IMA Journal of Numerical Analysis (2019) **00**, 1–29
doi:10.1093/imanum/dry092

Optimal proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems

BINGSHENG HE
Department of Mathematics, Southern University of Science and Technology, China, and Department of Mathematics, Nanjing University, China
hebma@nju.edu.cn

FENG MA
High-Tech Institute of Xi'an, Xi'an, 710025, Shaanxi, China
mafengnju@gmail.com

AND

XIAOMING YUAN*
Department of Mathematics, The University of Hong Kong, Hong Kong
*Corresponding author: xmyuan@hku.hk xmyuan@gmail.com

[Received on 4 March 2018; revised on 4 November 2018]

The augmented Lagrangian method (ALM) is fundamental in solving convex programming problems with linear constraints. The proximal version of ALM, which regularizes ALM's subproblem over the primal variable at each iteration by an additional positive-definite quadratic proximal term, has been well studied in the literature. In this paper we show that it is not necessary to employ a positive-definite quadratic proximal term for the proximal ALM and the convergence can be still ensured if the positive definiteness is relaxed to indefiniteness by reducing the proximal parameter. An indefinite proximal version of the ALM is thus proposed for the generic setting of convex programming problems with linear constraints. We show that our relaxation is optimal in the sense that the proximal parameter cannot be further reduced. The consideration of indefinite proximal regularization is particularly meaningful for generating larger step sizes in solving ALM's primal subproblems. When the model under discussion is separable in the sense that its objective function consists of finitely many additive function components without coupled variables, it is desired to decompose each ALM's subproblem over the primal variable in Jacobian manner, replacing the original one by a sequence of easier and smaller decomposed subproblems, so that parallel computation can be applied. This full Jacobian splitting version of the ALM is known to be not necessarily convergent, and it has been studied in the literature that its convergence can be ensured if all the decomposed subproblems are further regularized by sufficiently large proximal terms. But how small the proximal parameter could be is still open. The other purpose of this paper is to show the smallest proximal parameter for the full Jacobian splitting version of ALM for solving multi-block separable convex minimization models.

Keywords: convex programming; augmented Lagrangian method; proximal point algorithm; multi-block separable model; Jacobian splitting; parallel computation.

1. Introduction

We consider the generic convex minimization model with linear constraints

$$\min\{\boldsymbol{\theta}(\mathbf{x}) \mid \mathcal{A}\mathbf{x} = b,\ \mathbf{x} \in \mathcal{X}\}, \tag{1.1}$$

where $\boldsymbol{\theta} : \Re^n \to \Re$ is a closed proper convex, but not necessarily smooth function; $\mathcal{X} \subseteq \Re^n$ is a closed convex set; $\mathcal{A} \in \Re^{\ell \times n}$ and $b \in \Re^\ell$. The solution set of (1.1) is assumed to be nonempty throughout our discussion.

Let the Lagrangian function of (1.1) be defined as

$$L(\mathbf{x}, \lambda) = \boldsymbol{\theta}(\mathbf{x}) - \lambda^T(\mathcal{A}\mathbf{x} - b), \tag{1.2}$$

in which $\lambda \in \Re^\ell$ be the Lagrange multiplier. Moreover, we define the augmented Lagrangian function of the problem (1.1) as

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

with $\beta > 0$ the penalty parameter for the linear constraints. The augmented Lagrangian method (ALM) originally proposed in Hestenes (1969) and Powell (1969) for (1.1) reads as

$$\text{(ALM)} \begin{cases} \mathbf{x}^{k+1} = \arg\min\{\mathcal{L}_\beta(\mathbf{x}, \lambda^k) \mid \mathbf{x} \in \mathcal{X}\}, & \text{(1.4a)} \\ \lambda^{k+1} = \lambda^k - \beta(\mathcal{A}\mathbf{x}^{k+1} - b). & \text{(1.4b)} \end{cases}$$

The ALM plays a significant role in both theoretical study and algorithmic design for various convex programming models and the literature is too voluminous to list. A particularly insightful one is Rockafellar (1976a), showing that the ALM scheme (1.4) is an application of the well-known proximal point algorithm (PPA) that can date back to the seminal work (Martinet, 1970; Rockafellar, 1976b) to the dual problem of (1.1). Throughout, we also call $\mathbf{x}$ and λ the primal and dual variables; (1.4a) and (1.4b) the primal subproblem and the dual updating of ALM, respectively. For the penalty parameter β we fix it in our discussion for simplicity. Indeed, as our work (Chen *et al.*, 2016) shows, without loss of generality, we can fix it as 1 for theoretical discussion. We refer to Bertsekas (1982) for insightful discussions on how to determine this parameter for the sake of generating better numerical performance.

To implement the ALM (1.4) it is meaningful to discuss how to solve the primal subproblem (1.4a). An interesting strategy is to regularize the primal subproblem (1.4a) by a quadratic proximal term and accordingly consider the proximal version of ALM:

$$\text{(Proximal ALM)} \begin{cases} \mathbf{x}^{k+1} = \arg\min\{\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}} \mid \mathbf{x} \in \mathcal{X}\}, & \text{(1.5a)} \\ \lambda^{k+1} = \lambda^k - \beta(\mathcal{A}\mathbf{x}^{k+1} - b). & \text{(1.5b)} \end{cases}$$

In (1.5a) $\frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}}$ is the quadratic proximal regularization term and $\mathcal{D} \in \Re^{n \times n}$ is the proximal matrix that is usually required to be positive definite in the literature. The proximally regularized subproblem (1.5a) is in nature of the PPA in Martinet (1970) and Rockafellar (1976b). Analytically, because of the positive-definite quadratic proximal regularization, it is easy to establish the convergence of (1.5) under the positive-definiteness assumption of the proximal matrix $\mathcal{D}$; see, e.g., Section 2.3 for the convergence of a more general version of the proximal ALM (1.5) that allows for a more general step size in updating the dual variable λ.

From the algorithmic implementation perspective, the proximal ALM (1.5) is also interesting. It is straightforward to see that, by ignoring some constant terms in the objective, this subproblem has the

same solution as the following one:

$$\mathbf{x}^{k+1} = \arg\min \left\{ \boldsymbol{\theta}(\mathbf{x}) + \frac{\beta}{2} \left\| \mathcal{A}\mathbf{x} - \left(\frac{1}{\beta}\lambda^k + b \right) \right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}} \mid \mathbf{x} \in \mathcal{X} \right\}. \tag{1.6}$$

For a general objective function $\boldsymbol{\theta}(\mathbf{x})$ iterations are still required to iteratively approach to a solution point of the subproblem (1.6). But for some cases that often arise in data-driven applications, $\boldsymbol{\theta}(\mathbf{x})$ may be special enough so that its proximity operator, which is given by

$$\mathrm{Prox}_{\boldsymbol{\theta},\rho}(\mathbf{x}) := \mathrm{argmin} \left\{ \boldsymbol{\theta}(\mathbf{x}) + \frac{1}{2\rho}\|\mathbf{z} - \mathbf{x}\|^2 \mid \mathbf{z} \in \Re^n \right\}, \tag{1.7}$$

has a closed-form expression; in which $\rho > 0$ is a constant. Such a representative case is where $\boldsymbol{\theta}(\mathbf{x}) = \|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$. If the proximal matrix $\mathcal{D}$ in (1.5) is chosen as

$$\mathcal{D} = r \cdot I_n - \beta \mathcal{A}^T \mathcal{A} \tag{1.8}$$

then the proximally regularized ALM subproblem (1.5a) is specified as

$$\mathbf{x}^{k+1} = \arg\min \left\{ \boldsymbol{\theta}(\mathbf{x}) + \frac{r}{2} \left\| \mathbf{x} - \mathbf{x}^k - \frac{1}{r}\mathcal{A}^T\big(\lambda^k - \beta(\mathcal{A}\mathbf{x}^k - b)\big) \right\|^2 \mid \mathbf{x} \in \mathcal{X} \right\}, \tag{1.9}$$

which amounts to estimating the proximity operator of $\boldsymbol{\theta}(\mathbf{x})$ when $\mathcal{X} = \Re^n$. The implementation of (1.5) for such cases is thus extremely simple.

Hence, the linearized ALM, which is a special case of the proximal ALM (1.5) with $\mathcal{D}$ given in (1.8), reads as

$$\text{(Linearized ALM)} \begin{cases} \mathbf{x}^{k+1} = \arg\min \left\{ \boldsymbol{\theta}(\mathbf{x}) + \frac{r}{2}\left\| \mathbf{x} - \mathbf{x}^k - \frac{1}{r}\mathcal{A}^T\big(\lambda^k - \beta(\mathcal{A}\mathbf{x}^k - b)\big) \right\| \mid \mathbf{x} \in \mathcal{X} \right\}, & \text{(1.10a)} \\ \lambda^{k+1} = \lambda^k - \beta\big(\mathcal{A}\mathbf{x}^{k+1} - b\big). & \text{(1.10b)} \end{cases}$$

Recall that for the linearized ALM (1.10) the parameter r is required to satisfy the condition $r > \beta\|\mathcal{A}^T\mathcal{A}\|$ so as to ensure the positive definiteness of the matrix $\mathcal{D}$ given in (1.8), and hence the convergence of (1.10). We refer to Yang & Yuan (2013) and Section 2.3 for the detail of convergence analysis of the linearized ALM (1.10).

The parameter r also determines the step size for solving the $\mathbf{x}$-subproblem (1.10a) of the linearized ALM, and we prefer smaller values of r as long as the convergence of (1.10) can be guaranteed. For example, if the choice of (1.8) is relaxed to

$$\widetilde{\mathcal{D}} = \tau r I_n - \beta \mathcal{A}^T \mathcal{A} \quad \text{with} \quad r > \beta\|\mathcal{A}^T\mathcal{A}\| \text{ and } \tau \in (0, 1) \tag{1.11}$$

then the resulting primal problem of the proximal ALM (1.5a) still reduces to a problem analogous to that in the linearized ALM (1.10a); while obviously the quadratic proximal term $\frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\widetilde{\mathcal{D}}}$ with $\widetilde{\mathcal{D}}$ in (1.11) plays a lighter weight in the objective and thus the primal variable x can be updated with a larger step size. The efficiency of the linearized ALM with such a choice can be easily verified numerically, see, e.g., the appendix.

Therefore, the necessity of considering indefinite quadratic proximal regularization is illustrated by the linearized ALM context, and it inspires us to consider the possibility of further relaxing the positive-definiteness requirement of the proximal matrix $\mathcal{D}$ in (1.5a) for the general scenario of the proximal ALM (1.5). That is, we consider regularizing the ALM primal subproblem (1.4a) with a quadratic proximal term while its proximal matrix is not necessarily positive definite. Hence, instead of (1.5a), we solve the surrogate

$$\mathbf{x}^{k+1} = \arg\min\left\{\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_0} \;\middle|\; \mathbf{x} \in \mathcal{X}\right\}, \tag{1.12}$$

where the proximal matrix $\mathcal{D}_0$ is not necessarily positive definite. Note that we slightly abuse the notation of $\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_0}$ to denote the term $(\mathbf{x} - \mathbf{x}^k)^T\mathcal{D}_0(\mathbf{x} - \mathbf{x}^k)$ regardless of the indefiniteness of $\mathcal{D}_0$. For convenience of analysis we specify the structure of $\mathcal{D}_0$ as

$$\mathcal{D}_0 = \mathcal{D} - (1 - \tau)\beta\mathcal{A}^T\mathcal{A}, \tag{1.13}$$

where $\mathcal{D}$ is an arbitrarily positive-definite matrix in $\Re^{n\times n}$ and $\tau \in (0, 1)$. Obviously, $\mathcal{D}_0$ defined in (1.13) is not necessarily positive definite yet the convexity of the x-subproblem (1.12) with the choice (1.13) is kept. If we choose $\mathcal{D} = \tau(rI_n - \beta\mathcal{A}^T\mathcal{A})$, a smaller value of (1.8), then $\mathcal{D}_0$ given in (1.13) corresponds to the specific choice in (1.11), and the linearized ALM with larger step size is recovered. For this case, the indefinite proximal regularizer defined by (1.13) generates the same type of subproblems as that of (1.10a) in the linearized ALM, subject to just a fixed ratio of the proximal parameter. Here, we do not focus on how to choose $\mathcal{D}$. Instead, we are interested in the choice of τ for $\mathcal{D}_0$ given in (1.13). Since τ is the parameter determining the step size for solving the subproblem (1.12) hereafter we call τ the step size parameter for the primal ALM subproblem.

