Positive-Indefinite Proximal Augmented Lagrangian Method and its Application to Full Jacobian Splitting for Multi-block Separable Convex Minimization Problems

Bingsheng He[1] **Feng Ma**[2] **Xiaoming Yuan** [3]

September 15, 2016

Abstract. The augmented Lagrangian method (ALM) is fundamental for 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 positive-indefiniteness by reducing the proximal parameter. The positive-indefinite proximal ALM is thus proposed for the generic setting of convex programming problems with linear constraints. We show that our relaxation is optimal in sense of that the proximal parameter cannot be further reduced. The consideration of positive-indefinite proximal regularization is particularly meaningful for generating larger step sizes for solving the primal subproblems of ALM. When the model under discussion is separable in sense of that its objective function consists of finitely many additive function components without coupled variables, it is desired to decompose each ALM's primal subproblem in Jacobian manner, replacing the original primal subproblem by a sequence of easier and smaller decomposed subproblems, so that parallel computation can be applied. This full Jacobian splitting version of 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, convergence rate

1 Introduction

We consider the generic convex minimization model with linear constraints

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

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

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

[1]Department of Mathematics, Nanjing University, China, and Department of Mathematics, South University of Science and Technology of China. This author was supported by the NSFC Grant 11471156 and 91530115. Email: hebma@nju.edu.cn

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

[3]Corresponding author. Department of Mathematics, Hong Kong Baptist University, Hong Kong. This author was supported by a General Research Fund from Hong Kong Research Grants Council. Email: xmyuan@hkbu.edu.hk

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

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

with $\beta > 0$ the penalty parameter for the linear constraints. The augmented Lagrangian method originally proposed in [19, 21] for (1.1) reads as

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

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 [23], 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 [20, 24] to the dual problem of (1.1). Throughout, we also call $\boldsymbol{x}$ and λ the primal and dual variables; (1.4a) and (1.4b) the primal and dual subproblems of ALM, respectively. For the penalty parameter β, we fix it in our discussion for simplicity. Indeed, as our work [6] shows, without loss of generality, we can fix it as 1 for theoretical discussion. We refer to [1] 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 regularizing the primal subproblem (1.4a) by a quadratic proximal term and accordingly considering the proximal version of ALM:

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

In (1.5), $\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|_{\mathcal{D}}^2$ 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 [20, 24]. 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) which allows for a more general step size for updating the dual variable λ.

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

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

For a general objective function $\boldsymbol{\theta}(\boldsymbol{x})$, 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, $\boldsymbol{\theta}(\boldsymbol{x})$ may be special enough so that its proximity operator, which is given by

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

has a closed-form expression; in which $\rho > 0$ is a constant. Such a representative case is where $\boldsymbol{\theta}(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 := \sum_{i=1}^n |x_i|$. 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

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

which amounts to estimating the proximity operator of $\boldsymbol{\theta}(\boldsymbol{x})$ 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

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

Recall that for the linearized ALM (1.10), the parameter r is required to satisfy $r > \beta\|\mathcal{A}^T\mathcal{A}\|$ 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 [26] 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 $\boldsymbol{x}$-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

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

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 $\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}^k\|^2_{\mathcal{D}}$ with $\mathcal{D}$ 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.

Therefore, the necessity of considering positive-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 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

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

where the proximal matrix D_0 is not necessarily positive definite. Note that we slightly abuse the notation of $\|\boldsymbol{x}-\boldsymbol{x}^k\|^2_{\mathcal{D}_0}$ to denote the term $(\boldsymbol{x}-\boldsymbol{x}^k)^T D_0(\boldsymbol{x}-\boldsymbol{x}^k)$ regardless of the positive-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 $\Re^l$ and $\tau \in (0,1)$. Obviously, D_0 defined in (1.13) is not necessarily positive definite. If we choose $\mathcal{D}$ as (1.8), i.e., the linearized ALM is considered, then $\mathcal{D}_0$ given in (1.13) corresponds to the specific choice in (1.11). Hence, we do not focus on how to choose $\mathcal{D}$ here. 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 for 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., [8, 13], for some theoretical and experimental study 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 [9] and can be regarded as a splitting version of the ALM. Here we are interested in the more general scheme:

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

with the possibility of enlarging the step size for updating the dual variable. Recall that, as analyzed in [1, 24], in (1.4b), the dual variable λ is updated by a steepest accent step applied to the dual problem of (1.1). In [25], 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 positive-indefinite proximal ALM with a general step size for updating the dual variable:

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

where $\mathcal{D}_0$ is given in (1.13) with an arbitrarily given positive-definite matrix $\mathcal{D} \in \Re^l$ and $\tau \in (0, 1)$. It is abbreviated as PIDP-ALM hereafter. For the PIDP-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 PIDP-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 for 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:

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

Hence, for any given $\gamma \in (0, 2)$, we find a γ-depending lower bound of τ for the PIDP-ALM (1.15) via the restriction (1.16). 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 [26] 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. Therefore, though 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 PIDP-ALM (1.15) with the restriction (1.16) in Section 3 [4]. 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 PIDP-ALM (1.15) with

[4] 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 [14] posed on Optimization Online in July, 2016.

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 [18]. 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.

2.1 Variational inequality characterization of (1.1)

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

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

Alternatively, we can rewrite these inequalities as the variational inequalities:

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

or in the compact form

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

where

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

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

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

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, $\theta(x)$ and $f(x)$ be convex functions. If f is differentiable, and the solution set of the minimization problem* $\min\{\theta(x) + f(x) \,|\, x \in \mathcal{X}\}$ *is nonempty, then*

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

if and only if

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

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 for updating the dual variable and prove its convergence. Its iterative scheme reads as

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

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

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

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

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

By using (2.5b), it holds that

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

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

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

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

$$\lambda^{k+1}\in\Re^l,\ \ (\lambda-\lambda^{k+1})^T\{(\mathcal{A}\boldsymbol{x}^{k+1}-b) + \frac{1}{\gamma\beta}(\lambda^{k+1}-\lambda^k)\} \ge 0,\ \ \forall\lambda\in\Re^l. \quad (2.7b)$$

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

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

where

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

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

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

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

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

$$(\boldsymbol{x}^{k+1}-\boldsymbol{x}^*)^T\mathcal{D}(\boldsymbol{x}^k-\boldsymbol{x}^{k+1}) + (\boldsymbol{x}^{k+1}-\boldsymbol{x}^*)^T\Big(1-\frac{1}{\gamma}\Big)\mathcal{A}^T(\lambda^k-\lambda^{k+1})$$
$$+ (\lambda^{k+1}-\lambda^*)^T\frac{1}{\gamma\beta}(\lambda^k-\lambda^{k+1}) \ge 0, \quad \forall u^*\in\Omega^*.$$

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

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

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

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

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

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

or, equivalently,

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

This inequality essentially implies the convergence of the sequence generated by the scheme (2.5). The remaining part of the proof is subroutine and thus omitted; we refer to, e.g., [5] or [11] for a tutorial.

3 Convergence of the PIDP-ALM (1.15)

In this section, we prove the convergence for the PIDP-ALM (1.15) with the step size restriction (1.16). Note that the PIDP-ALM (1.15) differs from the scheme (2.5) in that the proximal matrix $\mathcal{D}_0$ in (1.15a) is positive 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.

3.1 Key theorem for convergence proof

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

Theorem 3.1 *Let $\{\boldsymbol{u}^k\}$ be the sequence generated by the PIDP-ALM* (1.15) *for the problem* (1.1). *Then for any $\tau \in (\frac{2+\gamma}{4}, 1)$, we have*

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

where

$$H = \begin{pmatrix} \mathcal{D}+(1-\tau)\beta\mathcal{A}^T\mathcal{A} & 0 \\ 0 & \frac{1}{\gamma\beta}I_l \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} \mathcal{D} & 0 \\ 0 & \frac{4\tau-\gamma-2}{\gamma^2\beta}I_l \end{pmatrix} \quad (3.2)$$

and $\mathcal{D}\in\Re^l$ is an arbitrarily positive-definite matrix and $\mathcal{D}_0$ is given by (1.13).

