Convergence Study on the Proximal Alternating Direction Method with Larger Step Size

Bingsheng He[1] **Feng Ma**[2]

February 12, 2017

Abstract. The alternating direction method of multipliers (ADMM) is a popular method for the separable convex programming with linear constraints, and the proximal ADMM is its important variant. Previous studies show that the relaxation factor $\gamma \in (0, \frac{1+\sqrt{5}}{2})$ by Fortin and Glowinski for the ADMM is also valid for the proximal ADMM. In this paper, we further demonstrate that the feasible region of γ depends on the proximal term added on the second subproblem, and can be enlarged when the proximal factor is positive. We derive the exact relationship between the relaxation factor γ and the proximal factor. Finally, we prove the global convergence and derive a worst-case $O(1/t)$ convergence rate in the ergodic sense for this generalized scheme.

Keywords. Alternating Direction Method of Multipliers, Convex Programming, Proximal Regularization, Contraction, Convergence Analysis.

Mathematics Subject Classification: 65K10, 90C25, 90C30.

1 Introduction

In this paper, we restrict our discussion to the canonical convex minimization model with separable structure in the generic setting:

$$\min\{\theta_1(x) + \theta_2(y) \mid Ax + By = b, \, x \in \mathcal{X}, y \in \mathcal{Y}\}, \tag{1.1}$$

where $A \in \Re^{m\times n_1}$, $B \in \Re^{m\times n_2}$, $b \in \Re^m$, $\mathcal{X} \subset \Re^{n_1}$ and $\mathcal{Y} \subset \Re^{n_2}$ are closed convex sets, $\theta_1 : \Re^{n_1} \to \Re$ and $\theta_2 : \Re^{n_2} \to \Re$ are convex (not necessarily smooth) functions. Throughout, the solution set of (1.1) is assumed to be nonempty; and the sets $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be simple.

The augmented Lagrangian function of (1.1) is defined as

$$\mathcal{L}_\beta(x, y, \lambda) = \theta_1(x) + \theta_2(y) - \lambda^T(Ax + By - b) + \frac{\beta}{2}\|Ax + By - b\|^2, \tag{1.2}$$

where λ is the Lagrange multiplier and $\beta > 0$ is a penalty parameter for the linear constraints. The augmented Lagrangian method (ALM) initially introduced in [25, 28] serves as a fundamental algorithm for solving (1.1). It reads as

$$\begin{cases} (x^{k+1}, y^{k+1}) = \arg\min\{\mathcal{L}_\beta(x, y, \lambda^k) \mid x \in \mathcal{X}, \; y \in \mathcal{Y}\}, & \text{(1.3a)} \\ \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). & \text{(1.3b)} \end{cases}$$

[1]Department of Mathematics, South University of Science and Technology of China, and Department of Mathematics, Nanjing University, China. This author was supported by the NSFC Grant 11471156. Email: hebma@nju.edu.cn

[2]Corresponding author. High-Tech Institute of Xi'an, Xi'an, 710025, Shaanxi, China. Email: mafengnju@gmail.com

In [29], it was shown that the ALM (1.3) is equivalent to the proximal point algorithm (PPA) to the dual problem of (1.1); and thus the relaxation technique developed for PPA in [15] can also be adapted to accelerate the original ALM. This results in the scheme:

$$\begin{cases} (x^{k+1}, y^{k+1}) = \arg\min\{\mathcal{L}_\beta(x, y, \lambda^k) \,|\, x \in \mathcal{X},\ y \in \mathcal{Y}\}, & \text{(1.4a)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b), & \text{(1.4b)} \end{cases}$$

where the parameter γ is chosen in the interval $(0, 2)$. Empirical studies show that the ALM (1.4) with over relaxation, i.e., $\gamma \in (1, 2)$, is often advantageous to speed up the performance.

From the computational perspective, ALM has a major drawback: it requires to minimize the augmented Lagrangian (1.2) with respect to (x, y) jointly, which can be prohibitively expensive when the problem data is large and dense. On the other hand, we have encountered many problems arising in sparse and low–rank optimization, the functions θ_1 and θ_2 usually have some specific properties, and a natural ideal is to exploit the underling structure of the problem (1.1) in the algorithmic design. The alternating directions method of multipliers (ADMM) proposed in [3, 14] is particularly suitable for this purpose. Given an initial iterate $(y^0, \lambda^0) \in \mathcal{Y} \times \Re^m$, ADMM generates the new iterate via the scheme:

$$\begin{cases} x^{k+1} = \arg\min\{\mathcal{L}_\beta(x, y^k, \lambda^k) \,|\, x \in \mathcal{X}\}, & \text{(1.5a)} \\ y^{k+1} = \arg\min\{\mathcal{L}_\beta(x^{k+1}, y, \lambda^k) \,|\, y \in \mathcal{Y}\}, & \text{(1.5b)} \\ \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). & \text{(1.5c)} \end{cases}$$

From the update procedure, we can see that ADMM updates the variables in a Gauss-Seidel type manner that each subproblem only involves one function in the original objective. Hence, the variables x and y are treated individually rather than together and the resulting subproblems can be carried out distributively. This feature makes the ADMM algorithm quite attractive for solving a wide range of large-scale problems, such as statistical learning, compressed sensing and semidefinite programming. We refer the readers to, e.g., [1, 7, 13], for recent surveys on ADMM.

There exists a variety of modifications to accelerate the convergence of ADMM, see e.g., [2, 6, 19, 20]. One popular and simple modification proposed by Fortin and Glowinski in [9, 10], and subsequently reported in [12], is to attach a relaxation factor to the Lagrange-multiplier-updating step in (1.5). This gives rise to the following scheme:

$$\begin{cases} x^{k+1} = \arg\min\{\mathcal{L}_\beta(x, y^k, \lambda^k) \,|\, x \in \mathcal{X}\}, & \text{(1.6a)} \\ y^{k+1} = \arg\min\{\mathcal{L}_\beta(x^{k+1}, y, \lambda^k) \,|\, y \in \mathcal{Y}\}, & \text{(1.6b)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b), & \text{(1.6c)} \end{cases}$$

where γ is a stepsize parameter that can be chosen in the interval $\left(0, \frac{1+\sqrt{5}}{2}\right)$. If choosing $\gamma > 1$ in (1.6c), the step size of updating the Lagrange multiplier is enlarged; and it is usually beneficial to induce faster convergence. We refer the reader to [12] and more in [5, 23] to the numerical verification. In addition, observe that the modification (1.6) differs from the original ADMM (1.5) only in the fact that the step size for updating the Lagrange multiplier can be larger than 1, and does not require additional computation. Thus it enjoys the same advantages as ADMM we have illustrated before.

The efficiency of ADMM relies on the fact that its subproblems have exact solutions or can be solved effectively to high precision. In many applications, the subproblems of ADMM may be not easy to evaluate, see e.g., [31, 32]. To treat the ADMM more sophisticatedly, it was suggested in [18] to regularize the subproblems with quadratic proximal terms. This leads to the proximal ADMM that reads as

$$\begin{cases} x^{k+1} = \arg\min\{\mathcal{L}_\beta(x, y^k, \lambda^k) + \|x - x^k\|_{D_1}^2 \mid x \in \mathcal{X}\}, & (1.7a) \\ y^{k+1} = \arg\min\{\mathcal{L}_\beta(x^{k+1}, y, \lambda^k) + \|y - y^k\|_{D_2}^2 \mid y \in \mathcal{Y}\}, & (1.7b) \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b), & (1.7c) \end{cases}$$

where D_1 and D_2 are given positive semidefinite matrices that play the role of the proximal metric. When D_1 and D_2 are appropriately chosen, the difficulty in solving the ADMM subproblems can be greatly alleviated. One particularly useful choice is the linearized ADMM [21, 32, 33] that choosing $D_2 = rI - \beta B^T B$ with $r > \beta\|B^T B\|$ to linearize the augmented terms. With this setting of D_i, the resulting subproblems reduce to estimating the resolvent operator of the function θ_i in the objective, and enjoy efficient solvability, see e.g., [4, 27, 31] for some applications.