On the other hand, it is equally interesting to investigate the step size in updating the dual variable λ in (1.4b). Conventionally, the step size is chosen as 1 in the original ALM scheme (1.4). But it has been observed in the literature that a larger step size may accelerate the convergence empirically. We refer to, e.g., Glowinski (1984) and He *et al.* (2016a), for some theoretical and experimental studies on how to use larger step sizes to update the dual variable in the context of alternating direction method of multipliers (ADMM), which was originally proposed in Glowinski & Marrocco (1975) and can be regarded as a splitting version of the ALM. Here we are interested in the more general scheme

$$\lambda^{k+1} = \lambda^k - \gamma\beta\big(\mathcal{A}\mathbf{x}^{k+1} - b\big),\ \gamma \in (0, 2), \tag{1.14}$$

with the possibility of enlarging the step size in updating the dual variable. Recall that, as analyzed in Bertsekas (1982) and Rockafellar (1976b), in (1.4b) the dual variable λ is updated by a steepest accent step applied to the dual problem of (1.1). In Tao & Yuan (2018) it is shown that the step size γ cannot be equivalent to or larger than 2; hence, it is sufficient to restrict $\gamma \in (0, 2)$. Hereafter, we call γ the step size parameter for the dual ALM subproblem.

Based on the discussion above we propose the following indefinite proximal ALM with a general step size for updating the dual variable:

$$\text{(IDP-ALM)}\ \begin{cases} \mathbf{x}^{k+1} = \arg\min\big\{\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_0} \;\big|\; \mathbf{x} \in \mathcal{X}\big\}, & \text{(1.15a)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(\mathcal{A}\mathbf{x}^{k+1} - b), \quad \gamma \in (0, 2), & \text{(1.15b)} \end{cases}$$

where $\mathcal{D}_0$ is given in (1.13) with an arbitrarily given positive-definite matrix $\mathcal{D} \in \Re^{n\times n}$ and $\tau \in (0, 1)$. It is abbreviated as IDP-ALM hereafter. For the IDP-ALM (1.15) we shall prove its convergence without any additional assumption on the model (1.1). Recall that the linearized ALM (1.10) is a special case of (1.15) where the matrix $\mathcal{D}$ is chosen as (1.8), $\tau = 1$ and $\gamma = 1$.

For the IDP-ALM (1.15) it is interesting to know the intrinsic relationship between these two step size parameters τ and γ in (1.15) to ensure its convergence. More precisely, how small could τ be when γ is fixed, and on the contrary, how large could γ be when τ is fixed? This is the principal question we want to answer in this paper. Intuitively, it is not hard to see that the primal and dual subproblems in (1.15) should be solved in counter natures in the sense that if the primal subproblem is solved more conservatively (using a larger value of τ) then it becomes rationale to employ a larger step size in updating the dual variable so that the dual subproblem is solved more aggressively, and vice versa. The mutually constrained nature of the step size parameters τ and γ will be rigorously formatted by the general formula

$$\tau > \frac{2+\gamma}{4}. \tag{1.16}$$

The formula (1.16) defines the 'trade-off' nature between τ and γ mathematically and quantitatively. According to (1.16), for any given $\gamma \in (0, 2)$, we can discern a γ-depending lower bound of τ for the IDP-ALM (1.15), and vice versa, a τ-depending upper bound of γ. Also, for the special case of $\gamma = 1$ that is the most popular case in the literature of the ALM and its proximal versions, the formula immediately implies that $\tau > 3/4$, instead of $\tau > 1$, is sufficiently good to determine the proximal regularization for the ALM. Hence, all existing proximal versions of the ALM can be immediately improved by using this smaller proximal term. Note that $\tau \geq 1$ is less interesting because of two reasons. (1) If $\tau \geq 1$ then the matrix $\mathcal{D}_0$ defined in (1.13) is positive definite and the scheme (1.15) with such $\mathcal{D}_0$ reduces to the regular proximal ALM (1.5). Hence, the convergence analysis for this case is much more trivial, as shown in Yang & Yuan (2013) and Section 2.3. (2) If $\tau \geq 1$ then it implies that the primal subproblem (1.15a) is solved conservatively with a too-small step size; hence, it is generally not preferable from numerical perspectives. Therefore, although we are mainly interested in the lower bound of τ given in (1.16) with which the convergence of (1.15) can be ensured, it is assumed by default that $\tau < 1$ in our discussion to be presented.

After summarizing some preliminaries in Section 2 we prove the convergence of the IDP-ALM (1.15) with the restriction (1.16) in Section 3.[1] In Section 4 we further show that the bound of τ given by (1.16) is optimal. That is, we construct an example to show that the scheme (1.15) is divergent for any $\tau < (2 + \gamma)/4$. Hence, the bound $(2 + \gamma)/4$ is optimal for the IDP-ALM (1.15) with guaranteed convergence. Then, in Section 5, we discuss the multi-block separable case of (1.1) where the objective function is the sum of finitely many additive function components without coupled variables and show how to improve our result in He *et al.* (2016c). This is the second purpose of this paper. Finally, some conclusions are drawn in Section 6.

2. Preliminaries

In this section we summarize some well-known preliminaries that will be used for further discussions and show some simple results.

[1] The partial result for the special case of $\gamma = 1$ and thus $\tau > 0.75$ has been released, via a different analysis, in our earlier preprint (He *et al.*, 2016b) posed on Optimization Online in July 2016.

2.1 *Variational inequality characterization of* (1.1)

The pair $(\mathbf{x}^*, \lambda^*)$ defined on $\mathcal{X} \times \Re^\ell$ is called a saddle point of the Lagrangian function (1.2) if it satisfies the inequalities

$$L_{\lambda \in \Re^\ell}(\mathbf{x}^*, \lambda) \le L(\mathbf{x}^*, \lambda^*) \le L_{\mathbf{x} \in \mathcal{X}}(\mathbf{x}, \lambda^*).$$

Alternatively, we can rewrite these inequalities as the variational inequalities

$$\begin{cases} \mathbf{x}^* \in \mathcal{X}, & \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^T(-\mathcal{A}^T\lambda^*) \ge 0, \ \forall \mathbf{x} \in \mathcal{X}, \\ \lambda^* \in \Re^\ell, & (\lambda - \lambda^*)^T(\mathcal{A}\mathbf{x}^* - b) \ge 0, \ \forall \lambda \in \Re^\ell, \end{cases} \tag{2.1}$$

or in the compact form

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

where

$$\mathbf{u} = \begin{pmatrix} \mathbf{x} \\ \lambda \end{pmatrix}, \quad F(\mathbf{u}) = \begin{pmatrix} -\mathcal{A}^T\lambda \\ \mathcal{A}\mathbf{x} - b \end{pmatrix} \quad \text{and} \quad \Omega = \mathcal{X} \times \Re^\ell. \tag{2.2b}$$

We denote by $\mathrm{VI}(\Omega, F, \boldsymbol{\theta})$ the variational inequality (2.2). Note that for the operator F defined in (2.2b) it is affine with a skew-symmetric matrix and thus we have

$$(\mathbf{u} - \boldsymbol{v})^T(F(\mathbf{u}) - F(\boldsymbol{v})) \equiv 0. \tag{2.3}$$

We also call (2.2) a monotone mixed variational inequality because the function $\boldsymbol{\theta}$ is convex and the operator F has the property (2.3). We denote by Ω^* the solution set of the variational inequality (2.2).

2.2 *A basic lemma*

The following lemma is basic and will be frequently used in our analysis. Its proof is elementary and thus omitted.

LEMMA 2.1 Let $\mathcal{X} \subset \Re^n$ be a closed convex set, $\theta(\mathbf{x})$ and $f(\mathbf{x})$ be convex functions. If f is differentiable, and the solution set of the minimization problem $\min\{\theta(\mathbf{x}) + f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\}$ is nonempty, then

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

if and only if

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

2.3 *Proximal ALM with a general step size for dual variable*

In this subsection we modify the proximal ALM (1.5) with a more general step size in updating the dual variable and prove its convergence. Its iterative scheme reads as

$$\text{(General Proximal ALM)} \begin{cases} \mathbf{x}^{k+1} = \arg\min\left\{\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}} \;\middle|\; \mathbf{x} \in \mathcal{X}\right\}, & (2.5a) \\ \lambda^{k+1} = \lambda^k - \gamma\beta(\mathcal{A}\mathbf{x}^{k+1} - b), \quad \gamma \in (0, 2), & (2.5b) \end{cases}$$

where the proximal matrix $\mathcal{D} \in \Re^{n\times n}$ is positive definite.

We first observe that the objective function of the primal subproblem (2.5a) is

$$\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}} = \boldsymbol{\theta}(\mathbf{x}) - (\lambda^k)^T(\mathcal{A}\mathbf{x} - b) + \frac{\beta}{2}\|\mathcal{A}\mathbf{x} - b\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}}.$$

According to Lemma 2.1 we have $\mathbf{x}^{k+1} \in \mathcal{X}$ and it is characterized by the variational inequality

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}\big(\mathbf{x}^{k+1}\big) + \big(\mathbf{x} - \mathbf{x}^{k+1}\big)^T\big\{-\mathcal{A}^T\big[\lambda^k - \beta(\mathcal{A}\mathbf{x}^{k+1} - b)\big] + \mathcal{D}\big(\mathbf{x}^{k+1} - \mathbf{x}^k\big)\big\} \geq 0,\ \forall\, \mathbf{x} \in \mathcal{X}. \quad (2.6)$$

By using (2.5b) it holds that

$$\lambda^k - \beta\big(\mathcal{A}\mathbf{x}^{k+1} - b\big) = \lambda^k + \frac{1}{\gamma}\big(\lambda^{k+1} - \lambda^k\big) = \lambda^{k+1} - \left(1 - \frac{1}{\gamma}\right)\big(\lambda^{k+1} - \lambda^k\big).$$

Substituting it into (2.6) it follows that $\mathbf{x}^{k+1} \in \mathcal{X}$ and

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}\big(\mathbf{x}^{k+1}\big) + \big(\mathbf{x} - \mathbf{x}^{k+1}\big)^T\left\{-\mathcal{A}^T\lambda^{k+1} + \mathcal{D}\big(\mathbf{x}^{k+1} - \mathbf{x}^k\big) + \left(1 - \frac{1}{\gamma}\right)\mathcal{A}^T\big(\lambda^{k+1} - \lambda^k\big)\right\} \geq 0,\ \forall\, \mathbf{x} \in \mathcal{X}. \quad (2.7a)$$

Let us rewrite the equality $\lambda^{k+1} = \lambda^k - \gamma\beta(\mathcal{A}\mathbf{x}^{k+1} - b)$ as the variational inequality

$$\lambda^{k+1} \in \Re^\ell,\ \ \big(\lambda - \lambda^{k+1}\big)^T\left\{\big(\mathcal{A}\mathbf{x}^{k+1} - b\big) + \frac{1}{\gamma\beta}\big(\lambda^{k+1} - \lambda^k\big)\right\} \geq 0,\ \ \forall\, \lambda \in \Re^\ell. \quad (2.7b)$$

Then, using the notation in (2.2), we can rewrite the inequalities (2.7) as

$$\mathbf{u}^{k+1} \in \Omega,\ \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}\big(x^{k+1}\big) + \big(\mathbf{u} - \mathbf{u}^{k+1}\big)^T\big\{F\big(\mathbf{u}^{k+1}\big) + P\big(\mathbf{u}^{k+1} - \mathbf{u}^k\big)\big\} \geq 0, \quad \forall\, \mathbf{u} \in \Omega, \quad (2.8)$$

where

$$P = \begin{pmatrix} \mathcal{D} & (1 - \frac{1}{\gamma})\mathcal{A}^T \\ 0 & \frac{1}{\gamma\beta}I_\ell \end{pmatrix}.$$

This is the variational inequality characterization of the $(k+1)$th iterate generated by the general proximal ALM (2.5).

