Double-proximal augmented Lagrangian methods with improved convergence condition [*]

Jianchao Bai[†] Shuang Rao[‡] Deren Han[§] Ruijun Sun[¶]

Abstract

In this paper, we propose a novel double-proximal augmented Lagrangian method (DP-ALM) for solving a family of linearly constrained convex minimization problems whose objective function is not necessarily smooth. This DP-ALM not only enjoys a flexible dual stepsize, but also contains a proximal subproblem with relatively smaller proximal parameter. By a new prediction-correction reformulation for this DP-ALM and similar variational characterizations for both the saddle-point of the problem and the generated sequences, we establish its global convergence and sublinear convergence rate in both ergodic and nonergodic senses. A toy example is taken to illustrate that the presented lower bound of proximal parameter is optimal (smallest). We also discuss a relaxed accelerated version as well as a linearized version of DP-ALM when the objective function has composite structures. Experiments results on solving two large-scale sparse optimization problems show that our proposed methods outperform some well-established methods. In the appendix, we briefly discuss a multi-block extended DP-ALM and its convergence condition.

Keywords: convex optimization, augmented Lagrangian method, proximal term, prediction-correction technique, convergence and complexity
Mathematics Subject Classification(2010): 65K10, 65Y20, 90C25

1 Introduction

Consider the following canonical convex optimization model

$$\min \theta(x) \qquad \text{s.t. } Ax = b, x \in \mathbb{X}, \tag{1}$$

where $\theta : \mathbb{R}^n \to \mathbb{R}$ is a proper lower semicontinuous convex function (possibly nonsmooth and non-strongly convex), $\mathbb{X} \subseteq \mathbb{R}^n$ is a simple closed convex set, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given matrix and vector, respectively. Many application problems can be formulated as the form of (1), such as the following typical examples:

[*] This work was supported by the National Natural Science Foundation of China (12471298, 12131004), the National Key R&D Program of China (2021YFA1003600), the Shaanxi Fundamental Science Research Project for Mathematics and Physics (23JSQ031), and the National Social Science Fund of China (22BGL11).

[†] School of Mathematics and Statistics & MOE Key Laboratory for Complexity Science in Aerospace, Northwestern Polytechnical University, Xi'an 710129, China (jianchaobai@nwpu.edu.cn).

[‡] School of Mathematics and Statistics, Northwestern Polytechnical University & MIIT Key Laboratory of Dynamics and Control of Complex Systems, Xi'an, 710129, China (raoshuang169@163.com).

[§] Corresponding author. LMIB of the Ministry of Education, School of Mathematical Sciences, Beihang University, Beijing, 100191, China. (handr@buaa.edu.cn).

[¶] School of Mathematics and Statistics, Northwestern Polytechnical University & MIIT Key Laboratory of Dynamics and Control of Complex Systems, Xi'an, 710129, China(rjsun@163.com).

- $\triangleright$ *Example 1.* The linearly constrained least-squares problem [28, 35] arises from scattered data approximation, fitting curves to data, and real-time signal processing, where $\theta(x) = \|Cx - d\|^2$; $C \in \mathbb{R}^{l\times n}$ is a large sparse matrix whose dimension satisfies $l > n$; the dimension of A satisfies $m \ll n$. A particular case of this example is $\theta(x) = 0$, then the problem reduces to an ill-conditioned linear equation.

- $\triangleright$ *Example 2.* The sparse optimization problem [23, 41] with $\theta(x) = \|x\|_1$ aims to find a sparse solution of the underdetermined system $Ax = b$. This problem arises from compressed sensing, statistical learning, machine learning, and graphical modeling, and the dimension of the coefficient matrix A satisfies $m \ll n$.

- $\triangleright$ *Example 3.* The decentralized composite optimization over networks [15] arises from multi-agent control, wireless communication, and machine learning, where

$$\theta(\mathbf{x}) = \sum_{i=1}^{N} f_i(x_i) + \sum_{i=1}^{N} g_i(x_i),\ \ A = \mathbf{I} - \mathbf{W},\ \ b = \mathbf{0}.$$

Here $\mathbf{x} = [x_1^\top, x_2^\top, \cdots, x_N^\top]^\top$, N denotes the number of computational agents, and $\mathbf{W} = W \otimes \mathbf{I}$ with W being a symmetric and doubly stochastic matrix. The loss function f_i is assumed to be smooth, while g_i is not necessarily smooth.

Except for these examples, the model (1) also applies in image processing and data processing [33, 39]. Throughout the context, we assume the solution set of (1) is nonempty.

1.1 Notations and definitions

Let $\mathbb{R}^n$ be the set of n-dimensional real column vectors equipped with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}$. The ℓ_1-norm of $x \in \mathbb{R}^n$ is defined as $\|x\|_1 = \sum_{i=1}^n |x_i|$ where x_i denotes the i-th element of x. We simply use bold $\mathbf{I}$ and $\mathbf{0}$ to represent the identity matrix and zero matrix (or vector), respectively. The symbol $\otimes$ denotes the so-called Kronecker product. Define $\|x\|_G = \sqrt{x^\top Gx}$ as the weighted G-norm, where $\top$ denotes the transpose operator and G is a symmetric positive definite matrix. The spectral radius of a square matrix A is denoted by $\rho(A)$, while $\lambda_{\max}(A)$ represents the largest eigenvalues of A. The proximity operator of a proper convex function $\theta(x) : \mathbb{X} \to \mathbb{R}$ with parameter $r > 0$ is defined as

$$\mathbf{prox}_{r,\theta}(\cdot) := \arg\min_{x\in\mathbb{X}} \Big\{\theta(x) + \frac{r}{2}\|x - \cdot\|^2\Big\}.$$

1.2 Related work

The Augmented Lagrangian Method (ALM, [22, 34]) is a benchmark method for solving equality constrained minimization problem (1), and it remains relatively popular yet vivid in recent years, cf. [4, 7, 18, 25, 26, 32, 42] to list a few. The framework of classic ALM for solving the problem (1) reads

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{X}} \Big\{L(x,\lambda^k) + \frac{\beta}{2}\|Ax - b\|^2\Big\}, \\ \lambda^{k+1} = \lambda^k - \gamma\beta\big(Ax^{k+1} - b\big), \end{cases} \tag{2}$$

where $\beta > 0$ is the quadratic penalty parameter for the equality constraint, $\gamma \in (0, 2)$ denotes the stepsize parameter of dual variable λ, and

$$L(x, \lambda) = \theta(x) - \langle\lambda, Ax - b\rangle$$

is the associated Lagrange function. Simple algebra shows that the x-subproblem in (2) amounts to $x^{k+1} = \arg\min_{x\in\mathbb{X}} \{\theta(x) + \frac{\beta}{2}\|Ax - b - \lambda^k/\beta\|^2\}$. When the coefficient matrix $A \neq \mathbf{I}$, solving this subproblem is still challenging even as difficult as the original.

To overcome the above obstacle, a powerful technique is to add a quadratic proximal term in the form of $\frac{1}{2}\|x - x^k\|_D^2$ to the ALM's subproblem. Here $D \in \mathbb{R}^{n\times n}$ is called the proximal matrix and usually required to be positive definite for the sake of convergence. If users choose $D = r\mathbf{I} - \beta A^\top A$ with $r > \beta\rho(A^\top A)$, then the modified subproblem can be simplified as the following proximity operator

$$x^{k+1} = \mathbf{prox}_{r,\theta}\big[x^k + A^\top[\lambda^k - \beta(Ax^k - b)]/r\big].$$

The function $\theta(x)$ may be special enough so that the above proximity operator has a closed-form solution. Such a representative is the case with $\theta = \|x\|_1$. Otherwise, one may exploit inner solvers or solve it inexactly, or use the computational formula [31] for accurately approximating the proximity operator. To investigate an indefinite proximal term, He, et al. [18] proposed an optimal proximal ALM as follows

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathbb{X}} \Big\{L(x,\lambda^k) + \frac{\beta}{2}\|Ax - b\|^2 + \frac{1}{2}\|x - x^k\|_{D_0}^2\Big\}, \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} - b), \end{cases} \tag{3}$$

where $D_0 = D_1 - (1-\tau)\beta A^\top A$ and D_1 is an arbitrarily positive-definite matrix in $\mathbb{R}^{n\times n}$. The global convergence of (3) and its sublinear convergence were established for any $\tau > \frac{2+\gamma}{4}$. Obviously, it follows from the region of τ that D_0 is not necessarily positive definite. The scheme (3) is called optimal proximal ALM since the proximal parameter can not be smaller than $\frac{2+\gamma}{4}$, that is, $\frac{2+\gamma}{4}$ is the optimal (smallest) lower bound.

Recently, Bai, et al. [4] studied a new double-penalty ALM with a relaxation step for convex optimization problems, where the key subproblem and the dual variable recursively take the following iterative scheme:

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathbb{X}} \Big\{L(x,\lambda^k) + \frac{\beta}{2}\|A(x - x^k)\|^2 + \frac{1}{2}\|x - x^k\|_D^2\Big\}, \\ \lambda^{k+1} = \lambda^k - \beta\big[A(2x^{k+1} - x^k) - b\big]. \end{cases} \tag{4}$$

The authors showed that the global convergence and sublinear convergence rate of (4) can be established for any arbitrarily positive-definite matrix D. Notice that the scheme (4) is different from most of ALM-type methods since the subproblem does not depend on the data b either. The two involved quadratic terms can be treated as different proximal terms: one involves the data matrix A, while the other does not involve it. Based on [4], a different dual-primal ALM was developed in [32] and was demonstrated to be efficient for solving large-scale basic pursuit problem and Lasso problem. Cui, et al. [10] also provided some gentle introductions to the recent advance in augmented Lagrangian methods for solving large-scale convex matrix optimization problems. Han [16] systematically reviewed the developments of the problem (1) and its multi-block extensions from ALM to its splitting methods. Recently, Birgin, et al. [2] studied the global convergence of a general augmented Lagrangian method based on a weak regularity condition which does not require boundedness of the set of Lagrange multipliers. For more details on accelerated versions of ALM, we refer to [25, 42].

In addition, some researchers also focused on stochastic ALM and nonconvex ALM, except for the above deterministic ALM-type work on convex programming problems. For example, a stochastic ALM based on a variant of the stochastic accelerated gradient method was presented in [3, Appendix A.1] for solving the case of (1) that $\theta(x)$ is a finite-sum of Lipschitz continuously differentiable convex functions. Li, et al. [26] developed a stochastic composite ALM by penalizing the constraints to formulate a quadratic penalty problem and employing semi-gradient for the value function. Bollapragada, et al. [7] also constructed an adaptive sampling ALM for linearly equality constrained optimization problems including a stochastic objective function, and they further established its sublinear convergence for convex objectives and linear convergence for strongly convex objectives. Advanced progresses on nonconvex ALM can be found in e.g. [1, 25, 45, 44].

1.3 The proposed algorithm and contributions

Mainly motivated by the interesting work [4, 18, 40] we propose the following double-proximal ALM-type method for solving the problem (1):

$$\text{(DP-ALM)}\quad \begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{X}} \left\{\theta(x) - \langle\lambda^k, Ax\rangle + \frac{\tau\beta}{2}\left\|A(x-x^k)\right\|^2 + \frac{\tau}{2}\left\|x-x^k\right\|_D^2\right\}, \\ \lambda^{k+1} = \lambda^k - \beta\big[\gamma(Ax^{k+1}-b) + A(x^{k+1}-x^k)\big], \end{cases} \tag{5}$$

where $D = r\mathbf{I} - \beta A^\top A$ with $r > \beta\rho(A^\top A)$ and the parameter

$$\tau > \frac{(\alpha-\frac{\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4}, \quad \forall\gamma\in(0,2), \alpha\in\mathbb{R}. \tag{6}$$

With the structure of D, the above x-subproblem can be simplified as that in (8a). Here, α can be treated as an auxiliary parameter. For using the so-called optimal parameter value approximating to 0.75 as pointed in [18], one can select $\alpha = \gamma/2$ and $\gamma = 1$. When applying the indefinite proximal point algorithm [24] to solve (1), it results in a similar dual update as that in our DP-ALM. We further reformulate (5) as the following prediction-correction framework, where

$$w^k = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix}, \quad \tilde{w}^k = \begin{pmatrix} \tilde{x}^k \\ \tilde{\lambda}^k \end{pmatrix} \quad \text{and} \quad M = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -(1-\alpha\gamma)\beta A & \gamma\mathbf{I} \end{bmatrix}. \tag{7}$$

A prediction-correction reformulation of DP-ALM (5).

Prediction Step:

$$\tilde{x}^k = \mathbf{prox}_{\tau r,\theta}\big[x^k + A^\top\lambda^k/(\tau r)\big]; \tag{8a}$$

$$\tilde{\lambda}^k = \lambda^k - \beta\big[A\tilde{x}^k - b + \alpha A(\tilde{x}^k - x^k)\big]; \tag{8b}$$

Correction Step:

$$w^{k+1} = w^k - M(w^k - \tilde{w}^k). \tag{9}$$

Main features and contributions of this paper are summarized as three aspects:

♣ **Flexibility of the algorithm.** If $(\tau,\gamma) = (1,1)$, our DP-ALM reduces to the double-penalty ALM in [4] without using the relaxation step, and both of them enjoy the same proximal subproblem. However, when $\alpha < 1$ or $\alpha > \gamma - 1$, we have

$$\frac{(\alpha-\frac{\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4} < 1,$$

which indicates that DP-ALM can enjoy a smaller proximal parameter τ than that in [4]. As discussed in the latter Section 4.1, $\frac{2+\gamma}{4}$ is the optimal lower bound of proximal parameter τ. Moreover, as γ approximates to zero, this bound will approximate to 1/2. Compared to the subproblem in the classical ALM, the subproblem in our DP-ALM not only maintains the merits of that in [4], but also allows a relatively smaller proximal parameter, while the dual update has an extra term $\alpha\beta A(x^{k+1}-x^k)$ that is not involved in (2) and (3). We point out that DP-ALM can reduce to some primal-dual methods as analyzed in the next subsection.

