SMAI-JCM
SMAI Journal of
Computational Mathematics

Block-wise Alternating Direction Method of Multipliers for Multiple-block Convex Programming and Beyond

Bingsheng He & Xiaoming Yuan

Volume 1 (2015), p. 145-174.

<http://smai-jcm.cedram.org/item?id=SMAI-JCM_2015__1__145_0>

SMAI Journal of Computational Mathematics
Vol. 1, 145-174 (2015)

Block-wise Alternating Direction Method of Multipliers for Multiple-block Convex Programming and Beyond

Bingsheng He [1]
Xiaoming Yuan [2]

[1] Department of Mathematics, South University of Science and Technology of China, and Department of Mathematics, Nanjing University, China
E-mail address: hebma@nju.edu.cn
[2] Department of Mathematics, Hong Kong Baptist University, Hong Kong
E-mail address: xmyuan@hkbu.edu.hk.

Abstract. The alternating direction method of multipliers (ADMM) is a benchmark for solving a linearly constrained convex minimization model with a two-block separable objective function; and it has been shown that its direct extension to a multiple-block case where the objective function is the sum of more than two functions is not necessarily convergent. For the multiple-block case, a natural idea is to artificially group the objective functions and the corresponding variables as two groups and then apply the original ADMM directly —- the block-wise ADMM is accordingly named because each of the resulting ADMM subproblems may involve more than one function in its objective. Such a subproblem of the block-wise ADMM may not be easy as it may require minimizing more than one function with coupled variables simultaneously. We discuss how to further decompose the block-wise ADMM's subproblems and obtain easier subproblems so that the properties of each function in the objective can be individually and thus effectively used, while the convergence can still be ensured. The generalized ADMM and the strictly contractive Peaceman-Rachford splitting method, two schemes closely relevant to the ADMM, will also be extended to the block-wise versions to tackle the multiple-block convex programming cases. We present the convergence analysis, including both the global convergence and the worst-case convergence rate measured by the iteration complexity, for these three block-wise splitting schemes in a unified framework.

Math. classification. 90C25, 90C06, 65K05.

Keywords. Convex programming, Operator splitting methods, Alternating direction method of multipliers, proximal point algorithm, Douglas-Rachford splitting method, Peaceman-Rachford splitting method, Convergence rate, Iteration complexity.

1. Introduction

We consider a separable convex minimization problem with linear constraints and its objective function is the sum of more than one function without coupled variables:

$$\min\Big\{\sum_{i=1}^{m}\theta_i(x_i)\mid\sum_{i=1}^{m}A_ix_i=b,\ x_i\in X_i, i=1,\cdots,m\Big\}, \tag{1.1}$$

where $\theta_i:\mathbb{R}^{n_i}\to\mathbb{R}$ $(i=1,\cdots,m)$ are convex (not necessarily smooth) closed functions; $A_i\in\mathbb{R}^{l\times n_i}$, $b\in\mathbb{R}^l$, and $X_i\subseteq\mathbb{R}^{n_i}$ $(i=1,\cdots,m)$ are convex sets. The solution set of (1.1) is assumed to be nonempty throughout our discussions in this paper. We also assume that matrices $A_i^TA_i$ for $(i=1,\dots,m)$ are all nonsingular.

Bingsheng He was supported by the NSFC grant: 11471156.
Xiaoming Yuan was supported by a General Research Fund from Hong Kong Research Grants Council.

Let the augmented Lagrangian function of (1.1) be

$$\mathcal{L}_\beta^m(x_1,x_2,\cdots,x_m,\lambda)=\sum_{i=1}^m\theta_i(x_i)-\lambda^T\Big(\sum_{i=1}^m A_ix_i-b\Big)+\frac{\beta}{2}\|\sum_{i=1}^m A_ix_i-b\|^2, \tag{1.2}$$

with $\lambda\in\mathbb{R}^l$ the Lagrange multiplier and $\beta>0$ a penalty parameter. For the special case of (1.1) with $m=2$, the alternating direction method of multipliers (ADMM) in [14] reads as

$$\begin{cases} x_1^{k+1}=\arg\min\{\mathcal{L}_\beta^2(x_1,x_2^k,\lambda^k)\mid x_1\in X_1\},\\ x_2^{k+1}=\arg\min\{\mathcal{L}_\beta^2(x_1^{k+1},x_2,\lambda^k)\mid x_2\in X_2\},\\ \lambda^{k+1}=\lambda^k-\beta(A_1x_1^{k+1}+A_2x_2^{k+1}-b).\end{cases} \tag{1.3}$$

Recently, the ADMM has found many efficient applications for a broad spectrum of applications in various fields such as machine learning, statistical learning, computer vision, wireless network, and so on. We refer the reader to [1, 7, 12] for some recent review papers on the ADMM. For the multiple-block case of (1.1) with $m\geq 3$, the direct extension of ADMM reads as

$$\begin{cases} x_1^{k+1}=\arg\min\{\mathcal{L}_\beta^m(x_1,x_2^k,\cdots,x_m^k,\lambda^k)\mid x_1\in X_1\},\\ \cdots\cdots\\ x_i^{k+1}=\arg\min\{\mathcal{L}_\beta^m(x_1^{k+1},\cdots,x_{i-1}^{k+1},x_i,x_{i+1}^k,\cdots,x_m^k,\lambda^k)\mid x_i\in X_i\},\\ \cdots\cdots\\ x_m^{k+1}=\arg\min\{\mathcal{L}_\beta^m(x_1^{k+1},\cdots,x_{m-1}^{k+1},x_m,\lambda^k)\mid x_m\in X_m\},\\ \lambda^{k+1}=\lambda^k-\beta(\sum_{i=1}^m A_ix_i^{k+1}-b).\end{cases} \tag{1.4}$$

The direct extension of ADMM scheme (1.4) indeed works empirically for some applications, as shown in, e.g. [32, 34]. However, it was shown in [3] that the scheme (1.4) is not necessarily convergent. In the literature, some surrogates of (1.4) have been well studied. For example, the schemes in [22, 21] suggest correcting the output of (1.4) appropriately to ensure the convergence; the scheme in [23] slightly changes the order of updating the Lagrange multiplier and twists some of the subproblems appropriately; the scheme in [26] suggests attaching a shrinking factor to the Lagrange multiplier updating step in (1.4).

Given the wide applicability of ADMM (1.3) for a two-block convex minimization model and the convergence difficulty of the direct extension of ADMM (1.4) for a multiple-block counterpart, in this paper we mainly answer the question of how to use the original ADMM scheme (1.3) directly for the multiple-block model (1.1) and to design an implementable algorithm with provable convergence which can individually take advantage of the properties of the functions in the objective of (1.1). Let us first elaborate on the difficulty by applying the original ADMM (1.3) to the model (1.1) with $m\geq 3$. Conceptually, we can group the m functions in the objective of (1.1) and accordingly all the variables as two groups; to which the original ADMM scheme (1.3) becomes applicable in a block-wise form. That is, we can rewrite the model (1.1) as

$$\min\left\{\sum_{i=1}^{m_1}\theta_i(x_i)+\sum_{j=1}^{m_2}\phi_j(y_j)\,\Big|\,\sum_{i=1}^{m_1}A_ix_i+\sum_{j=1}^{m_2}B_jy_j=b,\ x_i\in X_i,y_j\in Y_j\right\}, \tag{1.5}$$

where $m_1\geq 1$, $m_2\geq 1$ and $m=m_1+m_2$. Note that for the second group, we relabel (θ_i, x_i, A_i, X_i) for $i=m_1+1,\cdots,m$ in (1.1) as (ϕ_j, y_j, B_j, Y_j) with $j=1,2,\cdots,m_2$ in (1.5), respectively. This relabeling will significantly simplify our natation of presentation in the coming analysis and help us expose our idea more clearly. Furthermore, with the notation

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{m_1} \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{m_2} \end{pmatrix}, \quad \mathcal{A} = (A_1, \ldots, A_{m_1}), \qquad \mathcal{B} = (B_1, \ldots, B_{m_2}), \tag{1.6}$$

and

$$\vartheta(\boldsymbol{x}) = \sum_{i=1}^{m_1} \theta_i(x_i), \qquad \varphi(\boldsymbol{y}) = \sum_{j=1}^{m_2} \phi_j(y_j), \qquad \mathcal{X} = \prod_{i=1}^{m_1} X_i, \qquad \mathcal{Y} = \prod_{j=1}^{m_2} Y_j, \tag{1.7}$$

The reformulation (1.5) of the model (1.1) can be written as the block-wise form

$$\min\left\{\vartheta(\boldsymbol{x}) + \varphi(\boldsymbol{y}) \,\middle|\, \mathcal{A}\boldsymbol{x} + \mathcal{B}\boldsymbol{y} = b,\ \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}\right\}. \tag{1.8}$$

The Lagrange function of (1.8) is

$$L^2(\boldsymbol{x}, \boldsymbol{y}, \lambda) = \vartheta(\boldsymbol{x}) + \varphi(\boldsymbol{y}) - \lambda^T(\mathcal{A}\boldsymbol{x} + \mathcal{B}\boldsymbol{y} - b), \tag{1.9}$$

which defined on

$$\Omega = \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^l = X_1 \times \cdots \times X_{m_1} \times Y_1 \times \cdots \times Y_{m_2} \times \mathbb{R}^l. \tag{1.10}$$

The augmented Lagrangian function can be denoted by

$$\begin{aligned} \mathcal{L}^2_\beta(\boldsymbol{x}, \boldsymbol{y}, \lambda) &= L^2(\boldsymbol{x}, \boldsymbol{y}, \lambda) + \frac{\beta}{2}\|\mathcal{A}\boldsymbol{x} + \mathcal{B}\boldsymbol{y} - b\|^2 \\ &= \sum_{i=1}^{m_1} \theta_i(x_i) + \sum_{j=1}^{m_2} \phi_j(x_j) - \lambda^T\Big(\sum_{i=1}^{m_1} A_i x_i + \sum_{j=1}^{m_2} B_j y_j - b\Big) + \frac{\beta}{2}\Big\|\sum_{i=1}^{m_1} A_i x_i + \sum_{j=1}^{m_2} B_j y_j - b\Big\|^2. \end{aligned} \tag{1.11}$$

Then, conceptually, the original ADMM (1.3) is implementable to the block-wise reformulation (1.8) of the model (1.1). The resulting scheme, called a block-wise ADMM hereafter, reads as

$$\begin{cases} \boldsymbol{x}^{k+1} = \arg\min\{\mathcal{L}^2_\beta(\boldsymbol{x}, \boldsymbol{y}^k, \lambda^k) \mid \boldsymbol{x} \in \mathcal{X}\}, \\ \boldsymbol{y}^{k+1} = \arg\min\{\mathcal{L}^2_\beta(\boldsymbol{x}^{k+1}, \boldsymbol{y}, \lambda^k) \mid \boldsymbol{y} \in \mathcal{Y}\}, \\ \lambda^{k+1} = \lambda^k - \beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^{k+1} - b). \end{cases} \tag{1.12}$$

For a generic case of (1.1) with $m \geq 3$, however, solving the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in (1.12) is usually hard or even not feasible for many concrete applications of the abstract model (1.1). This is mainly due that such a subproblem is in a block-wise form and it may require minimizing more than one functions with variables coupled by the quadratic term in (1.11). Recall that for the ADMM (1.3) and its variants, all the functions in the objective of (1.1) are treated individually in their decomposed subproblems and thus these functions' properties can be effectively exploited in the algorithmic implementation. One representative case arising often in sparse and/or low-rank optimization models is that when a function θ_i is simple in the sense that its proximal operator can be evaluated explicitly, then the corresponding subproblem or its linearized version is easy enough to have a closed-form solution —- meaning the corresponding subproblem is completely exempted from any inner iteration. This feature is indeed the main reason to account for the efficient applications of ADMM-related schemes in a broad spectrum of applications. Therefore, although the scheme (1.12) represents a direct application of the original ADMM to the block-wise reformulation (1.8) and it makes theoretical senses, it is necessary to discuss how to solve its $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems efficiently. With the just-mentioned desire of taking advantage of the properties of θ_i's individually, we can consider further decomposing the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in (1.12) into m_1 and m_2 smaller subproblems (implementing decomposition to the quadratic term in (1.11) which couples the variables), respectively; so that each of them only has to deal with one

$\theta_i(x_i)$ in its objective. That is, for solving the multiple-block model (1.1) with $m \geq 3$, we can consider the following splitting version of the block-wise ADMM (1.12):

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_1 \in X_1\}, \\ \quad\quad \vdots \\ x_i^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^k, , x_2^k, \cdots, x_{i-1}^k, x_i, x_{i+1}^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_i \in X_i\}, \\ \quad\quad \vdots \\ x_{m_1}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^k, x_2^k, \cdots, x_{m_1-1}^k, x_{m_1}, \boldsymbol{y}^k, \lambda^k) \mid x_{m_1} \in X_{m_1}\}, \\ y_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1, y_2^k, \cdots, y_{m_2}^k, \lambda^k) \mid y_1 \in Y_1\}, \\ \quad\quad \vdots \\ y_j^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1^k, y_2^k, \cdots, y_{j-1}^k, y_j, y_{j+1}^k, \cdots, y_{m_2}^k, \lambda^k) \mid y_j \in Y_j\}, \\ \quad\quad \vdots \\ y_{m_2}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1^k, y_2^k, \cdots, x_{m_2-1}^k, y_{m_2}, \lambda^k) \mid y_{m_2} \in Y_{m_2}\}, \\ \lambda^{k+1} = \lambda^k - \beta(A\boldsymbol{x}^{k+1} + B\boldsymbol{y}^{k+1} - b). \end{cases} \tag{1.13}$$

Note that in (1.13), internally we consider the parallel (i.e., Jacobian) decomposition for the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems arising in (1.12). See Remark 3.5 for the discussion where the alternating (i.e., Gauss-Seidel) decomposition is implemented to these subproblems internally. With this consideration of parallel decomposition, it is possible to solve the smaller subproblems respectively decomposed by the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems simultaneously on parallel computation infrastructures, which is particularly of interest for the scenarios where large-dimension data is considered and distributed computation is requested. Thus, the scheme (1.13) is a mixture of alternating decomposition outside (i.e, the block-wise variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are updated alternatingly) and parallel decomposition inside (i.e., the individual variables in both the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems are updated in parallel). It is clear that the scheme (1.13) enjoys the same good feature as the ADMM-related schemes, because each of its subproblems could be very easy if the functions θ_i's are easy enough (e.g., their proximal operators can be evaluated explicitly). Despite of this nice feature, the scheme (1.13), however, is not necessarily convergent — see a counter example in the appendix.

