Numerical Algorithms
https://doi.org/10.1007/s11075-021-01069-x

ORIGINAL PAPER

A first-order inexact primal-dual algorithm for a class of convex-concave saddle point problems

Fan Jiang[1] · Zhongming Wu[2] · Xingju Cai[1] · Hongchao Zhang[3]

Received: 17 January 2020 / Accepted: 5 January 2021 / Published online: 16 March 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

In this paper, we study a first-order inexact primal-dual algorithm (I-PDA) for solving a class of convex-concave saddle point problems. The I-PDA, which involves a relative error criterion and generalizes the classical PDA, has the advantage of solving one subproblem inexactly when it does not have a closed-form solution. We show that the whole sequence generated by I-PDA converges to a saddle point solution with $\mathcal{O}(1/N)$ ergodic convergence rate, where N is the iteration number. In addition, under a mild calmness condition, we establish the global Q-linear convergence rate of the distance between the iterates generated by I-PDA and the solution set, and the R-linear convergence speed of the nonergodic iterates. Furthermore, we demonstrate that many problems arising from practical applications satisfy this calmness condition. Finally, some numerical experiments are performed to show the superiority and linear convergence behaviors of I-PDA.

Keywords Convex optimization · Saddle point problems · First-order primal-dual algorithm · Inexact · Nonergodic convergence · Linear convergence

✉ Hongchao Zhang
hozhang@math.lsu.edu

Fan Jiang
15905154902@163.com

Zhongming Wu
wuzm@nuist.edu.cn

Xingju Cai
caixingju@njnu.edu.cn

[1] School of Mathematical Sciences, Jiangsu Key Laboratory for NSLSCS, Nanjing Normal University, Nanjing 210023, China

[2] School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

[3] Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA

Mathematics subject classification (2010) 90C25 · 90C47 · 65K15

1 Introduction

In this paper, we propose a first-order inexact primal-dual algorithm (I-PDA) for solving the following saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\ \mathcal{L}(x,y) := f(x) + \langle Kx, y\rangle - g(y), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are two finite-dimensional real vector spaces endowed with the inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$, $K : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator with the operator norm $\|K\| = L$, $f : \mathcal{X} \to (-\infty,\infty]$ and $g : \mathcal{Y} \to (-\infty,\infty]$ are proper lower semicontinuous (l.s.c.) convex functions. The convex-concave saddle point problem (1) often arises from a wide range of applications such as finding a saddle point of the Lagrangian function for a convex optimization with linear constraints, image processing, and machine learning problems, see, e.g., [3, 4, 15, 18, 37]. Besides, it is well known that (1) is equivalent to the primal and dual problem:

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

where f^* and g^* are the Fenchel conjugate [32] of the functions f and g, respectively. Hence, problem (1) or its equivalent forms have been widely studied in the literature, see, e.g., [7–10, 26, 35].

The classical PDA for solving problem (1), which was designed by Chambolle and Pock [4] and He and Yuan [18], can be read as:

$$\begin{cases} x^{k+1} = \arg\min\left\{ f(x) + \langle Kx, y^k\rangle + \dfrac{1}{2\tau}\|x - x^k\|^2 \,\middle|\, x\in\mathcal{X}\right\}, & \text{(2a)}\\ \tilde{x}^{k+1} = x^{k+1} + \gamma(x^{k+1} - x^k), & \text{(2b)}\\ y^{k+1} = \arg\min\left\{ g(y) - \langle K\tilde{x}^{k+1}, y\rangle + \dfrac{1}{2\sigma}\|y - y^k\|^2 \,\middle|\, y\in\mathcal{Y}\right\}, & \text{(2c)} \end{cases}$$

where $\tau,\sigma > 0$ play the role of step sizes in the subproblems (2a) and (2c) respectively, and $\gamma \in [0,1]$ is a parameter. This scheme was mainly motivated by the classical Arrow-Hurwicz method [1] and the primal-dual hybrid gradient method [37] which is the special case of (2) with $\gamma = 0$. The convergence of PDA (2) has been well studied in [4, 5, 18]. Since then, many variants of PDA have been developed, such as extending the value range of γ [3, 18, 19], finding the suitable step sizes by line search strategy [21], solving the subproblems inexactly [20, 27], and solving the subproblems stochastically when the dual variable is separable [6]. In addition, there are some papers focusing on nonconvex settings [22, 33].

