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

Jianchao Bai[†] Shuang Rao[‡]

Abstract

In this paper, we consider a family of linearly constrained convex minimization problems whose objective function is not necessarily smooth. A preliminary double-proximal augmented Lagrangian method (DP-ALM) is developed, which enjoys a flexible dual stepsize and a proximal subproblem with relatively smaller proximal parameter. By a novel prediction-correction reformulation for the proposed DP-ALM and by similar variational characterizations for both the saddle-point of the problem and the generated sequences, we establish the global convergence of DP-ALM and its sublinear convergence rate in both ergodic and nonergodic senses. An toy example is taken to illustrate that the presented lower bound of proximal parameter is optimal. Besides, we briefly discuss a relaxed accelerated DP-ALM and the multi-block extension of DP-ALM as well as their convergence conditions.

Keywords: convex optimization, augmented Lagrangian method, proximal term, improved convergence condition, convergence and complexity

Mathematics Subject Classification(2010): 65K10, 65Y20, 90C25

1 Introduction

Many application problems can be formulated as the following optimization model

$$\min\{\theta(x) \mid Ax = b,\ x \in \mathbb{X}\}, \tag{1}$$

where $\theta : \mathbb{R}^n \to \mathbb{R}$ is a proper lower semicontinuous convex function (possibly nonsmooth, non-Lipschitz continuous, and non-strongly convex), $\mathbb{X} \subseteq \mathbb{R}^n$ is a closed convex set, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given. Here, we take several typical examples as follows:

- **Linear constrained least-squares problem** [18, 24] arising 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 large, sparse and its dimension satisfies $l > n$; the dimension of A satisfies $m \ll n$. A particular case of this example is $\theta(x) = 0$, that is, the problem reduces to an ill-conditioned linear equation.

[*] This work was supported by the Shaanxi Fundamental Science Research Project for Mathematics and Physics (23JSQ031).

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

[‡] Corresponding author. 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`).

- **Sparse optimization problem** [13, 27] where $\theta(x) = \|x\|_1$, which 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 $m \ll n$.

- **Decentralized composite optimization over networks** [7] arising 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 $\mathbf{x} = [x_1^\top, x_2^\top, \cdots, x_N^\top]^\top$ with N denoting the number of computational agents, and $\mathbf{W} = W \otimes \mathbf{I}$ with W being a symmetric and doubly stochastic matrix.

Except the above examples, the model (1) also arises in image processing, data processing and so forth. Throughout this paper, we assume the solution set of (1) is nonempty.

1.1 Notations and definitions

Denoted by $\mathbb{R}^n$ 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 $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is defined as $\|x\|_1 = \sum_{i=1}^n |x_i|$. 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 $\|x\|_G = \sqrt{x^\top G x}$ as a weighted G-norm, where $(\cdot)^\top$ denotes the transpose operator and G is a symmetric positive definite matrix. The distance from any point x to a closed convex set C in the sense of G-weighted norm is defined as $\mathrm{dist}_G(x, \mathcal{C}) := \min_{y \in \mathcal{C}} \|x - y\|_G$. When $G = \mathbf{I}$, $\|x\|_G$ reduces to the Euclidean norm and $\mathrm{dist}_G(x, \mathcal{C})$, simply denoted by $\mathrm{dist}(x, \mathcal{C})$, reduces to the Euclidean distance. 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

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

1.2 Related work

The Augmented Lagrangian Method (ALM, [12, 22]) is a benchmark method for solving equality constrained minimization problem (1), and it has remained relatively popular yet vivid in recent years, cf. [2, 4, 11, 16, 17, 23, 28] to list a few. The framework of classic ALM for solving (1) reads

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

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. Ignoring some constants, the x-subproblem of (2) amounts to $x^{k+1} = \arg\min_{x \in \mathbb{X}} \big\{ \theta(x) + \frac{\beta}{2}\big\|Ax - b - \lambda^k/\beta\big\|^2 \big\}$. When coefficient matrix $A \neq \mathbf{I}$ and the set $\mathbb{X} \neq \mathbb{R}^n$, solving this subproblem is still challenging even as difficult as solving (1) if without utilizing some linearization techniques or inner solvers.