In fact, the divergence of (1.13) can be intuitively understood: The decomposed $(x_1, x_2, \cdots, x_{m_1})$-subproblems (Resp., $(y_1, y_2, \cdots, y_{m_2})$-subproblems) in (1.13) represent a more implementable but inexact version of the block-wise $\boldsymbol{x}$-subproblem (Resp., $\boldsymbol{y}$-subproblem) in (1.12). This approximation, however, is not accurate enough because the block-wise $\boldsymbol{x}$-subproblem (Resp., $\boldsymbol{y}$-subproblem) in (1.12) is decomposed by m_1 (Resp., m_2) times in (1.13). Therefore, the convergence of the block-wise ADMM (1.12), which is indeed guaranteed under the condition that both of its block-wise subproblems must be solved exactly or inexactly but with certain requirement on the inexactness (see e.g. [18, 30]), does not necessarily hold for its splitting version (1.13). The counter example in the appendix is indeed a case of the model (1.1) with $m = 3$. Thus, we have $m_1 = 1$ and $m_2 = 2$ when implementing the splitting version (1.13), meaning that the x_i-subproblem in (1.13) is the same as that in (1.12) while only the $\boldsymbol{y}$-subproblem in (1.12) is approximated by two further decomposed subproblems in (1.13). For this case with the least extent of approximation, the scheme (1.13) is already shown not to be necessarily convergent by this counter example. This fact clearly shows the failure of convergence of (1.13) for a generic case where both $m_1 \geq 2$ and $m_2 \geq 2$. How to derive a convergence-guaranteed algorithm based on the scheme (1.13) is thus the emphasis of this paper.

We will present an algorithm based on the scheme (1.13) in Section 3, preceding by some preliminaries summarized in Section 2 for further analysis. Then, in Section 4, we prove the convergence for the new scheme and establish its worst-case convergence rate measured by the iteration complexity in both the ergodic and a nonergodic senses. In Sections 5 and 6, we apply similar analysis to the generalized ADMM in [6] and the strictly contractive Peaceman-Rachford splitting method in [20], respectively, and obtain their block-wise versions which are suitable for the model (1.1). The convergence analysis, including the convergence rate analysis, for these three block-wise splitting methods can be presented in a unified framework. Finally, we make some conclusions in Section 7.

2. Preliminaries

In this section, we summarize some known results in the literature for further analysis.

2.1. A variational inequality characterization

We denote by $\boldsymbol{x}^* = (x_1^*, \ldots, x_{m_1}^*)$ and $\boldsymbol{y}^* = (y_1^*, \ldots, y_{m_2}^*)$. Let $(\boldsymbol{x}^*, \boldsymbol{y}^*, \lambda^*)$ be a saddle point of the Lagrange function (1.9). Then, for any $\lambda \in \mathbb{R}^l, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}$, we have

$$L^2(\boldsymbol{x}^*, \boldsymbol{y}^*, \lambda) \le L^2(\boldsymbol{x}^*, \boldsymbol{y}^*, \lambda^*) \le L^2(\boldsymbol{x}, \boldsymbol{y}, \lambda^*).$$

Indeed, finding a saddle point of $L^2(\boldsymbol{x}, \boldsymbol{y}, \lambda)$ can be expressed as the following variational inequalities: $(x_1^*, \ldots, x_{m_1}^*, y_1^*, \ldots, y_{m_2}^*, \lambda^*) \in \Omega$ where Ω is defined in (1.10), such that

$$\left\{\begin{array}{llcll}
x_1^* \in X_1, & \theta_1(x_1) - \theta_1(x_1^*) & + & (x_1 - x_1^*)^T(-A_1^T\lambda^*) \ge 0, & \forall x_1 \in X_1,\\
 & & \vdots & & \\
x_{m_1}^* \in X_{m_1}, & \theta_{m_1}(x_{m_1}) - \theta_{m_1}(x_{m_1}^*) & + & (x_{m_1} - x_{m_1}^*)^T(-A_{m_1}^T\lambda^*) \ge 0, & \forall x_m \in X_m,\\
y_1^* \in Y_1, & \phi_1(y_1) - \phi_1(y_1^*) & + & (y_1 - y_1^*)^T(-B_1^T\lambda^*) \ge 0, & \forall y_1 \in Y_1,\\
 & & \vdots & & \\
y_{m_2}^* \in Y_{m_2}, & \phi_{m_2}(y_{m_2}) - \phi_{m_2}(y_{m_2}^*) & + & (y_{m_2} - y_{m_2}^*)^T(-B_{m_2}^T\lambda^*) \ge 0, & \forall y_{m_2} \in Y_{m_2},\\
\lambda^* \in \mathbb{R}^l, & (\lambda - \lambda^*)^T(\sum_{i=1}^{m_1} A_i x_i^* + \sum_{j=1}^{m_2} B_j y_j^* - b) \ge 0, & & & \forall \lambda \in \mathbb{R}^l.
\end{array}\right. \tag{2.1}$$

More compactly, the variational inequalities in (2.1) can be rewritten in a compact form

$$\mathrm{VI}(\Omega, F, \theta) \quad \boldsymbol{w}^* \in \Omega, \ \ f(\boldsymbol{u}) - f(\boldsymbol{u}^*) + (\boldsymbol{w} - \boldsymbol{w}^*)^T F(\boldsymbol{w}^*) \ge 0, \ \ \forall\, \boldsymbol{w} \in \Omega, \tag{2.2a}$$

with the definitions:

$$f(\boldsymbol{u}) = \sum_{i=1}^{m_1} \theta_i(x_i) + \sum_{j=1}^{m_2} \phi_j(y_j), \tag{2.2b}$$

$$\boldsymbol{u} = \begin{pmatrix} x_1 \\ \vdots \\ x_{m_1} \\ y_1 \\ \vdots \\ y_{m_2} \end{pmatrix}, \quad \boldsymbol{w} = \begin{pmatrix} x_1 \\ \vdots \\ x_{m_1} \\ y_1 \\ \vdots \\ y_{m_2} \\ \lambda \end{pmatrix}, \quad F(\boldsymbol{w}) = \begin{pmatrix} -A_1^T\lambda \\ \vdots \\ -A_{m_1}^T\lambda \\ -B_1^T\lambda \\ \vdots \\ -B_{m_2}^T\lambda \\ \sum_{i=1}^{m_1} A_i x_i + \sum_{j=1}^{m_2} B_j y_j - b \end{pmatrix} \tag{2.2c}$$

and Ω is given in (1.10) (also can be expressed as $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^l$). We also denote by Ω^* the set of all saddle points of $\mathcal{L}_\beta^2(\boldsymbol{x}, \boldsymbol{y}, \lambda)$.

Recall the notation defined in (1.6)-(1.7). For the variational inequality (2.2). we further have

$$f(\boldsymbol{u}) = \vartheta(\boldsymbol{x}) + \varphi(\boldsymbol{y}), \tag{2.3a}$$

where

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

This variational inequality indeed characterizes the first-order optimality condition of the block-wise reformulation (1.5) of the model (1.1). We need this variational inequality characterization for the upcoming theoretical analysis.

2.2. Some properties of the matrices $\mathcal{A}$ and $\mathcal{B}$

For the matrices $\mathcal{A}$ and $\mathcal{B}$ defined in (1.6), we have some properties which are useful for our analysis. We summarize them in the following lemma; its proof is omitted as it is trivial.

Lemma 2.1. *For the matrices defined $\mathcal{A}$ and $\mathcal{B}$ defined in* (1.6)*, we have the following conclusions:*

$$m_1 \cdot diag(\mathcal{A}^T\mathcal{A}) \succeq \mathcal{A}^T\mathcal{A} \qquad and \qquad m_2 \cdot diag(\mathcal{B}^T\mathcal{B}) \succeq \mathcal{B}^T\mathcal{B}, \tag{2.4}$$

where $diag(\mathcal{A}^T\mathcal{A})$ and $diag(\mathcal{B}^T\mathcal{B})$ are defined by

$$diag(\mathcal{A}^T\mathcal{A}) := \begin{pmatrix} A_1^TA_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{m_1}^TA_{m_1} \end{pmatrix} \quad and \quad diag(\mathcal{B}^T\mathcal{B}) := \begin{pmatrix} B_1^TB_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_{m_2}^TB_{m_2} \end{pmatrix},$$

respectively.

Lemma 2.2. *Let $\tau_i > m_i - 1$ for $i = 1, 2$. Then we have*

$$\mathcal{D}_A := (\tau_1 + 1)\beta diag(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} \succ 0 \tag{2.5}$$

and

$$\mathcal{D}_B := (\tau_2 + 1)\beta diag(\mathcal{B}^T\mathcal{B}) - \beta\mathcal{B}^T\mathcal{B} \succ 0, \tag{2.6}$$

where $diag(\mathcal{A}^T\mathcal{A})$ and $diag(\mathcal{B}^T\mathcal{B})$ are defined in Lemma 2.1.

Proof. We need only to show (2.5). It follows from (2.4) that

$$\mathcal{D}_A = (\tau_1 + 1)\beta \mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} \succeq \beta(\tau_1 + 1 - m_1)\mathrm{diag}(\mathcal{A}^T\mathcal{A}).$$

The positive definiteness of $\mathcal{D}_A$ follows from $(\tau_1 + 1 - m_1) > 0$ and $\mathrm{diag}(\mathcal{A}^T\mathcal{A}) \succ 0$ directly. □

3. A splitting version of the block-wise ADMM (1.12) for (1.1)

In this section, we present a splitting version of the block-wise ADMM (1.12) for the model (1.1) and give some remarks.

Algorithm 1: A splitting version of the block-wise ADMM for (1.1)

Initialization: Specify a grouping strategy for (1.1) and determine the integers m_1 and m_2. Choose the constants $\tau_1 > m_1 - 1$; $\tau_2 > m_2 - 1$ and $\beta > 0$. For a given iterate $\boldsymbol{w}^k = (\boldsymbol{x}^k, \boldsymbol{y}^k, \lambda^k) = (x_1^k, \ldots, x_{m_1}^k, y_1^k, \ldots, y_{m_2}^k, \lambda^k)$, the new iterate $\boldsymbol{w}^{k+1}$ is generated by the following steps.

$$\begin{cases} x_i^{k+1} = \arg\min\limits_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big) + \frac{\tau_1\beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \; i = 1, \cdots, m_1, \\ y_j^{k+1} = \arg\min\limits_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big(\boldsymbol{x}^{k+1}, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^k\big) + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}, \; j = 1, \cdots, m_2, \\ \lambda^{k+1} = \lambda^k - \beta\big(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^{k+1} - b\big). \end{cases} \tag{3.1}$$

Remark 3.1. As analyzed in the introduction, the scheme (1.13) is not necessarily convergent because the decomposed subproblems therein might not be accurate enough to approximate the $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in the block-wise ADMM (1.12). Compared to (1.13), the splitting version (3.1) proximally regularizes the decomposed subproblems in (1.13) but the proximally regularized subproblems are of the same difficulty as their original ones in (1.13). In fact, these added proximal terms play the role of controlling the proximity of the solutions of the decomposed subproblems to the solutions of the block-wise subproblems in (1.12), and the extent of this control is determined by the proximal coefficients τ_1 and τ_2. This is the intrinsic mechanism why the convergence of the scheme (3.1) can be sufficiently guaranteed by adding some proximal terms in its subproblems. The requirement $\tau_i > m_i - 1$ for $i = 1, 2$ is the specific mathematical condition for how large the mentioned proximity should be controlled to sufficiently lead to convergence. More specifically, we require this condition to guarantee the positive definiteness of the matrix G defined in (4.13) and thus ensure the strict contraction for the sequence generated by the scheme (3.1) with respect to the solution set.

