Generalized asymmetric forward-backward-adjoint algorithms for convex-concave saddle-point problem

Jianchao Bai* Yang Chen† Xue Yu‡ Hongchao Zhang§

Abstract

The convex-concave minimax problem, also known as the saddle-point problem, has been extensively studied from various aspects including the algorithm design, convergence condition and complexity. In this paper, we propose a generalized asymmetric forward-backward-adjoint algorithm (G-AFBA) to solve such a problem by utilizing both the proximal techniques and the extrapolation of primal-dual updates. Besides applying proximal primal-dual updates, G-AFBA enjoys a more relaxed convergence condition, namely, more flexible and possibly larger proximal stepsizes, which could result in significant improvements in numerical performance. We study the global convergence of G-AFBA as well as its sublinear convergence rate on both ergodic iterates and non-ergodic optimality error. The linear convergence rate of G-AFBA is also established under a calmness condition. By different ways of parameter and problem setting, we show that G-AFBA has close relationships with several well-established or new algorithms. We further propose an adaptive and a stochastic (inexact) versions of G-AFBA. Our numerical experiments on solving the robust principal component analysis problem and the 3D CT reconstruction problem indicate the efficiency of both the deterministic and stochastic versions of G-AFBA.

Keywords: Saddle-point problem, asymmetric forward-backward-adjoint algorithm, convergence and complexity, image processing

AMS subject classifications. 65K10, 65Y20, 90C25, 94A08

1 Introduction

Consider the following generic convex-concave saddle-point problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \mathcal{L}(x,y) := f(x) + \langle Kx, y\rangle - g(y), \tag{1.1}$$

where $f : \mathcal{X} \to (-\infty, \infty]$ and $g : \mathcal{Y} \to (-\infty, \infty]$ are proper lower semicontinuous convex functions (not necessarily smooth), $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional real Euclidean

*Corresponding author. School of Mathematics and Statistics & MOE Key Laboratory for Complexity Science in Aerospace, Northwestern Polytechnical University, Xi'an 710129, China (`jianchaobai@nwpu.edu.cn`).

†Corresponding author. School of Mathematics and Statistics, Northwestern Polytechnical University & MIIT Key Laboratory of Dynamics and Control of Complex Systems, Xi'an, 710129, China (`cy1202208@163.com`).

‡Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China (`xueyu_2019@ruc.edu.cn`).

§https://math.lsu.edu/∼hozhang, Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918 (`hozhang@math.lsu.edu`).

spaces, $K : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator. Let $K^\top$ denote the adjoint operator (or matrix transpose) of K, f^* and g^* denote the Fenchel conjugate [38] of f and g, respectively. Then, (1.1) amounts to the following primal and dual problems:

$$\min_{x \in \mathcal{X}} f(x) + g^*(Kx) \qquad \text{and} \qquad \min_{y \in \mathcal{Y}} f^*(-K^\top y) + g(y).$$

Due to these intrinsic relationships, the problem (1.1) covers a wide range of applications, including machine learning, signal and image processing, economics, statistics, see e.g. [9, 12, 21, 24, 28, 40, 48, 51] and the references therein. Throughout this paper, the solution set of (1.1) is assumed to be nonempty.

1.1 Notation

Let $\mathbb{R}^n$ be the set of n-dimensional Euclidean space equipped with an inner product $\langle \cdot, \cdot \rangle$ and Euclidean norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. Let $\mathbf{I}$ be the identity matrix and $\mathbf{0}$ be the zero matrix/vector. Given a positive definite self-adjoint linear operator or symmetric matrix H, we denote $\|x\|_H = \sqrt{\langle x, Hx \rangle} = \sqrt{x^\top H x}$ with the superscript $^\top$ representing transpose. Denote the Euclidean distance from $x \in \mathcal{C}$ to the closed convex set $\mathcal{C}$ by $\text{dist}(x, \mathcal{C}) = \min_{y \in \mathcal{C}} \|x - y\|$, and the G-weighted distance by $\text{dist}_G(x, \mathcal{C}) = \min_{y \in \mathcal{C}} \|x - y\|_G$ where G is a self-adjoint and positive definite linear operator. The notation $\rho(G)$ denotes the spectral radius of G, while $\lambda_{\min}(G)$ and $\lambda_{\max}(G)$ denote the minimum and maximum eigenvalues of G, respectively.

1.2 Related work

Due to the separable structure of f and g in (1.1), many effective algorithms are designed to treat them individually so as to make full use of the properties of each component objective function. An earlier yet simpler approach for solving (1.1) is the Arrow-Hurwicz method [1]:

$$\text{(PDHG)} \quad \begin{cases} x^{k+1} = \arg\min_{x \in \mathcal{X}} \mathcal{L}(x, y^k) + \frac{1}{2\tau} \left\| x - x^k \right\|^2, \\ y^{k+1} = \arg\max_{y \in \mathcal{Y}} \mathcal{L}(x^{k+1}, y) - \frac{1}{2\sigma} \left\| y - y^k \right\|^2, \end{cases} \tag{1.2}$$

where the positive parameters τ and σ are often regarded as the proximal primal and dual stepsizes. This Arrow-Hurwicz method was also called a primal-dual hybrid gradient method (PDHG) due to the earlier work [51], and it was described [50] as a proximal version of the traditional augmented Lagrangian method (ALM) for some canonical convex programming problems. O'Connor and Vandenberghe [36] showed that PDHG can be viewed as a special case of the Douglas-Rachford splitting algorithm [35] from the perspective of solving a monotone inclusion problem. Another related well-known algorithm based on (1.2) is proposed by Chambolle-Pock [9] (see e.g. [37]) by employing an extrapolation technique:

$$\begin{cases} x^{k+1} = \arg\min_{x \in \mathcal{X}} \mathcal{L}(x, y^k) + \frac{1}{2\tau} \left\| x - x^k \right\|^2, \\ y^{k+1} = \arg\max_{y \in \mathcal{Y}} \mathcal{L}(x^{k+1} + \alpha(x^{k+1} - x^k), y) - \frac{1}{2\sigma} \left\| y - y^k \right\|^2. \end{cases} \tag{1.3}$$

Here, $\alpha \in [0, 1]$ is an extrapolation stepsize. Clearly, (1.3) reduces to (1.2) when $\alpha = 0$. It was shown in [9] that (1.3) is closely related to the existing extragradient method [32] and a preconditioned version of the alternating direction method of multipliers (ADMM) [18]. The connection between (1.3) and the forward-backward splitting method [35] can be found in [42]. Although the scheme (1.3) applies a proximal technique, some counter-examples provided in [25] showed that when $\alpha = 0$, i.e. the PDHG method, it is not necessarily convergent. Moreover, the global convergence of (1.3) with $\alpha \in (0, 1)$ remains still unknown[1], although its global convergence with $\alpha = 0$ had been established [23] by assuming strong convexity on one of the objective functions. So far, the widely used scheme of (1.3) is the case with $\alpha = 1$:

$$\text{(CP-PPA)} \quad \begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}(x, y^k) + \frac{1}{2\tau}\|x - x^k\|^2, \\ y^{k+1} = \arg\max_{y\in\mathcal{Y}} \mathcal{L}(2x^{k+1} - x^k, y) - \frac{1}{2\sigma}\|y - y^k\|^2, \end{cases} \tag{1.4}$$

where the stepsize parameters τ and σ need to satisfy

$$\frac{1}{\tau\sigma} > L \quad \text{with} \quad L = \rho(K^\top K) \tag{1.5}$$

for ensuring global convergence of CP-PPA. Convergence of an adaptive version of (1.4) was investigated by Goldstein et al. [20]. More recently, He et al. [24] extended CP-PPA (1.4) to the following generalized version:

$$\text{(GCP-PPA)} \quad \begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}(x, y^k) + \frac{1}{2\tau}\|x - x^k\|^2, \\ \overline{y}^{k+1} = \arg\max_{y\in\mathcal{Y}} \mathcal{L}(x^{k+1} + \alpha(x^{k+1} - x^k), y) - \frac{1}{2\sigma}\|y - y^k\|^2, \\ y^{k+1} = \overline{y}^{k+1} - (1-\alpha)\sigma K(x^{k+1} - x^k), \end{cases} \tag{1.6}$$

where $\alpha \in [0, 1]$ is a parameter. GCP-PPA has global convergence when

$$\frac{1}{\tau\sigma} > (1 - \alpha + \alpha^2)L. \tag{1.7}$$

Obviously, when $\alpha = 1$ the above GCP-PPA reduces to CP-PPA, while for $\alpha \in [0, 1)$ an extrapolation step is used for the dual variable to ensure global convergence. Moreover, the stepsize requirement (1.7) is more relaxed than the condition (1.5). For example, when $\alpha = 0.5$, (1.7) only requires $\frac{1}{\tau\sigma} > 0.75L$. In addition, some stochastic and accelerated first-order methods have been also proposed for solving (1.1) when its objective function has certain structures or satisfies further smoothness conditions. For a much incomplete reference list, please see e.g. [11, 12, 26, 28, 33, 44, 47, 52].

As a generalization of (1.3), the Condat-Vũ scheme proposed independently in [14, 42] has attracted much attention in recent years and its convergence can be proved by casting the scheme into a forward-backward splitting method. However, the condition of involved parameters seems to be more restrictive than that of PDHG. Another interesting and closely related method is the asymmetric forward-backward-adjoint algorithm (AFBA) [33] for solving structured monotone inclusion problems, which was also studied and extended to solve the saddle-point problem (1.1) [46]. An inexact AFBA with absolute error criteria was further proposed in [30] to alleviate both theoretical and numerical difficulties of solving subproblems exactly. But, to our understanding, both the original AFBA and its inexact version have an even more conservative stepsize rule than that of the Condat-Vũ scheme. For a comprehensive survey on proximal splitting algorithms, we refer to [15] for more details.

1.3 The algorithm and contribution

Notice that the convergence condition of CP-PPA has been significantly improved by He et al. [24] through performing an extrapolation step on the y-variable along the

[1] Recently, its weak convergence was established in [2] when $\alpha > 1/2$ and $\tau\sigma L < 4/(1 + 2\alpha)$.

iterative difference of the x-variable. That is, the correction step of y-iterates uses the interactive information from x-iterates, which is different from the traditional way of performing correction steps along its own iterates. A natural and yet interesting question to investigate is whether the convergence condition (1.7) can be further improved by applying extrapolation steps on both the primal and dual updates. By this motivation, we propose the following generalized asymmetric forward-backward-adjoint algorithm:

$$\text{(G-AFBA)}\quad \begin{cases} \overline{x}^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top y^k\right\|^2, \\ \overline{y}^{k+1} = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma}\left\|y - y^k - \sigma K[\overline{x}^{k+1} + \alpha(\overline{x}^{k+1} - x^k)]\right\|^2, \\ x^{k+1} = \overline{x}^{k+1} - (1-\alpha)\mu\tau K^\top(\overline{y}^{k+1} - y^k), \\ y^{k+1} = \overline{y}^{k+1} + (1-\alpha)(1-\mu)\sigma K(\overline{x}^{k+1} - x^k), \end{cases} \tag{1.8}$$

where $\alpha, \mu \in [0,1]$, $\tau > 0$ and $\sigma > 0$ are algorithm parameters. To ensure the global convergence of G-AFBA, we require the primal-dual stepsize parameters (σ, τ) to satisfy

$$\frac{1}{\tau\sigma} > \frac{\alpha + (1-\mu+\mu^2)(1-\alpha)^2 + \sqrt{[\alpha - (1-\mu+\mu^2)(1-\alpha)^2]^2 + 4\alpha(1-\alpha)^2}}{2}L. \tag{1.9}$$

We now have the following comments on G-AFBA:

(I) **Flexibility of the algorithm.** Table 1 shows that G-AFBA is quite general and includes many well-established algorithms we have previously discussed as special cases. We refer to Sections 4-5 for more detailed discussions on the connections between G-AFBA and other related methods including the application of G-AFBA to multi-block convex programming, an adaptive version of G-AFBA, and a tailored stochastic G-AFBA for solving structured saddle-point problems from machine learning. The major difference between G-AFBA (1.8) and other existing PDHG-type methods is the two crossing extrapolation steps performed on the primal-dual variables, which can be also viewed as a correction step from our later analysis in a prediction-correction framework (see (3.2)). In fact, these two extrapolation steps can be also treated as backward and forward steps on the primal-dual variables.

Cases	Algorithms	Region of (τ, σ)
$\alpha = 1$	CP-PPA [9] & Reduced ALM	(1.5)
$(\alpha, \mu) = (0, 1)$	Exact version of Algorithm 2 [30]	(1.5)
$\alpha \in [0,1], \mu = 0$	GCP-PPA [24]	(1.7)
$\alpha, \mu \in [0,1]$	G-AFBA(ours)	(1.9)
$\alpha = 0, \mu \in [0,1]$	G1-AFBA(ours)	(4.4)

Table 1: Relationship between G-AFBA (1.8) and several methods.

(II) **Larger stepsize parameters.** Figure 1 visualizes the lower bound of $\frac{1}{\tau\sigma L}$ in (1.7) and (1.9) for ensuring global convergence, where Figure 1(a) is the same as Figure 1(b) but at different azimuth and elevation angles. As shown in Figure 1, the lower bound 0.75 of $\frac{1}{\tau\sigma L}$ with $\alpha = 0.5$ in (1.7) can be further improved by the lower bound given in (1.9). Hence, the current lower bound 0.75 on $\frac{1}{\tau\sigma L}$ for PDHG-type methods e.g. given in [24, 31, 34] is not tight, and possible

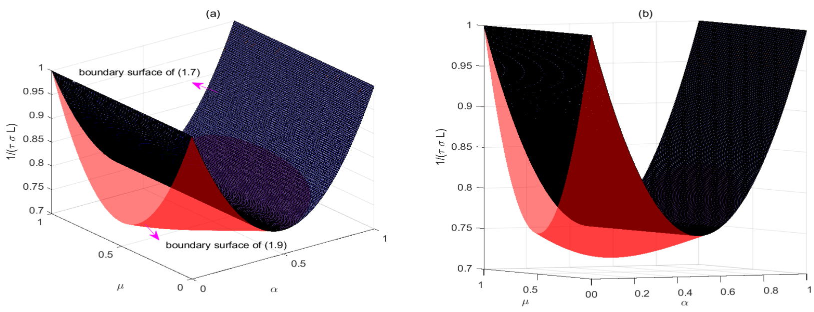

Figure 1: Visualization on the lower bound of $\frac{1}{\tau\sigma L}$ in (1.7) and (1.9).

larger stepsizes on σ and τ can be applied in G-AFBA without losing global convergence. For example, by setting $(\alpha, \mu) = (1/3, 1/2)$, the condition (1.9) reduces to $\frac{1}{\tau\sigma} > \frac{3+2\sqrt{3}}{9}L \approx 0.7182L$. Moreover, note that when $\mu = 0$, the condition (1.9) will reduce to (1.7) exactly matching the convergence condition of GCP-PPA.

(III) **Global convergence and various convergence rates.** Based on variational reformulations for both the saddle-point problem (1.1) and the iterative sequence of G-AFBA (1.8), we establish the global convergence of G-AFBA, its sublinear convergence rate in the sense of the primal-dual function value gap, the sublinear convergence rate of the optimality gap and the optimality error measured by the difference of two consecutive iterates. We also show the linear convergence of G-AFBA under proper regulation (calmness) conditions. We further propose an adaptive version of G-AFBA with similar convergence rate but often enjoying significantly better practical performance. In addition, we give a customized stochastic G-AFBA (SG-AFBA) for solving a structured (1.1) with large sample sizes from machine learning. In fact, by considering the sample size as one, SG-AFBA will reduce to an inexact deterministic G-AFBA which allows to solve one proximal mapping subproblem to an adaptive accuracy (see the discussion in Section 5). Our numerical experiments on solving two classes of image processing problems indicate that by allowing flexible choices of stepsizes σ and τ, G-AFBA and its variants can have better performance compared with some well-established methods.