To overcome the above obstacle, a meaningful and powerful technique is to add a quadratic proximal term in the form of $\frac{1}{2}\left\|x-x^k\right\|_D^2$, where $D\in\mathbb{R}^{n\times n}$ is called the proximal matrix. Consequently, the following proximal ALM has attracted much attention:

$$\left\{\begin{array}{l} x^{k+1}=\arg\min\limits_{x\in\mathbb{X}}\left\{L(x,\lambda^k)+\frac{\beta}{2}\|Ax-b\|^2+\frac{1}{2}\left\|x-x^k\right\|_D^2\right\},\\ \lambda^{k+1}=\lambda^k-\gamma\beta\bigl(Ax^{k+1}-b),\end{array}\right. \tag{3}$$

The proximal matrix D is usually required to be positive definite for the sake of convergence. If we choose $D=r\mathbf{I}-\beta A^\top A$ with $r>\beta\rho(A^\top A)$, then the subproblem of (3) can be simplified as the following proximity operator

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

The objective 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 $\theta=\|x\|_1$ as mentioned in the sparse signal recovery problem. Otherwise, one may exploit inner solvers or to solve it inexactly, or use the formula provided by [21] for accurately approximating the proximity operator. Recently, to investigate an indefinite proximal term, He, et al. [11] proposed the following optimal proximal ALM based on the scheme (3):

$$\left\{\begin{array}{l} x^{k+1}=\arg\min\limits_{x\in\mathbb{X}}\left\{L(x,\lambda^k)+\frac{\beta}{2}\left\|Ax-b\right\|^2+\frac{1}{2}\left\|x-x^k\right\|_{D_0}^2\right\},\\ \lambda^{k+1}=\lambda^k-\gamma\beta(Ax^{k+1}-b),\end{array}\right. \tag{4}$$

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

More recently, Bai, et al. [2] considered a new double-penalty ALM with a relaxation step for convex optimization problems, where the key subproblem and the dual variable obey the following iterations:

$$\left\{\begin{array}{l} x^{k+1}=\arg\min\limits_{x\in\mathbb{X}}\left\{L(x,\lambda^k)+\frac{\beta}{2}\left\|A(x-x^k)\right\|^2+\frac{1}{2}\left\|x-x^k\right\|_D^2\right\},\\ \lambda^{k+1}=\lambda^k-\beta\bigl[A(2x^{k+1}-x^k)-b\bigr].\end{array}\right. \tag{5}$$

The authors showed that the global convergence and sublinear convergence rate of (5) can be established for any arbitrarily positive-definite matrix D. Notice that the scheme (5) is different from most of ALM-type methods since the subproblem does not depend on the data b either. The involved two quadratic terms can be treated as different proximal terms: one involves the matrix A, and the other does not involve it. Based on [2], a different penalty dual-primal ALM was developed in [23] and was demonstrated to be efficiency on solving large-scale basic pursuit problem and Lasso problem. Cui, et al. [5] also provided some gentle introductions to the recent advance in augmented Lagrangian methods for solving large-scale convex matrix optimization problems. Han [8] systematically reviewed the developments of the problem (1) and its multi-block extensions from ALM to its related splitting methods. For more details on accelerated versions of ALM, we refer to [15, 16, 28].

Except the above deterministic ALM-type work on convex programming problems, some researchers also focused on studying stochastic ALM and nonconvex ALM. For example, a stochastic ALM based on a variant of stochastic accelerated gradient method was presented in [1, 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. [17] 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. [4] also constructed an adaptive sampling ALM for linearly equality constrained optimization problems with a stochastic objective function, and they further established its sublinear convergence for convex objectives and linear convergence for strongly convex objectives. Recent progresses on nonconvex ALM can be found in e.g. [16, 20].