Now, we can show the convergence of (2.5) easily. First, setting $\mathbf{u} = \mathbf{u}^*$ in (2.8) and using (2.3), we get

$$\big(\mathbf{u}^{k+1} - \mathbf{u}^*\big)^T P\big(\mathbf{u}^k - \mathbf{u}^{k+1}\big) \geq \boldsymbol{\theta}\big(x^{k+1}\big) - \boldsymbol{\theta}(\mathbf{x}^*) + \big(\mathbf{u}^{k+1} - \mathbf{u}^*\big)^T F(\mathbf{u}^*).$$

Notice that the right-hand side of the last inequality is non-negative. It follows from the definition of the matrix P that

$$\begin{aligned}\left(\mathbf{x}^{k+1}-\mathbf{x}^*\right)^T\mathcal{D}\left(\mathbf{x}^k-\mathbf{x}^{k+1}\right)+\left(\mathbf{x}^{k+1}-\mathbf{x}^*\right)^T\left(1-\frac{1}{\gamma}\right)\mathcal{A}^T\left(\lambda^k-\lambda^{k+1}\right)\\ +\left(\lambda^{k+1}-\lambda^*\right)^T\frac{1}{\gamma\beta}\left(\lambda^k-\lambda^{k+1}\right)\geq 0,\quad \forall\,\mathbf{u}^*\in\Omega^*.\end{aligned}$$

Using $\mathcal{A}\mathbf{x}^*=b$ we get

$$\left(\mathbf{x}^{k+1}-\mathbf{x}^*\right)^T\mathcal{D}\left(\mathbf{x}^k-\mathbf{x}^{k+1}\right)+\left(\lambda^{k+1}-\lambda^*\right)^T\frac{1}{\gamma\beta}\left(\lambda^k-\lambda^{k+1}\right)\geq(1-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1}-b\|^2,\ \forall\,\mathbf{u}^*\in\Omega^*.\quad(2.9)$$

Moreover, applying the identity $2\eta^T(\xi-\eta)=\|\xi\|^2-\|\eta\|^2-\|\xi-\eta\|^2$, it follows from (2.9) that

$$\begin{aligned}&\left(\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^*\|^2\right)-\left(\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^{k+1}-\lambda^*\|^2\right)\\ &\qquad\geq\left(\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^{k+1}\|^2\right)+2(1-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1}-b\|^2,\ \forall\,\mathbf{u}^*\in\Omega^*.\end{aligned}$$

Consequently, using $\lambda^k-\lambda^{k+1}=\gamma\beta(\mathcal{A}\mathbf{x}^{k+1}-b)$, we have

$$\begin{aligned}&\left(\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^{k+1}-\lambda^*\|^2\right)\\ &\leq\left(\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^*\|^2\right)-\left(\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{D}}^2+\left(\frac{2-\gamma}{\gamma}\right)\frac{1}{\gamma\beta}\|\lambda^k-\lambda^{k+1}\|^2\right),\ \forall\,\mathbf{u}^*\in\Omega^*,\end{aligned}$$

or, equivalently,

$$\begin{aligned}&\left(\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^{k+1}-\lambda^*\|^2\right)\\ &\leq\left(\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathcal{D}}^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^*\|^2\right)-\left(\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{D}}^2+(2-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1}-b\|^2\right),\ \forall\,\mathbf{u}^*\in\Omega^*.\end{aligned}$$

This inequality essentially implies the convergence of the sequence generated by the scheme (2.5). The remaining part of the proof is routine and thus omitted; we refer to, e.g., Blum & Oettli (1975) or He (2015) for a tutorial.

3. Convergence of the IDP-ALM (1.15)

In this section we prove the convergence for the IDP-ALM (1.15) with the step size restriction (1.16). Note that the IDP-ALM (1.15) differs from the scheme (2.5) in that the proximal matrix $\mathcal{D}_0$ in (1.15a) is indefinite. This difference makes the convergence proof of (1.15) much more challenging than that of

(2.5) as we just established in Section 2.3. Recall that the parameter τ is for defining the matrix $\mathcal{D}_0$ in (1.15a) through (1.13) for an arbitrarily given positive definite matrix $\mathcal{D} \in \Re^{n\times n}$.

3.1 *Main theorem for convergence proof*

The key theorem for proving the convergence of the IDP-ALM (1.15) is presented below.

THEOREM 3.1 Let $\{\mathbf{u}^k\}$ be the sequence generated by the IDP-ALM (1.15) for the problem (1.1). $\tau > 0$ is the parameter in the matrix D_0 that appears in the subproblem (1.15a). Then, for any $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$, we have

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\|_H^2 \leq \|\mathbf{u}^k - \mathbf{u}^*\|_H^2 - \|\mathbf{u}^k - \mathbf{u}^{k+1}\|_G^2, \quad \forall\, \mathbf{u}^* \in \Omega^*, \tag{3.1}$$

where

$$H = \begin{pmatrix} \mathcal{D} + (1-\tau)\beta\mathcal{A}^T\mathcal{A} & 0 \\ 0 & \frac{1}{\gamma\beta} I_\ell \end{pmatrix}, \qquad G = \begin{pmatrix} \mathcal{D} & 0 \\ 0 & \frac{4\tau-\gamma-2}{\gamma^2\beta} I_\ell \end{pmatrix}. \tag{3.2}$$

It is clear that the matrices H and G defined in (3.2) are positive definite for all $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$. Thus, the assertion (3.1) means the strict contraction of the sequence $\{\mathbf{u}^k\}$ generated by the IDP-ALM (1.15). Hence, with the assertion (3.1), it is easy to prove the convergence for the IDP-ALM (1.15). The rest of this section is focused on the proof of Theorem 3.1.

3.2 *Variational inequality characterization of* (1.15)

First of all we characterize the iterative scheme (1.15) by a variational inequality.

Since the objective function of the primal subproblem (1.15a) is

$$\mathcal{L}_\beta(\mathbf{x}, \lambda^k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathcal{D}_0}^2 = \boldsymbol{\theta}(\mathbf{x}) - (\lambda^k)^T(\mathcal{A}\mathbf{x} - b) + \frac{\beta}{2}\|\mathcal{A}\mathbf{x} - b\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathcal{D}_0}^2$$

it follows from Lemma 2.1 that $\mathbf{x}^{k+1} \in \mathcal{X}$ and it satisfies the variational inequality

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^{k+1}) + (\mathbf{x} - \mathbf{x}^{k+1})^T\{-\mathcal{A}^T\lambda^k + \beta\mathcal{A}^T(\mathcal{A}\mathbf{x}^{k+1} - b) + \mathcal{D}_0(\mathbf{x}^{k+1} - \mathbf{x}^k)\} \geq 0,\ \forall\, \mathbf{x} \in \mathcal{X}.$$

Then, the variational inequality characterization of the $(k+1)$th iterate generated by the IDP-ALM (1.15) can be summarized as the following lemma.

LEMMA 3.2 For given $\mathbf{u}^k = (\mathbf{x}^k, \lambda^k)$, $\mathbf{u}^{k+1} = (\mathbf{x}^{k+1}, \lambda^{k+1})$ is the output of the IDP-ALM (1.15) if and only if it satisfies

$$\begin{cases} \mathbf{x}^{k+1} \in \mathcal{X}, \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^{k+1}) + (\mathbf{x} - \mathbf{x}^{k+1})^T \\ \qquad \big\{-\mathcal{A}^T\lambda^k + \beta\mathcal{A}^T(\mathcal{A}\mathbf{x}^{k+1} - b) + \mathcal{D}_0(\mathbf{x}^{k+1} - \mathbf{x}^k)\big\} \geq 0,\ \forall\, \mathbf{x} \in \mathcal{X}, & \text{(3.3a)} \\ \lambda^{k+1} = \lambda^k - \gamma\beta(\mathcal{A}\mathbf{x}^{k+1} - b). & \text{(3.3b)} \end{cases}$$

3.3 *A prediction–correction expression*

We show that the IDP-ALM (1.15) can be expressed by a prediction–correction framework. This prediction–correction explanation is only for the convenience of theoretical analysis, and there is no need to follow this prediction–correction framework to implement the scheme (1.15) practically. For this purpose we define the artificial vector $\tilde{\mathbf{u}}^k = (\tilde{\mathbf{x}}^k, \tilde{\lambda}^k)$ by

$$\tilde{\mathbf{x}}^k = \mathbf{x}^{k+1} \quad \text{and} \quad \tilde{\lambda}^k = \lambda^k - \beta(\mathcal{A}\tilde{\mathbf{x}}^k - b), \tag{3.4}$$

where $\mathbf{x}^{k+1}$ is generated by (1.15a) with the given iterate $(\mathbf{x}^k, \lambda^k)$.

Using (3.4) the variational inequality (3.3a) can be written as

$$\tilde{\mathbf{x}}^k \in \mathcal{X},\ \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\tilde{\mathbf{x}}^k) + (\mathbf{x} - \tilde{\mathbf{x}}^k)^T\{-\mathcal{A}^T\tilde{\lambda}^k + \mathcal{D}_0(\tilde{\mathbf{x}}^k - \mathbf{x}^k)\} \geq 0,\ \forall\, \mathbf{x} \in \mathcal{X}. \tag{3.5a}$$

Notice that the equality $\tilde{\lambda}^k = \lambda^k - \beta(\mathcal{A}\tilde{\mathbf{x}}^k - b)$ can be written as the variational inequality

$$\tilde{\lambda}^k \in \Re^\ell,\ \ (\lambda - \tilde{\lambda}^k)^T\left\{(\mathcal{A}\tilde{\mathbf{x}}^k - b) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k)\right\} \geq 0,\ \ \forall\, \lambda \in \Re^\ell. \tag{3.5b}$$

Hence, it follows from (2.2) and (3.5) that $\tilde{\mathbf{u}}^k$ defined in (3.4) satisfies the variational inequality

$$\tilde{\mathbf{u}}^k \in \Omega,\ \ \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\tilde{\mathbf{x}}^k) + (\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\tilde{\mathbf{u}}^k) \geq (\mathbf{u} - \tilde{\mathbf{u}}^k)^T Q(\mathbf{u}^k - \tilde{\mathbf{u}}^k), \quad \forall\, \mathbf{u} \in \Omega, \tag{3.6a}$$

where

$$Q = \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{1}{\beta} I_\ell \end{pmatrix}. \tag{3.6b}$$

Further, using the notation (3.4), we have

$$\beta\big(\mathcal{A}\mathbf{x}^{k+1} - b\big) = \beta(\mathcal{A}\tilde{\mathbf{x}}^k - b) = \lambda^k - \tilde{\lambda}^k.$$

Thus, because of (1.15b), we obtain

$$\lambda^{k+1} = \lambda^k - \gamma\beta\big(\mathcal{A}\mathbf{x}^{k+1} - b\big) = \lambda^k - \gamma\big(\lambda^k - \tilde{\lambda}^k\big).$$

The iterate $\mathbf{u}^{k+1}$ generated by the IDP-ALM (1.15) can be viewed as the output by correcting $\tilde{\mathbf{u}}^k$ via the scheme

$$\mathbf{u}^{k+1} = \mathbf{u}^k - M(\mathbf{u}^k - \tilde{\mathbf{u}}^k), \tag{3.7a}$$

where

$$M = \begin{pmatrix} I & 0 \\ 0 & \gamma I_\ell \end{pmatrix}. \tag{3.7b}$$

Therefore, we can explain the iteration of IDP-ALM (1.15) as a prediction–correction framework, whose new iterate is generated by correcting the point $\tilde{\mathbf{u}}^k$ satisfying the variational inequality (3.6).

3.4 *Some matrices*

To proceed the convergence analysis more conveniently we need to define more matrices. First, we define

$$H_0 = \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{1}{\gamma\beta} I_\ell \end{pmatrix}, \tag{3.8}$$

where $\mathcal{D}_0$ is given in (1.13). Obviously, H_0 is symmetric, but not necessarily positive definite. Furthermore, it holds that

$$Q = H_0 M, \tag{3.9}$$

where Q and M are defined in (3.6b) and (3.7b), respectively.

Then, we define a symmetric matrix as

$$G_0 = Q^T + Q - M^T H_0 M. \tag{3.10}$$

Since $H_0 M = Q$, $M^T H_0 M = M^T Q$ and thus

$$M^T H_0 M = \begin{pmatrix} I & 0 \\ 0 & \gamma I_\ell \end{pmatrix} \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{1}{\beta} I_\ell \end{pmatrix} = \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{\gamma}{\beta} I_\ell \end{pmatrix}.$$

Using (3.6b) and the above equation we have

$$\begin{aligned} G_0 &= (Q^T + Q) - M^T H_0 M = \begin{pmatrix} 2\mathcal{D}_0 & 0 \\ 0 & \frac{2}{\beta} I_\ell \end{pmatrix} - \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{\gamma}{\beta} I_\ell \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{D}_0 & 0 \\ 0 & \frac{2-\gamma}{\beta} I_\ell \end{pmatrix}. \end{aligned} \tag{3.11}$$

Obviously, G_0 is not necessarily positive definite because $\mathcal{D}_0$ is not so. Again, we use $\|\mathbf{u}^k - \tilde{\mathbf{u}}^k\|^2_{G_0}$ to denote the term

$$\left(\mathbf{u}^k - \tilde{\mathbf{u}}^k\right)^T G_0 \left(\mathbf{u}^k - \tilde{\mathbf{u}}^k\right),$$

which is not necessarily non-negative.