Remark 3.2. Algebraically, the scheme (3.1) can be derived from the proximal version of the block-wise ADMM (1.12) with appropriate choices of proximal terms, see, e.g., [18] for an abstract discussion in the variational inequality context. But it should be mentioned that it is significantly different from the so-called linearized ADMM in the literature (e.g., [8, 35, 36]), which aims at linearizing the quadratic terms in ADMM's subproblems with a sufficiently large proximal parameter (e.g., it should be greater than $\beta \cdot \|A_i^T A_i\|$ if the x_i-subproblem in (3.1) is considered) and thus alleviating the linearized subproblems. In other words, the proximal parameters in current linearized ADMM literature are dependent on the involved matrices of the corresponding quadratic terms and they may need to be sufficiently large to ensure the convergence if the matrices happen to be ill-conditioned, see, e.g., [16]. In (3.1), however, the proximal parameters τ_1 and τ_2 are just constants independent of the matrices. Indeed, they just rely on the user-assigned numbers of blocks for regrouping the model (1.1); and could be small. More importantly, the idea of using proximal terms to regularize the subproblems in the block-wise ADMM (1.12) and obtaining (3.1) is just for decomposing each block-wise ADMM subproblem into smaller and simpler ones when the block-wise reformulation (1.5) of (1.1) is considered. We do can further discuss how to linearize the subproblems in (3.1) and obtain even easier subproblems, as mentioned at the end in Section 7. But our emphasis is just the scheme (3.1) and we do not discuss more elaborated versions of it.

Remark 3.3. Notice that if we use $\boldsymbol{x}^k$, instead of $\boldsymbol{x}^{k+1}$, for the y_j-subproblems in (1.13), it leads to a full Jacobian decomposition scheme of the application of the augmented Lagrangian method to the model (1.1); and its divergence was shown in [17] even for the case of $m = 2$. For this case, adding appropriate proximal terms to regularize the resulting subproblems can also guarantee the convergence, see [17] for detailed analysis. The scheme (3.1) differs from the scheme in [17] in that the block-wise variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are updated alternatingly, i.e., it is $\boldsymbol{x}^{k+1}$, not $\boldsymbol{x}^k$ for updating the $\boldsymbol{y}_j$-subproblems in (3.1). This fact indeed represents a more accurate approximation to the augmented Lagrangian method when it is applied to the model (1.1). This difference also brings us a significant difference in the requirement on the proximal coefficients τ_1 and τ_2: The conditions $\tau_1 > m_1 - 1$ and $\tau_2 > m_2 - 1$ in (3.1) are much less restrictive than that in [17] which is required to be greater than $m - 1$. For example, if we consider a case of (1.1) with $m = 20$, then the scheme in [17] requires the proximal coefficient to be greater than 19. Meanwhile, if we choose $m_1 = m_2 = 10$ to implement the scheme (3.1), then τ_1 and τ_2 are only required to be greater than 9. This is a very important difference because a larger proximal coefficient makes the proximal term play a more important role in the objective function and thus the proximity is controlled in a more conservative way and it is more likely to result in slower convergence.

Remark 3.4. Note that we only discuss grouping the variables and functions of (1.1) "blindly" as (1.5). For a particular application of (1.1), based on some known information or features, we may group the functions and variables more smartly and even discuss an optimal grouping strategy; but we would emphasize that how to group the variables and functions better for a particular application really depends on the particular structures and features of a given application itself. In this paper, we just provide the methodology and theoretical analysis to guarantee the convergence for the most general setting in form of (1.5).

Remark 3.5. As mentioned, the splitting scheme (3.1) is obtained by implementing the parallel (i.e., Jacobian) decomposition to the block-wise $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in (1.13). If we consider the alternating (i.e., Gauss-Seidel) decomposition for the same subproblems in (1.13), then it is easy to see that the direct extension of ADMM (1.4) is recovered. Similar as what we will analyze, we can also consider how to ensure the convergence for the scheme (1.4) by adding proximal regularization terms to its subproblems. For succinctness, we omit the detail of analysis.

4. Convergence analysis

In this section, we analyze the convergence for the splitting version (3.1) of the block-wise ADMM (1.12). We will prove the global convergence, and establish the worst-case convergence rate measured by both the ergodic and a nonergodic senses. First of all, we rewrite the iterative scheme of (3.1) as a form that is more favorable for our analysis.

4.1. A reformulation of (3.1)

In our analysis, we need the auxiliary variables

$$\tilde{\boldsymbol{x}}^k = \boldsymbol{x}^{k+1}, \qquad \tilde{\boldsymbol{y}}^k = \boldsymbol{y}^{k+1} \tag{4.1}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b). \tag{4.2}$$

Recall the notation $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{y}, \lambda) = (x_1, \cdots, x_{m_1}, y_1, \cdots, y_{m_2}, \lambda)$ with superscripts k and $k+1$; and further define $\tilde{\boldsymbol{w}} = (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\lambda}) = (\tilde{x}_1, \cdots, \tilde{x}_{m_1}, \tilde{y}_1, \cdots, \tilde{y}_{m_2}, \tilde{\lambda})$ with superscripts. Then, we can artificially rewrite the iterative scheme of (3.1) as the following prediction-correction framework.

Prediction.

$$\tilde{x}_i^k = \arg\min_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big) + \frac{\tau_1\beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \tag{4.3a}$$

$$\tilde{y}_j^k = \arg\min_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big(\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^k\big) + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}, \tag{4.3b}$$

$$\tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b\big). \tag{4.3c}$$

Correction.

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{4.4a}$$

where

$$M = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -\beta\mathcal{B} & I \end{pmatrix}, \tag{4.4b}$$

and $\tilde{\boldsymbol{w}}^k$ is the related sub-vector of the predictor $\tilde{\boldsymbol{w}}^k$ generated by (4.3).

The iterative scheme (4.3)-(4.4) can also be understood as a prediction-correction method. We would emphasize that it is the scheme (3.1) that will be implemented in practice; and the reformulation (4.3)-(4.4) is considered in our analysis because of two reasons. First, our upcoming convergence analysis is essentially based on the effort of showing that the sequence generated by (3.1) is strictly contractive with respect to the solution set of (1.1) and it turns out that the progress of proximity to the solution set at each iteration can be measured by the quantity $\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2$ where G is defined in (4.13), see the inequality (4.20) in Theorem 4.4. Second, the analysis we will consider in Sections 5 and 6 for two other methods can also be written as prediction-correction frameworks; and using this kind of reformulations can make us present the analysis for these three methods in a unified framework.

4.2. **Global convergence**

Indeed, to prove the global convergence of the scheme (3.1), we mainly need to prove two conclusions. We summarize them in the following two theorems respectively.

Theorem 4.1. *Let $\tilde{\boldsymbol{w}}^k$ be generated by* (4.3) *from the given vector $\boldsymbol{w}^k$. Then we have*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \;\; f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \;\; \forall \boldsymbol{w} \in \Omega, \tag{4.5}$$

where Q is defined by

$$Q = \begin{pmatrix} (\tau_1+1)\beta\, diag(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta\, diag(\mathcal{B}^T\mathcal{B}) & 0 \\ 0 & -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix}. \tag{4.6}$$

Proof. First, the x_i-subproblem in (4.3a) can be written as

$$\begin{aligned}
\tilde{x}_i^k &= \arg\min_{x_i\in X_i} \Big\{ \mathcal{L}_\beta^2[x_1^k,\ldots,x_{i-1}^k,x_i,x_{i+1}^k,\ldots,x_{m_1}^k,\boldsymbol{y}^k,\lambda^k] + \frac{\tau_1\beta}{2}\|A_i(x_i-x_i^k)\|^2 \Big\} \\
&\overset{(1.11)}{=} \arg\min_{x_i\in X_i} \left\{ \begin{array}{c} \theta_i(x_i) - (\lambda^k)^T A_i x_i + \frac{\beta}{2}\|A_i(x_i - x_i^k) + (\mathcal{A}\boldsymbol{x}^k + \mathcal{B}\boldsymbol{y}^k - b)\|^2 \\ +\frac{\tau_1\beta}{2}\|A_i(x_i-x_i^k)\|^2, \end{array} \right\},
\end{aligned}$$

in which some constant terms are ignored in its objective function. The first-order optimality condition of the above convex minimization problem can be written as

$$\begin{aligned} \tilde{x}_i^k \in X_i, \quad & \theta_i(x_i) - \theta_i(\tilde{x}_i^k) + (x_i - \tilde{x}_i^k)^T\{-A_i^T\lambda^k \\ & +\beta A_i^T[A_i(\tilde{x}_i^k - x_i^k) + (\mathcal{A}\boldsymbol{x}^k + \mathcal{B}\boldsymbol{y}^k - b)] + \tau_1\beta A_i^T A_i(\tilde{x}_i^k - x_i^k)\} \geq 0, \; \forall x_i \in X_i. \end{aligned} \tag{4.7}$$

Then, it follows from (4.3c) that

$$\lambda^k = \tilde{\lambda}^k + \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b).$$

Hence, based on (4.7), we have

$$\begin{aligned} \tilde{x}_i^k \in X_i, \quad & \theta_i(x_i) - \theta_i(\tilde{x}_i^k) + (x_i - \tilde{x}_i^k)^T\{-A_i^T(\tilde{\lambda}^k + \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b)) \\ & +\beta A_i^T[A_i(\tilde{x}_i^k - x_i^k) + (\mathcal{A}\boldsymbol{x}^k + \mathcal{B}\boldsymbol{y}^k - b)] + \tau_1\beta A_i^T A_i(\tilde{x}_i^k - x_i^k)\} \geq 0, \; \forall x_i \in X_i. \end{aligned}$$

Consequently, we have

$$\begin{aligned} \tilde{x}_i^k \in X_i, \quad & \theta_i(x_i) - \theta_i(\tilde{x}_i^k) + (x_i - \tilde{x}_i^k)^T\{-A_i^T\tilde{\lambda}^k \\ & +\beta A_i^T[A_i(\tilde{x}_i^k - x_i^k) - \mathcal{A}(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k)] + \tau_1\beta A_i^T A_i(\tilde{x}_i^k - x_i^k)\} \geq 0, \quad \forall x_i \in X_i, \end{aligned}$$

and it can be written as

$$\begin{aligned} \tilde{x}_i^k \in X_i, \quad & \theta_i(x_i) - \theta_i(\tilde{x}_i^k) + (x_i - \tilde{x}_i^k)^T\{-A_i^T\tilde{\lambda}^k \\ & -\beta A_i^T\mathcal{A}(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k) + (\tau_1+1)\beta A_i^T A_i(\tilde{x}_i^k - x_i^k)\} \geq 0, \quad \forall x_i \in X_i. \end{aligned}$$

Taking $i = 1, \ldots, m_1$ in the above variational inequality and summarizing them, we obtain that $\tilde{\boldsymbol{x}}^k \in \mathcal{X}$ and

$$\vartheta(\boldsymbol{x})-\vartheta(\tilde{\boldsymbol{x}}^k)+(\boldsymbol{x}-\tilde{\boldsymbol{x}}^k)^T\{-\mathcal{A}^T\tilde{\lambda}^k-\beta\mathcal{A}^T\mathcal{A}(\tilde{\boldsymbol{x}}^k-\boldsymbol{x}^k)+(\tau_1+1)\beta\text{diag}(\mathcal{A}^T\mathcal{A})(\tilde{\boldsymbol{x}}^k-\boldsymbol{x}^k)\} \geq 0, \ \forall \boldsymbol{x} \in \mathcal{X}. \quad (4.8)$$

Then, we deal with the y_j-subproblems in (4.3b). Again, the y_j-th subproblem can be written as

$$\begin{aligned} \tilde{y}_j^k &= \arg\min_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2(\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^k) + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\} \\ &\overset{(1.11)}{=} \arg\min_{y_j \in Y_j} \left\{ \begin{array}{c} \phi_j(y_j) - (\lambda^k)^T B_j y_j + \frac{\beta}{2}\|\mathcal{A}\tilde{\boldsymbol{x}}^k + B_j(y_j - y_j^k) + \mathcal{B}\boldsymbol{y}^k - b\|^2 \\ +\frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2, \end{array} \right\}. \end{aligned}$$

where some constant terms are also ignored in its objective function. The first-order optimality condition of the above convex minimization problem is given by

$$\begin{array}{l} \tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T\lambda^k \\ \qquad +\beta B_j^T[\mathcal{A}\tilde{\boldsymbol{x}}^k + B_j(\tilde{y}_j^k - y_j^k) + \mathcal{B}\boldsymbol{y}^k - b] + \tau_2\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \geq 0, \ \forall y_j \in Y_j. \end{array} \quad (4.9)$$

Again, using (4.3c), we have $\lambda^k = \tilde{\lambda}^k + \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b)$. Substituting it into (4.9), we obtain

$$\begin{array}{l} \tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T(\tilde{\lambda}^k + \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b)) \\ \qquad +\beta B_j^T[\mathcal{A}\tilde{\boldsymbol{x}}^k + B_j(\tilde{y}_j^k - y_j^k) + \mathcal{B}\boldsymbol{y}^k - b] + \tau_2\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \geq 0, \ \forall y_j \in Y_j. \end{array}$$

Consequently, we have

$$\begin{array}{l} \tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k \\ \qquad +\beta B_j^T[B_j(\tilde{y}_j^k - y_j^k)] + \tau_2\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \geq 0, \quad \forall y_j \in Y_j, \end{array}$$

and it can be written as

$$\tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k + (\tau_2 + 1)\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \geq 0, \quad \forall y_j \in Y_j.$$

Taking $j = 1, \ldots, m_2$ in the above variational inequality and summarizing them, we get

$$\tilde{\boldsymbol{y}}^k \in \mathcal{Y}, \ \varphi(\boldsymbol{y}) - \varphi(\tilde{\boldsymbol{y}}^k) + (\boldsymbol{y} - \tilde{\boldsymbol{y}}^k)^T\{-\mathcal{B}^T\tilde{\lambda}^k + (\tau_2 + 1)\beta\text{diag}(\mathcal{B}^T\mathcal{B})(\tilde{\boldsymbol{y}}^k - \boldsymbol{y}^k)\} \geq 0, \ \forall \boldsymbol{y} \in \mathcal{Y}. \quad (4.10)$$

Using (4.3c) again, we have

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

and it can be rewritten as

$$\tilde{\lambda}^k \in \mathbb{R}^l, \quad (\lambda - \tilde{\lambda}^k)^T\{(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\tilde{\boldsymbol{y}}^k - b) - \mathcal{B}(\tilde{\boldsymbol{y}}^k - \boldsymbol{y}^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k)\} \geq 0, \quad \forall \lambda \in \mathbb{R}^l. \quad (4.11)$$

Combining (4.8), (4.10) and (4.11) together and using the notations $F(\boldsymbol{w})$ and Q (see (2.3) and (4.6)), the assertion of this theorem is followed directly. □

Theorem 4.2. *For given matrices Q in* (4.6) *and M in* (4.4b), *let*

$$H = QM^{-1} \quad (4.12)$$

and

$$G = Q^T + Q - M^T HM. \quad (4.13)$$

Then, both H and G are positive definite.

