

# A GENERAL INERTIAL PROXIMAL POINT METHOD FOR MIXED VARIATIONAL INEQUALITY PROBLEM

CAIHUA CHEN\*, SHIQIAN MA<sup>†</sup>, AND JUNFENG YANG<sup>‡</sup>

**Abstract.** In this paper, we first propose a general inertial proximal point method for the mixed variational inequality (VI) problem. Based on our knowledge, without stronger assumptions, convergence rate result is not known in the literature for inertial type proximal point methods. Under certain conditions, we are able to establish the global convergence and a  $o(1/k)$  convergence rate result (under certain measure) of the proposed general inertial proximal point method. We then show that the linearized alternating direction method of multipliers (ADMM) for separable convex optimization with linear constraints is an application of a general proximal point method, provided that the algorithmic parameters are properly chosen. As byproducts of this finding, we establish global convergence and  $O(1/k)$  convergence rate results of the linearized ADMM in both ergodic and nonergodic sense. In particular, by applying the proposed inertial proximal point method for mixed VI to linearly constrained separable convex optimization, we obtain an inertial version of the linearized ADMM for which the global convergence is guaranteed. We also demonstrate the effect of the inertial extrapolation step via experimental results on the compressive principal component pursuit problem.

**Key words.** proximal point method, inertial proximal point method, mixed variational inequality, linearized alternating direction method of multipliers, inertial linearized alternating direction method of multipliers.

**AMS subject classifications.** 65K05, 65K10, 65J22, 90C25

**1. Introduction.** Let  $T : \mathfrak{R}^n \rightrightarrows \mathfrak{R}^n$  be a set-valued maximal monotone operator from  $\mathfrak{R}^n$  to its power set. The maximal monotone operator inclusion problem is to find  $w^* \in \mathfrak{R}^n$  such that

$$0 \in T(w^*). \tag{1.1}$$

Due to the mathematical generality of maximal monotone operators, the problem (1.1) is very inclusive and serves as a unified model for many problems of fundamental importance, for example, fixed point problem, variational inequality problem, minimization of closed proper convex functions, and their extensions. Therefore, it becomes extremely important in many cases to solve (1.1) in practical and efficient ways.

The classical proximal point method, which converts the maximal monotone operator inclusion problem to a fixed point problem of a firmly nonexpansive mapping via resolvent operators, is one of the most influential approaches for solving (1.1) and has been studied extensively both in theory and in practice. The proximal point method was originally proposed by Martinet [1] based on the work of Moreau [2] and was popularized by Rockafellar [3]. It turns out that the proximal point method is a very powerful algorithmic tool and contains many well known algorithms as special cases. In particular, it was shown that the classical augmented Lagrangian method for constrained optimization [4, 5], the Douglas-Rachford operator splitting method [6] and the alternating direction method of multipliers (ADMM, [7, 8]) are all applications of the

---

\*International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University, China (Email: [chchen@nju.edu.cn](mailto:chchen@nju.edu.cn)). Research supported in part by Natural Science Foundation of Jiangsu Province under project grant No. BK20130550 and the Natural Science Foundation of China NSFC grant 11371192.

<sup>†</sup>Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong (Email: [sqma@se.cuhk.edu.hk](mailto:sqma@se.cuhk.edu.hk)). This author was supported by Hong Kong Research Grants Council General Research Fund Early Career Scheme (Project ID: CUHK 439513).

<sup>‡</sup>Corresponding author (Email: [jfyang@nju.edu.cn](mailto:jfyang@nju.edu.cn)). Department of Mathematics, Nanjing University, China. This author was supported by the Natural Science Foundation of China NSFC-11371192 and a grant from Jiangsu Key Laboratory for Numerical Simulation of Large Scale Complex Systems. The work was done while this author was visiting the Chinese University of Hong Kong.

proximal point method, see [9, 10]. Various inexact, relaxed and accelerated variants of the proximal point method were also very well studied in the literature, see, e.g., [3, 10, 11].

The primary proximal point method for minimizing a differentiable function  $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$  can be interpreted as an implicit one-step discretization method for the ordinary differential equations

$$w' + \nabla f(w) = 0, \quad (1.2)$$

where  $w : \mathfrak{R} \rightarrow \mathfrak{R}^n$  is differentiable,  $w'$  denotes its derivative, and  $\nabla f$  is the gradient of  $f$ . Suppose that  $f$  is closed proper and convex and its minimum value is attained, then every solution trajectory  $\{w(t) : t \geq 0\}$  of the differential system (1.2) converges to a minimizer of  $f$  as  $t$  goes to infinity. Similar conclusion can be drawn for (1.1) by considering the evolution differential inclusion problem  $0 \in w'(t) + T(w(t))$  almost everywhere on  $\mathfrak{R}_+$ , provided that the operator  $T$  satisfies certain conditions, see e.g., [12].

The proximal point method is a one-step iterative method, i.e., each new iterate point does not depend on any iterate points already generated other than the current one. To speed up convergence, multi-step methods have been proposed in the literature by discretizing a second-order ordinary differential system of the form

$$w'' + \gamma w' + \nabla f(w) = 0, \quad (1.3)$$

where  $\gamma > 0$ . Studies in this direction can be traced back to at least [13] which examined the system (1.3) in the context of optimization. In the two-dimensional case, the system (1.3) characterizes roughly the motion of a heavy ball which rolls under its own inertial over the graph of  $f$  until friction stops it at a stationary point of  $f$ . The three terms in (1.3) denote, respectively, inertial force, friction force and gravity force. Therefore, the system (1.3) is usually referred to as the *heavy-ball with friction* (HBF) system. It is easy to show that the energy function  $E(t) = \frac{1}{2}\|w'(t)\|^2 + f(w(t))$  is always decreasing with time  $t$  unless  $w'$  vanishes, which implies that the HBF system is dissipative. It was proved in [14] that if  $f$  is convex and its minimum value is attained then each solution trajectory  $\{w(t) : t \geq 0\}$  of (1.3) converges to a minimizer of  $f$ . In theory the convergence of the solution trajectories of the HBF system to a stationary point of  $f$  can be faster than those of the first-order system (1.2), while in practice the second order inertial term  $w''$  can be exploited to design faster algorithms [15, 16]. Motivated by the properties of (1.3), an implicit discretization method was proposed in [14]. Specifically, given  $w^{k-1}$  and  $w^k$ , the next point  $w^{k+1}$  is determined via

$$\frac{w^{k+1} - 2w^k + w^{k-1}}{h^2} + \gamma \frac{w^{k+1} - w^k}{h} + \nabla f(w^{k+1}) = 0,$$

which results to an iterative algorithm of the form

$$w^{k+1} = (I + \lambda \nabla f)^{-1}(w^k + \alpha(w^k - w^{k-1})), \quad (1.4)$$

where  $\lambda = h^2/(1 + \gamma h)$  and  $\alpha = 1/(1 + \gamma h)$ . Note that (1.4) is nothing but a proximal point step applied to the extrapolated point  $w^k + \alpha(w^k - w^{k-1})$ , rather than  $w^k$  as in the classical proximal point method. Thus the resulting iterative scheme (1.4) is a two-step method and is usually referred as an inertial proximal point algorithm (PPA). Convergence properties of (1.4) were studied in [14] under some assumptions on the parameters  $\alpha$  and  $\lambda$ . Subsequently, this inertial technique was extended to solve the inclusion problem (1.1) of maximal monotone operators in [17]. See also [18] for approximate inertial PPA and [19, 20, 21] for some inertial type hybrid proximal algorithms. Recently, there are increasing interests in studying inertial type algorithms. Some latest references are inertial forward-backward splitting methods for certain

separable nonconvex optimization problems [22] and for strongly convex problems [23, 24], inertial versions of the Douglas-Rachford operator splitting method and the ADMM for maximal monotone operator inclusion problem [25, 26], and inertial forward-backward-forward method [27] based on Tseng's approach [28]. See also [29, 30].

**1.1. Contributions.** In this paper, we focus on the mixed variational inequality (VI) problem and study inertial PPA under a more general setting. In particular, a weighting matrix  $G$  in the proximal term is introduced. In our setting the matrix  $G$  is allowed to be positive semidefinite, as long as it is positive definite in the null space of a certain matrix. We establish its global convergence and a  $o(1/k)$  convergence rate result under certain conditions. To the best of our knowledge, without stronger assumptions, convergence rate result is not known in the literature for general inertial type proximal point methods. This general setting allows us to propose an inertial version of the linearized ADMM, a practical variant of the well-known ADMM which has recently found numerous applications [31]. We show that the linearized ADMM for separable convex optimization is an application of a general PPA to the primal-dual optimality conditions, as long as the parameters are properly chosen. As byproducts of this finding, we establish global convergence and  $O(1/k)$  convergence rate results of the linearized ADMM. Another aim of this paper is to study the effect of the inertial extrapolation step via numerical experiments. Finally, we connect inertial type algorithms with the popular accelerated methods pioneered by Nesterov [32] and give some concluding remarks.

The main reason that we restrict our analysis to mixed VI problem rather than the apparently more general problem (1.1) is because it is very convenient to represent the optimality conditions of linearly constrained separable convex optimization as mixed VI. In fact, our analysis for Theorems 1 and 2 can be generalized to the maximal monotone operator inclusion problem (1.1) without any difficulty.

**1.2. Notation.** We use the following notation. The standard inner product and  $\ell_2$  norm are denoted by  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$ , respectively. The sets of symmetric, symmetric positive semidefinite and symmetric positive definite matrices of order  $n$  are, respectively, denoted by  $S^n$ ,  $S_+^n$  and  $S_{++}^n$ . For any matrix  $A \in S_+^n$  and vectors  $u, v \in \mathfrak{R}^n$ , we let  $\langle u, v \rangle_A := u^T A v$  and  $\|u\|_A := \sqrt{\langle u, u \rangle_A}$ . The Frobenius norm is denoted by  $\| \cdot \|_F$ . The spectral radius of a square matrix  $M$  is denoted by  $\rho(M)$ .

**2. A general inertial PPA for mixed VI.** Let  $\Omega \subseteq \mathfrak{R}^n$  be a closed and convex set,  $\theta : \mathfrak{R}^n \rightarrow \mathfrak{R}$  be a closed proper convex function, and  $F : \mathfrak{R}^n \rightarrow \mathfrak{R}^n$  be a monotone mapping. In this paper, we consider the mixed VI problem: find  $w^* \in \Omega$  such that

$$\theta(w) - \theta(w^*) + \langle w - w^*, F(w^*) \rangle \geq 0, \quad \forall w \in \Omega. \quad (2.1)$$

Let  $G \in S_+^n$  and two sequences of parameters  $\{\alpha_k \geq 0 : k = 0, 1, 2, \dots\}$  and  $\{\lambda_k > 0 : k = 0, 1, 2, \dots\}$  be given. We study a general inertial PPA of the following form: given any  $w^0 = w^{-1} \in \mathfrak{R}^n$ , for  $k = 0, 1, 2, \dots$ , find  $w^{k+1} \in \Omega$  such that

$$\bar{w}^k := w^k + \alpha_k(w^k - w^{k-1}), \quad (2.2a)$$

$$\theta(w) - \theta(w^{k+1}) + \langle w - w^{k+1}, F(w^{k+1}) + \lambda_k^{-1} G(w^{k+1} - \bar{w}^k) \rangle \geq 0, \quad \forall w \in \Omega. \quad (2.2b)$$

We make the following assumptions.

ASSUMPTION 1. *The set of solutions of (2.1), denoted by  $\Omega^*$ , is nonempty.*

ASSUMPTION 2. *The mapping  $F$  is  $H$ -monotone in the sense that*

$$\langle u - v, F(u) - F(v) \rangle \geq \|u - v\|_H^2, \quad \forall u, v \in \mathfrak{R}^n, \quad (2.3)$$

where  $H \in S_+^n$ . Note that  $H = 0$  if  $F$  is monotone, and  $H \in S_{++}^n$  if  $F$  is strongly monotone.

ASSUMPTION 3. The sum of  $G$  and  $H$ , denoted by  $M$ , is positive definite, i.e.,  $M := G + H \in S_{++}^n$ .

Under Assumptions 2 and 3, it can be shown that  $w^{k+1}$  is uniquely determined in (2.2b). Therefore, the algorithm (2.2a)-(2.2b) is well defined. Clearly, the algorithm reduces to the classical PPA if  $G \in S_{++}^n$  and  $\alpha_k = 0$  for all  $k$ . It is called inertial PPA because  $\alpha_k$  can be greater than 0. We will impose conditions on  $\alpha_k$  to ensure global convergence of the inertial PPA framework (2.2). Our convergence results are extensions of those in [17].