♣ **Global convergence and various convergence rates.** Unlike most of prediction-correction reformulations, we reformulate DP-ALM as a novel prediction-correction framework with the aid of an auxiliary parameter, and finally establish the global convergence of DP-ALM and its various convergence rates, including the sublinear ergodic convergence rate in terms of the objective function value gap and the constraint violation, the sublinear nonergodic convergence rate in terms of the pointwise iterative residual and the first-order optimality error of the subproblems. Although our analysis firstly uses a distinctive update (8b), that is, the extra $\alpha\beta A(\tilde{x}^k - x^k)$ is exploited compared to the existing prediction steps on dual iterate, this can greatly simplify the whole convergence analysis, see its application in the next item. In addition, we further discuss the convergence and convergence conditions of an extended DP-ALM with the widely-used relaxation step, a linearized DP-ALM as well as a multi-block primal-dual splitting version.

♣ **The applicability of new reformulation technique.** Motivated by the above prediction-correction technique, we can apply this new technique to simply discuss the key convergence analysis of the existing scheme (3). Different from the original prediction $\tilde{\lambda}^k = \lambda^k - \beta(A\tilde{x}^k - b)$ as provided in [18], we now introduce the similar prediction steps to reformulate (3) as

$$\begin{cases} \tilde{x}^k = \arg\min\limits_{x\in\mathbb{X}}\left\{L(x,\lambda^k)+\frac{\beta}{2}\left\|Ax-b\right\|^2+\frac{1}{2}\left\|x-x^k\right\|_{D_0}^2\right\}, \\ \tilde{\lambda}^k = \lambda^k - \beta(A\tilde{x}^k-b) - \beta\alpha A(\tilde{x}^k - x^k), \\ \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} - \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \alpha\gamma\beta A & \gamma\mathbf{I} \end{bmatrix} \begin{pmatrix} x^k - \tilde{x}^k \\ \lambda^k - \tilde{\lambda}^k \end{pmatrix}. \end{cases}$$

Analogous to the analysis of sequel Lemma 2, we need to derive the condition to ensure the positive definiteness of the following two new matrices

$$S = \begin{bmatrix} D + (\tau - 1 - \alpha + \alpha^2\gamma)\beta A^\top A & \alpha\gamma A^\top \\ \alpha\gamma A & \frac{\gamma}{\beta}\mathbf{I} \end{bmatrix}$$

and

$$G = \begin{bmatrix} D + (\tau - 1 - \alpha - \alpha^2\gamma)\beta A^\top A & (1-\gamma)\alpha A^\top \\ (1-\gamma)\alpha A & \frac{2-\gamma}{\beta}\mathbf{I} \end{bmatrix}.$$

By a similar analysis to Proposition 1 in Section 2.2, the positive definiteness of these two matrices can be ensured if

$$\tau > \frac{(\alpha + \frac{2-\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4}, \quad \forall\gamma \in (0,2), \alpha \in \mathbb{R}.$$

The region of τ is similar to (6) and τ can be chosen approximating to 0.75 when $\alpha = \frac{\gamma-2}{2}$. That is, the optimal proximal parameter in [18, Section 4] can be obtained too, but here the techniques of convergence analysis are much simpler than before.

1.4 Connections with some primal-dual methods

This subsection aims to provide a concise analysis on the connections between our proposed method and several related primal-dual methods. First of all, the saddle-point reformulation of the problem (1) is

$$\min_{x\in\mathbb{X}}\max_{\lambda\in\mathbb{R}^m} L(x,\lambda) = \theta(x) - \langle\lambda, Ax-b\rangle. \tag{10}$$

- When applying the popular primal-dual hybrid gradient method (PDHG, [8]) to the problem (10), it involves the following iterations

$$\text{(PDHG)}\ \begin{cases} x^{k+1} = \mathbf{prox}_{\eta,\theta}\big[x^k + A^\top \lambda^k/\eta\big], \\ \lambda^{k+1} = \lambda^k - \frac{1}{\sigma}\big[(Ax^{k+1} - b) + A(x^{k+1} - x^k)\big], \end{cases}$$

  where $\eta > 0, \sigma > 0$ are proximal stepsizes satisfying $\eta\sigma > \rho(A^\top A)$. Note that our proposed DP-ALM includes PDHG as a special case if we choose $\tau r = \eta, \beta = \frac{1}{\sigma}$ and $\gamma = 1$. Moreover, it follows from the region of η and σ that $\frac{\tau r}{\beta} > \rho(A^\top A)$, which is much stricter than (17), meaning that our parameters can be chosen more flexible than that in PDHG.

- When applying the balanced ALM presented in [19] to the problem (1) or (10), it takes the following iterative scheme

$$\text{(B-ALM)}\ \begin{cases} x^{k+1} = \mathbf{prox}_{\varrho,\theta}\big[x^k + A^\top \lambda^k/\varrho\big], \\ \lambda^{k+1} = \lambda^k - \big(AA^\top/\varrho + \delta\mathbf{I}\big)^{-1}\big[(Ax^{k+1} - b) + A(x^{k+1} - x^k)\big], \end{cases}$$

  where ϱ, δ are any positive parameters. Our DP-ALM will directly become B-ALM if we select $\gamma = 1, \tau r = \varrho$ and $\beta = \frac{1}{\rho(A^\top A)/\varrho + \delta}$. Combining these relationships, the regions of r and τ in DP-ALM, we further have

$$\delta = \frac{1}{\beta} - \frac{\rho(A^\top A)}{\tau r} \geq \frac{1}{\beta} - \frac{\rho(A^\top A)}{\tau\beta\rho(A^\top A)} = \frac{1}{\beta}\big(1 - 1/\tau\big) \geq -\frac{1}{3\beta}$$

  which, comparing to $\delta > 0$ in [19], implies the flexibility of our method.

- When choosing $\tau r\mathbf{I} = D_1$, the x-subproblem in our DP-ALM amounts to that in the optimal proximal ALM (3). However, the dual update in DP-ALM enjoys an extra iteration $\alpha\beta A(\tilde{x}^k - x^k)$, which is different from the scheme in (3).

- When using the parameterized proximal point algorithm [29] to solve the problem (1) or (10), we have the following iterative scheme

$$\text{(P-PPA)}\ \begin{cases} x^{k+1} = \mathbf{prox}_{\sigma,\theta}\Big[x^k + \frac{1}{\sigma}\big\{A^\top(\lambda^k - \frac{1-t}{s}(Ax^k - b))\big\}\Big], \\ \lambda^{k+1} = \lambda^k - \frac{1}{s}\big[(Ax^{k+1} - b) + tA(x^{k+1} - x^k)\big], \end{cases}$$

  where $t \in \mathbb{R}$ is a scalar, $\sigma > 0, s > 0$ satisfy the condition $\sigma s > \rho(A^\top A)$. By choosing $\gamma = 1, \tau r = \sigma, \beta = \frac{1}{s}$ and $t = 1$, our DP-ALM is equivalent to this P-PPA. However, they are different when these relationships do not hold.

2 Technical preliminaries

In this section, a variational inequality is firstly provided to characterize the saddle-point of the constrained minimization problem (1). Then, the positive definiteness of two important block matrices is analyzed under proper conditions to ensure the global convergence of our proposed method.

2.1 Variational characterization

We begin with the following preliminary lemma about the variational characterization for the first-order optimality condition of composite convex minimization problems.

Lemma 1 *[17] Let $\Phi \subset \mathbb{R}^m$ be a closed convex set and let $f, h : \mathbb{R}^m \longrightarrow \mathbb{R}$ be two convex functions. In addition, h is differentiable. Suppose that the solution set of the problem $\min\{f(x) + h(x) \mid x \in \Phi\}$ is nonempty. Then,*

$$x^* = \arg\min\{f(x) + h(x) \mid x \in \Phi\}$$

if and only if

$$x^* \in \Phi,\ f(x) - f(x^*) + \big\langle x - x^*, \nabla h(x^*)\big\rangle \geq 0,\ \forall x \in \Phi.$$

Let $\Omega := \mathbb{X} \times \mathbb{R}^m$. Then, a point $(x^*, \lambda^*) \in \Omega$ is called the saddle point of (1) if

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*), \quad \forall x \in \mathbb{X}, \lambda \in \mathbb{R}^m.$$

Write these inequalities separately to have

$$\begin{cases} \theta(x) - \theta(x^*) + \big\langle x - x^*, -A^\top \lambda^*\big\rangle \geq 0, & \forall x \in \mathbb{X}, \\ \big\langle \lambda - \lambda^*, Ax^* - b\big\rangle \geq 0, & \forall \lambda \in \mathbb{R}^m, \end{cases}$$

which can be rewritten as the following mixed variational inequality

$$\mathrm{VI}(\theta, \mathcal{J}, \Omega): \quad \theta(x) - \theta(x^*) + \big\langle w - w^*, \mathcal{J}(w^*)\big\rangle \geq 0, \quad \forall w \in \Omega, \tag{11}$$

with

$$w = \begin{pmatrix} x \\ \lambda \end{pmatrix} \quad \text{and} \quad \mathcal{J}(w) = \begin{pmatrix} -A^\top \lambda \\ Ax - b \end{pmatrix}. \tag{12}$$

Note that the above operator $\mathcal{J}$ is affine with a skew-symmetric matrix, thus it holds

$$\big\langle w - \bar{w}, \mathcal{J}(w) - \mathcal{J}(\bar{w})\big\rangle \equiv 0, \quad \forall w, \bar{w} \in \Omega. \tag{13}$$

Since the solution set of (1) is nonempty, the solution set of $\mathrm{VI}(\theta, \mathcal{J}, \Omega)$, denoted by Ω^*, is also nonempty and can be characterized as

$$\Omega^* = \bigcap_{u \in \Omega} \Big\{\bar{u} \mid \theta(u) - \theta(\bar{u}) + \big\langle u - \bar{u}, \mathcal{J}(\bar{u})\big\rangle \geq 0\Big\}. \tag{14}$$

2.2 Basic matrices and properties

Since the matrix M defined in (7) is nonsingular for any $\gamma \in (0, 2)$, to simplify the convergence analysis of our DP-ALM, let's define

$$H = QM^{-1} \quad \text{and} \quad G = Q^\top + Q - M^\top HM, \tag{15}$$

where

$$Q = \begin{bmatrix} \tau r\mathbf{I} & A^\top \\ \alpha A & \frac{1}{\beta}\mathbf{I} \end{bmatrix}. \tag{16}$$

Now, we show that both H and G are positive definite under proper conditions.

Proposition 1 *For any $\gamma \in (0, 2)$ and τ satisfying (6), the matrices H and G defined by (15) are symmetric positive definite.*

Proof. First of all, we have from (6) that $\tau > \alpha$ because

$$\tau > \frac{(\alpha - \frac{\gamma}{2})^2}{2 - \gamma} + \frac{2 + \gamma}{4} \geq \alpha \Longleftrightarrow 4(\alpha - 1)^2 \geq 0 \text{ since } \gamma \in (0, 2).$$

Then, it follows from $\tau > \alpha$ that

$$\alpha\rho(A^\top A) < \frac{\tau\beta\rho(A^\top A)}{\beta} < \frac{\tau r}{\beta}, \tag{17}$$

which ensures the nonsingularity of matrix Q. Together with such property and the nonsingularity of M, we have $S := Q^\top M$ is also nonsingular and symmetric. With the notation S, the block matrices H and G given by (15) can be rewritten as

$$H = QS^{-1}Q^\top \quad \text{and} \quad G = Q^\top + Q - S.$$

Simple algebra shows

$$S = \begin{bmatrix} \tau r\mathbf{I} - (1-\alpha\gamma)\alpha\beta A^\top A & \alpha\gamma A^\top \\ \alpha\gamma A & \frac{\gamma}{\beta}\mathbf{I} \end{bmatrix}$$

and

$$G = \begin{bmatrix} \tau r\mathbf{I} + (1-\alpha\gamma)\alpha\beta A^\top A & (1+\alpha-\alpha\gamma)A^\top \\ (1+\alpha-\alpha\gamma)A & \frac{2-\gamma}{\beta}\mathbf{I} \end{bmatrix}. \tag{18}$$

Because the matrix S is symmetric, we have from the relationship $H = QS^{-1}Q^\top$ that H is also symmetric. Hence, to prove the positive definiteness of H, we only need to demonstrate the positive definiteness of S. Without loss of generality, suppose $m \leq n$ and let $A = V\Sigma U^\top$ be the singular value decomposition of A, where $V \in \mathbb{R}^{m\times m}$ and $U \in \mathbb{R}^{n\times n}$ are orthogonal matrices, $\Sigma = (\Sigma_m, \mathbf{0})$ is a diagonal matrix, and $\Sigma_m = \text{diag}(s_1, s_2, \ldots, s_m) \in \mathbb{R}^{m\times m}$ with $s_i \geq 0 (i = 1, 2, \ldots, m)$ being its singular values. Then, it follows that

$$A^\top A = U \begin{bmatrix} \Sigma_m^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^\top \quad \text{and} \quad AA^\top = V\Sigma_m^2 V^\top.$$

So, the matrix S can be decomposed as

$$S = \begin{bmatrix} U & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix} \underbrace{\begin{bmatrix} \tau r\mathbf{I} - (1-\alpha\gamma)\alpha\beta\Sigma_m^2 & \mathbf{0} & \alpha\gamma\Sigma_m \\ \mathbf{0} & \tau r\mathbf{I} & \mathbf{0} \\ \alpha\gamma\Sigma_m & \mathbf{0} & \frac{\gamma}{\beta}\mathbf{I} \end{bmatrix}}_{P} \begin{bmatrix} U & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix}^\top.$$

By advanced algebra computations (similar techniques can be found in [38, Page 16]), it can be demonstrated that the matrix P is positive definite if and only if

$$\left[\tau r - (1-\alpha\gamma)\alpha\beta s_i^2\right]\frac{\gamma}{\beta} - (\alpha\gamma s_i)^2 > 0, \quad \text{for all } i = 1, \cdots, m,$$

namely, $\left(\frac{\tau r}{\beta} - \alpha s_i^2\right)\gamma > 0$. Since $\gamma \in (0, 2)$ and $\rho(A^\top A) = \max_{i\in\{1,\ldots,m\}} s_i^2 > 0$, the matrix P is positive definite if $\frac{\tau r}{\beta} > \alpha\rho(A^\top A)$, which has been ensured by (6) or precisely (17). Consequently, the matrix H is positive definite if (6) holds.