In the previous work [30], the author extended the convergence result of ADMM to the proximal ADMM and proved that the relaxation parameter $\gamma \in (0, \frac{1+\sqrt{5}}{2})$ for the ADMM is also valid for the proximal ADMM (1.7). We refer the reader to [16, 20] for the iteration-complexity of this scheme. This paper further studies this issue and shows that the feasible region of γ depends on the proximal term equipped on the y-subproblem. More specially, we show that γ can be larger than $\frac{1+\sqrt{5}}{2}$ for the proximal ADMM (1.7) when D_2 is positive definite. For notational simplicity, we use the following proximal ADMM to conduct our analysis

$$\begin{cases} x^{k+1} = \arg\min\{\mathcal{L}_\beta(x, y^k, \lambda^k) \mid x \in \mathcal{X}\}, & (1.8a) \\ y^{k+1} = \arg\min\{\mathcal{L}_\beta(x^{k+1}, y, \lambda^k) + \frac{\tau\beta}{2}\|B(y - y^k)\|^2 \mid y \in \mathcal{Y}\}, & (1.8b) \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b), & (1.8c) \end{cases}$$

where $\tau \geq 0$ is a proximal factor. Our main purpose of this paper is to show that the proximal ADMM (1.8) is convergent for

$$\mathcal{D} = \left\{(\tau, \gamma) \mid \ \tau \geq 0 \ \& \ 0 < \gamma < \frac{1 - \tau + \sqrt{\tau^2 + 6\tau + 5}}{2}\right\}. \tag{1.9}$$

To the best of our knowledge, it is the first time that such a formulation is presented. We plot the specified domain $\mathcal{D}$ in Figure 1. As we can see, when the value of τ increases, the parameter γ can be chosen in the interval that is larger than $\left(0, \frac{1+\sqrt{5}}{2}\right)$. While when $\tau \to +\infty$, we have

$$\lim_{\tau \to +\infty} \frac{1 - \tau + \sqrt{\tau^2 + 6\tau + 5}}{2} = 2,$$

and thus γ can be chosen in the interval $(0, 2)$, which is very similar with the relaxation stepsize for the ALM (1.4). We will use this interesting phenomenon to establish a connection between ADMM and relaxed ALM (1.4) in Section 6. We remark that the convergent domain for (1.8) can be easily extended to the general proximal ADMM scheme (1.7). Indeed, if we want to estimate the feasible region of γ with a given matrix D_2, we only need to find a maximal value of τ such that $D_2 \succcurlyeq \tau\beta B^T B$.

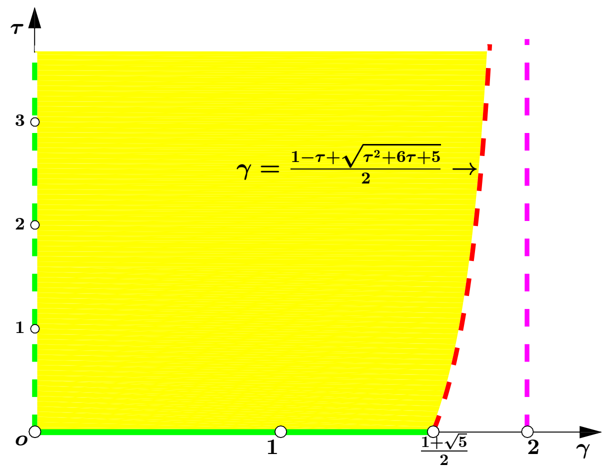

Fig. 1. convergent domain $\mathcal{D}$ for the proximal ADMM (1.8)

The rest of this paper is organized as follows. In Section 2, we characterize the model (1.1) as a variational inequality. We then interpret the proximal ADMM as a prediction-correction algorithm in Section 3 and show its convergence in Section 4. The convergence rate is analyzed in Section 5. The relationship between relaxed ALM and proximal ADMM is established in Section 6. Finally some conclusions are made in Section 7.

2 Variational inequality characterization of (1.1)

In this section, we present some preliminaries that will be used in our subsequent analysis. We first show how to characterize the optimality condition of the model (1.1) in the variational inequality (VI) context, which serves as a basis of the convergence analysis to be presented.

The Lagrangian function of the problem (1.1) is

$$L(x, y, \lambda) = \theta_1(x) + \theta_2(y) - \lambda^T(Ax + By - b), \tag{2.1}$$

which is defined on $\mathcal{X} \times \mathcal{Y} \times \Re^m$. In the above equality, (x, y) and λ are primal and dual variables, respectively. Let $\big((x^*, y^*), \lambda^*\big) \in \mathcal{X} \times \mathcal{Y} \times \Re^m$ be a saddle point of $L(x, y, \lambda)$. Then we have:

$$L_{\lambda\in\Re^m}(x^*, y^*, \lambda) \le L(x^*, y^*, \lambda^*) \le L_{x\in\mathcal{X}, y\in\mathcal{Y}}(x, y, \lambda^*).$$

Using (2.1), the above inequality can be expressed as

$$\left\{\begin{array}{lll} x^* \in \mathcal{X}, & \theta_1(x) - \theta_1(x^*) + (x - x^*)^T(-A^T\lambda^*) \ge 0, & \forall\, x \in \mathcal{X}, \\ y^* \in \mathcal{Y}, & \theta_2(y) - \theta_2(y^*) + (y - y^*)^T(-B^T\lambda^*) \ge 0, & \forall\, y \in \mathcal{Y}, \\ \lambda^* \in \Re^m, & (\lambda - \lambda^*)^T(Ax^* + By^* - b) \ge 0, & \forall\, \lambda \in \Re^m. \end{array}\right. \tag{2.2}$$

We can write (2.2) more compactly by the following variational inequality:

$$\text{VI}(\Omega, F, \theta) \qquad w^* \in \Omega, \quad \theta(u) - \theta(u^*) + (w - w^*)^T F(w^*) \ge 0, \quad \forall\, w \in \Omega, \tag{2.3a}$$

where

$$w = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad u = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad F(w) = \begin{pmatrix} -A^T\lambda \\ -B^T\lambda \\ Ax + By - b \end{pmatrix}, \tag{2.3b}$$

$$\theta(u) = \theta_1(x) + \theta_2(y), \qquad \text{and} \qquad \Omega = \mathcal{X} \times \mathcal{Y} \times \Re^m.$$

Note that the mapping $F(w)$ is affine with a skew-symmetric matrix, and thus is monotone with respect to Ω. The solution set of $\mathrm{VI}(\Omega, F, \theta)$ is denoted by Ω^*, and it is nonempty under the nonempty solution set assumption of (1.1).

Below we present a lemma to show how to characterize the optimality condition of an optimization model as a variational inequality. The proof is obvious and omitted here.

Proposition 2.1. Let $\theta(x) : \Re^n \to \Re$ and $f(x) : \Re^n \to \Re$ be convex functions defined on a closed convex set $\mathcal{X} \subset \Re^n$. Suppose $f(x)$ is differentiable and the solution set of the minimization problem $\min\{\theta(x) + f(x) \,|\, x \in \mathcal{X}\}$ is nonempty. Then,

$$x^* = \arg\min\{\theta(x) + f(x) \,|\, x \in \mathcal{X}\}, \tag{2.4a}$$

if and only if

$$x^* \in \mathcal{X}, \quad \theta(x) - \theta(x^*) + (x - x^*)^T \nabla f(x^*) \geq 0, \quad \forall\, x \in \mathcal{X}. \tag{2.4b}$$

3 A prediction-correction interpretation

In this section, we interpret the proximal ADMM (1.8) as a prediction-correction method in VI form. With this interpretation, we can easily present the convergence proof more compactly.