1.3 The proposed algorithm and contributions

In this paper, mainly motivated by the interesting work [2, 11] we propose the following novel double-proximal ALM-type method for solving the problem (1):

$$\text{(DP-ALM)} \quad \begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{X}} \left\{ \theta(x) - \langle \lambda^k, Ax\rangle + \frac{\tau\beta}{2}\left\|A(x-x^k)\right\|^2 + \frac{\tau}{2}\left\|x-x^k\right\|_D^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta\big[\gamma(Ax^{k+1}-b) + A(x^{k+1}-x^k)\big], \end{cases} \tag{6}$$

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

$$\tau > \frac{(\alpha-\frac{\gamma}{2})^2}{2-\gamma} + \frac{2+\gamma}{4}, \quad \forall \gamma \in (0,2), \alpha \in \mathbb{R}. \tag{7}$$

With the structure of D, the above x-subproblem can be simplified as that in (9a). Here, α can be treated as an auxiliary parameter. For using a smaller τ approximating to the so-called optimal parameter value 0.75 as in [11], one can select $\alpha = \gamma/2$ and $\gamma = 1$. We further describe DP-ALM (6) as the following prediction-correction framework, where

$$w^k = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix}, \quad \tilde{w}^k = \begin{pmatrix} \tilde{x}^k \\ \tilde{\lambda}^k \end{pmatrix} \quad \text{and} \quad M = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -(1-\alpha\gamma)\beta A & \gamma\mathbf{I} \end{bmatrix}. \tag{8}$$

A prediction-correction reformulation of DP-ALM (6).

Prediction Step:

$$\tilde{x}^k = \mathbf{prox}_{\tau r\theta}\big[x^k + A^\top\lambda^k/(\tau r)\big]; \tag{9a}$$

$$\tilde{\lambda}^k = \lambda^k - \beta\big[A\tilde{x}^k - b + \alpha A(\tilde{x}^k - x^k)\big]; \tag{9b}$$

Correction Step:

$$w^{k+1} = w^k - M(w^k - \tilde{w}^k). \tag{10}$$

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 [2] without using relaxation step, and both of them enjoy the same proximal subproblem. However, when $\alpha < 1$ or $\alpha > \gamma - 1$, we will have

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

  which indicates that DP-ALM could enjoy a smaller proximal parameter τ than the method in [2]. As discussed in Section 4, $\frac{2+\gamma}{4}$ is the optimal lower bound of proximal parameter τ. Moreover, as γ approximates to zero, the lower bound of τ will approximate to 1/2. Compared to the subproblem in the traditional ALM (2), the subproblem of DP-ALM not only maintains the merits of that in [2], but also allows relatively smaller proximal parameter, while the dual update has an extra term $-\beta A(x^{k+1}-x^k)$ that is not involved in (2) and (4). The preliminary method DP-ALM is further extended to the version with a relaxation step and the multi-block version whose subproblems are updated in parallel.

- **Global convergence and various convergence rates.** By reformulating DP-ALM as a prediction-correction version and by using variational characterizations for both the saddle-point of (1) and the involved iterative sequences, we establish the global convergence of DP-ALM and various convergence rate in detail, 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. Compared to the analysis in [11], our analysis uses a distinctive update as in (9a) with an auxiliary parameter $\alpha \in \mathbb{R}$, but greatly simplify the whole convergence analysis. Besides, we further discuss the convergence of an extended DP-ALM with popular relaxation step as well as its multi-block splitting version in the appendix.

- **The applicability of new techniques.** Motivated by the novel reformulation in Section 1.3, that is using a distinctive update, we can apply this new technique to simply discuss the key convergence analysis for the existing scheme (4). Different to the original prediction $\tilde{\lambda} = \lambda^k - \beta(A\tilde{x}^k - b)$ in [11], we can introduce the same auxiliary prediction update as in (9a) to reformulate (4) as

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

Similar to the analysis of Lemma 2, we need to derive the condition of ensuring 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

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

Analogous to the analysis of Proposition 1, 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 (7) and τ can be chosen approximating to 0.75 when $\alpha = \frac{\gamma-2}{2}$. That is, the optimal proximal parameter in [11, Section 4] can be obtained too, but here the techniques of convergence analysis is different from before.