1.4 Organization of the paper

In Section 2, we prepare some preliminaries that are used to analyze the convergence of G-AFBA. Section 3 is dedicated to analyzing the global convergence and sublinear/linear convergence rate of G-AFBA based on a prediction-correction framework. Section 4 shows the relationship of G-AFBA with some existing and new related methods. Section 5 provides an adaptive G-AFBA (aG-AFBA) and a customized stochastic G-AFBA (SG-AFBA). We finally present numerical comparisons of G-AFBA, aG-AFBA and SG-AFBA with some other well-known methods in Section 6.

2 Preliminaries

In this section, we first provide a variational formulation for the saddle-point problem (1.1). Then, we show some nice properties of certain block structured matrices which will play key roles in the theoretical analysis of G-AFBA.

2.1 Reformulation of the saddle-point

Let $\Omega := \mathcal{X} \times \mathcal{Y}$. We call a point $(x^*, y^*) \in \Omega$ the saddle-point of (1.1) if it satisfies

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

that is,

$$\begin{cases} f(x) - f(x^*) + \langle x - x^*, K^\top y^* \rangle \geq 0, & \forall x \in \mathcal{X}, \\ g(y) - g(y^*) + \langle y - y^*, -Kx^* \rangle \geq 0, & \forall y \in \mathcal{Y}. \end{cases} \tag{2.1}$$

These inequalities can be expressed as the following variational form

$$\mathrm{VI}(\theta, \mathcal{J}, \Omega): \ \theta(u) - \theta(u^*) + \langle u - u^*, \mathcal{J}(u^*) \rangle \geq 0, \quad \forall u \in \Omega, \tag{2.2}$$

where

$$u = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \theta(u) = f(x) + g(y), \quad \mathcal{J}(u) = \begin{pmatrix} K^\top y \\ -Kx \end{pmatrix}. \tag{2.3}$$

Notice that the above operator $\mathcal{J}(u)$ satisfies

$$\langle u - v, \mathcal{J}(u) - \mathcal{J}(v) \rangle \equiv 0, \quad \forall u, v \in \Omega.$$

In the convex optimization, u^* satisfies (2.2) if and only if u^* is a primal-dual solution of the problem (1.1). Because of the nonempty assumption on the solution set of (1.1), the solution set of $\mathrm{VI}(\theta, \mathcal{J}, \Omega)$, denoted by Ω^*, is also nonempty and can be characterized as (see [22])

$$\Omega^* = \bigcap_{u \in \Omega} \left\{ \overline{u} \mid \theta(u) - \theta(\overline{u}) + \langle u - \overline{u}, \mathcal{J}(\overline{u}) \rangle \geq 0 \right\}. \tag{2.4}$$

2.2 Some matrices and properties

In order to simplify and conveniently analyze the convergence of G-AFBA, we introduce the following matrices

$$Q = \begin{bmatrix} \frac{1}{\tau}\mathbf{I} & -K^\top \\ -\alpha K & \frac{1}{\sigma}\mathbf{I} \end{bmatrix}, \quad M = \begin{bmatrix} \mathbf{I} & -(1-\alpha)\mu\tau K^\top \\ (1-\alpha)(1-\mu)\sigma K & \mathbf{I} \end{bmatrix}. \tag{2.5}$$

Note that the matrix M is nonsingular for any $\mu \in [0, 1]$ and $\tau, \sigma > 0$. With these matrices, we define

$$H = QM^{-1} \quad \text{and} \quad G = Q^\top + Q - M^\top HM. \tag{2.6}$$

For the matrices H and G, the following properties hold.

Proposition 2.1 *For any parameters (τ, σ) satisfying (1.9), the matrices H and G defined in (2.6) are symmetric positive definite.*

Proof. First, notice that

$$\begin{aligned} &\tfrac{1}{(\tau\sigma)^2} + \left[(-1 + \mu - \mu^2)(1-\alpha)^2 - \alpha\right]\tfrac{L}{\tau\sigma} - (1-\alpha)^2(1-\mu)\mu\alpha L^2 > 0 \\ \Longleftrightarrow &\left[\tfrac{1}{\tau\sigma} + (1-\alpha)^2(1-\mu)\mu L\right]\left[\tfrac{1}{\tau\sigma} - \alpha L\right] > (1-\alpha)^2 \tfrac{L}{\tau\sigma}. \end{aligned}$$

Hence, for all (τ, σ) satisfying (1.9), we have $1/(\tau\sigma) > \alpha L$, which implies Q defined in (2.5) is nonsingular. Now, let us define $D = Q^\top M$. Then, D is nonsingular since M is nonsingular. In addition, the H and G defined in (2.6) can be written as

$$H = QD^{-1}Q^\top \quad \text{and} \quad G = Q^\top + Q - D. \tag{2.7}$$

By direct calculation, we can derive from (2.5) and (2.7) that

$$D = \begin{bmatrix} \frac{1}{\tau}\mathbf{I} - \alpha(1-\alpha)(1-\mu)\sigma K^\top K & -\big[\alpha + (1-\alpha)\mu\big] K^\top \\ -\big[\alpha + (1-\alpha)\mu\big] K & \frac{1}{\sigma}\mathbf{I} + (1-\alpha)\mu\tau K K^\top \end{bmatrix} \tag{2.8}$$

and

$$G = \begin{bmatrix} \frac{1}{\tau}\mathbf{I} + \alpha(1-\alpha)(1-\mu)\sigma K^\top K & \big[(1-\alpha)\mu - 1\big] K^\top \\ \big[(1-\alpha)\mu - 1\big] K & \frac{1}{\sigma}\mathbf{I} - (1-\alpha)\mu\tau K K^\top \end{bmatrix}. \tag{2.9}$$

Due to the symmetric property of D and the relationship $H = QD^{-1}Q^\top$, we also have H is symmetric. Hence, to show the positive definiteness of H, we only need to show D is positive definite. Without loss of generality, suppose K is an $m \times n (m \le n)$ dimensional operator matrix and let $K = V\Sigma U^\top$ be the singular value decomposition of K, where both $V \in \mathbb{R}^{m\times m}$ and $U \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma = (\Sigma_m, \mathbf{0})$ is a diagonal matrix with $\Sigma_m = \mathrm{diag}(s_1, s_2, \cdots, s_m) \in \mathbb{R}^{m\times m}$ and $s_i \ge 0 (i = 1, 2, \dots, m)$ being the singular values of K. Then, we have

$$K^\top K = U \begin{bmatrix} \Sigma_m^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^\top \quad \textit{and} \quad KK^\top = V\Sigma_m^2 V^\top.$$

Then, it follows from (2.8) that

$$D = \begin{bmatrix} U & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix} \underbrace{\begin{bmatrix} \frac{1}{\tau}\mathbf{I} - \alpha(1-\alpha)(1-\mu)\sigma\Sigma_m^2 & \mathbf{0} & -\big[\alpha + (1-\alpha)\mu\big]\Sigma_m \\ \mathbf{0} & \frac{1}{\tau}\mathbf{I} & \mathbf{0} \\ -\big[\alpha + (1-\alpha)\mu\big]\Sigma_m & \mathbf{0} & \frac{1}{\sigma}\mathbf{I} + (1-\alpha)\mu\tau\Sigma_m^2 \end{bmatrix}}_{P} \begin{bmatrix} U & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix}^\top.$$

By linear algebra calculations (e.g. see similar techniques in [43, Page 16]), we can show that the matrix P is positive definite if and only if

$$\Big(\frac{1}{\tau} - \alpha(1-\alpha)(1-\mu)\sigma s_i^2\Big)\Big(\frac{1}{\sigma} + (1-\alpha)\mu\tau s_i^2\Big) - \Big[\alpha + (1-\alpha)\mu\Big]^2 s_i^2 > 0$$

for all $i = 1, \dots, m$, which is equivalent to

$$\begin{aligned} &\frac{1}{(\tau\sigma)^2} + \big[(1-\mu)\mu(1-\alpha)^2 - \alpha\big]\frac{s_i^2}{\tau\sigma} - (1-\alpha)^2(1-\mu)\mu\alpha s_i^4 > 0 \\ \Longleftrightarrow\ & \Big[\frac{1}{\tau\sigma} + (1-\alpha)^2(1-\mu)\mu s_i^2\Big]\Big[\frac{1}{\tau\sigma} - \alpha s_i^2\Big] > 0. \end{aligned} \tag{2.10}$$

Since $L = \rho(K^\top K) = \rho(KK^\top) = \max_{i\in\{1,\dots,m\}} s_i^2 > 0$, $\alpha, \mu \in [0, 1]$ and $\sigma, \tau > 0$, we have from (2.10) that the matrix P is positive definite if $1/(\tau\sigma) > \alpha L$, which is ensured by the previous condition (1.9). So, from the above analysis, we have H is positive definite if (τ, σ) satisfies (1.9).

By the similar analysis and the representation of G in (2.9), we can show G is also positive definite if the condition (1.9) holds. The proof is completed. □

3 Convergence analysis

In this section, we first analyze the global convergence of G-AFBA and its sublinear convergence rate in the ergodic sense. We then study the sublinear convergence rate of G-AFBA in terms of both the difference of two consecutive iterations and the first-order optimality gap. We finally discuss the linear convergence of G-AFBA under a certain calmness conditions.

Now, observe that G-AFBA (1.8) can be equivalently written as the following prediction-correction framework, where M is given by (2.5), u^k and $\widetilde{u}^k$ are defined as

$$u^k = \begin{pmatrix} x^k \\ y^k \end{pmatrix} \quad \text{and} \quad \widetilde{u}^k = \begin{pmatrix} \widetilde{x}^k \\ \widetilde{y}^k \end{pmatrix},$$

and the proximal operator of a function h with parameter $\tau > 0$ is defined as

$$\operatorname{prox}_{\tau h}(y) := \arg\min_{x\in\mathcal{X}} \Big\{ h(x) + \frac{1}{2\tau}\|x-y\|^2 \Big\}.$$

Algorithm 3.1: A prediction-correction reformulation of G-AFBA.

Prediction Step:

$$\widetilde{x}^k = \operatorname{prox}_{\tau f}\big(x^k - \tau K^\top y^k\big); \tag{3.1a}$$

$$\widetilde{y}^k = \operatorname{prox}_{\sigma g}\big(y^k + \sigma K[\widetilde{x}^k + \alpha(\widetilde{x}^k - x^k)]\big); \tag{3.1b}$$

Correction Step:

$$u^{k+1} = u^k - M(u^k - \widetilde{u}^k). \tag{3.2}$$

3.1 Global convergence

The global convergence of G-AFBA will be analyzed based on the above prediction-correction reformulation.

Lemma 3.1 *Let $\{\widetilde{u}^k = (\widetilde{x}^k;\widetilde{y}^k)\}$ be the predictor sequence generated by (3.1a)-(3.1b) and $\{u^{k+1} = (x^{k+1};y^{k+1})\}$ be the corrector sequence generated by (3.2). Then, for any $u \in \Omega$, the following inequality*

$$\mathcal{L}(x,\widetilde{y}^k) - \mathcal{L}(\widetilde{x}^k, y) \geq (u-\widetilde{u}^k)^\top Q(u^k - \widetilde{u}^k) \tag{3.3}$$

holds[2]*, where Q is given by (2.5).*

Proof. We can derive from the first-order optimality condition of (3.1a) that

$$f(x) - f(\widetilde{x}^k) + \big\langle x - \widetilde{x}^k, K^\top y^k + \frac{1}{\tau}(\widetilde{x}^k - x^k)\big\rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Rearranging the above inequality to obtain

$$f(x) - f(\widetilde{x}^k) + \big\langle x - \widetilde{x}^k, K^\top \widetilde{y}^k\big\rangle \geq \Big\langle x - \widetilde{x}^k, \frac{1}{\tau}(x^k - \widetilde{x}^k) - K^\top(y^k - \widetilde{y}^k)\Big\rangle \tag{3.4}$$

for any $x \in \mathcal{X}$. Similarly, we have from (3.1b) that

$$g(y) - g(\widetilde{y}^k) + \big\langle y - \widetilde{y}^k, -K[\widetilde{x}^k + \alpha(\widetilde{x}^k - x^k)] + \frac{1}{\sigma}(\widetilde{y}^k - y^k)\big\rangle \geq 0, \quad \forall y \in \mathcal{Y},$$

which can be equivalently rewritten as

$$g(y) - g(\widetilde{y}^k) + \big\langle y - \widetilde{y}^k, -K\widetilde{x}^k\big\rangle \geq \Big\langle y - \widetilde{y}^k, -\alpha K(x^k - \widetilde{x}^k) + \frac{1}{\sigma}(y^k - \widetilde{y}^k)\Big\rangle \tag{3.5}$$

for any $y \in \mathcal{Y}$. Combining (3.4) and (3.5) completes the proof of (3.3). ☐

The following lemma shows that the sequence $\{\|u^* - u^k\|_H\}$ is strictly decreasing under the weighted norm $\|u\|_H = \sqrt{u^\top H u}$.

[2]Note that (3.3) is equivalent to $\theta(u) - \theta(\widetilde{u}^k) + \langle u - \widetilde{u}^k, \mathcal{J}(\widetilde{u}^k)\rangle \geq (u-\widetilde{u}^k)^\top Q(u^k - \widetilde{u}^k)$.

Lemma 3.2 *Under the condition (1.9), the sequences* $\{\widetilde{u}^k\}$ *and* $\{u^{k+1}\}$ *generated by G-AFBA satisfy*

$$\mathcal{L}(x,\widetilde{y}^k) - \mathcal{L}(\widetilde{x}^k,y) \geq \frac{1}{2}\big(\|u-u^{k+1}\|_H^2 - \|u-u^k\|_H^2\big) + \frac{1}{2}\|u^k-\widetilde{u}^k\|_G^2 \tag{3.6}$$

for any $u\in\Omega$*, where* H *and* G *are defined in (2.6). Moreover, we have*

$$\|u^*-u^k\|_H^2 \geq \|u^*-u^{k+1}\|_H^2 + \|u^k-\widetilde{u}^k\|_G^2, \quad \forall u^*\in\Omega^*. \tag{3.7}$$

Proof. According to (3.2) and the definition of H in (2.6), we have

$$(u-\widetilde{u}^k)^\top Q(u^k-\widetilde{u}^k) = (u-\widetilde{u}^k)^\top H(u^k-u^{k+1}). \tag{3.8}$$

Then, applying the identity

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

with $a=u$, $b=\widetilde{u}^k$, $c=u^k$ and $d=u^{k+1}$ to the right-hand side of (3.8) gives

$$\begin{aligned}
&(u-\widetilde{u}^k)^\top H(u^k-u^{k+1}) - \tfrac{1}{2}\Big\{\|u-u^{k+1}\|_H^2 - \|u-u^k\|_H^2\Big\}\\
&= \tfrac{1}{2}\Big\{\|u^k-\widetilde{u}^k\|_H^2 - \|u^{k+1}-\widetilde{u}^k\|_H^2\Big\}\\
&= \tfrac{1}{2}\Big\{\|u^k-\widetilde{u}^k\|_H^2 - \|u^{k+1}-u^k+(u^k-\widetilde{u}^k)\|_H^2\Big\}\\
&\overset{(3.2)}{=} \tfrac{1}{2}\Big\{\|u^k-\widetilde{u}^k\|_H^2 - \|(u^k-\widetilde{u}^k)-M(u^k-\widetilde{u}^k)\|_H^2\Big\}\\
&= \tfrac{1}{2}\Big\{(u^k-\widetilde{u}^k)^\top(Q^\top+Q-M^\top HM)(u^k-\widetilde{u}^k)\Big\} \overset{(2.6)}{=} \tfrac{1}{2}\|u^k-\widetilde{u}^k\|_G^2,
\end{aligned} \tag{3.9}$$

where the fourth equality exploits the relation $Q=HM$ and its symmetric property. Then, substituting (3.8) and (3.9) into (3.3) confirms the assertion (3.6).