THEOREM 1. Assume that Assumptions 1, 2 and 3 hold. Let  $\{w^k\}_{k=0}^\infty \subseteq \mathfrak{R}^n$  conforms to Algorithm (2.2a)-(2.2b). The parameters  $\{\alpha_k, \lambda_k\}_{k=0}^\infty$  satisfy, for all  $k$ ,  $0 \leq \alpha_k \leq \alpha$  for some  $\alpha \in [0, 1)$  and  $\lambda_k \geq \lambda$  for some  $\lambda > 0$ . If

$$\sum_{k=1}^{\infty} \alpha_k \|w^k - w^{k-1}\|_G^2 < \infty, \quad (2.4)$$

then the sequence  $\{w^k\}_{k=0}^\infty$  converges to some point in  $\Omega^*$  as  $k \rightarrow \infty$ .

*Proof.* First, we show that, for any  $w^* \in \Omega^*$ ,  $\lim_{k \rightarrow \infty} \|w^k - w^*\|_M$  exists. As a result,  $\{w^k\}_{k=0}^\infty$  is bounded and must have a limit point. Then, we show that any limit point of  $\{w^k\}_{k=0}^\infty$  must lie in  $\Omega^*$ . Finally, we establish the convergence of  $\{w^k\}_{k=0}^\infty$  to a point in  $\Omega^*$  as  $k \rightarrow \infty$ .

Let  $w^* \in \Omega^*$  be arbitrarily chosen and  $k \geq 0$ . It follows from setting  $w = w^* \in \Omega^*$  in (2.2b) and the  $H$ -monotonicity (2.3) of  $F$  that

$$\begin{aligned} \lambda_k^{-1} \langle w^{k+1} - w^*, w^{k+1} - \bar{w}^k \rangle_G &\leq \theta(w^*) - \theta(w^{k+1}) - \langle w^{k+1} - w^*, F(w^{k+1}) \rangle \\ &\leq \theta(w^*) - \theta(w^{k+1}) - \langle w^{k+1} - w^*, F(w^*) \rangle - \|w^{k+1} - w^*\|_H^2 \\ &\leq -\|w^{k+1} - w^*\|_H^2. \end{aligned} \quad (2.5)$$

Define  $\varphi_k := \|w^k - w^*\|_G^2$  and recall that  $\bar{w}^k = w^k + \alpha_k(w^k - w^{k-1})$ . Plug the identities

$$\begin{aligned} 2\langle w^{k+1} - w^*, w^{k+1} - w^k \rangle_G &= \varphi_{k+1} - \varphi_k + \|w^{k+1} - w^k\|_G^2, \\ 2\langle w^{k+1} - w^*, w^k - w^{k-1} \rangle_G &= \varphi_k - \varphi_{k-1} + \|w^k - w^{k-1}\|_G^2 + 2\langle w^{k+1} - w^k, w^k - w^{k-1} \rangle_G, \end{aligned}$$

into (2.5) and reorganize, we obtain

$$\begin{aligned} \psi_k &:= \varphi_{k+1} - \varphi_k - \alpha_k(\varphi_k - \varphi_{k-1}) \\ &\leq -\|w^{k+1} - w^k\|_G^2 + 2\alpha_k \langle w^{k+1} - w^k, w^k - w^{k-1} \rangle_G + \alpha_k \|w^k - w^{k-1}\|_G^2 - 2\lambda_k \|w^{k+1} - w^*\|_H^2 \\ &= -\|w^{k+1} - \bar{w}^k\|_G^2 + (\alpha_k^2 + \alpha_k) \|w^k - w^{k-1}\|_G^2 - 2\lambda_k \|w^{k+1} - w^*\|_H^2 \\ &\leq -\|w^{k+1} - \bar{w}^k\|_G^2 + 2\alpha_k \|w^k - w^{k-1}\|_G^2 - 2\lambda_k \|w^{k+1} - w^*\|_H^2 \\ &\leq -\|w^{k+1} - \bar{w}^k\|_G^2 + 2\alpha_k \|w^k - w^{k-1}\|_G^2, \end{aligned} \quad (2.6)$$

where the first inequality is due to (2.5) and the second follows from  $0 \leq \alpha_k < 1$ . Define

$$\theta_k := \varphi_k - \varphi_{k-1} \quad \text{and} \quad \delta_k := 2\alpha_k \|w^k - w^{k-1}\|_G^2.$$

Then, the inequality (2.6) implies that  $\theta_{k+1} \leq \alpha_k \theta_k + \delta_k \leq \alpha[\theta_k]_+ + \delta_k$ , where  $[t]_+ := \max\{t, 0\}$  for  $t \in \mathfrak{R}$ . Therefore, we have

$$[\theta_{k+1}]_+ \leq \alpha[\theta_k]_+ + \delta_k \leq \alpha^{k+1}[\theta_0]_+ + \sum_{j=0}^k \alpha^j \delta_{k-j}. \quad (2.7)$$

Note that by our assumption  $w^0 = w^{-1}$ . This implies that  $\theta_0 = [\theta_0]_+ = 0$  and  $\delta_0 = 0$ . Therefore, it follows from (2.7) that

$$\sum_{k=0}^{\infty} [\theta_k]_+ \leq \frac{1}{1-\alpha} \sum_{k=0}^{\infty} \delta_k = \frac{1}{1-\alpha} \sum_{k=1}^{\infty} \delta_k < \infty. \quad (2.8)$$

Here the second inequality is due to the assumption (2.4). Let  $\gamma_k := \varphi_k - \sum_{j=1}^k [\theta_j]_+$ . From (2.8) and  $\varphi_k \geq 0$ , it follows that  $\gamma_k$  is bounded below. On the other hand,

$$\gamma_{k+1} = \varphi_{k+1} - [\theta_{k+1}]_+ - \sum_{j=1}^k [\theta_j]_+ \leq \varphi_{k+1} - \theta_{k+1} - \sum_{j=1}^k [\theta_j]_+ = \varphi_k - \sum_{j=1}^k [\theta_j]_+ = \gamma_k,$$

i.e.,  $\gamma_k$  is nonincreasing. As a result,  $\{\gamma_k\}_{k=0}^{\infty}$  converges as  $k \rightarrow \infty$ , and the following limit

$$\lim_{k \rightarrow \infty} \varphi_k = \lim_{k \rightarrow \infty} \left( \gamma_k + \sum_{j=1}^k [\theta_j]_+ \right) = \lim_{k \rightarrow \infty} \gamma_k + \sum_{k=1}^{\infty} [\theta_k]_+$$

exists. That is,  $\lim_{k \rightarrow \infty} \|w^k - w^*\|_G$  exists for any  $w^* \in \Omega^*$ . Furthermore, it follows from the second “ $\leq$ ” of (2.6) and the definition of  $\theta_k$  and  $\delta_k$  that

$$\begin{aligned} \|w^{k+1} - \bar{w}^k\|_G^2 + 2\lambda_k \|w^{k+1} - w^*\|_H^2 &\leq \varphi_k - \varphi_{k+1} + \alpha_k (\varphi_k - \varphi_{k-1}) + \delta_k \\ &\leq \varphi_k - \varphi_{k+1} + \alpha [\theta_k]_+ + \delta_k. \end{aligned} \quad (2.9)$$

By taking sum over  $k$  and noting that  $\varphi_k \geq 0$ , we obtain

$$\sum_{k=1}^{\infty} (\|w^{k+1} - \bar{w}^k\|_G^2 + 2\lambda_k \|w^{k+1} - w^*\|_H^2) \leq \varphi_1 + \sum_{k=1}^{\infty} (\alpha [\theta_k]_+ + \delta_k) < \infty, \quad (2.10)$$

where the second inequality follows from (2.8) and assumption (2.4). Since  $\lambda_k \geq \lambda > 0$  for all  $k$ , it follows from (2.10) that

$$\lim_{k \rightarrow \infty} \|w^k - w^*\|_H = 0. \quad (2.11)$$

Recall that  $M = G + H$ . Thus,  $\lim_{k \rightarrow \infty} \|w^k - w^*\|_M$  exists. Since  $M$  is positive definite, it follows that  $\{w^k\}_{k=0}^{\infty}$  is bounded and must have at least one limit point.

Again from (2.10) we have

$$\lim_{k \rightarrow \infty} \|w^{k+1} - \bar{w}^k\|_G = 0.$$

Thus, the positive semidefiniteness of  $G$  implies that  $\lim_{k \rightarrow \infty} G(w^{k+1} - \bar{w}^k) = 0$ . On the other hand, for any fixed  $w \in \Omega$ , it follows from (2.2b) that

$$\theta(w) - \theta(w^k) + \langle w - w^k, F(w^k) \rangle \geq \lambda_{k-1}^{-1} \langle w^k - w, G(w^k - \bar{w}^{k-1}) \rangle. \quad (2.12)$$

Suppose that  $w^*$  is any limit point of  $\{w^k\}_{k=0}^{\infty}$  and  $w^{k_j} \rightarrow w^*$  as  $j \rightarrow \infty$ . Since  $\Omega$  is closed,  $w^* \in \Omega$ . Furthermore, by taking the limit over  $k = k_j \rightarrow \infty$  in (2.12) and noting that  $G(w^k - \bar{w}^{k-1}) \rightarrow 0$  and  $\lambda_{k-1} \geq \lambda > 0$ , we obtain

$$\theta(w) - \theta(w^*) + \langle w - w^*, F(w^*) \rangle \geq 0.$$

Since  $w$  can vary arbitrarily in  $\Omega$ , we conclude that  $w^* \in \Omega^*$ . That is, any limit point of  $\{w^k\}_{k=0}^\infty$  must also lie in  $\Omega^*$ .

Finally, we establish the uniqueness of limit points of  $\{w^k\}_{k=0}^\infty$ . Suppose that  $w_1^*$  and  $w_2^*$  are two limit points of  $\{w^k\}_{k=0}^\infty$  and  $\lim_{j \rightarrow \infty} w^{i_j} = w_1^*$ ,  $\lim_{j \rightarrow \infty} w^{k_j} = w_2^*$ . Assume that  $\lim_{k \rightarrow \infty} \|w^k - w_i^*\|_M = v_i$  for  $i = 1, 2$ . By taking the limit over  $k = i_j \rightarrow \infty$  and  $k = k_j \rightarrow \infty$  in the equality

$$\|w^k - w_1^*\|_M^2 - \|w^k - w_2^*\|_M^2 = \|w_1^* - w_2^*\|_M^2 + 2\langle w_1^* - w_2^*, w_2^* - w^k \rangle_M,$$

we obtain  $v_1 - v_2 = -\|w_1^* - w_2^*\|_M^2 = \|w_1^* - w_2^*\|_M^2$ . Thus,  $\|w_1^* - w_2^*\|_M = 0$ . Since  $M$  is positive definite, this implies that  $w_1^* = w_2^*$ . Therefore,  $\{w^k\}_{k=0}^\infty$  converges to some point in  $\Omega^*$  and the proof of the theorem is completed.  $\square$

We have the following remarks on the assumptions and results of Theorem 1.

REMARK 1. *In practice, it is not hard to select  $\alpha_k$  online such that the condition (2.4) is satisfied.*

REMARK 2. *If  $\alpha_k = 0$  for all  $k$ , then the condition (2.4) is obviously satisfied. In this case, we reestablished the convergence of the classical PPA under the weaker condition that  $G \in S_+^n$ , provided that  $\lambda_k \geq \lambda > 0$  and  $H + G \in S_{++}^n$ , e.g., when  $F$  is strongly monotone, i.e.,  $H \in S_{++}^n$ .*

REMARK 3. *Suppose that  $H = 0$  and  $G \in S_+^n$ , but  $G \notin S_{++}^n$ . Then, the sequence  $\{w^k\}_{k=0}^\infty$  may not be well defined since (2.2b) does not necessarily have a solution in general. In the case that  $\{w^k\}_{k=0}^\infty$  is indeed well defined (which is possible), the conclusion that  $\lim_{k \rightarrow \infty} \|w^k - w^*\|_G$  exists for any  $w^* \in \Omega^*$  still holds under condition (2.4). However, since  $G$  is only positive semidefinite, the boundedness of  $\{w^k\}_{k=0}^\infty$  cannot be guaranteed. If a limit point  $w^*$  of  $\{w^k\}_{k=0}^\infty$  does exist, then the conclusion  $w^* \in \Omega^*$  holds still. Moreover, suppose that  $w_1^*$  and  $w_2^*$  are any two limit points of  $\{w^k\}_{k=0}^\infty$ , then it holds that  $Gw_1^* = Gw_2^*$ .*

In the following theorem, we remove the assumption (2.4) by assuming that the sequence  $\{\alpha_k\}_{k=0}^\infty$  satisfies some additional easily implementable conditions. Moreover, we establish a  $o(1/k)$  convergence rate result for the general inertial proximal point method (2.2). The trick used here to improve convergence rate from  $O(1/k)$  to  $o(1/k)$  seems to be first introduced in [33, 34]. To the best of our knowledge, there is no convergence rate result known in the literature without stronger assumptions for inertial type proximal point methods.