2 Technical preliminaries

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

2.1 Variational characterization

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

Lemma 1 *[10] 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* $\min\{f(x) + h(x) \mid x \in \Phi\}$ *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 we have

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

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 (see [9])

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

2.2 Basic matrices and properties

Since the matrix M defined in (8) 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 (7), the matrices H and G defined by (15) are symmetric positive definite.*

Proof. First of all, we have from (7) 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 the matrix Q. Together with this property and the nonsingularity of M, we have $S := Q^\top M$ is also nonsingular and symmetric. With the notation S, the matrices H and G given by (15) can be rewritten as

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

In practice, 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 $\Sigma_m = \mathrm{diag}(s_1, s_2, \ldots, s_m) \in \mathbb{R}^{m\times m}$ 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 rewritten 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 [25, Page 16]), it can be demonstrated that the matrix P is positive definite if and only if

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

which is equivalent to $\left(\frac{\tau r}{\beta} - \alpha s_i^2\right)\gamma > 0$. Since $\gamma \in (0,2)$ and $\rho(A^\top A) = \max_{i\in\{1,\ldots,m\}} s_i^2 > 0$, then, the matrix P is positive definite if $\frac{\tau r}{\beta} > \alpha\rho(A^\top A)$, which has been ensured by (7) or precisely (17). Consequently, H is positive definite if (7) holds.

By analogous analysis for the matrix G in (18), it can be shown that G is positive definite if (7) holds. 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 of DP-ALM and the sublinear convergence rate of the optimality gap.

3.1 Global convergence

Based on the prediction-correction reformulation as in (9a)-(9b) and (10), 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 (7) it holds $\tilde{w}^k \in \Omega$ and*

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

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

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

Proof. According to Lemma 1, the first-order optimality condition of (9a) is $\tilde{x}^k \in \mathbb{X}$ and

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

for any $x \in \mathbb{X}$. Besides, the equality (9b) can be rewritten as

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

for any $\lambda \in \mathbb{R}^m$. Combine the last two relationships with the notations in (8), (12) and the matrix Q in (16) to obtain

$$\begin{aligned} \theta(x) - \theta(\tilde{x}^k) + \langle w - \tilde{w}^k, \mathcal{J}(\tilde{w}^k)\rangle &\geq (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) \\ &= (w - \tilde{w}^k)^\top H(w^k - w^{k+1}), \end{aligned} \quad (21)$$

where the equality uses the update in (10) 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 (21) gives

$$\begin{aligned} &(w - \tilde{w}^k)^\top H(w^k - w^{k+1}) - \frac{1}{2}\Big(\big\|w - w^{k+1}\big\|_H^2 - \big\|w - w^k\big\|_H^2\Big) \\ &= \frac{1}{2}\Big(\big\|w^k - \tilde{w}^k\big\|_H^2 - \big\|w^{k+1} - \tilde{w}^k\big\|_H^2\Big) \\ &= \frac{1}{2}\Big(\big\|w^k - \tilde{w}^k\big\|_H^2 - \big\|w^{k+1} - w^k + w^k - \tilde{w}^k\big\|_H^2\Big) \\ &\overset{(10)}{=} \frac{1}{2}\Big(\big\|w^k - \tilde{w}^k\big\|_H^2 - \big\|(w^k - \tilde{w}^k) - M(w^k - \tilde{w}^k)\big\|_H^2\Big) \\ &= \frac{1}{2}\big(w^k - \tilde{w}^k\big)^\top\big(Q^\top + Q - M^\top HM\big)\big(w^k - \tilde{w}^k\big) \overset{(15)}{=} \frac{1}{2}\big\|w^k - \tilde{w}^k\big\|_G^2. \end{aligned}$$

Substituting the last relationship into (21) 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. ■

Now, we are ready to establish the global convergence of our DP-ALM based on the aforementioned Lemma 2.

Theorem 1 *Let $\{w^{k+1}\}$ be the sequence generated by DP-ALM. Then, under the condition (7) we have*

$$\lim_{k\to\infty} \left\|w^k - w^{k+1}\right\| = 0 \tag{22}$$

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

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

Combine (23), (10) and the nonsingularity of M to confirm the result in (22).

By the uniformly boundness of $\{w^k\}$ and (10), 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 (21) 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 (23), 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 $\text{VI}(\theta, \mathcal{J}, \Omega)$ (11) and hence is also a solution point of the convex optimization problem (1).

