^2\right).\end{aligned} \tag{4.4}$$

Proof By the definition of $\xi_\mathbf{x}^{k+1} = \nabla_\mathbf{x}\mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$, we have

$$\xi_\mathbf{x}^{k+1} = \nabla f(\widehat{\mathbf{x}}^k) + A^\mathsf{T}\big(-\boldsymbol{\lambda}^k + \beta\widehat{\mathbf{r}}^{k+1}\big),$$

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

$$A^\mathsf{T}\boldsymbol{\lambda}^k = \nabla f(\widehat{\mathbf{x}}^k) - \xi_\mathbf{x}^{k+1} + \beta A^\mathsf{T}\widehat{\mathbf{r}}^{k+1},$$

which follows from $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\widehat{\mathbf{r}}^{k+1}$ that

$$sA^{\mathsf{T}}\boldsymbol{\lambda}^k = s\big(\nabla f(\widehat{\mathbf{x}}^k) - \xi_{\mathbf{x}}^{k+1}\big) + A^{\mathsf{T}}\big(\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\big).$$

So, we have

$$A^{\mathsf{T}}\boldsymbol{\lambda}^{k+1} = s\left(\nabla f(\widehat{\mathbf{x}}^k) - \xi_{\mathbf{x}}^{k+1}\right) + (1-s)A^{\mathsf{T}}\boldsymbol{\lambda}^k,$$

which by $\mathbf{d}_{\lambda}^k = \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k$ gives

$$A^{\mathsf{T}}\mathbf{d}_{\lambda}^k = s\boldsymbol{\delta}^k + (1-s)A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}, \tag{4.5}$$

where

$$\boldsymbol{\delta}^k = \nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1}) - \xi_{\mathbf{x}}^{k+1} + \xi_{\mathbf{x}}^k. \tag{4.6}$$

In the following we consider two cases on $s \in (0, 1]$ or $s \in (1, 2)$.

Case 1: $s \in (0, 1]$. It follows from (4.5) and the convexity of $\|\cdot\|^2$ that

$$\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 \le s\|\boldsymbol{\delta}^k\|^2 + (1-s)\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2.$$

By subtracting $(1-s)\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2$ and dividing s from both sides of the above inequality, we derive

$$\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 \le \|\boldsymbol{\delta}^k\|^2 + \frac{1-s}{s}\left(\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2\right). \tag{4.7}$$

Case 2: $s \in (1, 2)$. It follows from (4.5) that

$$\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 = (1-s)^2\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 + s^2\|\boldsymbol{\delta}^k\|^2 + 2s(1-s)\langle A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}, \boldsymbol{\delta}^k\rangle. \tag{4.8}$$

Then, by (4.8) and Cauchy-Schwartz inequality, for an $\nu > 0$ we have

$$\begin{aligned}\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 &\le (1-s)^2\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 + s^2\|\boldsymbol{\delta}^k\|^2 + s(s-1)\left(\nu\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 + \frac{1}{\nu}\|\boldsymbol{\delta}^k\|^2\right)\\ &= \left[(1-s)^2 + s(s-1)\nu\right]\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 + \left(s^2 + \frac{s(s-1)}{\nu}\right)\|\boldsymbol{\delta}^k\|^2. \end{aligned}\tag{4.9}$$

By choosing $\nu = (2-s)/s$, we have

$$(1-s)^2 + s(s-1)\nu = s-1 \qquad \text{and} \qquad s^2 + \frac{s(s-1)}{\nu} = \frac{s^2}{2-s}.$$

So, we have from from (4.9) that

$$\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 \le (s-1)\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 + \frac{s^2}{2-s}\|\boldsymbol{\delta}^k\|^2.$$

By subtracting $(s-1)\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2$ and dividing $2-s$ from both sides of the above inequality, we derive

$$\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 \le \frac{s^2}{(2-s)^2}\|\boldsymbol{\delta}^k\|^2 + \frac{s-1}{2-s}\Big(\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2\Big). \tag{4.10}$$

Now, combining (4.7) and (4.10) and noticing the definition of functions ψ_1 and ψ_2 in (4.3), we have

$$\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 \le \psi_1(s)\|\boldsymbol{\delta}^k\|^2 + \psi_2(s)\Big(\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2\Big). \tag{4.11}$$

In addition, by (3.6), (4.1), $\widehat{\mathbf{x}}^k - \widehat{\mathbf{x}}^{k-1} = \widehat{\mathbf{d}}_{\mathbf{x}}^k + \widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}$ and the definition of $\boldsymbol{\delta}^k$ in (4.6), we have

$$\begin{aligned}
\|\boldsymbol{\delta}^k\|^2 &= \|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1}) - \xi_{\mathbf{x}}^{k+1} + \xi_{\mathbf{x}}^k\|^2 \\
&\le \Big(L\|\widehat{\mathbf{d}}_{\mathbf{x}}^k + \widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + c_{\mathbf{x}}\beta\Big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\| + \|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^k\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|\Big)\Big)^2 \\
&\le \Big[(L + c_{\mathbf{x}}\beta)\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\| + L\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + c_{\mathbf{x}}\beta\Big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^k\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|\Big)\Big]^2 \\
&\le 2(L + c_{\mathbf{x}}\beta)^2\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 + 8L^2\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 + 8c_{\mathbf{x}}^2\beta^2\Big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 + \|\mathbf{d}_{\mathbf{y}}^k\|^2 + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|^2\Big).
\end{aligned} \tag{4.12}$$

Finally, the conclusion (4.4) follows from the above inequality and (4.11). □

Now, let us denote $\mathbf{w}^k = (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$, $\widehat{\mathbf{w}}^k = (\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ and define the *potential energy* functions as

$$\widehat{E}^{k+1} = \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) + \widehat{\Gamma}^k \quad \text{and} \quad E^{k+1} = \mathcal{L}_\beta(\mathbf{w}^{k+1}) + \Gamma^k, \tag{4.13}$$

where

$$\widehat{\Gamma}^k = \frac{8(1+\tau)\psi_1(s)c_{\mathbf{x}}^2\beta}{s\sigma_A}\Big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 + \|\mathbf{d}_{\mathbf{y}}^k\|^2\Big) + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2,$$

$$\Gamma_k = \widehat{\Gamma}^k + \frac{8(1+\tau)\psi_1(s)L^2}{s\beta\sigma_A}\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2$$

and τ is any constant satisfying $0 < \tau < \delta < 1$. Then, based on the previous lemma, we can derive the following potential energy reduction theorem.

Theorem 4.1 *Suppose the Assumptions 4.1–4.2 hold and the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 3.1 satisfy the conditions (3.3), (3.5) and (3.6). For any $\delta \in (0,1)$, let $\tau \in (0,\delta)$ be the constant in the potential energies E^k and $\widehat{E}^k$ defined in (4.13). If the parameters in Algorithm 3.1 are chosen such that*

$$\widehat{\mathcal{D}}_{\mathbf{x}} := \frac{1-\tau}{2(1+\tau)}\mathcal{D}_{\mathbf{x}} - \frac{\psi_1(s)\left[2(L/\beta + c_{\mathbf{x}})^2 + 8c_{\mathbf{x}}^2\right]}{s\sigma_A}\mathbf{I} \succeq \mathbf{0}, \tag{4.14}$$

$$\overline{\mathcal{D}}_{\mathbf{y}} := \frac{1-\tau}{16(1+\tau)}\mathcal{D}_{\mathbf{y}} - \frac{\psi_1(s)c_{\mathbf{x}}^2}{s\sigma_A}\mathbf{I} \succeq \mathbf{0}, \tag{4.15}$$

and

$$\widetilde{\mathcal{D}}_{\mathbf{x}} := \left(\frac{\delta-\tau}{1+\tau} - \frac{8\psi_1(s)(L/\beta)^2}{s\sigma_A}\right)\mathbf{I} \succeq \mathbf{0}. \tag{4.16}$$

Then, for all $k \geq 1$, we have

$$E^{k+1} \leq E^k - \frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2 - \frac{\tau\beta}{2}\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2 - \tau\beta\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2 \tag{4.17}$$

and

$$\widehat{E}^{k+1} \leq \widehat{E}^k - \frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2 - \frac{\tau\beta}{2}\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2 - \tau\beta\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2. \tag{4.18}$$

Proof First, by (3.3), (3.5) and (4.2), we have

$$\begin{aligned}
&\mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k)\\
&= \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)\\
&\quad -\mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) - \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)\\
&\leq \frac{1+\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2 - \frac{\beta}{2}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2 - \frac{\beta}{2}\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2\\
&\leq \frac{1+\tau}{s\beta\sigma_A}\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 - \frac{\beta}{2}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2 - \frac{\beta}{2}\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2.
\end{aligned} \tag{4.19}$$

In addition, by (4.4), we obtain

$$\begin{aligned}
&\frac{1+\tau}{s\beta\sigma_A}\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2 \\
&\leq \frac{(1+\tau)\psi_1(s)}{s\beta\sigma_A}\Big[2(L+c_{\mathbf{x}}\beta)^2\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 + 8c_{\mathbf{x}}^2\beta^2\left(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 + \|\mathbf{d}_{\mathbf{y}}^k\|^2 + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|^2\right)\\
&\quad +8L^2\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2\Big] + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\left(\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2\right).
\end{aligned} \tag{4.20}$$

Then, plugging (4.20) into (4.19), by (3.8) and $\widetilde{\mathbf{d}}_{\mathbf{x}}^k = \mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k$, we have

$$\begin{aligned}
&\mathcal{L}_\beta(\mathbf{w}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k)\\
&\leq \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2\\
&\leq \frac{8(1+\tau)\psi_1(s)c_{\mathbf{x}}^2\beta}{s\sigma_A}\left(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 - \|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|^2 - \|\mathbf{d}_{\mathbf{y}}^k\|^2\right)\\
&\quad + \frac{8(1+\tau)\psi_1(s)L^2}{s\beta\sigma_A}\left(\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 - \|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2\right)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2-\frac{\tau\beta}{2}\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2-\frac{\tau}{s\beta}\|\mathbf{d}_{\lambda}^k\|^2-\tau\beta\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2\\
&-(1+\tau)\beta\left(\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\widehat{\mathcal{D}}_{\mathbf{x}}}^2+8\|\mathbf{d}_{\mathbf{y}}^k\|_{\overline{\mathcal{D}}_{\mathbf{y}}}^2+\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|_{\widetilde{\mathcal{D}}_{\mathbf{x}}}^2\right)\\
&+\frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\left(\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\|^2-\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2\right),
\end{aligned}
\tag{4.21}
$$

where $0<\tau<\delta<1$, $\widehat{\mathcal{D}}_{\mathbf{x}}\succeq\mathbf{0}$, $\overline{\mathcal{D}}_{\mathbf{y}}\succeq\mathbf{0}$ and $\widetilde{\mathcal{D}}_{\mathbf{x}}$ are defined in (4.14), (4.15) and (4.16), respectively. Then, (4.17) follows from (4.21) and the definition of E^{k+1} in (4.13). Similarly, by (3.8) and $\widetilde{\mathbf{d}}_{\mathbf{x}}^k=\mathbf{x}^k-\widehat{\mathbf{x}}^{k-1}$, we have

$$
\begin{aligned}
\mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1})-\mathcal{L}_\beta(\widehat{\mathbf{w}}^k)&\le\mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1})-\mathcal{L}_\beta(\mathbf{w}^k)-\delta\beta\|\mathbf{x}^k-\widehat{\mathbf{x}}^{k-1}\|^2\\
&=\mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1})-\mathcal{L}_\beta(\mathbf{w}^k)-\delta\beta\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2.
\end{aligned}
$$

So, plugging (4.20) into (4.19), we can similarly derive by the definition of $\widehat{E}^{k+1}$ in (4.13) that (4.18) holds. □

4.2 Global convergence and sublinear convergence rate

We say $\mathbf{w}^*=(\mathbf{x}^*,\mathbf{y}^*,\boldsymbol{\lambda}^*)$ is a stationary point of the problem (1.1) if $\mathbf{0}\in\partial\mathcal{L}(\mathbf{w}^*)$, i.e.,

$$
\mathbf{0}=\nabla f(\mathbf{x}^*)-A^{\mathsf{T}}\boldsymbol{\lambda}^*,\quad \mathbf{0}\in\partial g(\mathbf{y}^*)-B^{\mathsf{T}}\boldsymbol{\lambda}^*\quad\text{and}\quad A\mathbf{x}^*+B\mathbf{y}^*=\mathbf{b}. \tag{4.22}
$$

Then, it is obvious that $\mathbf{w}^k=(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k)$ is a stationary point of (1.1) if $R^{k+1}=0$, where R^k is defined in (3.7). Hence, in the following global convergence theorem, we assume $R^k\neq0$ for all k and an infinite sequence $\{\mathbf{w}^k\}$ is generated by Algorithm 3.1. And, in the following, we denote

$$
\mathbf{r}^k:=A\mathbf{x}^k+B\mathbf{y}^k-\mathbf{b}\quad\text{and}\quad\mathbf{d}_{\mathbf{x}}^k:=\mathbf{x}^{k+1}-\mathbf{x}^k=\widetilde{\mathbf{d}}_{\mathbf{x}}^k+\widehat{\mathbf{d}}_{\mathbf{x}}^k. \tag{4.23}
$$

Theorem 4.2 *Suppose the Assumptions 4.1–4.2 hold and the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 3.1 satisfy the conditions (3.3), (3.4), (3.5) and (3.6). If the parameters in Algorithm 3.1 are chosen such that (4.14), (4.15) and (4.16) hold, and $\{\widehat{E}^k\}$ defined in (4.13) is bounded from below, then there exists a F^* such that*

$$
\lim_{k\to\infty}\mathcal{L}(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k)=\lim_{k\to\infty}\mathcal{L}_\beta(\mathbf{x}^k,\mathbf{y}^k,\boldsymbol{\lambda}^k)=\lim_{k\to\infty}E^k=\lim_{k\to\infty}\widehat{E}^k=F^*. \tag{4.24}
$$

In addition, we have

$$
\lim_{k\to\infty}dist(\mathbf{0},\partial\mathcal{L}(\mathbf{w}^k))=\lim_{k\to\infty}dist(\mathbf{0},\partial\mathcal{L}_\beta(\mathbf{w}^k))=0 \tag{4.25}
$$

and any limit point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$ is a stationary point of the problem (1.1).