First, for the iterate $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ generated by the proximal ADMM (1.8), we define an auxiliary vector $\tilde{w}^k = (\tilde{x}^k, \tilde{y}^k, \tilde{\lambda}^k)$ as

$$\tilde{x}^k = x^{k+1}, \qquad \tilde{y}^k = y^{k+1}, \tag{3.1a}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta(Ax^{k+1} + By^k - b). \tag{3.1b}$$

In (1.8), we can see that x^k is not involved to execute the new iteration while (y^k, λ^k) are essentially required. Following [1], we call x^k an *intermediate variable* and (y, λ) *essential variables*, respectively. Accordingly, for $w = (x, y, \lambda)$ and $w^k = (x^k, y^k, \lambda^k)$ generated by (1.8), we use the notations

$$v = \begin{pmatrix} y \\ \lambda \end{pmatrix}, \quad \text{and} \quad v^k = \begin{pmatrix} y^k \\ \lambda^k \end{pmatrix}$$

to denote the essential parts of w and w^k, respectively. We denote the essential part of w^* in Ω^* by use $v^* = (y^*, \lambda^*)$ and let $\mathcal{V}^*$ denote all the collection of v^*.

Now, we present a lemma to show how to generate the auxiliary vector $\tilde{w}^k$ with given v^k in the proximal ADMM (1.8) via using VI. It also measures the discrepancy between $\tilde{w}^k$ and a solution point of $\mathrm{VI}(\Omega, F, \theta)$.

Lemma 3.1. *For given $v^k = (y^k, \lambda^k)$, let w^{k+1} be generated by the proximal ADMM (1.8) and $\tilde{w}^k$ be defined by (3.1). Then, we have*

$$\tilde{w}^k \in \Omega, \;\; \theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(\tilde{w}^k) \geq (v - \tilde{v}^k)^T Q(v^k - \tilde{v}^k), \;\; \forall w \in \Omega, \tag{3.2a}$$

where

$$Q = \begin{pmatrix} (1+\tau)\beta B^T B & 0 \\ -B & \frac{1}{\beta} I_m \end{pmatrix}. \tag{3.2b}$$

Proof. According to (2.4), the optimality condition of the x-subproblem in (1.8) is

$$x^{k+1} \in \mathcal{X}, \;\; \theta_1(x) - \theta_1(x^{k+1}) + (x - x^{k+1})^T\{-A^T\lambda^k + \beta A^T(Ax^{k+1} + By^k - b)\} \ge 0, \;\; \forall\, x \in \mathcal{X}.$$

Using the auxiliary vector $\tilde{w}^k$ defined in (3.1), it can be further written as

$$\tilde{x}^k \in \mathcal{X}, \;\; \theta_1(x) - \theta_1(\tilde{x}^k) + (x - \tilde{x}^k)^T(-A^T\tilde{\lambda}^k) \ge 0, \;\; \forall\, x \in \mathcal{X}. \tag{3.3a}$$

Similarly, the optimality condition of the y-subproblem can be written as

$$\begin{aligned} y^{k+1} \in \mathcal{Y}, \;\; \theta_2(y) - \theta_2(y^{k+1}) + (y - y^{k+1})^T \Big\{ & -B^T\lambda^k + \beta B^T(Ax^{k+1} + By^{k+1} - b) \\ & +\tau\beta B^T B(y^{k+1} - y^k)\Big\} \ge 0, \;\; \forall y \in \mathcal{Y}. \end{aligned}$$

Now, consider the $\{\cdot\}$ term in the last inequality, we have

$$\begin{aligned} y^{k+1} \in \mathcal{Y}, \;\; \theta_2(y) - \theta_2(y^{k+1}) + (y - y^{k+1})^T \Big\{ & -B^T\big(\lambda^k - (Ax^{k+1} + By^k - b)\big) \\ & +(1+\tau)\beta B^T B(y^{k+1} - y^k)\Big\} \ge 0, \;\; \forall y \in \mathcal{Y}. \end{aligned}$$

Using the notations defined in (3.1), we obtain

$$\tilde{y}^k \in \mathcal{Y}, \;\; \theta_2(y) - \theta_2(\tilde{y}^k) + (y - \tilde{y}^k)^T\{-B^T\tilde{\lambda}^k + (1+\tau)\beta B^T B(\tilde{y}^k - y^k)\} \ge 0, \;\; \forall y \in \mathcal{Y}. \tag{3.3b}$$

For the definition of $\tilde{\lambda}^k$ given by (3.1), we have

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) + (1/\beta)(\tilde{\lambda}^k - \lambda^k) = 0,$$

and it can be written as

$$\tilde{\lambda}^k \in \Re^m, \;\; (\lambda - \tilde{\lambda}^k)^T\big\{(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) + (1/\beta)(\tilde{\lambda}^k - \lambda^k)\big\} \ge 0, \;\; \forall\, \lambda \in \Re^m. \tag{3.3c}$$

Combining (3.3a), (3.3b) and (3.3c), and using the notations of (2.3), the assertion of this lemma is proved. □

Lemma 3.2. *For given $v^k = (y^k, \lambda^k)$, let w^{k+1} be generated by the proximal ADMM* (1.8) *and $\tilde{w}^k$ be defined by* (3.1). *Then, we have*

$$v^{k+1} = v^k - M(v^k - \tilde{v}^k), \tag{3.4a}$$

where

$$M = \begin{pmatrix} I & 0 \\ -\gamma\beta B & \gamma I_m \end{pmatrix}. \tag{3.4b}$$

Proof. It follows from (1.8) and (3.1) that

$$\begin{aligned} \lambda^{k+1} &= \lambda^k - \gamma\beta(A\tilde{x}^k + B\tilde{y}^k - b) \\ &= \lambda^k - \gamma\big[\beta(A\tilde{x}^k + By^k - b) - \beta B(y^k - \tilde{y}^k)\big] \\ &= \lambda^k - \gamma(\lambda^k - \tilde{\lambda}^k) + \gamma\beta B(y^k - \tilde{y}^k). \end{aligned}$$

Together with $y^{k+1} = \tilde{y}^k$, we have the following relationship

$$\begin{pmatrix} y^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} y^k \\ \lambda^k \end{pmatrix} - \begin{pmatrix} I & 0 \\ -\gamma\beta B & \gamma I_m \end{pmatrix} \begin{pmatrix} y^k - \tilde{y}^k \\ \lambda^k - \tilde{\lambda}^k \end{pmatrix}.$$

The proof is complete. □

According to the above lemmas, the proximal ADMM (1.8) can be characterized as a prediction-correction method in VI framework which consists of the prediction step (3.2) and the correction step (3.4).

4 Contraction property

This section presents the contraction property of the proximal ADMM (1.8), which is the basis for the convergence proof. Towards this goal, we first define a matrix:

$$H = QM^{-1}. \tag{4.1}$$

We prove a simple fact for the matrix H.