Proof. First, we show the positive definiteness of the matrix H. For the matrix M defined in (4.4b), we have

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

Thus, according to the definition of the matrix H (see (4.12)), we have

$$\begin{aligned} H &= QM^{-1} \\ &= \begin{pmatrix} (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) & 0 \\ 0 & -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & \beta\mathcal{B} & I \end{pmatrix} \\ &= \begin{pmatrix} (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix}. \end{aligned} \tag{4.14}$$

The positive definiteness of H follows from (2.5) and (2.6) directly. Now, we turn to prove that the matrix G is positive definite. First, we have the identity

$$\begin{aligned} G &= Q^T + Q - M^T HM = M^T H + HM - M^T HM \\ &= H - (M^T - I)H(M - I). \end{aligned} \tag{4.15}$$

Since H is a block diagonal matrix, we use the notation $H = \mathrm{diag}(H_{11}, H_{22}, \frac{1}{\beta}I)$ and obtain

$$\begin{aligned} (M^T - I)H(M - I) &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta\mathcal{B}^T \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} H_{11} & 0 & 0 \\ 0 & H_{22} & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -\beta\mathcal{B} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{4.16}$$

Using (4.14), (4.15) and (4.16), we get

$$G = \begin{pmatrix} (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix}.$$

According to (2.5) and (2.6), each block of G is positive definite. The proof is complete. □

Based on the conclusions in Theorems 4.1 and 4.2, it is easy to analyze the convergence for the scheme (3.1). In the following two theorems, we first prove the global convergence for (3.1).

Theorem 4.3. *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the scheme* (3.1). *Let $\{\tilde{\boldsymbol{w}}^k\}$ be defined in* (4.3)*; H and G be defined in* (4.12) *and* (4.13)*, respectively. Then we have*

$$\begin{aligned} & f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \\ \geq\ & \frac{1}{2}(\|\boldsymbol{w} - \boldsymbol{w}^{k+1}\|_H^2 - \|\boldsymbol{w} - \boldsymbol{w}^k\|_H^2) + \frac{1}{2}\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2, \ \ \forall \boldsymbol{w} \in \Omega. \end{aligned} \tag{4.17}$$

Proof. Using $Q = HM$ (see (4.12)) and the relation (4.4a), the right-hand side of (4.5) can be written as

$$(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}).$$

Hence, we have

$$f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \ge (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}),\ \forall \boldsymbol{w} \in \Omega. \tag{4.18}$$

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 of (4.18) with

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

we obtain

$$(\boldsymbol{w}-\tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k-\boldsymbol{w}^{k+1}) = \frac{1}{2}(\|\boldsymbol{w}-\boldsymbol{w}^{k+1}\|_H^2 - \|\boldsymbol{w}-\boldsymbol{w}^k\|_H^2) + \frac{1}{2}(\|\boldsymbol{w}^k-\tilde{\boldsymbol{w}}^k\|_H^2 - \|\boldsymbol{w}^{k+1}-\tilde{\boldsymbol{w}}^k\|_H^2). \tag{4.19}$$

For the last term of (4.19), we have

$$\begin{aligned}
&\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^k\|_H^2 \\
&\quad= \quad \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^k - \boldsymbol{w}^{k+1})\|_H^2 \\
&\quad\overset{(5.7a)}{=} \quad \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 \\
&\quad= \quad 2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T HM(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T M^T HM(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \\
&\quad= \quad (\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T (Q^T + Q - M^T HM)(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \\
&\quad\overset{(4.13)}{=} \quad \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2.
\end{aligned}$$

Substituting (4.19) and the last equation into (4.18), the assertion of this theorem is proved. □

Then, with the assertion (4.17) and positive definiteness of the matrix G, we can show that the sequence $\{\boldsymbol{w}^k\}$ generated by the scheme (3.1) is strictly contractive with respect to Ω^*. We summarize this result in the following theorem.

Theorem 4.4. *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the scheme* (3.1). *Let $\{\tilde{\boldsymbol{w}}^k\}$ be defined in* (4.3)*; H and G be defined in* (4.12) *and* (4.13)*, respectively. Then we have*

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \le \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2, \quad \forall \boldsymbol{w}^* \in \Omega^*. \tag{4.20}$$

Proof. Setting $w = w^*$ in (4.17), we get

$$\|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \ge \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2 + 2\{f(\tilde{\boldsymbol{u}}^k) - f(\boldsymbol{u}^*) + (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\tilde{\boldsymbol{w}}^k)\}.$$

By using the optimality of w^* and the monotonicity of $F(\boldsymbol{w})$, we have

$$f(\tilde{\boldsymbol{u}}^k) - f(\boldsymbol{u}^*) + (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\tilde{\boldsymbol{w}}^k) \ge f(\tilde{\boldsymbol{u}}^k) - f(\boldsymbol{u}^*) + (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\boldsymbol{w}^*) \ge 0$$

and thus

$$\|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \ge \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2.$$

The assertion (4.20) follows directly. □

The global convergence of the scheme (3.1) can be easily proved based on the assertion in Theorem 4.4. We state it in the following theorem.

Theorem 4.5. *The sequence $\{\boldsymbol{w}^k\}$ generated by the splitting version of the block-wise ADMM* (3.1) *converges to a saddle point in Ω^*.*

Proof First, according to (4.20), it holds that $\{\boldsymbol{w}^k\}$ is bounded and

$$\lim_{k\to\infty} \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G = 0. \tag{4.21}$$

Thus, $\{\boldsymbol{w}^k\}$ (and $\{\tilde{\boldsymbol{w}}^k\}$) has a cluster point $\boldsymbol{w}^\infty$. Then, it follows from (4.5) that

$$\tilde{\boldsymbol{w}}^\infty \in \Omega, \quad f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^\infty) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^\infty)^T F(\tilde{\boldsymbol{w}}^\infty) \ge 0, \quad \forall \boldsymbol{w} \in \Omega,$$

and thus $\tilde{\boldsymbol{w}}^\infty$ is a solution point of VI(Ω, F, θ) in Ω^*. The proof is complete. □

4.3. Convergence rate in the ergodic sense

In this subsection, we establish a worst-case convergence rate measured by the iteration complexity in the ergodic sense for the scheme (3.1). Note that the same convergence rate was established in [24] for the original ADMM (1.3) and its linearized version. Here, we show that the scheme (3.1) enjoys the same worst-case convergence rate as the original ADMM even though it is a splitting version with m decomposed subproblems.

For this convergence rate analysis, we need to recall a characterization of the solution set Ω^* of VI (2.2), which is described in the following theorem. Its proof can be found in [9] (Theorem 2.3.5) or [24] (Theorem 2.1).

Theorem 4.6. *The solution set of VI(Ω, F, θ) is convex and it can be characterized as*

$$\Omega^* = \bigcap_{\boldsymbol{w}\in\Omega} \{\tilde{\boldsymbol{w}} \in \Omega : (f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}})) + (\boldsymbol{w} - \tilde{\boldsymbol{w}})^T F(\boldsymbol{w}) \ge 0\}. \tag{4.22}$$

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

$$f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}) + (\boldsymbol{w} - \tilde{\boldsymbol{w}})^T F(\boldsymbol{w}) \ge -\epsilon, \ \ \forall\, \boldsymbol{w} \in \mathcal{D}_{(\tilde{\boldsymbol{w}})},$$

where $\mathcal{D}_{(\tilde{\boldsymbol{w}})} = \{\boldsymbol{w} \in \Omega \,|\, \|\boldsymbol{w} - \tilde{\boldsymbol{w}}\| \le 1\}$[1].

In the following, we show that for a given $\epsilon > 0$, based on t iterations of the scheme (3.1), we can find $\tilde{\boldsymbol{w}} \in \mathcal{W}$ such that $\tilde{\boldsymbol{w}} \in \mathcal{W}$ and

$$\sup_{\boldsymbol{w}\in\mathcal{D}_{(\tilde{\boldsymbol{w}})}} \{f(\tilde{\boldsymbol{u}}) - f(\boldsymbol{u}) + (\tilde{\boldsymbol{w}} - \boldsymbol{w})^T F(\boldsymbol{w})\} \le \epsilon. \tag{4.23}$$

This means a worst-case $O(1/t)$ convergence rate is established for the scheme (3.1).

Prior to the proof, we emphasize that the conclusion in Theorem 4.3 is also very useful for establishing the worst-case $O(1/t)$ convergence rate for the scheme (3.1). In fact, using the monotonicity of F, we have

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

Substituting it into (4.17), we obtain

$$f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\boldsymbol{w}) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^k\|_H^2 \ge \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^{k+1}\|_H^2, \ \forall \boldsymbol{w} \in \Omega. \tag{4.24}$$

Note that the above assertion is hold for $G \succeq 0$.

Now, we establish the worst-case $O(1/t)$ convergence rate for the scheme (3.1) in the following theorem.

Theorem 4.7. *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the scheme* (3.1)*. Let $\{\tilde{\boldsymbol{w}}^k\}$ be defined in* (4.3) *and H be defined in* (4.12)*, respectively. For any integer $t > 0$, let*

$$\tilde{\boldsymbol{w}}_t = \frac{1}{t+1}\sum_{k=0}^{t} \tilde{\boldsymbol{w}}^k. \tag{4.25}$$

[1]We refer the reader to [29] (see (2.5) therein) for the definition of an ϵ-approximate solution

Then, we have $\tilde{\boldsymbol{w}}_t \in \Omega$ *and*

$$f(\tilde{\boldsymbol{u}}_t) - f(\boldsymbol{u}) + (\tilde{\boldsymbol{w}}_t - \boldsymbol{w})^T F(\boldsymbol{w}) \le \frac{1}{2(t+1)} \|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2, \quad \forall w \in \Omega. \tag{4.26}$$

Proof. First, it holds that $\tilde{\boldsymbol{w}}^k \in \Omega$ for all $k \ge 0$. Together with the convexity of X and $\mathbb{R}^l$, (4.25) implies that $\tilde{\boldsymbol{w}}_t \in \Omega$. Summarizing this inequality (4.24) over $k = 0, 1, \ldots, t$, we obtain

$$(t+1) f(\boldsymbol{u}) - \sum_{k=0}^{t} f(\tilde{\boldsymbol{u}}^k) + \Big((t+1)\boldsymbol{w} - \sum_{k=0}^{t} \tilde{\boldsymbol{w}}^k \Big)^T F(\boldsymbol{w}) + \frac{1}{2} \|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2 \ge 0, \quad \forall \boldsymbol{w} \in \Omega.$$

Using the notation $\tilde{\boldsymbol{w}}_t$, it can be written as

$$\frac{1}{t+1} \sum_{k=0}^{t} f(\tilde{\boldsymbol{u}}^k) - f(\boldsymbol{u}) + (\tilde{\boldsymbol{w}}_t - \boldsymbol{w})^T F(\boldsymbol{w}) \le \frac{1}{2(t+1)} \|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2, \quad \forall \boldsymbol{w} \in \Omega. \tag{4.27}$$

Since $f(\boldsymbol{u})$ is convex and

$$\tilde{\boldsymbol{u}}_t = \frac{1}{t+1} \sum_{k=0}^{t} \tilde{\boldsymbol{u}}^k,$$

we have that

$$f(\tilde{\boldsymbol{u}}_t) \le \frac{1}{t+1} \sum_{k=0}^{t} f(\tilde{\boldsymbol{u}}^k).$$

Substituting it into (4.27), the assertion of this theorem follows directly. □

Recall the definition (4.23). Theorem 4.7 thus indicates that the average of the first t iterates generated by the scheme (3.1) is an approximate solution of $\mathrm{VI}(\Omega, F, \theta)$ with an accuracy of $O(1/t)$. This clearly means a worst-case $O(1/t)$ convergence rate measured by the iteration complexity in the ergodic sense for the scheme (3.1).

4.4. Convergence rate in a nonergodic Sense

In this subsection, we establish a worst-case $O(1/\sqrt{t})$ convergence rate in a nonergodic sense for the scheme (3.1). Recall that the same convergence rate was established in [25] for the original ADMM (1.3) and its linearized version; we refer the reader to [33] for a more general study on the convergence rate of decomposition methods based on the proximal method of multipliers. Moreover, in general a worst-case nonergodic convergence rate measured by the iteration complexity is stronger than its ergodic counterpart.

We first need to prove the following lemma.