Now, by (23) and $\lim_{j\to\infty} w^{k_j} = w^\infty$, the sequence $\{w^{k_j}\}$ also converges to w^∞. By (20) again, we have

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

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

3.2 Ergodic convergence rate

Motivated by (14), for any $\epsilon > 0$, $\bar{w} \in \Omega^*$ is called an ϵ-approximate solution of $\text{VI}(\theta, \mathcal{J}, \Omega)$ with the accuracy 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 Lemma 2, we next establish the $\mathcal{O}(1/T)$ 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*

$$x_T := \frac{1}{T+1}\sum_{k=0}^{T}\tilde{x}^k \quad \text{and} \quad w_T := \frac{1}{T+1}\sum_{k=0}^{T}\tilde{w}^k. \tag{24}$$

Then, under the condition (7) we have

$$\theta(x_T) - \theta(x) + \big\langle w_T - w, \mathcal{J}(w)\big\rangle \le \frac{1}{2(1+T)}\big\|w - w^0\big\|_H^2, \quad \forall w \in \Omega. \tag{25}$$

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

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

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

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

Namely,

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

which, by the convexity of θ and the definition of x_T and w_T in (24), confirms the result in (25). ■

The above theorem shows that the average of first T iterates defined in (24) is an approximate solution of 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 [26]. To proceed, for any $\varsigma > 0$, let $\Gamma_\varsigma = \{\lambda \mid \|\lambda\| \le \varsigma\}$ and

$$\gamma_\varsigma = \inf_{x^* \in \mathbb{X}} \sup_{\lambda \in \Gamma_\varsigma} \Big\|\begin{pmatrix} x^* \\ \lambda \end{pmatrix} - \begin{pmatrix} x^0 \\ \lambda^0 \end{pmatrix}\Big\|_H^2. \tag{26}$$

Corollary 1 *Let γ_ς be defined in (26) and x_T be defined in (24). Then, for any $(x^*; \lambda^*) \in \Omega^*$ and $T > 0$, we have*

$$\big|\theta(x_T) - \theta(x^*)\big| \le \frac{\gamma_\varsigma}{2(1+T)} \quad \text{and} \quad \big\|Ax_T - b\big\| \le \frac{\gamma_\varsigma}{2(1+T)\big(1 + \|\lambda^*\|\big)}. \tag{27}$$

Proof. Set $w = (x^*; \lambda)$ into the inequality (25) to obtain

$$\begin{aligned} \theta(x_T) - \theta(x^*) + \big\langle w_T - w, \mathcal{J}(w)\big\rangle =& \theta(x_T) - \theta(x^*) - \lambda^\top(Ax_T - b) \\ \le& \frac{1}{2(1+T)}\Big\|\begin{pmatrix} x^* \\ \lambda \end{pmatrix} - \begin{pmatrix} x^0 \\ \lambda^0 \end{pmatrix}\Big\|_H^2, \end{aligned} \tag{28}$$

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

$$\theta(x_T) - \theta(x^*) + \varsigma\|Ax_T - b\| = \sup_{\lambda \in \Gamma_\varsigma}\big\{\theta(x_T) - \theta(x^*) - \lambda^\top(Ax_T - b)\big\} \le \frac{\gamma_\varsigma}{2(1+T)}. \tag{29}$$

By (28) again with (11), we have $\theta(x_T)-\theta(x^*)-\lambda^\top(Ax_T-b)\geq 0$, showing that

$$\theta(x_T)-\theta(x^*)\geq -\|\lambda^*\|\|Ax_T-b\|. \tag{30}$$

Then, take $\varsigma=2\|\lambda^*\|+1$ in (29) together with (30) to get

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

Rearrange the above inequality to confirm the second inequality in (27). Meanwhile, substitute the second inequality in (27) into (30) to obtain

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

which in turn confirms the first inequality in (27). ■

3.3 Nonergodic convergence rate

In this subsection, we will show the worst-case $\mathcal{O}(1/t)$ nonergodic convergence rate of DP-ALM in terms of pointwise iterative residual and optimality error based on the following preliminary lemma.