Lemma 4.1. *The matrix H defined in (4.1) is positive definite for any $\tau > 0$ and $\gamma > 0$ when the matrix B in (1.1) is full column rank.*

Proof. For the matrix M given in (3.4b), we have

$$M^{-1} = \begin{pmatrix} I & 0 \\ \beta B & \frac{1}{\gamma} I_m \end{pmatrix}.$$

Thus, it follows from (4.1) and (3.2b) that

$$\begin{aligned} H &= QM^{-1} = \begin{pmatrix} (1+\tau)\beta B^T B & 0 \\ -B & \frac{1}{\beta} I_m \end{pmatrix} \begin{pmatrix} I & 0 \\ \beta B & \frac{1}{\gamma} I_m \end{pmatrix} \\ &= \begin{pmatrix} (1+\tau)\beta B^T B & 0 \\ 0 & \frac{1}{\gamma\beta} I_m \end{pmatrix}. \end{aligned} \tag{4.2}$$

Because the matrix B in (1.1) is assumed to be full column rank, the positive definiteness of H follows immediately. □

According to (4.1), we have $Q = HM$. Recalling the correction step (3.4a), we can easily rewrite the right-hand side of (3.2a) as

$$(v - \tilde{v}^k)^T Q(v^k - \tilde{v}^k) = (v - \tilde{v}^k)^T H(v^k - v^{k+1}). \tag{4.3}$$

Since H is positive definite, applying the identity

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

to the right-hand side in (4.3) with

$$a = v, \quad b = \tilde{v}^k, \quad c = v^k, \quad \text{and} \quad d = v^{k+1},$$

we obtain

$$(v - \tilde{v}^k)^T H(v^k - v^{k+1}) = \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 - \|v - v^k\|_H^2\big) + \frac{1}{2}(\|v^k - \tilde{v}^k\|_H^2 - \|v^{k+1} - \tilde{v}^k\|_H^2). \tag{4.4}$$

Combining (4.4) with (4.3), we obtain

$$(v - \tilde{v}^k)^T Q(v^k - \tilde{v}^k) = \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 - \|v - v^k\|_H^2\big) + \frac{1}{2}(\|v^k - \tilde{v}^k\|_H^2 - \|v^{k+1} - \tilde{v}^k\|_H^2). \tag{4.5}$$

With the above analysis, we now present the following theorem, which is the basis for proving the strict contraction property of the scheme (1.8). It is also useful for estimating the convergence rate for the sequence generated by (1.8).

Theorem 4.2. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM* (1.8) *and $\tilde{w}^k$ be defined by* (3.1)*. Then we have*

$$\begin{aligned}
&\theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(\tilde{w}^k) \\
&\geq \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 - \|v - v^k\|_H^2\big) + \frac{1}{2}\|v^k - \tilde{v}^k\|_G^2, \;\; \forall w \in \Omega,
\end{aligned} \tag{4.6}$$

where H is defined in (4.1) *and*

$$G = Q^T + Q - M^T H M. \tag{4.7}$$

Proof. Substituting (4.5) into the right-hand side of (3.2a), we obtain

$$\begin{aligned}
&\theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(\tilde{w}^k) \\
&\geq \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 - \|v - v^k\|_H^2\big) + \frac{1}{2}\big(\|v^k - \tilde{v}^k\|_H^2 - \|v^{k+1} - \tilde{v}^k\|_H^2\big), \;\; \forall w \in \Omega,
\end{aligned} \tag{4.8}$$

The last term on the right-hand side of (4.8) can be represented as

$$\begin{aligned}
&\|v^k - \tilde{v}^k\|_H^2 - \|v^{k+1} - \tilde{v}^k\|_H^2 \\
&= \|v^k - \tilde{v}^k\|_H^2 - \|(v^k - \tilde{v}^k) - (v^k - v^{k+1})\|_H^2 \\
&\overset{(3.4a)}{=} \|v^k - \tilde{v}^k\|_H^2 - \|(v^k - \tilde{v}^k) - M(v^k - \tilde{v}^k)\|_H^2 \\
&= 2(v^k - \tilde{v}^k)^T H M (v^k - \tilde{v}^k) - (v^k - \tilde{v}^k)^T M^T H M (v^k - \tilde{v}^k) \\
&\overset{(4.1)}{=} (v^k - \tilde{v}^k)^T (Q^T + Q - M^T H M)(v^k - \tilde{v}^k).
\end{aligned}$$

Substituting this equation into (4.8) and recalling the definition of the matrix G given by (4.7), the assertion of this theorem is proved. □

Then, we have the following theorem.

Theorem 4.3. *For the sequence $\{w^k\}$ generated by the proximal ADMM* (1.8) *and $\tilde{w}^k$ be defined by* (3.1)*, we have*

$$\|v^{k+1} - v^*\|_H^2 \leq \|v^k - v^*\|_H^2 - \|v^k - \tilde{v}^k\|_G^2, \quad \forall v^* \in \mathcal{V}^*. \tag{4.9}$$

Proof. Setting $w = w^*$ in (4.6), we get

$$\|v^k - v^*\|_H^2 - \|v^{k+1} - v^*\|_H^2 \geq \|v^k - \tilde{v}^k\|_G^2 + 2\{\theta(\tilde{u}^k) - \theta(u^*) + (\tilde{w}^k - w^*)^T F(\tilde{w}^k)\}. \tag{4.10}$$

Then, using the optimality of w^* and the monotonicity of $F(w)$, we have

$$\theta(\tilde{u}^k) - \theta(u^*) + (\tilde{w}^k - w^*)^T F(\tilde{w}^k) \geq \theta(\tilde{u}^k) - \theta(u^*) + (\tilde{w}^k - w^*)^T F(w^*) \geq 0,$$

and thus

$$\|v^k - v^*\|_H^2 - \|v^{k+1} - v^*\|_H^2 \geq \|v^k - \tilde{v}^k\|_G^2. \tag{4.11}$$

The assertion (4.9) follows directly. □

Obviously, when the matrix G defined in (4.7) is positive definite, the inequality (4.9) indicates that the sequence $\{v^k\}$ generated by (1.8) is strictly contractive with respect to the solution set $\mathcal{V}^*$, and the convergence of the sequence can be easily established. Thus the symmetric matrix G

plays a key role in the convergence analysis. Now let us make a further investigation of the positive definiteness for G. Since $HM = Q$ (see (4.1)), we have $M^T HM = M^T Q$. Note that

$$M^T Q = \begin{pmatrix} I & -\gamma\beta B^T \\ 0 & \gamma I_m \end{pmatrix} \begin{pmatrix} (1+\tau)\beta B^T B & 0 \\ -B & \frac{1}{\beta} I_m \end{pmatrix} = \begin{pmatrix} (1+\tau+\gamma)\beta B^T B & -\gamma B^T \\ -\gamma B & \frac{\gamma}{\beta} I_m \end{pmatrix}.$$

Using (3.2b) and the above equation, we have

$$\begin{aligned} G &= (Q^T + Q) - M^T HM \\ &= \begin{pmatrix} 2(1+\tau)\beta B^T B & -B^T \\ -B & \frac{2}{\beta} I_m \end{pmatrix} - \begin{pmatrix} (1+\tau+\gamma)\beta B^T B & -\gamma B^T \\ -\gamma B & \frac{\gamma}{\beta} I_m \end{pmatrix} \\ &= \begin{pmatrix} (1+\tau-\gamma)\beta B^T B & -(1-\gamma) B^T \\ -(1-\gamma) B & \frac{1}{\beta}(2-\gamma) I_m \end{pmatrix}. \end{aligned} \tag{4.12}$$

According to (4.12), when the matrix B is full column rank and $\gamma \in (0, 1]$, the matrix G is positive definite, and the proximal ADMM (1.8) is convergent under Theorem 4.3. In fact, the convergence of the commonly used ADMM and its proximal part in case of $\gamma = 1$ is obtained by Theorem 4.3 ; see e.g., [24]. However, for the Glowinski's case (e.g., $\gamma > 1$), G may not be positive definite, and the contraction property for (1.8) can not be obtained by only using (4.9). The remaining task is to further investigate the term $\|v^k - \tilde{v}^k\|_G^2$. This is the main difficulty of the convergence analysis for (1.8).