When the proximal operators of f and g are easy to compute, PDA (2) is efficient. However, when applying PDA (2) to solve some problems in practical applications, such as the ℓ_1 regularized sparse recovery problem [24, 34, 36] and the constrained TV-ℓ_2 image restoration problem [16, 23], one of the subproblems in the PDA (2)

usually does not possess closed-form solutions and some inner-iterative methods should be introduced to evaluate the proximal operator [20]. Therefore, for practical use of PDA, it is important to guarantee the effectiveness of PDA with approximate solutions of subproblems in (2) while still ensure global convergence and convergence rate as those of exact PDA. Along this line of research, Rasch and Chambolle [27] introduced four types of approximation for computing the proximal operator based on certain absolute error condition. Instead of solving the subproblems directly, they assumed that the dual problems of these subproblems can be solved by some iterative methods to a summable error tolerance. Global convergence and convergence rate of the proposed methods were then analyzed under different combinations of approximate subproblem solutions. Recently, Jiang et al. [20] proposed two types of inexact criteria for PDA, namely the absolute and relative error criterion. The absolute error criterion constructs an absolute summable tolerance sequence before implementing the method, while the relative one involves a single parameter ranging in $[0, 1)$. When these criteria are satisfied, it is shown that any cluster point of the generated iterates will be a solution of (1). However, this convergence result is weaker than that of the standard exact PDA where the whole generated sequence can converge.

In the literature, there are many variants and applications of either exact or inexact versions of PDA. However, we do not see any inexact PDA using relative error criterion theoretically ensures the convergence of the whole iterate sequence as guaranteed by the exact PDA. In addition, only a few works studied the linear convergence rate. It has shown in [4, 5, 18] that when $\gamma = 1$, the primal-dual gap of the ergodic sequence generated by exact PDA (2) enjoys an $\mathcal{O}(1/N)$ convergence rate, where N is the iteration number. Chambolle and Pock [4, 5] showed that when f (or g) is strongly convex, the $\mathcal{O}(1/N^2)$ convergence rates for the nonergodic sequence and primal-dual gap of ergodic sequence can be obtained by dynamic selection of the combination parameter γ at each iteration. Moreover, when both f and g are strongly convex, the R-linear convergence rates for the nonergodic iterates and primal-dual gap of the ergodic iterates can be obtained. Malitsky and Pock [21] showed that the previous convergence rate results can be also maintained under proper line search strategy, except the linear convergence rate. Rasch and Chambolle [27] proved that all the convergence rate results can be achieved by solving the subproblems inexactly under the same strongly convex assumptions on the objective functions. However, there are some drawbacks in the existing linear convergence results. Firstly, the existing linear convergence is mainly based on the strong convexity of the objective function [4, 5, 27], which is not satisfied by many problems in practical applications. Secondly, existing results only establish the R-linear convergence rate [4, 5, 27], which is weaker than the Q-linear convergence rate that we will establish for the inexact PDA (I-PDA) developed in this paper. Thirdly, the current linear convergence rate of inexact PDA with strongly convex assumptions on the objective function is only established under the absolute summable error criterion, while we will show the linear convergence of our I-PDA a relative error criterion under a mild calmness assumption.