By an analogous analysis for the matrix G in (18), it can be shown that G is positive definite if (6) holds. Then, the proof is completed. ■

3 Convergence analysis

In this section, we analyze the global convergence of DP-ALM and its sublinear convergence rate in the ergodic and nonergodic sense. We also discuss a possible stopping criterion and the sublinear convergence rate of the optimality gap in a remark.

3.1 Global convergence

Based on the prediction-correction reformulation as in (8a)-(8b) and (9), we next show that the sequence $\{w^* - w^k\}$ is monotone decreasing under the H-weighted norm.

Lemma 2 *Let $\{\tilde{w}^k\}$ and $\{w^{k+1}\}$ be the predictor sequence and corrector sequence generated by DP-ALM, respectively. Then, under the condition (6) it holds $\tilde{w}^k \in \Omega$ and*

$$\theta(x) - \theta(\tilde{x}^k) + \left\langle w - \tilde{w}^k, \mathcal{J}(w) \right\rangle \geq \frac{1}{2}\Big(\left\|w - w^{k+1}\right\|_H^2 - \left\|w - w^k\right\|_H^2\Big) + \frac{1}{2}\left\|w^k - \tilde{w}^k\right\|_G^2 \quad (19)$$

for any $w \in \Omega$, where H and G are given by (15). Moreover, we have

$$\left\|w^* - w^k\right\|_H^2 \geq \left\|w^* - w^{k+1}\right\|_H^2 + \left\|w^k - \tilde{w}^k\right\|_G^2, \quad \forall w^* \in \Omega^*. \quad (20)$$

Proof. According to Lemma 1, the first-order optimality condition of (8a) is $\tilde{x}^k \in \mathbb{X}$ and

$$\theta(x) - \theta(\tilde{x}^k) + \left\langle x - \tilde{x}^k, -A^\top \tilde{\lambda}^k + \tau\beta A^\top A(\tilde{x}^k - x^k) + \tau D(\tilde{x}^k - x^k) + A^\top(\tilde{\lambda}^k - \lambda^k)\right\rangle \geq 0$$

for any $x \in \mathbb{X}$. Besides, the equality (8b) can be rewritten as

$$\Big\langle \lambda - \tilde{\lambda}^k, A\tilde{x}^k - b + \alpha A(\tilde{x}^k - x^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k)\Big\rangle = 0 \quad (21)$$

for any $\lambda \in \mathbb{R}^m$. Combine the last two relationships with the notations in (7), (12) and the matrix Q in (16) to get

$$\begin{aligned} \theta(x) - \theta(\tilde{x}^k) + \left\langle w - \tilde{w}^k, \mathcal{J}(\tilde{w}^k)\right\rangle &\geq (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) \\ &= (w - \tilde{w}^k)^\top H(w^k - w^{k+1}), \end{aligned} \quad (22)$$

where the equality uses the update in (9) and the equality in the left-hand-side of (15). Now, applying the identity

$$(a-b)^\top H(c-d) = \frac{1}{2}\Big(\|a-d\|_H^2 - \|a-c\|_H^2\Big) + \frac{1}{2}\Big(\|c-b\|_H^2 - \|d-b\|_H^2\Big)$$

with $a = w, b = \tilde{w}^k, c = w^k$ and $d = w^{k+1}$ to the right-hand side of (22) gives

$$\begin{aligned} &(w - \tilde{w}^k)^\top H(w^k - w^{k+1}) - \frac{1}{2}\Big(\left\|w - w^{k+1}\right\|_H^2 - \left\|w - w^k\right\|_H^2\Big) \\ &= \frac{1}{2}\Big(\left\|w^k - \tilde{w}^k\right\|_H^2 - \left\|w^{k+1} - \tilde{w}^k\right\|_H^2\Big) \\ &= \frac{1}{2}\Big(\left\|w^k - \tilde{w}^k\right\|_H^2 - \left\|w^{k+1} - w^k + w^k - \tilde{w}^k\right\|_H^2\Big) \\ &\overset{(9)}{=} \frac{1}{2}\Big(\left\|w^k - \tilde{w}^k\right\|_H^2 - \left\|(w^k - \tilde{w}^k) - M(w^k - \tilde{w}^k)\right\|_H^2\Big) \\ &= \frac{1}{2}\big(w^k - \tilde{w}^k\big)^\top\big(Q^\top + Q - M^\top HM\big)\big(w^k - \tilde{w}^k\big) \overset{(15)}{=} \frac{1}{2}\left\|w^k - \tilde{w}^k\right\|_G^2. \end{aligned}$$

Substituting the last relationship into (22) together with (13) confirms the assertion (19).

Finally, setting $w = w^*$ in (19) and using (11) leads to

$$\left\|w^* - w^k\right\|_H^2 - \left\|w^* - w^{k+1}\right\|_H^2 - \left\|w^k - \tilde{w}^k\right\|_G^2 \geq 0.$$

So, (20) follows directly. The proof is completed. ■

We are ready to establish the global convergence of DP-ALM based on Lemma 2.

Theorem 1 *Let $\{w^{k+1}\}$ be the sequence generated by DP-ALM. Then, under the condition (6) we have*

$$\lim_{k\to\infty} \left\|w^k - w^{k+1}\right\| = 0 \tag{23}$$

and there exists a $w^\infty \in \Omega^*$ such that $\lim\limits_{k\to\infty} w^k = w^\infty$.

Proof. It follows from (20) and the positive definiteness of G and H that the sequence $\{w^k\}$ is uniformly bounded and

$$\lim_{k\to\infty} \left\|w^k - \tilde{w}^k\right\| = 0. \tag{24}$$

Combine (24), (9) and the nonsingularity of M to confirm the result in (23).

By the uniformly boundness of $\{w^k\}$ and (9), the sequence $\{\tilde{w}^k\}$ is also uniformly bounded and has at least one limit point $w^\infty = (x^\infty; \lambda^\infty) \in \Omega^*$. Suppose that $\{\tilde{w}^{k_j}\}$ is a subsequence converging to w^∞. Then, it follows from (22) that

$$\theta(x) - \theta(\tilde{x}^{k_j}) + \left\langle w - \tilde{w}^{k_j}, \mathcal{J}(\tilde{w}^{k_j})\right\rangle \geq (w - \tilde{w}^{k_j})^\top Q(w^{k_j} - \tilde{w}^{k_j}), \quad \forall w \in \Omega,$$

which, together with (24), the lower semicontinuity of $\theta(x)$ and the continuity of $\mathcal{J}(w)$, implies

$$\theta(x) - \theta(x^\infty) + \left\langle w - w^\infty, \mathcal{J}(w^\infty)\right\rangle \geq 0, \quad \forall w \in \Omega.$$

In other words, w^∞ is a solution point of VI$(\theta, \mathcal{J}, \Omega)$ and hence is also a solution point of the convex optimization problem (1).

Besides, by (24) and $\lim_{j\to\infty} w^{k_j} = w^\infty$, the sequence $\{w^{k_j}\}$ also converges to w^∞. Then, by (20) again we have

$$\left\|w^\infty - w^{k_j}\right\|_H \geq \left\|w^\infty - w^k\right\|_H \quad \text{for all } k \geq k_j.$$

Hence, the whole sequence $\{w^k\}$ converges to w^∞. ■

3.2 Ergodic convergence rate

Motivated by (14), $\bar{w} \in \Omega^*$ is called an ϵ-approximate solution of VI$(\theta, \mathcal{J}, \Omega)$ with the accuracy $\epsilon > 0$ if it holds

$$\theta(x) - \theta(\bar{x}) + \left\langle w - \bar{w}, \mathcal{J}(w)\right\rangle \geq -\epsilon, \quad \forall w \in \mathcal{B}_{\bar{w}} = \left\{w \in \Omega \mid \|w - \bar{w}\| \leq 1\right\}.$$

To analyze the convergence rate of DP-ALM in terms of the iteration complexity for $\{w^k\}$, we need to show that for given $\epsilon > 0$, after T-th iterations, DP-ALM is able to find a point $\tilde{w} \in \Omega$ such that

$$\sup_{w\in\mathcal{B}_{\bar{w}}} \left\{\theta(\bar{x}) - \theta(x) + \left\langle \bar{w} - w, \mathcal{J}(w)\right\rangle\right\} \leq \epsilon = \mathcal{O}(1/T).$$

Based on Lemma 2, we next establish such an ergodic convergence rate of DP-ALM.

Theorem 2 *Let $\{\tilde{w}^k\}$ be the sequence generated by DP-ALM and H be defined in (15). For any integer number $T > 0$, let*

$$x_T := \frac{1}{T+1}\sum_{k=0}^{T} \tilde{x}^k \quad \text{and} \quad w_T := \frac{1}{T+1}\sum_{k=0}^{T} \tilde{w}^k. \tag{25}$$

Then, under the condition (6) we have

$$\theta(x_T) - \theta(x) + \left\langle w_T - w, \mathcal{J}(w)\right\rangle \leq \frac{1}{2(1+T)}\left\|w - w^0\right\|_H^2, \quad \forall w \in \Omega. \tag{26}$$

Proof. Combing the positive definiteness of G, we can rewrite (19) as

$$\theta(\tilde{x}^k) - \theta(x) + \big\langle \tilde{w}^k - w, \mathcal{J}(w) \big\rangle \leq \frac{1}{2}\Big(\big\|w - w^k\big\|_H^2 - \big\|w - w^{k+1}\big\|_H^2 \Big), \quad \forall w \in \Omega.$$

Summarizing this inequality over $k = 0, 1, ...T$ results in

$$\sum_{k=0}^{T} \theta(\tilde{x}^k) - (1+T)\theta(x) + \Big\langle \sum_{k=0}^{T} \tilde{w}^k - (1+T)w, \mathcal{J}(w) \Big\rangle \leq \frac{1}{2}\big\|w - w^0\big\|_H^2.$$

Namely,

$$\frac{1}{1+T}\sum_{k=0}^{T} \theta(\tilde{x}^k) - \theta(x) + \Big\langle \frac{1}{1+T}\sum_{k=0}^{T} \tilde{w}^k - w, \mathcal{J}(w) \Big\rangle \leq \frac{1}{2(1+T)}\big\|w - w^0\big\|_H^2,$$

which, by the convexity of θ and the definition of x_T and w_T in (25), confirms the result in (26). ■

The above theorem shows that the average of the first T iterates defined in (25) is an approximate solution of $\mathrm{VI}(\theta, \mathcal{J}, \Omega)$ with the rate of $\mathcal{O}(1/T)$. In what follows, a more compact result based on Theorem 2 will be provided, showing that both the objective value gap and the constraint violation will decrease in the order of $\mathcal{O}(1/T)$. Similar theory can be found in [40]. To proceed, for any $\varsigma > 0$, let $\Gamma_\varsigma = \{\lambda \mid \|\lambda\| \leq \varsigma\}$ and

$$\gamma_\varsigma = \inf_{x^* \in \mathbb{X}} \sup_{\lambda \in \Gamma_\varsigma} \Big\| \begin{pmatrix} x^* \\ \lambda \end{pmatrix} - \begin{pmatrix} x^0 \\ \lambda^0 \end{pmatrix} \Big\|_H^2. \tag{27}$$

Corollary 1 *Let γ_ς be defined in (27) and x_T be defined in (25). Then, for any $(x^*; \lambda^*) \in \Omega^*$ and $T > 0$, we have*

$$\big|\theta(x_T) - \theta(x^*)\big| \leq \frac{\gamma_\varsigma}{2(1+T)} \quad \textit{and} \quad \big\|Ax_T - b\big\| \leq \frac{\gamma_\varsigma}{2(1+T)\big(1 + \|\lambda^*\|\big)}. \tag{28}$$

Proof. Set $w = (x^*; \lambda)$ into the inequality (26) to obtain

$$\begin{aligned} \theta(x_T) - \theta(x^*) + \big\langle w_T - w, \mathcal{J}(w) \big\rangle =& \theta(x_T) - \theta(x^*) - \lambda^\top (Ax_T - b) \\ \leq& \frac{1}{2(1+T)} \Big\| \begin{pmatrix} x^* \\ \lambda \end{pmatrix} - \begin{pmatrix} x^0 \\ \lambda^0 \end{pmatrix} \Big\|_H^2, \end{aligned} \tag{29}$$

where the equality uses $Ax^* = b$. Then, we deduce from the last inequality that

$$\theta(x_T) - \theta(x^*) + \varsigma\|Ax_T - b\| = \sup_{\lambda \in \Gamma_\varsigma} \big\{\theta(x_T) - \theta(x^*) - \lambda^\top (Ax_T - b)\big\} \leq \frac{\gamma_\varsigma}{2(1+T)}. \tag{30}$$