Proof If $\{\widehat{E}^k\}$ is bounded from below, we obtain from (4.18) that

$$\begin{aligned} & c\sum_{k=1}^{K}\left\{\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|_{\mathcal{D}_{\mathbf{x}}}^2+\|\mathbf{d}_{\mathbf{y}}^k\|_{\mathcal{D}_{\mathbf{y}}}^2+\|\mathbf{d}_{\lambda}^k\|^2+\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2\right\} \\ & \leq \widehat{E}^1-\widehat{E}^{K+1}\leq \widehat{E}^1-\overline{P}, \end{aligned} \tag{4.26}$$

where $c=\min\{\tau\beta/2,\tau/(s\beta)\}>0$ and $\overline{P}$ is the lower bound of E^k. Then, (4.26), $\mathcal{D}_{\mathbf{x}}\succ\mathbf{0}$ and $\mathcal{D}_{\mathbf{y}}\succ\mathbf{0}$ imply that

$$\lim_{k\to\infty}\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|=0,\quad \lim_{k\to\infty}\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|=0,\quad \lim_{k\to\infty}\|\mathbf{d}_{\mathbf{y}}^k\|=0\quad\text{and}\quad \lim_{k\to\infty}\|\mathbf{d}_{\lambda}^k\|=0. \tag{4.27}$$

In addition, by (4.27), $\mathbf{d}_{\lambda}^k=-s\beta\widehat{\mathbf{r}}^{k+1}$ and the definition of R^k in (3.7), we have

$$\lim_{k\to\infty}\|\widehat{\mathbf{r}}^k\|=0\quad\text{and}\quad \lim_{k\to\infty}R^k=\lim_{k\to\infty}(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|+\|\mathbf{d}_{\mathbf{y}}^{k-1}\|+\|\widehat{\mathbf{r}}^k\|)=0. \tag{4.28}$$

So, we have from $\mathbf{r}^k=\widehat{\mathbf{r}}^k+A\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}$, $\|\mathbf{d}_{\mathbf{x}}^k\|\leq\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|+\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|$ (4.27) and (4.28) that

$$\lim_{k\to\infty}\|\mathbf{r}^k\|=0\qquad\text{and}\qquad\lim_{k\to\infty}\|\mathbf{d}_{\mathbf{x}}^k\|=0, \tag{4.29}$$

where $\mathbf{r}^k$ and $\mathbf{d}_{\mathbf{x}}^k$ are defined in (4.23). By (4.17), we have $\{\widehat{E}^k\}_{k=1}^{\infty}$ is a monotonically nonincreasing sequence, which together with the assumption that $\{\widehat{E}^k\}$ being bounded from below implies $\lim_{k\to\infty}\widehat{E}^k=F^*$ for some F^*. Then, it follows from the definition of E^k, (4.27) and (4.29) that (4.24) holds.

Now, by direct calculation, we have

$$\begin{aligned} \partial_{\mathbf{x}}\mathcal{L}_\beta(\mathbf{w}^k) &= \partial_{\mathbf{x}}\mathcal{L}(\mathbf{w}^k)+\beta A^{\mathsf{T}}\mathbf{r}^k=\nabla f(\mathbf{x}^k)-A^{\mathsf{T}}\boldsymbol{\lambda}^k+\beta A^{\mathsf{T}}\mathbf{r}^k \\ &= \nabla_{\mathbf{x}}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^{k-1})-A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}+(\nabla f(\mathbf{x}^k)-\nabla f(\widehat{\mathbf{x}}^{k-1})), \\ \partial_{\mathbf{y}}\mathcal{L}_\beta(\mathbf{w}^k) &= \partial_{\mathbf{y}}\mathcal{L}(\mathbf{w}^k)+\beta B^{\mathsf{T}}\mathbf{r}^k=\partial_{\mathbf{y}}g(\mathbf{y}^k)-B^{\mathsf{T}}\boldsymbol{\lambda}^k+\beta B^{\mathsf{T}}\mathbf{r}^k \\ &= \partial_{\mathbf{y}}\mathcal{L}_\beta(\mathbf{x}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^{k-1})-B^{\mathsf{T}}(\mathbf{d}_{\lambda}^{k-1}-\beta A\mathbf{d}_{\mathbf{x}}^{k-1}), \\ \partial_{\lambda}\mathcal{L}_\beta(\mathbf{w}^k) &= \partial_{\lambda}\mathcal{L}(\mathbf{w}^k)=-\mathbf{r}^k. \end{aligned} \tag{4.30}$$

Then, it follows from (3.4), (3.6), (4.27) and (4.29) that (4.25) holds. In addition, for any limiting point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$, it follows from (4.25) and the definition of the *limiting-subdifferential* $\partial\mathcal{L}(\mathbf{w}^*)$ that (4.22) holds. Hence, $\mathbf{w}^*$ is a stationary point of (1.1). □

From Theorem 4.2 and (4.28), we can see that for any limiting stationary point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$, we have $\mathcal{L}(\mathbf{x}^*,\mathbf{y}^*,\boldsymbol{\lambda}^*)=F(\mathbf{x}^*,\mathbf{y}^*)=f(\mathbf{x}^*)+g(\mathbf{y}^*)=F^*$. In addition, we can observe from (4.26) that

$$\min_{k\in\{1,\ldots,K\}}\left\{\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2+\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2+\|\mathbf{d}_{\mathbf{y}}^k\|^2+\|\widehat{\mathbf{r}}^{k+1}\|^2\right\}=\mathcal{O}(1/K),$$

which together with (3.4) and (3.6) implies

$$\min_{k\in\{1,\ldots,K\}}\left\{\mathrm{dist}(\mathbf{0},\partial\mathcal{L}(\mathbf{w}^k))\right\}=\mathcal{O}(1/\sqrt{K}).$$

In Theorem 4.2, we assume the parameters in Algorithm 3.1 are chosen such that the potential energy sequence $\{\widehat{E}^k\}$ is uniformly bounded from below. The following theorem gives a sufficient condition to ensure the uniform lower bound of $\{\widehat{E}^k\}$, which in turn also implies the uniform lower bound of $\{E^k\}$ since $\lim_{k\to\infty}\|\widehat{\mathbf{d}}_k\|=\lim_{k\to\infty}\|\widetilde{\mathbf{d}}_k\|=0$.

Theorem 4.3 *Suppose there exists a constant $\overline{\beta}>0$ such that*

$$\inf\left\{f(\widehat{\mathbf{x}}^{k-1})+g(\mathbf{y}^k)+\frac{\overline{\beta}}{2}\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b}\|^2\right\}=:\overline{P}>-\infty. \tag{4.31}$$

Then, under the conditions of Theorem 4.1 and $\beta\geq\overline{\beta}$, we have $\widehat{E}^k\geq\overline{P}$ for all $k\geq 1$.

Proof Since $\beta\geq\overline{\beta}$, it follows from $\boldsymbol{\lambda}^k=\boldsymbol{\lambda}^{k-1}-s\beta(A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b})$ and (4.31) that

$$\begin{aligned}
&\mathcal{L}_\beta(\widehat{\mathbf{w}}^k)=\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)\\
&\geq f(\widehat{\mathbf{x}}^{k-1})+g(\mathbf{y}^k)-(\boldsymbol{\lambda}^k)^{\mathsf{T}}(A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b})+\frac{\overline{\beta}}{2}\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b}\|^2\\
&\geq\overline{P}+\frac{1}{s\beta}(\boldsymbol{\lambda}^k)^{\mathsf{T}}(\boldsymbol{\lambda}^k-\boldsymbol{\lambda}^{k-1})\\
&=\overline{P}+\frac{1}{2s\beta}\left(\|\boldsymbol{\lambda}^k\|^2-\|\boldsymbol{\lambda}^{k-1}\|^2+\|\boldsymbol{\lambda}^k-\boldsymbol{\lambda}^{k-1}\|^2\right).
\end{aligned}$$

Hence, by the definition of $\widehat{E}^k$ in (4.13) and the above inequality, we have

$$\sum_{k=1}^{\infty}\left(\widehat{E}^k-\overline{P}\right)\geq\sum_{k=1}^{\infty}\left(\mathcal{L}_\beta(\widehat{\mathbf{w}}^k)-\overline{P}\right)\geq-\frac{1}{s\beta}\|\boldsymbol{\lambda}^0\|^2. \tag{4.32}$$

By Theorem 4.1, $\{\widehat{E}^k\}_{k=1}^{\infty}$ is monotonically decreasing. So, if there exists a $\overline{k}\geq 1$ such that $\widehat{E}^{\overline{k}}<\overline{P}$, we will have $\widehat{E}^k<\overline{P}$ for all $k>\overline{k}$, which implies $\sum_{k=1}^{\infty}\left(\widehat{E}^k-\overline{P}\right)=-\infty$. This will contradict (4.32). Hence, we have $\widehat{E}^k\geq\overline{P}$ for all k. □

Remark 4.1 The condition (4.31) in Theorem 4.3 is obviously satisfied if

$$\inf\ f(\mathbf{x})+g(\mathbf{y})+\frac{\overline{\beta}}{2}\|A\mathbf{x}+B\mathbf{y}-\mathbf{b}\|^2>-\infty \tag{4.33}$$

for all $\mathbf{x}$ and $\mathbf{y}$. And in many applications, the function $F(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{y})$ is uniformly bounded from below and therefore, (4.33) holds. For example, in statistical learning both the graph-guided fused lasso model [42] and the smoothly clipped absolute deviation (SCAD) model [66] have nonnegative objective function value.

4.3 Linear convergence rate

In this subsection, we discuss the linear convergence of $\{E^k\}$ and $\{\mathbf{w}^k\}$ under proper conditions. Let Ω^* be the set of all stationary points of the problem (1.1) satisfying (4.22), i.e.,

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

Note that Ω^* is a closed set. In the following, let us denote $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*) \in \Omega^*$. For studying linear convergence, we need the following additional assumption.