In this paper, we propose a new I-PDA which solves one of the subproblems inexactly to an adaptive accuracy relative to the total optimality error of the original

problem. We show that this I-PDA will still maintain the same global convergence and convergence rate of exact PDA although one of the subproblems is solved inexactly. Without loss of generality, we assume that the proximal operator of f possesses a closed-form solution, i.e., the exact solution of the subproblem (2a) can be obtained, while some iterative methods should be applied to compute the proximal operator of g, i.e., the subproblem (2c) can only be solved inexactly to the required adaptive accuracy. Unlike the convergence result in [20], we show the whole iterate sequence generated by I-PDA will converge to a saddle point solution of (1) and the primal-dual function value gap at the ergodic iterates possesses a $\mathcal{O}(1/N)$ convergence rate. Under a mild calmness condition, we further establish the global Q-linear convergence rate for the distance between iterates generated by I-PDA and the solution set, and the R-linear convergence for the nonergodic iterates. Moreover, we show that many practical problems in applications actually satisfy the calmness condition, although the function f or g in the objective function is not strongly convex. Some numerical experiments on these practical problems are also performed to demonstrate the effectiveness and linear convergence rate of I-PDA.

The rest of this paper is organized as follows. In Section 2, we introduce some notations and recall some basic concepts and results. In Section 3, we present the framework of I-PDA with a relative error criterion and analyze its global convergence. Under a mild calmness condition, the Q-linear convergence and R-linear convergence properties of the iterates generated by I-PDA are discussed in Section 4. In Section 5, we provide some practical examples in applications that satisfy the calmness condition. Some numerical experiments are conducted in Section 6 to demonstrate the efficiency and linear convergence rate of I-PDA. Finally, we draw some conclusions in Section 7.

2 Preliminaries

In this section, we summarize some basic concepts that will be useful in the subsequent sections and recall the first-order optimality condition of problem (1). Besides, we formalize the inexact solution of the subproblem.

2.1 Notations and basic concepts

We use $\mathbf{N}$, $\mathbf{R}_+$, and $\mathbf{R}^n$ to denote the set of natural number, nonnegative real number, and n-dimensional Euclidean space, respectively. For a real number c and a set V, cV is defined by $cV := \{cv \mid v \in V\}$. For a function $f : \mathcal{X} \to \mathbf{R} \cup \{\infty\}$, the domain of f is defined by $\mathrm{dom} f := \{x \in \mathcal{X} \mid f(x) < \infty\}$. f is lower semicontinuous (l.s.c.) if $f(x) \le \liminf_{y \to x} f(y)$ and it is proper if $\mathrm{dom} f \neq \emptyset$. The Fenchel conjugate [32] of a function $f : \mathcal{X} \to [-\infty, \infty]$ is denoted by f^*, that is:

$$f^*(v) := \sup_{x \in \mathcal{X}} \{\langle v, x\rangle - f(x)\}.$$

For a proper, convex and l.s.c. function $f : \mathcal{X} \to (-\infty, \infty]$, its subdifferential at x is denoted by $\partial f(x) = \{d \mid f(z) \geq f(x) + \langle z - x, d\rangle,\ \forall\, z \in \mathcal{X}\}$, and for any $y \in \mathcal{X}$ and $\sigma > 0$, its proximal operator [25] $\mathrm{prox}_{\sigma f}$ is given by:

$$\mathrm{prox}_{\sigma f}(y) = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2\sigma}\|x - y\|^2 \right\}.$$

If f is the indicator function δ_C of the closed convex set C, then $\mathrm{prox}_f(\cdot) = \Pi_C(\cdot)$, the projection operator onto the set C. For a linear operator K, its adjoint operator is denoted as K^*. If S is a self-adjoint (not necessarily positive definite) linear operator, we use $\|x\|_S^2$ to denote $\langle x, Sx\rangle$. For a closed convex set $C \subset \mathcal{X}$, we denote $\mathrm{dist}(x, C) = \min_{z \in C}\{\|x - z\|\}$ and $\mathrm{dist}_G(x, C) = \min_{z \in C}\{\|x - z\|_G\}$ when G is a self-adjoint and positive definite linear operator. We also use I to denote the identity operator. For a self-adjoint and positive definite linear operator G, we say a sequence $\{u^k\} \subset \mathcal{U}$ converge to $\hat{u} \in \mathcal{U}$ Q-linearly under G-norm, if there exist a scalar $\xi \in (0, 1)$ and $\bar{k} \in \mathbf{N}$ such that:

$$\|u^{k+1} - \hat{u}\|_G \leq \xi \|u^k - \hat{u}\|_G, \quad \forall\, k \geq \bar{k}.$$

Moreover, if there exists a nonnegative scalar sequence $\{w_k\}$ such that:

$$\|u^k - \hat{u}\|_G \leq w_k,$$

where $\{w_k\}$ converges to zero Q-linearly, we say the sequence $\{u^k\}$ converge to $\hat{u}$ R-linearly under G-norm.