THEOREM 2. *Assume that Assumptions 1, 2 and 3 hold. Suppose that the parameters  $\{\alpha_k, \lambda_k\}_{k=0}^\infty$  satisfy, for all  $k$ ,  $0 \leq \alpha_k \leq \alpha_{k+1} \leq \alpha < \frac{1}{3}$  and  $\lambda_k \geq \lambda$  for some  $\lambda > 0$ . Let  $\{w^k\}_{k=0}^\infty$  be the sequence generated by Algorithm (2.2a)-(2.2b). Then, we have the following results.*

1.  $\{w^k\}_{k=0}^\infty$  converges to some point in  $\Omega^*$  as  $k \rightarrow \infty$ ;
2. For any  $w^* \in \Omega^*$  and positive integer  $k$ , it holds that

$$\min_{0 \leq i \leq k-1} \|w^{i+1} - \bar{w}^i\|_G^2 \leq \frac{\left(1 + \frac{2}{1-3\alpha}\right) \|w^0 - w^*\|_G^2}{k}. \quad (2.13)$$

Moreover, it holds as  $k \rightarrow \infty$  that

$$\min_{0 \leq i \leq k-1} \|w^{i+1} - \bar{w}^i\|_G^2 = o\left(\frac{1}{k}\right). \quad (2.14)$$

*Proof.* Let  $w^* \in \Omega^*$  be arbitrary fixed and, for all  $k \geq 0$ , retain the notation  $\varphi_k = \|w^k - w^*\|_G^2$ ,

$$\psi_k = \varphi_{k+1} - \varphi_k - \alpha_k (\varphi_k - \varphi_{k-1}) \quad \text{and} \quad \theta_k = \varphi_k - \varphi_{k-1}.$$

It follows from the first “ $\leq$ ” in (2.6) and  $\lambda_k \geq 0$  that

$$\begin{aligned}\psi_k &\leq -\|w^{k+1} - w^k\|_G^2 + 2\alpha_k \langle w^{k+1} - w^k, w^k - w^{k-1} \rangle_G + \alpha_k \|w^k - w^{k-1}\|_G^2 \\ &\leq -\|w^{k+1} - w^k\|_G^2 + \alpha_k (\|w^{k+1} - w^k\|_G^2 + \|w^k - w^{k-1}\|_G^2) + \alpha_k \|w^k - w^{k-1}\|_G^2 \\ &= -(1 - \alpha_k) \|w^{k+1} - w^k\|_G^2 + 2\alpha_k \|w^k - w^{k-1}\|_G^2,\end{aligned}\tag{2.15}$$

where the second “ $\leq$ ” follows from the Cauchy-Schwartz inequality. Define

$$\mu_k := \varphi_k - \alpha_k \varphi_{k-1} + 2\alpha_k \|w^k - w^{k-1}\|_G^2.$$

From  $0 \leq \alpha_k \leq \alpha_{k+1} \leq \alpha < \frac{1}{3}$ , the fact that  $\varphi_k \geq 0$  and (2.15), we have

$$\begin{aligned}\mu_{k+1} - \mu_k &= \varphi_{k+1} - \alpha_{k+1} \varphi_k + 2\alpha_{k+1} \|w^{k+1} - w^k\|_G^2 - (\varphi_k - \alpha_k \varphi_{k-1} + 2\alpha_k \|w^k - w^{k-1}\|_G^2) \\ &\leq \psi_k + 2\alpha_{k+1} \|w^{k+1} - w^k\|_G^2 - 2\alpha_k \|w^k - w^{k-1}\|_G^2 \\ &\leq -(1 - \alpha_k) \|w^{k+1} - w^k\|_G^2 + 2\alpha_{k+1} \|w^{k+1} - w^k\|_G^2 \\ &\leq -(1 - 3\alpha) \|w^{k+1} - w^k\|_G^2 \\ &\leq 0.\end{aligned}\tag{2.16}$$

Thus,  $\mu_{k+1} \leq \mu_k$  for all  $k \geq 0$ . Note that  $w^0 = w^{-1}$  by our assumption. It follows from the definitions of  $\mu_k$  and  $\varphi_k$  that  $\mu_0 = (1 - \alpha_0) \varphi_0 \leq \varphi_0 := \|w^0 - w^*\|_G^2$ . Therefore, we have

$$-\alpha \varphi_{k-1} \leq \varphi_k - \alpha \varphi_{k-1} \leq \varphi_k - \alpha_k \varphi_{k-1} \leq \mu_k \leq \mu_0 \leq \varphi_0.\tag{2.17}$$

Further take into account (2.16), we obtain

$$\varphi_k \leq \alpha \varphi_{k-1} + \varphi_0 \leq \alpha^k \varphi_0 + \varphi_0 \sum_{j=0}^{k-1} \alpha^j \leq \alpha^k \varphi_0 + \frac{\varphi_0}{1 - \alpha}.\tag{2.18}$$

The second last “ $\leq$ ” in (2.16) implies that  $(1 - 3\alpha) \|w^{k+1} - w^k\|_G^2 \leq \mu_k - \mu_{k+1}$  for  $k \geq 0$ . Together with (2.17) and (2.18), this implies

$$(1 - 3\alpha) \sum_{j=0}^k \|w^{j+1} - w^j\|_G^2 \leq \mu_0 - \mu_{k+1} \leq \varphi_0 + \alpha \varphi_k \leq \alpha^{k+1} \varphi_0 + \frac{\varphi_0}{1 - \alpha} \leq 2\varphi_0,\tag{2.19}$$

where the second inequality is due to  $\mu_0 \leq \varphi_0$  and  $-\alpha \varphi_k \leq \mu_{k+1}$ , the next one follows from (2.18), and the last one is due to  $\alpha < 1/3$ . By taking the limit  $k \rightarrow \infty$ , we obtain

$$\frac{1}{2} \sum_{k=1}^{\infty} \delta_k = \sum_{k=1}^{\infty} \alpha_k \|w^k - w^{k-1}\|_G^2 \leq \alpha \sum_{k=1}^{\infty} \|w^k - w^{k-1}\|_G^2 \leq \frac{2\varphi_0 \alpha}{1 - 3\alpha} := C_1 < \infty.\tag{2.20}$$

The convergence of  $\{w^k\}_{k=0}^{\infty}$  to a solution point in  $\Omega^*$  follows from the proof of Theorem 1.

It follows from (2.9) that, for  $i \geq 0$ ,  $\|w^{i+1} - \bar{w}^i\|_G^2 \leq \varphi_i - \varphi_{i+1} + \alpha[\theta_i]_+ + \delta_i$ , from which we obtain

$$\sum_{i=0}^{k-1} \|w^{i+1} - \bar{w}^i\|_G^2 \leq \varphi_0 - \varphi_k + \alpha \sum_{i=1}^{k-1} [\theta_i]_+ + \sum_{i=1}^{k-1} \delta_i \leq \varphi_0 + \alpha C_2 + 2C_1,\tag{2.21}$$

where  $C_1$  is defined in (2.20) and  $C_2$  is defined as

$$C_2 := \frac{2C_1}{1 - \alpha} \geq \frac{1}{1 - \alpha} \sum_{i=1}^{\infty} \delta_i \geq \sum_{i=1}^{\infty} [\theta_i]_+.$$

Here the first “ $\geq$ ” follows from the definition of  $C_1$  in (2.20) and the second one follows from (2.8). Direct calculation shows that

$$\varphi_0 + \alpha C_2 + 2C_1 = \left[ 1 + \left( \frac{2\alpha}{1-\alpha} + 2 \right) \frac{2\alpha}{1-3\alpha} \right] \varphi_0 \leq \left( 1 + \frac{2}{1-3\alpha} \right) \varphi_0, \quad (2.22)$$

where the “ $\leq$ ” follows from  $\alpha < 1/3$ . The estimate (2.13) follows immediately from (2.21) and (2.22). The  $o(1/k)$  result (2.14) follows from

$$\frac{k-1}{2} \min_{0 \leq i \leq k-1} \|w^{i+1} - \bar{w}^i\|_G^2 \leq \sum_{i=\lfloor \frac{k-1}{2} \rfloor}^{k-1} \|w^{i+1} - \bar{w}^i\|_G^2, \quad (2.23)$$

where  $\lfloor (k-1)/2 \rfloor$  denotes the greatest integer no greater than  $(k-1)/2$ , and the fact that the right-hand-side of (2.23) converges to 0 as  $k \rightarrow \infty$  because  $\sum_{i=0}^{\infty} \|w^{i+1} - \bar{w}^i\|_G^2 < \infty$ .  $\square$

REMARK 4. Note that  $w^{k+1}$  is obtained via a proximal point step from  $\bar{w}^k$ . Thus, the equality  $w^{k+1} = \bar{w}^k$  implies that  $w^{k+1}$  is already a solution of (2.1) (even if  $G$  is only positive semidefinite, see (2.2b)). In this sense, the error estimate given in (2.13) can be viewed as a convergence rate result of the general inertial proximal point method (2.2). In particular, (2.13) implies that, to obtain an  $\varepsilon$ -optimal solution in the sense that  $\|w^{k+1} - \bar{w}^k\|_G^2 \leq \varepsilon$ , the upper bound of iterations required by (2.2) is

$$\frac{\left( 1 + \frac{2}{1-3\alpha} \right) \|w^0 - w^*\|_G^2}{\varepsilon}.$$

REMARK 5. In general Hilbert space, weak convergence of  $\{w^k\}_{k=0}^{\infty}$  to a point in  $\Omega^*$  can still be guaranteed under similar assumptions. The analysis is similar to that of Theorems 1 and 2 by using a well-known result, called Opial’s lemma [35], in functional analysis of Banach space.

**3. Inertial linearized ADMM.** In this section, we prove that under suitable conditions the linearized ADMM is an application of PPA with weighting matrix  $G \in S_{++}^n$ . As byproducts of this result, we establish convergence, ergodic and nonergodic convergence rate results for linearized ADMM within the PPA framework. Furthermore, an inertial version of the linearized ADMM is proposed, whose convergence is guaranteed by Theorems 1 and 2.

Let  $f : \mathfrak{R}^{n_1} \rightarrow \mathfrak{R}$  and  $g : \mathfrak{R}^{n_2} \rightarrow \mathfrak{R}$  be closed convex functions,  $\mathcal{X} \subseteq \mathfrak{R}^{n_1}$  and  $\mathcal{Y} \subseteq \mathfrak{R}^{n_2}$  be closed convex sets. Consider linearly constrained separable convex optimization problem of the form

$$\min_{x,y} \{f(x) + g(y) : \text{s.t. } Ax + By = b, x \in \mathcal{X}, y \in \mathcal{Y}\}, \quad (3.1)$$

where  $A \in \mathfrak{R}^{m \times n_1}$ ,  $B \in \mathfrak{R}^{m \times n_2}$  and  $b \in \mathfrak{R}^m$  are given. We assume that the set of KKT points of (3.1) is nonempty. Under very little assumptions, see, e.g., [36], (3.1) is equivalent to the mixed variational inequality problem (2.1) with  $\Omega$ ,  $w$ ,  $\theta$  and  $F$  given, respectively, by  $\Omega := \mathcal{X} \times \mathcal{Y} \times \mathfrak{R}^m$ ,

$$w = \begin{pmatrix} x \\ y \\ p \end{pmatrix}, \quad \theta(w) := f(x) + g(y), \quad F(w) = \begin{pmatrix} 0 & 0 & -A^T \\ 0 & 0 & -B^T \\ A & B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ p \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}. \quad (3.2)$$

Since the coefficient matrix defining  $F$  is skew-symmetric,  $F$  is monotone, and thus Assumption 2 is satisfied with  $H = 0$ . Let  $\beta > 0$  and define the Lagrangian and the augmented Lagrangian functions, respectively, as

$$\mathcal{L}(x, y, p) := f(x) + g(y) - \langle p, Ax + By - b \rangle, \quad (3.3a)$$

$$\bar{\mathcal{L}}(x, y, p) := \mathcal{L}(x, y, p) + \frac{\beta}{2} \|Ax + By - b\|^2. \quad (3.3b)$$

Given  $(y^k, p^k)$ , the classical ADMM in “ $x - p - y$ ” order iterates as follows:

$$x^{k+1} = \arg \min_{x \in \mathcal{X}} \bar{\mathcal{L}}(x, y^k, p^k), \quad (3.4a)$$

$$p^{k+1} = p^k - \beta(Ax^{k+1} + By^k - b), \quad (3.4b)$$

$$y^{k+1} = \arg \min_{y \in \mathcal{Y}} \bar{\mathcal{L}}(x^{k+1}, y, p^{k+1}). \quad (3.4c)$$

Note that here we still use the latest value of each variable in each step of the alternating computation. Therefore, it is equivalent to the commonly seen ADMM in “ $y - x - p$ ” order in a cyclic sense. We use the order “ $x - p - y$ ” because the resulting algorithm can be easily explained as a PPA-like algorithm applied to the primal-dual optimality conditions, see [37].