Set $u=u^*$ in (3.6) and use (2.1) with $(x,y)=(\widetilde{x}^k,\widetilde{y}^k)$ to obtain

$$\|u^*-u^k\|_H^2 - \|u^*-u^{k+1}\|_H^2 - \|u^k-\widetilde{u}^k\|_G^2 \geq 2\big[\mathcal{L}(\widetilde{x}^k,y^*) - \mathcal{L}(x^*,\widetilde{y}^k)\big] \geq 0.$$

Then, (3.7) follows directly. The proof is complete. □

In what follows, based on Lemma 3.2, we are ready to prove the global convergence of G-AFBA.

Theorem 3.1 *Under the condition (1.9), the sequence* $\{u^{k+1}\}$ *generated by G-AFBA converges to the solution point of (1.1).*

Proof. First, it follows from (3.7) in Lemma 3.2 and the positive definiteness of G and H that the sequence $\{u^k\}$ is bounded and

$$\lim_{k\to\infty}\|u^k-\widetilde{u}^k\| = 0. \tag{3.10}$$

As a result, the sequence $\{\widetilde{u}^k\}$ is also bounded and has at least one limit point u^∞. Let $\{\widetilde{u}^{k_j}\}$ be a subsequence converging to u^∞. Then, it follows from (3.3) that

$$\theta(u) - \theta(\widetilde{u}^{k_j}) + \big\langle u-\widetilde{u}^{k_j}, \mathcal{J}(\widetilde{u}^{k_j})\big\rangle \geq (u-\widetilde{u}^{k_j})^\top Q(u^{k_j}-\widetilde{u}^{k_j}), \quad \forall u\in\Omega,$$

which, together with (3.10), the lower semicontinuity of $\theta(u)$ and the continuity of $\mathcal{J}(u)$, implies

$$\theta(u) - \theta(u^\infty) + \big\langle u-u^\infty, \mathcal{J}(u^\infty)\big\rangle \geq 0, \quad \forall u\in\Omega.$$

That is to say, u^∞ is a solution point of (2.2) and hence is a solution point of (1.1).

Now, by (3.10) and $\lim_{j\to\infty} u^{k_j} = u^\infty$, the sequence $\{u^{k_j}\}$ also converges to u^∞. For any $k\geq k_j$, we can deduce from (3.7) that $\|u^\infty-u^{k_j}\|_H \geq \|u^\infty-u^k\|_H$. So, the whole sequence $\{u^k\}$ converges to u^∞. The proof is complete. □

3.2 Sublinear rate of convergence

In this section, we analyze the worst-case $\mathcal{O}(1/T)$ convergence rate of G-AFBA in the ergodic sense in terms of the optimality error measured by both the difference of two consecutive iterates and the first-order optimality gap, respectively, where T denotes the iteration number. First, it is obvious that (2.1) can be also expressed as

$$\mathcal{L}(x, y^*) - \mathcal{L}(x^*, y) \geq 0, \quad \forall (x, y) \in \Omega.$$

Hence, by (2.4), $\overline{u} = (\overline{x}; \overline{y})$ is often called an ϵ-approximate solution point of VI$(\theta, \mathcal{J}, \Omega)$ (2.2) with the accuracy $\epsilon > 0$ if it satisfies

$$\mathcal{L}(\overline{x}, y) - \mathcal{L}(x, \overline{y}) \leq \epsilon, \quad \forall u \in \mathcal{B}_{\overline{u}} = \{u \in \Omega \mid \|u - \overline{u}\| \leq 1\}.$$

In the following, we will demonstrate that, after T iterations, G-AFBA is able to find a point $\overline{u}$ such that

$$\sup_{u \in \mathcal{B}_{\overline{u}}} \big\{\mathcal{L}(\overline{x}, y) - \mathcal{L}(x, \overline{y})\big\} \leq \mathcal{O}(1/T). \tag{3.11}$$

Theorem 3.2 *Let $\{\widetilde{u}^k\}$ be the predictor sequence generated by (3.1a)-(3.1b) and $\{u^k\}$ be the corrector sequence generated by (3.2). For any integers $T > 0$ and $\kappa \geq 0$, let*

$$x_T = \frac{1}{T}\sum_{k=\kappa}^{T+\kappa} \widetilde{x}^k \quad \textit{and} \quad y_T = \frac{1}{T}\sum_{k=\kappa}^{T+\kappa} \widetilde{y}^k. \tag{3.12}$$

Then, under the condition (1.9) we have

$$\mathcal{L}(x_T, y) - \mathcal{L}(x, y_T) \leq \frac{1}{2(T+1)}\big\|u - u^\kappa\big\|_H^2, \quad \forall u \in \Omega, \tag{3.13}$$

where H is defined in (2.6).

Proof. The inequality (3.6) together with the positive definiteness of G implies

$$\mathcal{L}(\widetilde{x}^k, y) - \mathcal{L}(x, \widetilde{y}^k) \leq \frac{1}{2}\Big\{\big\|u - u^k\big\|_H^2 - \big\|u - u^{k+1}\big\|_H^2\Big\} \tag{3.14}$$

for any $u \in \Omega$. Sum the last inequality over $k = \kappa, \kappa+1, \cdots, T+\kappa$ to obtain

$$\sum_{k=\kappa}^{T+\kappa} \big[\mathcal{L}(\widetilde{x}^k, y) - \mathcal{L}(x, \widetilde{y}^k)\big] \leq \frac{1}{2}\big\|u - u^\kappa\big\|_H^2,$$

which, by the convexity of f, g, the definitions of x_T and y_T in (3.12), gives

$$(T+1)\big[\mathcal{L}(x_T, y) - \mathcal{L}(x, y_T)\big] \leq \frac{1}{2}\big\|u - u^\kappa\big\|_H^2.$$

Hence, (3.13) holds. The proof is complete. $\square$

Theorem 3.2 implies that under a more flexible condition (1.9), we have (3.11) holds, i.e., the primal-dual function value gap converges to zero with the worst-case $\mathcal{O}(1/T)$ ergodic rate. A similar result to (3.13) in the sense of expectation can be found in [4]. We next show that $\{\|u^k - u^{k+1}\|_H^2\}$, which also measures the optimality error, monotonically goes to zero with the worst-case $\mathcal{O}(1/T)$ convergence rate. The following lemma confirms that the sequence $\{\|u^k - u^{k+1}\|_H^2\}$ decreases monotonically.

Lemma 3.3 *Under the condition (1.9), the sequence $\{u^k\}$ generated by (3.2) satisfies*

$$\big\|u^k - u^{k+1}\big\|_H^2 \geq \big\|u^{k+1} - u^{k+2}\big\|_H^2. \tag{3.15}$$

Proof. It follows from (3.3) with $u = \widetilde{u}^{k+1}$ that

$$\mathcal{L}(\widetilde{x}^{k+1}, \widetilde{y}^k) - \mathcal{L}(\widetilde{x}^k, \widetilde{y}^{k+1}) \geq (\widetilde{u}^{k+1} - \widetilde{u}^k)^\top Q(u^k - \widetilde{u}^k). \tag{3.16}$$

Similarly, (3.3) holds at the $(k+1)$-th iteration, that is,

$$\mathcal{L}(x, \widetilde{y}^{k+1}) - \mathcal{L}(\widetilde{x}^{k+1}, y) \geq (u - \widetilde{u}^{k+1})^\top Q(u^{k+1} - \widetilde{u}^{k+1}), \quad \forall u \in \Omega,$$

which, by setting $u = \widetilde{u}^k$, results in

$$\mathcal{L}(\widetilde{x}^k, \widetilde{y}^{k+1}) - \mathcal{L}(\widetilde{x}^{k+1}, \widetilde{y}^k) \geq (\widetilde{u}^k - \widetilde{u}^{k+1})^\top Q(u^{k+1} - \widetilde{u}^{k+1}). \tag{3.17}$$

Combining (3.16) and (3.17), we have

$$(\widetilde{u}^k - \widetilde{u}^{k+1})^\top Q\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\} \geq 0. \tag{3.18}$$

Then, adding the equality

$$\begin{aligned} &\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\}^\top Q\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\} \\ =&\frac{1}{2}\big\|u^k - \widetilde{u}^k - (u^{k+1} - \widetilde{u}^{k+1})\big\|^2_{(Q^\top + Q)} \end{aligned} \tag{3.19}$$

to both sides of (3.18) leads to

$$\begin{aligned} &\tfrac{1}{2}\big\|u^k - \widetilde{u}^k - (u^{k+1} - \widetilde{u}^{k+1})\big\|^2_{(Q^\top + Q)} \\ &\leq (u^k - u^{k+1})^\top Q\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\} \\ &\overset{(3.2)}{=} (u^k - \widetilde{u}^k)^\top M^\top Q\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\} \\ &\overset{(2.5)}{=} (u^k - \widetilde{u}^k)^\top M^\top HM\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\}. \end{aligned}$$

Using this relationship, the identity $\|a\|_H^2 - \|b\|_H^2 = 2a^\top H(a-b) - \|a-b\|_H^2$ with $a = M(u^k - \widetilde{u}^k)$ and $b = M(u^{k+1} - \widetilde{u}^{k+1})$ and $u^k - u^{k+1} = M(u^k - \widetilde{u}^k)$, we have

$$\begin{aligned} &\big\|u^k - u^{k+1}\big\|_H^2 - \big\|u^{k+1} - u^{k+2}\big\|_H^2 \\ &= \big\|M(u^k - \widetilde{u}^k)\big\|_H^2 - \big\|M(u^{k+1} - \widetilde{u}^{k+1})\big\|_H^2 \\ &= 2(u^k - \widetilde{u}^k)^\top M^\top HM\big\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\big\} - \big\|M\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\}\big\|_H^2 \\ &\geq \big\|u^k - \widetilde{u}^k - (u^{k+1} - \widetilde{u}^{k+1})\big\|^2_{(Q^\top + Q)} - \big\|M\{(u^k - \widetilde{u}^k) - (u^{k+1} - \widetilde{u}^{k+1})\}\big\|_H^2 \\ &\overset{(2.6)}{=} \big\|u^k - \widetilde{u}^k - (u^{k+1} - \widetilde{u}^{k+1})\big\|_G^2 \geq 0, \end{aligned}$$

where the last inequality follows from the positive definiteness of G. We complete the proof. □

Theorem 3.3 *Suppose the condition (1.9) holds. Then, for any integers $T > 0$ and $\kappa \geq 0$, there exists a constant $c_0 > 0$ such that the sequence $\{u^{k+1}\}$ generated by G-AFBA satisfies*

$$\big\|u^{T+\kappa} - u^{T+\kappa+1}\big\|_H^2 \leq \frac{1}{(T+1)c_0}\big\|u^\kappa - u^*\big\|_H^2, \quad \forall u^* \in \Omega^*. \tag{3.20}$$

Proof. First, by the positive definiteness of G and $M^\top HM$, there exists a constant c_0 such that $G - c_0 M^\top HM$ is positive definite. Hence, we have

$$\big\|u^k - \widetilde{u}^k\big\|_G^2 \geq c_0\big\|M(u^k - \widetilde{u}^k)\big\|_H^2 = c_0\big\|u^k - u^{k+1}\big\|_H^2.$$

Then, it follows from inequality (3.7) that

$$\|u^{k+1} - u^*\|_H^2 \le \|u^k - u^*\|_H^2 - c_0\|u^k - u^{k+1}\|_H^2, \quad \forall u^* \in \Omega^*. \tag{3.21}$$

Summing (3.21) over $k = \kappa, \kappa+1, \cdots, T+\kappa$, it follows from the monotonicity of $\{\|u^k - u^{k+1}\|_H^2\}$ given in (3.15) that

$$\|u^\kappa - u^*\|_H^2 \ge \sum_{k=\kappa}^{T+\kappa} c_0\|u^k - u^{k+1}\|_H^2 \ge (1+T)c_0\|u^{T+\kappa} - u^{T+\kappa+1}\|_H^2$$

for any $u^* \in \Omega^*$, which leads to (3.20) immediately. □

For any given $\epsilon > 0$, Theorem 3.3 shows that the proposed G-AFBA (1.8) needs at most $\lceil c/\epsilon \rceil$ iterations to ensure $\|u^k - u^{k+1}\|_H^2 \le \epsilon$, where $c = \inf_{u^* \in \Omega^*} \|u^0 - u^*\|_H^2 / c_0$. Recall that u^{k+1} is a solution point of $\mathrm{VI}(\theta, \mathcal{J}, \Omega)$ if and only if $\|u^k - u^{k+1}\| = 0$. Hence, Theorem 3.3 indicates that $\|u^k - u^{k+1}\|_H$, which can be used as a measure of optimality error, converges to zero sublinearly. Moreover, let $d^k := (d_x^k, d_y^k)$, where

$$d_x^k = \frac{1}{\tau}(x^k - \widetilde{x}^k) - K^\top(y^k - \widetilde{y}^k) \quad \text{and} \quad d_y^k = \frac{1}{\sigma}(y^k - \widetilde{y}^k) - \alpha K(x^k - \widetilde{x}^k).$$

Since the optimality conditions in (3.4) and (3.5) are equivalent to

$$\begin{cases} f(x) - f(\widetilde{x}^k) + \langle x - \widetilde{x}^k, K^\top \widetilde{y}^k - d_x^k \rangle \ge 0, & \forall x \in \mathcal{X}, \\ g(y) - g(\widetilde{y}^k) + \langle y - \widetilde{y}^k, -K\widetilde{x}^k - d_y^k \rangle \ge 0, & \forall y \in \mathcal{Y}, \end{cases}$$

we have from finite-dimensional Euclidean spaces of $\mathcal{X}$ and $\mathcal{Y}$ that

$$d_x^k - K^\top \widetilde{y}^k \in \partial f(\widetilde{x}^k) \quad \text{and} \quad d_y^k + K\widetilde{x}^k \in \partial g(\widetilde{y}^k).$$

Hence, $\|d^k\|$ also measures the first-order optimality error. Notice that $d^k = Q(u^k - \widetilde{u}^k) = H(u^k - u^{k+1})$. So,

$$\|d^k\| = \|H(u^k - u^{k+1})\| \le \sqrt{\lambda_{\max}(H)}\|u^k - u^{k+1}\|_H,$$

which, by Theorem 3.3, implies $\|d^k\|$ also goes to zero in a sublinear rate.

3.3 Linear rate of convergence

For any $u = (x; y) \in \Omega$, we define the KKT mapping as

$$R(u) := \begin{pmatrix} x - \mathrm{prox}_f(x - K^\top y) \\ y - \mathrm{prox}_g(y + Kx) \end{pmatrix} \tag{3.22}$$

which is Lipschitz continuous on Ω because the proximal operator of a proper convex function is Lipschitz continuous with unit Lipschitz constant. Furthermore, given any $u \in \Omega$, we have $u \in \Omega^*$ if and only if $R(u) = 0$. Hence, $\Omega^* = \{u \in \Omega \mid R(u) = 0\}$.

In this subsection, under a calmness condition (see (3.23)), we establish the Q-linear convergence of $\{\mathrm{dist}_H(u^k, \Omega^*)\}$ to zero, where $\mathrm{dist}_H(u^k, \Omega^*) = \min_{u \in \Omega^*} \|u - u^k\|_H$, and the R-linear convergence of $\{u^k\}$ to a $u^\infty \in \Omega^*$. Similar conditions had been used for the linear convergence of ADMM and the inexact primal-dual algorithm, cf. [3, 29] to list a few.