The pair $(\hat{x}, \hat{y})$ defined on $\mathcal{X} \times \mathcal{Y}$ is called a saddle point of problem (1) if it satisfies the following inequalities:

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

Alternatively, we can rewrite these inequalities as:

$$\begin{cases} f(x) - f(\hat{x}) + \langle x - \hat{x}, K^*\hat{y}\rangle \geq 0, & \forall\, x \in \mathcal{X}, \\ g(y) - g(\hat{y}) + \langle y - \hat{y}, -K\hat{x}\rangle \geq 0, & \forall\, y \in \mathcal{Y}. \end{cases} \tag{3}$$

Note that the inequality system (3) on $(\hat{x}, \hat{y})$ can be also reformulated as the following KKT system:

$$\begin{cases} 0 \in \partial f(\hat{x}) + K^*\hat{y}, \\ 0 \in \partial g(\hat{y}) - K\hat{x}. \end{cases} \tag{4}$$

We denote the solution set to the KKT system (4) by $\widehat{\mathcal{U}}$ and assume $\widehat{\mathcal{U}}$ is nonempty in this paper.

Let $\mathcal{U} := \mathcal{X} \times \mathcal{Y}$ and $u := (x, y) \in \mathcal{U}$. For any $u \in \mathcal{U}$, we define the KKT mapping $R : \mathcal{U} \to \mathcal{U}$ as:

$$R(u) := \begin{pmatrix} x - \mathrm{prox}_f(x - K^*y) \\ y - \mathrm{prox}_g(y + Kx) \end{pmatrix}. \tag{5}$$

Since the proximal operator of a proper convex function is Lipschitz continuous with unit Lipschitz constant, the mapping $R(\cdot)$ is continuous on $\mathcal{U}$. Obviously, for any $u \in \mathcal{U}$, we have $u \in \widehat{\mathcal{U}}$ if and only if $R(u) = 0$.

Now we recall the definition of locally upper Lipschitz continuity [29].

Definition 1 Let $B_{\mathcal{Y}}$ be the unit ball in $\mathcal{Y}$. Then, the multivalued mapping $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is locally upper Lipschitz continuous at $x^0 \in \mathcal{X}$ with modulus $\kappa_0 > 0$, if there exists a neighborhood V of x^0 such that:

$$F(x) \subseteq F(x^0) + \kappa_0 \|x - x^0\| B_{\mathcal{Y}}, \quad \forall x \in V.$$

For a multivalued mapping $F : \mathcal{X} \rightrightarrows \mathcal{Y}$, it is said to be piecewise polyhedral, if its graph, denoted as Gph F, is the union of finitely many polyhedral sets. Robinson [30] showed that if F is piecewise polyhedral, then it is locally upper Lipschitz continuous at any $x^0 \in \mathcal{X}$ with modulus κ_0 independent of x^0.

A proper l.s.c. convex function $f : \mathcal{X} \to (-\infty, \infty]$ is called piecewise linear-quadratic if its domain is the union of finitely many polyhedral sets and f is an affine or a quadratic function on each of these polyhedral sets. A piecewise linear mapping is also piecewise polyhedral. Furthermore, we summarize several useful results in the following lemma, whose proof can be found in [31].

Lemma 1 *Let $f : \mathcal{X} \to (-\infty, \infty]$ be a proper l.s.c. convex function. Then f is piecewise linear-quadratic if and only if the graph of ∂f is piecewise polyhedral. f is piecewise linear-quadratic if and only if f^* is piecewise linear-quadratic. Moreover, f is piecewise linear-quadratic function if and only if the proximal mapping of f is piecewise linear.*

The following definition of calmness is given in [11].