Lemma 3 *Let M and H be given by (8) and (15), respectively. Then, the iterates $\{w^k\}$ and $\{\tilde{w}^k\}$ generated by DP-ALM satisfy*

$$(w^k-\tilde{w}^k)^\top M^\top HM\{(w^k-\tilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\}\geq \frac{1}{2}\big\|(w^k-\tilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\big\|^2_{Q+Q^\top}. \tag{31}$$

Proof. Setting $w=\tilde{w}^{k+1}$ in (21) results in

$$\theta(\tilde{x}^{k+1})-\theta(\tilde{x}^k)+\big\langle \tilde{w}^{k+1}-\tilde{w}^k,\mathcal{J}(\tilde{w}^k)+Q(\tilde{w}^k-w^k)\big\rangle\geq 0. \tag{32}$$

Meanwhile, it follows from the inequality (21) with $k:=k+1$ that

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

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

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

Combine (32) and (33) together with the property in (13) to achieve

$$(\tilde{w}^k-\tilde{w}^{k+1})^\top Q\{(w^k-\tilde{w}^k)-(w^{k+1}-\tilde{w}^{k+1})\}\geq 0. \tag{34}$$

Now, by adding the identity

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

to both sides of (34), we can get

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

which immediately ends the proof of (31) based on (10) and the matrix H in (15). ■

Theorem 3 *Let M and H be given by (8) and (15), respectively. Then, for any integer $t > 0$ there exits a constant $c > 0$ such that the sequences $\{w^k\}$ and $\{\widetilde{w}^k\}$ generated by DP-ALM satisfy*

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

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

$$\left\|w^{k+1} - w^*\right\|_H^2 \leq \left\|w^k - w^*\right\|_H^2 - c\left\|M(w^k - \tilde{w}^k)\right\|_H^2, \ \ \forall w^* \in \Omega^*, \tag{35}$$

which shows

$$c\sum_{k=0}^{t}\left\|M(w^k - \tilde{w}^k)\right\|_H^2 \leq \left\|w^0 - w^*\right\|_H^2 \tag{36}$$

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

$$\|a\|_H^2 - \|b\|_H^2 = 2a^\top H(a-b) - \|a-b\|_H^2, \tag{37}$$

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

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

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

$$\left\|M(w^k - \tilde{w}^k)\right\|_H^2 \geq \left\|M(w^{k+1} - \tilde{w}^{k+1})\right\|_H^2,$$

showing that

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

Substituting it into (36) ends the proof. ■

Remark 1 *For any given $\epsilon > 0$, Theorem 3 indicates that our DP-ALM needs at most $\lceil c_1/\epsilon\rceil$ iterations to guarantee $\left\|M(w^k - \tilde{w}^k)\right\|_H^2 \leq \epsilon$, where $c_1 = \inf_{w^*\in\Omega^*}\left\|w^0 - w^*\right\|_H^2/c$. According to Theorem 1 and Theorem 3, $\left\|w^k - w^{k+1}\right\|$ can be used as a stopping condition of DP-ALM. In addition, by letting $d^k = (d_x^k; d_\lambda^k)$ where*

$$\left\{\begin{array}{l} d_x^k = \tau\beta A^\top A(\tilde{x}^k - x^k) + \tau D(\tilde{x}^k - x^k) + A^\top(\tilde{\lambda}^k - \lambda^k), \\ d_\lambda^k = \alpha A(\tilde{x}^k - x^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k), \end{array}\right. \tag{38}$$

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

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

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 (38) 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 (10), implies $\|d^k\|$ goes to zero in a sublinear rate.

4 Optimality of the formula (7)

In Section 1.3, we have provided a region of the proximal parameter τ to guarantee the convergence of DP-ALM. Moreover, it should be larger than $\frac{2+\gamma}{4}$ for any $\gamma \in (0, 2)$. Then, it is interesting to ask whether or not the lower bound $\frac{2+\gamma}{4}$ is optimal (smallest). In this section, we take an example to illustrate that this bound is optimal and it is impossible to find a lower bound smaller than $\frac{2+\gamma}{4}$.