By (29) again with (11), we have $\theta(x_T) - \theta(x^*) - \lambda^\top (Ax_T - b) \geq 0$, showing that

$$\theta(x_T) - \theta(x^*) \geq -\big\|\lambda^*\big\|\big\|Ax_T - b\big\|. \tag{31}$$

Then, take $\varsigma = 2\|\lambda^*\| + 1$ in (30) together with (31) to get

$$\big(1 + \big\|\lambda^*\big\|\big)\big\|Ax_T - b\big\| \leq \theta(x_T) - \theta(x^*) + \big(1 + 2\big\|\lambda^*\big\|\big)\big\|Ax_T - b\big\| \leq \frac{\gamma_\varsigma}{2(1+T)}.$$

Rearrange the above inequality to confirm the second inequality in (28). Meanwhile, substitute the second inequality in (28) into (31) to obtain

$$\theta(x_T) - \theta(x^*) \geq -\frac{\gamma_\varsigma}{2(1+T)},$$

which in turn confirms the first inequality in (28). ■

Remark 1 *Let* $P = \begin{bmatrix} \alpha A & \mathbf{0} \\ \mathbf{0} & \gamma \mathbf{I}/\beta \end{bmatrix} M^{-1} = \begin{bmatrix} A/\gamma & \mathbf{0} \\ (1/\gamma - \alpha)A & \mathbf{I}/(\beta\gamma) \end{bmatrix}$. *Next, we analyze the worst-case* $\mathcal{O}(1/T)$ *convergence rate of* $\|Ax_T - b\|$ *from a different viewpoint. Combine the definition of* x_T*, (8b) and (9) to have*

$$Ax_T - b = \frac{1}{1+T}\sum_{k=0}^{T}(A\tilde{x}^k - b) = \frac{1}{1+T}P(w^0 - w^{1+T}),$$

which, by the triangle inequality and (20), implies

$$\|Ax_T - b\| \leq \frac{1}{1+T}\frac{\rho(P)}{\gamma\rho(H)}\Big(\|w^0 - w^*\|_H + \|w^* - w^{1+T}\|_H\Big) \leq \frac{1}{1+T}\frac{2\rho(P)}{\gamma\rho(H)}\|w^0 - w^*\|_H.$$

Here, $\rho(P) = \frac{1}{\gamma}\max\{|\lambda(A)|, 1/\beta\}$ *and* $\lambda(A)$ *is the eigenvalue of* A*. Although the terms in the right-hand-side of the above result and in the second inequality of (28) are different, both of them ensure the ergodic sublinear convergence rate of the constraint violation.*

3.3 Nonergodic convergence rate

Before showing the worst-case $\mathcal{O}(1/t)$ nonergodic convergence rate of DP-ALM in terms of pointwise iterative residual and optimality error, we need to analyze the following preliminary lemma.

Lemma 3 *Let* M *and* H *be given by (7) and (15), respectively. Then, the iterates* $\{w^k\}$ *and* $\{\tilde{w}^k\}$ *generated by DP-ALM satisfy*

$$(w^k - \tilde{w}^k)^\top M^\top H M\{(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\} \geq \frac{1}{2}\|(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\|^2_{Q+Q^\top}. \tag{32}$$

Proof. Setting $w = \tilde{w}^{k+1}$ in (22) results in

$$\theta(\tilde{x}^{k+1}) - \theta(\tilde{x}^k) + \langle \tilde{w}^{k+1} - \tilde{w}^k, \mathcal{J}(\tilde{w}^k) + Q(\tilde{w}^k - w^k)\rangle \geq 0. \tag{33}$$

Meanwhile, it follows from the inequality (22) with $k := k+1$ that

$$\theta(x) - \theta(\tilde{x}^{k+1}) + \langle w - \tilde{w}^{k+1}, \mathcal{J}(\tilde{w}^{k+1}) + Q(\tilde{w}^{k+1} - w^{k+1})\rangle \geq 0,$$

which, by letting $w = \tilde{w}^k$, gives

$$\theta(\tilde{x}^k) - \theta(\tilde{x}^{k+1}) + \langle \tilde{w}^k - \tilde{w}^{k+1}, \mathcal{J}(\tilde{w}^{k+1}) + Q(\tilde{w}^{k+1} - w^{k+1})\rangle \geq 0. \tag{34}$$

Combine (33) and (34) together with the property in (13) to achieve

$$(\tilde{w}^k - \tilde{w}^{k+1})^\top Q\{(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\} \geq 0. \tag{35}$$

Then, by adding the identity

$$\begin{aligned}&\{(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\}^\top Q\{(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\} \\ =&\frac{1}{2}\|(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\|^2_{Q+Q^\intercal}\end{aligned}$$

to both sides of (35), we can get

$$(w^k - w^{k+1})^\top Q\{(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\} \geq \frac{1}{2}\|(w^k - \tilde{w}^k) - (w^{k+1} - \tilde{w}^{k+1})\|^2_{Q+Q^\intercal},$$

which immediately completes the proof based on (9) and the matrix H in (15). ■

Theorem 3 *Let M and H be given by (7) and (15), respectively. Then, for any integer $t>0$ there exists a constant $c>0$ such that the sequences $\{w^k\}$ and $\{\widetilde{w}^k\}$ generated by DP-ALM satisfy*

$$\big\|M(w^k-\tilde{w}^k)\big\|_H^2 \le \frac{1}{(t+1)c}\big\|w^0-w^*\big\|_H^2, \quad \forall w^*\in\Omega^*.$$

Proof. According to Proposition 1 and (20), there exists a constant $c>0$ such that

$$\big\|w^{k+1}-w^*\big\|_H^2 \le \big\|w^k-w^*\big\|_H^2 - c\big\|M(w^k-\tilde{w}^k)\big\|_H^2, \;\; \forall w^*\in\Omega^*, \tag{36}$$

which shows

$$c\sum_{k=0}^{t}\big\|M(w^k-\tilde{w}^k)\big\|_H^2 \le \big\|w^0-w^*\big\|_H^2 \tag{37}$$

for any integer $t>0$. In addition, by applying the following identity

$$\|a\|_H^2-\|b\|_H^2 = 2a^\top H(a-b)-\|a-b\|_H^2, \tag{38}$$

with $a=M(w^k-\tilde{w}^k)$ and $b=M(w^{k+1}-\tilde{w}^{k+1})$, we have

$$\begin{aligned}
&\big\|M(w^k-\tilde{w}^k)\big\|_H^2-\big\|M(w^{k+1}-\tilde{w}^{k+1})\big\|_H^2\\
=&2(w^k-\tilde{w}^k)^\top M^\top HM\big\{(w^k-\tilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\big\}\\
&-\big\|M(w^k-\tilde{w}^k)-M(w^{k+1}-\tilde{w}^{k+1})\big\|_H^2\\
\ge&\big\|(w^k-\tilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\big\|_{Q+Q^\top}^2-\big\|M(w^k-\tilde{w}^k)-M(w^{k+1}-\tilde{w}^{k+1})\big\|_H^2\\
=&\big\|(w^k-\widetilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\big\|_G^2\ge 0,
\end{aligned}$$

where the first inequality follows from Lemma 3 and the final equality uses the definition of G in (15). As a result,

$$\big\|M(w^k-\tilde{w}^k)\big\|_H^2 \ge \big\|M(w^{k+1}-\tilde{w}^{k+1})\big\|_H^2,$$

showing that

$$(t+1)\big\|M(w^t-\tilde{w}^t)\big\|_H^2 \le \sum_{k=0}^{t}\big\|M(w^k-\tilde{w}^k)\big\|_H^2.$$

Substituting it into (37) ends the proof. ■

Remark 2 *For any given $\epsilon>0$, Theorem 3 indicates that our DP-ALM needs at most $[c_1/\epsilon]$ iterations to guarantee $\big\|M(w^k-\tilde{w}^k)\big\|_H^2\le\epsilon$, where $c_1=\inf_{w^*\in\Omega^*}\big\|w^0-w^*\big\|_H^2/c$. According to Theorem 1, w^{k+1} is a solution of $\mathrm{VI}(\theta,\mathcal{J},\Omega)$ if $\big\|w^k-w^{k+1}\big\|=0$, equivalently, $\big\|M(w^k-\tilde{w}^k)\big\|_H=0$. Hence, a possible stopping criterion is $\big\|w^k-w^{k+1}\big\|\le\epsilon$. Because of the monotonicity of $\{\|M(w^k-\tilde{w}^k)\|_H^2\}$ and the result in Theorem 3, by using [12, Lemma 1.1] we can refine the worst-case convergence in Theorem 3 from $\mathcal{O}(1/t)$ to $o(1/t)$. In addition, by letting $d^k=(d_x^k;d_\lambda^k)$ where*

$$\left\{\begin{array}{l}
d_x^k=\tau\beta A^\top A(\tilde{x}^k-x^k)+\tau D(\tilde{x}^k-x^k)+A^\top(\tilde{\lambda}^k-\lambda^k),\\
d_\lambda^k=\alpha A(\tilde{x}^k-x^k)+\frac{1}{\beta}(\tilde{\lambda}^k-\lambda^k),
\end{array}\right. \tag{39}$$

we have $A\tilde{x}^k-b+d_\lambda^k=\mathbf{0}$ and

$$\theta(x)-\theta(\tilde{x}^k)+\big\langle x-\tilde{x}^k,-A^\top\tilde{\lambda}^k+d_x^k\big\rangle\ge 0, \quad \forall x\in\mathbb{X},$$

or equivalently $A^\top \tilde{\lambda}^k - d_x^k \in \partial\theta(\tilde{x}^k) + \mathcal{N}_{\mathbb{X}}(\tilde{x}^k)$. Here, $\|d_x^k\|$ measures the first-order optimality error, $\mathcal{N}_{\mathbb{X}}(x)$ denotes the normal cone of $\mathbb{X}$ at x, and $\partial f(x)$ denotes the subdifferential of f at x. Notice that (39) can be rewritten as $d^k = Q(u^k - \tilde{u}^k) = H(u^k - u^{k+1})$. So,

$$\|d^k\| = \|H(u^k - u^{k+1})\| \leq \lambda_{\max}(H)\|u^k - u^{k+1}\|_H,$$

which, by Theorem 3 and (9), implies $\|d^k\|$ converges to zero in a sublinear rate.

4 More discussions on DP-ALM

In this section, we first take an example to illustrate that the lower bound $\frac{2+\gamma}{4}$ implied in Section 1.3 is optimal(smallest) and it is impossible to find a lower bound smaller than $\frac{2+\gamma}{4}$. Then, we briefly discuss a relaxed version and a linearized version of the fundamental DP-ALM as well as their convergence properties.

4.1 Optimality of the formula (6)

Consider the example mentioned in [18, Section 4], that is, the simplest equation $x = 0$ in $\mathbb{R}$, and we will show that DP-ALM is not necessarily convergent when $\tau < \frac{2+\gamma}{4}$. Clearly, $x = 0$ is a special case of the model (1):

$$\min_{\mathbb{R}} \big\{ 0 \cdot x \mid x = 0 \big\}. \tag{40}$$

Without loss of generality we take $\beta = 1$. Then, DP-ALM for solving (40) reads

$$\left\{ \begin{array}{l} x^{k+1} = \arg\min\limits_{x\in\mathbb{R}} \big\{ -x\lambda^k + \frac{\tau}{2}\big(x - x^k\big)^2 + \frac{\tau(r-1)}{2}\big(x - x^k\big)^2 \big\} = \frac{\lambda^k}{\tau r} + x^k, \\ \lambda^{k+1} = \lambda^k - \big(\gamma x^{k+1} + x^{k+1} - x^k\big) = \frac{\tau r - 1 - \gamma}{\tau r}\lambda^k - \gamma x^k. \end{array} \right. \tag{41}$$

By setting $\bar{\tau} = \tau r$, we can rewrite the above updates as

$$w^{k+1} = \varphi(\bar{\tau}) w^k \qquad \text{with} \qquad \varphi(\bar{\tau}) = \begin{bmatrix} 1 & \frac{1}{\bar{\tau}} \\ -\gamma & \frac{\bar{\tau} - 1 - \gamma}{\bar{\tau}} \end{bmatrix}.$$

Let $f_1(\bar{\tau})$, $f_2(\bar{\tau})$ be the two eigenvalues of the matrix $\varphi(\bar{\tau})$. Simple algebra shows

$$f_1(\bar{\tau}) = 1 + \frac{-1 - \gamma + \sqrt{(1+\gamma)^2 - 4\gamma\bar{\tau}}}{2\bar{\tau}} \quad \text{and} \quad f_2(\bar{\tau}) = 1 + \frac{-1 - \gamma - \sqrt{(1+\gamma)^2 - 4\gamma\bar{\tau}}}{2\bar{\tau}}.$$

For the function $f_2(\bar{\tau})$, we have $f_2\Big(\frac{2+\gamma}{4}\Big) = -1$ and

$$f_2'(\bar{\tau}) = \frac{1}{4\bar{\tau}^2}\left(\frac{4\gamma\bar{\tau}}{\sqrt{(1+\gamma)^2 - 4\gamma\bar{\tau}}} + 2(1+\gamma) + 2\sqrt{(1+\gamma)^2 - 4\gamma\bar{\tau}} \right).$$

So, for any $\gamma \in (0, 2)$ and $\bar{\tau} \in \big(0, \frac{2+\gamma}{4}\big)$, we obtain $(1+\gamma)^2 - 4\gamma\bar{\tau} > 0$ and $f_2'(\bar{\tau}) > 0$. Consequently,

$$f_2(\bar{\tau}) < f_2\Big(\frac{2+\gamma}{4}\Big) = -1, \quad \text{for any } \bar{\tau} \in \Big(0, \frac{2+\gamma}{4}\Big).$$

Since here $r > \beta = 1$, combine the definition of $\bar{\tau}$ and its region to have $\tau \in \big(0, \frac{2+\gamma}{4}\big)$. So, for any $\tau \in \big(0, \frac{2+\gamma}{4}\big)$, the matrix $\varphi(\bar{\tau})$ has an eigenvalue less than -1. Therefore, the iterative scheme in (41), that is the application of DP-ALM to the problem (40), is not necessarily convergent for any $\tau \in \big(0, \frac{2+\gamma}{4}\big)$. In other words, $\frac{2+\gamma}{4}$ is the smallest lower bound of τ to ensure the convergence of DP-ALM.