Theorem 3.4 *Let $\{\widetilde{u}^k\}$ be the predictor sequence generated by (3.1a)-(3.1b) and $\{u^k\}$ be the corrector sequence generated by (3.2). Suppose the condition (1.9) holds. Then, we have the following properties:*

(i) There exists a saddle-point $u^\infty = (x^\infty; y^\infty) \in \Omega^*$ such that

$$\lim_{k\to\infty} \widetilde{u}^k = \lim_{k\to\infty} u^{k+1} = u^\infty.$$

(ii) If R^{-1} is calm at the origin for u^∞ with modulus $\theta > 0$, that is,

$$dist(u, \Omega^*) \le \theta \|R(u)\|, \quad \forall u \in \big\{u \in \Omega \big| \|u - u^\infty\| \le r\big\}, \tag{3.23}$$

for some $r > 0$, then there exist a $\xi \in (0, 1)$ such that

$$\mathrm{dist}_H(u^{k+1}, \Omega^*) \le \xi \mathrm{dist}_H(u^k, \Omega^*) \tag{3.24}$$

for all $k \ge 0$. Moreover, the sequence $\{\|u^k - u^\infty\|\}$ converges to zero R-linearly.

Proof. First, property (i) directly follows from Theorem 3.1. So, there exists an integer $\overline{k} > 0$ such that

$$\|u^k - u^\infty\| \le r, \quad \forall k \ge \overline{k}. \tag{3.25}$$

From the optimality conditions of (3.1a)-(3.1b), we can derive

$$\begin{cases} \widetilde{x}^k = \mathrm{prox}_f \Big[\widetilde{x}^k - \Big(\frac{1}{\tau}\big(\widetilde{x}^k - x^k\big) + K^\top y^k\Big)\Big], \\ \widetilde{y}^k = \mathrm{prox}_g \Big[\widetilde{y}^k - \Big(\frac{1}{\sigma}(\widetilde{y}^k - y^k) - K(\widetilde{x}^k + \alpha(\widetilde{x}^k - x^k))\Big)\Big]. \end{cases} \tag{3.26}$$

Combine (3.26) and the definition of $R(\cdot)$ in (3.22) to obtain

$$\begin{aligned} &\|R(\widetilde{u}^k)\|^2 = \big\|\widetilde{x}^k - \mathrm{prox}_f(\widetilde{x}^k - K^\top \widetilde{y}^k)\big\|^2 + \big\|\widetilde{y}^k - \mathrm{prox}_g(\widetilde{y}^k + K\widetilde{x}^k)\big\|^2 \\ &\le \big\| - \tfrac{1}{\tau}(\widetilde{x}^k - x^k) + K^\top(\widetilde{y}^k - y^k)\big\|^2 + \big\|\alpha K(\widetilde{x}^k - x^k) - \tfrac{1}{\sigma}(\widetilde{y}^k - y^k)\big\|^2 \\ &\le 2\big(\alpha^2 L + \tfrac{1}{\tau^2}\big)\|x^k - \widetilde{x}^k\|^2 + 2\big(L + \tfrac{1}{\sigma^2}\big)\|y^k - \widetilde{y}^k\|^2 \\ &\le \kappa_1 \|u^k - \widetilde{u}^k\|^2, \end{aligned}$$

where first inequality uses the non-expansive property of $\mathrm{prox}_f(\cdot)$ and $\mathrm{prox}_g(\cdot)$, and

$$\kappa_1 = 2 \max\left\{\alpha^2 L + \frac{1}{\tau^2}, L + \frac{1}{\sigma^2}\right\}. \tag{3.27}$$

So, it follows from the last inequality and (3.23) that for all $k \ge \overline{k}$,

$$\mathrm{dist}(\widetilde{u}^k, \Omega^*) \le \theta\sqrt{\kappa_1}\|u^k - \widetilde{u}^k\|. \tag{3.28}$$

Then, by triangle inequality and (3.28), for all $k \ge \overline{k}$, we have

$$\begin{aligned} &\frac{1}{\sqrt{\lambda_{\max}(H)}} \mathrm{dist}_H(u^k, \Omega^*) \le \mathrm{dist}(u^k, \Omega^*) \le \mathrm{dist}(\widetilde{u}^k, \Omega^*) + \|u^k - \widetilde{u}^k\| \\ \le \quad &(1 + \theta\sqrt{\kappa_1})\|u^k - \widetilde{u}^k\| \le \frac{1 + \theta\sqrt{\kappa_1}}{\sqrt{\lambda_{\min}(G)}}\|u^k - \widetilde{u}^k\|_G. \end{aligned} \tag{3.29}$$

Since (3.7) holds for any $u^* \in \Omega^*$, for all $k \ge 0$ we have

$$\mathrm{dist}_H^2(u^{k+1}, \Omega^*) \le \mathrm{dist}_H^2(u^k, \Omega^*) - \|u^k - \widetilde{u}^k\|_G^2, \tag{3.30}$$

which together with (3.29) gives

$$\mathrm{dist}_H(u^{k+1}, \Omega^*) \le \sqrt{1 - \frac{1}{(1 + \theta\sqrt{\kappa_1})^2}\frac{\lambda_{\min}(G)}{\lambda_{\max}(H)}} \mathrm{dist}_H(u^k, \Omega^*) \tag{3.31}$$

for all $k \geq \overline{k}$. Finally, (3.30) and (3.31) implies there exists a $\xi \in (0,1)$ such that (3.24) holds, that is, the sequence $\{\mathrm{dist}_H(u^k, \Omega^*)\}$ converges to zero Q-linearly.

Now, let $d^k = u^{k+1} - u^k$. We have from (3.30) and triangle inequality that

$$\begin{aligned} \left\|d^k\right\|_H &= \left\|u^{k+1} - u^k\right\|_H \leq \mathrm{dist}_H(u^k, \Omega^*) + \mathrm{dist}_H((u^{k+1}, \Omega^*) \\ &\leq 2\mathrm{dist}_H(u^k, \Omega^*) \overset{(3.24)}{\leq} 2\xi^k \mathrm{dist}_H(u^0, \Omega^*). \end{aligned}$$

Hence, we have from $u^\infty = u^k + \sum_{j=k}^{\infty} d^j$ that

$$\begin{aligned} \left\|u^k - u^\infty\right\|_H &\leq \textstyle\sum_{j=k}^{\infty} \left\|d^j\right\|_H \leq 2\mathrm{dist}_H(u^0, \Omega^*) \sum_{j=k}^{\infty} \xi^j \\ &= 2\mathrm{dist}_H(u^0, \Omega^*)\xi^k \textstyle\sum_{j=0}^{\infty} \xi^j = \xi^k \left(2\mathrm{dist}_H(u^0, \Omega^*)\frac{1}{1-\xi}\right), \end{aligned}$$

which implies the sequence $\{\|u^k - u^\infty\}$ converges to zero R-linearly. □

Theorem 3.4 shows linear convergence of G-AFBA under the calmness condition. In practice, it is not easy to check whether the calmness condition (3.23) holds or not. However, when the mapping R defined by (3.22) is piecewise polyhedral, or equivalently, R^{-1} is piecewise polyhedral, we know (e.g. see [39]) there exist two constants $\beta, \eta > 0$ such that

$$\mathrm{dist}(u, \Omega^*) \leq \beta\|R(u)\|, \quad \forall u \in \left\{u \in \Omega \middle| \|R(u)\| \leq \eta\right\}. \tag{3.32}$$

When $R(u) > \eta$, for all $\|u - u^\infty\| \leq r$ with some $r > 0$, we have

$$\mathrm{dist}(u, \Omega^*) \leq \|u - u^\infty\| \leq r < \frac{r}{\eta}\|R(u)\|. \tag{3.33}$$

So, given any $r > 0$, we have from (3.32) and (3.33) that the calmness condition (3.23) holds with $\theta = \max\{\beta, r/\eta\}$. Moreover, by Theorem 3.1, there exists a $\overline{r} > 0$ such that $\|u^k - u^\infty\| \leq \overline{r}$ for all $k \geq 0$. Hence, when the mapping R defined by (3.22) is piecewise polyhedral, for $\{u^k\}$ generated by G-AFBA, we have $\mathrm{dist}(u^k, \Omega^*) \leq \overline{\theta}\|R(u^k)\|$ for some $\overline{\theta} > 0$. Furthermore, by Theorem 3.4, we have $\{\mathrm{dist}_H(u^k, \Omega^*)\}$ converges to zero Q-linearly and $\{\|u^k - u^\infty\}$ converges to zero R-linearly. Here, we want to mention that linear convergence had been also discussed when assuming certain strongly convexity on the objective function (see e.g. [10, 11]).

4 Connections between (1.8) and other related methods

In this section, we discuss in a bit more detail on the connections between G-AFBA (1.8) and some existing and new related algorithms.

- **Case 1 (CP-PPA in [9] and a reduced ALM).** When $\alpha = 1$, G-AFBA (1.8) will reduce to

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top y^k\right\|^2, \\ y^{k+1} = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma}\left\|y - y^k - \sigma K(2\widetilde{x}^k - x^k)\right\|^2, \end{cases}$$

  which is CP-PPA proposed in [9]. When $\alpha = 1$ and $g = 0$, the problem (1.1) is equivalent to

$$\min f(x) \quad \text{s.t. } Kx = \mathbf{0},\ x \in \mathcal{X}, \tag{4.1}$$

and G-AFBA (1.8) recovers a ALM-type method

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top \lambda^k\right\|^2, \\ \lambda^{k+1} = \lambda^k + \sigma K(2x^{k+1} - x^k). \end{cases}$$

Note that two different parameters τ and σ are exploited here, which is different from the standard augmented Lagrangian method for solving (4.1).

- **Case 2 (Exact version of [30, Algorithm 2]).** When $(\alpha, \mu) = (0, 1)$, G-AFBA reduces to

$$\begin{cases} \overline{x}^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top y^k\right\|^2, \\ y^{k+1} = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma}\left\|y - y^k - \sigma K\overline{x}^{k+1}\right\|^2, \\ x^{k+1} = \overline{x}^{k+1} - \tau K^\top (y^{k+1} - y^k), \end{cases} \tag{4.2}$$

which is the exact version of [30, Algorithm 2] by setting the iterative relative error to zero. For this case, the condition (1.9) reduces to $1/(\sigma\tau) > L$, which matches the condition given in [30].

- **Case 3 (A subclass of G-AFBA).** By setting $\alpha = 0$, G-AFBA reduces to

$$\text{(G1-AFBA)} \quad \begin{cases} \overline{x}^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top y^k\right\|^2, \\ \overline{y}^{k+1} = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma}\left\|y - y^k - \sigma K\overline{x}^{k+1}\right\|^2, \\ x^{k+1} = \overline{x}^{k+1} - \mu\tau K^\top (\overline{y}^{k+1} - y^k), \\ y^{k+1} = \overline{y}^{k+1} + (1-\mu)\sigma K(\overline{x}^{k+1} - x^k). \end{cases} \tag{4.3}$$

One may consider (4.3) as an extension of (4.2), since (4.3) applies an additional extrapolation step on the y-iterate, while the x^{k+1}-iterate in (4.3) can be written as

$$x^{k+1} = \overline{x}^{k+1} - \tau K^\top (\overline{y}^{k+1} - y^k) + (1-\mu)\tau K^\top (\overline{y}^{k+1} - y^k).$$

Interestingly, with $\alpha = 0$, the condition (1.9) for convergence reduces to

$$\frac{1}{\tau\sigma} > (1 - \mu + \mu^2)L. \tag{4.4}$$

Clearly, $(1 - \mu + \mu^2) \le 1$ for any $\mu \in [0, 1]$ and when $\mu = 0.5$, it becomes $\frac{1}{\tau\sigma} > 0.75L$. The condition (4.4) seems similar to the condition (1.7) for ensuring convergence of GCP-PPA [24]. However, we can see from (4.3) that G1-AFBA is completely a different method from GCP-PPA (1.6).

- **Case 4 (GCP-PPA [24]).** When $\mu = 0$, G-AFBA reduces to

$$\begin{cases} x^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau}\left\|x - x^k + \tau K^\top y^k\right\|^2, \\ \overline{y}^{k+1} = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma}\left\|y - y^k - \sigma K[x^{k+1} + \alpha(x^{k+1} - x^k)]\right\|^2, \\ y^{k+1} = \overline{y}^{k+1} + (1-\alpha)\sigma K(x^{k+1} - x^k), \end{cases} \tag{4.5}$$

which is the method (1.6) proposed in [24]. As mentioned in the introduction, in this case the condition (1.9) will reduce to (1.7), which is exactly the condition derived in [24] for the convergence of GCP-PPA. Moreover, as pointed in [24], GCP-PPA is equivalent to CP-PPA for solving the the convex programming $\min\{f(x) \mid Kx = b, x \in \mathcal{X}\}$.

- **Case 5 (G-AFBA for multi-block problem).** Consider the following saddle-point problem with multi-block structure:

$$\min_{x\in\mathbb{R}^n}\max_{\lambda\in\mathbb{R}^m} \mathcal{L}(x,\lambda) := \sum_{i=1}^{q} f_i(x_i) + \langle Kx, \lambda\rangle - \langle b, \lambda\rangle, \tag{4.6}$$

where each f_i, $i = 1,\dots,q$, is a proper lower semicontinuous convex function, $x = (x_1,\cdots,x_q)^\top$ with $x_i \in \mathbb{R}^{n_i}$, $K = (A_1,\cdots,A_q)$ is given with $A_i \in \mathbb{R}^{m\times n_i}$ and $n = \sum_{i=1}^{q} n_i$. Clearly, the problem (4.6) is a special case of (1.1) and is the dual problem of the following multi-block separable convex optimization problem

$$\min\left\{\sum_{i=1}^{q} f_i(x_i)\Big|\ \sum_{i=1}^{q} A_i x_i = b,\ x_i \in \mathbb{R}^{n_i}\right\}. \tag{4.7}$$

Applying G-AFBA (1.8) to (4.6) results in the following operator splitting method:

$$\begin{cases} \overline{x}_i^{k+1} = \arg\min_{x_i\in\mathbb{R}^{n_i}} f_i(x_i) + \frac{1}{2\tau}\left\|x_i - x_i^k + \tau {A_i}^\top\lambda^k\right\|^2,\ i = 1,\cdots,q,\\ \overline{\lambda}^{k+1} = \lambda^k + \sigma\sum_{i=1}^{q} A_i\left[\overline{x}_i^{k+1} + \alpha(\overline{x}_i^{k+1} - x_i^k)\right] - b,\\ x_i^{k+1} = \overline{x}_i^{k+1} - (1-\alpha)\mu\,\tau {A_i}^\top(\overline{\lambda}^{k+1} - \lambda^k),\ i = 1,\cdots,q,\\ \lambda^{k+1} = \overline{\lambda}^{k+1} + (1-\alpha)(1-\mu)\,\sigma\sum_{i=1}^{q} A_i(\overline{x}_i^{k+1} - x_i^k). \end{cases} \tag{4.8}$$

Note that the above scheme (4.8) updates the primal variable x_i in parallel and is different from the proximal ADMM proposed [16] for solving (4.7). However, by our previous analysis, the scheme (4.8) will enjoy all the convergent properties we discussed before.

5 More extensions

In this section, we would give an adaptive and a stochastic versions of G-AFBA, and we briefly discuss their convergence properties.

5.1 Extension to an adaptive G-AFBA

Our adaptive G-AFBA (see Algorithm 5.1) as well as its convergence theory are motivated from an adaptive PDHG (a-PDHG) developed in [20]. In fact, a-PDHG can be considered as a special case of aG-AFBA, which is almost identical to G-AFBA except using adaptive stepsizes (τ_k,σ_k). In particular, in Algorithm 5.1, the stepsizes (τ_k,σ_k) are adjusted according to the ratio between the PrimalError(k) (error related to x-variable at the k-th iteration) and DualError(k) (error related to y-variable at the k-th iteration), which can be defined/chosen by the user in various ways such that problem (1.1) is solved as long as $\max\{\text{PrimalError}(k), \text{DualError}(k)\} = 0$. The goal is to adaptively adjust the stepsizes (τ_k,σ_k) so that both the primal error and the dual error can be reduced in a balanced way. Hence, the overall acceleration of Algorithm 5.1 can be achieved. Moreover, it is not difficult to show (one may see [20] for details) the stepsizes (τ_k,σ_k) in Algorithm 5.1 satisfy the following conditions:

(A1) Both $\{\tau_k\}$ and $\{\sigma_k\}$ are positive, bounded, and the product $\tau_k\sigma_k = \tau_0\sigma_0 := C_{\tau\sigma}$ satisfying (1.9);

(A2) The sequence $\{\phi_k\}$ is summable, where $\phi_k = \max\left\{\frac{\tau_k - \tau_{k+1}}{\tau_k}, \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}, 0\right\}$.