Consider the example in [11, 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) as:

$$\min_{\mathbb{R}} \big\{0 \cdot x \mid x = 0\big\}. \tag{39}$$

Without loss of generality we take $\beta = 1$. Then, DP-ALM for solving (39) reads

$$\begin{cases} x^{k+1} = \arg\min\limits_{x \in \mathbb{R}} \big\{ -x\lambda^k + \frac{\tau}{2}\big(x - x^k\big)^2 + \frac{\tau(r-1)}{2}\big(x - x^k\big)^2\big\} = \frac{\lambda^k}{\tau r} + x^k, \\ \lambda^{k+1} = \lambda^k - \big(\gamma x^{k+1} + x^{k+1} - x^k\big) = \frac{\tau r - 1 - \gamma}{\tau r}\lambda^k - \gamma x^k. \end{cases} \tag{40}$$

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 $\varphi(\bar{\tau})$. Then, 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

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

Then, 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. Theretofore, the iterative scheme in (40), that is the application of DP-ALM to the problem (39), 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.

5 Relaxed version of DP-ALM

This section aims to extend the previous DP-ALM to the following relaxed accelerated version and provide a concise discussion on its convergence:

$$\text{(RP-ALM)} \quad \begin{cases} \hat{x}^k = \arg\min\limits_{x \in \mathbb{X}} \Big\{\theta(x) - \langle\lambda^k, Ax\rangle + \frac{\tau\beta}{2}\big\|A(x - x^k)\big\|^2 + \frac{\tau}{2}\big\|x - x^k\big\|_D^2\Big\}, \\ \hat{\lambda}^k = \lambda^k - \beta\gamma(A\hat{x}^k - b) - \beta A(\hat{x}^k - x^k), \\ \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \eta \begin{pmatrix} \hat{x}^k - x^k \\ \hat{\lambda}^k - \lambda^k \end{pmatrix}, \end{cases} \tag{41}$$

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

Analogous to the aforementioned analysis in Lemma 2, it is not difficult (for details, see e.g. [2, 6]) to show

$$\theta(x)-\theta(\tilde{x}^k)+\langle w-\tilde{w}^k, \mathcal{J}(w)\rangle \geq \frac{1}{2\eta}\Big(\big\|w-w^{k+1}\big\|_H^2-\big\|w-w^k\big\|_H^2\Big)+\frac{1}{2}\big\|w^k-\tilde{w}^k\big\|_G^2 \quad (42)$$

and

$$\big\|w^*-w^k\big\|_H^2 \geq \big\|w^*-w^{k+1}\big\|_H^2+\eta\big\|w^k-\tilde{w}^k\big\|_G^2. \quad (43)$$

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 $\{\big\|w^*-w^k\big\|_H^2\}$, we just need to investigate 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

$$\tau > \frac{[\alpha+\frac{2-2\eta-2\gamma\eta+\gamma\eta^2}{2}]^2}{(2-\eta)(2-\gamma\eta)}+\frac{\eta(\eta^2\gamma-4\eta\gamma-2\eta+4\gamma+4)}{4(2-\eta)} \quad (44)$$

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 (44) reduces to the previous in (7); when $\gamma = 1$ and $\tau = 1$, the proposed RP-ALM reduces to our proposed method in [2]. 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 (44) amounts to

$$\tau > \frac{\alpha^2\gamma\eta-\alpha\eta}{2-\eta}+\frac{[(1-\gamma\eta)\alpha+1)]^2}{(2-\eta)(2-\gamma\eta)}. \quad (45)$$

We can conclude that:

- If $\alpha = 0$, this region reduces to $\tau > \frac{1}{(2-\eta)(2-\gamma\eta)}$. By selecting $\eta \to 0$, we have that the lower bound of τ approximates to $1/4$ which is half of bound $1/2$ as shown by (7). 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 and [3, 11, 14] as well as the region $\frac{13-2\sqrt{13}}{9}$ in [19].

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 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.

Finally, we have from (42) and (43) 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.