Lemma 4.8. *Let* $\{\boldsymbol{w}^k\}$ *be the sequence generated by the scheme* (3.1)*. Let* $\{\tilde{\boldsymbol{w}}^k\}$ *be defined in* (4.3)*; M, H and Q be defined in* (4.4b), (4.12) *and* (4.6)*, respectively. Then we have*

$$(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T M^T H M \{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\} \ge \frac{1}{2} \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_{(Q^T+Q)}^2. \tag{4.28}$$

Proof. First, setting $\boldsymbol{w} = \tilde{\boldsymbol{w}}^{k+1}$ in (4.5) gives us

$$f(\tilde{\boldsymbol{u}}^{k+1}) - f(\tilde{\boldsymbol{u}}^k) + (\tilde{\boldsymbol{w}}^{k+1} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \ge (\tilde{\boldsymbol{w}}^{k+1} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k). \tag{4.29}$$

Note that (4.5) is also true for $k := k+1$. Thus, we have

$$f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^{k+1}) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^{k+1})^T F(\tilde{\boldsymbol{w}}^{k+1}) \ge (\boldsymbol{w} - \tilde{\boldsymbol{w}}^{k+1})^T Q(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1}), \forall\, \boldsymbol{w} \in \Omega.$$

Setting $\boldsymbol{w} = \tilde{\boldsymbol{w}}^k$ in the above inequality, we obtain

$$f(\tilde{\boldsymbol{u}}^k) - f(\tilde{\boldsymbol{u}}^{k+1}) + (\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1})^T F(\tilde{\boldsymbol{w}}^{k+1}) \ge (\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1}) Q(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1}). \tag{4.30}$$

Combining (4.29) and (4.30) and using the monotonicity of F, we get

$$(\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1})^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\} \geq 0. \tag{4.31}$$

Adding the term

$$\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\}^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\}$$

to the both sides of (4.31), and using $\boldsymbol{w}^T Q\boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T(Q^T + Q)\boldsymbol{w}$, we obtain

$$(\boldsymbol{w}^k - \boldsymbol{w}^{k+1})^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\} \geq \frac{1}{2}\|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_{(Q^T+Q)}.$$

Substituting $(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) = M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ into the left-hand side of the last inequality and using $Q = HM$, we obtain (4.28) and the lemma is proved. □

Now, we are ready to prove an important inequality which will play a crucial role in the coming analysis for the worst-case convergence rate in a nonergodic sense. We summarize it in the following theorem.

Theorem 4.9. *Let* $\{\boldsymbol{w}^k\}$ *be the sequence generated by the scheme* (3.1)*. Let* $\{\tilde{\boldsymbol{w}}^k\}$ *be defined in* (4.3)*; M and H be defined in* (4.4b) *and* (4.12)*, respectively. Then we have*

$$\|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_H \leq \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H, \quad \forall\, k > 0,$$

and consequently (4.4a)

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^{k+2}\|_H \leq \|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H, \quad \forall\, k > 0. \tag{4.32}$$

Proof. Setting $a = M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ and $b = M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})$ in the identity

$$\|a\|^2_H - \|b\|^2_H = 2a^T H(a-b) - \|a-b\|^2_H,$$

we obtain

$$\begin{aligned}&\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|^2_H - \|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_H \\ &\quad = \; 2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T M^T HM[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})] - \|M[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})]\|^2_H.\end{aligned}$$

Inserting (4.28) into the first term of the right-hand side of the last equality, we obtain

$$\begin{aligned}&\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|^2_H - \|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_H \\ &\quad \geq \quad \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_{(Q^T+Q)} - \|M[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})]\|^2_H \\ &\quad \overset{(4.13)}{=} \quad \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_G \geq 0,\end{aligned}$$

where the last inequality is because of the positive definiteness of the matrix $(Q^T + Q) - M^T HM \succeq 0$. The assertion (4.32) follows from the last inequality and (4.4a) immediately. □

Note that it follows from $G \succ 0$ and Theorem 4.4 that there exists a constant $c_0 > 0$ such that

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2_H \leq \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2_H - c_0\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|^2_H, \quad \forall \boldsymbol{w}^* \in \Omega^*,$$

and consequently due to $M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) = (\boldsymbol{w}^k - \boldsymbol{w}^{k+1})$,

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|^2_H \leq \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2_H - c_0\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|^2_H, \quad \forall \boldsymbol{w}^* \in \Omega^*. \tag{4.33}$$

Now, with (4.33) and (4.32), we are ready to establish a worst-case $O(1/t)$ convergence rate in a nonergodic sense for the scheme (3.1).

Theorem 4.10. *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the scheme* (3.1) *and H be defined in* (4.12)*. The, for any integer $t > 0$, we have*

$$\|\boldsymbol{w}^t - \boldsymbol{w}^{t+1}\|_H^2 \le \frac{1}{(t+1)c_0}\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H^2, \quad \forall \boldsymbol{w}^* \in \Omega^*, \tag{4.34}$$

with $c_0 > 0$.

Proof. First, it follows from (4.33) that

$$\sum_{k=0}^{\infty} c_0\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 \le \|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H^2, \quad \forall\, \boldsymbol{w}^* \in \Omega^*. \tag{4.35}$$

According to Theorem 4.9, the sequence $\{\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2\}$ is monotonically non-increasing. Therefore, we have

$$(t+1)\|\boldsymbol{w}^t - \boldsymbol{w}^{t+1}\|_H^2 \le \sum_{k=0}^{t} \|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2. \tag{4.36}$$

The assertion (4.34) follows from (4.35) and (4.36) immediately. □

Let $d := \inf\{\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H \,|\, \boldsymbol{w}^* \in \Omega^*\}$. Then, for any given $\epsilon > 0$, Theorem 4.10 shows that the scheme (3.1) needs at most $\lfloor d^2/c_0\epsilon \rfloor$ iterations to ensure that $\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 \le \epsilon$. Recall (4.18). It indicates that $\boldsymbol{w}^k$ is a solution point of $\mathrm{VI}(\Omega, F, \theta)$ if $\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 = 0$. A worst-case $O(1/\sqrt{t})$ convergence rate in a nonergodic sense is thus established for the scheme (3.1).

5. A splitting version of the block-wise generalized ADMM

In this section, we consider the generalized ADMM in [6] and propose a splitting version of the block-wise generalized ADMM for solving the model (1.1). The convergence analysis similar as those in Section 4 will be conducted.

5.1. Algorithm

In [10], it was shown that the ADMM scheme (1.3) is an application of the DRSM in [5, 27] to the dual of (1.1) with $m = 2$; and it was further shown in [6] that the DRSM is an application of the PPA in [28]. These two conclusions mean that the ADMM scheme (1.3) is a special case of the PPA. Thus, it was suggested in [6] to apply the relaxed PPA in [15] and accordingly the generalized PPA was proposed.

Based on our analysis in Section 3, it is easy to propose the block-wise version of the generalized ADMM for the grouped model (1.5). For this purpose, we first remark that we can rewrite the ADMM scheme (1.3) as

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \lambda^k) \mid x_1 \in X_1\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^k - b), \\ x_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2, \lambda^{k+1}) \mid x_2 \in X_2\}, \end{cases} \tag{5.1}$$

which is equivalent to (1.3) in cyclical sense. Based on this representation, it was explained in [2] that the ADMM scheme (5.1) is indeed an application of the PPA in [28] with a customized proximal coefficient in metric form to the model (1.1) with $m = 2$. Hence, applying the acceleration scheme in [15] to the specific PPA (5.1), the generalized ADMM in [6] can be expressed as

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \lambda^k) \mid x_1 \in X_1\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^k - b), \\ x_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2, \lambda^{k+1}) \mid x_2 \in X_2\}, \\ w^{k+1} := w^k - \alpha(w^k - w^{k+1}), \end{cases} \tag{5.2}$$

with $\alpha \in (0, 2)$. Recall the block-wise ADMM (1.12). Then, the block-wise generalized ADMM when the scheme (5.2) is applied to the the grouped two-block model (1.5) can be written as

$$\begin{cases} \tilde{\boldsymbol{x}}^k = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}, \boldsymbol{y}^k, \lambda^k) \mid \boldsymbol{x} \in \mathcal{X}\}, \\ \tilde{\lambda}^k = \lambda^k - \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b) \\ \tilde{\boldsymbol{y}}^k = \arg\min\{\mathcal{L}_\beta^2(\tilde{\boldsymbol{x}}^k, \boldsymbol{y}, \tilde{\lambda}^k) \mid \boldsymbol{y} \in \mathcal{Y}\}, \\ \boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \end{cases} \tag{5.3}$$

with $\alpha \in (0, 2)$. Similar as (1.13), we can consider further decomposing the block-wise $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in (5.3) and obtain a splitting version of the generalized ADMM (5.3) as the following:

$$\begin{cases} \tilde{x}_1^k = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_1 \in X_1\}, \\ \tilde{x}_2^k = \arg\min\{\mathcal{L}_\beta^2(x_1^k, x_2, x_3^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_2 \in X_2\}, \\ \qquad \vdots \\ \tilde{x}_{m_1}^k = \arg\min\{\mathcal{L}_\beta^2(x_1^k, x_2^k, \cdots, x_{m_1-1}^k, x_{m_1}, \boldsymbol{y}^k, \lambda^k) \mid x_{m_1} \in X_{m_1}\}, \\ \tilde{\lambda}^k = \lambda^k - \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b), \\ \tilde{y}_1^k = \arg\min\{\mathcal{L}_\beta^2(\tilde{\boldsymbol{x}}^k, y_1, y_2^k, \cdots, y_{m_2}^k, \tilde{\lambda}^k) \mid y_1 \in Y_1\}, \\ \tilde{y}_2^k = \arg\min\{\mathcal{L}_\beta^2(\tilde{\boldsymbol{x}}^k, y_1^k, y_2, y_3^k, \cdots, y_{m_2}^k, \tilde{\lambda}^k) \mid y_2 \in Y_2\}, \\ \qquad \vdots \\ \tilde{y}_{m_2}^k = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^k, y_1^k, y_2^k, \cdots, x_{m_2-1}^k, y_{m_2}, \tilde{\lambda}^k) \mid y_{m_2} \in Y_{m_2}\}, \\ \boldsymbol{w}^{k+1} := \boldsymbol{w}^k - \alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k). \end{cases} \tag{5.4}$$

Recall that it has been mentioned that the splitting version of the block-wise ADMM scheme (1.13), which is a special case of (5.4) with $\alpha = 1$, is not necessarily convergent. Therefore, we have to investigate how to ensure the convergence for (5.4). It turns out that its convergence can also be guaranteed if the subproblems in (5.4) are proximally regularized, just like the splitting version of the block-wise ADMM (3.1). We summarize the splitting version of the block-wise generalized ADMM in the following.

Algorithm 2: A splitting version of the block-wise generalized ADMM for (1.1)

Initialization: Specify a grouping strategy for (1.1) and determine the integers m_1 and m_2. Choose the constants $\tau_1 > m_1 - 1$; $\tau_2 > m_2 - 1$, $\alpha \in (0,2)$ and $\beta > 0$. For a given iterate $\boldsymbol{w}^k = (x_1^k, \ldots, x_{m_1}^k, y_1^k, \ldots, y_{m_2}^k, \lambda^k) = (\boldsymbol{x}^k, \boldsymbol{y}^k, \lambda^k)$, the new iterate $\boldsymbol{w}^{k+1}$ is generated by the following steps.

$$\begin{cases} \tilde{x}_i^k = \arg\min\limits_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big) + \frac{\tau_1 \beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \ i = 1, \cdots, m_1, \\ \tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b\big), \\ \tilde{y}_j^k = \arg\min\limits_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big(\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \tilde{\lambda}^k\big) + \frac{\tau_2 \beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}, \ j = 1, \cdots, m_2, \\ \boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k). \end{cases} \tag{5.5}$$

5.2. Convergence analysis

Similarly, we can rewrite the splitting version (5.4) of the block-wise generalized ADMM as a prediction-correction form and conduct the convergence analysis analogously as that in Section 3.

Prediction.

$$\tilde{x}_i^k = \arg\min_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big) + \frac{\tau_1 \beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \tag{5.6a}$$

$$\tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b\big), \tag{5.6b}$$

$$\tilde{y}_j^k = \arg\min_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big(\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \tilde{\lambda}^k\big) + \frac{\tau_2 \beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}. \tag{5.6c}$$

Correction.

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - M_1(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{5.7a}$$

where

$$M_1 = \alpha I, \quad \alpha \in (0,2) \tag{5.7b}$$

and $\tilde{\boldsymbol{w}}^k$ is the related sub-vector of the predictor $\tilde{\boldsymbol{w}}^k$ generated by (5.6).

With the prediction-correction reformulation (5.6)-(5.7), we can establish the same convergence results as those in Section 3. As mentioned, Theorems 4.1 and 4.2 are the basis for the convergence analysis. In the following, we prove some conclusions similar as the scheme (5.5).

Theorem 5.1. *Let $\tilde{\boldsymbol{w}}^k$ be generated by* (5.6) *from a given $\boldsymbol{w}^k$. Then we have*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \ \ f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q_1(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\} \geq 0, \ \ \forall \boldsymbol{w} \in \Omega, \tag{5.8}$$

where

$$Q_1 = \begin{pmatrix} (\tau_1+1)\beta diag(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta diag(\mathcal{B}^T\mathcal{B}) & -\mathcal{B}^T \\ 0 & -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \tag{5.9}$$

is a symmetric matrix.