4.2 Relaxed version of DP-ALM

This subsection aims to extend the previous DP-ALM to the following relaxed accelerated version and provide a concise convergence analysis:

$$\text{(RP-ALM)}\quad \begin{cases} \hat{x}^k = \arg\min\limits_{x\in\mathbb{X}} \left\{\theta(x) - \langle \lambda^k, Ax\rangle + \frac{\tau\beta}{2}\left\|A(x-x^k)\right\|^2 + \frac{\tau}{2}\left\|x-x^k\right\|_D^2\right\}, \\ \hat{\lambda}^k = \lambda^k - \beta\gamma(A\hat{x}^k - b) - \beta A(\hat{x}^k - x^k), \\ \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \eta \begin{pmatrix} \hat{x}^k - x^k \\ \hat{\lambda}^k - \lambda^k \end{pmatrix}, \end{cases} \tag{42}$$

where $\eta \in (0,2)$ denotes the relaxation factor satisfying $\gamma\eta \in (0,2)$, and the rest parameters are the same as before. When $\eta = 1$, RP-ALM reduces to the previous DP-ALM.

Analogous to the aforementioned analysis in Lemma 2, it is not difficult (for details, see e.g. [4]) to show

$$\theta(x) - \theta(\tilde{x}^k) + \left\langle w - \tilde{w}^k, \mathcal{J}(w)\right\rangle \geq \frac{1}{2\eta}\left(\left\|w - w^{k+1}\right\|_H^2 - \left\|w - w^k\right\|_H^2\right) + \frac{1}{2}\left\|w^k - \tilde{w}^k\right\|_G^2 \tag{43}$$

and

$$\left\|w^* - w^k\right\|_H^2 \geq \left\|w^* - w^{k+1}\right\|_H^2 + \eta\left\|w^k - \tilde{w}^k\right\|_G^2. \tag{44}$$

Here

$$\tilde{x}^k = \hat{x}^k, \quad G = Q^\top + Q - \eta M^\top H M,$$

and the rest notations are the same as before. To ensure the monotonicity of the sequence $\{\left\|w^* - w^k\right\|_H^2\}$, we just need to derive the condition to ensure the positive definiteness of G. Similar to the analysis in Proposition 1, G is positive definite if the proximal parameter τ satisfies

$$\tau > \frac{[\alpha + \frac{2-2\eta-2\gamma\eta+\gamma\eta^2}{2}]^2}{(2-\eta)(2-\gamma\eta)} + \frac{\eta(\eta^2\gamma - 4\eta\gamma - 2\eta + 4\gamma + 4)}{4(2-\eta)} \tag{45}$$

for any $\alpha \in [0,1)$. The region of τ indicates

$$\tau > \frac{[\alpha + \frac{2-2\eta-2\gamma\eta+\gamma\eta^2}{2}]^2}{(2-\eta)(2-\gamma\eta)} + \frac{\eta(\eta^2\gamma - 4\eta\gamma - 2\eta + 4\gamma + 4)}{4(2-\eta)} \geq \alpha,$$

which in turn guarantees the positive definiteness of the matrix H. When $\eta = 1$, the inequality (45) reduces to the previous in (6); when $\gamma = 1$ and $\tau = 1$, our RP-ALM reduces to the method in [4]. However, the parameter τ in RP-ALM could be smaller than 1, and hence our RP-ALM is more general than the previous. Besides, simple algebra shows that (45) amounts to

$$\tau > \frac{\alpha^2\gamma\eta - \alpha\eta}{2-\eta} + \frac{[(1-\gamma\eta)\alpha + 1)]^2}{(2-\eta)(2-\gamma\eta)}. \tag{46}$$

We observe that:

- If $\alpha = 0$, this region reduces to $\tau > \frac{1}{(2-\eta)(2-\gamma\eta)}$. By selecting $\eta \to 0$, the lower bound of τ approximates to $1/4$ which is half of the bound $1/2$ as shown by (6). This lower bound seems to be the smallest one in the literature.

- If $\gamma\eta = 1$, this region reduces to $\tau > \frac{(\alpha - \frac{\eta}{2})^2}{2-\eta} + \frac{2+\eta}{4}$. By selecting $\alpha = \frac{\eta}{2}$ and $\eta \to 1/2$, the lower bound of τ approximates to $5/8$, which is also smaller than $3/4$ as discussed in Section 4.1 and [6, 18] as well as the region $\frac{13-2\sqrt{13}}{9}$ in [30].

The above discussions indicate that the region of proximal parameter τ could be significantly reduced by exploiting a relaxed acceleration step in the original algorithm. Exactly, the lower bound of τ in the relaxed method will be a half of that in the method without the relaxation step. We guess this conjecture holds for other first-order proximal methods such as the proximal point method, proximal alternating direction method and primal-dual hybrid gradient method. But, as we tested in experiments, these settings seem not better than other parameter settings when applying DP-ALM and RP-ALM to solve some sparse optimization problems.

Finally, we have from (43) and (44) that RP-ALM converges globally with sublinear ergodic/nonergodic convergence rates, whose proofs are similar to the analysis in Section 3 and thus are omitted for the sake of conciseness.

4.3 Linearized version of DP-ALM

In this subsection, we consider a composite case of the problem (1), that is,

$$\min \theta(x) := f(x) + g(x) \qquad \text{s.t. } Ax = b, x \in \mathbb{X}, \tag{47}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function whose gradient is Lipschitz continuous with constant L_f, g is a proper lower semicontinuous convex function (possibly nonsmooth), and the rest notations follow the same meanings as that in (1). For the problem (47), let's consider the following linearized DP-ALM (abbreviated as LP-ALM):

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{X}} \left\{ g(x) + \langle x, \nabla f(x^k) - A^\top \lambda^k \rangle + \frac{\tau\beta}{2} \|A(x - x^k)\|^2 + \frac{\tau}{2} \|x - x^k\|_D^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta\big[\gamma(Ax^{k+1} - b) + A(x^{k+1} - x^k)\big]. \end{cases} \tag{48}$$

Since $D = r\mathbf{I} - \beta A^\top A$ with $r > \beta\rho(A^\top A)$, the key subproblem in (48) amounts to $x^{k+1} = \mathbf{prox}_{\tau r,g}\big[x^k + (A^\top \lambda^k - \nabla f(x^k))/(\tau r)\big]$. Note that the above algorithm will reduce to the Condat-Vu method [11, 37] by forcing $r = \gamma = 1$.

Next, we briefly analyze the parameter region to ensure the convergence of LP-ALM. Similar to the reformulation in Section 1.3, we denote the output of (48) as $\tilde{x}^k$ and $\tilde{\lambda}^k$ respectively, together with a correction step (9). Then, by the first-order optimality condition of the subproblem in (48), we have

$$h(x) - h(\tilde{x}^k) + \langle x - \tilde{x}^k, \nabla f(x^k) - A^\top \tilde{\lambda}^k + \tau r(\tilde{x}^k - x^k) + A^\top(\tilde{\lambda}^k - \lambda^k) \rangle \geq 0$$

for any $x \in \mathbb{X}$. Note that, the convexity of θ_1 and its Lipschitz continuity implies

$$\begin{aligned} \langle x - \tilde{x}^k, \nabla f(x^k) \rangle &= \langle x - x^k, \nabla f(x^k) \rangle + \langle x^k - \tilde{x}^k, \nabla f(x^k) \rangle \\ &\leq f(x) - f(x^k) + f(x^k) - f(\tilde{x}^k) + \frac{L_f}{2} \|x^k - \tilde{x}^k\|^2 \\ &= f(x) - f(\tilde{x}^k) + \frac{L_f}{2} \|x^k - \tilde{x}^k\|^2. \end{aligned}$$

Hence, combining the notation $\theta(x) := f(x) + g(x)$ it holds that

$$\theta(x) - \theta(\tilde{x}^k) + \langle x - \tilde{x}^k, -A^\top \tilde{\lambda}^k + \tau r(\tilde{x}^k - x^k) + A^\top(\tilde{\lambda}^k - \lambda^k) \rangle \geq -\frac{L_f}{2} \|x^k - \tilde{x}^k\|^2,$$

which, together with the previous equality (21), gives

$$\theta(x) - \theta(\tilde{x}^k) + \langle w - \tilde{w}^k, \mathcal{J}(\tilde{w}^k) \rangle \geq (w - \tilde{w}^k)^\top H(w^k - w^{k+1}) - \frac{L_f}{2} \|x^k - \tilde{x}^k\|^2,$$

where the notations $w, \tilde{w}^k, \mathcal{J}, H$ are the same as before. Analogous to the proof of Lemma 2, we deduce

$$\left\|w^* - w^k\right\|_H^2 \geq \left\|w^* - w^{k+1}\right\|_H^2 + \left\|w^k - \tilde{w}^k\right\|_{\bar{G}}^2, \quad \forall w^* \in \Omega^*. \tag{49}$$

Here $\bar{G} = G - \mathrm{diag}\big(L_f/2\mathbf{I}, \mathbf{0}\big)$ and G is given by (15). Similar to the analysis in Proposition 1, the block matrix G is positive definite if

$$\tau > \frac{(\alpha - \frac{\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4} + \frac{L_f}{2\beta\rho(A^\top A)}, \quad \forall \gamma \in (0,2), \alpha \in \mathbb{R}.$$

The inequality in (49) indicates the global convergence of LP-ALM (48).

5 Numerical experiments

In this section, we apply the proposed augmented Lagrangian methods to solve two kinds of large-scale sparse optimization problems, aiming to evaluate the performance and robustness of our methods on synthetic and public data. All experiments are implemented in MATLAB R2020b (64-bit) and performed on a PC with Windows 11 operating system, with an AMD Ryzen 7 8845H w/Radeon 780M Graphics and 64GB RAM.

5.1 Signal recovery problem

Consider the sparse signal recovery problem, as stated in Example 2, with an original signal $x_{\text{orig}} \in \mathbb{R}^n$ containing $m/50$ spikes with amplitude ±1. The measurement matrix $A \in \mathbb{R}^{m\times n}$ is drawn firstly from the standard norm distribution $\mathcal{N}(0,1)$ and then each of its columns is normalized; the vector $b = A * x_{\text{orig}} + 0.01 * \mathrm{randn}(m,1)$. More details on the problem data can be found in the codes of [5, Section 4.1]. Applying our preliminary DP-ALM (5) to this problem results in the following iterations:

$$\left\{\begin{array}{l} x^{k+1} = \mathbf{prox}_{\tau r, \|x\|_1}\left[x^k + A^\top \lambda^k/(\tau r)\right], \\ \lambda^{k+1} = \lambda^k - \beta\left[\gamma(Ax^{k+1} - b) + A(x^{k+1} - x^k)\right]. \end{array}\right. \tag{50}$$

Applying our RP-ALM to Example 2 results in the above iterations plus a relaxation step as in (42), where $\mathbf{prox}_{\tau r, \|x\|_1}(\cdot)$ can be explicitly obtained by the built-in MATLAB function "`wthresh`". Note that the classical ALM (2) can not be used for solving Example 2 since the resulting subproblem is as difficult as the original and has no closed-form-solution, while some linearized ALM-type methods including (3) and ours can be used directly. We compare our algorithms DP-ALM, RP-ALM with the following existing algorithms with tuned parameters involved:

- Optimal Proximal ALM (OP-ALM, [18]) with parameters $(\beta, \tau, \gamma) = (3, 0.751, 1)$;
- Generalized Primal-Dual Algorithm (G-PDA, [21]) with parameters $(r, s, v) = (\sqrt{0.75\rho(A^\top A)}/v, v\sqrt{0.75\rho(A^\top A)}, 0.1)$;
- Customized PPA (C-PPA, [20]) with parameters $(\gamma, r, s) = (1.8, 8, \frac{1.01}{r}\rho(A^\top A))$;
- Parameterized PPA (P-PPA, [29]) with parameters $(t, \sigma, s) = (-1, 8, \frac{1.01}{\sigma}\rho(A^\top A))$.

The parameters of DP-ALM use $(\beta, \gamma, \tau, r) = (23, 1.9, \frac{2+\gamma}{4} + 10^{-3}, \beta\rho(A^\top A)(1 + 10^{-3}))$, $D = r\mathbf{I} - \beta A^\top A$ with $r = 1.001\beta\rho(A^\top A)$; and the parameters of RP-ALM use $(\beta, \gamma, \eta, \tau, r)$ $= (23, 1.9, 1.06, \frac{\eta(\eta^2\gamma - 4\eta\gamma - 2\eta + 4\gamma + 4)}{4(2-\eta)} + 10^{-3}, \beta\rho(A^\top A)(1 + 10^{-3}))$. All of these parameters are tuned to satisfy their convergence region by a for-loop and then relatively reasonable values are selected when costing smaller iteration numbers and CPU time.