Assumption 4.3 (a) For any $\xi \geq \inf_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w})$, there exist $\epsilon > 0$ and $\kappa > 0$ such that

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

whenever $\operatorname{dist}(\mathbf{0}, \partial \mathcal{L}_\beta(\mathbf{w})) \leq \epsilon$ and $\mathcal{L}_\beta(\mathbf{w}) \leq \xi$.
(b) Ω^* is nonempty and there exists $\omega^* > 0$ such that $\|\mathbf{w}_1 - \mathbf{w}_2\| \geq \omega^*$ whenever $\mathbf{w}_1$, $\mathbf{w}_2 \in \Omega^*$ and $F(\mathbf{x}_1, \mathbf{y}_1) \neq F(\mathbf{x}_2, \mathbf{y}_2)$.
(c) Function g is locally weakly convex near

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

that is, there exist $\varepsilon, \sigma > 0$ such that for any $\mathbf{y}_1, \mathbf{y}_2$ with $\operatorname{dist}(\mathbf{y}_1, \Omega^*_{\mathbf{y}}) \leq \epsilon$, $\operatorname{dist}(\mathbf{y}_2, \Omega^*_{\mathbf{y}}) \leq \epsilon$ and $\|\mathbf{y}_1 - \mathbf{y}_2\| \leq \varepsilon$ and for any $\boldsymbol{\nu} \in \partial g(\mathbf{y}_2)$, it has

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

We have the following comments on Assumption 4.3. Assumption 4.3 (a) is a local error bound condition and [57, Lemma 7] provides certain sufficient conditions to ensure this assumption when analyzing linear convergence rate of a nonconvex algorithm. Similar local error bound conditions have been often used in the convergence rate analysis of many algorithms [7, 45, 50, 51, 58, 62]. Assumption 4.3 (b) essentially requires that the isocost surface of F restricted on Ω^* are properly separated. For more examples and discussions on functions satisfying the error bound conditions and the isocost properties, one may refer to references [57, 58, 62, 70]. Assumption 4.3 (c) requires that g is locally weakly convex near the projection of the stationary point set Ω onto the $\mathbf{y}$-coordinates. Convex functions and Lipschitz continuously differential functions obviously satisfies this requirement. For more properties on weakly convex functions as well as its relations to lower-$\mathcal{C}^2$ functions, one may refer to references [1, 52, 53, 59].

We now give the following linear convergence theorem on the energy sequence $\{E^k\}$. The linear convergence of energy sequence $\{\widehat{E}^k\}$ can be similarly proved.

Theorem 4.4 *Suppose the conditions in Theorem 4.2 and Assumption 4.3 hold. Then, for the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 3.1, we have*

(i) $\lim_{k\to\infty} dist(\mathbf{w}^k, \Omega^*) = 0$;
(ii) if $\{\mathbf{w}^k\}$ has at least one cluster point, then for all k sufficiently large,

$$0 \le E^{k+1} - F^* \le \theta(E^k - F^*), \tag{4.34}$$

where $\theta \in (0, 1)$ is some constant, E^k is defined in (4.13) and $F^* = \lim_{k\to\infty} E^k$ is defined in (4.24).

Proof By (4.24) and (4.25), there exists a $\zeta \ge \inf_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w})$ such that $\mathcal{L}_\beta(\mathbf{w}^k) \le \zeta$ for all k and $\lim_{k\to\infty} dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) = 0$. Hence, conclusion (i) follows from Assumption 4.3 (a) with $\xi = \zeta$.

We now prove conclusion (ii). For any iterate $\mathbf{w}^k$, let us define a $\overline{\mathbf{w}}^k \in \Omega^*$ such that $\text{dist}(\mathbf{w}^k, \Omega^*) = \|\mathbf{w}^k - \overline{\mathbf{w}}^k\|$. Since Ω^* is closed, such $\overline{\mathbf{w}}^k$ exists. Then, by conclusion (i), we have

$$\lim_{k\to\infty} \|\mathbf{w}^k - \overline{\mathbf{w}}^k\| = 0. \tag{4.35}$$

In addition, we have from (4.27) and $\|\mathbf{w}^k - \mathbf{w}^{k-1}\| \le \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\| + \|\mathbf{d}_{\lambda}^{k-1}\|$ that

$$\lim_{k\to\infty} \|\mathbf{w}^k - \mathbf{w}^{k-1}\| = 0. \tag{4.36}$$

Therefore, we have from $\|\overline{\mathbf{w}}^k - \overline{\mathbf{w}}^{k-1}\| \le \|\overline{\mathbf{w}}^k - \mathbf{w}^k\| + \|\mathbf{w}^k - \mathbf{w}^{k-1}\| + \|\mathbf{w}^{k-1} - \overline{\mathbf{w}}^{k-1}\|$, (4.35) and (4.36) that

$$\lim_{k\to\infty} \|\overline{\mathbf{w}}^k - \overline{\mathbf{w}}^{k-1}\| = 0.$$

So, by Assumption 4.3 (b) and $\overline{\mathbf{w}}^k \in \Omega$, there exists a constant $\overline{F}^*$ such that

$$\mathcal{L}_\beta(\overline{\mathbf{w}}^k) = \mathcal{L}_\beta(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k, \overline{\boldsymbol{\lambda}}^k) = F(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k) = \overline{F}^* \tag{4.37}$$

for all k sufficiently large. Now, suppose $\{\mathbf{w}^k\}$ has a cluster point $\mathbf{w}^*$, i.e., there exists a subsequence $\{\mathbf{w}^{k_i}\}$ converging to $\mathbf{w}^*$. Then, we have from Theorem 4.2 that $\mathbf{w}^* \in \Omega$, and in addition, by (4.35), we have

$$\lim_{i\to\infty} \|\overline{\mathbf{w}}^{k_i} - \mathbf{w}^*\| \le \lim_{i\to\infty} \left(\|\overline{\mathbf{w}}^{k_i} - \mathbf{w}^{k_i}\| + \|\mathbf{w}^{k_i} - \mathbf{w}^*\|\right) = 0.$$

Hence, we have from (4.37), $\mathbf{w}^* \in \Omega$ and Assumption 4.3 (b) again that $\mathcal{L}_\beta(\mathbf{w}^*) = \overline{F}^*$. So, by the lower semicontinuity of the function $\mathcal{L}_\beta(\cdot)$, we have

$$\overline{F}^* = \mathcal{L}_\beta(\mathbf{w}^*) \le \lim_{i\to\infty} \mathcal{L}_\beta(\mathbf{w}^{k_i}) = F^*, \tag{4.38}$$

where $F^* = \lim_{k\to\infty} E^k = \lim_{k\to\infty} \mathcal{L}_\beta(\mathbf{w}^k)$ is defined in Theorem 4.2.

By the definition of $\mathcal{L}_\beta(\mathbf{x},\mathbf{y},\boldsymbol{\lambda})$ in (1.3) and the update of $\boldsymbol{\lambda}^k$ in Algorithm 3.1, we have

$$\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda})=\frac{1}{s\beta}(\boldsymbol{\lambda}-\boldsymbol{\lambda}^k)^{\mathsf{T}}(\boldsymbol{\lambda}^{k-1}-\boldsymbol{\lambda}^k), \quad (4.39)$$

$$\begin{aligned}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda})-\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y},\boldsymbol{\lambda})&=g(\mathbf{y}^k)-g(\mathbf{y})+\boldsymbol{\lambda}^{\mathsf{T}}B(\mathbf{y}-\mathbf{y}^k) \quad (4.40)\\ &+\frac{\beta}{2}\left(\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b}\|^2-\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}-\mathbf{b}\|^2\right),\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y},\boldsymbol{\lambda})-\mathcal{L}_\beta(\mathbf{x},\mathbf{y},\boldsymbol{\lambda})&=f(\widehat{\mathbf{x}}^{k-1})-f(\mathbf{x})+\boldsymbol{\lambda}^{\mathsf{T}}A(\mathbf{x}-\widehat{\mathbf{x}}^{k-1})\\ &+\frac{\beta}{2}\left(\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}-\mathbf{b}\|^2-\|A\mathbf{x}+B\mathbf{y}-\mathbf{b}\|^2\right). \quad (4.41)\end{aligned}$$

Then, by setting $(\mathbf{x},\mathbf{y},\boldsymbol{\lambda})=\overline{\mathbf{w}}^k$ in (4.39), (4.40) and (4.41), for all k sufficiently large, we have from (4.37) and (4.38) that

$$\begin{aligned}&\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-F^*\\ &\le\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-\overline{F}^*=\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_\beta(\overline{\mathbf{x}}^k,\overline{\mathbf{y}}^k,\overline{\boldsymbol{\lambda}}^k)\\ &\le\frac{1}{s\beta}(\overline{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k)^{\mathsf{T}}(\boldsymbol{\lambda}^{k-1}-\boldsymbol{\lambda}^k)+\frac{L}{2}\|\overline{\mathbf{x}}^k-\widehat{\mathbf{x}}^{k-1}\|^2\\ &+\frac{1}{2s^2\beta}\|\mathbf{d}_\lambda^{k-1}\|^2+g(\mathbf{y}^k)-g(\overline{\mathbf{y}}^k)+\langle B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k,\overline{\mathbf{y}}^k-\mathbf{y}^k\rangle, \quad (4.42)\end{aligned}$$

where the inequality comes from Lipschitz continuity of f, $A^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k=\nabla f(\overline{\mathbf{x}}^k)$, $A\overline{\mathbf{x}}^k+B\overline{\mathbf{y}}^k=\mathbf{b}$ and $\mathbf{d}_\lambda^{k-1}=-s\beta\widehat{\mathbf{r}}^k$. From (3.4), there exists a $\xi_{\mathbf{y}}^k\in\partial_{\mathbf{y}}\mathcal{L}_\beta(\mathbf{x}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^{k-1})$, i.e.,

$$\boldsymbol{\nu}^k:=\xi_{\mathbf{y}}^k+B^{\mathsf{T}}\boldsymbol{\lambda}^{k-1}-\beta B^{\mathsf{T}}(A\mathbf{x}^{k-1}+B\mathbf{y}^k-\mathbf{b})\in\partial g(\mathbf{y}^k)$$

with $\|\xi_{\mathbf{y}}^k\|\le c_{\mathbf{y}}\beta\|\mathbf{d}_{\mathbf{y}}^{k-1}\|$. So, we have

$$\begin{aligned}\|\boldsymbol{\nu}^k-B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k\|&\le\|\xi_{\mathbf{y}}^k\|+\|B^{\mathsf{T}}(\boldsymbol{\lambda}^{k-1}-\overline{\boldsymbol{\lambda}}^k)\|+\beta\|B^{\mathsf{T}}(A\mathbf{x}^{k-1}+B\mathbf{y}^k-\mathbf{b})\|\\ &\le c_{\mathbf{y}}\beta\|\mathbf{d}_{\mathbf{y}}^{k-1}\|+\|B\|\big(\|\mathbf{d}_\lambda^{k-1}\|+\|\boldsymbol{\lambda}^k-\overline{\boldsymbol{\lambda}}^k\|\big)+\beta\|B\|\big(\|\widehat{\mathbf{r}}^k\|+\|A\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|\big).\end{aligned} \quad (4.43)$$

Now, by (4.35), we have $\lim_{k\to\infty}\|\mathbf{y}_k-\overline{\mathbf{y}}_k\|=0$ and $\lim_{k\to\infty}\operatorname{dist}(\mathbf{y}^k,\Omega_y)=0$. Hence, it follows from Assumption 4.3 (c) that

$$g(\overline{\mathbf{y}}^k)\ge g(\mathbf{y}^k)+\langle\boldsymbol{\nu}^k,\overline{\mathbf{y}}^k-\mathbf{y}^k\rangle-\sigma\|\overline{\mathbf{y}}^k-\mathbf{y}^k\|^2$$

for all k sufficiently large, where $\sigma > 0$ is a constant, which implies

$$\begin{aligned} & g(\mathbf{y}^k) - g(\overline{\mathbf{y}}^k) + \langle B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k\rangle \\ &= g(\mathbf{y}^k) - g(\overline{\mathbf{y}}^k) + \langle \boldsymbol{\nu}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k\rangle + \langle B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k\rangle \\ &\le \sigma\|\overline{\mathbf{y}}^k - \mathbf{y}^k\|^2 + \|B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k\|\|\overline{\mathbf{y}}^k - \mathbf{y}^k\|. \end{aligned}$$