Moreover, according to (3.11) and (1.13), we have

$$H_0 = \begin{pmatrix} \mathcal{D} - (1-\tau)\beta \mathcal{A}^T A & 0 \\ 0 & \frac{1}{\gamma\beta} I_\ell \end{pmatrix}, \tag{3.12}$$

and

$$G_0 = \begin{pmatrix} \mathcal{D} - (1-\tau)\beta \mathcal{A}^T A & 0 \\ 0 & \frac{2-\gamma}{\beta} I_\ell \end{pmatrix}, \tag{3.13}$$

where $\mathcal{D} \in \Re^{n\times n}$ is an arbitrarily given positive-definite matrix.

3.5 *Some inequalities*

Then, based on the prediction–correction explanation in the last subsection, we prove several lemmas and theorems to prepare for the convergence proof for the IDP-ALM (1.15).

THEOREM 3.3 Let $\{\mathbf{u}^k\}$ be the sequence generated by the IDP-ALM (1.15) for the problem (1.1) and $\tilde{\mathbf{u}}^k$ be defined by (3.4). Then we have $\tilde{\mathbf{u}}^k \in \Omega$ and

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\tilde{\mathbf{x}}^k) + (\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\mathbf{u}) \geq \frac{1}{2}\big(\|\mathbf{u} - \mathbf{u}^{k+1}\|_{H_0}^2 - \|\mathbf{u} - \mathbf{u}^k\|_{H_0}^2\big) + \frac{1}{2}\|\mathbf{u}^k - \tilde{\mathbf{u}}^k\|_{G_0}^2, \ \forall\, \mathbf{u} \in \Omega, \quad (3.14)$$

where G_0 is defined in (3.10).

Proof. Recall that $(\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\tilde{\mathbf{u}}^k) = (\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\mathbf{u})$ (see (2.3)). The left-hand side of (3.6a) equals

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\tilde{\mathbf{x}}^k) + (\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\mathbf{u}).$$

Using $Q = H_0 M$ (see (3.9)) and the relation (3.7a) the right-hand side of (3.6a) can be written as

$$(\mathbf{u} - \tilde{\mathbf{u}}^k)^T H_0 (\mathbf{u}^k - \mathbf{u}^{k+1}),$$

and hence we have

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\tilde{\mathbf{x}}^k) + (\mathbf{u} - \tilde{\mathbf{u}}^k)^T F(\mathbf{u}) \geq (\mathbf{u} - \tilde{\mathbf{u}}^k)^T H_0 (\mathbf{u}^k - \mathbf{u}^{k+1}), \ \forall\, \mathbf{u} \in \Omega. \quad (3.15)$$

Applying the identity

$$(a-b)^T H_0 (c-d) = \frac{1}{2}\Big\{\|a-d\|_{H_0}^2 - \|a-c\|_{H_0}^2\Big\} + \frac{1}{2}\Big\{\|c-b\|_{H_0}^2 - \|d-b\|_{H_0}^2\Big\},$$

to the right-hand side of (3.15) with

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

we thus obtain

$$\begin{aligned}&(\mathbf{u} - \tilde{\mathbf{u}}^k)^T H_0 (\mathbf{u}^k - \mathbf{u}^{k+1}) \\ &= \frac{1}{2}\big(\|\mathbf{u} - \mathbf{u}^{k+1}\|_{H_0}^2 - \|\mathbf{u} - \mathbf{u}^k\|_{H_0}^2\big) + \frac{1}{2}\big(\|\mathbf{u}^k - \tilde{\mathbf{u}}^k\|_{H_0}^2 - \|\mathbf{u}^{k+1} - \tilde{\mathbf{u}}^k\|_{H_0}^2\big). \quad (3.16)\end{aligned}$$

For the last term of the right-hand side of (3.16) we have

$$\begin{aligned}
&\|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{H_0}^2-\|\mathbf{u}^{k+1}-\tilde{\mathbf{u}}^k\|_{H_0}^2\\
&\quad= \|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{H_0}^2-\|(\mathbf{u}^k-\tilde{\mathbf{u}}^k)-(\mathbf{u}^k-\mathbf{u}^{k+1})\|_{H_0}^2\\
&\quad\overset{(3.7a)}{=} \|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{H_0}^2-\|(\mathbf{u}^k-\tilde{\mathbf{u}}^k)-M(\mathbf{u}^k-\tilde{\mathbf{u}}^k)\|_{H_0}^2\\
&\quad= 2(\mathbf{u}^k-\tilde{\mathbf{u}}^k)^TH_0M(\mathbf{u}^k-\tilde{\mathbf{u}}^k)-(\mathbf{u}^k-\tilde{\mathbf{u}}^k)^TM^THM(\mathbf{u}^k-\tilde{\mathbf{u}}^k)\\
&\quad= (\mathbf{u}^k-\tilde{\mathbf{u}}^k)^T(Q^T+Q-M^TH_0M)(\mathbf{u}^k-\tilde{\mathbf{u}}^k)\\
&\quad\overset{(3.10)}{=} \|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{G_0}^2. \qquad (3.17)
\end{aligned}$$

Substituting (3.16) and (3.17) into (3.15) the assertion of this theorem is proved. □

LEMMA 3.4 Let $\{\mathbf{u}^k\}$ be the sequence generated by the IDP-ALM (1.15) for the problem (1.1) and $\tilde{\mathbf{u}}^k$ be defined by (3.4). Then we have

$$\|\mathbf{u}^{k+1}-\mathbf{u}^*\|_{H_0}^2\le\|\mathbf{u}^k-\mathbf{u}^*\|_{H_0}^2-\|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{G_0}^2. \qquad (3.18)$$

Proof. Setting $\mathbf{u}$ in (3.14) as an arbitrarily fixed $\mathbf{u}^*\in\Omega^*$ we get

$$\begin{aligned}
&\|\mathbf{u}^k-\mathbf{u}^*\|_{H_0}^2-\|\mathbf{u}^{k+1}-\mathbf{u}^*\|_{H_0}^2-\|\mathbf{u}^k-\tilde{\mathbf{u}}^k\|_{G_0}^2\\
&\quad\ge 2\big(\boldsymbol{\theta}(\tilde{\mathbf{x}}^k)-\boldsymbol{\theta}(\mathbf{x}^*)+(\tilde{\mathbf{u}}^k-\mathbf{u}^*)^TF(\mathbf{u}^*)\big),\quad\forall\,\mathbf{u}^*\in\Omega^*.
\end{aligned}$$

Because of the optimality, the right-hand side of the last inequality is non-negative and the lemma is proved. □

THEOREM 3.5 Let $\{\mathbf{u}^k\}$ be the sequence generated by the IDP-ALM (1.15) for the problem (1.1). Then, for any $\tau\in(0,1)$, we have

$$\begin{aligned}
&\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathcal{D}}^2+(1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1}-b\|^2+\frac{1}{\gamma\beta}\|\lambda^{k+1}-\lambda^*\|^2\\
&\le\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathcal{D}}^2+(1-\tau)\beta\|\mathcal{A}\mathbf{x}^k-b\|^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^*\|^2\\
&\quad-\Big(\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathcal{D}}^2+(4\tau-\gamma-2)\beta\|\mathcal{A}\mathbf{x}^{k+1}-b\|^2\Big). \qquad (3.19)
\end{aligned}$$

Proof. According to (3.18) and the structure of H_0 and G_0 we have

$$\begin{aligned}
&\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_{\mathcal{D}}^2-(1-\tau)\beta\|\mathcal{A}(\mathbf{x}^{k+1}-\mathbf{x}^*)\|^2+\frac{1}{\gamma\beta}\|\lambda^{k+1}-\lambda^*\|^2\\
&\le\|\mathbf{x}^k-\mathbf{x}^*\|_{\mathcal{D}}^2-(1-\tau)\beta\|\mathcal{A}(\mathbf{x}^k-\mathbf{x}^*)\|^2+\frac{1}{\gamma\beta}\|\lambda^k-\lambda^*\|^2\\
&\quad-\Big(\|\mathbf{x}^k-\tilde{\mathbf{x}}^k\|_{\mathcal{D}}^2-(1-\tau)\beta\|\mathcal{A}(\mathbf{x}^k-\tilde{\mathbf{x}}^k)\|^2+\frac{2-\gamma}{\beta}\|\lambda^k-\tilde{\lambda}^k\|^2\Big).
\end{aligned}$$

Because $\tilde{\mathbf{x}}^k = \mathbf{x}^{k+1}$, $\mathcal{A}\mathbf{x}^* = b$ and $\lambda^k - \tilde{\lambda}^k = \beta(\mathcal{A}\mathbf{x}^{k+1} - b)$ (see (3.4)) it follows that

$$\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathcal{D}}^2 - (1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^{k+1} - \lambda^*\|^2 \\
\leq\ & \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathcal{D}}^2 - (1-\tau)\beta\|\mathcal{A}\mathbf{x}^k - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^k - \lambda^*\|^2 \\
&-\Big(\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{D}}^2 + (2-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2\Big) + (1-\tau)\beta\|\mathcal{A}(\mathbf{x}^k - \mathbf{x}^{k+1})\|^2. \qquad (3.20)
\end{aligned}$$

Using the inequality $\|\xi - \eta\|^2 \leq 2\|\xi\|^2 + 2\|\eta\|^2$ with $\xi = \mathcal{A}\mathbf{x}^k - b$ and $\eta = \mathcal{A}\mathbf{x}^{k+1} - b$ we get

$$\|\mathcal{A}(\mathbf{x}^k - \mathbf{x}^{k+1})\|^2 \leq 2\|\mathcal{A}\mathbf{x}^k - b\|^2 + 2\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2.$$

Substituting it into the right-hand side of (3.20) we obtain

$$\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathcal{D}}^2 - (1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^{k+1} - \lambda^*\|^2 \\
\leq\ & \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathcal{D}}^2 + (1-\tau)\beta\|\mathcal{A}\mathbf{x}^k - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^k - \lambda^*\|^2 \\
&-\Big(\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{D}}^2 + (2-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1} - \boldsymbol{b}\|^2\Big) + 2(1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - \boldsymbol{b}\|^2.
\end{aligned}$$

Adding the term $2(1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2$ to both sides of the above inequality we get

$$\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathcal{D}}^2 + (1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^{k+1} - \lambda^*\|^2 \\
\leq\ & \Big(\|\mathbf{x}^k - \mathbf{x}^*\|_{\mathcal{D}}^2 + (1-\tau)\beta\|\mathcal{A}\mathbf{x}^k - b\|^2 + \frac{1}{\gamma\beta}\|\lambda^k - \lambda^*\|^2\Big) \\
&-\Big(\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{D}}^2 + (2-\gamma)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2\Big) + 4(1-\tau)\beta\|\mathcal{A}\mathbf{x}^{k+1} - b\|^2.
\end{aligned}$$

The assertion (3.19) follows from the last inequality immediately. □

3.6 *Convergence proof*

Now we are ready to prove Theorem 3.1 that essentially implies the convergence of the IDP-ALM (1.15).

Proof of Theorem 3.1. According to $b = \mathcal{A}\mathbf{x}^*$ and $\mathcal{A}\mathbf{x}^{k+1} - b = \frac{1}{\gamma\beta}(\lambda^k - \lambda^{k+1})$, (3.19) can be written as

$$\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_{\mathcal{D}}^2 + (1-\tau)\beta\|\mathcal{A}(\mathbf{x}^{k+1} - \mathbf{x}^*)\|^2 + \frac{1}{\gamma\beta}\|\lambda^{k+1} - \lambda^*\|^2 \\
\leq\ & \|\mathbf{x}^k - \mathbf{x}^*\|_{\mathcal{D}}^2 + (1-\tau)\beta\|\mathcal{A}(\mathbf{x}^k - \mathbf{x}^*)\|^2 + \frac{1}{\gamma\beta}\|\lambda^k - \lambda^*\|^2 \\
&-\Big(\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{D}}^2 + \frac{4\tau - \gamma - 2}{\gamma^2\beta}\|\lambda^k - \lambda^{k+1}\|^2\Big).
\end{aligned}$$

Recall the definitions of the matrices H and G in (3.2). The assertion (3.1) follows immediately. □

Recall that the positive definiteness of the matrices H and G defined in (3.2) is ensured if $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$. Now, based on the key inequality (3.1) in Theorem 3.1, we can prove the convergence theorem for the IDP-ALM (1.15).