Size	**DP-ALM**			**RP-ALM**		
(m,n)	Iter	CPU	Equ_err	Iter	CPU	Equ_err
(1000,3000)	*164*	**0.71**	9.88e-6	**154**	*0.77*	9.89e-6
(2000,6000)	*219*	**3.66**	9.94e-6	**197**	*3.81*	9.78e-6
(3000,9000)	*232*	**8.45**	9.73e-6	**216**	*9.05*	9.84e-6
(4000,12000)	*246*	**15.69**	9.94e-6	**228**	*16.92*	9.88e-6
(5000,15000)	*263*	**26.06**	9.93e-6	**239**	*27.39*	9.96e-6
(6000,18000)	*278*	**39.60**	9.90e-6	**256**	*42.38*	9.96e-6
(7000,21000)	*295*	**56.23**	9.99e-6	**271**	*59.92*	9.93e-6
(8000,24000)	*301*	**75.21**	9.97e-6	**274**	*79.56*	9.92e-6
(9000,27000)	*305*	**98.15**	9.99e-6	**286**	*107.40*	9.96e-6
(10000,30000)	*325*	**127.09**	9.99e-6	**298**	*135.22*	9.87e-6
Size	**C-PPA**			**G-PDA**		
(m,n)	Iter	CPU	Equ_err	Iter	CPU	Equ_err
(1000,3000)	1000	1.66	9.93e-6	460	11.24	9.96e-6
(2000,6000)	1255	10.93	9.98e-6	584	11.35	9.99e-6
(3000,9000)	1329	23.99	9.97e-6	639	26.68	9.97e-6
(4000,12000)	1522	48.51	9.95e-6	697	51.96	9.99e-6
(5000,15000)	1613	79.06	9.99e-6	751	85.87	9.99e-6
(6000,18000)	1775	125.09	9.99e-6	796	131.16	9.99e-6
(7000,21000)	1780	166.87	9.99e-6	828	182.91	9.98e-6
(8000,24000)	1804	222.76	9.99e-6	881	256.34	9.98e-6
(9000,27000)	1902	310.34	9.99e-6	893	337.03	9.95e-6
(10000,30000)	1934	370.47	9.94e-6	898	408.16	9.96e-6
Size	**P-PPA**			**OP-ALM**		
(m,n)	Iter	CPU	Equ_err	Iter	CPU	Equ_err
(1000,3000)	1636	12.79	9.99e-6	601	2.97	9.93e-6
(2000,6000)	2060	63.22	9.98e-6	755	14.88	9.93e-6
(3000,9000)	2174	142.74	9.98e-6	791	33.96	9.99e-6
(4000,12000)	2559	300.92	9.99e-6	930	68.81	9.97e-6
(5000,15000)	2631	471.54	9.99e-6	945	107.82	9.99e-6
(6000,18000)	2905	754.96	9.98e-6	1058	174.69	9.99e-6
(7000,21000)	2891	1005.89	9.99e-6	1063	234.78	9.95e-6
(8000,24000)	2920	1333.88	9.99e-6	1072	338.54	9.99e-6
(9000,27000)	3113	1821.88	9.99e-6	1130	424.11	9.99e-6
(10000,30000)	3118	2236.69	9.99e-6	1136	515.07	9.99e-6

Table 1: Comparison of different algorithms for signal recovery problem.

Table 1 reports the final results of the above algorithms for solving signal recovery problem with different sizes of A, including the iteration numbers (Iter), the CPU time in seconds (CPU), and the equality constraint error (Equ_err). All tests are terminated when the stopping criterion $\text{Equ\_err}(k) = \|Ax^k - b\|^2 < \epsilon$ is satisfied under the maximal iteration numbers 5000, where $0 < \epsilon \ll 1$ is a given tolerance. Figure 1 also deficits four comparison curves of log(Equ_err) vs the iteration numbers under tolerances $\epsilon \in \{10^{-5}, 10^{-7}\}$, respectively. In addition, Figures 2-3 show the recovery quality of sparse signal by different algorithms when the problem sizes are $(m, n) = (1000, 3000)$ and $(m, n) = (10000, 30000)$, respectively.

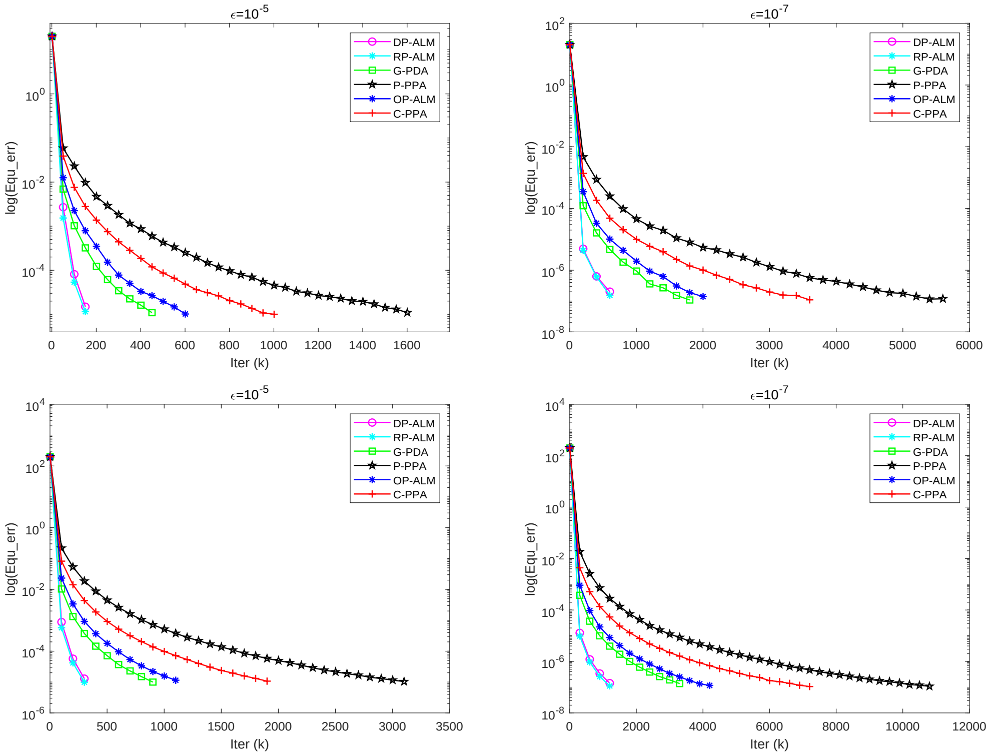

Figure 1: Comparison curves of log(Equ_err) by different algorithms for solving signal recovery problem with $(m, n) = (1000, 3000)$(top) and $(m, n) = (10000, 30000)$(bottom).

Firstly, the reported results in Table 1 demonstrate that all algorithms are feasible to solve the sparse signal recovery problems as the problem size increases, especially for large-scale problems (note that the dimension of signal increases from 3000 to 30000). Secondly, it can be seen from Table 1 and Figure 1 that the proposed algorithms DP-ALM and RP-ALM significantly outperform other well-established methods in terms of the iteration number and CPU time, which is perhaps due to the double proximal terms and the merits of dual update proposed in this paper. Moreover, performance of our two methods have been demonstrated by Figure 1 whenever higher or lower tolerance is required. Thirdly, our relaxed version, that is RP-ALM, performs significantly better than the fundamental algorithm DP-ALM, which demonstrates the accelerated performance of using relaxation step. Last but not the least, as can be seen from Figures 2-3, once the positions of the non-zero elements in the reconstructed signal are accurately identified, the resulting signal is capable of precisely replicating the number of spikes and closely resembling the original signal.

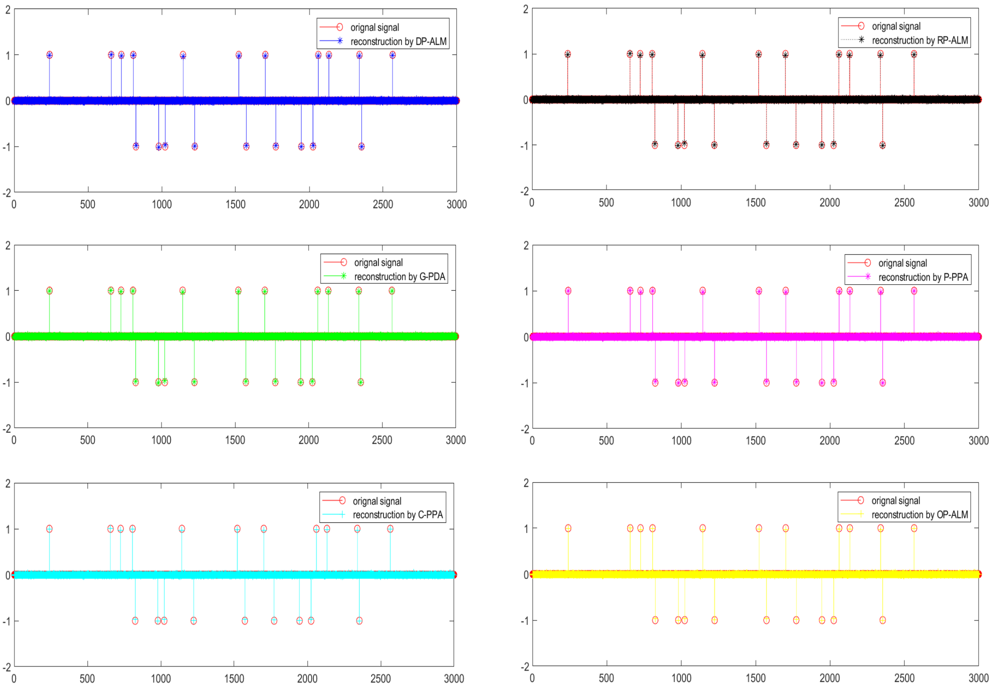

Figure 2: Comparison between the original signal and reconstructed signal by different algorithms for solving signal recovery problem with $(m, n) = (1000, 3000)$.

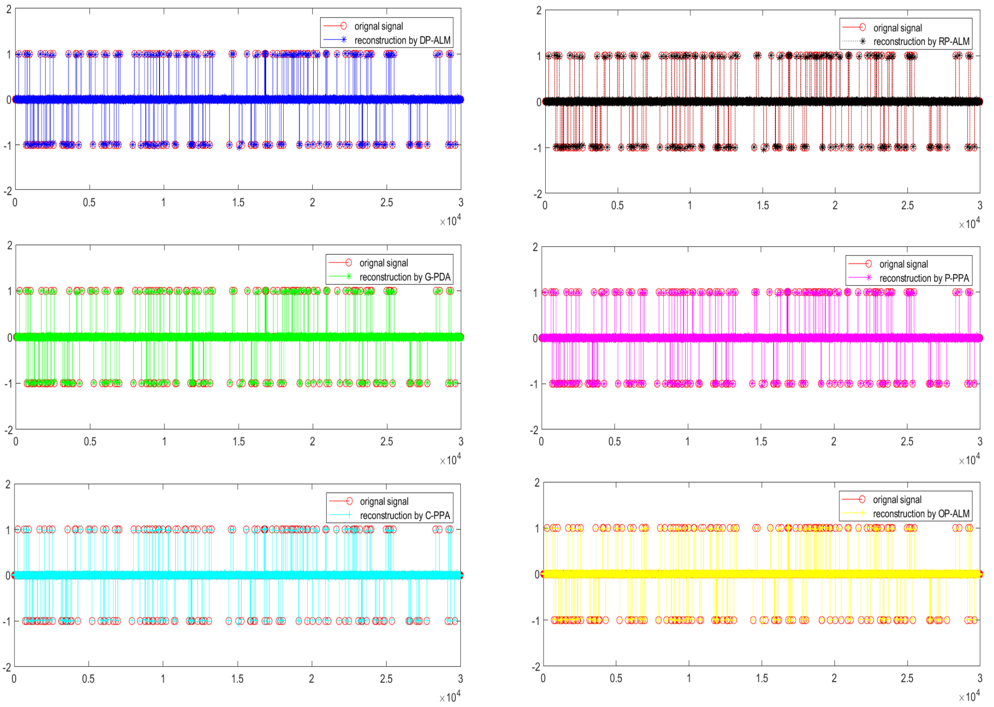

Figure 3: Comparison between the original signal and reconstructed signal by different algorithms for solving signal recovery problem with $(m, n) = (10000, 30000)$.

5.2 Decentralized composite optimizition over networks

Consider an undirected and connected network $\mathcal{G}(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{v_1,\cdots,v_N\}$ denotes the vertex set, and the edge set $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$ specifies the connectivity in the network, namely, a communication link between agents i and j exists iff $(i,j)\in\mathcal{E}$. Denote $x_i\in\mathbb{R}^n$ be the decision variable held by the agent i. Then, the problem in the form of Example 3 is built by introducing $x_1=\cdots=x_N$ that is equivalent to $(\mathbf{I}-\mathbf{W})\mathbf{x}=\mathbf{0}$. In the following experiments, we take $g_i(x_i)=\nu_{1,i}\|x_i\|+\frac{\nu_{2,i}}{2}\|x_i\|^2$ as the default regularizer term in Example 3 with $\nu_{1,i}$ and $\nu_{2,i}$ being the regularization parameters, while two different kinds of loss functions are considered:

- $f_i(x_i)=\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{2}\|\mathbf{a}_{ij}^{\mathsf{T}}x_i-\mathbf{b}_{ij}\|^2$ for decentralized linear regression;
- $f_i(x_i)=\frac{1}{m_i}\sum_{j=1}^{m_i}\ln\left(1+e^{-(\mathbf{a}_{ij}^{\mathsf{T}}x_i)\mathbf{b}_{ij}}\right)$ for decentralized logistic regression.