Given  $(x^k, y^k, p^k)$  and two parameters  $\tau, \eta > 0$ , the iteration of linearized ADMM in “ $x - p - y$ ” order appears as

$$u^k = A^T(Ax^k + By^k - b), \quad (3.5a)$$

$$x^{k+1} = \arg \min_{x \in \mathcal{X}} f(x) - \langle p^k, Ax \rangle + \frac{\beta}{2\tau} \|x - (x^k - \tau u^k)\|^2, \quad (3.5b)$$

$$p^{k+1} = p^k - \beta(Ax^{k+1} + By^k - b), \quad (3.5c)$$

$$v^k = B^T(Ax^{k+1} + By^k - b), \quad (3.5d)$$

$$y^{k+1} = \arg \min_{y \in \mathcal{Y}} g(y) - \langle p^{k+1}, By \rangle + \frac{\beta}{2\eta} \|y - (y^k - \eta v^k)\|^2. \quad (3.5e)$$

In the following, we prove that under suitable assumptions  $(x^{k+1}, y^{k+1}, p^{k+1})$  generated by (3.5) conforms to the classical PPA with an appropriate symmetric and positive definite weighting matrix  $G$ .

**THEOREM 3.** *Given  $w^k = (x^k, y^k, p^k) \in \Omega$ , then  $w^{k+1} = (x^{k+1}, y^{k+1}, p^{k+1})$  generated by the linearized ADMM framework (3.5) satisfies*

$$w^{k+1} \in \Omega, \theta(w) - \theta(w^{k+1}) + \langle w - w^{k+1}, F(w^{k+1}) + G(w^{k+1} - w^k) \rangle \geq 0, \forall w \in \Omega, \quad (3.6)$$

where

$$G = \begin{pmatrix} \beta \left( \frac{1}{\tau} I - A^T A \right) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\beta}{\eta} I & -B^T \\ \mathbf{0} & -B & \frac{1}{\beta} I \end{pmatrix}. \quad (3.7)$$

Here  $I$  denotes identity matrix of appropriate size.

*Proof.* The optimality conditions of (3.5b) and (3.5e) imply that

$$f(x) - f(x^{k+1}) + (x - x^{k+1})^T \left\{ -A^T p^k + \frac{\beta}{\tau} (x^{k+1} - x^k) + \beta A^T (Ax^k + By^k - b) \right\} \geq 0, \forall x \in \mathcal{X},$$

$$g(y) - g(y^{k+1}) + (y - y^{k+1})^T \left\{ -B^T p^{k+1} + \frac{\beta}{\eta} (y^{k+1} - y^k) + \beta B^T (Ax^{k+1} + By^k - b) \right\} \geq 0, \forall y \in \mathcal{Y}.$$

By noting (3.5c), the above relations can be rewritten as

$$f(x) - f(x^{k+1}) + (x - x^{k+1})^T \left\{ -A^T p^{k+1} + \beta \left( \frac{1}{\tau} I - A^T A \right) (x^{k+1} - x^k) \right\} \geq 0, \forall x \in \mathcal{X}, \quad (3.8a)$$

$$g(y) - g(y^{k+1}) + (y - y^{k+1})^T \left\{ -B^T p^{k+1} + \frac{\beta}{\eta} (y^{k+1} - y^k) - B^T (p^{k+1} - p^k) \right\} \geq 0, \forall y \in \mathcal{Y}. \quad (3.8b)$$

Note that (3.5c) can be equivalently represented as

$$(p - p^{k+1})^T \left\{ (Ax^{k+1} + By^{k+1} - b) - B(y^{k+1} - y^k) + \frac{1}{\beta}(p^{k+1} - p^k) \right\} \geq 0, \quad \forall p \in \mathfrak{R}^m. \quad (3.9)$$

By the notation defined in (3.2), we see that the addition of (3.8a), (3.8b) and (3.9) yields (3.6), with  $G$  defined in (3.7).  $\square$

REMARK 6. *Clearly, the matrix  $G$  defined in (3.7) is symmetric and positive definite provided that the parameters  $\tau$  and  $\eta$  are reasonably small. In particular,  $G$  is positive definite if  $\tau < 1/\rho(A^T A)$  and  $\eta < 1/\rho(B^T B)$ . Using similar analysis, it is easy to verify that  $w^{k+1} = (x^{k+1}, y^{k+1}, p^{k+1})$  generated by the ADMM framework (3.4) conforms to (3.6) with  $G$  defined by*

$$G = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta B^T B & -B^T \\ \mathbf{0} & -B & \frac{1}{\beta} I \end{pmatrix}, \quad (3.10)$$

which is clearly never positive definite. See [37] for details.

For the linearized ADMM framework (3.5), we have the following convergence results. Their proofs are given in the Appendix for convenience of readers. Similar convergence analysis and complexity results can be found in [38, 39], and also [40], where a unified analysis of the proximal method of multipliers is given.

THEOREM 4. *Assume that  $0 < \tau < 1/\rho(A^T A)$  and  $0 < \eta < 1/\rho(B^T B)$ . Let  $\{w^k = (x^k, y^k, p^k)\}_{k=0}^\infty$  be generated by the linearized ADMM framework (3.5) from any starting point  $w^0 = (x^0, y^0, p^0)$ . The following results hold.*

1. *The sequence  $\{w^k = (x^k, y^k, p^k)\}_{k=0}^\infty$  converges to a solution of (2.1), i.e., there exists  $w^* = (x^*, y^*, p^*) \in \Omega^*$  such that  $\lim_{k \rightarrow \infty} w^k = w^*$ . Moreover,  $(x^*, y^*)$  is a solution of (3.1).*
2. *For any fixed integer  $k > 0$ , define  $\bar{w}^k := \frac{1}{k+1} \sum_{i=0}^k w^{i+1}$ . Then, it holds that*

$$\bar{w}^k \in \Omega, \quad \theta(w) - \theta(\bar{w}^k) + (w - \bar{w}^k)^T F(w) \geq -\frac{\|w - w^0\|_G^2}{2(k+1)}, \quad \forall w \in \Omega, \quad (3.11)$$

or, equivalently,

$$\bar{w}^k = (\bar{x}^k, \bar{y}^k, \bar{p}^k) \in \Omega, \quad \mathcal{L}(\bar{x}^k, \bar{y}^k, p) - \mathcal{L}(x, y, \bar{p}^k) \leq \frac{\|w - w^0\|_G^2}{2(k+1)}, \quad \forall w = (x, y, p) \in \Omega. \quad (3.12)$$

Here  $\mathcal{L}$  is the Lagrangian function defined in (3.3a).

3. *After  $k > 0$  iterations, we have*

$$\|w^k - w^{k-1}\|_G^2 \leq \frac{\|w^0 - w^*\|_G^2}{k}. \quad (3.13)$$

Moreover, it holds as  $k \rightarrow \infty$  that

$$\|w^k - w^{k-1}\|_G^2 = o(1/k). \quad (3.14)$$

REMARK 7. *It is not hard to show that the set of solutions  $\Omega^*$  of the mixed VI problem (2.1) can be expressed as the intersection of  $\Omega_w := \{\bar{w} \in \Omega \mid \theta(w) - \theta(\bar{w}) + (w - \bar{w})^T F(w) \geq 0\}$  for all  $w \in \Omega$ , i.e.,*

$$\Omega^* = \bigcap_{w \in \Omega} \Omega_w = \bigcap_{w \in \Omega} \{\bar{w} \in \Omega \mid \theta(w) - \theta(\bar{w}) + (w - \bar{w})^T F(w) \geq 0\}.$$

See, e.g., [41]. Therefore, the result (3.11) essentially assures that after  $k$  iterations an approximate solution  $\bar{w}^k$  with accuracy  $O(1/k)$  can be found. On the other hand, it is easy to show that  $w^* = (x^*, y^*, p^*) \in \Omega^*$  if and only if

$$\mathcal{L}(x^*, y^*, p) - \mathcal{L}(x, y, p^*) \leq 0, \quad \forall w = (x, y, p) \in \Omega. \quad (3.15)$$

Thus, (3.12) can be viewed as an approximation to the optimality condition (3.15). Since  $\bar{w}^k$  is the average of all the points generated in the first  $(k+1)$  iterations, the result (3.11) or (3.12) is usually called an ergodic convergence rate.

REMARK 8. It is easy to see from (3.6) that  $w^{k+1}$  must be a solution if  $w^{k+1} = w^k$ . As such, the difference of two consecutive iterations can be viewed in some sense as a measure of how close the current point is to the solution set. Therefore, the result (3.13) estimates the convergence rate of  $w^k$  to the solution set using the measure  $\|w^k - w^{k-1}\|_G^2$ .

REMARK 9. We note that all the results given in Theorem 4 remain valid if the conditions on  $\tau$  and  $\eta$  are relaxed to  $0 < \tau \leq 1/\rho(A^T A)$  and  $0 < \eta \leq 1/\rho(B^T B)$ , respectively. The proof is a little bit complicated and we refer interested readers to [42, 43].

Now we state the inertial version of the linearized ADMM, which is new to the best of our knowledge. Given  $\beta, \tau, \eta > 0$ , a sequence  $\{\alpha_k \geq 0\}_{k=0}^\infty$ ,  $(x^k, y^k, p^k)$  and  $(x^{k-1}, y^{k-1}, p^{k-1})$ , the inertial linearized ADMM iterates as follows:

$$(\bar{x}^k, \bar{y}^k, \bar{p}^k) = (x^k, y^k, p^k) + \alpha_k(x^k - x^{k-1}, y^k - y^{k-1}, p^k - p^{k-1}) \quad (3.16a)$$

$$u^k = A^T(A\bar{x}^k + B\bar{y}^k - b), \quad (3.16b)$$

$$x^{k+1} = \arg \min_{x \in \mathcal{X}} f(x) - \langle \bar{p}^k, Ax \rangle + \frac{\beta}{2\tau} \|x - (\bar{x}^k - \tau u^k)\|^2, \quad (3.16c)$$

$$p^{k+1} = \bar{p}^k - \beta(Ax^{k+1} + B\bar{y}^k - b), \quad (3.16d)$$

$$v^k = B^T(Ax^{k+1} + B\bar{y}^k - b), \quad (3.16e)$$

$$y^{k+1} = \arg \min_{y \in \mathcal{Y}} g(y) - \langle p^{k+1}, By \rangle + \frac{\beta}{2\eta} \|y - (\bar{y}^k - \eta v^k)\|^2. \quad (3.16f)$$

The following convergence result is a consequence of Theorems 2 and 3.

THEOREM 5. Let  $G$  be defined in (3.7) and  $\{(x^k, y^k, p^k)\}_{k=0}^\infty \subseteq \mathfrak{R}^n$  be generated by (3.16) from any starting point  $(x^0, y^0, p^0) = (x^{-1}, y^{-1}, p^{-1})$ . Suppose that  $0 < \tau < 1/\rho(A^T A)$ ,  $0 < \eta < 1/\rho(B^T B)$  and  $\{\alpha_k\}_{k=0}^\infty$  satisfies, for all  $k$ ,  $0 \leq \alpha_k \leq \alpha_{k+1} \leq \alpha < \frac{1}{3}$ . Then, the sequence  $\{(x^k, y^k, p^k)\}_{k=0}^\infty$  converges to some point in  $\Omega^*$ , the set of solutions of (2.1), as  $k \rightarrow \infty$ . Moreover, it holds that

$$\min_{0 \leq i \leq k-1} \|(x^{i+1}, y^{i+1}, p^{i+1}) - (\bar{x}^i, \bar{y}^i, \bar{p}^i)\|_G^2 = o\left(\frac{1}{k}\right). \quad (3.17)$$

**4. Numerical Results.** In this section, we present numerical results to compare the performance of the linearized ADMM (3.5) (abbreviated as LADMM) and the proposed inertial linearized ADMM (3.16) (abbreviated as iLADMM). Both algorithms were implemented in MATLAB. All the experiments were performed with Microsoft Windows 8 and MATLAB v7.13 (R2011b), running on a 64-bit Lenovo laptop with an Intel Core i7-3667U CPU at 2.00 GHz and 8 GB of memory.