THEOREM 3.6 Let $\{\mathbf{u}^k\}$ be the sequence generated by the IDP-ALM (1.15) for the problem (1.1). Then, for any $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$, the sequence $\{\mathbf{u}^k\}$ converges to a $\mathbf{u}^\infty$, which is a solution point of the variational inequality (2.2).

Proof. First, it follows from (3.1) that

$$\sum_{k=1}^{\infty} \|\mathbf{u}^k - \mathbf{u}^{k+1}\|_G^2 \leq \|\mathbf{u}^0 - \mathbf{u}^*\|_H^2, \quad \forall\, \mathbf{u}^* \in \Omega^*.$$

Consequently, we have

$$\lim_{k\to\infty} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|_{\mathcal{D}} = 0 \qquad \text{and} \qquad \lim_{k\to\infty} \|\lambda^k - \lambda^{k+1}\| = 0. \tag{3.21}$$

For any fixed $\mathbf{u}^* \in \Omega^*$ and $k \geq 1$ we have

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\|_H^2 \leq \|\mathbf{u}^0 - \mathbf{u}^*\|_H^2 \tag{3.22}$$

and thus $\{\mathbf{u}^k\}$ in a bounded set. Let $\mathbf{u}^\infty$ be a cluster point of $\{\mathbf{u}^k\}$ and $\{\mathbf{u}^{k_j}\}$ be the subsequence converging to $\mathbf{u}^\infty$. According to (3.15) $\mathbf{u}^\infty$ is a solution point of the variational inequality (2.2). Since $\mathbf{u}^\infty$ is a solution point it follows from (3.1) that

$$\|\mathbf{u}^{k+1} - \mathbf{u}^\infty\|_H^2 \;\leq\; \|\mathbf{u}^k - \mathbf{u}^\infty\|_H^2 - \|\mathbf{u}^k - \mathbf{u}^{k+1}\|_G^2. \tag{3.23}$$

Note that $\mathbf{u}^\infty$ is also the limit point of $\{\mathbf{u}^{k_j}\}$. Together with (3.21) it is impossible for the sequence $\{\mathbf{u}^k\}$ to have more than one cluster point. Thus the sequence $\{\mathbf{u}^k\}$ converges to $\mathbf{u}^\infty$ and the proof is complete. □

4. Optimality of the formula (1.16)

Recall we relate the step size parameters τ and γ by the formula (1.16) for the IDP-ALM (1.15). That is, for any given $\gamma \in (0, 2)$, τ is required to be $\frac{2+\gamma}{4} < \tau < 1$. It is interesting to ask whether or not the lower bound $\frac{2+\gamma}{4}$ is optimal (smallest), especially given the preference of seeking values of τ as small as possible. In this section we show that the lower bound $\frac{2+\gamma}{4}$ is optimal and it is not possible to find a lower bound of τ smaller than $\frac{2+\gamma}{4}$.

Let us consider the simplest equation $x = 0$ in $\Re$ and show that the IDP-ALM (1.15) is not necessarily convergent when $\tau < \frac{2+\gamma}{4}$. Obviously, $x = 0$ is a special case of the model (1.1) as

$$\min\{0 \cdot x \mid x = 0,\ x \in \Re\}. \tag{4.1}$$

Without loss of generality we take $\beta = 1$ and thus the augmented Lagrangian function of the problem (4.1) is

$$\mathcal{L}(x, \lambda) = -\lambda x + \frac{1}{2}\|x\|^2.$$

The iterative scheme of the IDP-ALM (1.15) for (4.1) is

$$\begin{cases} x^{k+1} = \arg\min\{-x\lambda^k + \frac{1}{2}\|x\|^2 + \frac{1}{2}\|x - x^k\|^2_{\mathcal{D}_0} \mid x \in \Re\}, & \text{(4.2a)} \\ \lambda^{k+1} = \lambda^k - \gamma\, x^{k+1}. & \text{(4.2b)} \end{cases}$$

Since $\beta = 1$ and $\mathcal{A}^T\mathcal{A} = 1$ it follows from (1.13) that

$$\mathcal{D}_0 = \mathcal{D} - (1 - \tau).$$

Let us take $\mathcal{D} = \delta,\ \forall\, \delta > 0$. We thus have $\mathcal{D}_0 = (\delta + \tau) - 1$ and the recursion (4.2) becomes

$$\begin{cases} -\lambda^k + x^{k+1} + ((\delta + \tau) - 1)(x^{k+1} - x^k) = 0, \\ \lambda^{k+1} = \lambda^k - \gamma\, x^{k+1}. \end{cases} \tag{4.3}$$

We thus just need to study the iterative sequence $\{\mathbf{u}^k = (x^k, \lambda^k)\}$. For any given $\tau < \frac{2+\gamma}{4}$ there exists $\delta > 0$ such that $\delta + \tau < \left(\frac{2+\gamma}{4}\right)$ holds. Setting $\alpha = \delta + \tau$ the iterative scheme for $u = (x, \lambda)$ can be written as

$$\begin{cases} \alpha x^{k+1} = \lambda^k + (\alpha - 1)x^k, \\ \lambda^{k+1} = \lambda^k - \gamma\, x^{k+1}. \end{cases} \tag{4.4}$$

With elementary manipulations we get

$$\begin{cases} x^{k+1} = \dfrac{\alpha - 1}{\alpha}x^k + \dfrac{1}{\alpha}\lambda^k, \\ \lambda^{k+1} = \dfrac{\gamma(1 - \alpha)}{\alpha}x^k + \dfrac{\alpha - \gamma}{\alpha}\lambda^k, \end{cases} \tag{4.5}$$

which can be written as

$$\mathbf{u}^{k+1} = P(\alpha)\mathbf{u}^k \qquad \text{with} \qquad P(\alpha) = \frac{1}{\alpha}\begin{pmatrix} \alpha - 1 & 1 \\ \gamma(1 - \alpha) & \alpha - \gamma \end{pmatrix}. \tag{4.6}$$

To find the eigenvalues of the matrix $P(\alpha)$ let $\det(P(\alpha) - \lambda I) = 0$ and we have (see, e.g., pp. 262–263 in Strang (2006))

$$\begin{aligned} \det(P(\alpha) - \lambda I) &= \left(1 - \frac{1}{\alpha} - \lambda\right)\left(1 - \frac{\gamma}{\alpha} - \lambda\right) - \gamma\left(\frac{1}{\alpha^2} - \frac{1}{\alpha}\right) \\ &= \lambda^2 + \frac{1}{\alpha}(\gamma + 1 - 2\alpha)\lambda + 1 - \frac{1}{\alpha} = 0. \end{aligned}$$

Let $f_1(\alpha)$ and $f_2(\alpha)$ be the two eigenvalues of the matrix $P(\alpha)$. Then we have

$$f_1(\alpha) = \frac{(2\alpha - 1 - \gamma) + \sqrt{(1+\gamma)^2 - 4\gamma\alpha}}{2\alpha},$$

and

$$f_2(\alpha) = \frac{(2\alpha - 1 - \gamma) - \sqrt{(1+\gamma)^2 - 4\gamma\alpha}}{2\alpha}.$$

For the function $f_2(\alpha)$ we have

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

and

$$\begin{aligned} f_2'(\alpha) &= \frac{1}{4\alpha^2}\left(\left(2 - \frac{-4\gamma}{2\sqrt{(1+\gamma)^2 - 4\gamma\alpha}}\right)2\alpha - 2\left((2\alpha - 1 - \gamma) - \sqrt{(1+\gamma)^2 - 4\gamma\alpha}\right)\right) \\ &= \frac{1}{4\alpha^2}\left(\frac{4\gamma\alpha}{\sqrt{(1+\gamma)^2 - 4\gamma\alpha}} + 2(1+\gamma) + 2\sqrt{(1+\gamma)^2 - 4\gamma\alpha}\right). \end{aligned}$$

For any $\gamma > 0$ and $\alpha \in \left(0, \frac{2+\gamma}{4}\right)$ we have $(1+\gamma)^2 - 4\gamma\alpha > 0$, and $f_2'(\alpha) > 0$. Consequently, it follows that

$$f_2(\alpha) < f_2\left(\frac{2+\gamma}{4}\right) = -1, \quad \forall\, \alpha \in \left(0, \frac{2+\gamma}{4}\right).$$

That is, for any $\alpha \in \left(0, \frac{2+\gamma}{4}\right)$, the matrix $P(\alpha)$ in (4.6) has an eigenvalue less than -1. Hence, the iterative scheme (4.5), i.e., the application of the IDP-ALM (1.15) to the problem (4.1), is not necessarily convergent for any $\tau \in \left(0, \frac{2+\gamma}{4}\right)$. Thus, $\frac{2+\gamma}{4}$ is the smallest lower bound for τ to ensure the convergence of the IDP-ALM (1.15).

5. Application to full Jacobian splitting

In this section we focus on the multi-block separable case of (1.1) whose objective function is the sum of finitely many additive function components without coupled variables and discuss how to apply our previous analysis to the full Jacobian splitting version of ALM, and consequently improve the result in our previous work (He *et al.*, 2016c).

5.1 *Multi-block model*

When concrete applications are considered the abstract model (1.1) can often be specified as the multi-block separable case, where the objective function can be expressed as the sum of m ($m \geq 2$) additive

function components without coupled variables

$$\begin{array}{ll}\min & \sum_{i=1}^{m}\theta_i(x_i)\\ & \sum_{i=1}^{m}A_ix_i=b;\\ & x_i\in X_i,\quad i=1,\cdots,m,\end{array}\tag{5.1}$$

where $\theta_i:\Re^{n_i}\to\Re$ $(i=1,\ldots,m)$ are closed proper convex functions; $X_i\subseteq\Re^{n_i}$ $(i=1,\ldots,m)$ are closed convex sets; $A_i\in\Re^{\ell\times n_i}$ $(i=1,\ldots,m)$ are given matrices; $b\in\Re^{\ell}$ is a given vector and $\sum_{i=1}^{m}n_i=n$. As (1.1) the solution set of (5.1) is assumed to be nonempty.

5.2 *Direct application of the ALM* (1.4)

Note that the multi-block separable model (5.1) corresponds to the generic model (1.1) with the following specifications:

$$\mathbf{x}=\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_m\end{pmatrix},\quad \boldsymbol{\theta}(\mathbf{x})=\sum_{i=1}^{m}\theta_i(x_i),\tag{5.2a}$$

and

$$\mathcal{A}=(A_1,A_2,\ldots,A_m),\qquad \mathcal{X}=X_1\times X_2\times\cdots\times X_m.\tag{5.2b}$$

Accordingly, the augmented Lagrangian function (1.3) can be specified as

$$\mathcal{L}_\beta(x_1,\ldots,x_m,\lambda)=\sum_{i=1}^{m}\theta_i(x_i)-\lambda^T\left(\sum_{i=1}^{m}A_ix_i-b\right)+\frac{\beta}{2}\left\|\sum_{i=1}^{m}A_ix_i-b\right\|^2.\tag{5.3}$$

Moreover, the generic ALM scheme (1.4), if applied straightforwardly to the well-structured form (5.1), is specified as

$$\left\{\begin{array}{rcl}\left(x_1^{k+1},\ldots,x_m^{k+1}\right) &=& \arg\min\left\{\mathcal{L}_\beta(x_1,\ldots,x_m,\lambda^k)\,\middle|\,x_i\in X_i,\ i=1,\cdots,m\right\},\\ \lambda^{k+1} &=& \lambda^k-\beta\left(\sum_{i=1}^{m}A_ix_i^{k+1}-b\right).\end{array}\right.\tag{5.4}$$

5.3 *Splitting versions of the ALM suitable for* (5.1)

The implementation of the direct application of the ALM (5.4), however, is usually not preferable, because the resulting primal problem in (5.4) has a complicated objective function with highly correlated variables and the function components are not treated individually. A useful strategy to improve the implementability of ALM for the separable case (5.1) is to split the primal ALM subproblem in (5.4), in either Jacobian or Gauss–Seidel manner. The resulting iterative schemes are featured by that only one function component is involved in each decomposed subproblems, exactly like the well-studied incremental type methods in, e.g., Bertsekas (2011, 2015). This line of research, which could be called augmented Lagrangian–based splitting algorithms, has gained much attention from the community.