Hence, by (4.42), (4.43), $\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}^{k-1}\|^2 \le 2\big(\|\overline{\mathbf{x}}^k - \mathbf{x}^k\|^2 + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2\big)$ and $\mathbf{d}_{\lambda}^{k-1} = -s\beta\widehat{\mathbf{r}}^k$, there exit two constants $c_1 > 0$ and $c_2 > 0$ such that

$$\begin{aligned} \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) - F^* &\le \frac{1}{s\beta}(\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k)^{\mathsf{T}}(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k) + \frac{L}{2}\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}^{k-1}\|^2 \\ &\quad + \frac{1}{2s^2\beta}\|\mathbf{d}_{\lambda}^{k-1}\|^2 + \sigma\|\overline{\mathbf{y}}^k - \mathbf{y}^k\|^2 + \|B^{\mathsf{T}}\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k\|\|\overline{\mathbf{y}}^k - \mathbf{y}^k\| \\ &\le c_1\big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|^2 + \|\mathbf{d}_{\lambda}^{k-1}\|^2 + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2\big) + c_2\|\mathbf{w}^k - \overline{\mathbf{w}}^k\|^2 \end{aligned} \tag{4.44}$$

for all k sufficiently large. By (3.4), (3.6), (4.30), $\mathbf{d}_{\lambda}^{k-1} = -s\beta\widehat{\mathbf{r}}^k$, $\mathbf{r}^k = \widehat{\mathbf{r}}^k + A\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}$, and $\mathbf{d}_{\mathbf{x}}^{k-1} = \widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1} + \widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}$, we have

$$\begin{aligned} & dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) \\ &\le \|\nabla_{\mathbf{x}}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1}) - A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\| + \|\nabla f(\mathbf{x}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1})\| + \|\mathbf{r}^k\| \\ &\quad + dist\left(B^{\mathsf{T}}(\mathbf{d}_{\lambda}^{k-1} - \beta A\mathbf{d}_{\mathbf{x}}^{k-1}), \partial_{\mathbf{y}}\mathcal{L}_\beta(\mathbf{x}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1})\right) \\ &\le c_{\mathbf{x}}\beta(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|) + \|A^{\mathsf{T}}\mathbf{d}_{\lambda}^{k-1}\| + c_{\mathbf{y}}\beta\|\mathbf{d}_{\mathbf{y}}^{k-1}\| + \|B^{\mathsf{T}}(\mathbf{d}_{\lambda}^{k-1} - \beta A\mathbf{d}_{\mathbf{x}}^{k-1})\| \\ &\quad + L\|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \frac{1}{s\beta}\|\mathbf{d}_{\lambda}^{k-1}\| + \|A\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\| \\ &\le c_3(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\| + \|\mathbf{d}_{\lambda}^{k-1}\| + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|), \end{aligned}$$

where $c_3 = \max\big\{(c_{\mathbf{x}} + \|B^{\mathsf{T}}A\|)\beta, (c_{\mathbf{x}} + c_{\mathbf{y}})\beta, 1/(s\beta) + \|A\| + \|B\|, L + \|A\| + \beta\|B^{\mathsf{T}}A\|\big\} > 0$. So, by Assumption 4.3 (a), we have

$$\begin{aligned} \|\mathbf{w}^k - \overline{\mathbf{w}}^k\| = dist(\mathbf{w}^k, \Omega) &\le \kappa\, dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) \\ &\le \kappa c_3\big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\| + \|\mathbf{d}_{\mathbf{y}}^{k-1}\| + \|\mathbf{d}_{\lambda}^{k-1}\| + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|\big) \end{aligned}$$

for all k sufficiently large, which together with (4.44) gives

$$\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) - F^* \le \overline{c}\big(\|\widehat{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2 + \|\mathbf{d}_{\mathbf{y}}^{k-1}\|^2 + \|\mathbf{d}_{\lambda}^{k-1}\|^2 + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^{k-1}\|^2\big), \tag{4.45}$$

where $\overline{c} = c_1 + 4c_2c_3^2\kappa^2$. Hence, defining $d^k := \|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 + \|\mathbf{d}_{\mathbf{y}}^k\|^2 + \|\mathbf{d}_{\lambda}^k\|^2 + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2$, it follows from the definition of E^k in (4.13), (3.8) and (4.45) that

$$E^{k+1} - F^* \le \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2 - F^* + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\|A^{\mathsf{T}}\mathbf{d}_{\lambda}^k\|^2$$

$$+\frac{8(1+\tau)\psi_1(s)\beta}{s\sigma_A}\big(c_{\mathbf{x}}^2(\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2+\|\mathbf{d}_{\mathbf{y}}^k\|^2)+(L/\beta)^2\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2\big)$$
$$\le \gamma\, d^k, \tag{4.46}$$

where $\gamma = \bar{c}+\max\big\{(8(1+\tau)\psi_1(s)(c_{\mathbf{x}}^2\beta^2+L^2),\,(1+\tau)\|A\|^2\psi_2(s)\big\}/(s\beta\sigma_A)$. Additionally, we have by (4.17), $\mathcal{D}_{\mathbf{x}} \succ \mathbf{0}$ and $\mathcal{D}_{\mathbf{y}} \succ \mathbf{0}$ that $E^k \ge F^*$ for all $k \ge 1$ and

$$E^{k+1} \le E^k - \overline{\gamma} d^k, \tag{4.47}$$

where $\overline{\gamma} = \min\{\frac{\tau\beta}{2}\sigma_{\mathcal{D}_{\mathbf{x}}}, \frac{\tau\beta}{2}\sigma_{\mathcal{D}_{\mathbf{y}}}, \frac{\tau}{s\beta}, \tau\beta\} > 0$, $\sigma_{\mathcal{D}_{\mathbf{x}}} > 0$ and $\sigma_{\mathcal{D}_{\mathbf{y}}} > 0$ are the smallest eigenvalue of $\mathcal{D}_{\mathbf{x}}$ and $\mathcal{D}_{\mathbf{y}}$, respectively. Thus, by (4.46) and (4.47), for k sufficiently large, we have $0 \le E^{k+1} - F^* \le \theta(E^k - F^*)$, where $\theta = \gamma/(\gamma+\overline{\gamma}) \in (0,1)$. □

Based on the linear convergence result in the previous theorem, we can establish the following linear convergence of the sequence $\{\mathbf{w}^k\}$.

Theorem 4.5 *Suppose the conditions in Theorem 4.2 and Assumption 4.3 hold. If the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 3.1 has one cluster point, then $\{\mathbf{w}^k\}$ converges R-linearly to a stationary point of the problem (1.1).*

Proof We have from $\mathcal{D}_{\mathbf{x}}, \mathcal{D}_{\mathbf{y}} \succ \mathbf{0}$, (4.17) and $E^k \ge F^*$ for all $k \ge 1$ that

$$\begin{aligned}
\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\|^2 &\le \frac{2}{\tau\beta\sigma_{\mathcal{D}_{\mathbf{x}}}}(E^k - E^{k+1}) \le M_1(E^k - F^*),\\
\|\mathbf{d}_{\mathbf{y}}^k\|^2 &\le \frac{2}{\tau\beta\sigma_{\mathcal{D}_{\mathbf{y}}}}(E^k - E^{k+1}) \le M_1(E^k - F^*),\\
\|\mathbf{d}_{\lambda}^k\|^2 &\le \frac{s\beta}{\tau}(E^k - E^{k+1}) \le M_1(E^k - F^*),\\
\|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\|^2 &\le \frac{1}{\tau\beta}(E^k - E^{k+1}) \le M_1(E^k - F^*)
\end{aligned} \tag{4.48}$$

where $M_1 = \max\{2/(\tau\beta\sigma_{\mathcal{D}_{\mathbf{x}}}), 2/(\tau\beta\sigma_{\mathcal{D}_{\mathbf{y}}}), s\beta/\tau, 1/(\tau\beta)\}$. In addition, by Theorem 4.4, there exists a constant $M_2 > 0$ such that $0 \le E^k - F^* \le M_2\theta^k$ for all $k \ge 0$, where $\theta \in (0,1)$ is the constant in (4.34). Hence, it follows from (4.48) that

$$\|\widehat{\mathbf{d}}_{\mathbf{x}}^k\| \le Mq^k, \qquad \|\mathbf{d}_{\mathbf{y}}^k\| \le Mq^k, \qquad \|\mathbf{d}_{\lambda}^k\| \le Mq^k \quad \text{and} \quad \|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\| \le Mq^k,$$

where $M = \sqrt{M_1 M_2}$ and $q = \sqrt{\theta} \in (0,1)$. Therefore, we have

$$\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \le \|\widehat{\mathbf{d}}_{\mathbf{x}}^k\| + \|\widetilde{\mathbf{d}}_{\mathbf{x}}^k\| + \|\mathbf{d}_{\mathbf{y}}^k\| + \|\mathbf{d}_{\lambda}^k\| \le 4\,Mq^k.$$

Then, for any $m_2 > m_1 \ge 1$, we have

$$\|\mathbf{w}^{m_2} - \mathbf{w}^{m_1}\| \le \sum_{k=m_1}^{m_2-1}\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \le \frac{4\,M}{1-q}q^{m_1},$$

which implies the sequence $\{\mathbf{w}^k\}$ is a Cauchy sequence and hence convergent. Suppose $\{\mathbf{w}^k\}$ converges to $\mathbf{w}^*$. Letting $m_2 \to \infty$ in the above inequality, we have

$$\|\mathbf{w}^* - \mathbf{w}^{m_1}\| \leq \frac{4M}{1-q} q^{m_1},$$

which shows $\{\mathbf{w}^k\}$ converges R-linearly to $\mathbf{w}^*$. Finally, Theorem 4.2 ensures that $\mathbf{w}^*$ is a stationary point of (1.1). □

5 Inexact subproblem solution

Depending on various (e.g. smooth, convex and sparse) properties of the function g, one can design different algorithms to solve the $\mathbf{y}$-subproblem (3.1) inexactly to find $\mathbf{y}^{k+1}$ satisfying the conditions (3.3) and (3.4). Here, in this subsection, we just propose a gradient method with extrapolation to find an inexact solution satisfying (3.5) and (3.6) of the $\mathbf{x}$-subproblem. Note that the $\mathbf{x}$-subproblem (3.2) is equivalent to

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{n_\mathbf{x}}} \Phi^k(\mathbf{x}) &:= f(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathcal{D}_\mathbf{x}^k} + \mathbf{x}^\mathsf{T}\mathbf{p}^k + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{A^\mathsf{T}A} \\ &= h^k(\mathbf{x}) + \phi^k(\mathbf{x}), \end{aligned} \tag{5.1}$$

where $\mathbf{p}^k = -A^\mathsf{T}\big[\boldsymbol{\lambda}^k - \beta(A\mathbf{x}^k + B\mathbf{y}^{k+1} - \mathbf{b})\big]$, $\phi^k(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{p}^k + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{A^\mathsf{T}A}$ and

$$h^k(\mathbf{x}) = f(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathcal{D}_\mathbf{x}^k}. \tag{5.2}$$

In this section, we make the following assumptions.

Assumption 5.1 (a) The optimal value of the $\mathbf{x}$-subproblem is bounded from below, i.e., $\Phi^* = \min_{\mathbf{x}\in\mathbb{R}^{n_\mathbf{x}}} \Phi^k(\mathbf{x}) > -\infty$, where the function Φ^k is defined in (5.1).
(b) There exist constants $L_1 > 0$ and $L_2 > 0$ such that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_\mathbf{x}}$, it holds

$$-\frac{L_1}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \leq f(\mathbf{z}_2) - f(\mathbf{z}_1) - \langle \nabla f(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1\rangle \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2.$$

Obviously, by (4.1) we have $\max\{L_1, L_2\} \leq L$.

Assumption 5.2 The proximal matrix $\mathcal{D}_\mathbf{x}^k$ chosen in the $\mathbf{x}$-subproblem is positive definite and upper bounded, i.e.,

$$\overline{\eta}\mathbf{I} \succeq \mathcal{D}_\mathbf{x}^k \succeq \eta\mathbf{I} \qquad \text{for some } \overline{\eta} \geq \eta > 0. \tag{5.3}$$

Under Assumptions 5.1 and 5.2, it follows from the definition h^k in (5.2) that

$$-\frac{\mu}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \leq h^k(\mathbf{z}_2) - h^k(\mathbf{z}_1) - \langle \nabla h^k(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1\rangle \leq \frac{\Lambda}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \tag{5.4}$$

Initialization: Choose $\Theta > \Lambda$; Set $\check{\mathbf{x}}_1 = \mathbf{x}_1 = \mathbf{x}^k$ and $\tau = 1 - \sqrt{\frac{\Theta-\mu}{\Theta+\mu}}$.
For $t = 1, 2, 3, \ldots$
Set $\beta_t = \max\{\overline{\beta}_t, \tau\}$, where $\overline{\beta}_t = 2/(t+1)$.
$\widehat{\mathbf{x}}_t = \beta_t \check{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$.
Set $\gamma_t = \beta_t \Theta(t+1)/t$.
$\check{\mathbf{x}}_{t+1} = \arg\min \left\{ \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2} \|\mathbf{x} - \check{\mathbf{x}}_t\|^2 + \phi(\mathbf{x}) \right\}$.
$\mathbf{x}_{t+1} = \beta_t \check{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$.
end

Algorithm 5.1 A unified proximal gradient (UPG) method for solving $\mathbf{x}$-subproblem (5.1)

for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_\mathbf{x}}$, where $\mu = \max\{L_1 - \beta\eta, 0\}$ and $\Lambda = L_2 + \beta\overline{\eta}$.