Here, any agent i holds its own training data $(\mathbf{a}_{ij},\mathbf{b}_{ij})\in\mathbb{R}^n\times\{-1,1\}, j=1,\cdots,m_i$ including sample vectors $\mathbf{a}_{ij}$ and corresponding classes $\mathbf{b}_{ij}$. Due to the composite structure of Example 3, it is challenging to obtain an analytical solution when using DP-ALM. However, the extended linearized version (i.e., LP-ALM) in Section 4.3 can be applied to transform the problem into a more tractable form, that is,

$$\min_{\mathbf{x}}\theta(\mathbf{x})=\underbrace{\sum_{i=1}^{N}\Big(f_i(x_i)+\frac{\nu_{2,i}}{2}\|x_i\|^2\Big)}_{:=\varphi(\mathbf{x})}+\underbrace{\sum_{i=1}^{N}\nu_{1,i}\|x_i\|}_{:=\phi(\mathbf{x})},\quad \text{s.t. } \underbrace{(\mathbf{I}-\mathbf{W})}_{A}\mathbf{x}=\mathbf{0}. \tag{51}$$

As a result, applying LP-ALM to the composite minimization problem (51) reads:

$$\begin{cases}\mathbf{x}^{k+1}=\mathbf{prox}_{\tau r,\phi}\big[\mathbf{x}^k+\frac{-\nabla\varphi(\mathbf{x}^k)+(\mathbf{I}-\mathbf{W})^\top\lambda^k}{\tau r}\big],\\ \lambda^{k+1}=\lambda^k-\beta[\gamma(\mathbf{I}-\mathbf{W})\mathbf{x}^{k+1}+(\mathbf{I}-\mathbf{W})(\mathbf{x}^{k+1}-\mathbf{x}^k)].\end{cases}$$

The proximal operator ϕ admits a closed-form solution by the proximal operator of $\|\cdot\|$. Similarly, the related version of LP-ALM (denoted by RLP-ALM) results in the above iterations plus a relaxation step.

We will compare the customized methods LP-ALM and RLP-ALM with two advanced methods D-iPGM [15] and NIDS [27] for solving (51), where the mixing matrix W involved in $\mathbf{W}=W\otimes\mathbf{I}$ is generated by the Metropolis-Hastings rule [36, Sec. 2.4]. We initially establish $N=100$ agents and then uniformly and randomly distribute the sample data across these agents, and finally run these agents through a randomly generated connected network with $\frac{0.1N(N-1)}{2}$ undirected edges. Three public datasets (see Table 2) from the LIBSVM website are used for this experiments. The tuned parameters in our LP-ALM and RLP-ALM are listed in Table 2, and we choose $\nu_{1,i}=0.01,\nu_{2,i}=1$ as default regularization parameters. The relaxation parameter η in RLP-ALM is taken as 1.8 for `ijcnn1`, 0.92 for `a7a` and 0.95 for `covtype`, respectively. The initial point is set to be $\mathbf{x}=\mathbf{0}$. The involved parameters in D-iPGM and NIDS follow the original settings: $[\tau,\beta]=[1.1602,0.4310]$ for `ijcnn1`, $[\tau,\beta]=[0.0327,15.2905]$ for `a7a` and $[\tau,\beta]=[0.7626,0.6557]$ for `covtype`.

It is noteworthy that only one round of communication is involved in each iteration of the comparison algorithms. The amount of information exchanging over the network is directly proportional to the number of iterations. Therefore, in the performance evaluation, we only record the number of iterations as a metric. The number of iterations for each experiment is the same, five experiments are performed, and the average is taken as our results. Figures 4-5 plot iterative error $\|\mathbf{x}^k-\mathbf{x}^*\|/\|\mathbf{x}^*\|$ and constraint violation

Dataset	Number of samples	Dimensionality	β	γ	τ	r
`ijcnn1`	55500	22	1.80	1.10	1.55	3.25
`a7a`(linear)	16100	122	3.50	1.95	1.73	5.78
`a7a`(logistic)	16100	122	2.05	1.95	2.07	3.39
`covtype`	581012	44	2.11	1.98	1.38	3.50

Table 2: Real-world datasets and algorithmic parameters used in the experiments.

and $\|(\mathbf{I}-\mathbf{W})\mathbf{x}^k\|$, versus numbers, where $\mathbf{x}^\star$ denotes the solution of (51) using centralized approaches. We can see that both LP-ALM and RLP-ALM perform better than D-iPGM and NIDS, and RLP-ALM performs sometimes significantly better than LP-ALM. This fact suggests that the relaxed accelerated step may improve the algorithms in some particular problems.

6 Concluding remarks

In this article, several variants of the proximal augmented Lagrangian method have been developed for solving linearly constrained convex programming problems. We have analyzed the connections between our proposed method and other well-established methods in the literature, and we also established the global convergence and ergodic/nonergodic convergence rates of the fundamental method called DP-ALM. A notable lightspot is that these convergence results are analyzed based on a novel prediction technique and hence the involved proximal parameter can enjoy the smallest lower bound. This DP-ALM is also extended to relaxed version, linearized version and multi-block splitting version. Preliminary experiments on testing two large-scale sparse minimization problems verify the performance and robustness of our methods. In this future work, we wish to extend the proposed methods to the general separable nonconvex minimization problem subject to linear constraints [14] and the general nonsmooth nonconvex-linear minimax optimization problem [43].

Appendix: a multi-block extension

For theoretical interests, in this appendix we will extend DP-ALM to solve the following multiple-block separable convex optimization problem

$$\min\left\{\theta(x) := \sum_{i=1}^{p}\theta_i(x_i) \;\Big|\; \sum_{i=1}^{p} A_i x_i = b,\ x_i \in \mathbb{X}_i\right\}, \tag{52}$$

where $\theta_i : \mathbb{R}^{n_i} \to \mathbb{R}(i = 1, 2, \cdots, p)$ are proper lower semicontinuous convex functions (possibly nonsmooth, non-Lipschitz continuous, and non-strongly convex), $\mathbb{X}_i \subseteq \mathbb{R}^{n_i}$ are closed convex sets, $A_i \in \mathbb{R}^{m\times n_i}$ and $b \in \mathbb{R}^m$ are given. Note that problem (52) includes the generic unconstrained problem [9] $\min_{x\in\mathbb{R}^n}\{f(x) + g(Ax) + h(x)\}$ as a special case.

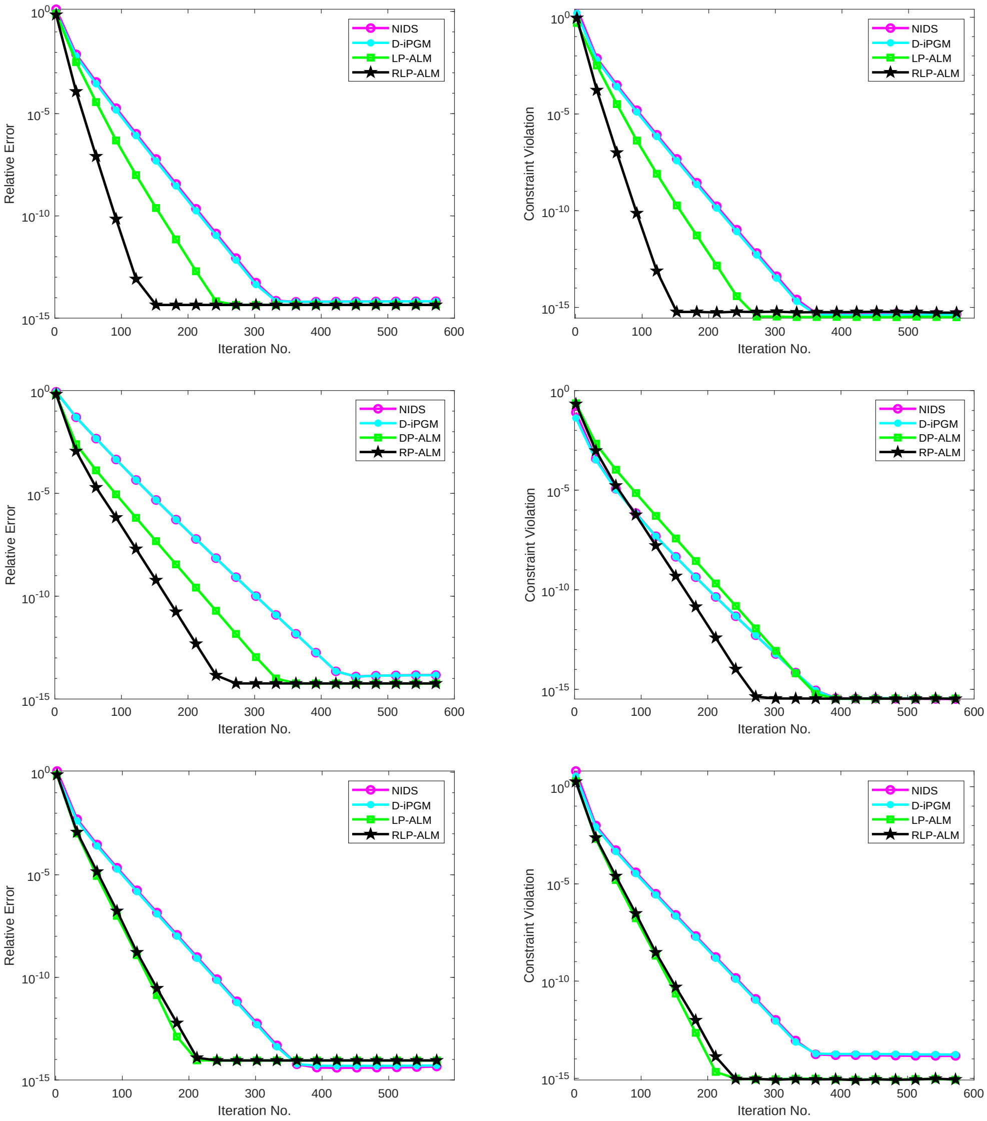

Figure 4: Convergence curves of the relative error and constraint violation when $f_i(x_i) = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{2}\|\mathbf{a}_{ij}^{\mathsf{T}}x_i - \mathbf{b}_{ij}\|^2$ among datasets ijcnn1, a7a and covtype (from top to bottom).

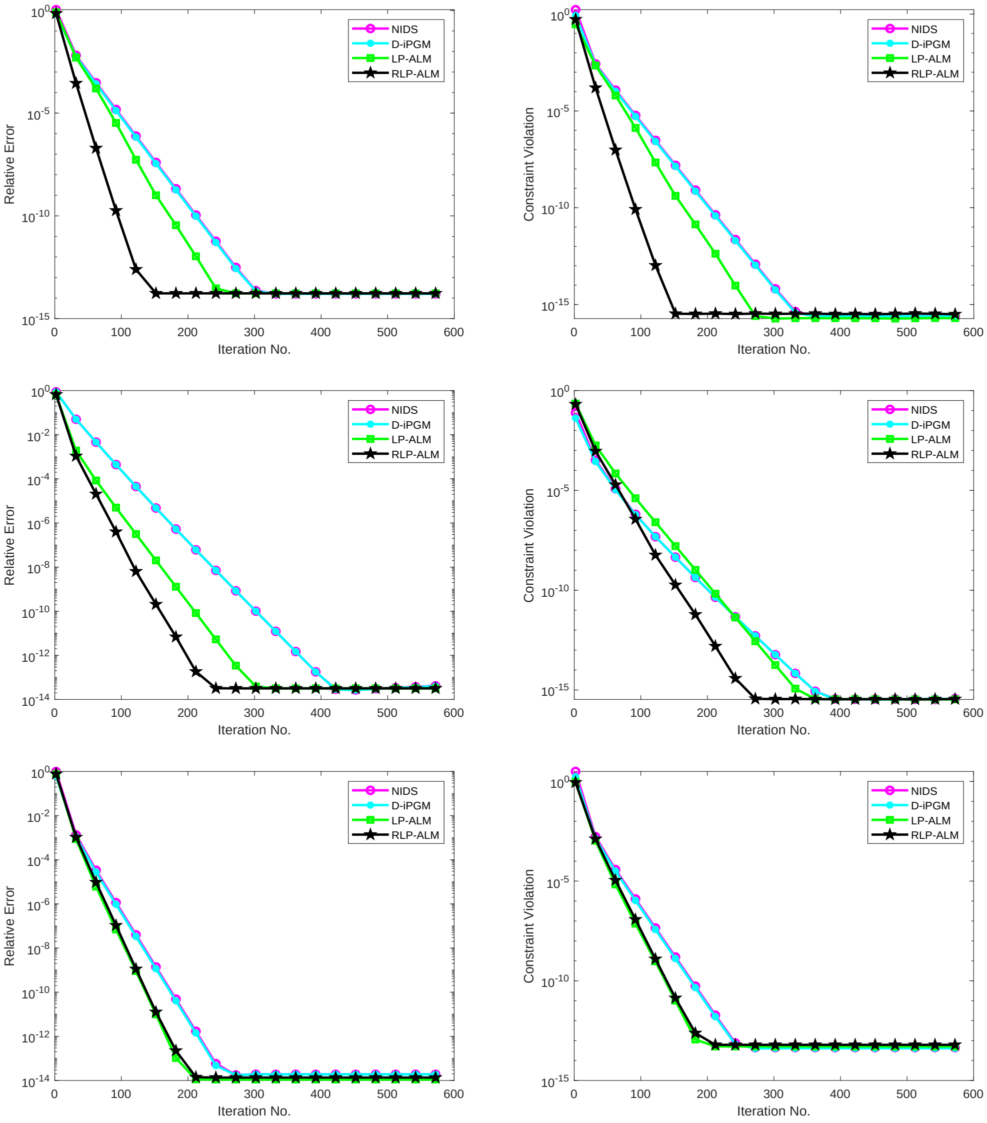

Figure 5: Convergence curves of the relative error and constraint violation when $f_i(x_i) = \frac{1}{m_i}\sum_{j=1}^{m_i} \ln\left(1+e^{-(\mathbf{a}_{ij}^{\mathsf{T}} x_i)\mathbf{b}_{ij}}\right)$ among datasets ijcnn1, a7a and covtype (from top to bottom).