Proof. Note that the x_i-subproblems in (5.6a) and the $\tilde{\lambda}^k$ update in (5.6b) are the same as those in (4.3a) and (4.3c), respectively. Then, as the proof of Theorem 4.1, see (4.8)), we have $\tilde{\boldsymbol{x}}^k \in \mathcal{X}$ and

$$\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{x} - \tilde{\boldsymbol{x}}^k)^T\{-\mathcal{A}^T\tilde{\lambda}^k - \beta\mathcal{A}^T\mathcal{A}(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k) + (\tau_1+1)\beta\text{diag}(\mathcal{A}^T\mathcal{A})(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k)\} \geq 0, \ \forall \boldsymbol{x} \in \mathcal{X}. \tag{5.10}$$

For the y_j-subproblems in (5.6c), it follows that

$$\begin{aligned}\tilde{y}_j^k &= \arg\min_{y_j\in Y_j}\Big\{\mathcal{L}_\beta^2[\tilde{\boldsymbol{x}}^k, y_1^k,\ldots,y_{j-1}^k, y_j, y_{j+1}^k,\ldots,y_{m_2}^k,\tilde{\lambda}^k] + \frac{\tau_1\beta}{2}\|B_j(y_j-y_j^k)\|^2\Big\}\\ &\overset{(1.11)}{=} \arg\min_{y_j\in Y_j}\left\{\begin{array}{c}\phi_j(y_j)-(\tilde{\lambda}^k)^TB_jy_j+\frac{\beta}{2}\|\mathcal{A}\tilde{\boldsymbol{x}}^k+B_j(y_j-y_j^k)+\mathcal{B}\boldsymbol{y}^k-b\|^2\\ +\frac{\tau_2\beta}{2}\|B_j(y_j-y_j^k)\|^2\end{array}\right\},\end{aligned}$$

where some constant terms are ignored in its objective function. The first-order optimality condition of the above convex minimization problem is

$$\begin{aligned}\tilde{y}_j^k\in Y_j,\quad &\phi_j(y_j)-\phi_j(\tilde{y}_j^k)+(y_j-\tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k\\ &+\beta B_j^T[\mathcal{A}\tilde{\boldsymbol{x}}^k+B_j(\tilde{y}_j^k-y_j^k)+\mathcal{B}\boldsymbol{y}^k-b]+\tau_2\beta B_j^TB_j(\tilde{y}_j^k-y_j^k)\}\ge 0,\ \forall y_j\in Y_j.\end{aligned}$$

Again, using $(\mathcal{A}\tilde{\boldsymbol{x}}^k+\mathcal{B}\boldsymbol{y}^k-b)=\frac{1}{\beta}(\lambda^k-\tilde{\lambda}^k)$, it follows from the last inequality that

$$\begin{aligned}\tilde{y}_j^k\in Y_j,\quad &\phi_j(y_j)-\phi_j(\tilde{y}_j^k)+(y_j-\tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k\\ &+\beta B_j^T[B_j(\tilde{y}_j^k-y_j^k)+\tfrac{1}{\beta}(\lambda^k-\tilde{\lambda}^k)]+\tau_2\beta B_j^TB_j(\tilde{y}_j^k-y_j^k)\}\ge 0,\quad \forall y_j\in Y_j,\end{aligned}\tag{5.11}$$

and it can be written as $\tilde{y}_j^k\in Y_j$ and

$$\phi_j(y_j)-\phi_j(\tilde{y}_j^k)+(y_j-\tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k+(\tau_2+1)\beta B_j^TB_j(\tilde{y}_j^k-y_j^k)-B_j^T(\tilde{\lambda}^k-\lambda^k)\}\ge 0,\ \forall y_j\in Y_j.$$

Taking $j=1,\ldots,m_2$ in the above variational inequality and summarizing them, we have $\tilde{\boldsymbol{y}}^k\in\mathcal{Y}$ and

$$\varphi(\boldsymbol{y})-\varphi(\tilde{\boldsymbol{y}}^k)+(\boldsymbol{y}-\tilde{\boldsymbol{y}}^k)^T\{-\mathcal{B}^T\tilde{\lambda}^k+(\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B})(\tilde{\boldsymbol{y}}^k-\boldsymbol{y}^k)-\mathcal{B}^T(\tilde{\lambda}^k-\lambda^k)\}\ge 0,\ \forall\boldsymbol{y}\in\mathcal{Y}.\tag{5.12}$$

It follows from (5.6b) that

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

which can be rewritten as

$$\tilde{\lambda}^k\in\mathbb{R}^l,\ (\lambda-\tilde{\lambda}^k)^T\{(\mathcal{A}\tilde{\boldsymbol{x}}^k+\mathcal{B}\tilde{\boldsymbol{y}}^k-b)-\mathcal{B}(\tilde{\boldsymbol{y}}^k-\boldsymbol{y}^k)+\frac{1}{\beta}(\tilde{\lambda}^k-\lambda^k)\}\ge 0,\ \forall\lambda\in\mathbb{R}^l.\tag{5.13}$$

Combining (5.10), (5.12) and (5.13), and using the notations $F(\boldsymbol{w})$ and Q_1 (see (2.3b) and (5.9)), the assertion of this theorem is followed directly. □

Theorem 5.2. *Let the matrices M_1 and Q_1 be defined in* (5.7b) *and* (5.9)*. Then, we have*

(1). *The matrix Q_1 is positive definite.*

(2). *Both the matrices defined below*

$$H_1 := Q_1M_1^{-1}\qquad and\qquad G_1 := Q_1^T+Q_1-M_1^TH_1M_1$$

are positive definite.

Proof. It is clear that Q_1 is symmetric. The first diagonal block of Q_1 equals $\mathcal{D}_A$ (see (2.5)) and thus is positive definite. Now, we observe the rest diagonal part of Q_1. Again, using (2.4) and $\tau_2>m_2-1$, we obtain

$$\begin{pmatrix} (\tau_2+1)\beta\text{diag}(\mathcal{B}^T\mathcal{B}) & -\mathcal{B}^T \\ -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix}$$
$$= \begin{pmatrix} (\tau_2+1-m_2)\beta\text{diag}(\mathcal{B}^T\mathcal{B}) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} m_2\beta\text{diag}(\mathcal{B}^T\mathcal{B}) & -\mathcal{B}^T \\ -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix}$$
$$\succeq \begin{pmatrix} (\tau_2+1-m_2)\beta\text{diag}(\mathcal{B}^T\mathcal{B}) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \beta\mathcal{B}^T\mathcal{B} & -\mathcal{B}^T \\ -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \succ 0.$$

Thus, the matrix Q_1 is positive definite.

Because $M_1 = \alpha I$ and $\alpha \in (0, 2)$ (see (5.7b)), we have

$$H_1 = Q_1 M_1^{-1} = \frac{1}{\alpha} Q_1$$

and

$$G_1 = Q_1^T + Q_1 - M_1^T H_1 M_1 = (2-\alpha)Q_1.$$

Thus, both the matrices H_1 and G_1 are positive definite. The proof is complete. □

Then, with the conclusions in Theorems 5.1 and 5.2, the proofs for the global convergence and the worst-case convergence rates in both the ergodic and nonergodic senses for the scheme (5.5) are identical as those in Sections 4.2-4.4. We thus omit them. We would reiterate again that this is the reason we artificially rewrite the scheme (3.1) as the prediction-correction form (4.3)-(4.4) in the analysis, because this prediction-correction form turns out to be a unified framework for analyzing the convergence for all the three methods discussed in this paper and using this framework helps us present the convergence results for the schemes (5.5) and (6.5) in the next section in succinctness.

6. A splitting version of the block-wise strictly contractive Peaceman-Rachford splitting method

As mentioned, the ADMM (1.3) is an application of the DRSM in [5, 27] to the dual of the model (1.1) with $m = 2$. Since the Peaceman-Rachford splitting method (PRSM) in [31, 27] is as popular as the DRSM in the PDE literature, it is interesting to apply the PRSM to the dual of (1.1) with $m = 2$. The resulting scheme is

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \lambda^k) \mid x_1 \in X_1\}, \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \beta(A_1x_1^{k+1} + A_2x_2^k - b), \\ x_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2, \lambda^{k+\frac{1}{2}}) \mid x_2 \in X_2\}, \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b). \end{cases} \tag{6.1}$$

which differs from the ADMM scheme (1.3) in that the Lagrange multiplier is updated once the x_1-subproblem is solved and the updated multiplier is used in the x_2-subproblem. Thus, compared with the ADMM scheme (1.3), the PRSM scheme (6.1) offers the same set of advantages. Indeed, according to [11], the RRSM is usually faster than the ADMM whenever it is indeed convergent. The PRSM scheme (6.1), however, according to [11] again (see also [13]), "is less 'robust' in that it converges under more restrictive assumptions than ADMM". This is because there is a significant difference between them — the sequence generated by the ADMM scheme (1.3) is strictly contractive while that by the

PRSM scheme (6.1) is only contractive, with respect to the solution set of (1.1). Thus, the convergence of (6.1) is not guaranteed. We refer to [4, 20] for details.

To tackle the lack of strict contraction of (6.1), the so-called strictly contractive PRSM was proposed in [20]:

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \lambda^k) \mid x_1 \in X_1\}, \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \alpha\beta(A_1 x_1^{k+1} + A_2 x_2^k - b), \\ x_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2, \lambda^{k+\frac{1}{2}}) \mid x_2 \in X_2\}, \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \alpha\beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{6.2}$$

with $\alpha \in (0,1)$. The relaxation factor α ensures that the sequence generated by (6.2) is strictly contractive; thus the convergence can be proved.

In this section, we consider how to use the strictly contractive PRSM scheme (6.2) to solve the model (1.1) and propose a splitting version of the block-wise strictly contractive PRSM scheme for (1.1).

6.1. **Algorithm**

Similarly, we can apply the strictly contractive PRSM (6.2) to the grouped model (1.5); and using the notation in (1.6)-(1.7), we obtain the following block-wise version of the strictly contractive PRSM for the model (1.1):

$$\begin{cases} \boldsymbol{x}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}, \boldsymbol{y}^k, \lambda^k) \mid \boldsymbol{x} \in \mathcal{X}\}, \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \alpha\beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b), \\ \boldsymbol{y}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, \boldsymbol{y}, \lambda^{k+\frac{1}{2}}) \mid \boldsymbol{y} \in \mathcal{Y}\}, \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \alpha\beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^{k+1} - b). \end{cases} \tag{6.3}$$

Furthermore, as (1.13), we can consider decomposing the block-wise $\boldsymbol{x}$- and $\boldsymbol{y}$-subproblems in (6.3) and obtain a splitting version of the block-wise strictly contractive PRSM (6.3):

$$\begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1, x_2^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_1 \in X_1\}, \\ x_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^k, x_2, x_3^k, \cdots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k) \mid x_2 \in X_2\}, \\ \qquad \vdots \\ x_{m_1}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(x_1^k, x_2^k, \cdots, x_{m_1-1}^k, x_{m_1}, \boldsymbol{y}^k, \lambda^k) \mid x_{m_1} \in X_{m_1}\}, \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \alpha\beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b), \\ y_1^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1, y_2^k, \cdots, y_{m_2}^k, \lambda^{k+\frac{1}{2}}) \mid y_1 \in Y_1\}, \\ y_2^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1^k, y_2, y_3^k, \cdots, y_{m_2}^k, \lambda^{k+\frac{1}{2}}) \mid y_2 \in Y_2\}, \\ \qquad \vdots \\ y_{m_2}^{k+1} = \arg\min\{\mathcal{L}_\beta^2(\boldsymbol{x}^{k+1}, y_1^k, y_2^k, \cdots, x_{m_2-1}^k, y_{m_2}, \lambda^{k+\frac{1}{2}}) \mid y_{m_2} \in Y_{m_2}\}, \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \alpha\beta(A\boldsymbol{x}^{k+1} + B\boldsymbol{y}^{k+1} - b). \end{cases} \tag{6.4}$$

In [19], a counter example was given to show that the scheme (6.4) is not necessarily convergent even for the case $m_1 = 1$ and $m_2 = 2$. Therefore, we also consider regularizing the subproblems in (6.4) to ensure the convergence. This is the algorithm summarized below.

Algorithm 3: A splitting version of the block-wise strictly contractive PRSM for (1.1)

Initialization: Specify a grouping strategy for (1.1) and determine the integers m_1 and m_2. Choose the constants $\tau_1 > m_1 - 1$; $\tau_2 > m_2 - 1$; $\alpha \in (0,1)$ and $\beta > 0$. For a given iterate $\boldsymbol{w}^k = (x_1^k, \ldots, x_{m_1}^k, y_1^k, \ldots, y_{m_2}^k, \lambda^k) = (\boldsymbol{x}^k, \boldsymbol{y}^k, \lambda^k)$, the new iterate $\boldsymbol{w}^{k+1}$ is generated by the following steps.

$$\begin{cases} x_i^{k+1} = \arg\min\limits_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big) + \frac{\tau_1\beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \ i = 1, \ldots, m_1, \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \alpha\beta\big(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b\big), \\ y_j^{k+1} = \arg\min\limits_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big(\boldsymbol{x}^{k+1}, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^{k+\frac{1}{2}}\big) + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}, \ j = 1, \cdots, m_2, \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \alpha\beta\big(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^{k+1} - b\big). \end{cases} \tag{6.5}$$

6.2. **Convergence analysis**

As Sections 3 and 5, we conduct the convergence analysis for the splitting version of the strictly contractive PRSM (6.5) in a unified framework. For this purpose, we again rewrite the scheme (6.5) as a prediction-correction form.