Lemma 4.4. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM* (1.8) *and $\tilde{w}^k$ be defined by* (3.1)*. Then we have*

$$\begin{aligned} \|v^k - \tilde{v}^k\|_G^2 &= (1+\tau)\beta\|B(y^k - y^{k+1})\|^2 + (2-\gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 \\ &\quad + 2\beta(Ax^{k+1} + By^{k+1} - b)^T B(y^k - y^{k+1}). \end{aligned} \tag{4.13}$$

Proof. Because $G = \begin{pmatrix} (1+\tau-\gamma)\beta B^T B & -(1-\gamma) B^T \\ -(1-\gamma) B & \frac{1}{\beta}(2-\gamma) I_m \end{pmatrix}$ (see (4.12)), $v = \begin{pmatrix} y \\ \lambda \end{pmatrix}$ and $\tilde{y}^k = y^{k+1}$, we have

$$\begin{aligned} \|v^k - \tilde{v}^k\|_G^2 &= (1+\tau-\gamma)\beta\|B(y^k - y^{k+1})\|^2 - 2(1-\gamma)(y^k - y^{k+1})^T B^T(\lambda^k - \tilde{\lambda}^k) \\ &\quad + (2-\gamma)\frac{1}{\beta}\|\lambda^k - \tilde{\lambda}^k\|^2. \end{aligned} \tag{4.14}$$

Notice that (see (3.1))

$$\lambda^k - \tilde{\lambda}^k = \beta(Ax^{k+1} + By^k - b) = \beta\big\{(Ax^{k+1} + By^{k+1} - b) + B(y^k - y^{k+1})\big\}.$$

Thus, we have

$$\begin{aligned} \|\lambda^k - \tilde{\lambda}^k\|^2 &= \beta^2\|(Ax^{k+1} + By^{k+1} - b) + B(y^k - y^{k+1})\|^2 \\ &= \beta^2\|Ax^{k+1} + By^{k+1} - b\|^2 + 2\beta^2(Ax^{k+1} + By^{k+1} - b)^T B(y^k - y^{k+1}) \\ &\quad + \beta^2\|B(y^k - y^{k+1})\|^2. \end{aligned} \tag{4.15}$$

Substituting (4.15) into the right-hand side of (4.14), we obtain

$$\begin{aligned}
\|v^k - \tilde{v}^k\|_G^2 &= (1+\tau-\gamma)\beta\|By^k - By^{k+1}\|^2 - 2(1-\gamma)\beta(y^k - y^{k+1})^T B^T(Ax^{k+1} + By^{k+1} - b) \\
&\quad -2(1-\gamma)\beta\|By^k - By^{k+1}\|^2 + (2-\gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 \\
&\quad +(2-\gamma)\beta\|By^k - By^{k+1}\|^2 + 2(2-\gamma)\beta(y^k - y^{k+1})^T B^T(Ax^{k+1} + By^{k+1} - b) \\
&= (1+\tau)\beta\|By^k - By^{k+1}\|^2 + (2-\gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 \\
&\quad +2\beta(y^k - y^{k+1})^T B^T(Ax^{k+1} + By^{k+1} - b).
\end{aligned}$$

This completes the proof. □

In the following we will treat the crossing term in the right-hand side of (4.13), namely,

$$\beta(Ax^{k+1} + By^{k+1} - b)^T B(y^k - y^{k+1}).$$

For given $\tau > 0$, we restrict the parameter γ to satisfy

$$\gamma > 0 \quad \text{and} \quad \gamma + \frac{(1-\gamma)^2}{1+\tau} < 2. \tag{4.16}$$

The solution of the inequalities (4.16) is

$$0 < \gamma < \frac{1-\tau+\sqrt{\tau^2+6\tau+5}}{2}, \tag{4.17}$$

which specifies the domain of (τ, γ) defined in (1.9) and depicted in Figure 1. For each pair of τ and γ chosen in (1.9), it is trivial to select a constant T such that

$$T \in \Big(\gamma + \frac{(1-\gamma)^2}{1+\tau}, 2\Big). \tag{4.18}$$

In the following, we will use the constant T to make our analysis.

Lemma 4.5. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM (1.8). Then, for any $(\tau,\gamma) \in \mathcal{D}$ (1.9), we have*

$$\begin{aligned}
&\beta(Ax^{k+1} + By^{k+1} - b)^T B(y^k - y^{k+1}) \\
&\geq -\frac{(1-\gamma)^2}{2(T-\gamma)}\beta\|B(y^{k+1} - y^k)\|^2 - \frac{T-\gamma}{2}\beta\|Ax^k + By^k - b\|^2 \\
&\quad +\frac{\tau}{2}\beta\|B(y^{k+1} - y^k)\|^2 - \frac{\tau}{2}\beta\|B(y^k - y^{k-1})\|^2,
\end{aligned} \tag{4.19}$$

where T is a constant defined in (4.18).

Proof. Note that the optimality condition of the y-subproblem (1.8b) is

$$\begin{aligned}
y^{k+1} \in \mathcal{Y},\ \ \theta_2(y) - \theta_2(y^{k+1}) + (y - y^{k+1})^T \Big\{ &-B^T\lambda^k + \beta B^T(Ax^{k+1} + By^{k+1} - b) \\
&+\tau\beta B^T B(y^{k+1} - y^k)\Big\} \geq 0,\ \ \forall y \in \mathcal{Y}.
\end{aligned} \tag{4.20}$$

Analogously, for the previous iteration, we have

$$\begin{aligned}
y^k \in \mathcal{Y},\ \ \theta_2(y) - \theta_2(y^k) + (y - y^k)^T \Big\{ &-B^T\lambda^{k-1} + \beta B^T(Ax^k + By^k - b) \\
&+\tau\beta B^T B(y^k - y^{k-1})\Big\} \geq 0,\ \ \forall y \in \mathcal{Y}.
\end{aligned} \tag{4.21}$$

Setting $y = y^k$ and $y = y^{k+1}$ in (4.20) and (4.21), respectively, and then adding them, we get

$$(y^k - y^{k+1})^T B^T \Big\{(\lambda^{k-1} - \lambda^k) - \beta(Ax^k + By^k - b) + \beta(Ax^{k+1} + By^{k+1} - b) \\ -\tau\beta B(y^k - y^{k-1}) + \tau\beta B(y^{k+1} - y^k)\Big\} \geq 0. \quad (4.22)$$

Note that in the $\{k-1\}$-th iteration (see (1.8c)), we have

$$\lambda^k = \lambda^{k-1} - \gamma\beta(Ax^k + By^k - b). \quad (4.23)$$

Substituting (4.23) into (4.22) and with a simple manipulation, we have

$$(y^k - y^{k+1})^T B^T \Big\{\beta(Ax^{k+1} + By^{k+1} - b) - (1-\gamma)\beta(Ax^k + By^k - b) \\ -\tau\beta B(y^k - y^{k-1}) + \tau\beta B(y^{k+1} - y^k)\Big\} \geq 0.$$

Then, we get

$$\begin{aligned} &(y^k - y^{k+1})^T \beta B^T (Ax^{k+1} + By^{k+1} - b) \\ \geq\ & (1-\gamma)(y^k - y^{k+1})^T \beta B^T (Ax^k + By^k - b) \\ &+\tau\beta(y^k - y^{k+1})^T B^T B(y^k - y^{k-1}) + \tau\beta\|B(y^{k+1} - y^k)\|^2. \end{aligned} \quad (4.24)$$