Since we focus on solving the $\mathbf{x}$-subproblem, where the outer iteration number k is fixed, for notation simplicity, in the following of this section we simply denote Φ^k, h^k, ϕ^k and Λ^k as Φ, h, ϕ and Λ, respectively. Then, our algorithm for solving (5.1) is described in Algorithm 5.1, which is a generalization of the accelerated gradient method proposed in [33] for solving convex subproblems of ADMM to the case when f is not necessarily convex.

Theorem 5.1 *Suppose Assumptions 5.1 and 5.2 hold. Then, for the sequence* $\{\mathbf{x}_t\}$ *generated by Algorithm 5.1, we have*

$$\lim_{t\to\infty} \|\nabla\Phi(\mathbf{x}_t)\| = \lim_{t\to\infty} \|\nabla\Phi(\widehat{\mathbf{x}}_t)\| = 0. \tag{5.5}$$

Proof First, apparently, by the definitions in (5.4), we have $\Lambda > \mu \geq 0$ since $\eta > 0$. When $\mu = 0$, we have h is a convex function, and it follows from Algorithm 5.1 that $\tau = 0$ and $\beta_t = \overline{\beta}_t$ for all $t \geq 1$. In this case, Algorithm 5.1 will just reduce to a standard accelerated gradient method (see algorithms developed in [33, 34]) for solving convex composite optimization which guarantees $\lim_{t\to\infty} \Phi(\mathbf{x}_t) = \lim_{t\to\infty} \Phi(\widehat{\mathbf{x}}_t) = \Phi^* > -\infty$. Hence, (5.5) holds.

In the following, we discuss the convergence of Algorithm 5.1 when $\mu > 0$. From the updates of $\mathbf{x}_{t+1}$ and $\widehat{\mathbf{x}}_t$, we have

$$\beta_t(\check{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1-\beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) = \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t \mathbf{s}_t, \tag{5.6}$$

where $\mathbf{s}_t = \check{\mathbf{x}}_{t+1} - \check{\mathbf{x}}_t$. Then, by (5.4) and (5.6), the following relations hold

$$\begin{aligned} h(\mathbf{x}_{t+1}) &\leq h(\widehat{\mathbf{x}}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \rangle + \frac{\Lambda}{2} \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \\ &= h(\widehat{\mathbf{x}}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \rangle + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\Lambda\beta_t^2}{2} \|\mathbf{s}_t\|^2 \\ &\leq h(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\Lambda\beta_t^2}{2} \|\mathbf{s}_t\|^2. \end{aligned} \tag{5.7}$$

Furthermore, by (5.6), (5.7), $\mathbf{x}_{t+1} = \beta_t \check{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$ and the convexity of function ϕ, we have

$$\begin{aligned}
\Phi(\mathbf{x}_{t+1}) &= h(\mathbf{x}_{t+1}) + \phi(\mathbf{x}_{t+1}) \\
&\le \beta_t\big[h(\mathbf{x}_t) + \big\langle \nabla h(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x}_t\big\rangle + \phi(\check{\mathbf{x}}_{t+1})\big] + (1-\beta_t)\big[h(\mathbf{x}_t) + \phi(\mathbf{x}_t)\big] \\
&\quad + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2}{2}\|\mathbf{s}_t\|^2 \\
&= \beta_t\left[h(\mathbf{x}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + \phi(\check{\mathbf{x}}_{t+1})\right] \\
&\quad + (1-\beta_t)\Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2 - \gamma_t\beta_t}{2}\|\mathbf{s}_t\|^2. \qquad (5.8)
\end{aligned}$$

Now, it follows from

$$\check{\mathbf{x}}_{t+1} = \arg\min\left\{\langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}\rangle + \frac{\gamma_t}{2}\|\mathbf{x} - \check{\mathbf{x}}_t\|^2 + \phi(\mathbf{x})\right\}$$

and $\mathbf{s}_t = \check{\mathbf{x}}_{t+1} - \check{\mathbf{x}}_t$ that

$$\begin{aligned}
&\langle \nabla h(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + \phi(\check{\mathbf{x}}_{t+1}) \\
&\quad \le \frac{\gamma_t}{2}\big(\|\mathbf{x}_t - \check{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|^2\big) + \phi(\mathbf{x}_t) - \frac{1}{2}\|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|_{\mathcal{M}}^2, \qquad (5.9)
\end{aligned}$$

where $\mathcal{M} = \beta A^{\mathsf{T}} A$ and

$$\nabla h(\widehat{\mathbf{x}}_t) + \gamma_t \mathbf{s}_t + \nabla\phi(\check{\mathbf{x}}_{t+1}) = \mathbf{0}. \qquad (5.10)$$

By (5.8) and (5.9), we have

$$\begin{aligned}
\Phi(\mathbf{x}_{t+1}) &\le \beta_t\left[h(\mathbf{x}_t) + \frac{\gamma_t}{2}\left(\|\mathbf{x}_t - \check{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|^2\right) + \phi(\mathbf{x}_t) - \frac{1}{2}\|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|_{\mathcal{M}}^2\right] \\
&\quad + (1-\beta_t)\Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2 - \gamma_t\beta_t}{2}\|\mathbf{s}_t\|^2 \\
&\le \Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\beta_t\gamma_t}{2}\left(\|\mathbf{x}_t - \check{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|^2\right) \\
&\quad - \frac{\beta_t}{2}\|\mathbf{x}_t - \check{\mathbf{x}}_{t+1}\|_{\mathcal{M}}^2 - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2, \qquad (5.11)
\end{aligned}$$

where the last inequality follows from

$$\gamma_t\beta_t - \Lambda\beta_t^2 = \beta_t^2\Theta(t+1)/t - \Lambda\beta_t^2 \ge (\Theta - \Lambda)\beta_t^2.$$

Now, note that

$$\check{\mathbf{x}}_t - \mathbf{x}_t = \frac{1}{\beta_t}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) \quad \text{and} \quad \check{\mathbf{x}}_{t+1} - \mathbf{x}_t = \frac{1}{\beta_t}(\mathbf{x}_{t+1} - \mathbf{x}_t). \qquad (5.12)$$

Then, we have from (5.11) that

$$\Phi(\mathbf{x}_{t+1}) \le \Phi(\mathbf{x}_t) + \frac{\mu + \gamma_t/\beta_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\gamma_t/\beta_t}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \frac{\beta_t}{2}\|\check{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2. \tag{5.13}$$

For $t \ge 2$, by (5.12), we obtain

$$\begin{aligned}\widehat{\mathbf{x}}_t - \mathbf{x}_t &= \beta_t(\check{\mathbf{x}}_t - \mathbf{x}_t) = \beta_t(\check{\mathbf{x}}_t - \mathbf{x}_{t-1} + \mathbf{x}_{t-1} - \mathbf{x}_t)\\ &= \beta_t\left(\frac{1}{\beta_{t-1}}(\mathbf{x}_t - \mathbf{x}_{t-1}) + \mathbf{x}_{t-1} - \mathbf{x}_t\right)\\ &= \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1}),\end{aligned} \tag{5.14}$$

where $\theta_t = \frac{\beta_t}{\beta_{t-1}}(1 - \beta_{t-1})$. In addition, by defining $\beta_0 = 1$ and $\mathbf{x}_0 = \mathbf{x}_1$, we can see (5.14) holds for all $t \ge 1$. Hence, for $t \ge 1$ it follows from (5.13) that

$$\Phi(\mathbf{x}_{t+1}) \le \Phi(\mathbf{x}_t) + \frac{(\gamma_t/\beta_t + \mu)\theta_t^2}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\gamma_t/\beta_t}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \frac{\beta_t}{2}\|\check{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2. \tag{5.15}$$

Since $\gamma_t/\beta_t = \Theta(t+1)/t$, we have

$$\gamma_t/\beta_t - \gamma_{t+1}/\beta_{t+1} = \Theta/(t^2 + t) > 0.$$

So, we have from (5.15) that

$$\begin{aligned}&\Phi(\mathbf{x}_{t+1}) + \frac{\eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\\ &\le \Phi(\mathbf{x}_t) + \frac{\eta_t}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\gamma_{t+1}/\beta_{t+1} - \eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\\ &\quad - \frac{\beta_t}{2}\|\check{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2,\end{aligned} \tag{5.16}$$

where $\eta_t = (\gamma_t/\beta_t + \mu)\theta_t^2$.

Now, by the choice of β_t in Algorithm 5.1 and $\mu > 0$, we have

$$\beta_t = \max\{\overline{\beta}_t, \tau\}, \quad \text{where} \quad \tau = 1 - \sqrt{(\Theta - \mu)/(\Theta + \mu)} > 0. \tag{5.17}$$

So, for all $t \ge 1$, we have $\beta_t/\beta_{t-1} \le 1$ and

$$\theta_t = \beta_t/\beta_{t-1}(1 - \beta_{t-1}) \le 1 - \beta_{t-1} \le \sqrt{(\Theta - \mu)/(\Theta + \mu)} < 1. \tag{5.18}$$

Then, by (5.18) and $\gamma_t/\beta_t = \Theta(t+1)/t > \Theta$, for all $t \geq 1$, we have

$$\begin{aligned}\gamma_t/\beta_t - \eta_t &= \gamma_t/\beta_t - (\gamma_t/\beta_t + \mu)\theta_t^2 = \gamma_t/\beta_t(1-\theta_t^2) - \mu\theta_t^2 \\ &\geq \Theta(1-\theta_t^2) - \mu\theta_t^2 = \Theta - (\Theta+\mu)\theta_t^2 \geq \Theta - (\Theta+\mu)\frac{\Theta-\mu}{\Theta+\mu} = \mu.\end{aligned} \tag{5.19}$$

Hence, it follows from (5.16), (5.17) and (5.19) that

$$\begin{aligned}&\Phi(\mathbf{x}_{t+1}) + \frac{\eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\leq \Phi(\mathbf{x}_t) + \frac{\eta_t}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\mu}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad - \frac{\beta_t}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_{\mathcal{M}}^2 - \frac{(\Theta-\Lambda)\tau^2}{2}\|\mathbf{s}_t\|^2\end{aligned} \tag{5.20}$$

for all $t \geq 1$. Since $\Phi(\mathbf{x})$ is bounded from below by Assumption 5.1, we can obtain from (5.20), $\mu > 0$, $\tau > 0$ and $\Theta > \Lambda$ that

$$\sum_{t=\bar{t}}^{\infty}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 < \infty \quad \text{and} \quad \sum_{t=\bar{t}}^{\infty}\|\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t\|^2 = \sum_{t=t_0}^{\infty}\|\mathbf{s}_t\|^2 < \infty,$$

which implies

$$\lim_{t\to\infty}\|\mathbf{x}_{t+1} - \mathbf{x}_t\| = 0 \quad \text{and} \quad \lim_{t\to\infty}\|\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t\| = 0. \tag{5.21}$$

Since $\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t(\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t)$, we have from (5.21) that $\lim_{t\to\infty}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\| = 0$. Then, we have from (5.12) that

$$\lim_{t\to\infty}\|\breve{\mathbf{x}}_t - \mathbf{x}_t\| \leq 1/\tau \lim_{t\to\infty}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\| = 0. \tag{5.22}$$

Therefore, (5.5) follows from (5.10), (5.21), (5.22) and the Lipschitz continuity of ∇f and $\nabla\phi$. □

By Theorem 5.1, any cluster point of $\{\mathbf{x}_t\}$ will be a stationary point of the $\mathbf{x}$-subproblem (5.1). Now suppose $\liminf_{t\to\infty}\|\mathbf{x}_t - \mathbf{x}^k\| > 0$. Otherwise, $\mathbf{x}^k$ is a stationary point of the $\mathbf{x}$-subproblem. We now discuss that the sequence $\{\mathbf{x}_t\}$ generated by Algorithm 5.1 will essentially satisfy the conditions (3.5) and (3.6). First, since $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}}\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \beta\mathcal{D}_{\mathbf{x}}^k(\mathbf{x} - \mathbf{x}^k)$ and $\lim_{t\to\infty}\nabla\Phi(\mathbf{x}_t) = \mathbf{0}$. the condition (3.6) will be satisfied by setting $\widehat{\mathbf{x}}^k = \mathbf{x}_t$ for any $c_{\mathbf{x}} > \overline{\eta}$ and all t sufficiently large. Second, since $\mathbf{x}_0 = \mathbf{x}_1 = \mathbf{x}^k$, we have from (5.16) that

$$\Phi(\widehat{\mathbf{x}}_k) = \Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_1) = \Phi(\mathbf{x}^k)$$

for $t \geq 1$. Note that $\Phi(\widehat{\mathbf{x}}^k) \leq \Phi(\mathbf{x}^k)$ is equivalent to

$$\frac{\beta}{2}\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|_{\mathcal{D}_{\mathbf{x}}^k} + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \leq \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k).$$

So, with the choice of $\mathcal{D}_{\mathbf{x}}^k$ satisfying (5.3), the condition (3.5) holds with $\mathcal{D}_{\mathbf{x}} = \eta\mathbf{I}$ by setting $\widehat{\mathbf{x}}^k = \mathbf{x}_t$ for all $t \geq 1$.