First of all, we definite the auxiliary variables

$$\tilde{\boldsymbol{x}}^k = \boldsymbol{x}^{k+1}, \qquad \tilde{\boldsymbol{y}}^k = \boldsymbol{y}^{k+1} \tag{6.6}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b\big). \tag{6.7}$$

Then, we have

$$\begin{aligned} \lambda^{k+\frac{1}{2}} &= \lambda^k - \alpha\beta\big(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^k - b\big) = \lambda^k - \alpha(\lambda^k - \tilde{\lambda}^k) \\ &= \tilde{\lambda}^k + (\alpha - 1)(\tilde{\lambda}^k - \lambda^k). \end{aligned} \tag{6.8}$$

With these notations, we rewrite the scheme (6.5) as a prediction-correction form, in order to use the same analytic framework used in Sections 3 and 5 to analyze the convergence for the (6.5).

Prediction.

$$\tilde{x}_i^k = \arg\min_{x_i \in X_i} \Big\{ \mathcal{L}_\beta^2\big[x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_{m_1}^k, \boldsymbol{y}^k, \lambda^k\big] + \frac{\tau_1\beta}{2}\|A_i(x_i - x_i^k)\|^2 \Big\}, \tag{6.9a}$$

$$\tilde{y}_j^k = \arg\min_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2\big[\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^{k+\frac{1}{2}}\big] + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\}, \tag{6.9b}$$

where

$$\lambda^{k+\frac{1}{2}} = \tilde{\lambda}^k + (\alpha - 1)(\tilde{\lambda}^k - \lambda^k) \tag{6.9c}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b\big). \tag{6.9d}$$

Correction.

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - M_2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{6.10a}$$

where

$$M_2 = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -\alpha\beta\mathcal{B} & 2\alpha I \end{pmatrix}, \tag{6.10b}$$

and $\tilde{\boldsymbol{w}}^k$ is the related sub-vector of the predictor $\tilde{\boldsymbol{w}}^k$ generated by (6.9).

Note that (6.10a) is obtained by using (6.6), (6.8) and the following reasoning:

$$\begin{array}{rcl}\lambda^{k+1} & = & \lambda^{k+\frac{1}{2}} - \alpha\beta(\mathcal{A}\boldsymbol{x}^{k+1} + \mathcal{B}\boldsymbol{y}^{k+1} - b) \\ & = & \tilde{\lambda}^k + (1-\alpha)(\lambda^k - \tilde{\lambda}^k) - \alpha\beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\tilde{\boldsymbol{y}}^k - b) \\ & = & \tilde{\lambda}^k + (1-\alpha)(\lambda^k - \tilde{\lambda}^k) - \alpha[\beta\mathcal{B}(\tilde{\boldsymbol{y}}^k - \boldsymbol{y}^k) + \beta(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b)] \\ & = & \tilde{\lambda}^k + (1-\alpha)(\lambda^k - \tilde{\lambda}^k) - [-\alpha\beta\mathcal{B}(\boldsymbol{y}^k - \tilde{\boldsymbol{y}}^k) + \alpha(\lambda^k - \tilde{\lambda}^k)] \\ & = & \lambda^k - [-\alpha\beta\mathcal{B}(\boldsymbol{y}^k - \tilde{\boldsymbol{y}}^k) + 2\alpha(\lambda^k - \tilde{\lambda}^k)]. \end{array} \tag{6.11}$$

With the prediction-correction reformulation (6.9)-(6.10), we can also establish the same convergence results as those in Sections 3 and 5. As mentioned, Theorems 4.1 and 4.2 are the basis for the convergence analysis. In the following, we prove some conclusions similar as the scheme (5.5).

Theorem 6.1. *Let $\tilde{\boldsymbol{w}}^k$ be generated by* (6.9) *from the given vector $\boldsymbol{w}^k$. Then we have*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \ \ f(\boldsymbol{u}) - f(\tilde{\boldsymbol{u}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \ge (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q_2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \ \ \forall \boldsymbol{w} \in \Omega, \tag{6.12}$$

where the matrix Q_2 is defined by

$$Q_2 = \begin{pmatrix} (\tau_1+1)\beta\, diag(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A} & 0 & 0 \\ 0 & (\tau_2+1)\beta\, diag(\mathcal{B}^T\mathcal{B}) & -\alpha\mathcal{B}^T \\ 0 & -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix}. \tag{6.13}$$

Proof. Since the x_i-subproblems in (6.9a) are the same as those in (4.3a), see (4.8), we have $\tilde{\boldsymbol{x}}^k \in \mathcal{X}$ and

$$\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{x} - \tilde{\boldsymbol{x}}^k)^T\{-\mathcal{A}^T\tilde{\lambda}^k - \beta\mathcal{A}^T\mathcal{A}(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k) + (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A})(\tilde{\boldsymbol{x}}^k - \boldsymbol{x}^k)\} \ge 0, \ \forall \boldsymbol{x} \in \mathcal{X}. \tag{6.14}$$

For the y_j-subproblems in (6.9b), we have

$$\begin{array}{rcl}\tilde{y}_j^k & = & \arg\min\limits_{y_j \in Y_j} \Big\{ \mathcal{L}_\beta^2(\tilde{\boldsymbol{x}}^k, y_1^k, \ldots, y_{j-1}^k, y_j, y_{j+1}^k, \ldots, y_{m_2}^k, \lambda^{k+\frac{1}{2}}) + \dfrac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \Big\} \\ & \overset{(1.11)}{=} & \arg\min\limits_{y_j \in Y_j} \left\{ \begin{array}{c} \phi_j(y_j) - (\lambda^{k+\frac{1}{2}})^T B_j y_j + \frac{\beta}{2}\|\mathcal{A}\tilde{\boldsymbol{x}}^k + B_j(y_j - y_j^k) + \mathcal{B}\boldsymbol{y}^k - b\|^2 \\ + \frac{\tau_2\beta}{2}\|B_j(y_j - y_j^k)\|^2 \end{array} \right\}, \end{array}$$

in which some constant terms in the objective function are ignored. The first-order optimality condition of the above convex minimization problem is

$$\begin{array}{l}\tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T\lambda^{k+\frac{1}{2}} \\ \qquad\qquad + \beta B_j^T[\mathcal{A}\tilde{\boldsymbol{x}}^k + B_j(\tilde{y}_j^k - y_j^k) + \mathcal{B}\boldsymbol{y}^k - b] + \tau_2\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \ge 0, \ \forall y_j \in Y_j. \end{array}$$

Recall (6.9c) and use $(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\boldsymbol{y}^k - b) = \frac{1}{\beta}(\lambda^k - \tilde{\lambda}^k)$. Then, it follows from the last inequality that

$$\begin{array}{l}\tilde{y}_j^k \in Y_j, \quad \phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T(\tilde{\lambda}^k + (\alpha-1)(\tilde{\lambda}^k - \lambda^k)) \\ \qquad +\beta B_j^T[B_j(\tilde{y}_j^k - y_j^k) + \frac{1}{\beta}(\lambda^k - \tilde{\lambda}^k)] + \tau_2\beta B_j^T B_j(\tilde{y}_j^k - y_j^k)\} \geq 0, \ \forall y_j \in Y_j.\end{array}$$

Consequently, we have $\tilde{y}_j^k \in Y_j$ and

$$\phi_j(y_j) - \phi_j(\tilde{y}_j^k) + (y_j - \tilde{y}_j^k)^T\{-B_j^T\tilde{\lambda}^k + (\tau_2+1)\beta B_j^T B_j(\tilde{y}_j^k - y_j^k) - \alpha B_j^T(\tilde{\lambda}^k - \lambda^k)\} \geq 0, \ \forall y_j \in Y_j.$$

Taking $j = 1, \ldots, m_2$ in the above variational inequality and summarizing them, we have $\tilde{\boldsymbol{y}}^k \in \mathcal{Y}$ and

$$\varphi(\boldsymbol{y}) - \varphi(\tilde{\boldsymbol{y}}^k) + (\boldsymbol{y} - \tilde{\boldsymbol{y}}^k)^T\{-\mathcal{B}^T\tilde{\lambda}^k + (\tau_2+1)\beta \text{diag}(\mathcal{B}^T\mathcal{B})(\tilde{\boldsymbol{y}}^k - \boldsymbol{y}^k) - \alpha\mathcal{B}^T(\tilde{\lambda}^k - \lambda^k)\} \geq 0, \tag{6.15}$$

for all $\boldsymbol{y} \in \mathcal{Y}$. Then, using (6.9d), we have

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

which can be rewritten as

$$\tilde{\lambda}^k \in \mathbb{R}^l, \quad (\lambda - \tilde{\lambda}^k)^T\{(\mathcal{A}\tilde{\boldsymbol{x}}^k + \mathcal{B}\tilde{\boldsymbol{y}}^k - b) - \mathcal{B}(\tilde{\boldsymbol{y}}^k - \boldsymbol{y}^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k)\} \geq 0, \quad \forall\, \lambda \in \mathbb{R}^l. \tag{6.16}$$

Combining (6.14), (6.15) and (6.16), and using the notations $F(\boldsymbol{w})$ and Q_2 (see (2.3b) and (6.13)), the assertion of this theorem is proved immediately. □

Theorem 6.2. *Let the matrices M_2 and Q_2 be defined in* (6.10b) *and* (6.13)*, respectively. Then, both the matrices*

$$H_2 = Q_2 M_2^{-1} \tag{6.17}$$

and

$$G_2 = Q_2^T + Q_2 - M_2^T H_2 M_2 \tag{6.18}$$

are positive definite.

Proof. First, we check the positive definiteness of the matrix H_2. For the matrix M_2 defined in (6.10b), we have

$$M_2^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & \frac{1}{2}\beta\mathcal{B} & \frac{1}{2\alpha}I \end{pmatrix}.$$

Note that the left-upper diagonal part of Q_2 equals $\mathcal{D}_A$ (see (6.13) and (2.5)). Thus, according to the definitions of the matrices H_2 and Q_2 (see (6.17) and (6.13)), we have

$$\begin{aligned} H_2 &= Q_2 M_2^{-1} \\ &= \begin{pmatrix} \mathcal{D}_A & 0 & 0 \\ 0 & (\tau_2+1)\beta\text{diag}(\mathcal{B}^T\mathcal{B}) & -\alpha\mathcal{B}^T \\ 0 & -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & \frac{1}{2}\beta\mathcal{B} & \frac{1}{2\alpha}I \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{D}_A & 0 & 0 \\ 0 & (\tau_2+1)\beta\text{diag}(\mathcal{B}^T\mathcal{B}) - \frac{1}{2}\alpha\beta\mathcal{B}^T\mathcal{B} & -\frac{1}{2}\mathcal{B}^T \\ 0 & -\frac{1}{2}\mathcal{B} & \frac{1}{2\alpha\beta}I \end{pmatrix}, \end{aligned}$$

which shows that H_2 is symmetric. The positive definiteness of H_2 follows from $\mathcal{D}_A \succ 0$ and

$$\begin{aligned}
&\begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \frac{1}{2}\alpha\beta\mathcal{B}^T\mathcal{B} & -\frac{1}{2}\mathcal{B}^T \\ -\frac{1}{2}\mathcal{B} & \frac{1}{2\alpha\beta}I \end{pmatrix} \\
= & \begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \alpha\beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & \frac{1}{\beta}I \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \alpha\beta\mathcal{B}^T\mathcal{B} & -\mathcal{B}^T \\ -\mathcal{B} & \frac{1}{\alpha\beta}I \end{pmatrix} \\
\succeq & \begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & \frac{1}{\beta}I \end{pmatrix} \overset{(2.6)}{=} \begin{pmatrix} \mathcal{D}_B & 0 \\ 0 & \frac{1}{\beta}I \end{pmatrix} \succ 0.
\end{aligned}$$

Now, we turn to check the positive definiteness of the matrix G_2. Note that

$$Q_2^T + Q_2 = \begin{pmatrix} 2\mathcal{D}_A & 0 & 0 \\ 0 & 2(\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) & -(1+\alpha)\mathcal{B}^T \\ 0 & -(1+\alpha)\mathcal{B} & \frac{2}{\beta}I \end{pmatrix}$$

and

$$\begin{aligned}
M_2^T H_2 M_2 &= Q_2^T M_2 \\
&= \begin{pmatrix} \mathcal{D}_A & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) & -\mathcal{B}^T \\ 0 & -\alpha\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -\alpha\beta\mathcal{B} & 2\alpha I \end{pmatrix} \\
&= \begin{pmatrix} \mathcal{D}_A & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) + \alpha\beta\mathcal{B}^T\mathcal{B} & -2\alpha\mathcal{B}^T \\ 0 & -2\alpha\mathcal{B} & \frac{2\alpha}{\beta}I \end{pmatrix}.
\end{aligned}$$

From the definition of G_2 (see (6.18)), it follows that

$$\begin{aligned}
G_2 &= Q_2^T + Q_2 - M_2^T H_2 M_2 \\
&= \begin{pmatrix} \mathcal{D}_A & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \alpha\beta\mathcal{B}^T\mathcal{B} & -(1-\alpha)\mathcal{B}^T \\ 0 & -(1-\alpha)\mathcal{B}^T & \frac{2(1-\alpha)}{\beta}I \end{pmatrix}.
\end{aligned}$$

To show that G_2 is positive definite, we need only to verify

$$\begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \alpha\beta\mathcal{B}^T\mathcal{B} & -(1-\alpha)\mathcal{B}^T \\ -(1-\alpha)\mathcal{B}^T & \frac{2(1-\alpha)}{\beta}I \end{pmatrix} \succ 0. \tag{6.19}$$

Because $\alpha \in (0,1)$, we have

$$\begin{aligned}
&\begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \alpha\beta\mathcal{B}^T\mathcal{B} & -(1-\alpha)\mathcal{B}^T \\ -(1-\alpha)\mathcal{B}^T & \frac{2(1-\alpha)}{\beta}I \end{pmatrix} \\
&= \begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & \frac{1-\alpha}{\beta}I \end{pmatrix} + (1-\alpha)\begin{pmatrix} \beta\mathcal{B}^T\mathcal{B} & -\mathcal{B}^T \\ -\mathcal{B} & \frac{1}{\beta}I \end{pmatrix} \\
&\succeq \begin{pmatrix} (\tau_2+1)\beta\mathrm{diag}(\mathcal{B}^T\mathcal{B}) - \beta\mathcal{B}^T\mathcal{B} & 0 \\ 0 & \frac{1-\alpha}{\beta}I \end{pmatrix} \overset{(2.6)}{=} \begin{pmatrix} \mathcal{D}_B & 0 \\ 0 & \frac{1}{\beta}I \end{pmatrix} \succ 0.
\end{aligned}$$

The proof is complete. □

Then, with the conclusions in Theorems 6.1 and 6.2, the proofs for the global convergence and the worst-case convergence rates in both the ergodic and nonergodic senses for the scheme (6.5) are identical as those in Sections 4.2-4.4. We thus omit them. Recall the worst-case convergence rates in both the ergodic and nonergodic senses have been established in [20] for the original strictly contractive PRSM (6.1). With the prediction-correction reformulation (6.9)-(6.10) and the unified analytic framework presented in Sections 4.3-4.4, we can also establish the same convergence results for the splitting version of the block-wise strictly contractive PRSM (6.5).