Particularly, the mentioned ADMM originally proposed in Glowinski & Marrocco (1975) is such a case for (5.1) with $m = 2$ and the primal subproblem in (5.4) is decomposed in the Gauss–Seidel manner. Later, it is shown in Chen *et al.* (2016) that the Gauss–Seidel decomposition cannot be straightforwardly extended to the case of (5.1) with $m \geq 3$; thus, specific strategies are required to design Gauss–Seidel type augmented Lagrangian–based splitting algorithms for (5.1) with $m \geq 3$, see our previous work (He *et al.*, 2012, 2015a,b, 2017) for instances.

5.4 *Full Jacobian splitting versions of ALM*

On the other hand, it is interesting to consider decomposing the primal ALM subproblem in (5.4) in Jacobian manner so that the resulting subproblems can be solved in parallel. More precisely, applying the full Jacobian splitting to the primal subproblem in (5.4), we obtain the scheme

$$\begin{cases} \begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1\}; \\ \qquad\qquad \vdots \\ x_i^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \mid x_i \in X_i\}; \\ \qquad\qquad \vdots \\ x_m^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) \mid x_m \in X_m\}; \end{cases} & \text{(5.5a)} \\ \lambda^{k+1} = \lambda^k - \beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big). & \text{(5.5b)} \end{cases}$$

We call (5.5) the full Jacobian splitting version of ALM for the multi-block separable convex minimization model (5.1). It enjoys the feature that all the x_i-subproblems can be solved in parallel, and this is an important feature when large- or huge-scale data is under consideration and parallel computing infrastructures are available.

However, it is shown in He *et al.* (2015a) that the convergence of (5.5) is not guaranteed even for the case of (5.1) with $m = 2$. Therefore, despite the favorable feature eligible for parallel computation, the scheme (5.5) should be meticulously modified to guarantee the convergence. In our previous work (He *et al.*, 2016c) (see also Deng *et al.* (2017)) it is suggested to regularize all the decomposed subproblems over primal variables by sufficiently large proximal terms

$$\begin{cases} \begin{cases} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \mid x_1 \in X_1\}; \\ \qquad\qquad \vdots \\ x_i^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_i(x_i - x_i^k)\|^2 \mid x_i \in X_i\}; \\ \qquad\qquad \vdots \\ x_m^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) + \frac{s\beta}{2}\|A_m(x_m - x_m^k)\|^2 \mid x_m \in X_m\}; \end{cases} & \text{(5.6a)} \\ \lambda^{k+1} = \lambda^k - \beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big), & \text{(5.6b)} \end{cases}$$

in which $s > 0$ is the proximal parameter. We call (5.6) the proximally regularized full Jacobian splitting version of ALM. The convergence of (5.6) is proved in He *et al.* (2016c) under the condition of $s \geq m-1$; see Theorem 3.1 therein.

It is easy to see that ignoring some constant terms in the objective functions we can rewrite the proximally regularized full Jacobian splitting version of ALM (5.6) as

$$\begin{cases} \begin{cases} x_1^{k+1} = \arg\min\limits_{x_1\in X_1}\left\{\theta_1(x_1) - (\lambda^k)^T A_1 x_1 + \frac{(s+1)\beta}{2}\|A_1(x_1 - x_1^k) + \frac{1}{s+1}(\mathcal{A}\mathbf{x}^k - b)\|^2\right\};\\ \qquad\vdots\\ x_i^{k+1} = \arg\min\limits_{x_i\in X_i}\left\{\theta_i(x_i) - (\lambda^k)^T A_i x_i + \frac{(s+1)\beta}{2}\|A_i(x_i - x_i^k) + \frac{1}{s+1}(\mathcal{A}\mathbf{x}^k - b)\|^2\right\};\\ \qquad\vdots\\ x_m^{k+1} = \arg\min\limits_{x_m\in X_m}\left\{\theta_m(x_m) - (\lambda^k)^T A_m x_m + \frac{(s+1)\beta}{2}\|A_m(x_m - x_m^k) + \frac{1}{s+1}(\mathcal{A}\mathbf{x}^k - b)\|^2\right\}; \end{cases} & (5.7a)\\ \lambda^{k+1} = \lambda^k - \beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big). & (5.7b) \end{cases}$$

5.5 *Relationship to some existing methods*

In the literature the full Jacobian splitting version of ALM (5.6) has been studied via different perspectives. Here we summarize its relationship to some existing methods.

First, by introducing the auxiliary variable $\boldsymbol{y} = \{y_1, y_2, \ldots, y_m\}$, the model (5.1) can be reformulated as the following two-block separable convex minimization model

$$\begin{aligned} &\min \sum_{i=1}^m \theta_i(x_i),\\ &\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix}\begin{pmatrix} x_1\\ x_2\\ \vdots\\ x_m \end{pmatrix} - \begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_m \end{pmatrix} = 0, \qquad (5.8)\\ &x_i \in X_i,\ (i = 1, 2, \ldots, m), \quad \mathcal{Y} = \left\{\boldsymbol{y} = (y_1, \ldots, y_m) \mid \sum_{i=1}^m y_i = b\right\}. \end{aligned}$$

Then, it was suggested in Wang *et al.* (2015) (see Algorithms 2 and 3 therein) to apply the original ADMM in Glowinski & Marrocco (1975) to the model (5.8) by regarding $\sum_{i=1}^m \theta_i(x_i)$ as the first block of function and the second block is null; $(x_1, x_2, \cdots, x_m)$ is the first block of variable and $\mathcal{Y}$ the second. It is analyzed in He *et al.* (2016c) (see 4.14 therein) that the resulting scheme is exactly the proximally regularized full Jacobian splitting version of ALM (5.6) with $s = m - 1$, or it can be equivalently presented as

$$\begin{cases} x_i^{k+1} = \arg\min\Big\{\theta_i(x_i) - (\lambda^k)^T A_i x_i + \frac{\alpha}{2}\|A_i(x_i - x_i^k) + \frac{1}{m}\big(\sum_{i=1}^m A_i x_i^k - b\big)\|^2 \,\Big|\, x_i \in X_i\Big\},\\ \qquad i = 1, \cdots, m, & (5.9a)\\ \lambda^{k+1} = \lambda^k - \frac{\alpha}{m}\big(\sum_{i=1}^m A_i x_i^{k+1} - b\big), & (5.9b) \end{cases}$$

where $\alpha > 0$ is a parameter playing the penalty role for the linear constraints in (5.8).

Later, it is pointed out in Bertsekas (2015) that the scheme (5.9) is essentially the same as the earlier scheme proposed in Bertsekas & Tsitsiklis (1989) (also see (2.33)–(2.34) in Bertsekas (2015)) with a notation difference of the Lagrange multiplier. Let us recall the scheme in Bertsekas & Tsitsiklis (1989):

$$\begin{cases} x_i^{k+1} = \arg\min\left\{\theta_i(x_i) + (\lambda^k)^T A_i x_i + \frac{\alpha}{2}\left\|A_i(x_i - x_i^k) + \frac{1}{m}\left(\sum_{j=1}^m A_j x_j^k - b\right)\right\|^2 \,\middle|\, x_i \in X_i\right\}, \\ \qquad\qquad i = 1, \cdots, m, \end{cases} \tag{5.10a}$$

$$\lambda^{k+1} = \lambda^k + \frac{\alpha}{m}\left(\sum_{i=1}^m A_i x_i^{k+1} - b\right). \tag{5.10b}$$

Indeed, it is clear that the scheme (5.10) corresponds to the special case of Algorithm 8.1 in Gu *et al.* (2014) with $\beta = \frac{\alpha}{m}$ and $\gamma = 1$. It is also easy to see that the proximally regularized full Jacobian splitting version of ALM (5.7) with the special choice of $s = m - 1$ becomes identical with (5.9) if we choose $\alpha = m\beta$.

5.6 *How small could* s *be?*

As mentioned, though the restriction $s \geq m - 1$ is sufficient to ensure the convergence of (5.6), it is preferable to further relax this restriction and thus find smaller lower bounds of s to render larger step sizes for the decomposed subproblems over the primal variables in (5.6). This request is more important when m is larger. Our second purpose is to answer the question of what the smallest value of s is in (5.6) to guarantee the convergence. Indeed, we consider a more general scheme of the proximally regularized full Jacobian splitting version of ALM (5.6) as the following:

$$\begin{cases} x_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \,\middle|\, x_1 \in X_1\right\}; \\ \qquad\qquad \vdots \\ x_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_i(x_i - x_i^k)\| \,\middle|\, x_i \in X_i\right\}; \\ \qquad\qquad \vdots \\ x_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) + \frac{s\beta}{2}\|A_m(x_m - x_m^k)\| \,\middle|\, x_m \in X_m\right\}; \end{cases} \tag{5.11a}$$

$$\lambda^{k+1} = \lambda^k - \gamma\beta\left(\sum_{i=1}^m A_i x_i^{k+1} - b\right), \ \gamma \in (0, 2). \tag{5.11b}$$

Then, we shall show that the step size parameters s and γ in (5.11) can be related by the formula

$$s > \tau m - 1, \qquad \tau \in \left(\frac{2+\gamma}{4}, 1\right), \tag{5.12}$$

which improves the result $s \geq m - 1$ in our previous work (He *et al.*, 2016c) to ensure the convergence. Indeed, the analysis in Section 3 can be exactly used to derive the restriction (5.12) for the scheme (5.11); the detail is provided in the next subsection. Also, because of the result in Section 4, $(2 + \gamma)/4$ is the smallest lower bound of τ to ensure the convergence of (5.11).

5.7 *Convergence proof*

To derive the improved result (5.12) in ensuring the convergence of (5.11) we use the notation in (5.2) and rewrite (5.11) as

$$\begin{cases} \begin{cases} x_1^{k+1} = \arg\min\limits_{x_1\in X_1} \left\{ \begin{array}{c} \theta_1(x_1) - (\lambda^k)^T\big(A_1(x_1 - x_1^k) + (\mathcal{A}\mathbf{x}^k - b)\big) \\ + \frac{\beta}{2}\|A_1(x_1 - x_1^k) + (\mathcal{A}\mathbf{x}^k - b)\|^2 + \frac{s\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \end{array} \right\}; \\ \vdots \\ x_i^{k+1} = \arg\min\limits_{x_i\in X_i} \left\{ \begin{array}{c} \theta_i(x_i) - (\lambda^k)^T\big(A_i(x_i - x_i^k) + (\mathcal{A}\mathbf{x}^k - b)\big) \\ + \frac{\beta}{2}\|A_i(x_i - x_i^k) + (\mathcal{A}\mathbf{x}^k - b)\|^2 + \frac{s\beta}{2}\|A_i(x_i - x_i^k)\|^2 \end{array} \right\}; \\ \vdots \\ x_m^{k+1} = \arg\min\limits_{x_m\in X_m} \left\{ \begin{array}{c} \theta_m(x_m) - (\lambda^k)^T\big(A_m(x_m - x_m^k) + (\mathcal{A}\mathbf{x}^k - b)\big) \\ + \frac{\beta}{2}\|A_m(x_m - x_m^k) + (\mathcal{A}\mathbf{x}^k - b)\|^2 + \frac{s\beta}{2}\|A_m(x_m - x_m^k)\|^2 \end{array} \right\}; \end{cases} \quad (5.13a) \\ \lambda^{k+1} = \lambda^k - \gamma\beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big), \quad \gamma \in (0,2). \quad (5.13b) \end{cases}$$

For each $i \in \{1, 2, \ldots, m\}$, applying Lemma 2.1 to the x_i-subproblem in (5.13a), we get $x_i^{k+1} \in X_i$ and

$$\theta_i(x_i) - \theta_i\big(x_i^{k+1}\big) + \big(x_i - x_i^{k+1}\big)^T \left(\begin{array}{c} -A_i^T\lambda^k + \beta A_i^T\big(A_i(x_i^{k+1} - x_i^k) + (\mathcal{A}\mathbf{x}^k - b)\big) \\ +s\beta A_i^T A_i(x_i^{k+1} - x_i^k) \end{array} \right) \geq 0, \ \forall\, x_i \in X_i,$$

which can be rewritten as

$$x_i^{k+1} \in X_i, \ \theta_i(x_i) - \theta_i\big(x_i^{k+1}\big) + \big(x_i - x_i^{k+1}\big)^T \left(\begin{array}{c} -A_i^T\lambda^k + \beta A_i^T(\mathcal{A}\mathbf{x}^k - b) \\ + (1+s)\beta A_i^T A_i(x_i^{k+1} - x_i^k) \end{array} \right) \geq 0, \ \forall\, x_i \in X_i.$$

Considering the above inequality for $i = 1, 2, \cdots, m$, and using the notation in (5.2), we obtain

$$\mathbf{x}^{k+1} \in \mathcal{X}, \ \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^{k+1}) + (\mathbf{x} - \mathbf{x}^{k+1})^T \left(\begin{array}{c} -\mathcal{A}^T\lambda^k + \beta\mathcal{A}^T(\mathcal{A}\mathbf{x}^k - b) \\ (1+s)\beta\,\mathrm{diag}(\mathcal{A}^T\mathcal{A})(\mathbf{x}^{k+1} - \mathbf{x}^k) \end{array} \right) \geq 0, \ \forall\, \mathbf{x} \in \mathcal{X}, \quad (5.14)$$

where

$$\mathrm{diag}(\mathcal{A}^T\mathcal{A}) = \begin{pmatrix} A_1^TA_1 & 0 & \cdots & 0 \\ 0 & A_2^TA_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_m^TA_m \end{pmatrix}.$$

We summarize the assertion (5.14) in the following lemma.