**4.1. Compressive principal component pursuit.** In our experiments, we focused on the compressive principal component pursuit problem [44], which aims to recover low-rank and sparse components from

compressive or incomplete measurements. Let  $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^q$  be a linear operator,  $L_0$  and  $S_0$  be, respectively, low-rank and sparse matrices of size  $m \times n$ . The incomplete measurements are given by  $b = \mathcal{A}(L_0 + S_0)$ . Under certain technical conditions, such as  $L_0$  is  $\mu$ -incoherent, the support of  $S_0$  is randomly distributed with nonzero probability  $\rho$  and the signs of  $S_0$  conform to Bernoulli distribution, it was proved in [44] that the low-rank and the sparse components  $L_0$  and  $S_0$  can be exactly recovered with high probability via solving the convex optimization problem

$$\min_{L, S} \{ \|L\|_* + \lambda \|S\|_1 : \text{s.t. } \mathcal{A}(L + S) = b \}, \quad (4.1)$$

as long as the range space of the adjoint operator  $\mathcal{A}^*$  is randomly distributed according to the Haar measure and its dimension  $q$  is in the order  $O((\rho mn + mr) \log^2 m)$ . Here  $\lambda = 1/\sqrt{m}$  is a constant,  $\|L\|_*$  and  $\|S\|_1$  denote the nuclear norm of  $L$  (sum of all singular values) and the  $\ell_1$  norm of  $S$  (sum of absolute values of all components), respectively. Note that to determine a rank  $r$  matrix, it is sufficient to specify  $(m + n - r)r$  elements. Let the number of nonzeros of  $S_0$  be denoted by  $\text{nnz}(S_0)$ . Without considering the distribution of the support of  $S_0$ , we define the *degree of freedom* of the pair  $(L_0, S_0)$  by

$$\text{dof} := (m + n - r)r + \text{nnz}(S_0). \quad (4.2)$$

The augmented Lagrangian function of (4.1) is given by

$$\bar{\mathcal{L}}(L, S, p) := \|L\|_* + \lambda \|S\|_1 - \langle p, \mathcal{A}(L + S) - b \rangle + \frac{\beta}{2} \|\mathcal{A}(L + S) - b\|^2.$$

One can see that the minimization of  $\bar{\mathcal{L}}$  with respect to either  $L$  or  $S$ , with the other two variables being fixed, does not have closed form solution. To avoid inner loop for iteratively solving ADMM-subproblems, the linearized ADMM framework (3.5) and its inertial version (3.16) can obviously be applied. Note that it is necessary to linearize both ADMM-subproblems in order to avoid inner loops. Though the iterative formulas of LADMM and inertial LADMM for solving (4.1) can be derived very easily based on (3.5) and (3.16), we elaborate them below for clearness and subsequent references. Let  $(L^k, S^k, p^k)$  be given. The LADMM framework (3.5) for solving (4.1) appears as

$$U^k = \mathcal{A}^*(\mathcal{A}(L^k + S^k) - b), \quad (4.3a)$$

$$L^{k+1} = \arg \min_L \|L\|_* - \langle p^k, \mathcal{A}(L) \rangle + \frac{\beta}{2\tau} \|L - (L^k - \tau U^k)\|_F^2, \quad (4.3b)$$

$$p^{k+1} = p^k - \beta(\mathcal{A}(L^{k+1} + S^k) - b), \quad (4.3c)$$

$$V^k = \mathcal{A}^*(\mathcal{A}(L^{k+1} + S^k) - b), \quad (4.3d)$$

$$S^{k+1} = \arg \min_S \lambda \|S\|_1 - \langle p^{k+1}, \mathcal{A}(S) \rangle + \frac{\beta}{2\eta} \|S - (S^k - \eta V^k)\|_F^2. \quad (4.3e)$$

The inertial LADMM framework (3.16) for solving (4.1) appears as

$$(\bar{L}^k, \bar{S}^k, \bar{p}^k) = (L^k, S^k, p^k) + \alpha_k(L^k - L^{k-1}, S^k - S^{k-1}, p^k - p^{k-1}), \quad (4.4a)$$

$$U^k = \mathcal{A}^*(\mathcal{A}(\bar{L}^k + \bar{S}^k) - b), \quad (4.4b)$$

$$L^{k+1} = \arg \min_L \|L\|_* - \langle \bar{p}^k, \mathcal{A}(L) \rangle + \frac{\beta}{2\tau} \|L - (\bar{L}^k - \tau U^k)\|_F^2, \quad (4.4c)$$

$$p^{k+1} = \bar{p}^k - \beta(\mathcal{A}(L^{k+1} + \bar{S}^k) - b), \quad (4.4d)$$

$$V^k = \mathcal{A}^*(\mathcal{A}(L^{k+1} + \bar{S}^k) - b), \quad (4.4e)$$

$$S^{k+1} = \arg \min_S \lambda \|S\|_1 - \langle p^{k+1}, \mathcal{A}(S) \rangle + \frac{\beta}{2\eta} \|S - (\bar{S}^k - \eta V^k)\|_F^2. \quad (4.4f)$$

Note that the subproblems (4.3b) (or (4.4c)) and (4.3e) (or (4.4f)) have closed form solutions given, respectively, by the shrinkage operators of matrix nuclear norm and vector  $\ell_1$  norm, see, e.g., [45, 46]. The main computational cost per iteration of both algorithms is one singular value decomposition (SVD) required in solving the  $L$ -subproblem.

**4.2. Generating experimental data.** In our experiments, we set  $m = n$  and tested different ranks of  $L_0$  (denoted by  $r$ ), sparsity levels of  $S_0$  (i.e.,  $\text{nnz}(S_0)/(mn)$ ) and sample ratios (i.e.,  $q/(mn)$ ). The low-rank matrix  $L_0$  was generated by  $\text{randn}(m, r) * \text{randn}(r, n)$  in MATLAB. The support of  $S_0$  is randomly determined by uniform distribution, while the values of its nonzeros are uniformly distributed in  $[-10, 10]$ . Such type of synthetic data are roughly those tested in [44]. As for the linear operator  $\mathcal{A}$ , we tested three types of linear operators, i.e., two-dimensional partial DCT (discrete cosine transform), FFT (fast Fourier transform) and WHT (Walsh-Hadamard transform). The rows of these transforms are selected uniformly at random.

**4.3. Parameters, stopping criterion and initialization.** The model parameter  $\lambda$  was set to  $1/\sqrt{m}$  in our experiments, which is determined based on the exact recoverability theory in [44]. As for the other parameters ( $\beta$ ,  $\tau$  and  $\eta$ ) common to LADMM and iLADMM, we used the same set of values and adaptive rules in all the tests. Now we elaborate how the parameters are chosen. Since  $\mathcal{A}$  contains rows of orthonormal transforms, it holds that  $\mathcal{A}\mathcal{A}^* = \mathcal{I}$ , the identity operator. Therefore, it holds that  $\rho(\mathcal{A}^*\mathcal{A}) = 1$ . We set  $\tau = \eta = 0.99$ , which satisfies the convergence requirement specified in Theorems 4 and 5. The penalty parameter  $\beta$  was initialized at  $0.1q/\|b\|_1$  and was tuned at the beginning stage of the algorithm. Specifically, we tuned  $\beta$  within the first 30 iterations according to the following rule:

$$\beta_{k+1} = \begin{cases} \max(0.5\beta_k, 10^{-3}), & \text{if } r_k < 0.1; \\ \min(2\beta_k, 10^2), & \text{if } r_k > 5; \\ \beta_k, & \text{otherwise,} \end{cases} \quad \text{where } r_k := \frac{\beta_k \|\mathcal{A}(L^k + S^k) - b\|^2}{2s_k(\|L^k\|_* + \lambda\|S^k\|_1)}.$$

Here  $s_k$  is a parameter attached to the objective function  $\|L\|_* + \lambda\|S\|_1$  and was chosen adaptively so that the quadratic term  $\frac{\beta}{2}\|\mathcal{A}(L+S) - b\|^2$  and the objective term  $\|L\|_* + \lambda\|S\|_1$  remain roughly in the same order. Note that the choice of  $\beta$  does not have much theory and is usually determined via numerical experiments, see, e.g., [47] for the influence of different  $\beta$ 's in linearized ADMM for matrix completion problem. The extrapolation parameter  $\alpha_k$  for iLADMM was set to 0.28 and held constant in all our experiments. Note that this value of  $\alpha_k$  is determined based on experiments and may be far from optimal. How to select  $\alpha_k$  adaptively to achieve stable and faster convergence remains a research issue. Here our main goal is to illustrate the effect of the extrapolation steps. We also present some numerical results to compare the performance of iLADMM with different constant strategies for  $\alpha_k$ .

It is easy to see from (3.6) that if two consecutive iterates generated by proximal point method are identical then a solution is already obtained. Since LADMM is an application of a general PPA, we terminated it by the following rule

$$\frac{\|(L^{k+1}, S^{k+1}, p^{k+1}) - (L^k, S^k, p^k)\|}{1 + \|(L^k, S^k, p^k)\|} < \varepsilon, \quad (4.5)$$

where  $\varepsilon > 0$  is a tolerance parameter. Here  $\|(L, S, p)\| := \sqrt{\|L\|_F^2 + \|S\|_F^2 + \|p\|^2}$ . Since iLADMM generates the new point  $(L^{k+1}, S^{k+1}, p^{k+1})$  by applying proximal point method to  $(\bar{L}^k, \bar{S}^k, \bar{p}^k)$ , we used the same stopping rule as (4.5) except that  $(L^k, S^k, p^k)$  is replaced by  $(\bar{L}^k, \bar{S}^k, \bar{p}^k)$ . That is

$$\frac{\|(L^{k+1}, S^{k+1}, p^{k+1}) - (\bar{L}^k, \bar{S}^k, \bar{p}^k)\|}{1 + \|(\bar{L}^k, \bar{S}^k, \bar{p}^k)\|} < \varepsilon. \quad (4.6)$$

TABLE 4.1  
*Results of  $\text{rank}(L_0) = 5$ :  $\varepsilon = 10^{-5}$ , average results of 10 random trials.*

$m = n = 1024$				LADMM			iLADMM			iter2 iter1
$r$	$k/m^2$	$(q/m^2, q/\text{dof})$	$\mathcal{A}$	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter1	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter2	
5	1%	(40%, 20.26)	pdct	2.24e-5	4.45e-5	260.6	1.46e-5	3.97e-5	199.4	0.77
			pfft	1.68e-5	3.78e-5	223.8	1.35e-5	4.00e-5	173.3	0.77
			pwht	2.92e-5	5.24e-5	271.8	1.73e-5	3.99e-5	206.1	0.76
		(60%, 30.39)	pdct	1.11e-5	2.24e-5	170.3	1.24e-5	2.65e-5	123.1	0.72
			pfft	1.14e-5	2.24e-5	134.6	1.13e-5	2.67e-5	98.4	0.73
			pwht	1.23e-5	2.32e-5	160.9	1.12e-5	2.37e-5	119.3	0.74
		(80%, 40.52)	pdct	1.20e-5	1.28e-5	89.7	8.51e-6	1.52e-5	61.5	0.69
			pfft	3.68e-6	1.00e-5	67.2	7.06e-6	1.44e-5	42.2	0.63
			pwht	8.10e-6	1.09e-5	88.3	6.80e-6	1.22e-5	60.4	0.68
	5%	(40%, 6.70)	pdct	1.03e-5	4.07e-5	409.7	1.02e-5	3.72e-5	321.5	0.78
			pfft	1.32e-5	4.82e-5	349.5	1.10e-5	3.80e-5	278.4	0.80
			pwht	1.04e-5	4.14e-5	408.8	1.04e-5	3.74e-5	320.4	0.78
		(60%, 10.04)	pdct	8.54e-6	2.68e-5	198.5	5.35e-6	1.99e-5	149.5	0.75
			pfft	7.92e-6	3.13e-5	188.4	5.03e-6	1.84e-5	136.9	0.73
			pwht	8.12e-6	2.75e-5	204.0	5.19e-6	1.86e-5	150.1	0.74
		(80%, 13.39)	pdct	9.39e-6	1.93e-5	107.9	4.32e-6	1.39e-5	77.9	0.72
			pfft	5.94e-6	1.78e-5	99.7	3.54e-6	1.40e-5	70.1	0.70
			pwht	5.62e-6	1.50e-5	107.6	4.10e-6	1.43e-5	77.4	0.72
	10%	(40%, 3.64)	pdct	2.58e-2	7.19e-2	—	1.20e-5	4.08e-5	879.0	—
			pfft	1.53e-5	5.21e-5	757.1	1.21e-5	4.02e-5	563.3	0.74
			pwht	2.50e-2	6.93e-2	—	1.21e-5	4.02e-5	872.7	—
		(60%, 5.47)	pdct	8.58e-6	3.03e-5	344.9	7.64e-6	2.50e-5	243.9	0.71
			pfft	9.19e-6	3.23e-5	309.5	7.27e-6	2.42e-5	239.8	0.77
			pwht	8.49e-6	3.08e-5	345.5	7.77e-6	2.50e-5	243.8	0.71
(80%, 7.29)		pdct	8.02e-6	2.13e-5	170.7	6.45e-6	1.71e-5	123.8	0.73	
		pfft	7.82e-6	1.79e-5	160.7	5.98e-6	1.39e-5	116.0	0.72	
		pwht	8.12e-6	2.00e-5	171.0	6.54e-6	1.83e-5	124.0	0.73	

In our experiments, we initialize all variables  $L$ ,  $S$  and  $p$  at zeros.