Algorithm 5.1: An adaptive G-AFBA (aG-AFBA).

Initialization: Choose $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ and (τ_0, σ_0) satisfying (1.9), set $\alpha, \mu \in [0, 1]$, $\theta_0 = \eta = 0.95$ and $\gamma_1 > 1 > \gamma_2 > 0$, given $\epsilon > 0$.

for $k = 0, 1, \cdots$

1. $\widetilde{x}^k = \arg\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\tau_k}\left\|x - x^k + \tau_k K^\top y^k\right\|^2$;
2. $\widetilde{y}^k = \arg\min_{y\in\mathcal{Y}} g(y) + \frac{1}{2\sigma_k}\left\|y - y^k - \sigma_k K[\widetilde{x}^k + \alpha(\widetilde{x}^k - x^k)]\right\|^2$;
3. $x^{k+1} = \widetilde{x}^k - (1-\alpha)\mu\tau_k K^\top(\widetilde{y}^k - y^k)$;
4. $y^{k+1} = \widetilde{y}^k + (1-\alpha)(1-\mu)\sigma_k K(\widetilde{x}^k - x^k)$;
5. **if** DualError$(k) > \gamma_1$ PrimalError(k), **then**
6. $\quad\tau_{k+1} = \tau_k(1-\theta_k), \sigma_{k+1} = \sigma_k/(1-\theta_k), \theta_{k+1} = \theta_k\eta$;
7. **else if** DualError$(k) < \gamma_2$ PrimalError(k), **then**
8. $\quad\tau_{k+1} = \tau_k/(1-\theta_k), \sigma_{k+1} = \sigma_k(1-\theta_k), \theta_{k+1} = \theta_k\eta$;
9. **end if**
10. **if** $\max\{\text{DualError}, \text{PrimalError}\} \le \epsilon$, break;

end for

Return (x^{k+1}, y^{k+1}).

To analyze global convergence of aG-AFBA, analogous to the previous analysis in Section 3, let us define the following matrices:

$$Q_k = \begin{bmatrix} \frac{1}{\tau_k}\mathbf{I} & -K^\top \\ -\alpha K & \frac{1}{\sigma_k}\mathbf{I} \end{bmatrix}, \quad M_k = \begin{bmatrix} \mathbf{I} & -(1-\alpha)\mu\tau_k K^\top \\ (1-\alpha)(1-\mu)\sigma_k K & \mathbf{I} \end{bmatrix},$$

$$H_k = Q_k M_k^{-1} = \begin{bmatrix} \mathcal{T}_k & -(1-\mu+\alpha\mu)K^\top \\ -(1-\mu+\alpha\mu)K & \Sigma_k \end{bmatrix} \begin{bmatrix} \mathbf{C}_x^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_\mathbf{y}^{-1} \end{bmatrix},$$

where

$$\mathcal{T}_k = \frac{1}{\tau_k}\Big(\mathbf{I} + (1-\alpha)(1-\mu)C_{\tau\sigma}K^\top K\Big), \quad \Sigma_k = \frac{1}{\sigma_k}\Big(\mathbf{I} - (1-\alpha)\alpha\mu C_{\tau\sigma}KK^\top\Big),$$

and

$$\mathbf{C}_x = \mathbf{I} + (1-\alpha)^2(1-\mu)\mu C_{\tau\sigma}K^\top K, \quad \mathbf{C}_y = \mathbf{I} + (1-\alpha)^2(1-\mu)\mu C_{\tau\sigma}KK^\top.$$

In addition, we define

$$P_k = \begin{bmatrix} \mathcal{T}_k & \mathbf{0} \\ \mathbf{0} & \Sigma_k \end{bmatrix} \begin{bmatrix} \mathbf{C}_x^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_\mathbf{y}^{-1} \end{bmatrix} = \begin{bmatrix} \mathcal{T}_k\mathbf{C}_x^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_k\mathbf{C}_y^{-1} \end{bmatrix}.$$

Both $\mathbf{C}_x$ and $\mathbf{C}_y$ are symmetric positive definite, so does their inverse. By Proposition 2.1, H_k is symmetric and positive definite. Hence, we have $\mathcal{T}_k\mathbf{C}_x^{-1}$ and $\Sigma_k\mathbf{C}_y^{-1}$ are symmetric and positive definite, and $(K^\top\mathbf{C}_y^{-1})^\top = K\mathbf{C}_x^{-1}$.

Based on the above preparations, we next show that the sequence $\{u^k - u^*\}$ is upper bounded, which is essential for deriving the convergence rate of aG-AFBA.

Lemma 5.1 *Suppose the parameters τ_k and σ_k in aG-AFBA satisfy the assumptions (A1)-(A2). Then, we have*

$$\|u^k - u^*\|_{P_k}^2 \le c_u \tag{5.1}$$

for some upper bound $c_u > 0$.

Proof. Since $\tau_k\sigma_k = C_{\tau\sigma}$ satisfies (1.9), by Proposition 2.1, H_k is positive definite. In addition, there exists an $\epsilon \in (0,1)$ independent of k such that $(\widetilde{\tau}_k, \widetilde{\sigma}_k)$ still satisfies (1.9), where $\widetilde{\tau}_k = \tau_k/(1-\epsilon)$ and $\widetilde{\sigma}_k = \sigma_k/(1-\epsilon)$. Hence, similar to Proposition 2.1, $H_k - \epsilon P_k$ is still positive definite for all $k \geq 0$. Since H_k is positive definite, for any $u \in \mathcal{X} \times \mathcal{Y}$ we have

$$u^\top H_k u \geq 0 \Longleftrightarrow u^\top P_k u \geq 2(1-\mu+\alpha\mu) y^\top K \mathbf{C}_x^{-1} x.$$

Then, we have from $H_k - \epsilon P_k$ being positive definite that

$$c_H \|u\|_{H_k}^2 \geq 2(1-\mu+\alpha\mu) y^\top K \mathbf{C}_x^{-1} x, \tag{5.2}$$

for any $k \geq 0$ and any $u \in \mathcal{X}\times\mathcal{Y}$, where $c_H = 1/\epsilon$. By taking $u = (x^{k+1}-x^*, y^{k+1}-y^*)$ and H_{k+1} in the above inequality, we have

$$2(1-\mu+\alpha\mu)(y^{k+1}-y^*)^\top K \mathbf{C}_x^{-1}(x^{k+1}-x^*) \leq c_H \|u^{k+1}-u^*\|_{H_{k+1}}^2.$$

So, from the above inequality, $(K^\top \mathbf{C}_y^{-1})^\top = K\mathbf{C}_x^{-1}$, (3.7) and (5.2), we have

$$\begin{aligned}
&\|u^k - u^*\|_{H_k}^2 \geq \|u^{k+1}-u^*\|_{H_k}^2 \\
=&\|u^{k+1}-u^*\|_{P_k}^2 - 2(1-\mu+\alpha\mu)(y^{k+1}-y^*)^\top K\mathbf{C}_x^{-1}(x^{k+1}-x^*) \\
\geq&\delta_k \|u^{k+1}-u^*\|_{P_{k+1}}^2 - 2(1-\mu+\alpha\mu)(y^{k+1}-y^*)^\top K\mathbf{C}_x^{-1}(x^{k+1}-x^*) \\
=&\delta_k \|u^{k+1}-u^*\|_{H_{k+1}}^2 - 2(1-\delta_k)(1-\mu+\alpha\mu)(y^{k+1}-y^*)^\top K\mathbf{C}_x^{-1}(x^{k+1}-x^*) \\
\geq&\delta_k \|u^{k+1}-u^*\|_{H_{k+1}}^2 - c_H(1-\delta_k)\|u^{k+1}-u^*\|_{H_{k+1}}^2 \\
=&\left\{1-\phi_k\left[1+c_H\right]\right\}\|u^{k+1}-u^*\|_{H_{k+1}}^2,
\end{aligned}$$

where the second inequality uses

$$\mathcal{T}_k = \frac{\tau_{k+1}}{\tau_k}\mathcal{T}_{k+1}, \quad \Sigma_k = \frac{\sigma_{k+1}}{\sigma_k}\Sigma_{k+1}, \quad \delta_k := 1-\phi_k = \min\left\{\frac{\tau_{k+1}}{\tau_k}, \frac{\sigma_{k+1}}{\sigma_k}, 1\right\},$$

and the second equality uses the relationship between H_k and P_k.

Since the sequence $\{\phi_k\}$ is summable, we have $\phi_k(1+c_H) \in (0,1)$ for all k sufficiently large. Hence, we assume, without loss of generality, that $\phi_k(1+c_H) \in (0,1)$ for all k. Then, it follows that

$$\|u^0 - u^*\|_{H_0}^2 \geq \prod_{j=0}^{k-1}\left\{1-\phi_j(1+c_H)\right\}\|u^k-u^*\|_{H_k}^2. \tag{5.3}$$

Since $\sum_{j=0}^{\infty}\phi_j < \infty$, we have $\prod_{j=0}^{\infty}\left\{1-\phi_j(1+c_2)\right\} \geq 1/c_1$ for some $c_1 > 0$. So, we have from (5.3) that

$$\|u^n - u^*\|_{H_n}^2 \leq c_1\|u^0-u^*\|_{H_0}^2,$$

which together with $H_n - \epsilon P_n$ being positive definite gives (5.1). □

Lemma 5.2 *Let $c_u > 0$ be given by Lemma 5.1 and $c_\phi = \sum_{k=0}^{\infty}\phi_k$. Then, under the assumptions (A1) and (A2), we have*

$$\sum_{k=1}^{n}\left(\|u^k-u\|_{H_k}^2 - \|u^k-u\|_{H_{k-1}}^2\right) \leq 2c_\phi\left(c_u + c_p\|u-u^*\|^2\right),$$

where c_p is a constant such that $\|u-u^*\|_{P_k}^2 \leq c_p\|u-u^*\|^2$.

Proof. Since $\mathcal{T}_k\mathbf{C}_x^{-1}$ and $\Sigma_k\mathbf{C}_y^{-1}$ are positive definite, it follows from the definition of ϕ_k that

$$\begin{cases} \mathcal{T}_k\mathbf{C}_x^{-1} - \mathcal{T}_{k-1}\mathbf{C}_x^{-1} = \dfrac{\tau_{k-1}-\tau_k}{\tau_{k-1}}\mathcal{T}_k\mathbf{C}_x^{-1} \preceq \phi_{k-1}\mathcal{T}_k\mathbf{C}_x^{-1}, \\ \Sigma_k\mathbf{C}_y^{-1} - \Sigma_{k-1}\mathbf{C}_y^{-1} = \dfrac{\sigma_{k-1}-\sigma_k}{\sigma_{k-1}}\Sigma_k\mathbf{C}_y^{-1} \preceq \phi_{k-1}\Sigma_k\mathbf{C}_y^{-1}. \end{cases}$$

Then, by the definitions of H_k and P_k and (5.1), we have

$$\begin{aligned} &\sum_{k=1}^{n}\left(\|u^k-u\|_{H_k}^2 - \|u^k-u\|_{H_{k-1}}^2\right) \\ =&\sum_{k=1}^{n}\left(\|x^k-x\|_{(\mathcal{T}_k-\mathcal{T}_{k-1})\mathbf{C}_x^{-1}}^2 + \|y^k-y\|_{(\Sigma_k-\Sigma_{k-1})\mathbf{C}_y^{-1}}^2\right) \\ \leq&\sum_{k=1}^{n}\phi_{k-1}\|u^k-u\|_{P_k}^2 \leq 2\sum_{k=1}^{n}\phi_{k-1}\left(\|u^k-u^*\|_{P_k}^2 + \|u-u^*\|_{P_k}^2\right) \\ \leq&2c_\phi\big(c_u + c_p\|u-u^*\|^2\big) < \infty, \end{aligned}$$

where the last inequality follows from the definition of c_p such that $\|u-u^*\|_{P_k}^2 \leq c_p\|u-u^*\|^2$ for all k. Note that such c_p exists by Assumption (A1). □

Theorem 5.1 *Let $x_t^N = \frac{1}{t}\sum_{k=0}^{t-1}\tilde{x}^k$, $y_t^N = \frac{1}{t}\sum_{k=0}^{t-1}\tilde{y}^k$. Then, under the conditions given in Lemma 5.2, for any $u \in \mathcal{X}\times\mathcal{Y}$ we have*

$$\mathcal{L}(x_t^N, y) - \mathcal{L}(x, y_t^N) \leq \frac{\|u-u_0\|_{H_0}^2 + 2c_\phi c_u + c_\phi c_p\|u-u^*\|^2}{2t}, \tag{5.4}$$

where c_ϕ, c_u and c_p are the same constants given in Lemma 5.2.

Proof. Summing (3.14) over $k = 0, 1, \cdots, t-1$ together with Lemma 5.2, we obtain

$$\begin{aligned} &2\sum_{k=0}^{t-1}\left[\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)\right] \\ \leq&\|u-u_0\|_{H_0}^2 - \|u-u_t\|_{H_t}^2 + \sum_{k=1}^{t}\left(\|u-u_k\|_{H_k}^2 - \|u-u_k\|_{H_{k-1}}^2\right) \\ \leq&\|u-u_0\|_{H_0}^2 - \|u-u_t\|_{H_t}^2 + 2c_\phi c_u + 2c_\phi c_p\|u-u^*\|^2, \end{aligned}$$

which, by the convexity of f, g, the definitions of x_t^N and y_t^N, yields

$$\mathcal{L}(x_t^N, y) - \mathcal{L}(x, y_t^N) \leq \frac{\|u-u_0\|_{H_0}^2 - \|u-u_t\|_{H_t}^2 + 2c_\phi c_u + 2c_\phi c_p\|u-u^*\|^2}{2t}$$

and immediately gives (5.4). □

5.2 Extension to a stochastic G-AFBA

Now, let us consider the following case of special structured (1.1):

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x) + \langle Kx, y\rangle - g(y), \quad \text{where} \quad f(x) = \frac{1}{N}\sum_{j=1}^{N} f_j(x) \tag{5.5}$$

is an average of N Lipschitz continuously differentiable real-valued convex functions f_j, $j = 1, \ldots, N$, i.e., there exists a $\nu > 0$ such that

$$\|\nabla f_j(x_1) - \nabla f_j(x_2)\| \leq \nu \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathcal{X}.$$

Problem (5.5) often arises from machine learning applications, e.g. [4, 6], where N denotes the sample size and $f_j(x)$ corresponds to the empirical loss on the j-th sample data. A major difficulty for solving (5.5) in machine learning applications is that the sample size N can be huge so that it is computationally prohibitive to evaluate either the function value f or its gradient at each iteration. Hence, in this subsection, by extending the previous analysis of deterministic G-AFBA, we aim to develop a stochastic version of G-AFBA, see Algorithm 5.2, for solving the structured problem (5.5). In the following, we briefly discuss the convergence properties of SG-AFBA following a similar approach proposed in [4].

Algorithm 5.2: A stochastic G-AFBA (SG-AFBA).

Initialization: choose (τ, σ) satisfying (1.9), $\alpha, \mu \in [0, 1]$ and initialize $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, $\breve{x}^0 = x^0$.
for $k = 0, 1, \cdots$
1. Choose $m_k > 0, \vartheta_k > 0$, and compute $h^k = x^k - \tau K^\top y^k$;
2. $(\widetilde{x}^k, \breve{x}^{k+1}) = \mathbf{xsub}(x^k, \breve{x}^k, \vartheta_k, m_k, h^k)$;
3. $\widetilde{y}^k = \arg\min_{y \in \mathcal{Y}} g(y) + \frac{1}{2\sigma} \left\| y - y^k - \sigma K[\widetilde{x}^k + \alpha(\widetilde{x}^k - x^k)] \right\|^2$;
4. $x^{k+1} = \widetilde{x}^k - (1-\alpha)\mu\, \tau K^\top (\widetilde{y}^k - y^k)$;
5. $y^{k+1} = \widetilde{y}^k + (1-\alpha)(1-\mu)\, \sigma K(\widetilde{x}^k - x^k)$;

end for
Return (x^{k+1}, y^{k+1}).