6 Numerical experiments

In this section, we would like to evaluate the performance of Algorithm 3.1 on a sparse optimization problem and a nonconvex quadratic programming problem, where the Assumption 4.3 is known to be satisfied [50, 51, 57]. First, our convergence theory requires that the parameters in Algorithm 3.1 are chosen such that (4.14) and (4.15) hold and $\{\widehat{E}^k\}$ defined in (4.13) is bounded from below. However, the condition (4.14) depends on the Lipschitz constant L, which is usually unknown for general nonlinear function f and a poor estimate of its value may severely deteriorate the algorithm performance. On the other hand, a closer inspection on the convergence proof (see inequality (4.12)) reveals that the convergence results still hold as long as

$$\|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1})\| \leq L\big(\|\widehat{\mathbf{d}}^k\| + \|\widetilde{\mathbf{d}}^{k-1}\|\big) \tag{6.1}$$

holds for all k sufficiently large. Here, L may be some constant smaller than the true Lipschitz constant. Hence, in numerical experiments, we gradually estimate the Lipschitz constant by starting with some $L^0 > 0$ and for $k = 0, 1, \ldots$, update L^k as

$$L^{k+1} = \begin{cases} \rho L^k, & \text{if } \|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1})\| > L^k\big(\|\widehat{\mathbf{d}}^k\| + \|\widetilde{\mathbf{d}}^{k-1}\|\big), \\ L^k, & \text{otherwise}, \end{cases} \tag{6.2}$$

where $\rho > 1$ is some parameter. By this way, since ∇f is Lipschitz continuous, we see that L^k can only be increased finite number of times. Hence, L^k will remain as a constant L such that (6.1) will hold for all k sufficiently large. Under the above choice of L^k, we dynamically update β by $\beta^k = L^k/c_\beta$ at the k-th iteration for some $c_\beta \in (0, 1)$. We require that L^0 and c_β are chosen such that for all $\beta \geq \beta^0 = L^0/c_\beta$, the functions $\mathcal{L}_{\mathbf{y}}^k(\cdot)$ and $\mathcal{L}_{\mathbf{y}}^k(\cdot)$ are bounded from below and (4.33) holds with $\overline{\beta} = \beta^0$. Hence, we can always solve the subproblems inexactly as required by Algorithm 3.1, and $\{\widehat{E}^k\}$ (also $\{E^k\}$) will be bounded from below by Theorem 4.3. So, to ensure global convergence, by Theorem 4.2 and the above setting, we only need to require c_β and the parameters in Algorithm 3.1 are chosen such that

$$\begin{aligned} &\frac{\varphi(\tau)}{2}\mathcal{D}_{\mathbf{x}}^k - \frac{\psi_1(s)\left[2(c_\beta + c_{\mathbf{x}})^2 + 8c_{\mathbf{x}}^2\right]}{s\sigma_A}\mathbf{I} \succeq \mathbf{0}, \\ &\frac{\varphi(\tau)}{16}\mathcal{D}_{\mathbf{y}}^k - \frac{\psi_1(s)c_{\mathbf{x}}^2}{s\sigma_A}\mathbf{I} \succeq \mathbf{0} \quad \text{and} \quad \frac{\delta - \tau}{1+\tau} - \frac{8\psi_1(s)c_\beta^2}{s\sigma_A} \geq 0 \end{aligned} \tag{6.3}$$

for some $\tau \in (0, \delta)$, where $\varphi(\tau) = (1-\tau)/(1+\tau)$. In our numerical experiments, the parameters are chosen as

$$c_\beta = c_{\mathbf{x}} = \frac{1}{14},\ \mathcal{D}_{\mathbf{x}}^k = \mathcal{D}_{\mathbf{y}}^k = \frac{1}{6}\mathbf{I},\ s = 1,\ \rho = 1.01,\ \eta = 1.2,\ \text{and}\ \delta = 0.1.$$

The above choices of parameters satisfy the condition (6.3) with τ sufficiently small in $(0, \delta)$, since $\sigma_A = 1$ in our experiments. Furthermore, all of the forthcoming experiments are implemented in MATLAB R2019b (64-bit) with starting point $(\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ and performed on a PC with Windows 10 operating system, an Intel i7-8565U CPU and 16GB RAM.

6.1 The SCAD penalty problem

Recall the following smoothly clipped absolute deviation (SCAD) penalty problem from statistical learning [20, 66]:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := \frac{1}{2}\|H\mathbf{x} - \mathbf{u}\|^2 + \sum_{i=1}^n p_\kappa(|\mathbf{x}_i|),$$

where $H \in \mathbb{R}^{m\times n}$, $\mathbf{u} \in \mathbb{R}^m$ and the nonconvex SCAD penalty $p_\kappa(\cdot)$ is defined as

$$p_\kappa(\theta) := \begin{cases} \kappa\theta, & \theta \le \kappa, \\ \frac{-\theta^2 + 2c\kappa\theta - \kappa^2}{2(c-1)}, & \kappa < \theta \le c\kappa, \\ \frac{(c+1)\kappa^2}{2}, & \theta > c\kappa, \end{cases}$$

with $c > 2$ and $\kappa > 0$ being the knots of the quadratic spline function. Clearly, the above problem can be reformulated as a special case of (1.1):

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|H\mathbf{x} - \mathbf{u}\|^2 + \sum_{i=1}^n p_\kappa(|\mathbf{y}_i|) \qquad \text{subject to} \quad \mathbf{x} - \mathbf{y} = \mathbf{0}. \tag{6.4}$$

Then, (6.4) is in the format of (1.1) with $f(\mathbf{x}) = \frac{1}{2}\|H\mathbf{x}-\mathbf{u}\|^2$, $g(\mathbf{y}) = \sum_{i=1}^n p_\kappa(|\mathbf{y}_i|)$, $A = \mathbf{I}$, $B = -\mathbf{I}$ and $\mathbf{b} = \mathbf{0}$. Applying I-ADMM Algorithm 3.1 and UPG Algorithm 5.1 with $\mathcal{D}_{\mathbf{y}}^k = \eta_{\mathbf{y}}\mathbf{I}$ and $D_x^k = \eta_{\mathbf{x}}\mathbf{I}$, we have the following updates:

$$\begin{cases} \mathbf{y}^{k+1} = \arg\min_{\mathbf{y}\in\mathbb{R}^n} \sum_{i=1}^n p_\kappa(|\mathbf{y}_i|) + \frac{(1+\eta_{\mathbf{y}})\beta}{2}\left\|\mathbf{y} - \frac{\mathbf{x}^k + \eta_{\mathbf{y}}\mathbf{y}^k - \boldsymbol{\lambda}^k/\beta}{1+\eta_{\mathbf{y}}}\right\|^2, \\ \check{\mathbf{x}}_{t+1} = \frac{1}{\gamma_t + \beta}\left\{\boldsymbol{\lambda}^k + \beta(\eta_{\mathbf{x}}\mathbf{x}^k + \mathbf{y}^{k+1}) - (H^\top H + \beta\eta_{\mathbf{x}}\mathbf{I})\widehat{\mathbf{x}}_t + \gamma_t\check{\mathbf{x}}_t + H^\top\mathbf{u}\right\}, \end{cases}$$

where the $\mathbf{y}$-subproblem has a closed form solution [20, 66].

We chose $\beta^0 = L^0/c_\beta = 1$ in this experiment, which ensures that the $\mathbf{x}$-subproblem is bounded from below since the function f here is nonnegative. We compare I-ADMM with several well-known algorithms for solving the SCAD penalty problem including

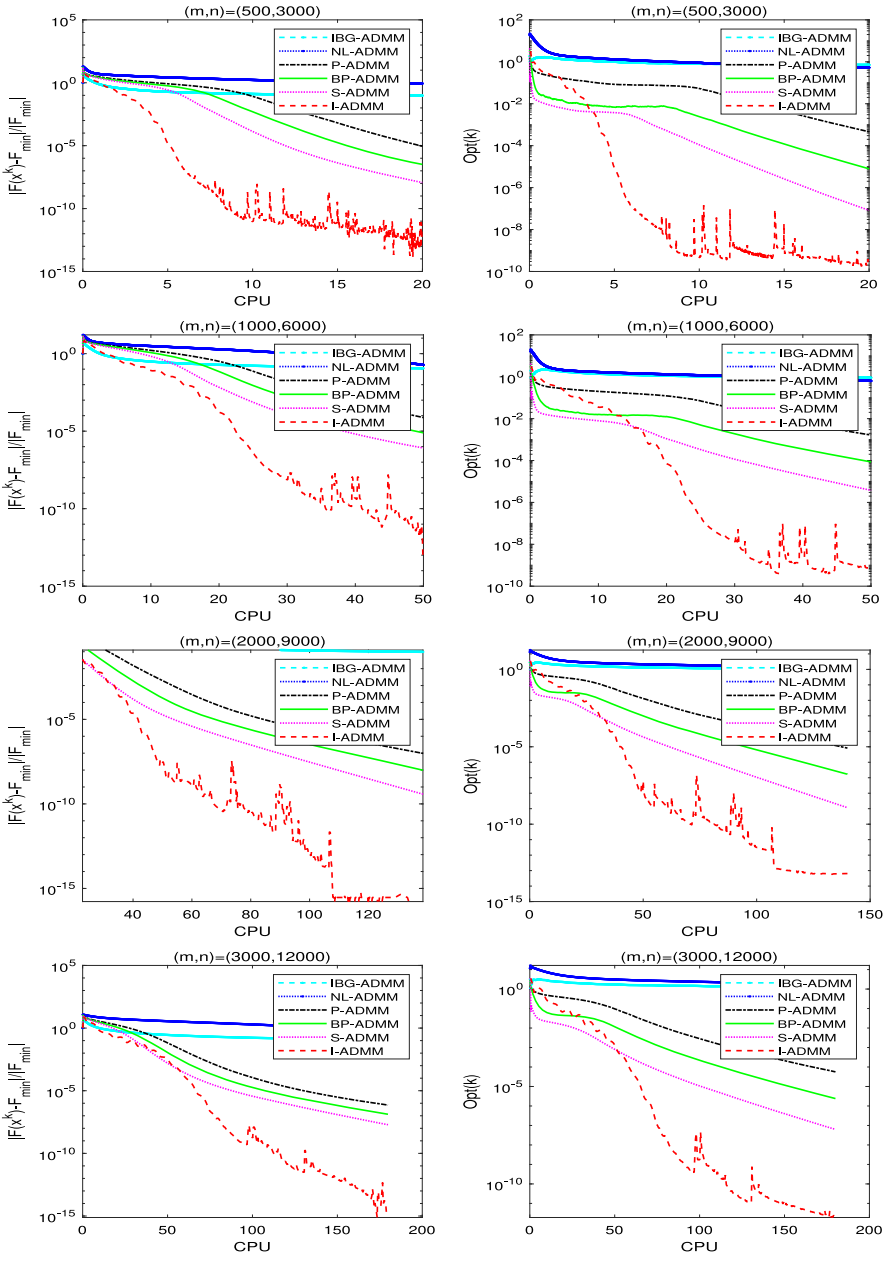

Fig. 1 Numerical comparison of different algorithms for the SCAD penalty problem

NL-ADMM [60], P-ADMM [47], BP-ADMM (Algorithm 2, [11]), S-ADMM [45] and IBG-ADMM [66], where

- NL-ADMM uses the tuned value $\beta = 300$ and $s = 1.6$ as the dual stepsize;
- P-ADMM uses $\beta = 5.1L$ as the penalty value according to [47, Example 1];
- BP-ADMM uses $t_k = \beta$ which is 1.2 times the maximal value satisfying the involved conditions (14) and (15) in [11] (also see [11, Assumption 1]);
- S-ADMM uses the tuned stepsizes $(\alpha, \theta) = (0.05, 1.2)$ and the penalty parameter is chosen to be larger than the maximal eigenvalue of the involved quadratic function (see [45, Assumption 3.1]);
- IBG-ADMM [66] solves (6.4) by introducing variable $\mathbf{y} = H\mathbf{x} - \mathbf{u}$ (see [66, Section 4.2] for more details on the implementation and parameter settings).