LEMMA 5.1 For given $\mathbf{u}^k = (\mathbf{x}^k, \lambda^k)$, $\mathbf{u}^{k+1} = (\mathbf{x}^{k+1}, \lambda^{k+1})$ is the output of (5.11) if and only if it satisfies

$$\begin{cases} \mathbf{x}^{k+1} \in \mathcal{X}, \ \boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^{k+1}) + (\mathbf{x} - \mathbf{x}^{k+1})^T \\ \qquad \left\{ -\mathcal{A}^T\lambda^k + \beta\mathcal{A}^T(\mathcal{A}\mathbf{x}^k - b) + (1+s)\beta \mathrm{diag}(\mathcal{A}^T\mathcal{A})(\mathbf{x}^{k+1} - \mathbf{x}^k) \right\} \geq 0, \ \forall\, \mathbf{x} \in \mathcal{X}, & (5.15a) \\ \lambda^{k+1} = \lambda^k - \gamma\beta(\mathcal{A}\mathbf{x}^{k+1} - b). & (5.15b) \end{cases}$$

Hence, for the specific scheme (5.11), the matrix $(1+s)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A})$ in (5.15a) plays the same role as the matrix $\mathcal{D}_0$ in (3.3a). Moreover, notice that (5.15a) in Lemma 5.1 can be written as $\mathbf{x}^{k+1} \in \mathcal{X}$ and

$$\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}(\mathbf{x}^{k+1}) + (\mathbf{x} - \mathbf{x}^{k+1})^T \begin{pmatrix} -\mathcal{A}^T\lambda^k + \beta\mathcal{A}^T(\mathcal{A}\mathbf{x}^{k+1} - b) \\ + \left[(1+s)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A}\right](\mathbf{x}^{k+1} - \mathbf{x}^k) \end{pmatrix} \geq 0, \ \forall\, \mathbf{x} \in \mathcal{X}.$$

Recall that our analysis requires relating the matrix $\mathcal{D}_0$ to a positive-definite matrix $\mathcal{D}$ via the scheme (1.13). Thus, for the specific scheme (5.11) with $\mathcal{D}_0 = (1+s)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \beta\mathcal{A}^T\mathcal{A}$, if we set

$$\mathcal{D} = (1+s)\beta\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \tau\beta\mathcal{A}^T\mathcal{A}$$

then it holds

$$\mathcal{D}_0 = \mathcal{D} - (1-\tau)\beta\mathcal{A}^T\mathcal{A}.$$

Then, according to (1.13), we just need to choose s to guarantee

$$(1+s)\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \tau\mathcal{A}^T\mathcal{A} \succ 0, \quad \text{for } \tau \in \left(\frac{2+\gamma}{4}, 1\right). \tag{5.16}$$

With the assertion (5.16) the remaining part of the proof for the scheme (5.11) is the same as what we have presented in Section 3. Hence, we say the convergence result in this section for (5.11) is an application of the previous analysis in Section 3 for the general IDP-ALM (1.15).

To fulfill (5.16) notice that

$$\begin{aligned} &(1+s)\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \tau\mathcal{A}^T\mathcal{A} \\ &= \mathrm{diag}(\mathcal{A}^T)\left[(1+s)\begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & I & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I \end{pmatrix} - \tau\begin{pmatrix} I & I & \cdots & I \\ I & I & \cdots & I \\ \vdots & \ddots & \ddots & I \\ I & \cdots & I & I \end{pmatrix}\right]\mathrm{diag}(\mathcal{A}), \end{aligned}$$

where

$$\mathrm{diag}(\mathcal{A}^T) = \begin{pmatrix} A_1^T & 0 & \cdots & 0 \\ 0 & A_2^T & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_m^T \end{pmatrix} \quad \text{and} \quad \mathrm{diag}(\mathcal{A}) = \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_m \end{pmatrix}.$$

Thus, when each A_i is assumed to be full column-rank in (5.1), for $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$, we have

$$(1+s)\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \tau\mathcal{A}^T\mathcal{A} \succ 0 \quad \Leftrightarrow \quad (1+s)I_m - \tau e_m e_m^T \succ 0,$$

where e_m is the column vector in $\Re^m$ with all elements as 1. In other words, when each A_i is full column-rank in (5.1) and $\tau \in \left(\frac{2+\gamma}{4}, 1\right)$, we have

$$(1+s)\mathrm{diag}(\mathcal{A}^T\mathcal{A}) - \tau\mathcal{A}^T\mathcal{A} \succ 0 \quad \Leftrightarrow \quad s > \tau m - 1.$$

Hence, the convergence of the sequence $\{(x_1^k, x_2^k, \cdots, x_m^k, \lambda^k)\}$ generated by (5.11) with the step size restriction (5.12) is ensured. In particular, for the scheme (5.6) that corresponds to (5.11) with $\gamma = 1$, its convergence is guaranteed whenever $s > \frac{3}{4}m - 1$, which improves the result $s \geq m - 1$ in our previous work (He *et al.*, 2016c).

5.8 *Remarks*

Meanwhile, it is worth mentioning that if the assumption of the full column-rank of A_i does not hold for (5.1), then it is easy to slightly modify our analysis to establish the convergence of the sequence $\{(A_1x_1^k, A_2x_2^k, \cdots, A_mx_m^k, \lambda^k)\}$, instead of $\{(x_1^k, x_2^k, \cdots, x_m^k, \lambda^k)\}$, for the scheme (5.11) with the step size restriction (5.12). Discussing this different analysis is trivial and it is not the scope of this paper; we thus omit the detail and refer to, e.g., He *et al.* (2015b) for similar analysis.

6. Conclusions

In this paper we study the proximal version of the ALM for convex programming problems and show for the first time that the proximal term can be positive indefinite. This result makes it possible to solve the subproblems by larger step sizes, and hence accelerates the proximal version of ALM. We find the optimal choice of the proximal term in the sense that the convergence being guaranteed while any smaller one being divergent; our result significantly differs from a number of so-called 'semi-proximal' ALM-based methods in the literature that can only relax the proximal term's positive definiteness requirement to positive semi-definiteness under more assumptions on the model *per se*. We also extend this result to the full Jacobian splitting version of the ALM in He *et al.* (2016c) for solving a multi-block separable convex minimization problem, where the objective function is the sum of finitely many additive function components without coupled variables.

Funding

National Natural Science Foundation of China (11871029 to B.H., 11701564 to F.M.); General Research Fund from Hong Kong Research Grants Council (12300317 to X.Y.).

References

Bertsekas, D. P. (1982) *Constrained Optimization and Lagrange Multiplier Methods*. Rheinboldt, W. (ed.). New York: Academic Press.

Bertsekas, D. P. (2011) Incremental proximal methods for large scale convex optimization. *Math. Programming*, **129**, 163–195.

Bertsekas, D. P. (2015) *Incremental aggregated proximal and augmented Lagrangian algorithms*. arXiv: 1509:09257v1.

BERTSEKAS, D. P. & TSITSIKLIS, J. N. (1989) *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall.

BLUM, E. & OETTLI, W. (1975) *Mathematische Optimierung, Econometrics and Operations Research XX*. Berlin: Springer.

CAI, J. F., CANDÈS, E. J. & SHEN, Z. W. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, **20**, 1956–1982.

CHEN, C. H., HE, B. S. & YUAN, X. M. (2012) Matrix completion via an alternating direction method. *IMA J. Numer. Anal.*, **32**, 227–245.

CHEN, C. H., HE, B. S., YE, Y. Y. & YUAN, X. M. (2016) The direct extension of ADMM for multi-block convex minimization problems is not necessary convergent. *Math. Programming*, **155**, 57–79.

DENG, W., LAI, M.-J., PENG, Z. and YIN, W. (2017) Parallel multi-block ADMM with $o(1/k)$ convergence. *J. Sci. Comput.*, **71**, 712–736.

GLOWINSKI, R. (1984) *Numerical Methods for Nonlinear Variational Problems*. Vijayasundaram, G. (ed.). New York, Berlin, Heidelberg, Tokyo: Springer.

GLOWINSKI, R. & MARROCCO, A. (1975) Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problémes non linéaires. *R.A.I.R.O.*, **R2**, 41–76.

GU, G. Y., HE, B. S. & YUAN, X. M. (2014) Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach. *Comput. Optim. Appl.*, **59**, 135–161.

HE, B. S. (2015) PPA-like contraction methods for convex optimization: a framework using variational inequality approach. *J. Oper. Res. Soc. China*, **3**, 391–420.

HE, B. S., HOU, L. S. & YUAN, X. M. (2015a) On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. *SIAM J. Optim.*, **25**, 2274–2312.

HE, B. S., TAO, M. & YUAN, X. M. (2015b) A splitting method for separable convex programming. *IMA J. Numer. Anal.*, **31**, 394–426.

HE, B. S., MA, F. & YUAN, X. M. (2016a) Convergence study on the symmetric version of ADMM with larger step sizes. *SIAM J. Imaging Sci.*, **9**, 1467–1501.

HE, B. S., MA, F. & YUAN, X. M. (2016b) Optimal linearized alternating direction method of multipliers via positive-indefinite proximal regularization for convex programming. Manuscript.

HE, B. S., XU, H. K. & YUAN, X. M. (2016c) On the proximal Jacobian decomposition of ALM for multiple-block separable convex minimization problems and its relationship to ADMM. *J. Sci. Comput.*, **66**, 1204–1217.

HE, B. S., TAO, M. & YUAN, X. M. (2012) Alternating direction method with Gaussian back substitution for separable convex programming. *SIAM J. Optim.*, **22**, 313–340.

HE, B. S., TAO, M. & YUAN, X. M. (2017) Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. *Math. Oper. Res.*, **42**, 662–691.

HE, B. S., YUAN, X. M. & ZHANG, W. X. (2013) A customized proximal point algorithm for convex minimization with linear constraints. *Comput. Optim. Appl.*, **56**, 559–572.

HESTENES, M. R. (1969) Multiplier and gradient methods. *J. Optim. Theory Appl.*, **4**, 303–320.

LARSEN, R. M. (2004) Propack-software for large and sparse SVD calculations [online]. Available at sun.stanford.edu/~rmunk/PROPACK, pp. 2008–2009.

MA, S. Q., GOLDFARB, D. & CHENA, L. F. (2011) Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Programming Ser. A.*, **128**, 321–353.

MARTINET, B. (1970) Regularisation, d'inéquations variationelles par approximations succesives. *Rev. Francaise d'Inform. Recherche Oper.*, **4**, 154–159.

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

ROCKAFELLAR, R. T. (1976a) Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, **1**, 877–898.

Rockafellar, R. T. (1976b) Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, **14**, 877–898.

Starck, J. L., Murtagh, F. & Fadili, J. M. (2010) *Sparse Image and Signal Processing, Wavelets, Curvelets, Morphological Diversity*. Cambridge: Cambridge University Press.

Strang, G. (2006) *Linear Algebra and its Applications*, 4th edn. Independence, Kentucky, USA: Cengage Learning Press.

Tao, M. & Yuan, X. M. (2018) The generalized proximal point algorithm with step size of 2 is not necessarily convergent. *Comput. Optim. Appl.*, **70**, 827–839.

Wang, X., Hong, M., Ma, S. & Luo, Z. Q. (2015) Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers. *Pacific J. Optim.*, **11**, 645–667.

Yang, J. F. & Yuan, X. M. (2013) Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.*, **82**, 301–329.