$(\mathbf{x}^+, \breve{\mathbf{x}}^+) = \mathbf{xsub}(x_1, \breve{x}_1, \vartheta_k, m_k, h^k)$
for $t = 1, 2, \ldots, m_k$
1. Randomly select $\xi_t \in \{1, 2, \ldots, N\}$ with uniform probability;
2. $\beta_t = 2/(t+1)$, $\quad \gamma_t = 2/(t\vartheta_k)$, $\quad \widehat{x}_t = \beta_t \breve{x}_t + (1-\beta_t)x_t$;
3. $d_t = \widehat{g}_t + e_t$, where $\widehat{g}_t = \nabla f_{\xi_t}(\widehat{x}_t)$ and e_t is a random vector satisfying $\mathbb{E}[e_t] = \mathbf{0}$;
4. $\breve{x}_{t+1} = \arg\min_{x \in \mathcal{X}} \langle d_t, x \rangle + \frac{\gamma_t}{2} \left\| x - \breve{x}_t \right\|^2 + \frac{1}{2\tau} \left\| x - h^k \right\|^2$;
5. $x_{t+1} = \beta_t \breve{x}_{t+1} + (1-\beta_t)x_t$;

end for
Return $(\mathbf{x}^+, \breve{\mathbf{x}}^+) = (\mathbf{x}_{m_k+1}, \breve{\mathbf{x}}_{m_k+1})$.

We first need to obtain a variational inequality analogous to (3.3) for establishing the convergence of SG-AFBA. Note that the $\breve{x}_{t+1}$-subproblem in step 4 of subroutine **xsub** amounts to

$$\breve{x}_{t+1} = \arg\min_{x \in \mathcal{X}} \langle d_t + K^\top y^k, x \rangle + \frac{\gamma_t}{2} \left\| x - \breve{x}_t \right\|^2 + \frac{1}{2\tau} \left\| x - x^k \right\|^2.$$

Hence, almost same to the proof of [4, Lemma 3.1], we have the following lemma.

Lemma 5.3 *Let us define* $\Gamma_t = 2/(t(t+1))$ *and*

$$\phi_k(x) = f(x) + \psi_k(x), \qquad \textit{where} \;\; \psi_k(x) = \frac{1}{2\tau} \left\| x - x^k \right\|^2 + \langle K^\top y^k, x \rangle. \tag{5.6}$$

Then, for any $x \in \mathcal{X}$ and k with $\vartheta_k \in (0, 1/\nu)$, we have

$$\frac{1}{\Gamma_t}\left[\phi_k(x_{t+1}) - \phi_k(x)\right] \le \begin{cases} \theta_1, & t = 1, \\ \frac{1}{\Gamma_{t-1}}\left[\phi_k(x_t) - \phi_k(x)\right] + \theta_t, & t \ge 2, \end{cases} \tag{5.7}$$

where for all $t \ge 1$,

$$\theta_t = \frac{1}{\vartheta_k}\left[\|x - \breve{x}_t\|^2 - \|x - \breve{x}_{t+1}\|^2\right] - \frac{t}{2\tau}\|x - \breve{x}_{t+1}\|^2 + t\langle \boldsymbol{\delta}_t, \breve{x}_t - x\rangle + \frac{\vartheta_k t^2}{4}\frac{\|\boldsymbol{\delta}_t\|^2}{(1 - \vartheta_k\nu)}, \tag{5.8}$$

and $\boldsymbol{\delta}_t = \nabla f(\widehat{x}_t) - d_t$.

Based on Lemma 5.3, we further establish the following result.

Lemma 5.4 *Let $\boldsymbol{\delta}_t$ be defined in Lemma 5.3, and suppose $\vartheta_k \in (0, 1/\nu)$. Then the iterates generated by SG-AFBA satisfy*

$$f(x) - f(\widetilde{x}^k) + \langle x - \widetilde{x}^k, K^\top y^k + \frac{1}{\tau}(\widetilde{x}^k - x^k)\rangle \ge \zeta^k, \tag{5.9}$$

for all $x \in \mathcal{X}$, where

$$\begin{aligned}\zeta^k \quad = \quad & \frac{2}{m_k(m_k+1)}\left[\frac{1}{\vartheta_k}\left(\left\|x - \breve{x}^{k+1}\right\|^2 - \left\|x - \breve{x}^k\right\|^2\right)\right. \\ & \left. - \sum_{t=1}^{m_k} t\langle \boldsymbol{\delta}_t, \breve{x}_t - x\rangle - \frac{\vartheta_k}{4(1 - \vartheta_k\nu)}\sum_{t=1}^{m_k} t^2\|\boldsymbol{\delta}_t\|^2\right].\end{aligned} \tag{5.10}$$

Proof. Let $T = m_k$. Summing (5.7) over $1 \le t \le T$ and recalling that $\breve{x}^k = \breve{x}_1$, $\widetilde{x}^k = x_{T+1}$, and $\breve{x}^{k+1} = \breve{x}_{T+1}$, we obtain

$$\begin{aligned}\frac{1}{\Gamma_T}\left[\phi_k(\widetilde{x}^k) - \phi_k(x)\right] \le \sum_{t=1}^{T}\theta_t = \frac{1}{\vartheta_k}\left[\left\|x - \breve{x}^k\right\|^2 - \left\|x - \breve{x}^{k+1}\right\|^2\right] \\ -\frac{1}{2\tau}\sum_{t=1}^{T} t\|x - \breve{x}_{t+1}\|^2 + \sum_{t=1}^{T} t\langle \boldsymbol{\delta}_t, \breve{x}_t - x\rangle + \frac{\vartheta_k}{4(1 - \vartheta_k\nu)}\sum_{t=1}^{T} t^2\|\boldsymbol{\delta}_t\|^2\end{aligned} \tag{5.11}$$

for any $x \in \mathcal{X}$, where θ_t is defined in (5.8). Dividing $x_{t+1} = \beta_t\breve{x}_{t+1} + (1 - \beta_t)x_t$ by Γ_t and exploiting the identity $\beta_t/\Gamma_t = t$ yields $(1/\Gamma_t)x_{t+1} = (1/\Gamma_{t-1})x_t + t\breve{x}_{t+1}$. Sum this equality over $2 \le t \le T$ and recall $\Gamma_1 = \beta_1 = 1$ to obtain

$$\begin{aligned}\widetilde{x}^k \quad & = \quad x_{T+1} = \Gamma_T\left\{\frac{1}{\Gamma_1}x_2 + \sum_{t=2}^{T} t\breve{x}_{t+1}\right\} = \Gamma_T\left\{x_2 - \breve{x}_2 + \sum_{t=1}^{T} t\breve{x}_{t+1}\right\} \\ & = \quad \Gamma_T\left\{\left[\beta_1\breve{x}_2 + (1 - \beta_1)x_1\right] - \breve{x}_2 + \sum_{t=1}^{T} t\breve{x}_{t+1}\right\} = \sum_{t=1}^{T}(t\Gamma_T)\breve{x}_{t+1}.\end{aligned} \tag{5.12}$$

Since $\Gamma_T\sum_{t=1}^{T} t = 1$ and $\|z - x\|^2$ is convex in z, it follows from (5.12) that

$$\left\|\widetilde{x}^k - x\right\|^2 \le \sum_{t=1}^{T}(t\Gamma_T)\|\breve{x}_{t+1} - x\|^2, \qquad \forall x \in \mathcal{X}.$$

Plug the last inequality into (5.11) to obtain

$$\begin{aligned}\tfrac{1}{\Gamma_T}\left[\phi_k(\widetilde{x}^k) - \phi_k(x) + \tfrac{1}{2\tau}\left\|\widetilde{x}^k - x\right\|^2\right] \le \tfrac{1}{\vartheta_k}\left[\left\|x - \breve{x}^k\right\|^2 - \left\|x - \breve{x}^{k+1}\right\|^2\right] \\ + \sum_{t=1}^{T} t\langle \boldsymbol{\delta}_t, \breve{x}_t - x\rangle + \tfrac{\vartheta_k}{4(1 - \vartheta_k\nu)}\sum_{t=1}^{T} t^2\|\boldsymbol{\delta}_t\|^2.\end{aligned} \tag{5.13}$$

Now, by the definitions of ϕ_k and ψ_k in (5.6), we have

$$\begin{cases} \phi_k(\widetilde{x}^k) - \phi_k(x) = f(\widetilde{x}^k) - f(x) + \psi_k(\widetilde{x}^k) - \psi_k(x), \\ \psi_k(\widetilde{x}^k) - \psi_k(x) = \langle K^\top y^k, \widetilde{x}^k - x \rangle + \frac{1}{2\tau}\left[\|\widetilde{x}^k - x^k\|^2 - \|x - x^k\|^2\right]. \end{cases}$$

The identity $(\mathbf{a}-\mathbf{b})^\top(\mathbf{a}-\mathbf{c}) = \frac{1}{2}\left\{\|\mathbf{a}-\mathbf{c}\|^2 - \|\mathbf{c}-\mathbf{b}\|^2 + \|\mathbf{a}-\mathbf{b}\|^2\right\}$ with $\mathbf{a} = \widetilde{x}^k$, $\mathbf{b} = x^k$, and $\mathbf{c} = x$ implies that

$$\frac{1}{2}\left[\|\widetilde{x}^k - x^k\|^2 - \|x - x^k\|^2 + \|\widetilde{x}^k - x\|^2\right] = (\widetilde{x}^k - x^k)^\top(\widetilde{x}^k - x).$$

Insert all these relations in (5.13) and make the substitutions $T = m_k$ and $\Gamma_T = 2/(T(T+1))$ with simple transformation to obtain (5.9). $\square$

Now, replacing the inequality (3.4) by (5.9), under the condition (1.9), we will have from the same proofs of Lemmas 3.1-3.2 that

$$\theta(u)-\theta(\widetilde{u}^k)+\langle u-\widetilde{u}^k, \mathcal{J}(u)\rangle \geq \frac{1}{2}\left(\|u-u^{k+1}\|_H^2-\|u-u^k\|_H^2\right)+\frac{1}{2}\|u^k-\widetilde{u}^k\|_G^2+\zeta^k, \quad (5.14)$$

where H and G are positive definite matrices defined in (2.6). With the help of (5.14), we have the following theorem.

Theorem 5.2 *Let $u_T = (x_T, y_T)$ be defined in (3.12). If for some integers $T > 0$ and $\kappa \geq 0$, the following conditions hold for all $k \in [\kappa, \kappa+T]$: (I) $\vartheta_k \in (0, 1/(2\nu)]$ and the sequence $\{\vartheta_k m_k(m_k+1)\}$ is nondecreasing; (II) $\mathbb{E}(\|\boldsymbol{\delta}_t\|^2) \leq \varsigma^2$ for some $\varsigma > 0$, where $\boldsymbol{\delta}_t$ is defined in Lemma 5.3. Then, under condition (1.9), for any $u \in \Omega$, it has*

$$\begin{aligned} &\mathbb{E}\big[\theta(u_T) - \theta(u) + \langle u_T - u, \mathcal{J}(u)\rangle\big] \qquad (5.15)\\ &\leq \frac{1}{2(1+T)}\left\{\varsigma^2 \sum_{k=\kappa}^{\kappa+T} \vartheta_k m_k + \frac{4}{m_\kappa(m_\kappa+1)\vartheta_\kappa}\|x - \breve{x}^\kappa\|^2 + \|u - u^\kappa\|_H^2\right\}. \end{aligned}$$

Proof. Summing the inequality (5.14) over k between κ and $\kappa+T$, using the convexity of θ and the definition of u_T, we can obtain

$$\theta(u_T) - \theta(u) + \langle u_T - u, \mathcal{J}(u)\rangle \leq \frac{1}{1+T}\left\{\frac{1}{2}\|u - u^\kappa\|_H^2 - \sum_{k=\kappa}^{\kappa+T} \zeta^k\right\}. \quad (5.16)$$

By assumption (I), the sequence $\{\vartheta_k m_k(m_k+1)\}$ is nondecreasing for $k \in [\kappa, \kappa+T]$, which implies

$$\sum_{k=\kappa}^{\kappa+T} \frac{1}{m_k(m_k+1)\vartheta_k}\left(\|x - \breve{x}^k\|^2 - \|x - \breve{x}^{k+1}\|^2\right) \leq \frac{\|x - \breve{x}^\kappa\|^2}{m_\kappa(m_\kappa+1)\vartheta_\kappa}. \quad (5.17)$$

The definition of $\boldsymbol{\delta}_t$ in Lemma 5.3 gives

$$\boldsymbol{\delta}_t = \nabla f(\widehat{x}_t) - d_t = \nabla f(\widehat{x}_t) - \nabla f_{\xi_t}(\widehat{x}_t) - e_t.$$

Then, because the random variable $\xi_t \in \{1, 2, \ldots, N\}$ is chosen with uniform probability and $\mathbb{E}[e_t] = \mathbf{0}$, it holds that $\mathbb{E}[\boldsymbol{\delta}_t] = \mathbf{0}$. Thus, since $\boldsymbol{\delta}_t$ only depends on the index ξ_t while $\breve{x}_t$ depends on ξ_{t-1}, ξ_{t-2}, $\ldots$, we have $\mathbb{E}\left[\langle \boldsymbol{\delta}_t, \breve{x}_t - x\rangle\right] = 0$. Then, it follows from $\mathbb{E}(\|\boldsymbol{\delta}_t\|^2) \leq \varsigma^2$ from assumption (II) and $m_k \geq 1$ that

$$\mathbb{E}\left[\sum_{t=1}^{m_k} t^2\|\boldsymbol{\delta}_t\|^2\right] \leq \frac{\varsigma^2 m_k(m_k+1)(2m_k+1)}{6} \leq m_k^2(m_k+1)\left(\frac{\varsigma^2}{2}\right).$$

So, by ζ^k defined in (5.10) and the condition $\vartheta_k \leq 1/(2\nu)$, we have

$$-\mathbb{E}\left[\sum_{k=\kappa}^{\kappa+T} \zeta^k\right] \leq \frac{2\|x - \breve{x}^\kappa\|^2}{m_\kappa(m_\kappa + 1)\vartheta_\kappa} + \frac{\varsigma^2}{2}\sum_{k=\kappa}^{\kappa+T} \vartheta_k m_k.$$

Applying the expectation operator to (5.16) together with this bound completes the proof. □

Theorem 5.3 *Suppose the conditions in Theorem 5.2 hold. Let*

$$\vartheta_k = \min\left\{\frac{c_1}{m_k(m_k+1)}, c_2\right\} \quad \text{and} \quad m_k = \max\left\{\lceil c_3 k^\varrho \rceil, m\right\},$$

where $c_1, c_2, c_3 > 0$, $\varrho \geq 1$ are constants and $m > 0$ is a given integer. Then, for every $u^* = (x^*, y^*) \in \Omega^*$ and $u_T = (x_T, y_T)$ being defined in (3.12), we have

$$\left|\mathbb{E}\left[\mathcal{L}(x_T, y^*) - \mathcal{L}(x^*, y_T)\right]\right| = \left|\mathbb{E}\left[\theta(u_T) - \theta(u^*)\right]\right| = E_\varrho(T), \tag{5.18}$$

where $E_\varrho(T) = \mathcal{O}(1/T)$ for $\varrho > 1$ and $E_\varrho(T) = \mathcal{O}(T^{-1}\log T)$ for $\varrho = 1$.