Same as those used in [66], the parameters in function p_κ is set as $(c, \kappa) = (3.7, 0.1)$. We first generated a matrix $\overline{H}$ with each component $\overline{H}_{ij} \sim \mathcal{N}(0, 1)$. We then normalize each column of $\overline{H}$ and take it as H. We take $\mathbf{x}^* \in \mathbb{R}^m$ to be a random sparse vector with the density $100/n$ and then set $\mathbf{u} = H\mathbf{x}^* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 100/n)$. The following optimality error $\text{Opt(k)} := \max\left\{\|\mathbf{x}^k - \mathbf{y}^k\|, \|H^\mathsf{T}(H\mathbf{x}^k - \mathbf{u}) - \boldsymbol{\lambda}^k\|\right\}$ is used for the iterates generated by different comparison algorithms, while a different $\text{Opt(k)} = \max\left\{\|H\mathbf{x}^k - \mathbf{y}^k - \mathbf{u}\|, \|\mathbf{y}^k + \boldsymbol{\lambda}^k\|\right\}$ is used for the iterates generated by IBG-ADMM, since it solves the problem (6.4) in a different setup format.

Table 1 reports numerical results of the aforementioned comparison algorithms when a certain CPU time budget is reached, where $F(\mathbf{x}^k)$(end) and Opt(end) denote the function value and optimality error at the last iteration. Figure 1 depicts the convergence curves of $|F(\mathbf{x}^k) - F_{\min}|/|F_{\min}|$ and Opt(k) versus CPU time, where $F_{\min}$ is the minimum of the objective values obtained by all the comparison algorithms. We can see from Table 1 that I-ADMM performs significantly better than other comparison algorithms with respect to the iteration number and the objective function value, and could always obtain a higher accurate solution in terms of optimality error. This efficiency is due to the adaptive inexact subproblem solution, the expansion linesearch step and the adaptive way for updating the Lipschitz constant in (6.2).

6.2 The nonconvex quadratic programming problem

In this subsection, we consider the following Nonconvex Quadratic Programming (NQP) problem

$$\min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^m} \ \frac{1}{2}\mathbf{x}^\mathsf{T}G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} \quad \text{subject to} \quad A\mathbf{x} = \mathbf{y},\ \mathbf{v} \le \mathbf{y} \le \mathbf{u},\ \mathbf{e}^\mathsf{T}\mathbf{y} = c, \tag{6.5}$$

where a symmetric matrix $G \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{v} \le \mathbf{u} \in \mathbb{R}^m$ are given matrices and vectors, respectively, $\mathbf{e}$ is the vector of ones and the scalar c satisfies $\mathbf{e}^\mathsf{T}\mathbf{v} \le c \le \mathbf{e}^\mathsf{T}\mathbf{u}$. When $A = \mathbf{I}$, the problem (6.5) will reduce to a quadratic programming problem with simplex constraints, which includes the example problems in [62, Section 4.1] and has many applications. Note that since efficient projection on the feasible set of (6.5), which is a polyhedron, is in general nontrivial, NQP is not easily solved by the algorithms which require repeated projections on a polyhedron,

Table 1 Numerical results of different algorithms for the SCAD penalty problem

Size	CPU	IBG-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	1857	2.404969	7.1530e−1
(1000,6000)	50	1033	2.779820	9.0591e−1
(2000,9000)	140	1076	3.334330	1.0787e+0
(3000,12000)	180	749	3.901173	1.2769e+0
Size	CPU	NL-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	2321	4.110444	5.3002e−1
(1000,6000)	50	2277	2.966996	6.2893e−1
(2000,9000)	140	1322	3.797333	9.6753e−1
(3000,12000)	180	883	5.445761	1.5238e+0
Size	CPU	P-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	2260	2.193902	4.5738e−4
(1000,6000)	50	1202	2.495977	1.6469e−3
(2000,9000)	140	1311	3.039743	8.2070e−6
(3000,12000)	180	868	3.500552	5.6349e−5
Size	CPU	BP-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	2271	2.193884	7.8562e−6
(1000,6000)	50	1185	2.495809	8.5903e−5
(2000,9000)	140	1301	3.039743	1.6709e−7
(3000,12000)	180	857	3.500550	2.3991e−6
Size	CPU	S-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	2360	2.193883	8.2222e−8
(1000,6000)	50	1213	2.495792	3.7610e−6
(2000,9000)	140	1340	3.039743	1.1609e−9
(3000,12000)	180	847	3.500550	6.3096e−8
Size	CPU	I-ADMM		
(m,n)	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
(500,3000)	20	**843**	**2.193883**	**1.9621e−10**
(1000,6000)	50	**360**	**2.495790**	**7.1638e−10**
(2000,9000)	140	**440**	**3.039743**	**6.4663e−14**
(3000,12000)	180	**341**	**3.500550**	**1.8932e−12**

Bold values indicates the smallest value in that categary

Table 2 Numerical results of different algorithms for solving the NQP problem

Size	CPU	P-ADMM		
n	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
2000	200	24327	3.024108	3.0892e−2
3000	300	14641	1.119980	7.0825e−2
4000	400	11168	1.072323	1.4532e−1
5000	500	9045	0.985398	2.3115e−1
Size	CPU	BP-ADMM		
n	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
2000	200	18357	3.024114	3.7827e−2
3000	300	11786	1.119972	7.6823e−2
4000	400	8430	1.072409	1.7569e−1
5000	500	7328	0.985195	2.4323e−1
Size	CPU	S-ADMM		
n	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
2000	200	19603	3.024074	2.0892e−3
3000	300	14843	1.119909	3.6842e−2
4000	400	10069	1.072095	4.7554e−2
5000	500	9115	0.983785	1.3257e−1
Size	CPU	I-ADMM		
n	time(s)	Iter	$F(\mathbf{x}^k)$(end)	Opt(end)
2000	200	**8180**	**3.020470**	**3.7338e−4**
3000	300	**6296**	**1.119848**	**6.2395e−4**
4000	400	**3929**	**1.072086**	**5.1192e−5**
5000	500	**3987**	**0.983581**	**5.9107e−4**

such as the proximal gradient method with extrapolation [62] or the projected gradient methods [9, 32].

Note that the problem (6.5) can be also rewritten in the format of (1.1) as

$$\min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^m} \ \frac{1}{2}\mathbf{x}^{\mathsf{T}}G\mathbf{x}-\mathbf{g}^{\mathsf{T}}\mathbf{x}+\delta_{\mathcal{C}}(\mathbf{y}) \qquad \text{subject to} \quad A\mathbf{x}=\mathbf{y}, \tag{6.6}$$

where $\delta_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}=\{\mathbf{y}\in\mathbb{R}^m : \mathbf{v}\le\mathbf{y}\le\mathbf{u},\ \mathbf{e}^{\mathsf{T}}\mathbf{y}=c\}$, i.e., $\delta_{\mathcal{C}}(\mathbf{y})=0$ if $\mathbf{y}\in\mathcal{C}$; $\delta_{\mathcal{C}}(\mathbf{y})=\infty$, otherwise. Applying I-ADMM Algorithm 3.1 and UPG Algorithm 5.1 to the problem (6.6) with $\mathcal{D}_{\mathbf{y}}^k=\eta_{\mathbf{y}}\mathbf{I}$ and $D_x^k=\eta_{\mathbf{x}}\mathbf{I}$ involves solving the following subproblems:

$$\mathbf{y}^{k+1}=\arg\min_{\mathbf{y}\in\mathbb{R}^n}\delta_{\mathcal{C}}(\mathbf{y})+\frac{(1+\eta_{\mathbf{y}})\beta}{2}\|\mathbf{y}-\mathbf{q}\|^2 \ \text{ and } \ (\frac{\gamma_t}{\beta}\mathbf{I}+A^{\mathsf{T}}A)\check{\mathbf{x}}_{t+1}=\mathbf{b},$$

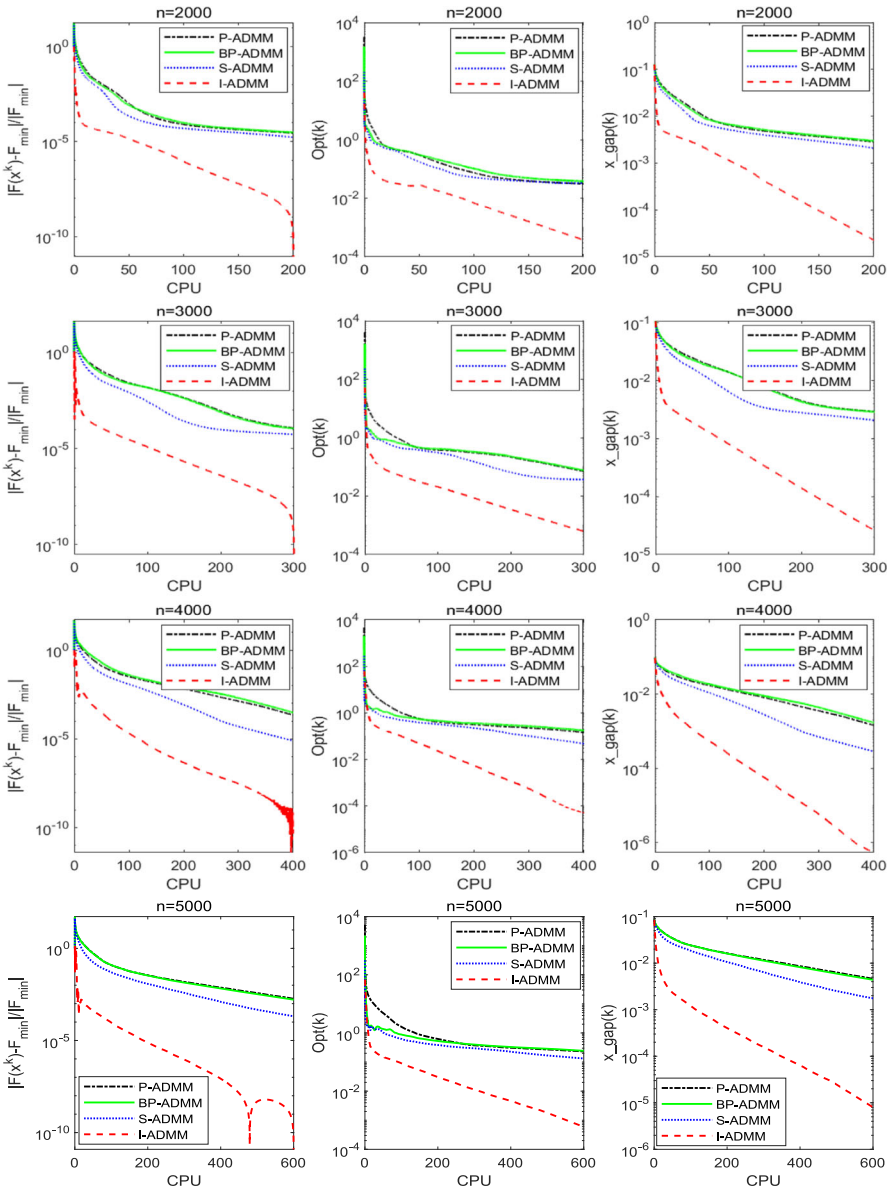

Fig. 2 Numerical comparison of different algorithms for solving the NQP problem

where $\mathbf{q} := \frac{A\mathbf{x}^k+\eta_{\mathbf{y}}\mathbf{y}^k-\boldsymbol{\lambda}^k/\beta}{1+\eta_{\mathbf{y}}}$ and $\mathbf{b} := \frac{1}{\beta}A^{\mathsf{T}}\boldsymbol{\lambda}^k + \eta_{\mathbf{x}}\mathbf{x}^k + A^{\mathsf{T}}\mathbf{y}^{k+1} - (\eta_{\mathbf{x}}\mathbf{I} + \frac{1}{\beta}G)\widehat{\mathbf{x}}_t + \frac{1}{\beta}(\gamma_t\breve{\mathbf{x}}_t + \mathbf{g})$. Observe that the above $\mathbf{y}$-subproblem, which needs projection on a simplex, has no closed-form solution. Hence, we solve it inexactly by the method developed in [16, 56] using the stopping criteria (3.3) and (3.4) with $c_{\mathbf{y}} = 0.1$. In addition, when $m \ll n$, the Sherman-Morrison-Woodbury Formula should be used to

solve $\breve{\mathbf{x}}_{t+1}$ as

$$\breve{\mathbf{x}}_{t+1} = \frac{\beta}{\gamma_t}\mathbf{b} - \frac{\beta^2}{\gamma_t^2}A^{\mathsf{T}}\Big(\mathbf{I} + \frac{\beta}{\gamma_t}AA^{\mathsf{T}}\Big)^{-1}A\mathbf{b}.$$