7. Conclusions

In this paper, we focus on how to apply the original alternating direction method of multipliers (ADMM) to a multiple-block convex minimization model with linear constraints and an objective function in form or more than two functions without coupled variables. If we artificially regroup the variables and functions as two groups and directly apply the original ADMM to the regrouped two-block model, the resulting block-wise subproblems usually should be further decomposed in order to yield easier and solvable subproblems. Then, certain proximal terms are needed to regularize these further decomposed subproblems to ensure the convergence. Accordingly, a splitting version of the block-wise ADMM is derived; and its convergence and worst-case convergence rate measured by the iteration complexity are analyzed. We also extend the analysis to the generalized ADMM and the strictly contractive Peaceman-Rachford splitting method, two methods that are closely relevant to the ADMM; and derive two splitting versions of their block-wise forms. The convergence analysis for these three methods are presented in a unified framework.

The splitting schemes proposed in this paper are to some extent of prototype sense, because no specific properties of the objective functions or structures are assumed in the model (1.1). In fact, the proposed schemes include a number of specific algorithms — even for a given scenario of the model (1.1) with a fixed value of m, different combinations of the two integers m_1 and m_2 for grouping result in different specific algorithms for the proposed three types of splitting schemes. Thus, there is no space to report the elaborated numerical results. In fact, the efficiency of some of the specific algorithms derivable from the proposed schemes have already been well demonstrated. For instance, the special case of (6.4) with $m_1 = 1$ and $m_2 = 2$ has been tested in [19] for some popular applications arising in image processing and machine learning areas.

Some more sophisticated designs based on the algorithms developed in this paper can be conducted for some more concrete scenarios of the abstract model (1.1) where some properties of the functions or some structures of the model are specified. For example, one can further consider how to embed the linearization technique with the decomposed subproblems in (3.1), (5.5) and (6.5) in order to yield really easy subproblems with closed-form solutions for some specifical cases of the model (1.1). Also, one may consider extending the analysis in this paper to the online and stochastic contexts which appear to have wide applications in some areas such as machine learning. This paper is an example in the separable convex programming context of showing how to construct implementable algorithms for more complicated models based on some existing algorithms which are suitable for relatively simpler models. The methodology used in the algorithmic design and the roadmap used in the analysis represent a new development on designing operator splitting methods in the convex programming context, and they might be useful for other algorithmic designs in different contexts.

Appendix A. A divergence example of (1.13)

In this appendix, we show that the splitting version (1.13) of the block-wise ADMM, if without additional proximal regularization in its subproblems, is not necessarily convergent. The example showing its divergence is inspired by the counter example in [3] showing the divergence of the direct extension of ADMM (1.4); and we thank Juwei Lu for providing this example.

We consider the system of linear equations

$$A_1x_1 + A_2x_2 + A_3x_3 = 0, \tag{A.1}$$

where $A_1, A_2, A_3 \in \mathbb{R}^4$ are linearly independent such that the matrix (A_1, A_2, A_3) is full rank and x_1, x_2, x_3 are all in $\mathbb{R}$. This is a special case of the model (1.1) with $m = 3$, $\theta_1(x_1) = \theta_2(x_2) = \theta_3(x_3) = 0$, $l = 4$, $n_1 = n_2 = n_3 = 1$, $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathbb{R}$; and the coefficients matrices are A_1, A_2 and A_3, respectively. Obviously, the system of linear equation (A.1) has the unique solution $x_1^* = x_2^* = x_3^* = 0$. In particular, we consider

$$(A_1, A_2, A_3) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{pmatrix}. \tag{A.2}$$

We consider the scheme (1.13) with $m_1 = 1$ and $m_2 = 2$ and apply it to the homogeneous system of linear equation (A.1). The resulting scheme can be written as

$$\begin{cases} A_1^T(\beta(A_1{x_1}^{k+1} + A_2x_2^k + A_3x_3^k) - \lambda^k) = 0 \\ A_2^T(\beta(A_1{x_1}^{k+1} + A_2{x_2}^{k+1} + A_3x_3^k) - \lambda^k) = 0 \\ A_3^T(\beta(A_1{x_1}^{k+1} + A_2{x_2}^k + A_3{x_3}^{k+1}) - \lambda^k) = 0 \\ \alpha\beta(A_1x_1^{k+1} + A_2x_2^{k+1} + A_3x_3^{k+1}) + \lambda^{k+1} - \lambda^k = 0. \end{cases} \tag{A.3}$$

It follows from the first equation in (A.3) that

$$x_1^{k+1} = \frac{1}{A_1^TA_1}(-A_1^TA_2x_2^k - A_1^TA_3x_3^k + A_1^T\lambda^k/\beta). \tag{A.4}$$

For ease of notation, let us denote $\mu^k = \lambda^k/\beta$. Then, plugging (A.4) into the rest equations in (A.3), we derive that

$$\begin{pmatrix} A_2^TA_2 & 0 & 0 \\ 0 & A_3^TA_3 & 0 \\ \alpha A_2 & \alpha A_3 & I \end{pmatrix} \begin{pmatrix} x_2^{k+1} \\ x_3^{k+1} \\ \mu^{k+1} \end{pmatrix}$$
$$= \left[\begin{pmatrix} 0 & -A_2^TA_3 & A_2^T \\ -A_3^TA_1 & 0 & A_3^T \\ 0 & 0 & I \end{pmatrix} - \frac{1}{A_1^TA_1} \begin{pmatrix} A_2^TA_1 \\ A_3^TA_1 \\ \alpha A_1 \end{pmatrix} (-A_1^TA_2, -A_1^TA_3, A_1^T) \right] \begin{pmatrix} x_2^k \\ x_3^k \\ \mu^k \end{pmatrix}.$$

Let

$$L_2 = \begin{pmatrix} A_2^TA_2 & 0 & 0 \\ 0 & A_3^TA_3 & 0 \\ \alpha A_2 & \alpha A_3 & I \end{pmatrix},$$

$$R_2 = \begin{pmatrix} 0 & -A_2^TA_3 & A_2^T \\ -A_3^TA_1 & 0 & A_3^T \\ 0 & 0 & I \end{pmatrix} - \frac{1}{A_1^TA_1} \begin{pmatrix} A_2^TA_1 \\ A_3^TA_1 \\ \alpha A_1 \end{pmatrix} \left(-A_1^TA_2, -A_1^TA_3, A_1^T\right),$$

and denote

$$M_2 = L_2^{-1}R_2.$$

Then, the scheme (A.3) can be written compactly as

$$\begin{pmatrix} x_2^k \\ x_3^k \\ \mu^k \end{pmatrix} = M_1^k \begin{pmatrix} x_2^0 \\ x_3^0 \\ \mu^0 \end{pmatrix}.$$

Obviously, if the the spectral radius of M_1, denoted by $\rho(M_1) := |\lambda_{\max}(M_1)|$ (the largest eigenvalue of M_1), is not smaller than 1, then the sequence generated by the scheme above is not possible to converge to the solution point $(x_1^*, x_2^*, x_3^*) = (0, 0, 0)$ of the system (A.1). It can be numerically shown that $\rho(M_1) \geq 1$ for different choices of α varying from 0 to 1 with the equal distance of 0.02; and it is monotonically increasing with respect to $\alpha \in (0, 1)$. Therefore, the sequence generated by the scheme above is not convergent to the solution point of the system (A.1). It is thus illustrated that the scheme (1.13) is not necessarily convergent.

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. *Foun. Trends Mach. Learn.*, 3(1):1–122, 2011.

[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):2179–2186, 2013.

[3] C. H. Chen, B. S. He, Y. Y. Ye, and X. M. Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Math. Program.*, 155(1-2):57–79, 2016.

[4] E. Corman and X. M. Yuan. A generalized proximal point algorithm and its convergence rate. *SIAM J. Optim*, 24(4):1614–1638, 2014.

[5] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.*, 82(2):421–439, 1956.

[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(1-3):293–318, 1992.

[7] J. Eckstein and W. Yao. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports*, 32, 2012.

[8] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.*, 3(4):1015–1046, 2010.

[9] F. Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I.* Springer Science & Business Media, 2003.

[10] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and Its Applications*, pages 299 – 331. Elsevier, 1983.

[11] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer-Verlag Berlin Heidelberg, 1984.

[12] R. Glowinski. On alternating direction methods of multipliers: a historical perspective. In *Modeling, Simulation and Optimization for Science and Technology*, pages 59–82. Springer, 2014.

[13] R. Glowinski, T. Kärkkäinen, and K. Majava. On the convergence of operator-splitting methods. In Y. Kuznetsov, P. Neittanmaki, and O. Pironneau, editors, *Numerical Methods for Scienfic computing, Variational Problems and Applications.* CIMNE, 2003.

[14] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):41–76, 1975.

[15] E. G. Gol′shteĭn and N. V. Tret′yakov. Modified lagrangians in convex programming and their generalizations. In *Point-to-Set Maps and Mathematical Programming*, pages 86–97. Springer, 1979.

[16] W. W. Hager, C. Ngo, M. Yashtini, and H. Zhang. An alternating direction approximate newton algorithm for ill-conditioned inverse problems with application to parallel mri. *Journal of the Operations Research Society of China*, 3(2):139–162, 2015.

[17] B. S. He, L. S. Hou, and X. M. Yuan. On full jacobian decomposition of the augmented lagrangian method for separable convex programming. *SIAM J. Optim*, 25(4):2274–2312, 2015.

[18] B. S. He, L. Z. Liao, D. Han, and H. Yang. A new inexact alternating directions method for monotone variational inequalities. *Math. Program.*, 92(1):103–118, 2002.

[19] B. S. He, H. Liu, J. Lu, and X. M. Yuan. Application of the strictly contractive peaceman-rachford splitting method to multi-block convex programming. In R. Glowinksi, S. Osher, and W. Yin, editors, *Operator Splitting Methods and Applications.* to appear.

[20] 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(3):1011–1040, 2014.

[21] B. S. He, M. Tao, and X. M. Yuan. Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming.

[22] B. S. He, M. Tao, and X. M. Yuan. Alternating direction method with gaussian back substitution for separable convex programming. *SIAM J. Optim*, 22(2):313–340, 2012.

[23] B. S. He, M. Tao, and X. M. Yuan. A splitting method for separable convex programming. *IMA J. Numer. Anal.*, 35(1):394–426, 2015.

[24] B. S. He and X. M. Yuan. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012.

[25] B. S. He and X. M. Yuan. On non-ergodic convergence rate of douglas–rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

[26] M. Hong and Z. Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Math. Program.*, to appear.

[27] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[28] B. Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle, série rouge*, 4(3):154–158, 1970.

[29] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1):125–161, 2013.

[30] F. Ng, M. K.and Wang and X. M. Yuan. Inexact alternating direction methods for image recovery. *SIAM J. Sci. Comput.*, 33(4):1643–1668, 2011.

[31] D. H. Peaceman and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *SIAM J. Appl. Math.*, 3(1):28–41, 1955.

[32] Y.G. Peng, A. Ganesh, J. Wright, W. L. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Anal. Mach. Intel.*, 34(11):2233–2246, 2012.

[33] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim*, 24(1):269–297, 2014.

[34] M. Tao and X. M. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM J. Optim*, 21(1):57–81, 2011.

[35] X. F. Wang and X. M. Yuan. The linearized alternating direction method of multipliers for dantzig selector. *SIAM J. Sci. Comput.*, 34(5):A2792–A2811, 2012.

[36] J. F. Yang and X. M. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.*, 82(281):301–329, 2013.