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

3.2 Variational inequality characterization

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

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

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

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

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

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

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

3.3 A prediction-correction expression

We show that the PIDP-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 $\tilde{\boldsymbol{u}}^k = (\tilde{\boldsymbol{x}}^k, \tilde{\lambda}^k)$ by

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

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

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

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

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

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

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

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

where

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

Further, using the notation (3.4), we have

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

Thus, because of (1.15b), we obtain

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

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

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

where

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

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

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

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

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

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

Using (3.6b) and the above equation, we have

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

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

$$(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k)^T G_0 (\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k),$$

which is not necessarily non-negative.

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

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

and

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

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

3.4 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 PIDP-ALM (1.15).

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

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

where G_0 is defined in (3.10)*.*

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

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

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

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

and hence we have

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

Applying the identity

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

to the right-hand side of (3.15) with

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

we thus obtain

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

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

$$\begin{aligned} & \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k\|_{H_0}^2 - \|\boldsymbol{u}^{k+1} - \tilde{\boldsymbol{u}}^k\|_{H_0}^2 \\ = \ & \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k\|_{H_0}^2 - \|(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k) - (\boldsymbol{u}^k - \boldsymbol{u}^{k+1})\|_{H_0}^2 \\ \overset{(3.7a)}{=} \ & \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k\|_{H_0}^2 - \|(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k) - M(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k)\|_{H_0}^2 \\ = \ & 2(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k)^T H_0 M (\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k) - (\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k)^T M^T H M (\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k) \\ = \ & (\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k)^T (Q^T + Q - M^T H_0 M)(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k) \\ \overset{(3.10)}{=} \ & \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^k\|_{G_0}^2. \end{aligned} \tag{3.17}$$

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

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

$$\|\boldsymbol{u}^{k+1}-\boldsymbol{u}^*\|_{H_0}^2 \le \|\boldsymbol{u}^k-\boldsymbol{u}^*\|_{H_0}^2 - \|\boldsymbol{u}^k-\tilde{\boldsymbol{u}}^k\|_{G_0}^2. \tag{3.18}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.5 Convergence proof

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

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

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

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 $\tau\in(\frac{2+\gamma}{4},1)$. Now, based on the key inequality (3.1) in Theorem 3.1, we can prove the convergence theorem for the PIDP-ALM (1.15).

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

Proof. First, it follows from (3.1) that

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

Consequently, we have

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

For any fixed $\boldsymbol{u}^*\in\Omega^*$ and $k\ge1$, we have

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

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

$$\|\boldsymbol{u}^{k+1}-\boldsymbol{u}^\infty\|_H^2\ \le\ \|\boldsymbol{u}^k-\boldsymbol{u}^\infty\|_H^2-\|\boldsymbol{u}^k-\boldsymbol{u}^{k+1}\|_G^2.\tag{3.23}$$

Note that $\boldsymbol{u}^\infty$ is also the limit point of $\{\boldsymbol{u}^{k_j}\}$. Together with (3.21), it is impossible for the sequence $\{\boldsymbol{u}^k\}$ to have more than one cluster point. Thus the sequence $\{\boldsymbol{u}^k\}$ converges to $\boldsymbol{u}^\infty$ 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 PIDP-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 PIDP-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

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

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

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

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 $\{u^k = (x^k, \lambda^k)\}$. For any given $\tau < \frac{2+\gamma}{4}$, there exists $\delta > 0$ such that $\delta + \tau < \big(\frac{2+\gamma}{4}\big)$ 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

$$\begin{cases} x^{k+1} = \dfrac{\alpha - 1}{\alpha}x^k + \dfrac{1}{\alpha}\lambda^k, \\[2mm] \lambda^{k+1} = \dfrac{\gamma(1 - \alpha)}{\alpha}x^k + \dfrac{\alpha - \gamma}{\alpha}\lambda^k, \end{cases} \tag{4.5}$$

which can be written as

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

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

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

and

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

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

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

That is, for any $\alpha \in \big(0, \frac{2+\gamma}{4}\big)$, 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 PIDP-ALM (1.15) to the problem (4.1), is not necessarily convergent for any $\tau \in \big(0, \frac{2+\gamma}{4}\big)$. Thus, $\frac{2+\gamma}{4}$ is the smallest lower bound for τ to ensure the convergence of the PIDP-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 [18].