For the constant T selected by (4.18), we have $T - \gamma > 0$. Using the Cauchy-Schwarz inequality to the crossing terms in the right-hand side of (4.24), we have

$$\begin{aligned} &(1-\gamma)(y^k - y^{k+1})^T \beta B^T (Ax^k + By^k - b) \\ \geq\ & -\frac{(1-\gamma)^2}{2(T-\gamma)}\beta\|B(y^{k+1} - y^k)\|^2 - \frac{T-\gamma}{2}\beta\|Ax^k + By^k - b\|^2, \end{aligned} \quad (4.25)$$

and

$$\tau\beta(y^k - y^{k+1})^T B^T B(y^k - y^{k-1}) \geq -\frac{\tau}{2}\beta\|B(y^{k+1} - y^k)\|^2 - \frac{\tau}{2}\beta\|B(y^k - y^{k-1})\|^2. \quad (4.26)$$

Substituting (4.25) and (4.26) into (4.24), this lemma is proved. □

Based on Lemma 4.4 and Lemma 4.5, we now establish the following results:

Lemma 4.6. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM (1.8) and $\tilde{w}^k$ be defined by (3.1). Then, for any $(\tau, \gamma) \in \mathcal{D}$ (1.9), we have*

$$\begin{aligned} \|v^k - \tilde{v}^k\|_G^2 \ \geq\ & \Big[(1+\tau) - \frac{(1-\gamma)^2}{T-\gamma}\Big]\beta\|B(y^k - y^{k+1})\|^2 + (2-T)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 \\ &+(T-\gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 - (T-\gamma)\beta\|Ax^k + By^k - b\|^2 \\ &+\tau\beta\|B(y^{k+1} - y^k)\|^2 - \tau\beta\|B(y^k - y^{k-1})\|^2, \end{aligned} \quad (4.27)$$

where T is a constant defined in (4.18).

Proof. Substituting (4.19) into (4.13), we obtain the assertion (4.27) immediately. □

5 Convergence analysis

In this section, we show the global convergence and estimate the convergence rate in terms of the iteration complexity for the proximal ADMM (1.8).

5.1 Global convergence

Theorem 5.1. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM (1.8). Then, for any $(\tau,\gamma)\in\mathcal{D}$ (1.9), we have*

$$\begin{aligned}&\|v^{k+1}-v^*\|_H^2+(T-\gamma)\beta\|Ax^{k+1}+By^{k+1}-b\|^2+\tau\beta\|B(y^{k+1}-y^k)\|^2\\ \le\ &\big(\|v^k-v^*\|_H^2+(T-\gamma)\beta\|Ax^k+By^k-b\|^2+\tau\beta\|B(y^k-y^{k-1})\|^2\big)\\ -\ &\beta\Big\{\big[(1+\tau)-\frac{(1-\gamma)^2}{T-\gamma}\big]\|B(y^k-y^{k+1})\|^2+(2-T)\|Ax^{k+1}+By^{k+1}-b\|^2\Big\},\ \forall v^*\in\mathcal{V}^*,\end{aligned}\tag{5.1}$$

where T is a constant defined in (4.18).

Proof. The assertion follows from (4.9) and (4.27) directly. □

We summarize the global convergence of (1.8) in the following theorem

Theorem 5.2. *For the sequence $\{w^k\}$ generated by the proximal ADMM (1.8), we have*

$$\lim_{k\to\infty}\big(\|B(y^k-y^{k+1})\|^2+\|Ax^{k+1}+By^{k+1}-b\|^2\big)=0.\tag{5.2}$$

Moreover, if the matrix B is assumed to be full column rank, then the sequence $\{v^k\}$ converges to a solution point $v^\infty\in\mathcal{V}^*$.

Proof. According to (4.18), we have

$$\begin{cases}1+\tau-\dfrac{(1-\gamma)^2}{T-\gamma}>0,\\[2mm] 2-T>0.\end{cases}$$

Let (y^0,λ^0) be the initial point and x^1 be the computational result of (y^0,λ^0). Using (5.1), we have

$$\begin{aligned}&\beta\sum_{k=1}^{\infty}\Big(\big[1+\tau-\frac{(1-\gamma)^2}{T-\gamma}\big]\|B(y^k-y^{k+1})\|^2+(2-T)\|Ax^{k+1}+By^{k+1}-b\|^2\Big)\\ \le\ &\big(\|v^1-v^*\|_H^2+(T-\gamma)\beta\|Ax^1+By^1-b\|^2+\tau\beta\|B(y^1-y^0)\|^2\big).\end{aligned}$$

The assertion (5.2) is proved.

Furthermore, we know from (5.1) that the sequence $\{v^k\}$ is bounded. Thus it must have a finite cluster point v^∞, and there exists a subsequence $\{v^{k_j}\}$ convergent to this point. We will prove that v^∞ is a solution point of (2.3). Let $\tilde{x}^\infty$ be induced by (1.8a) with given $(y^\infty,\lambda^\infty)$. Recall the matrix B is assumed to be full column rank. From (5.2), we immediately have

$$B(y^\infty-\tilde{y}^\infty)=0\qquad\text{and}\qquad\lambda^\infty-\tilde{\lambda}^\infty=0.$$

Then, it follows from (3.2a) that

$$\tilde{w}^\infty\in\Omega,\ \ \theta(u)-\theta(\tilde{u}^\infty)+(w-\tilde{w}^\infty)^TF(\tilde{w}^\infty)\ge0,\ \ \forall w\in\Omega,\tag{5.4}$$

we obtain that $\tilde{w}^\infty=w^\infty$ is a solution point of (2.3). On the other hand, since (5.1) holds for any solution point of (2.3) and $w^\infty\in\Omega^*$, we have

$$\begin{aligned}&\|v^{k+1}-v^\infty\|_H^2+(T-\gamma)\beta\|Ax^{k+1}+By^{k+1}-b\|^2+\tau\beta\|B(y^{k+1}-y^k)\|^2\\ &\le\|v^k-v^\infty\|_H^2+(T-\gamma)\beta\|Ax^k+By^k-b\|^2+\tau\beta\|B(y^k-y^{k-1})\|^2.\end{aligned}\tag{5.5}$$

Because $\lim_{k\to\infty}\|Ax^k+By^k-b\|^2=0$ and $\lim_{k\to\infty}\|B(y^k-y^{k-1})\|^2=0$, the sequence $\{v^k\}$ cannot have another cluster point and thus it converges to a solution point $v^*=v^\infty\in\mathcal{V}^*$. □

5.2 Convergence rate

In the following, we present a theorem to characterize the solution set of VI (2.3). With this theorem, we can easily estimate the convergence rate in terms of the iteration complexity in the ergodic sense. Its proof can be found in [8] (Theorem 2.3.5) or [24] (Theorem 2.1), here we omit it.