Appendix: a multi-block extension

In this appendix, we will extend DP-ALM to solve the following multiple block separable convex optimization problem

$$\min\left\{\theta(x) := \sum_{i=1}^{p}\theta_i(x_i) \;\Big|\; \sum_{i=1}^{p} A_i x_i = b,\ x_i \in \mathbb{X}_i\right\}, \tag{46}$$

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. Our multi-block extension of DP-ALM (denoted by DP-mALM) for solving (46) reads

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1\in\mathbb{X}_1}\left\{\theta_1(x_1) - \langle\lambda^k, A_1x_1\rangle + \frac{\tau\beta}{2}\left\|A_1(x_1 - x_1^k)\right\|^2 + \frac{\tau}{2}\left\|x_1 - x_1^k\right\|_{D_1}^2\right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2\in\mathbb{X}_2}\left\{\theta_2(x_2) - \langle\lambda^k, A_2x_2\rangle + \frac{\tau\beta}{2}\left\|A_2(x_2 - x_2^k)\right\|^2 + \frac{\tau}{2}\left\|x_2 - x_2^k\right\|_{D_2}^2\right\}, \\ \quad\vdots \\ x_p^{k+1} = \arg\min\limits_{x_p\in\mathbb{X}_p}\left\{\theta_p(x_p) - \langle\lambda^k, A_px_p\rangle + \frac{\tau\beta}{2}\left\|A_p(x_p - x_p^k)\right\|^2 + \frac{\tau}{2}\left\|x_p - x_p^k\right\|_{D_p}^2\right\}, \\ \lambda^{k+1} = \lambda^k - \beta\Big[\gamma\Big(\sum\limits_{i=1}^{p} A_ix_i^{k+1} - b\Big) + \sum\limits_{i=1}^{p} A_i(x_i^{k+1} - x_i^k)\Big], \end{cases} \tag{47}$$

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 (47) are updated in parallel and are similar to the updating way in [2], but the dual variable updates differently from the previous.

By denoting

$$\tilde{x}_i^k = x_i^{k+1}(i = 1, 2, \cdots, p),\quad \tilde{\lambda}^k = \lambda^k - \beta\Big[\sum_{i=1}^{p} A_i\tilde{x}_i^k - b + \alpha\sum_{i=1}^{p} A_i\Big(\tilde{x}_i^k - x_i^k\Big)\Big], \tag{48}$$

the previous inequality (21) 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 (48) and the update of λ^{k+1} in (47), we can obtain the previous relationship (10) with

$$M = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ -(1-\gamma\alpha)\beta A_1 & \cdots & -(1-\gamma\alpha)\beta A_p & \gamma\mathbf{I} \end{bmatrix}. \tag{49}$$

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 difficulty 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 (7) when $p = 1$. Finally, the relaxed version of (47) with the relaxation step as in (41) is still convergent, and similar analysis can date back to Section 5.

References

[1] 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)

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

[3] J. Bai, Y. Chen, X. Yu, H. Zhang, *A generalized asymmetric forward-backward-adjoint algorithm for convex-concave saddle-point problem*, (2024) https://optimization-online.org/wp-content/uploads/2023/09/PDA_BCm.pdf

[4] 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)

[5] 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)

[6] G. Gu, B. He, X. Yuan, *Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach*, Comput. Optim. Appl., 59: 135–161, (2014)

[7] 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)

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

[9] B. He, X. Yuan, *On the $\mathcal{O}(1/n)$ convergence rate of the Douglas-Rachford alternating direction method*, SIAM J. Numer. Anal., 50: 700–709, (2012)

[10] 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)

[11] 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)

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

[13] 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)

[14] F. Jiang, Z. Zhang, H. He, *Solving saddle point problems: a landscape of primal-dual algorithm with larger stepsizes*, J. Global Optim., 85: 821–846, (2023)

[15] Y. Ke, C. Ma, *An accelerated augmented Lagrangian method for linearly constrained convex programming with the rate of convergence $\mathcal{O}(1/k^2)$*, Appl. Math. J. Chinese Univ., 32: 117–126, (2017)

[16] 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)

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

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

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

[20] J. Melo, R. Monteiro, H. Wang, *A proximal augmented Lagrangian method for linearly constrained nonconvex composite optimization problems*, J. Optim. Theory Appl., 202: 388–420, (2024)

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

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

[23] X. Qu, G. Yu, J.Liu, J. Chen, Z. Liu, *A new penalty dual-primal augmented Lagrangian method and its extensions*, Taiwanese J. Math., (2024) https://doi.org/10.11650/tjm/240603

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

[25] 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)

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

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

[28] 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., (2024) https://doi.org/10.1007/s11075-024-01796-x