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:

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

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:

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

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 sepcified as

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

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

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

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 functions 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., [2, 3]. This line of research, which could be called augmented-Lagrangian-based splitting algorithms, has gained a lot of attention from the community. Particularly, the mentioned ADMM originally proposed in [9] 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 [6] 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 [12, 15, 16, 17, 18] 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:

$$\left\{\begin{array}{l}\left\{\begin{array}{l} x_1^{k+1} = \arg\min\big\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_m^k, \lambda^k) \;\big|\; x_1 \in X_1\big\}; \\ \quad\quad\vdots \\ x_i^{k+1} = \arg\min\big\{\mathcal{L}_\beta(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k)\big| x_i \in X_i\big\}; \\ \quad\quad\vdots \\ x_m^{k+1} = \arg\min\big\{\mathcal{L}_\beta(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) \;\big|\; x_m \in X_m\big\}; \end{array}\right. \quad (5.5a) \\ \lambda^{k+1} = \lambda^k - \beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big). \quad (5.5b) \end{array}\right.$$

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 [12] 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 appropriately modified to guarantee the convergence. In our previous work [18] (see also [7]), it is suggested to regularize all the decomposed subproblems over primal variables by sufficiently large proximal terms:

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

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 [18] 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

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

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:

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

Then, it was suggested in [22] (see Algorithms 2 and 3 therein) to apply the original ADMM in [9] 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 [18] (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

$$\left\{ \begin{array}{ll} x_i^{k+1} = \arg\min\big\{\theta_i(x_i) - (\lambda^k)^T A_i x_i + \dfrac{\alpha}{2}\|A_i(x_i - x_i^k) + \dfrac{1}{m}\big(\sum\limits_{i=1}^m A_i x_i^k - b\big)\|^2 \mid x_i \in X_i\big\}, & \text{(5.9a)} \\ \qquad i = 1, \cdots, m. & \\ \lambda^{k+1} = \lambda^k - \dfrac{\alpha}{m}\big(\sum\limits_{i=1}^m A_i x_i^{k+1} - b\big), & \text{(5.9b)} \end{array} \right.$$

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

Later, it is pointed out in [3] that the scheme (5.9) is essentially the same as the earlier scheme proposed in [4] (also see (2.33)-(2.34) in [3]) with a notation difference of the Lagrange multiplier. Let us recall the scheme in [4]:

$$\left\{ \begin{array}{ll} x_i^{k+1} = \arg\min\big\{\theta_i(x_i) + (\lambda^k)^T A_i x_i + \dfrac{\alpha}{2}\big\|A_i(x_i - x_i^k) + \dfrac{1}{m}\big(\sum\limits_{j=1}^m A_j x_j^k - b\big)\big\|^2 \mid x_i \in X_i\big\}, & \text{(5.10a)} \\ \qquad i = 1, \cdots, m. & \\ \lambda^{k+1} = \lambda^k + \dfrac{\alpha}{m}\big(\sum\limits_{i=1}^m A_i x_i^{k+1} - b\big). & \text{(5.10b)} \end{array} \right.$$

Indeed, it is clear that the scheme (5.10) corresponds to the special case of Algorithm 8.1 in [10] 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

$$\left\{\begin{array}{l}\left\{\begin{array}{l} x_1^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \mid x_1 \in X_1\}; \\ \quad\vdots \\ x_i^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_i(x_i - x_i^k)|x_i \in X_i\}; \\ \quad\vdots \\ x_m^{k+1} = \arg\min\{\mathcal{L}_\beta(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) + \frac{s\beta}{2}\|A_m(x_m - x_m^k) \mid x_m \in X_m\}; \end{array}\right. \quad (5.11a) \\ \lambda^{k+1} = \lambda^k - \gamma\beta\Big(\sum_{i=1}^m A_i x_i^{k+1} - b\Big),\ \gamma \in (0,2). \quad (5.11b)\end{array}\right.$$

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

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

which improves the result $s \geq m-1$ in our previous work [18] 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 Proof

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

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

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

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

which can be rewritten as

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

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

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

where

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

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

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

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

Hence, for the specific scheme (5.11), the matrix $(1+s)\beta\text{diag}(\mathcal{A}^T\mathcal{A})$ 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 $\boldsymbol{x}^{k+1} \in \mathcal{X}$ and

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

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 $\mathcal{D}_0 = (1+s)\beta\text{diag}(\mathcal{A}^T\mathcal{A})$, if we set

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

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

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

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 PIDP-ALM (1.15).