Proof. The proof is same as that of [4, Theorem 4.2] and thus is omitted here. □

Notice that, when considering the sample size $N = 1$ and setting $e_t = 0$, SG-AFBA will reduce to a deterministic algorithm to solve (1.1), while applying the subroutine **xsub** to solve the prediction step (3.1a) inexactly. This inexact G-AFBA will be particularly useful when the function f is not simple so that it is expensive or there is no closed-form solution for calculating the prediction step (3.1a) exactly.

6 Numerical experiments

6.1 Robust principal component analysis

The robust principal component analysis problem, which arises from video surveillance and face recognition [5, 8, 31, 41, 49] etc., aims at recovering the low-rank and sparse components of a given matrix. Such a problem is often modeled [13] as

$$\min_{X,Y\in\mathbb{R}^{m\times n}} \left\{\|X\|_* + \lambda\|Y\|_1 \mid X + Y = C\right\}, \tag{6.1}$$

where C is the given data, $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear norm (the sum of all singular values) and the l_1-norm (the sum of absolute values of all entries) of a matrix, respectively, and $\lambda > 0$ is a weight parameter. Clearly, (6.1) can be reformulated as the following saddle-point problem

$$\min_{X,Y\in\mathbb{R}^{m\times n}} \max_{Z\in\mathbb{R}^{m\times n}} \|X\|_* + \lambda\|Y\|_1 + \langle X + Y, Z\rangle - \langle C, Z\rangle. \tag{6.2}$$

We will test the proposed G-AFBA in (1.8), aG-AFBA in Algorithm 5.1 and G1-AFBA (that is G-AFBA with $\alpha = 0$ as shown in (4.3)) with other comparison algorithms by solving (6.2) with $\lambda = 1/\sqrt{\max(m,n)}$ as suggested in [8] and four real data sets: Hall airport video containing 300 144×176 frames, ShoppingMall video containing 350 256×320 frames, Bootstrap video containing 200 120×160 frames, and Lobby video containing 200 128×160 frames. We use $(\alpha, \mu) = (1/3, 1/2)$ as default values for G-AFBA, $(\gamma_1, \gamma_2) = (1.5, 0.96)$ for aG-AFBA, $(\alpha, \mu) = (0, 1/2)$ for

G1-AFBA and we choose $(\tau, \sigma) = (c_1/\sqrt{\iota}, c_2/\sqrt{\iota})$ to satisfy the condition (1.9), where $c_1, c_2 > 0$ are some constants satisfying $c_1 c_2 < 1$ and

$$\iota = \frac{\alpha + (1-\mu+\mu^2)(1-\alpha)^2 + \sqrt{[\alpha-(1-\mu+\mu^2)(1-\alpha)^2]^2 + 4\alpha(1-\alpha)^2}}{2} L$$

with $L = 2$. After tuning the parameters through the for loop (similar technique is used in the comparative methods), we set $c_1 = 0.2$ and $c_2 = 0.95/c_1$ for G-AFBA, aG-AFBA and G1-AFBA, respectively, for this set of testing problems. The following are several comparison algorithms where the parameters are also tuned and chosen to obtain the best possible performance:

- Dual-Primal Balanced ALM (DP-BALM) with involved parameters $(\beta_1, \beta_2, \alpha, \delta) = (10, 10, 1, 10^{-3})$, which is suggested in [45, Section 5.2.2];
- Generalized PDHG (G-PDHG) with $(\tau, \sigma) = (c_1/\sqrt{0.75L}, c_2/\sqrt{0.75L})$, where parameters (c_1, c_2) use the same setting as our G-AFBA to satisfy the condition $\frac{1}{\tau\sigma} > 0.75L$, which gives much better performance than the original setting given in [31, Section 5.4];
- PDHG (1.2) with $(\tau, \sigma) = (c_1/\sqrt{L}, c_2/\sqrt{L})$ and $(c_1, c_2) = (7.0711, 0.1245)$;
- aPDHG with the tuned $(\gamma_1, \gamma_2) = (8, 2)$ and the same (τ, σ) used in PDHG as the initial values, since these values give better performance than the suggested setting in [20];
- GCP-PPA (1.6) [24] with $(\alpha, \mu) = (1/2, 0)$ and the same (c_1, c_2) as those for G-AFBA, to satisfy the convergence condition (1.7).
- Extended G-AFBA (eG-AFBA) [46] with parameters $(c_1, c_2) = (0.9899, 0.2121)$ to satisfy the involved condition $\frac{1}{\tau\sigma} > L/4$.

All experiments are implemented in MATLAB R2018a and performed on a PC with Windows 10 operating system, with an Intel i7-8700K CPU and 16GB RAM. All algorithms start with initial iteration $(X, Y, Z) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ and are terminated when $\max\{\text{PrimalError}(k), \text{DualError}(k)\} < 10^{-4}$ is satisfied, where

$$\begin{aligned} \text{PrimalError}(k) &:= \frac{\|X^{k+1} - X^k\|_F + \|Y^{k+1} - Y^k\|_F}{\tau_k(\|X^k\|_F + \|Y^k\|_F + 1)} \quad \text{and} \\ \text{DualError}(k) &:= \frac{\|X^{k+1} + Y^{k+1} - C\|_F}{\|C\|_F}. \end{aligned}$$

Here, $\tau_k > 0$ is the primal stepsize used at the k-th iteration by each comparison method. Similar stopping criterions can be found in e.g. [31, 41, 49].

Table 2 reports the number of iterations (Iter), the computing time in seconds (Time(s)), the PrimalError and DualError at the last iterate of the algorithms. Figure 2 also visualizes the background and foreground separations of the 10th frames of Hall airport, the 259th frames of ShoppingMall, the 194th frames of Bootstrap, and the 80th frames of Lobby, respectively. The results obtained by eG-AFBA, DP-BALM and PDHG are not showed since they take significantly more iterations and CPU time than others. The computing results of Table 2 demonstrate that aG-AFBA performs the best among all the comparison algorithms in terms of CPU time and the iteration number; G-AFBA is slightly better than its special case G1-AFBA and usually better than other comparison algorithms. Although there are more relaxed stepsize requirements of eG-AFBA for ensuring convergence, eG-AFBA

Data	Methods	Iter	Time(s)	PrimalError	DualError
Hall airport	G-AFBA	101	50.91	9.91e-5	5.69e-5
	aG-AFBA	**78**	**39.56**	9.45e-5	6.93e-5
	G1-AFBA	104	52.96	9.99e-5	5.50e-5
	eG-AFBA	189	110.46	9.77e-5	9.94e-5
	GCP-PPA	120	55.84	9.82e-5	3.51e-5
	DP-BALM	267	149.03	9.97e-5	7.13e-6
	PDHG	170	99.67	9.98e-5	3.14e-6
	a-PDHG	109	73.34	9.88e-5	7.85e-5
	G-PDHG	121	58.70	9.95e-5	3.58e-5
ShoppingMall	G-AFBA	120	283.20	9.79e-5	9.70e-5
	aG-AFBA	**101**	**225.36**	9.72e-5	9.80e-5
	G1-AFBA	124	289.97	9.96e-5	9.35e-5
	eG-AFBA	275	754.83	7.56e-5	9.92e-5
	GCP-PPA	131	298.28	9.76e-5	8.07e-5
	DP-BALM	173	445.26	9.99e-5	1.34e-5
	PDHG	146	322.40	9.78e-5	2.60e-5
	a-PDHG	112	290.81	8.44e-5	9.98e-5
	G-PDHG	133	304.64	9.76e-5	8.06e-5
Bootstrap	G-AFBA	101	22.49	9.89e-5	2.87e-5
	aG-AFBA	**91**	**20.89**	9.79e-5	9.59e-5
	G1-AFBA	104	23.59	9.92e-5	2.76e-5
	eG-AFBA	171	44.86	9.97e-5	9.27e-5
	GCP-PPA	119	24.19	9.89e-5	1.67e-5
	DP-BALM	296	68.03	9.97e-5	7.03e-6
	PDHG	181	42.99	9.94e-5	2.41e-6
	a-PDHG	166	35.62	9.94e-5	7.07e-5
	G-PDHG	120	24.17	9.98e-5	1.68e-5
Lobby	G-AFBA	103	26.61	9.90e-5	1.32e-5
	aG-AFBA	**91**	**23.64**	6.89e-5	9.82e-5
	G1-AFBA	107	27.47	9.97e-5	1.37e-5
	eG-AFBA	188	55.02	9.35e-5	9.93e-5
	GCP-PPA	129	29.74	9.83e-5	6.67e-6
	DP-BALM	359	93.01	9.95e-5	7.96e-6
	PDHG	213	47.76	9.99e-5	2.57e-6
	a-PDHG	152	42.39	9.81e-5	3.17e-5
	G-PDHG	130	30.12	9.87e-5	6.53e-6

Table 2: Numerical results of different algorithms for solving Problem (6.2).

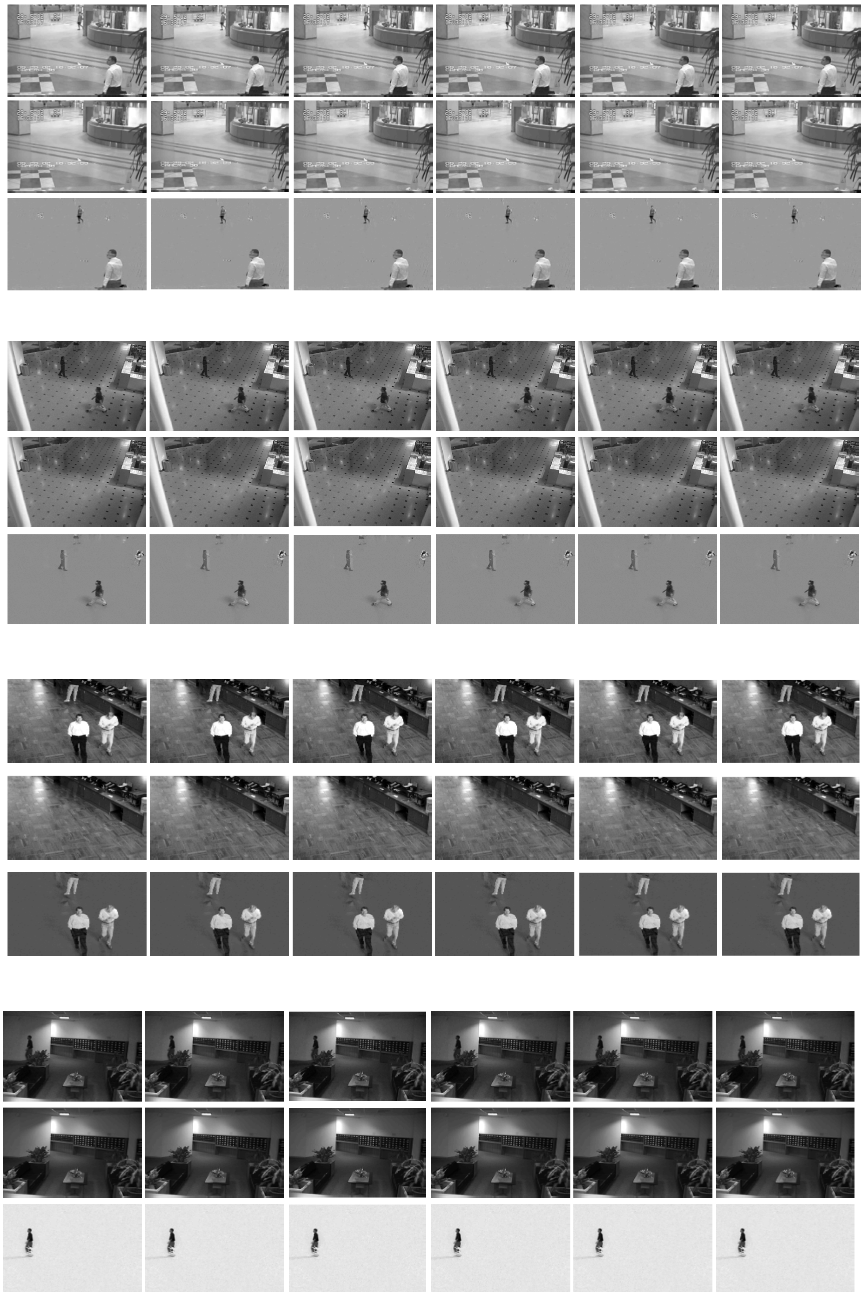

Figure 2: Background and foreground separations of the 10th frame(rows 1-3) of Hall airport, the 259th frame(rows 4-6) of ShoppingMall, the 194th frame(rows 7-9) of Bootstrap, and the 80th frame(rows 10-12) of Lobby. From left to right: G-AFBA, G1-AFBA, GCP-PPA, G-PDHG, a-PDHG aG-AFBA, respectively.

seems to take more iteration numbers and CPU time. We think this may be due to the different strategies used by the correction step of eG-AFBA that requires inversion of a matrix. Besides, the two adaptive methods (a-PDHG and aG-AFBA) clearly improve the performance of its original version, which verifies the effectiveness of adaptively adjusting the proximal stepsizes. Figure 3 depicts the convergence curves of PrimalError(k) and ItError(k) $:= \|u^k - u^*\|/(\|u^*\| + 1)$ obtained by G-AFBA on the four data sets, where $u^* = (X^*, Y^*, Z^*)$ is the approximate solution obtained by running G-AFBA after 300 iterations, which demonstrates the convergence rates in Theorem 3.3 and (3.24), respectively.

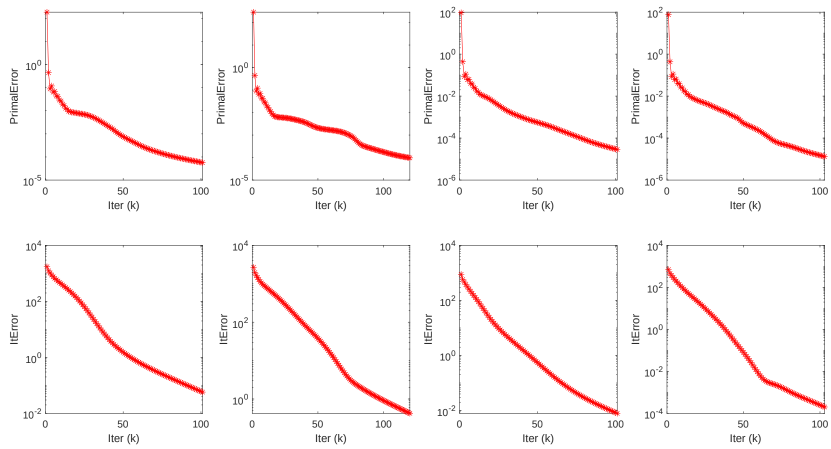

Figure 3: Convergence curves of PrimalError(k) and $\|u^k - u^*\|$ obtained by G-AFBA.

6.2 3D CT reconstruction problem

The 3D CT reconstruction problem is a crucial problem in medical imaging and plays a vital role in diagnosis, treatment planning, and research [7, 19]. The problem with TV-L_1 regularization is formulated as the following

$$\begin{array}{ll}\min\limits_{x,y} \frac{1}{N}\sum_{j=1}^{N}(\mathcal{R}_j x - b_j)^2 + \lambda\|y\|_1 \\ \text{s.t.} \quad\quad \nabla x = y,\end{array} \tag{6.3}$$

where $\lambda > 0$ is a weight parameter, $\mathcal{R}$ is the Radon transform generated by the cone beam scanning geometry [19], b is the observed noisy input data, and ∇ is a discrete gradient operator. The primal-dual formulation of (6.3), as a special case of (5.5), can be written as

$$\min_{x,y}\max_{z} \frac{1}{N}\sum_{j=1}^{N}(\mathcal{R}_j x - b_j)^2 + \lambda\|y\|_1 + \langle \nabla x, z\rangle - \langle y, z\rangle. \tag{6.4}$$

When N is sufficiently large, e.g. $N = 131,334,144$ in our numerical experiment, the computation of the prediction step (3.1a) of applying G-AFBA to solve (6.4) becomes prohibitively expensive. Hence, we would apply the stochastic gradient based SG-AFBA, that is Algorithm 5.2, to solve (6.4) with $\lambda = 0.1$. We set $(\alpha, \mu) = (1/2, 0)$, $(\tau, \sigma) = (10^2, 10^{-7})$ and $m_k = 10$ for SG-AFBA. Hence, in this case, SG-AFBA is in fact a stochastic version of GCP-PPA. The reconstructed image quality is usually evaluated by the Peak Signal-to-Noise Ratio (PSNR):

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{d_x \times d_y \times d_z}{\mathrm{MSE}}\right) \quad \text{with} \quad \mathrm{MSE} = \|x - \widetilde{x}\|^2,$$

where x and $\widetilde{x}$ are the original and reconstructed 3D images, respectively. We also denote the relative error by $\mathrm{Res} = \|x - \widetilde{x}\|/\|x\|$.