A. Some preliminary numerical results

The exclusive focus of this paper is the finding of the optimal choice of the proximal term for the IDP-ALM (1.15) and the corresponding rigorous theory for the convergence. Mathematically, these theoretical results close the discussion of finding the optimal proximal terms for a class of proximal versions of the ALM (1.4), and numerically, the acceleration by this optimal proximal term can be easily verified by various applications studied in the literature. Indeed, by just multiplying the proximal parameter with the constant 0.75, the efficiency of the proximal version of the ALM can be immediately improved. In the literature there are already a huge volume of references affirmatively showing the efficiency of the ALM and its various splitting versions for specific applications of both the canonical model (1.1) and the separable model (5.1). For completeness, below we just provide some preliminary numerical results that can show the efficiency of the IDP-ALM (1.15). Our primary purpose is numerically showing the immediate improvement by simply reducing the proximal parameter in the IDP-ALM (1.15). It is worthwhile mentioning that it is not reasonable to expect a dramatic improvement by this nearly computationally free modification of the step size. But the theory found in this paper is generally applicable to a wide range of applications. Also, instead of comparing the IDP-ALM (1.15) with a number of other algorithms, we just use the well-tested linearized ALM (1.10) as a benchmark because its advantage to a number of other algorithms has been well studied in the literature already. We also skip the numerical results for testing the optimality proximal term in the scheme (5.6) because of the similarity and succinctness.

Our codes were written in MATLAB R2015b and implemented in a Lenovo personal computer with a 2.8GHz Intel Core i7 CPU and 8GB memory.

B. Nuclear-norm minimization problem

We first test the nuclear-norm relaxation model of the matrix completion problem

$$\min\{\|X\|_* \mid X_{i,j} = M_{i,j},\ (i,j) \in \Omega\}, \tag{B.1}$$

where $\|X\|_*$ is the nuclear norm of $X \in \Re^{n\times n}$ defining as the sum of all singular values of X, $\Omega = \{(i,j),\quad i,j \in \{1,\cdots,n\}\}$ is an index set with cardinality p, $M \in \Re^{n\times n}$ is the unknown matrix to be completed; $M_{i,j}$ with $(i,j) \in \Omega$ represent the sampled (known) entries of M. In the literature there are many algorithms applicable to the model (B.1), e.g., Cai *et al.* (2010), Ma *et al.* (2011) and Chen *et al.* (2012).

The model (B.1) is a trivial extension of (1.1) with matrix variables; thus, the IDP-ALM (1.15) can be applied. To implement the IDP-ALM (1.15) it is noted that the constraint in (B.1) is given by a

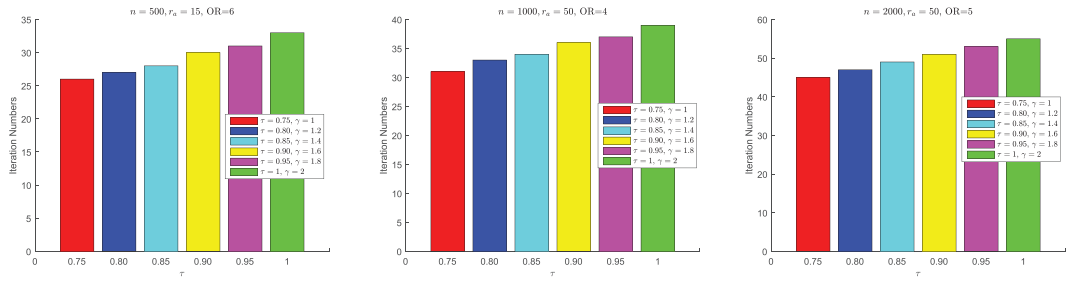


FIG. B1. Iteration numbers of (1.15) with different τ for (B1) with randomly generated data.

projection operator $P_\Omega : \Re^{n\times n} \to \Re^{n\times n}$. Hence, $\|\mathcal{A}^T\mathcal{A}\| = 1$ because the mapping $\mathcal{A}$ reduces to the projection operator P_Ω in (B.1). Note that for the resulting subproblem at each iteration of the IDP-ALM (1.15) its closed-form solution is given by

$$\left(I + \frac{1}{r}\partial(\|X\|_*)\right)^{-1}(C) := U\mathrm{diag}\left(\max\left\{\sigma - \frac{1}{r}\mathbf{1}_{r_a}, \mathbf{0}_{r_a}\right\}\right)V^T,\ \forall\, C \in \Re^{n\times n},$$

where $U\mathrm{diag}(\sigma)V^T$ denotes the singular value decomposition of the matrix C and r_a is the rank of C. We implement the PROPACK library (Larsen, 2004) in computing the singular values in our experiments.

We follow the popular way in the literature, e.g., Yang & Yuan (2013) and Cai *et al.* (2010), to generate some synthetic data. That is, the solution X^* with rank r_a is generated as a product of $M_L M_R^T$, where $M_L, M_R \in \Re^{n\times r_a}$ are generated to have independent and identically distributed Gaussian entries. The set Ω is sampled uniformly at random from cardinality p. The sampling ratio defined as the ratio between the numbers of samples and entries in the matrix is denoted by 'SR', i.e., SR=p/n^2. The oversampling ratio defined as the ratio between the numbers of samples and the degrees of freedom in the matrix of rank r_a is denoted by 'OR', i.e., OR $= p/r_a(2n - r_a)$. Since $\|P_\Omega^T P_\Omega\| = 1$ and smaller values of r are preferable we set $r = 1.001 * \beta$ for implementing the IDP-ALM (1.15). Moreover, $\beta = 1/7 * \sqrt{n}$ is set throughout. As Cai *et al.* (2010) the stopping criterion is

$$\frac{\|X_\Omega^k - X_\Omega^*\|_F}{\|X_\Omega^*\|_F} \le 10^{-4}, \tag{B.2}$$

where $\|X\|_F$ is the Frobenius norm.

Recall that our focus is verifying the improvement by values of $\tau \in (0.75, 1)$ for the IDP-ALM (1.15). We report the results of the cases where $\tau = \{0.75, 0.8, 0.85, 0.9, 0.95, 1\}$. For each value of τ, according to (1.16), we choose $\gamma = 4\tau - 2$. In Figure B1 we show some comparisons in terms of the iteration number for the IDP-ALM (1.15) with different values of τ. For succinctness three cases of n, r_a and OR are reported. The best performance with $\tau = 0.75$ is demonstrated by these bar graphs, and our theoretical assertion regarding the optimality of $\tau = 0.75$ is verified.

In Table B1 we list the comparison between the optimal choice ($\tau = 0.75, \gamma = 1$) in the IDP-ALM (1.15) and the benchmark linearized ALM (1.10) that corresponds to the scheme (1.15), but with $\tau = 1$. The number of iterations (It.), computing time in seconds (CPU) and relative error (RErr), is compared. According to this table the acceleration of the IDP-ALM (1.15) with the optimal choice of ($\tau = 0.75, \gamma = 1$) is clearly shown.

TABLE B1 *Numerical results of indefinite linearized ALM for (B1)*

$n \times n$ matrix				Linearized ALM			Optimal Proximal ALM ($\tau = 0.75, \gamma = 1$)		
n	r_a	OR	SR	It.	CPU	RErr	It.	CPU	RErr
500	5	6	0.12	92	2.32	9.27e−05	78	1.99	9.02e−05
500	10	5	0.20	56	1.79	9.24e−05	45	1.49	9.54e−05
500	50	3	0.57	29	1.73	9.30e−05	22	1.44	8.69e−05
1000	10	6	0.12	89	17.33	8.82e−05	70	13.76	9.49e−05
1000	50	4	0.39	41	16.00	8.48e−05	31	10.06	8.19e−05
1000	100	3	0.57	34	22.06	9.36e−05	26	15.71	7.97e−05
2000	10	6	0.06	142	92.35	9.95e−05	121	77.08	9.99e−05
2000	50	5	0.25	58	102.00	9.92e−05	44	73.76	9.86e−05
2000	100	4	0.39	47	129.06	9.22e−05	36	94.61	8.05e−05
5000	10	6	0.02	270	1603.84	9.88e−05	244	1349.79	9.54e−05
5000	50	5	0.10	117	2213.15	9.61e−05	89	1788.87	9.76e−05
5000	100	4	0.16	98	3137.25	9.92e−05	75	2192.54	9.84e−05

C. Wavelet-based image inpainting problem

Then, we test a wavelet-based image inpainting problem. Let $x \in \Re^l$ be a clean image and $b \in \Re^l$ an observed image with some missing pixels. For the case where the unknown image x has a sparse representation under a wavelet dictionary $W \in \Re^{l\times n}$, that is, $x = W\mathbf{x}$ where $\mathbf{x}$ is a sparse vector, the following wavelet-based model is widely studied:

$$\min \{\|\mathbf{x}\|_1 \mid BW\mathbf{x} = b\}, \tag{C.1}$$

where $\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$, $B \in \Re^{l\times l}$ is a diagonal matrix whose elements are either 0 or 1 as indices of the pixels (the missing pixels correspond to a value of 0 and the kept pixels correspond to a value of 1). For more details, see, e.g., Starck *et al.* (2010). The model (C.1) is a special case of the model (1.1) with $\mathcal{A} = BW$. For succinctness we only consider the reflective boundary condition. Obviously, since the binary matrix B is diagonal and the dictionary W has the property $WW^T = I$, we also have $\|\mathcal{A}^T\mathcal{A}\| = 1$ when $\mathcal{A}$ reduces to BW in (C.1).

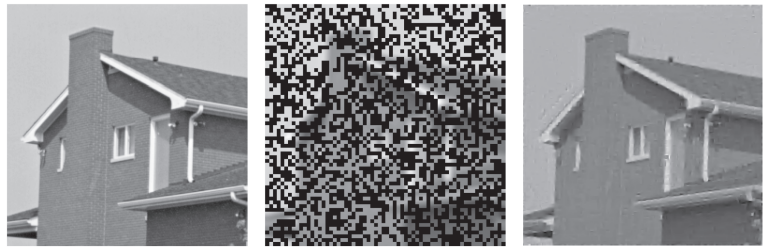
We test the 256 × 256 'house' image (see Fig. C2(a)), and the setting for experiments is similar as that in He *et al.* (2013). More specifically, the original image is first blurred by an out-of-focus kernel with a radius of 5, and then a mask operator S is added such that 60% pixels are corrupted uniformly at random. The corrupted image is shown in Fig. C2(b). We use the signal-to-noise ratio (SNR) to measure the quality of the reconstructed image

$$\text{SNR} := 20\log_{10}\frac{\|x - \bar{x}\|}{\|x\|},$$

where x is the reconstructed image.

Recall that $\|\mathcal{A}^T\mathcal{A}\| = 1$ when $\mathcal{A} = BW$. Then, to implement the IDP-ALM (1.15), we set $\beta = 1.2$ and then $r = 1.201$ in our experiments. The stopping criterion is

$$\frac{\|x^{k+1} - x^k\|}{\|x^{k+1}\|} < 10^{-3}.$$



(a) (b) (c)

FIG. C2. (a) Clean image; (b) Corrupted image; (c) Restored image by (1.15) with $\tau = 0.75$ and $\gamma = 1$.

TABLE C2 *Numerical results of the IDP-ALM (1.15) for (C1)*

(τ and γ)	It.	CPU	SNR
$\tau = 0.75, \gamma = 1$	102	12.85	25.20
$\tau = 0.80, \gamma = 1.2$	105	12.64	25.18
$\tau = 0.85, \gamma = 1.4$	107	12.53	25.14
$\tau = 0.90, \gamma = 1.6$	110	13.12	25.12
$\tau = 0.95, \gamma = 1.8$	112	13.43	25.08
$\tau = 1, \gamma = 1$ (i.e., Linearized ALM (1.10))	115	13.44	25.05

Note that the closed-form solution of the resulting subproblem is given by

$$\left(I + \frac{1}{r}\partial(\|x\|_1)\right)^{-1}(c) := \text{sign}(c) \circ \left(\max\left\{|c| - \frac{1}{r}, 0\right\}\right), \ \forall\, c \in \Re^n,$$

where sign$(\cdot)$ is the signum function and the operator '$\circ$' stands for component wise scalar multiplication.

In Table C2 we report some results for the IDP-ALM (1.15) with different values of τ and γ. Note that the value of γ is also calculated via the formula (1.16). Again, the benchmark case for comparison is the linearized ALM (1.10), i.e., the case of (1.15) with $\tau = 1$ and $\gamma = 1$. It is shown that values of τ in the interval $(0.75, 1)$ accelerate the performance of the IDP-ALM (1.15) and the optimal value 0.75 is still the fastest. The reconstructed image is shown in Fig. C2(c).