Theorem 5.3. *The solution set of* $VI(\Omega, F, \theta)$ *is convex and it can be characterized as*

$$\Omega^* = \bigcap_{w\in\Omega} \big\{\tilde{w} \in \Omega : \big(\theta(u) - \theta(\tilde{u})\big) + (w - \tilde{w})^T F(w) \geq 0\big\}. \tag{5.6}$$

Therefore, for a given accuracy $\epsilon > 0$, $\tilde{w} \in \Omega$ is called an ϵ-approximate solution point of $\mathrm{VI}(\Omega, F, \theta)$ if it satisfies

$$\theta(u) - \theta(\tilde{u}) + (w - \tilde{w})^T F(w) \geq -\epsilon, \;\; \forall\, w \in \mathcal{D}_{(\tilde{w})},$$

where

$$\mathcal{D}_{(\tilde{w})} = \{w \in \Omega \,|\, \|w - \tilde{w}\| \leq 1\}.$$

To estimate the convergence rate in terms of the iteration complexity for a sequence $\{w^k\}$, we need to show that for given $\epsilon > 0$, after t iterations, this sequence can offer a point $\tilde{w} \in \Omega$ such that

$$\tilde{w} \in \Omega \quad \text{and} \quad \sup_{w\in\mathcal{D}_{(\tilde{w})}} \big\{\theta(\tilde{u}) - \theta(u) + (\tilde{w} - w)^T F(w)\big\} \leq \epsilon. \tag{5.7}$$

Before we present the analysis for estimating the convergence rate for (1.8), we notice that it follows from the monotonicity of F that

$$(w - \tilde{w}^k)^T F(w) \geq (w - \tilde{w}^k)^T F(\tilde{w}^k).$$

Thus, substituting this inequality into (4.6), we obtain

$$\theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(w) \geq \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 - \|v - v^k\|_H^2\big) + \frac{1}{2}\|v^k - \tilde{v}^k\|_G^2, \;\; \forall w \in \Omega. \tag{5.8}$$

Notice that (see (4.27))

$$\begin{aligned}\|v^k - \tilde{v}^k\|_G^2 \;\; \geq \;\; & (T - \gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 - (T - \gamma)\beta\|Ax^k + By^k - b\|^2 \\ & +\tau\beta\|B(y^{k+1} - y^k)\|^2 - \tau\beta\|B(y^k - y^{k-1})\|^2.\end{aligned}$$

Substituting it in (5.8), we have

$$\begin{aligned}& \theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(w) \\ & \qquad + \frac{1}{2}\big(\|v - v^k\|_H^2 + (T - \gamma)\beta\|Ax^k + By^k - b\|^2 + \tau\beta\|B(y^k - y^{k-1})\|^2\big) \\ & \geq \;\; \frac{1}{2}\big(\|v - v^{k+1}\|_H^2 + (T - \gamma)\beta\|Ax^{k+1} + By^{k+1} - b\|^2 + \tau\beta\|B(y^{k+1} - y^k)\|^2\big), \;\; \forall w \in \Omega.\end{aligned} \tag{5.9}$$

This enables us to derive the assertions in the following theorem.

Theorem 5.4. *Let the sequence $\{w^k\}$ be generated by the proximal ADMM (1.8) and $\tilde{w}^k$ be defined by (3.1). Then, for $(\tau,\gamma)\in\mathcal{D}$ (1.9) and any integer number $t>0$, we have*

$$\begin{aligned}&\theta(\tilde{u}_t)-\theta(u)+(\tilde{w}_t-w)^TF(w)\\ &\quad\le\ \frac{1}{2t}\big(\|v-v^1\|_H^2+(T-\gamma)\beta\|Ax^1+By^1-b\|^2+\tau\beta\|B(y^1-y^0)\|^2\big)\quad\forall w\in\Omega,\end{aligned}\tag{5.10a}$$

where

$$\tilde{w}_t=\frac{1}{t}\sum_{k=1}^{t}\tilde{w}^k.\tag{5.10b}$$

Proof. First, we rewrite the results in (5.9) in our desired forms:

$$\begin{aligned}&\theta(\tilde{u}^k)-\theta(u)+(\tilde{w}^k-w)^TF(w)\\ &\quad+\frac{1}{2}\big(\|v-v^{k+1}\|_H^2+(T-\gamma)\beta\|Ax^{k+1}+By^{k+1}-b\|^2+\tau\beta\|B(y^{k+1}-y^k)\|^2\big)\\ &\quad\le\frac{1}{2}\big(\|v-v^k\|_H^2+(T-\gamma)\beta\|Ax^k+By^k-b\|^2+\tau\beta\|B(y^k-y^{k-1})\|^2\big),\quad\forall w\in\Omega,\end{aligned}\tag{5.11}$$

Summarizing the inequalities (5.11) over $k=1,\ldots,t$, we obtain

$$\begin{aligned}&\sum_{k=1}^{t}\theta(\tilde{u}^k)-t\theta(u)+\Big(\sum_{k=1}^{t}\tilde{w}^k-tw\Big)^TF(w)\\ &\quad\le\ \frac{1}{2}\big(\|v-v^1\|_H^2+(T-\gamma)\beta\|Ax^1+By^1-b\|^2+\tau\beta\|B(y^1-y^0)\|^2\big),\quad\forall w\in\Omega.\end{aligned}\tag{5.12}$$

Then, using the notation of $\tilde{w}_t$ in (5.10b), the last inequality can be written as

$$\begin{aligned}&\frac{1}{t}\sum_{k=1}^{t}\theta(\tilde{u}^k)-\theta(u)+(\tilde{w}_t-w)^TF(w)\\ &\quad\le\ \frac{1}{2t}\big(\|v-v^1\|_H^2+(T-\gamma)\beta\|Ax^1+By^1-b\|^2+\tau\beta\|B(y^1-y^0)\|^2\big),\quad\forall w\in\Omega.\end{aligned}\tag{5.13}$$

It follows from the definition of $\tilde{w}_t$ in (5.10b) that

$$\tilde{u}_t=\frac{1}{t}\sum_{k=1}^{t}\tilde{u}^k.$$

Since $\theta(u)$ is convex, it follows that

$$\theta(\tilde{u}_t)\le\frac{1}{t}\sum_{k=1}^{t}\theta(\tilde{u}^k).$$

Substituting it into (5.13), the assertion (5.10) follows directly. □

Recall (5.7), we can see from (5.10) that after t iterations, the proximal ADMM (1.8) is able to generate an approximate solution point (i.e., $\tilde{w}_t$ given in (5.10b)) with an accuracy of $O(1/t)$. A worse-case $O(1/t)$ convergence rate in the ergodic sense is thus established for the proximal ADMM (1.8).

6 Bridges between ADMM and ALM

In this section, we establish a bridge between the relaxation factor of ADMM and ALM by using (1.9). Consider the following convex problems with linear constraints

$$\min\{\theta_1(x) \mid Ax = b,\, x \in \mathcal{X}\}. \tag{6.1}$$

Although it is not separable, we can also employ ADMM to solve it. To make ADMM applicable, by introducing a new variable y, the problem (6.2) converts to the following separable convex problem

$$\min\{\theta_1(x) + \mathcal{I}(y) \mid Ax + y = b,\, x \in \mathcal{X}, y \in \Re^m\}, \tag{6.2}$$

where $\mathcal{I}(\cdot)$ is the indicator function that is defined as

$$\mathcal{I}(y) = \begin{cases} 0, & y = 0, \\ +\infty, & y \neq 0. \end{cases} \tag{6.3}$$

We set the initial point $(y^0, \lambda^0) = (0, 0)$. Since $y^k = 0$ for all $k \in \{0, \cdots \infty\}$ in the ADMM iteration, the ADMM (1.6) by Fortin and Glowinski for solving (6.2) reduces to

$$\begin{cases} x^{k+1} = \arg\min\big\{\theta_1(x) - (\lambda^k)^T Ax + \dfrac{\beta}{2}\|Ax - b\|^2 \;\big|\; x \in \mathcal{X}\big\}, & \text{(6.4a)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} - b), & \text{(6.4b)} \end{cases}$$

where $\gamma \in \big(0, \frac{1+\sqrt{5}}{2}\big)$.

The iteration scheme (6.4) is actually the relaxed ALM for solving (6.1). However, it is well known that the relaxation parameter γ of ALM can be chosen in the interval $\big(0, 2\big)$. Thus Glowinski's ADMM is still theoretically insufficient to build bridges between ADMM and ALM.