**4.4. Experimental results.** Recall that the matrix size is  $m \times n$ , the number of measurements is  $q$ , the rank of  $L_0$  is  $r$ , and the degree of freedom of the pair  $(L_0, S_0)$  is defined in (4.2). In our experiments, we tested  $m = n = 1024$ . Let  $k$  be the number of nozeros of  $S_0$ . We tested four different ranks for  $L_0$ , three levels of sparsity for  $S_0$  and four levels of sample ratios. Specifically, in our experiments we tested  $r \in \{5, 10, 15, 20\}$ ,  $k/m^2 \in \{0.01, 0.05, 0.10\}$  and  $q/m^2 \in \{0.4, 0.6, 0.8\}$ .

Let  $(L, S)$  be the recovered solution. For each setting, we report the relative errors of  $L$  and  $S$  to the true low-rank and sparse matrices  $L_0$  and  $S_0$ , i.e.,  $\|L - L_0\|_F / \|L_0\|_F$  and  $\|S - S_0\|_F / \|S_0\|_F$ , and the number of iterations to meet the condition (4.5) or (4.6), which are denoted by iter1 and iter2 for LADMM and iLADMM, respectively. We terminated both algorithms if the number of iterations reached 1000 but the stopping rule (4.5) or (4.6) still did not hold. For each problem scenario, we run 10 random trials for both algorithms and report the averaged results. Detailed experimental results for  $\varepsilon = 10^{-5}$  and  $r = 5, 10, 15$  and 20 are given in Tables 4.1-4.4, respectively. In each table, dashed line “—” represents that the maximum iteration number was reached.

It can be seen from Tables 4.1-4.4 that iLADMM is generally faster than LADMM to obtain solutions satisfying the aforementioned conditions. Specifically, within our setting the numbers of iterations consumed

TABLE 4.2  
*Results of rank( $L_0$ ) = 10:  $\varepsilon = 10^{-5}$ , average results of 10 random trials.*

$m = n = 1024$				LADMM			iLADMM				
$r$	$k/m^2$	$(q/m^2, q/\text{dof})$	$\mathcal{A}$	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter1	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter2	$\frac{\text{iter2}}{\text{iter1}}$	
10	1%	(40%, 13.59)	pdct	2.20e-5	5.42e-5	340.5	1.74e-5	5.33e-5	253.0	0.74	
			pfft	1.87e-5	5.09e-5	300.1	1.65e-5	5.46e-5	226.8	0.76	
			pwht	2.22e-5	5.50e-5	350.5	1.91e-5	5.54e-5	257.4	0.73	
		(60%, 20.38)	pdct	2.12e-5	3.85e-5	198.1	1.20e-5	3.16e-5	141.6	0.71	
			pfft	1.03e-5	2.75e-5	163.1	9.56e-6	3.07e-5	117.0	0.72	
			pwht	1.84e-5	3.61e-5	191.0	1.32e-5	3.38e-5	135.9	0.71	
		(80%, 27.18)	pdct	7.82e-6	1.55e-5	101.1	7.16e-6	1.60e-5	66.6	0.66	
			pfft	8.63e-6	1.67e-5	79.3	6.36e-6	1.70e-5	47.0	0.59	
			pwht	7.43e-6	1.40e-5	97.8	6.74e-6	1.56e-5	64.5	0.66	
		5%	(40%, 5.76)	pdct	9.22e-6	3.91e-5	536.2	1.41e-5	4.95e-5	406.5	0.76
				pfft	1.30e-5	5.29e-5	448.9	1.15e-5	4.31e-5	346.6	0.77
				pwht	9.16e-6	3.93e-5	537.4	1.42e-5	5.08e-5	407.7	0.76
	(60%, 8.64)		pdct	7.86e-6	3.18e-5	241.9	9.93e-6	3.28e-5	183.4	0.76	
			pfft	8.45e-6	3.55e-5	223.3	9.87e-6	3.24e-5	170.1	0.76	
			pwht	1.42e-5	3.71e-5	244.6	9.32e-6	3.21e-5	184.4	0.75	
	(80%, 11.52)		pdct	8.78e-6	2.17e-5	120.4	4.36e-6	1.78e-5	84.5	0.70	
			pfft	6.74e-6	2.82e-5	109.8	4.22e-6	1.76e-5	76.3	0.69	
			pwht	8.55e-6	2.01e-5	121.2	4.61e-6	1.68e-5	85.1	0.70	
	10%		(40%, 3.35)	pdct	6.84e-2	2.24e-1	—	2.54e-2	8.67e-2	—	—
				pfft	1.09e-5	4.23e-5	965.2	1.26e-5	4.39e-5	702.5	0.73
				pwht	6.79e-2	2.21e-1	—	2.48e-2	8.38e-2	—	—
		(60%, 5.02)	pdct	8.02e-6	3.27e-5	398.4	7.75e-6	2.76e-5	299.1	0.75	
			pfft	9.53e-6	3.58e-5	352.3	7.75e-6	2.75e-5	266.9	0.76	
			pwht	8.13e-6	3.27e-5	397.6	7.70e-6	2.75e-5	298.4	0.75	
(80%, 6.70)		pdct	8.49e-6	2.57e-5	185.9	7.08e-6	2.08e-5	132.6	0.71		
		pfft	8.62e-6	2.59e-5	172.8	6.85e-6	1.84e-5	121.2	0.70		
		pwht	8.49e-6	2.43e-5	185.9	7.06e-6	2.21e-5	132.4	0.71		

by iLADMM range, roughly, from 60%–80% of those consumed by LADMM. If we take into account all the tests (except those cases where either LADMM or iLADMM failed to terminate within 1000 iterations, e.g.,  $(r, k/m^2, q/m^2) = (5, 0.1, 40\%)$  and  $\mathcal{A}$  is partial DCT), the overall average number of iterations used by iLADMM is about 74% of that used by LADMM. Note that in some cases iLADMM obtained satisfactory results within the number of allowed iterations (1000 in our setting), while LADMM did not. For example,  $(r, k/m^2, q/m^2) = (5, 0.1, 40\%)$  and  $\mathcal{A}$  is partial DCT or partial WHT. In most cases, the recovered matrices  $L$  and  $S$  are close to the true low-rank and sparse components  $L_0$  and  $S_0$ , respectively. The relative errors are usually in the order  $10^{-5}$ – $10^{-6}$ . For some cases, the recovered solutions are not of high quality (relative errors are large), which is mainly because the number of samples are small relative to the degree of freedom of  $(L_0, S_0)$ . This can be seen from the values of  $q/\text{dof}$  listed in the tables. Roughly speaking, the recovered solutions are satisfactory (say, relative errors are less than  $10^{-3}$ ) provided that  $q/\text{dof}$  is no less than 3.5.

We note that the per iteration cost of both LADMM and iLADMM for the compressive principal pursuit model (4.1) is dominated by one SVD and thus is roughly identical. The extra cost of the extrapolation inertial step in (4.4a) is negligible compared to the computational load of SVD. This is the main reason that we only reported the number of iterations but not CPU time consumed by both algorithms. The inertial technique actually accelerates the original algorithm to a large extent but without increasing the

TABLE 4.3  
 Results of  $\text{rank}(L_0) = 15$ :  $\varepsilon = 10^{-5}$ , average results of 10 random trials.

$m = n = 1024$				LADMM			iLADMM			iter2 iter1
$r$	$k/m^2$	$(q/m^2, q/\text{dof})$	$\mathcal{A}$	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter1	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter2	
15	1%	(40%, 10.23)	pdct	1.87e-5	5.29e-5	418.6	1.78e-5	5.87e-5	310.5	0.74
			pfft	1.78e-5	5.23e-5	367.5	1.57e-5	5.68e-5	276.1	0.75
			pwht	1.59e-5	4.86e-5	415.5	1.59e-5	5.49e-5	311.1	0.75
		(60%, 15.35)	pdct	2.01e-5	4.31e-5	223.7	1.26e-5	3.59e-5	158.2	0.71
			pfft	1.53e-5	3.73e-5	193.9	1.16e-5	3.77e-5	137.0	0.71
			pwht	2.20e-5	4.55e-5	225.1	1.20e-5	3.47e-5	159.0	0.71
		(80%, 20.47)	pdct	8.30e-6	1.72e-5	112.7	6.86e-6	1.67e-5	72.7	0.64
			pfft	1.06e-5	1.98e-5	90.5	7.07e-6	2.08e-5	51.2	0.57
			pwht	1.17e-5	2.28e-5	117.4	8.37e-6	1.95e-5	76.5	0.65
	5%	(40%, 5.06)	pdct	2.08e-5	7.79e-5	674.1	1.33e-5	5.12e-5	506.5	0.75
			pfft	1.28e-5	5.46e-5	548.2	1.60e-5	5.84e-5	414.0	0.76
			pwht	2.07e-5	7.40e-5	673.6	1.34e-5	5.20e-5	505.0	0.75
		(60%, 7.59)	pdct	7.46e-6	3.45e-5	282.4	1.02e-5	3.79e-5	209.0	0.74
			pfft	9.17e-6	3.94e-5	256.7	1.12e-5	3.91e-5	190.8	0.74
			pwht	6.90e-6	3.41e-5	282.6	1.02e-5	3.80e-5	210.3	0.74
		(80%, 10.12)	pdct	6.76e-6	2.29e-5	134.1	5.00e-6	2.02e-5	93.3	0.70
			pfft	7.24e-6	2.66e-5	121.3	4.96e-6	2.01e-5	82.5	0.68
			pwht	7.42e-6	2.06e-5	136.5	5.58e-6	2.02e-5	93.3	0.68
	10%	(40%, 3.10)	pdct	1.01e-1	3.50e-1	—	6.15e-2	2.25e-1	—	—
			pfft	2.61e-2	8.19e-2	—	1.40e-5	4.98e-5	888.4	—
			pwht	1.02e-1	3.51e-1	—	6.22e-2	2.27e-1	—	—
		(60%, 4.65)	pdct	7.98e-6	3.37e-5	455.3	7.99e-6	2.93e-5	336.5	0.74
			pfft	9.79e-6	3.87e-5	394.3	8.41e-6	3.07e-5	294.9	0.75
			pwht	7.78e-6	3.35e-5	454.9	7.90e-6	2.86e-5	336.6	0.74
(80%, 6.20)		pdct	8.96e-6	2.81e-5	201.9	7.75e-6	2.16e-5	143.2	0.71	
		pfft	9.38e-6	3.12e-5	186.0	7.63e-6	2.53e-5	129.0	0.69	
		pwht	8.93e-6	2.75e-5	202.6	7.86e-6	2.48e-5	142.5	0.70	

total computational cost.

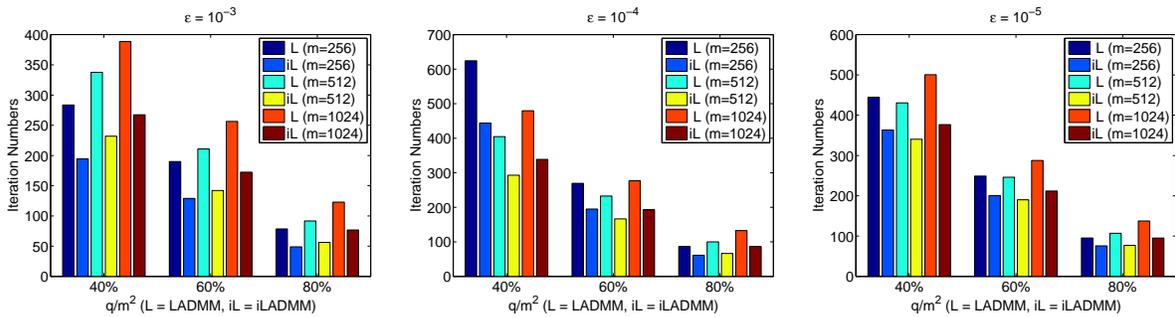


FIG. 4.1. Comparison results on different matrix sizes and stopping tolerance: Average results of 10 random trials ( $m = n = 256, 512, 1024$ , and from left to right  $\varepsilon = 10^{-3}, 10^{-4}, 10^{-5}$ , respectively).