To fulfill (5.16), notice that

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

where

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

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

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

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

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

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 [18].

5.8 Remarks

Meanwhile, it worths to mention 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., [16], for similar analysis.

6 Conclusions

In this paper, we propose the positive-indefinite proximal version of the augmented Lagrangian method (ALM) for convex programming problems that allows the ALM's primal subproblem at each iteration to be regularized by a quadratic proximal term induced with a positive-indefinite matrix. The consideration of positive-indefinite quadratic proximal regularization makes particular sense in generating larger step sizes for solving the primal subproblems and thus potentially resulting in better numerical performance. We also allow the dual variable to be updated by a general step size; and the mutually-constrained roles of the step sizes for respectively updating the primal and dual variables are rigorously formatted by a formula. This formula is shown to be optimal by examples. The convergence of the positive-indefinite proximal ALM with this step size restriction is proved. Then, we consider the multi-block separable convex minimization case where the objective function is the sum of finitely many additive function components without coupled variables, and revisit the full Jacobian splitting version of the ALM in [18] whose subproblems over primal variables are all required to be proximally regularized. We show that the proximal parameter for this proximally regularized full Jacobian splitting version of ALM can be further reduced from 1 to the tight value of 0.75.

References

[1] D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods, Academic Press, New York (1982).

[2] D. P. Bertsekas, Incremental proximal methods for large scale convex optimization, Math. Program., 129 (2011), pp. 163-195.

[3] D. P. Bertsekas, Incremental aggregated proximal and augmented Lagrangian algorithms, arViv: 1509:09257v1, 2015.

[4] D. P. Bertsekas and J. N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J., 1989.

[5] E. Blum and W. Oettli, Mathematische Optimierung, Econometrics and Operations Research XX, Springer Verlag, 1975.

[6] 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 necessary convergent, Math. Program., 155 (2016), pp. 57-79.

[7] W. Deng, M.-J. Lai, Z. Peng and W. Yin, Parallel multi-block ADMM with $o(1/k)$ convergence, manuscrip.

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

[9] R. Glowinski and A. Marrocco, *Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problèmes non linéaires*, R.A.I.R.O., R2 (1975), pp. 41-76.

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

[11] B. S. He, PPA-like contraction methods for convex optimization: a framework using variational inequality approach, J. Oper. Res. Soc. China, (3) (2015), pp. 391-420.

[12] 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)(2015), pp. 2274-2312.

[13] B. S. He, F. Ma and X. M. Yuan, Convergence study on the symmetric version of ADMM with larger step sizes, SIAM J. Imaging Sci., to appear.

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

[15] B. S. He, M. Tao and X. M. Yuan, Alternating direction method with Gaussian back substitution for separable convex programming, SIAM J. Optim., 22 (2012), pp. 313-340.

[16] B. S. He, M. Tao and X. M. Yuan, A splitting method for separable convex programming, IMA J. Numer. Analy., 31 (2015), pp. 394-426.

[17] 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, Math. Oper. Res., to appear.

[18] B. S. He, H. K. Xu and X. M. Yuan, On the proximal Jacobian decomposition of ALM for multiple-block separable convex minimization problems and its relationship to ADMM, J. Sci. Comput., 66 (2016), pp. 1204-1217.

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

[20] B. Martinet, Regularisation, d'inéquations variationelles par approximations succesives, Rev. Francaise d'Inform. Recherche Oper., 4 (1970), pp. 154-159.

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

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

[23] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, Math. Oper. Res., 1(1976), pp. 877-898.

[24] R. T. Rockafellar, Monotone operators and the proximal point algorithm, SIAM J. Cont. Optim., 14 (1976), pp. 877-898.

[25] M. Tao and X. M. Yuan, The generalized proximal point algorithm with step size of 2 is not necessarily convergent, manuscript, May 2016.

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