Now we show how to repair this theoretically incompleteness. If we set the initial point $(y^0, \lambda^0) = (0, 0)$, the proximal ADMM (1.8) for solving (6.2) iterates as

$$\begin{cases} x^{k+1} = \arg\min\big\{\theta_1(x) - (\lambda^k)^T Ax + \dfrac{\beta}{2}\|Ax + y^k - b\|^2 \;\big|\; x \in \mathcal{X}\big\}, & \text{(6.5a)} \\ y^{k+1} = \arg\min\big\{\mathcal{I}(y) - (\lambda^k)^T y + \dfrac{\tau\beta}{2}\|y - y^k\|^2\big\}, & \text{(6.5b)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + y^{k+1} - b), & \text{(6.5c)} \end{cases}$$

Let $\tau \to +\infty$. Then $\lim\limits_{\tau \to +\infty} \dfrac{1 - \tau + \sqrt{\tau^2 + 6\tau + 5}}{2} = 2$. Since $y^k = 0$ for all $k \in \{0, \cdots \infty\}$ in (6.5), the algorithm (6.5) reduces to

$$\begin{cases} x^{k+1} = \arg\min\big\{\theta_1(x) - (\lambda^k)^T Ax + \dfrac{\beta}{2}\|Ax - b\|^2 \;\big|\; x \in \mathcal{X}\big\}, & \text{(6.6a)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} - b), & \text{(6.6b)} \end{cases}$$

where $\gamma \in \big(0, 2\big)$.

Thus the proximal ADMM (1.8) with $\tau \to +\infty$ for solving (6.2) reduces to the relaxed ALM (1.4) for solving (6.1).

7 Conclusions

In this paper, we revisit the proximal alternating direction method of multipliers (ADMM). We show that the feasible region of the relaxation factor γ attached on the dual updating step of the proximal ADMM depends on the proximal term equipped on the y-subproblem; and can be enlarged when the proximal factor is positive. We derive the exact relationship between these two terms. At last, we provide a simple proof of the $O(1/t)$ convergence rate for the proximal ADMM in ergodic sense. This work also verifies the intuitive argument which we highlighted in [22]: If the primal subproblem is solved more conservatively (using a larger value of τ), we can employ a larger step size for updating the dual variable so that the dual subproblem is solved more aggressively; and vice versa.

During the finalization of this paper, we learned of a very recent work [17], in which the authors obtained a partially similar result for the proximal ADMM. They found that when D_2 is a sufficiently large positive multiple of the identity, γ can be chosen in the interval $(0, 2)$. However, our studies differs from theirs in two important aspects. First, the analysis in [17] requires some restrictive assumptions for θ_2 and β, while our theoretical results are obtained without any further assumptions. Secondly, we obtained an exact connection between the proximal term and the relaxation stepsize. As a byproduct, for a given positive matrix of D_2, we can determine a feasible interval of γ that is larger than $\frac{1+\sqrt{5}}{2}$, which is more general than the results in [17].

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2010), pp. 1-122.

[2] X. J. Cai, G. Y. Gu, B. S. He and X. M. Yuan, *A proximal point algorithm revisit on the alternating direction method of multipliers*, Sci. China Math., 56 (10), pp. 2179-2186.

[3] T. F. Chan and R. Glowinski, *Finite Element Approximation and Iterative Solution of a Class of Mildly Nonlinear Elliptic Equations*, Technical Report STAN-CS-78-674, Computer Science Department, Stanford University, 1978.

[4] R. H. Chan, M. Tao and X. M. Yuan, *Linearized alternating direction method of multipliers for constrained linear least-squares problem*, East Asian J. Appl. Math., 2(2012), pp. 326-341.

[5] C. H. Chen, B. S. He and X. M. Yuan, *Matrix completion via alternating direction method*, IMA J. Numer. Analy., 32 (2012), pp. 227-245.

[6] J. Eckstein and D.P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293-318.

[7] J. Eckstein and W. Yao, *Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives*, Pacific J. Optim., 11(2015), pp. 619-644.

[8] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity problems, Volume I*, Springer Series in Operations Research, Springer-Verlag, 2003.

[9] M. Fortin and R. Glowinski (eds), *Méthodes de Lagrangien Augmenté: Application à la Résolution Numérique de Problèmes aux Limites*, Dunod, Paris (1982)

[10] M. Fortin and R. Glowinski (eds), *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, North Holland, Amsterdam, The Netherlands (1983)

[11] D. Gabay, *Applications of the method of multipliers to variational inequalities*, Augmented Lagrange Methods: Applications to the Solution of Boundary-valued Problems, M. Fortin and R. Glowinski, eds., North Holland, Amsterdam, The Netherlands, (1983), pp. 299–331.

[12] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.

[13] R. Glowinski, *On alternating direction methods of multipliers: A historical perspective*, In Modeling, Simulation and Optimization for Science and Technology, W. Fitzgibbon, Y.A. Kuznetsov, P. Neittaanmki & O. Pironneau, eds., Computational Methods in Applied Sciences, Vol. 34, Springer, Dordrecht, 2014, 59-82.

[14] R. Glowinski, and A. Marrocco, *Sur l'approximation par éléments finis d'ordre un et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires*, Revue Fr. Autom. Inform. Rech. Opér., Anal. Numér. 2 (1975), pp. 41–76.

[15] E. G. Gol'shtein and N. V. Tret'yakov, *Modified Lagrangians in convex programming and their generalizations*, Point-to-Set Maps and Mathematical Programming, pp. 86-97, Springer Berlin Heidelberg, 1979.

[16] Max L.N. Goncalves, J. G. Melo and R. D. C. Monteiro, *Extending the ergodic convergence rate of the proximal ADMM*, ArXiv: 1611.02903, Nov 2016.

[17] Max L.N. Goncalves, J. G. Melo and R. D. C. Monteiro, *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, Optimization Online, Feb 2016.

[18] B. S. He, L. Z. Liao, D. R. Han, and H. Yang, *A new inexact alternating directions method for monotone variational inequalities*, Math. Program., 92(2002), pp. 103-118.

[19] B. S. He, H. Liu, Z. R. Wang and X. M. Yuan, *A strictly contractive Peaceman-Rachford splitting method for convex programming*, SIAM J. Optim., 24(2014), pp. 1011-1040.

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

[21] B. S. He, F. Ma and X. M. Yuan, *Linearized alternating direction method of multipliers via positive-indefinite proximal regularization for convex programming*, manuscript, July, 2016.

[22] B. S. He, F. Ma and X. M. Yuan, *Positive-indefinite proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems*, manuscript, September, 2016.

[23] B. S. He, M. H. Xu, and X. M. Yuan, *Solving large-scale least squares covariance matrix problems by alternating direction methods*, SIAM J. Matrix Anal. Appli., 32(2011), pp. 136-152.

[24] B. S. He and X. M. Yuan, *On the $O(1/t)$ convergence rate of the Douglas-Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700-709.

[25] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appli., 4(1969), pp. 303-320.

[26] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16(1979), pp. 964-979.

[27] S. Q. Ma, *Alternating proximal gradient method for convex minimization*, J. Sci. Comput., 68(2016), pp. 546-572.

[28] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, In Optimization edited by R. Fletcher, pp. 283-298, Academic Press, New York, 1969.

[29] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res. 1(1976), pp. 97-116.

[30] M. H. Xu, *Proximal alternating directions method for structured variational inequalities*, J. Optim. Theory Appli., 134(2007), pp. 107-117.

[31] X. F. Wang and X. M. Yuan, *The linearized alternating direction method of multipliers for Dantzig selector*, SIAM J. Sci. Comput., 34 (2012), pp. 2792–2811.

[32] J. F. Yang and X. M. Yuan, *Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization*, Math. Comp., 82 (2013), pp. 301-329.

[33] X. Q. Zhang, M. Burger and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, J. Sci. Comput., 46 (2010), pp. 20–46.