To better understand the behavior of iLADMM relative to LADMM, we also tested different matrix sizes ( $m = n = 256, 512$  and  $1024$ ) with different levels of stopping tolerance ( $\varepsilon = 10^{-3}, 10^{-4}$  and  $10^{-5}$  in (4.5)). For each case, we tested  $r \in \{5, 10, 15, 20\}$  and  $k/m^2 \in \{0.01, 0.05, 0.10\}$  for a fixed  $q$  such that  $q/m^2 \in \{0.4, 0.6, 0.8\}$ . For each  $q$ , we accumulated the iteration numbers for different  $(r, k)$  and the three

TABLE 4.4  
 Results of  $\text{rank}(L_0) = 20$ :  $\varepsilon = 10^{-5}$ , average results of 10 random trials.

$m = n = 1024$				LADMM			iLADMM			iter2 iter1	
$r$	$k/m^2$	$(q/m^2, q/\text{dof})$	$\mathcal{A}$	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter1	$\frac{\ L-L_0\ _F}{\ L_0\ _F}$	$\frac{\ S-S_0\ _F}{\ S_0\ _F}$	iter2		
20	1%	(40%, 8.22)	pdct	2.21e-5	6.28e-5	506.4	1.83e-5	6.41e-5	374.9	0.74	
			pfft	1.83e-5	6.62e-5	434.6	1.81e-5	6.93e-5	327.7	0.75	
			pwht	1.55e-5	5.48e-5	500.9	1.62e-5	6.12e-5	375.2	0.75	
		(60%, 12.33)	pdct	1.19e-5	3.38e-5	249.8	1.15e-5	3.83e-5	176.1	0.70	
			pfft	1.12e-5	3.32e-5	218.2	1.00e-5	3.88e-5	155.9	0.71	
			pwht	2.11e-5	4.71e-5	253.6	1.57e-5	4.68e-5	176.4	0.70	
		(80%, 16.43)	pdct	1.29e-5	2.84e-5	123.8	7.86e-6	2.31e-5	78.5	0.63	
			pfft	1.03e-5	2.47e-5	100.4	5.63e-6	2.08e-5	56.6	0.56	
			pwht	1.20e-5	2.50e-5	128.2	7.84e-6	2.51e-5	81.2	0.63	
		5%	(40%, 4.51)	pdct	1.89e-5	7.60e-5	837.6	1.72e-5	6.57e-5	616.0	0.74
				pfft	1.22e-5	5.35e-5	655.8	1.73e-5	6.64e-5	488.6	0.75
				pwht	1.89e-5	7.65e-5	832.9	1.73e-5	6.52e-5	613.1	0.74
	(60%, 6.77)		pdct	7.29e-6	3.62e-5	325.5	1.04e-5	4.31e-5	237.8	0.73	
			pfft	9.55e-6	4.30e-5	292.2	1.17e-5	4.06e-5	214.4	0.73	
			pwht	7.11e-6	3.66e-5	323.8	1.09e-5	4.24e-5	236.9	0.73	
	(80%, 9.02)		pdct	1.10e-5	2.85e-5	149.2	5.43e-6	2.28e-5	101.9	0.68	
			pfft	7.68e-6	3.13e-5	133.7	5.26e-6	2.47e-5	88.6	0.66	
			pwht	9.83e-6	2.67e-5	151.4	6.33e-6	2.30e-5	102.8	0.68	
	10%		(40%, 2.88)	pdct	1.32e-1	4.65e-1	—	9.35e-2	3.52e-1	—	—
				pfft	6.03e-2	1.98e-1	—	1.25e-2	4.32e-2	—	—
				pwht	1.32e-1	4.62e-1	—	9.35e-2	3.49e-1	—	—
		(60%, 4.33)	pdct	7.23e-6	3.27e-5	517.0	1.22e-5	4.09e-5	375.3	0.73	
			pfft	9.63e-6	4.08e-5	441.0	8.89e-6	3.32e-5	325.2	0.74	
			pwht	7.41e-6	3.33e-5	516.5	1.24e-5	4.26e-5	374.9	0.73	
(80%, 5.77)		pdct	9.20e-6	3.05e-5	219.3	8.41e-6	2.43e-5	154.1	0.70		
		pfft	9.81e-6	3.23e-5	200.1	8.14e-6	2.79e-5	137.3	0.69		
		pwht	9.23e-6	3.16e-5	219.7	8.46e-6	2.49e-5	154.0	0.70		

types of linear operators and took an average finally. The results are summarized in Figure 4.1. Again, these results are average of 10 random trials for each case. From the results we can see that iLADMM is faster and terminates earlier than LADMM with different levels of stopping tolerance. Roughly speaking, iLADMM reduced the cost of LADMM by about 30%.

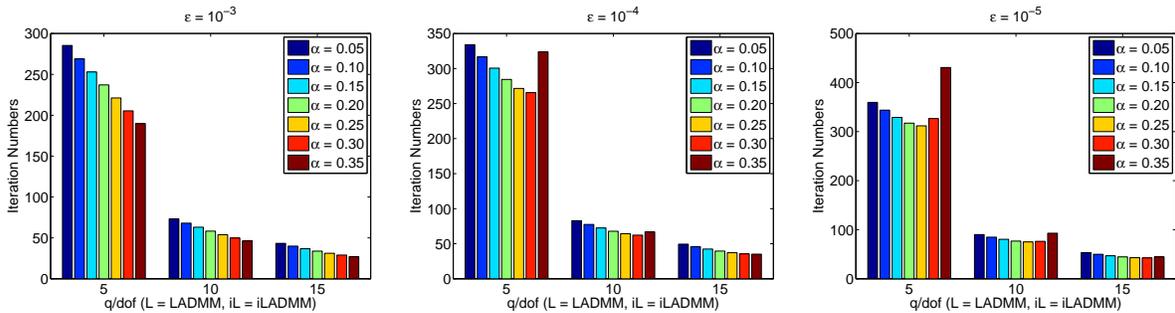


FIG. 4.2. Comparison results on different  $\alpha_k \equiv \alpha$  and stopping tolerance: Average results of 10 random trials ( $m = n = 512$ ,  $\alpha$  ranges from 0.05 to 0.35, and from left to right  $\varepsilon = 10^{-3}, 10^{-4}, 10^{-5}$ , respectively).

We also run iLADMM with various constant strategies for  $\alpha_k$ . In particular, we set  $m = n = 512$  and

tested different values of  $q$  such that  $q/\text{dof} \in \{5, 10, 15\}$ . For each case, we varied  $r \in \{5, 10, 15, 20\}$  and  $k/m^2 \in \{0.01, 0.05, 0.10\}$  for the three types of aforementioned measurement matrices. We accumulated the number of iterations and took an average finally. The detailed average results of 10 random trials for  $\alpha_k \equiv \alpha$  from 0.05 to 0.35 are given in Figure 4.2.

From the results in Figure 4.2 we see that, for the tested 7 values of  $\alpha$ , iLADMM is slightly faster if  $\alpha$  is larger, provided that  $\alpha$  does not exceed 0.3. We have also observed that for  $\alpha > 0.3$  iLADMM either slows down or performs not very stable, especially when  $q/\text{dof}$  is small. This is the main reason that we set  $\alpha_k$  a constant value that is near 0.3 but not larger.

**5. Concluding remarks.** In this paper, we proposed and analyzed a general inertial proximal point method within the setting of mixed VI problem (2.1). The proposed method adopts a weighting matrix and allows more flexibility. Our convergence results require weaker conditions in the sense that the weighting matrix  $G$  does not necessarily be positive definite, as long as the function  $F$  is  $H$ -monotone and  $G$  is positive definite in the null space of  $H$ . The convergence analysis can be easily adapted to the monotone inclusion problem (1.1). We also showed that the linearized ADMM for linearly constrained separable convex optimization problem is a proximal point method applied to the primal-dual optimality conditions, as long as the parameters are reasonably small. As byproducts of this finding, we established with standard analytic techniques for proximal point method the global convergence and convergence rate results of LADMM. This proximal reformulation also allows us to propose an inertial version of LADMM, whose convergence is guaranteed under suitable conditions. Our preliminary implementation of the algorithm and extensive experimental results on compressive principal component pursuit problem have shown that the inertial LADMM is generally faster than the original LADMM. Though in a sense the acceleration is not very significant, we note that the inertial LADMM does not require any additional and unnegligible computational cost either.

Throughout our experiments the extrapolation steplength  $\alpha_k$  held constant. How to select  $\alpha_k$  adaptively based on the current information such that the overall algorithm performs more efficiently and stable is a practically very important question and deserves further investigation. Another theoretical issue is to investigate worst-case complexity analysis for general inertial type algorithms. In fact, complexity results of inertial type algorithms for minimizing closed proper convex functions already exist in the literature. The pioneering work in this direction is due to Nesterov [32], where the algorithm can also be viewed in the perspective of inertial algorithms. Refined analyses for more general problems can be found in [48, 11]. Let  $f : \mathfrak{R}^n \rightarrow \mathfrak{R}$  be a closed proper convex function and be bounded below. Based on [32, 48, 11], the following algorithm can be studied. Let  $x^0 \in \mathfrak{R}^n$  be given. Set  $x_0 = x^{-1}$ ,  $t_0 = 1$  and  $k = 0$ . For  $k \geq 0$  the algorithm iterates as

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad (5.1a)$$

$$\bar{w}^k = w^k + \frac{t_k - 1}{t_{k+1}}(w^k - w^{k-1}), \quad (5.1b)$$

$$w^{k+1} = \arg \min_w f(w) + \frac{1}{2\lambda_k} \|w - \bar{w}^k\|^2. \quad (5.1c)$$

Using analyses similar to those in [32, 11, 48], one can show that the sequence  $\{w^k\}_{k=0}^\infty$  satisfies

$$f(w^k) - \min_{w \in \mathfrak{R}^n} f(w) = O(1/k^2).$$

Algorithm (5.1) is nothing but an inertial PPA with steplength  $\alpha_k = \frac{t_k - 1}{t_{k+1}}$ . It is interesting to note that  $\alpha_k$  is monotonically increasing as  $k \rightarrow \infty$  and converges to 1, which is much larger than the upper bound condition

$\alpha < 1/3$  required in Theorem 2. Also note that the convergence for (5.1) is measured by the objective residue. Without further assumptions on  $f$ , it seems difficult to establish convergence of the sequence  $\{w^k\}_{k=0}^\infty$ , see, e.g., [11]. In comparison, our results impose smaller upper bound on  $\alpha_k$  but guarantee the convergence of the sequence of iterates  $\{w^k\}_{k=0}^\infty$ . Even though, there seems to be certain gap between the classical results [32, 48, 11] for minimizing closed proper convex functions and the results presented in the present paper. Further research in this direction is interesting.