Our multi-block extension of DP-ALM (denoted by DP-mALM) for solving (52) reads

$$\left\{\begin{array}{l} x_1^{k+1} = \arg\min\limits_{x_1\in\mathbb{X}_1} \left\{\theta_1(x_1) - \langle\lambda^k, A_1x_1\rangle + \frac{\tau\beta}{2}\left\|A_1(x_1 - x_1^k)\right\|^2 + \frac{\tau}{2}\left\|x_1 - x_1^k\right\|_{D_1}^2\right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2\in\mathbb{X}_2} \left\{\theta_2(x_2) - \langle\lambda^k, A_2x_2\rangle + \frac{\tau\beta}{2}\left\|A_2(x_2 - x_2^k)\right\|^2 + \frac{\tau}{2}\left\|x_2 - x_2^k\right\|_{D_2}^2\right\}, \\ \vdots \\ x_p^{k+1} = \arg\min\limits_{x_p\in\mathbb{X}_p} \left\{\theta_p(x_p) - \langle\lambda^k, A_px_p\rangle + \frac{\tau\beta}{2}\left\|A_p(x_p - x_p^k)\right\|^2 + \frac{\tau}{2}\left\|x_p - x_p^k\right\|_{D_p}^2\right\}, \\ \lambda^{k+1} = \lambda^k - \beta\Big[\gamma\Big(\sum\limits_{i=1}^p A_ix_i^{k+1} - b\Big) + \sum\limits_{i=1}^p A_i(x_i^{k+1} - x_i^k)\Big], \end{array}\right. \tag{53}$$

where $D_i = r_i\mathbf{I} - \beta A_i^\top A_i$ with $r_i > \beta\rho(A_i^\top A_i)$. Note that the subproblems in (53) are updated in parallel and are similar to the updating way in [4], but the dual variable updates differently from the previous.

By denoting

$$\tilde{x}_i^k = x_i^{k+1}(i = 1, 2, \cdots, p), \quad \tilde{\lambda}^k = \lambda^k - \beta\Big[\sum_{i=1}^p A_i\tilde{x}_i^k - b + \alpha\sum_{i=1}^p A_i\Big(\tilde{x}_i^k - x_i^k\Big)\Big], \tag{54}$$

the previous inequality (22) still holds for any $w \in \Omega := \mathbb{X}_1 \times \cdots \mathbb{X}_p \times \mathbb{R}^m$, but with

$$w = \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \\ \lambda \end{pmatrix}, \mathcal{J}(w) = \begin{pmatrix} -A_1^\top\lambda \\ -A_2^\top\lambda \\ \vdots \\ -A_p^\top\lambda \\ \sum_{i=1}^p A_ix_i - b \end{pmatrix}, Q = \begin{bmatrix} \tau r_1\mathbf{I} & \cdots & \mathbf{0} & A_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \tau r_p\mathbf{I} & A_p^\top \\ \alpha A_1 & \cdots & \alpha A_p & \frac{1}{\beta}\mathbf{I} \end{bmatrix}.$$

By making use of the notations in (54) and the update of λ^{k+1} in (53), we can obtain the previous relationship (9) with

$$M = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ -(1-\gamma\alpha)\beta A_1 & \cdots & -(1-\gamma\alpha)\beta A_p & \gamma\mathbf{I} \end{bmatrix}. \tag{55}$$

Obviously, the notations in this section are multi-block extension of the previous, so the convergence theories in Section 3 can be similarly established if both

$$H = QM^{-1} \quad \text{and} \quad G = Q^\top + Q - M^\top HM$$

are positive definite. Analogous to the analysis in proving Proposition 1, we need to analyze the conditions to ensure the positive definiteness of $S = Q^\top M$, namely,

$$S = \begin{bmatrix} \tau r_1\mathbf{I} & & & \alpha\gamma A_1^\top \\ & \ddots & & \vdots \\ & & \tau r_p\mathbf{I} & \alpha\gamma A_p^\top \\ \alpha\gamma A_1 & \cdots & \alpha\gamma A_p & \frac{\gamma}{\beta}\mathbf{I} \end{bmatrix} - (1-\alpha\gamma)\alpha\beta\begin{bmatrix} \mathcal{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where

$$\mathcal{A} = (A_1\ A_2\ \cdots A_p)^\top(A_1\ A_2\ \cdots A_p).$$

For all $r_i > \beta\rho(A_i^\top A_i)$, it is not difficult to verify that S is positive definite if $\tau > p\alpha$. Besides, we can ensure the positive definiteness of the matrix

$$G = \begin{bmatrix} \tau r_1 \mathbf{I} & & & (1+\alpha-\alpha\gamma)A_1^\top \\ & \ddots & & \vdots \\ & & \tau r_p \mathbf{I} & (1+\alpha-\alpha\gamma)A_p^\top \\ (1+\alpha-\alpha\gamma)A_1 & \cdots & (1+\alpha-\alpha\gamma)A_p & \frac{2-\gamma}{\beta}\mathbf{I} \end{bmatrix} + (1-\alpha\gamma)\alpha\beta \begin{bmatrix} \mathcal{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

for any

$$\tau > p\Big[\frac{(1+\alpha-\alpha\gamma)^2}{2-\gamma} - (1-\alpha\gamma)\alpha\Big] = p\Big[\frac{(\alpha-\frac{\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4}\Big], \quad \forall \alpha \in \mathbb{R},$$

which also ensures $\tau > p\alpha$. Obviously, this new region reduces to the previous region in (6) when $p = 1$. Finally, the relaxed version of (53) with the relaxation step as in (42) is still convergent, and similar analysis can date back to Section 4.2.

References

[1] E. Birgin, J. Martínez, *Complexity and performance of an augmented Lagrangian algorithm*, Optim. Methods Soft. 35: 885-920, (2020)

[2] E. Birgin, G. Haeser, N. Maculan, L. Ramirez, *On the global convergence of a general class ofaugmented Lagrangian methods*, http-s://www.ime.usp.br/ egbirgin/publications/bhmama2024-alframework.pdf, (2024)

[3] J. Bai, D. Han, H. Sun, H. Zhang, *Convergence on a symmetric accelerated stochastic ADMM with larger stepsizes*, CSIAM Trans. Appl. Math., 3: 448-479, (2022)

[4] J. Bai, L. Jia, Z. Peng, *A new insight on augmented Lagrangian method with applications in machine learning*, J. Sci. Comput., 99: 53, (2024)

[5] J. Bai, K. Guo, J. Liang, Y. Jing, H. So, Accelerated symmetric ADMM and its applications in large-scale signal processing, J. Comput. Math., 42(6): 1605-1626, (2024)

[6] J. Bai, Y. Chen, X. Yu, H. Zhang, *Generalized asymmetric forward-backward-adjoint algorithms for convex-concave saddle-point problem*, Optimization Online, https://optimization-online.org/?p=24126, (2024)

[7] R. Bollapragada, C. Karamanli, et al., *An adaptive sampling augmented Lagrangian method for stochastic optimization with deterministic constraints*, Comput. Math. Appl., 149: 239-258, (2023)

[8] A. Chambolle, T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40: 120-145, (2011)

[9] A. Chambolle, T. Pock, *On the ergodic convergence rates of a first-order primaldual algorithm*, Math. Program., 159: 253-287, (2016)

[10] Y. Cui, C. Ding, X. Li, X. Zhao, *Augmented Lagrangian methods for convex matrix optimization problems*, J. Oper. Res. Soc. China, 10: 305-342, (2022)

[11] L. Condat, *A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158: 460-479, (2013)

[12] W. Deng, M. Lai, Z. Peng, W. Yin, *Parallel multi-block ADMM with $o(1/k)$ convergence*, J. Sci. Comput., 71: 712-736, (2017)

[13] D, Donoho, Y. Tsaig, *Fast solution of l_1-norm minimization problems when the solution may be sparse*, IEEE Trans. Inform., 54: 4789-4812, (2008)

[14] K. Guo, D. Han, T. Wu, *Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints*, INT J. Comput. Math., 94: 1653-1669, (2017)

[15] L. Guo, X. Shi, J. Cao, Z. Wang, *Decentralized inexact proximal gradient method with network-independent stepsizes for convex composite optimization*, IEEE Trans. Signal Process., 71: 786-801, (2023)

[16] D. Han, *A survey on some recent developments of alternating direction method of multipliers*, J. Oper. Res. Soc. China, 10: 1-52, (2022)

[17] B. He, F. Ma, X. Yuan, *Convergence study on the symmetric version of ADMM with larger step sizes*, SIAM J. Imaging Sci., 9, 1467-1501, (2016)

[18] B. He, F. Ma, X. Yuan, *Optimal proximal augmented Lagrangian method and its application to fullJacobian splitting for multi-block separable convex minimization problems*, IMA J. Numer. Anal., 40: 1188-1216, (2020)

[19] B. He, X. Yuan, *Balanced augmented Lagrangian method for convex programming*, arXiv:2108.08554, (2021)

[20] B, He. Yuan. X, Zhang. W, *A customized proximal point algorithm for convex minimization with linear constraints*, Compout. Optim. Appl., 56: 559-572, (2013)

[21] He. B, Ma. F, Xu. S, Yuan. X, *A generalized primal-dual algorithm with improved convergence condition for saddle point problems*, SIAM J. Imaging Sci., 15: 1157-1183, (2022)

[22] M. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4: 303-320, (1969)

[23] Y. Hu, C. Li, K. Meng, J. Qin, X. Yang, *Group sparse optimization via $\ell_{p,q}$ regularization*, J. Mach. Learn. Res., 18: 1-52, (2017)

[24] F. Jiang, X. Cai, D. Han, *The indefinite proximal point algorithms for maximalmonotone operators*, Optimization, 70: 1759-1790, (2021)

[25] W. Kong, R. Monteiro, *An accelerated inexact dampened augmented Lagrangian method for linearly-constrained nonconvex composite optimization problems*, Comput. Optim. Appl., 85: 509-545, (2023)

[26] Y. Li, M. Zhao, W. Chen, Z. Wen, *A stochastic composite augmented Lagrangian method for reinforcement learning*, SIAM J. Optim., 33: 921-949, (2023)

[27] Z. Li, W. Shi, M. Yan, *A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates*, IEEE Trans. Signal Process., 67: 4494-4506, (2019)

[28] C. Loan, *On the method of weighting for equality-constrained least-squares problems*, SIAM J. Numer. Anal., 22: 851-864, (1985)

[29] F. Ma, M. Ni, *A class of customized proximal point algorithms for linearly constrained convex optimization*, Comput. Appl. Math., 37: 896-911, (2018)

[30] Y. Ma, J. Bai, H. Sun, *An inexact ADMM with proximal-indefinite term and larger stepsize*, Appl. Numer. Math., 184: 542-566, (2023)

[31] S. Osher, H. Heaton, S. Fung, *A Hamilton-Jacobi-based proximal operator*, PANS, 120, e2220469120, (2023)

[32] X. Ou, G. Yu, J. Liu, J. Chen, Z. Liu, *A new penalty dual-primal augmented Lagrangian method and its extensions*, Taiwanese J. Math., 28: 1223-1244, (2024)

[33] J. Peng, H. Wang, et al, *Stable local-smooth principal component pursuit*, SIAM J. Imaging Sci., 17: 1182-1205, (2024)

[34] M. Powell, *A method for nonlinear constraints in minimization problems*, Optimization (R. Fletcher ed.). New York: Academic Press, pp. 283-298, (1969)

[35] J. Scott, M. Tuma, *Solving large linear least squares problems with linear equality constraints*, BIT Numer. Math., 62: 1765-1787, (2022)

[36] W. Shi, Q. Ling, G. Wu, W. Yin, *EXTRA: An exact first-order algorithm for decentralized consensus optimization*, SIAM J. Optim., 25: 944-966, (2015)

[37] B. Vu, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Adv. Comput. Math. 38: 667-681, (2013)

[38] N. Wang, J. Li, *A class of preconditioners based on symmetric-triangular decomposition and matrix splitting for generalized saddle point problems*, IMA J. Numer. Anal., 43: 2998-3025, (2023)

[39] S. Xu, X. Cao, et al, *Hyperspectral image denoising by asymmetric noise modeling*, IEEE Trans. Geosci. Remote Sens., 60: 5545214, (2022)

[40] Y. Xu, *Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming*, Math. program., 185: 199-244, (2021)

[41] J. Yang, Y. Zhang, *Alternating direction algorithms for ℓ_1-problems in compressive sensing*, SIAM J. Sci. Comput., 33: 250-278, (2011)

[42] T. Zhang, Y. Xia, S. Li, $\mathcal{O}(1/k^2)$ *convergence rates of (dual-primal) balanced augmented Lagrangian methods for linearly constrained convex programming*, Numer. Algor., 98: 325-345, (2025)

[43] H. Zhang, Z. Xu, *An alternating proximal gradient algorithm for nonsmooth nonconvex-linear minimax problems with coupled linear constraints*, J. Oper. Res. Soc. China, doi: 10.1007/s40305-024-00550-3, (2024)

[44] J. Zhang, Z. Luo, *A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization*, SIAM J. Optim., 32: 2319-2346, (2022)

[45] D. Zhu, L. Zhao, S. Zhang, *A first-order primal-dual method for nonconvex constrained optimization based on the augmented Lagrangian*, Math. Oper. Res., 49: 125-150, (2024)