For comparison purpose, we solve the reformulation problem (6.4) by the deterministic Generalized ADMM (G-ADMM, [17]) and 5 stochastic gradient-based methods: stochastic ADMM (sto-ADMM, [27]), stochastic ADMM based on the popular SARAH gradient estimator (called SARAH-ADMM, [7]) and the SAGA gradient estimator (called SAGA-ADMM, [7]), PDHG (1.2) and CP-PPA (1.4). All experiments are run in MATLAB R2019a on a high-performance computational cluster with a Tesla V100 GPU and 192GB memory. For each algorithm, we run 3 times to solve (6.4) with a 2000-second time budget for each run.

Methods	PSNR	Res
sto-ADMM	24.8068 ± 0.0013	0.4099 ± 6.29e-05
G-ADMM	24.8493 ± 0.0059	0.4079 ± 2.79e-04
SARAH-ADMM	24.9106 ± 0.0041	0.4051 ± 1.93e-04
SAGA-ADMM	24.8810 ± 0.0017	0.4064 ± 7.72e-05
PDHG	25.0356 ± 0.0396	0.3993 ± 1.82e-03
CP-PPA	24.9976 ± 0.0719	0.4010 ± 3.32e-03
SG-AFBA	**25.1245** ± 0.1256	**0.3952** ± 5.74e-03

Table 3: The mean and standard deviation of PSNR and Res from solving (6.3).

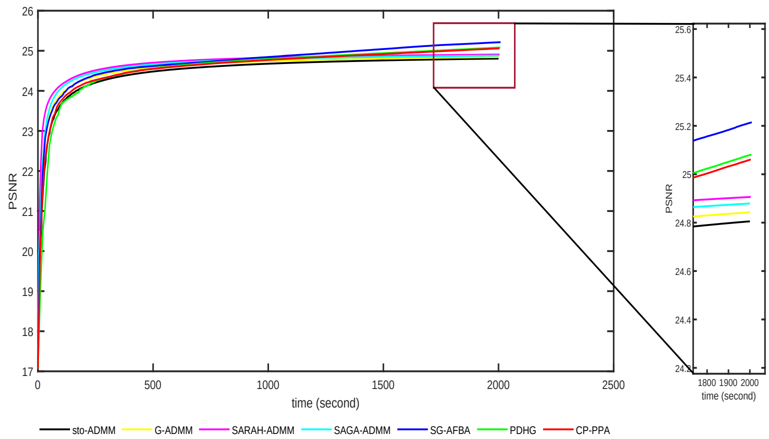

Figure 4: Comparison of different algorithms for solving (6.3).

Table 3 shows the mean and standard deviation of the final PSNR and Res obtained by each algorithm over 3 independent runs. We can see from Table 3 that SG-AFBA has overall better performance, achieving the highest PSNR and the lowest relative error Res, although it has relatively larger standard deviation on the PSNR value. In addition, both PDHG and CP-PPA perform better than other ADMM-type methods from the final obtained PSNR. Figure 4 shows the average convergence curve of PSNR of each algorithm within 2000 seconds. From Figure 4 we see that although SARAH-ADMM converges faster than other algorithms at the beginning iterations (see the left-hand-side of Figure 4), SG-AFBA seems to generate the best final result. Figures 5 and 6 visualize the 7th and 58th slices of the reconstructed 3D CT image, respectively. It shows that the images reconstructed by SG-AFBA are closer to the ground truth compared to other algorithms. Taking the 7th slice of the reconstructed 3D CT image as an example, many blurry circle contours can be observed in the images reconstructed by comparative algorithms sto-ADMM, SAGA-ADMM, SARAH-ADMM and G-ADMM. However, these circular contours are not clear in the images reconstructed by our SG-AFBA. Similar observations can be also seen from the 58th slice.

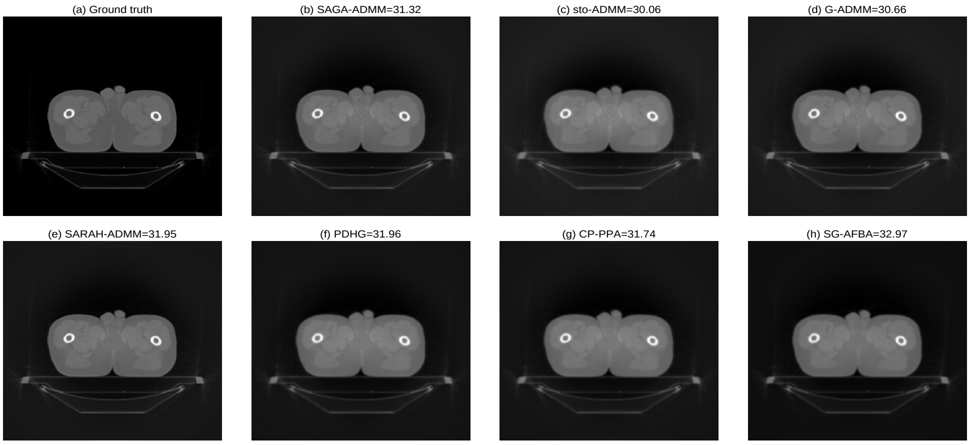

Figure 5: Final reconstruction images of different methods for the **7th** slice.

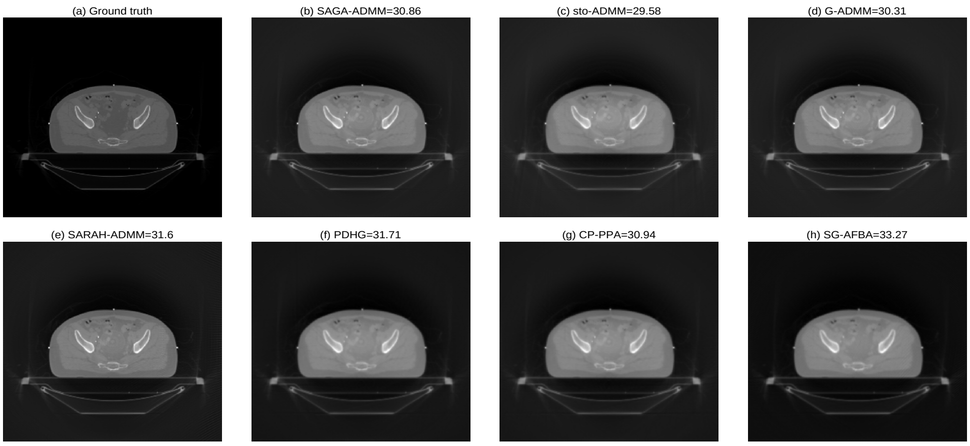

Figure 6: Final reconstruction images of different methods for the **58th** slice.

Acknowledgements The authors would like to thank the anonymous referees for providing very constructive comments, which have significantly improved the quality of the paper.

Funding This research was supported by the National Natural Science Foundation of China (12471298, 12171479), the Shaanxi Fundamental Science Research Project for Mathematics and Physics (23JSQ031), and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001).

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest The authors have not disclosed any competing interests.

References

[1] K. Arrow, L. Hurwicz, H. Uzawa, *Studies in Linear and Non-linear Programming*, Stanford Mathematical Studies in the Social Sciences, vol. II, Stanford University Press, Stanford (1958).

[2] S. Banert, M. Upadhyaya, P. Giselsson, *The Chambolle-Pock method converges weakly with $\theta > 1/2$ and $\tau\sigma\|L\|^2 < 4/(1+2\theta)$*, arXiv:2309.03998v1, (2023).

[3] J. Bai, X. Chang, J. Li, F. Xu, *Convergence revisit on generalized symmetric ADMM*, Optimization, 70 (2021), pp. 149–168.

[4] J. Bai, W. Hager, H. Zhang, *An inexact accelerated stochastic ADMM for separable convex optimization*, Comput. Optim. Appl., 81 (2022), pp. 479–518.

[5] J. Bai, L, Jia, Z, Peng, *A new insight on augmented Lagrangian method with applications in machine learning*, J. Sci. Comput., 99 (2024), 53.

[6] J. Bai, F. Bian, X. Chang, L. Du, *Accelerated stochastic Peaceman-Rachford method for empirical risk minimization*, J. Oper. Res. Soc. China, 11 (2023), pp. 783–807.

[7] F. Bian, J. Liang, X. Zhang, *A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization*, Inverse Problems, 37 (2021), 075009.

[8] E. Candes, X. Li, Y. Ma, J. Wright, *Robust principal component analysis?* J. ACM, 58 (2011), pp. 1–37.

[9] A. Chambolle, T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[10] A. Chambolle, T. Pock, *On the ergodic convergence rates of a first-order primal-dual algorithm*, Math. Program., 159 (2016), pp. 253–287.

[11] A. Chambolle, M. Ehrhardt, P. Richtarik, C. Schonlieb, *Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications*, SIAM J. Optim., 28 (2018), pp. 2783–2808.

[12] X. Chang, J. Yang, H. Zhang, *Golden ratio primal-dual algorithm with linesearch*, SIAM J. Optim., 32 (2022), pp. 1584–1613.

[13] V. Chandrasekaran, S. Sanghavi, P. Parrilo, A. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM J. Optim., 21 (2011), pp. 572–596.

[14] L. Condat, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.

[15] L. Condat, D. Kitahara, A. Contreras, A. Hirabayashi, *proximal splitting algorithms for convex optimization: a tour of recent advances, with new twists*, SIAM Rev., 65 (2023), pp. 375–435.

[16] W. Deng, W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*, J. Sci. Comput., 66 (2016), pp. 889–916.

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

[18] D. Gabay, B. Mercier, *A dual algorithm for the solution of nonlinear variational problems viafinite-element approximations*, Comput. Math. Appl., 2 (1976), pp. 17–40.

[19] H. Gao, *Fast parallel algorithms for the x-ray transform and its adjoint*, Med. Phys., 39 (2012), pp. 7110–7120.

[20] T. Goldstein, M. Li, X. Yuan, *Adaptive primal-dual splitting methods for statistical learning and image processing*, NeurIPS, (2015), pp. 2089–2097.

[21] S. Hayden, O. Stanley, *A low patch-rank interpretation of texture*, SIAM J. Imaging Sci., 6 (2013), pp. 226–262.

[22] B. He, X. Yuan, *On the $o(1/n)$ convergence rate of the douglas-rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.

[23] B. He, Y. You, X. Yuan, *On the convergence of primal-dual hybrid gradient algorithm*, SIAM J. Imaging Sci., 7 (2014), pp. 2526–2537.

[24] B. He, F. Ma, S. Xu, X. Yuan, *A generalized primal-dual algorithm with improved convergence condition for saddle point problems*, SIAM J. Imaging Sci., 15 (2022), p-p. 1157–1183.

[25] B. He, S. Xu, X. Yuan, *On convergence of the Arrow-Hurwicz method for saddle point problems*, J. Math. Imaging Vision, 64 (2022), pp. 662–671.

[26] X. He, N. Huang, Y. Fang, *Non-ergodic convergence rate of an inertial accelerated primalCdual algorithm for saddle point problems*, Commun. Nonlinear Sci. Numer. Simulat., 140 (2024), 108289.

[27] F. Huang, S. Chen, *Mini-batch stochastic ADMMs for nonconvex nonsmooth optimization*, arXiv: 1802.03284, (2019).

[28] F. Jiang, X. Cai, Z. Wu, D. Han, *Approximate first-order primal-dual algorithms for saddle point problems*, Math. Comput., 90 (2021), pp. 1227–1262.

[29] F. Jiang, Z. Wu, X. Cai, H. Zhang, *A first-order inexact primal-dual algorithm for a class of convex-concave saddle point problems*, Numer. Algor., 88 (2021), pp. 1109–1136.

[30] F. Jiang, X. Cai, D. Han, *Inexact asymmetric forward-backward-adjoint splitting algorithms for saddle point problems*, Numer. Algor., 94 (2023), pp. 479–509.

[31] F. Jiang, Z. Zhang, H. He, *Solving saddle point problems: a landscape of primal-dual algorithm with larger stepsizes*, J. Global Optim., 85 (2023), pp. 821–846.

[32] G. Korpelevič, *An extragradient method for finding saddle points and for other problems*, Èkon. Mat. Metody, 12 (1976), pp. 747–756.

[33] P. Latafat, P. Patrinos, *Asymmetric forward-backward-adjoint splitting for solving monotone inclusions involving three operators*, Comput. Optim. Appl., 68 (2017), pp. 57–93.

[34] Z. Li, M. Yan, *New convergence analysis of a primal-dual algorithm with large stepsizes*, Adv. Comput. Math., 47 (2021), pp. 1-20.

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

[36] D. O'Connor, L. Vandenberghe, *On the equivalence of the primal-dual hybrid gradient method and Douglas-Rachford splitting*, Math. Program., 179 (2020), pp. 85–108.

[37] B. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phy., 4 (1964), pp. 1–17.

[38] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, (2015).

[39] S. Robinson, *Some Continuity Properties of Polyhedral Multifunctions*, Mathematical Programming at Oberwolfach, Springer, Berlin, (1981), pp. 206–214.

[40] H. Sun, X. Tai, J. Yuan, *Efficient and convergent preconditioned ADMM for the potts models*, SIAM J. Sci. Comput. 43 (2021), pp. B455-B478.

[41] M. Tao, X. Yuan, *Recovering Low-Rank and Sparse Components of Matrices from Incomplete and Noisy Observations*, SIAM J. Optim. 21 (2011), pp. 57–81.

[42] B. Vũ, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Adv. Comput. Math., 38 (2013), pp. 667–681.

[43] N. Wang, J. Li, *A class of preconditioners based on symmetric-triangular decomposition and matrix splitting for generalized saddle point problems*, IMA J. Numer. Anal., 43 (2023), pp. 2998–3025.

[44] W. Xian, F. Huang, Y. Zhang, H. Huang. *A faster decentralized algorithm for nonconvex minimax problems*, NeurIPS 2021, https://openreview.net/forum?id=rjIjkiyAJao.

[45] S. Xu, *A dual-primal balanced augmented Lagrangian method for linearly constrained convex programming*, J. Appl. Math. Comput. 69 (2023), pp. 1015–1035.

[46] S. Xu. *A search direction inspired primal-dual method for saddle point problems*, Optimization Online, (2020), https://optimization-online.org/2019/11/7491/.

[47] Z. Xu, H. Zhang, Y. Xu, G. Lan, *A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems*, Math. Program., 201 (2023), pp. 635–706.

[48] J. Yang, Y. Zhang, *Alternating direction algorithms for ℓ_1-problems in compressive sensing*, SIAM J. Sci. Comput., 33 (2011), pp. 250–278.

[49] L. Yang, T. Pong, X. Chen, *Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction*, SIAM J. Imaging Sci., 10 (2017), pp. 74–110.

[50] X. Zhang, M. Burger, S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, J. Sci. Comput., 46 (2011), pp. 20–46.

[51] M. Zhu, T. Chan, *An efficient primal-dual hybrid gradient algorithm for total variation image restoration*, CAM Report 08-34, UCLA, Los Angeles, CA, (2008).

[52] Y. Zhu, D. Liu, Q. Dinh, *New primal-dual algorithms for a class of nonsmooth and nonlinear convex-concave minimax problems*, SIAM J, Optim., 32 (2022), pp. 2580–2611.