#### REFERENCES

- [1] B. Martinet, “Régularisation d’inéquations variationnelles par approximations successives,” *Rev. Française Informat. Recherche Opérationnelle*, vol. 4, no. Ser. R-3, pp. 154–158, 1970.
- [2] J.-J. Moreau, “Proximité et dualité dans un espace hilbertien,” *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.
- [3] R. T. Rockafellar, “Monotone operators and the proximal point algorithm,” *SIAM J. Control Optimization*, vol. 14, no. 5, pp. 877–898, 1976.
- [4] M. R. Hestenes, “Multiplier and gradient methods,” *J. Optimization Theory Appl.*, vol. 4, pp. 303–320, 1969.
- [5] M. J. D. Powell, “A method for nonlinear constraints in minimization problems,” in *Optimization (Sympos., Univ. Keele, Keele, 1968)*. London: Academic Press, 1969, pp. 283–298.
- [6] J. Douglas, Jr. and H. H. Rachford, Jr., “On the numerical solution of heat conduction problems in two and three space variables,” *Trans. Amer. Math. Soc.*, vol. 82, pp. 421–439, 1956.
- [7] R. Glowinski and A. Marrocco, “Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité, d’une classe de problèmes de Dirichlet non linéaires,” *R.A.I.R.O., R2*, vol. 9, no. R-2, pp. 41–76, 1975.
- [8] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation,” *Computers and Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.
- [9] R. T. Rockafellar, “Augmented Lagrangians and applications of the proximal point algorithm in convex programming,” *Math. Oper. Res.*, vol. 1, no. 2, pp. 97–116, 1976.
- [10] J. Eckstein and D. P. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Math. Programming*, vol. 55, no. 3, Ser. A, pp. 293–318, 1992.
- [11] O. Güler, “New proximal point algorithms for convex minimization,” *SIAM J. Optim.*, vol. 2, no. 4, pp. 649–664, 1992.
- [12] R. E. Bruck, Jr., “Asymptotic convergence of nonlinear contraction semigroups in Hilbert space,” *J. Funct. Anal.*, vol. 18, pp. 15–26, 1975.
- [13] B. T. Poljak, “Some methods of speeding up the convergence of iterative methods,” *Ž. Vychisl. Mat. i Mat. Fiz.*, vol. 4, pp. 791–803, 1964.
- [14] F. Alvarez, “On the minimizing property of a second order dissipative system in Hilbert spaces,” *SIAM J. Control Optim.*, vol. 38, no. 4, pp. 1102–1119 (electronic), 2000.
- [15] F. Aluffi-Pentini, V. Parisi, and F. Zirilli, “Algorithm 617. DAFNE: a differential-equations algorithm for nonlinear equations,” *ACM Trans. Math. Software*, vol. 10, no. 3, pp. 317–324, 1984.
- [16] A. S. Antipin, “Minimization of convex functions on convex sets by means of differential equations,” *Differential Equations*, vol. 30, no. 9, pp. 1475–1486, 1652, 1994.
- [17] F. Alvarez and H. Attouch, “An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping,” *Set-Valued Anal.*, vol. 9, no. 1-2, pp. 3–11, 2001, wellposedness in optimization and related topics (Gargnano, 1999).
- [18] A. Moudafi and E. Elissabeth, “Approximate inertial proximal methods using the enlargement of maximal monotone operators,” *International Journal of Pure and Applied Mathematics*, vol. 5, no. 3, pp. 283–299, 2003.
- [19] F. Alvarez, “Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space,” *SIAM J. Optim.*, vol. 14, no. 3, pp. 773–782 (electronic), 2004.
- [20] P.-E. Maingé and A. Moudafi, “A proximal method for maximal monotone operators via discretization of a first order dissipative dynamical system,” *J. Convex Anal.*, vol. 14, no. 4, pp. 869–878, 2007.
- [21] P.-E. Maingé and N. Merabet, “A new inertial-type hybrid projection-proximal algorithm for monotone inclusions,” *Applied Mathematics of Computation*, vol. 215, pp. 3149–3162, 2010.
- [22] P. Ochs, Y. Chen, T. Brox, and T. Pock, “ipiano: Inertial proximal algorithm for non-convex optimization,” *manuscript*, 2014.
- [23] P. Ochs, T. Brox, and T. Pock, “ipiasco: Inertial proximal algorithm for strongly convex optimization,” *manuscript*, 2014.
- [24] H. Attouch, J. Peypouquet, and P. Redont, “A dynamical approach to an inertial forward-backward algorithm for convex

- minimization,” *SIAM J. Optim.*, vol. 24, no. 1, pp. 232–256, 2014.
- [25] R. I. Bot, E. R. Csetnek, and C. Hendrich, “Inertial douglas-rachford splitting for monotone inclusion problems,” *arXiv preprint arXiv:1403.3330*, 2014.
- [26] R. I. Bot and E. R. Csetnek, “An inertial alternating direction method of multipliers,” *arXiv preprint arXiv:1404.4582*, 2014.
- [27] —, “An inertial tseng’s type proximal algorithm for nonsmooth and nonconvex optimization problems,” *arXiv preprint arXiv:1406.0724*, 2014.
- [28] P. Tseng, “A modified forward-backward splitting method for maximal monotone mappings,” *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 431–446, 2000.
- [29] R. I. Bot and E. R. Csetnek, “An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems,” *arXiv preprint arXiv:1402.5291*, 2014.
- [30] —, “A hybrid proximal-extragradient algorithm with inertial effects,” *arXiv preprint arXiv:1407.0214*, 2014.
- [31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [32] Y. E. Nesterov, “A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ ,” *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.
- [33] W. Deng, M.-J. Lai, and W. Yin, “On the  $o(1/k)$  convergence and parallelization of the alternating direction method of multipliers,” *arXiv preprint arXiv:1312.3040*, 2013.
- [34] D. Davis and W. Yin, “Convergence rate analysis of several splitting schemes,” *arXiv preprint arXiv:1406.4834*, 2014.
- [35] Z. Opial, “Weak convergence of the sequence of successive approximations for nonexpansive mappings,” *Bull. Amer. Math. Soc.*, vol. 73, pp. 591–597, 1967.
- [36] J. Eckstein, “Splitting methods for monotone operators with applications to parallel optimization,” Ph.D. dissertation, Massachusetts Institute of Technology, 1989.
- [37] X. Cai, G. Gu, B. He, and X. Yuan, “A proximal point algorithm revisit on the alternating direction method of multipliers,” *Sci. China Math.*, vol. 56, no. 10, pp. 2179–2186, 2013.
- [38] B. He and X. Yuan, “On the  $O(1/n)$  convergence rate of the Douglas-Rachford alternating direction method,” *SIAM J. Numer. Anal.*, vol. 50, no. 2, pp. 700–709, 2012.
- [39] —, “On non-ergodic convergence rate of douglas-rachford alternating direction method of multipliers,” Tech. rep., Nanjing University, Tech. Rep., 2012.
- [40] R. Shefi and M. Teboulle, “Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization,” *SIAM J. Optim.*, vol. 24, no. 1, pp. 269–297, 2014.
- [41] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vol. I*, ser. Springer Series in Operations Research. Springer-Verlag, New York, 2003.
- [42] M. Fazel, T. K. Pong, D. F. Sun, and P. Tseng, “Hankel matrix rank minimization with applications to system identification and realization,” *SIAM J. Matrix Analysis and Applications*, to appear, 2013.
- [43] C. Chen, “Numerical algorithms for a class of matrix norm approximation problems,” Ph.D. dissertation, Nanjing University, 2012.
- [44] J. Wright, A. Ganesh, K. Min, and Y. Ma, “Compressive principal component pursuit,” *Information and Inference*, vol. 2, no. 1, pp. 32–68, 2013.
- [45] S. Ma, D. Goldfarb, and L. Chen, “Fixed point and Bregman iterative methods for matrix rank minimization,” *Math. Program.*, vol. 128, no. 1-2, Ser. A, pp. 321–353, 2011.
- [46] J. Yang and Y. Zhang, “Alternating direction algorithms for  $\ell_1$ -problems in compressive sensing,” *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.
- [47] J. Yang and X. Yuan, “Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization,” *Mathematics of Computation*, vol. 82, no. 2, pp. 301–329, 2013.
- [48] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

**Appendix A. Proof of Theorem 4.** First, we sketch the proof of convergence of the sequence  $\{w^k\}$  to a solution of (2.1). Clearly, the matrix  $G$  defined in (3.7) is symmetric and positive definite under our assumptions that  $\tau < 1/\rho(A^T A)$  and  $\eta < 1/\rho(B^T B)$ . Let  $w^* \in \Omega^*$  be arbitrary fixed. It follows from setting

$w = w^*$  in (3.6) that

$$\begin{aligned} \langle w^{k+1} - w^*, G(w^k - w^{k+1}) \rangle &\geq \theta(w^{k+1}) - \theta(w^*) + \langle w^{k+1} - w^*, F(w^{k+1}) \rangle \\ &\geq \theta(w^{k+1}) - \theta(w^*) + \langle w^{k+1} - w^*, F(w^*) \rangle \\ &\geq 0, \end{aligned}$$

where the second “ $\geq$ ” follows from the monotonicity of  $F$ . Therefore, we obtain

$$\begin{aligned} \|w^{k+1} - w^*\|_G^2 &= \|w^k - w^*\|_G^2 - \|w^k - w^{k+1}\|_G^2 - 2\langle w^{k+1} - w^*, G(w^k - w^{k+1}) \rangle \\ &\leq \|w^k - w^*\|_G^2 - \|w^k - w^{k+1}\|_G^2. \end{aligned} \quad (\text{A.1})$$

Since  $G$  is positive definite, this implies that measured by  $G$ -norm the sequence  $\{w^k\}$  is strictly contractive with respect to  $\Omega^*$  unless  $w^k = w^{k+1}$  in which case  $w^k$  is already a solution. The convergence of  $\{w^k\}$  to some solution  $w^* \in \Omega^*$  follows directly from standard analyses for PPA and the key inequality (A.1). We omit the details.

Second, we prove (3.11). Let  $w^{i+1} \in \Omega$  be generated via (3.5). It follows from the monotonicity of  $F$  and (3.6) that, for any  $w \in \Omega$ , there holds

$$\begin{aligned} \theta(w) - \theta(w^{i+1}) + (w - w^{i+1})^T F(w) &\geq \theta(w) - \theta(w^{i+1}) + (w - w^{i+1})^T F(w^{i+1}) \\ &\geq (w - w^{i+1})^T G(w^i - w^{i+1}). \end{aligned}$$

By noting the relation  $2(w - w^{i+1})^T G(w^i - w^{i+1}) \geq \|w - w^{i+1}\|_G^2 - \|w - w^i\|_G^2$ , we obtain

$$\theta(w) - \theta(w^{i+1}) + (w - w^{i+1})^T F(w) \geq \frac{1}{2} (\|w - w^{i+1}\|_G^2 - \|w - w^i\|_G^2), \forall w \in \Omega.$$

Take sum over  $i = 0, 1, \dots, k$  and divide both sides by  $(k+1)$ , we get

$$\theta(w) - \frac{1}{k+1} \sum_{i=0}^k \theta(w^{i+1}) + \left( w - \frac{1}{k+1} \sum_{i=0}^k w^{i+1} \right)^T F(w) \geq -\frac{\|w - w^0\|_G^2}{2(k+1)}, \forall w \in \Omega. \quad (\text{A.2})$$

The conclusion (3.11) follows directly from (A.2) by noting the definition of  $\bar{w}^k$  and the fact that

$$\frac{1}{k+1} \sum_{i=0}^k \theta(w^{i+1}) \geq \theta \left( \frac{1}{k+1} \sum_{i=0}^k w^{i+1} \right) = \theta(\bar{w}^k).$$

The equivalence of (3.12) and (3.11) can be verified directly from the notation defined in (3.2) and the definition of  $\mathcal{L}$  in (3.3a).

Finally, we prove (3.13) and (3.14). Since (3.6) holds for all  $k$ , it also holds for  $k := k-1$ , i.e.,

$$\theta(w) - \theta(w^k) + \langle w - w^k, F(w^k) + G(w^k - w^{k-1}) \rangle \geq 0, \forall w \in \Omega. \quad (\text{A.3})$$

By setting  $w = w^k$  and  $w = w^{k+1}$  in (3.6) and (A.3), respectively, and taking an addition, we obtain

$$\langle G(w^{k+1} - w^k), (w^k - w^{k-1}) - (w^{k+1} - w^k) \rangle \geq \langle w^{k+1} - w^k, F(w^{k+1}) - F(w^k) \rangle \geq 0.$$

In addition, by taking into account the fact that

$$\|w^k - w^{k-1}\|_G^2 - \|w^{k+1} - w^k\|_G^2 \geq 2\langle G(w^{k+1} - w^k), (w^k - w^{k-1}) - (w^{k+1} - w^k) \rangle,$$

we obtain  $\|w^k - w^{k-1}\|_G \geq \|w^{k+1} - w^k\|_G$ , i.e.,  $\|w^k - w^{k-1}\|_G$  is monotonically nonincreasing with respect to  $k$ . By further considering (A.1), we obtain

$$k\|w^k - w^{k-1}\|_G^2 \leq \sum_{i=0}^{k-1} \|w^{i+1} - w^i\|_G^2 \leq \sum_{i=0}^{k-1} (\|w^i - w^*\|_G^2 - \|w^{i+1} - w^*\|_G^2) \leq \|w^0 - w^*\|_G^2,$$

which implies the relation (3.13). By using the trick introduced in [33, 34], we can derive the  $o(1/k)$  result (3.14). Specifically, we have

$$\frac{k}{2}\|w^k - w^{k-1}\|_G^2 \leq \sum_{i=\lfloor \frac{k}{2} \rfloor}^k \|w^i - w^{i-1}\|_G^2, \quad (\text{A.4})$$

where  $\lfloor k/2 \rfloor$  denotes the greatest integer no greater than  $k/2$ . The result (3.14) follows by further considering  $\sum_{k=0}^{\infty} \|w^{k+1} - w^k\|_G^2 < \infty$  and thus the right-hand-side of (A.4) converges to 0 as  $k \rightarrow \infty$ .