In our numerical experiments, A is always generated to be an orthogonal matrix, i.e. $A^{\mathsf{T}}A = \mathbf{I}$. Note that even for A being an orthogonal matrix, projection on the feasible set of problem (6.5) is in general still nontrivial. Specifically, similar to the way of generating the problem data in [62], we randomly generate G, $\mathbf{g}$, A, and set $\mathbf{v}$, $\mathbf{u}$ and c by the following MATLAB codes:

```
D=randn(n); Z=zeros(n,n);
for i=1:n Z(i,i)=10*(rand(1)−0.1); end
G=D'*Z*D; g=randn(n,1); U=randn(n,n);
A=orth(U)'; u=10*ones(n,1); v=zeros(n,1); c=5;
```

We set $\beta^0 = L^0/c_\beta = 2|\min\{\lambda_{\min}(G), 0\}| + 1$ to ensure that the $\mathbf{x}$-subproblem is bounded from below, where $\lambda_{\min}(G)$ is the minimum eigenvalue of G. We compare I-ADMM with the aforementioned algorithms P-ADMM, BP-ADMM and S-ADMM. The rest two algorithms IBG-ADMM and NL-ADMM are not compared since their performance is much worse for solving this test problem. The numerical experiment results including the number of iterations, the final objective function value $F(\mathbf{x}^k)$ and the final optimality error Opt(k) $:=$ $\max\big(\|A\mathbf{x}^k - \mathbf{y}^k\|, \|G\mathbf{x}^k - \mathbf{g} - A^{\mathsf{T}}\boldsymbol{\lambda}^k\|, \|\mathbf{y}^k - \mathcal{P}_{\mathcal{C}}(\mathbf{y}^k - \boldsymbol{\lambda}^k)\|\big)$ along with problem dimension n are reported in Table 2. Here, $\mathcal{P}_{\mathcal{C}}(\cdot)$ denotes projection onto the convex set $\mathcal{C}$. In Fig. 2, we also plot $|F(\mathbf{x}^k) - F_{\min}|/|F_{\min}|$, Opt(k) and $\mathbf{x}$_gap(k) $:=$ $\frac{\|\mathbf{x}^k - \mathbf{x}^*\|}{1+\|\mathbf{x}^*\|}$ against the CPU time with $n = 2000, 3000, 4000, 5000$ respectively, where $F_{\min}$ is the minimum objective value obtained by all the algorithms, and $\mathbf{x}^*$ denotes the approximate optimal solution obtained by I-ADMM under twice of the CPU time budget. From Table 2 and Fig. 2, we can again see that I-ADMM converges much faster and obtains a higher accuracy solution than other comparison algorithms under the same CPU time budget. Besides, Fig. 2 clearly shows the linear convergence behavior of the optimality error Opt(k) and the iteration error $\mathbf{x}$_gap(k) generated by I-ADMM for solving the NQP problem (6.5).

7 Conclusion

We have developed an inexact alternating direction method of multipliers with an expansion line search step for solving a class of separable nonconvex and nonsmooth structured optimization with linear constraints. This I-ADMM solves each subproblem inexactly to an adaptive accuracy and allows a larger range of dual stepsize. Under proper assumptions, the global convergence and linear convergence rate of I-ADMM have been established. In addition, a unified proximal gradient method with momentum acceleration is proposed to solve the smooth but possibly nonconvex subproblem inexactly. By allowing adaptive inexact subproblem solution, the expansion linesearch step and the adaptive way for updating the Lipschitz constant, our proposed I-ADMM performs significantly better than other state-of-the-art algorithms

for solving some nonconvex quadratic programming problems and nonconvex sparse optimization problems from statistical learning.

Data Availability The data for test problems in the paper is randomly generated and described in the paper.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest to this work.

References

1. Aussel, D., Daniilidis, A., Thibault, L.: Subsmooth sets: functional characterizations and related concepts. Trans. Am. Math. Soc. **357**, 1275–1301 (2004)
2. Bai, J., Li, J., Xu, F., Zhang, H.: Generalized symmetric ADMM for separable convex optimization. Comput. Optim. Appl. **70**, 129–170 (2018)
3. Bai, J., Hager, W.W., Zhang, H.: An inexact accelerated stochastic ADMM for separable composite convex optimization. Comput. Optim. Appl. **81**, 479–518 (2022)
4. Bai, J., Han, D., Sun, H., Zhang, H.: Convergence on a symmetric accelerated stochastic ADMM with larger stepsizes. CSIAM Trans. Appl. Math. **3**, 448–479 (2022)
5. Bai, J., Bian, F., Chang, X., Du, L.: Accelerated stochastic Peaceman–Rachford method for empirical risk minimization. J. Oper. Res. Soc. China **11**, 783–807 (2023)
6. Barber, R., Sidky, E.: Convergence for nonconvex ADMM with applications to CT imaging. J. Mach. Learn. Res. **25**, 1–46 (2024)
7. Beck, A., Teboulle, M.: A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization. Math. Oper. Res. **31**, 398–417 (2006)
8. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
9. Birgin, E.G., Martínez, J.M., Raydan, M.: Algorithm 813: SPG-software for convex-constrained optimization. ACM Trans. Math. Softw. **27**, 340–349 (2001)
10. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2010)
11. Bot, R., Nguyen, D.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. **45**, 682–712 (2020)
12. Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**, 39–73 (2017)
13. Candès, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**, 1–37 (2011)
14. Chen, C., Li, M., Liu, X., Ye, Y.: Extended ADMM and BCD for nonseparable convex minimization models with quadratic coupling terms: convergence analysis and insights. Math. Program. **173**, 37–77 (2019)
15. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**, 81–101 (1994)
16. Dai, Y.H., Fletcher, R.: New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds. Math. Program. **106**, 403–421 (2006)
17. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Valued Var. Anal. **25**, 829–858 (2017)
18. Eckstein, J., Bertsekas, D.: On the Douglas–Rachford splitting method and the proximal point algorithm for the maximal monotone operators. Math. Program. **55**, 293–318 (1992)
19. Eckstein, J., Silva, P.: A practical relative error criterion for augmented Lagrangians. Math. Program. **141**, 319–348 (2013)
20. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

21. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: Fortin, M., Glowinski, R. (eds.) Augmented Lagrange Methods: Applications to the Solution of Boundary-Valued Problems, pp. 299–331. North Holland, Amsterdam (1983)
22. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
23. Gao, X., Xu, Y., Zhang, S.: Randomized primal–dual proximal block coordinate updates. J. Oper. Res. Soc. China **7**, 205–250 (2019)
24. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**, 59–99 (2016)
25. Ghadimi, S., Lan, G., Zhang, H.: Generalized uniformly optimal methods for nonlinear programming. J. Sci. Comput. **79**, 1854–1881 (2019)
26. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
27. Goldfarb, D., Ma, S.: Fast multiple-splitting algorithms for convex optimization. SIAM J. Optim. **22**, 533–556 (2012)
28. Gol'shtein, E., Tret'yakov, N.: Modified Lagrangians in convex programming and their generalizations. Point-to-Set Maps and Mathematical Programming, pp. 86–97 (1979)
29. Goncalves, M., Melo, J., Monteiro, R.: Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. Pac. J. Optim. **15**, 379–398 (2019)
30. Guo, K., Han, D., Wu, T.: Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. Int. J. Comput. Math. **94**, 1653–1669 (2017)
31. Guo, K., Han, D., Wang, D., Wu, T.: Convergence of ADMM for multi-block nonconvex separable optimization models. Front. Math. China **12**, 1139–1162 (2017)
32. Hager, W.W., Zhang, H.: An active set algorithm for nonlinear optimization with polyhedral constraints. Sci. China Math. **59**, 1525–1542 (2016)
33. Hager, W.W., Zhang, H.: Inexact alternating direction multiplier methods for separable convex optimization. Comput. Optim. Appl. **73**, 201–235 (2019)
34. Hager, W.W., Zhang, H.: Convergence rates for an inexact ADMM applied to separable convex optimization. Comput. Optim. Appl. **77**, 729–754 (2020)
35. Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**, 227–238 (2012)
36. Han, D., Yuan, X.G., Zhang, W.: An augmented-Lagrangian-based parallel splitting method for separable convex minimization with applications to image processing. Math. Comput. **83**, 2263–2291 (2014)
37. He, B., Liao, L., Han, D., Yan, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**, 103–118 (2002)
38. He, B., Tao, M., Xu, M., Yuan, X.: An alternating direction-based contraction method for linearly constrained separable convex programming problems. Optimization **62**, 573–596 (2013)
39. He, B., Yuan, X.: On the $o(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
40. He, B., Yuan, X., Zhang, W.: A customized proximal point algorithm for convex minimization with linear constraints. Comput. Optim. Appl. **56**, 559–572 (2013)
41. Hestenes, M.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
42. Huang, F., Chen, C., Lu, Z.: Stochastic alternating direction method of multipliers with variance reduction for nonconvex optimizations (2017). arXiv:1610.02758v5
43. Hong, M., Luo, Z., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. **26**, 337–364 (2016)
44. Jia, Z., Gao, X., Cai, X., Han, D.: Local linear convergence of the alternating direction method of multipliers for nonconvex separable optimization problems. J. Optim. Theory Appl. **188**, 1–25 (2021)
45. Jia, Z., Gao, X., Cai, X., Han, D.: The convergence rate analysis of the symmetric ADMM for the nonconvex separable optimization problems. J. Ind. Manag. Optim. **17**, 1943–1971 (2021)
46. Kim, S., Xing, E.: Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. **5**, 1–18 (2009)
47. Li, G., Pong, T.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**, 2434–2460 (2015)
48. Li, M., Sun, D., Toh, K.: A convergent 3-block semi-proximal ADMM for for convex minimization problems with one strongly convex block. Asia Pac. J. Oper. Res. **32**, 1–19 (2015)

49. Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multiblock variables. SIAM J. Optim. **25**, 1478–1497 (2015)
50. Luo, Z., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**, 408–425 (1992)
51. Luo, Z., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
52. Rockafellar, R.: Generalized directional derivatives and subgradients of nonconvex functions. Can. J. Math. **32**, 257–280 (1980)
53. Rockafellar, R.: Favorable classes of Lipschitz continuous functions in subgradient optimization. In: Nurminski, E. (ed.) Nondifferential Optimization. Pergamon Press, New York (1982)
54. Rockafellar, R., Wets, R.: Variational Analysis. Springer, New York (1998)
55. Solodov, M., Svaiter, B.: An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. Math. Oper. Res. **25**, 214–230 (2000)
56. Tavakoli, R., Zhang, H.: A nonmonotone spectral projected gradient method for large-scale topology optimization. Numer. Algebra Control Optim. **2**, 395–412 (2012)
57. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
58. Tseng, P., Yun, S.: A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Comput. Optim. Appl. **47**, 179–206 (2010)
59. Vial, J.: Strong and weak convexity of sets and functions. Math. Oper. Res. **8**(2), 231–259 (1983)
60. Qiao, L., Zhang, B., Su, J., Lu, X.: Linearized alternating direction method of multipliers for constrained nonconvex regularization optimization. ALML **63**, 97–109 (2016)
61. Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. **78**, 29–63 (2019)
62. Wen, B., Cheng, X., Pong, T.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**, 124–145 (2017)
63. Wen, Z., Yang, C., Liu, X., Marchesini, S.: Alternating direction methods for classical and ptychographic phase retrieval. Inverse Prob. **28**, 115010 (2012)
64. Yang, J., Zhang, Y.: Alternating direction algorithms for l_1-problems in compressive sensing. SIAM J. Comput. **33**, 250–278 (2011)
65. Wu, Z., Li, M., Wang, D., Han, D.: A symmetric alternating direction method of multipliers for separable nonconvex minimization problems. Asia Pac. J. Oper. Res. **34**, 1750030 (2017)
66. Xu, J., Chao, M.: An inertial Bregman generalized alternating direction method of multipliers for nonconvex optimization. J. Appl. Math. Comput. **68**, 1757–1783 (2022)
67. Xu, Z., Chang, X., Xu, F., Zhang, H.: $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. **23**, 1013–1027 (2012)
68. Yashtini, M.: Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization. J. Glob. Optim. **84**, 913–939 (2022)
69. Yang, L., Pong, T., Chen, X.: Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground. SIAM J. Imaging Sci. **10**, 74–110 (2017)
70. Zhou, Z., So, A.: A unified approach to error bounds for structured convex optimization problems. Math. Program. **165**, 689–728 (2017)