Definition 2 Let $(x^0, y^0) \in \text{Gph } F$. The multivalued mapping $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is calm at x^0 for y^0 with modulus $\kappa_0 \geq 0$, if there exists a neighborhood V of x^0 and a neighborhood W of y^0 such that:

$$F(x) \cap W \subseteq F(x^0) + \kappa_0 \|x - x^0\| B_{\mathcal{Y}}, \quad \forall x \in V.$$

If $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is the subdifferential of a convex piecewise linear-quadratic function f, it follows from Lemma 1 that F is piecewise polyhedral. Then, as discussed in [30], we know that F is locally upper Lipschitz continuous at any $x^0 \in \mathcal{X}$ with modulus κ_0 independent of x^0. Furthermore, according to Definitions 1 and 2, we can deduce that for any $(x^0, y^0) \in \text{Gph } F$, F is calm at x^0 for y^0 with modulus $\kappa_0 > 0$ independent of the choice of (x^0, y^0).

2.2 Inexact subproblem solution

We assume that there exists an iterative method $\mathscr{G}$ which can be used to solve the proximal mapping related to the y-subproblem in our I-PDA. Formally, we have the following assumption.

Assumption 1 *Suppose $\mathscr{G}$ is an iterative method having the following properties: for any $\bar{y} \in \mathcal{Y}$ and $\sigma > 0$, $\mathscr{G}$ can generate an infinite sequence $(y^l, e^l) \in \mathcal{Y} \times \mathcal{Y}$, $l = 0, 1, 2, \ldots$, satisfying:*

$$\lim_{l\to\infty} e^l = 0 \quad \textit{and} \quad e^l \in \partial_y \left[g(y) + \frac{1}{2\sigma}\|y - \bar{y}\|^2 \right]_{y=y^l}.$$

Note that Assumption 1 implies there exists an iterative method $\mathscr{G}$ that can be used to solve the y-subproblem in our I-PDA to any required accuracy (more details can be seen in Algorithm 1). Similar assumptions are also used in [13, 14, 20]. However, if the proximal mapping in the y-subproblem of I-PDA has a closed-form solution or can be solved exactly easily, we can regard the subproblem solution is simply given by the first iteration by $\mathscr{G}$, i.e., $y^1 = \mathrm{prox}_{\tau g}(\bar{y})$ and $e^1 = 0$.

Note that the iterates $\{y^l\}$ generated by $\mathscr{G}$ converge to $\mathrm{prox}_{\tau g}(\bar{y})$. In fact, it follows from Assumption 1 that $y^l = \mathrm{prox}_{\tau g}(\bar{y} + \sigma e^l)$. Since the proximal operator of a proper convex l.s.c. function is nonexpansive, we have $\|y^l - \mathrm{prox}_{\tau g}(\bar{y})\| \le \sigma\|e^l\|$. Combining this with $\lim_{l\to\infty} e^l = 0$, we obtain that the sequence $\{y^l\}$ generated by $\mathscr{G}$ converges to $\mathrm{prox}_{\tau g}(\bar{y})$.

3 An inexact primal-dual algorithm

In this section, we first propose our inexact PDA (I-PDA) with a relative-error criterion for solving the y-subproblem. Then, we show the global convergence and give the convergence rate result of the proposed algorithm.

Throughout this paper, we assume the solution set of problem (1) is nonempty and the parameters in Algorithm 1 satisfy $\tau\sigma L^2 < 1$. We first denote the self-adjoint operators $H : \mathcal{Y} \to \mathcal{Y}$ and $G : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$, respectively, as:

$$H := \left(\frac{1}{\sigma} I - \tau K K^*\right)^{-1} \quad \text{and} \quad G := \begin{pmatrix} \frac{1}{\tau} I & -K^* \\ -K & \frac{1}{\sigma} I \end{pmatrix}. \tag{6}$$

Then, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we define $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ as:

$$\varphi(x, y) := \frac{1}{\tau}\|x\|^2 - 2\langle x, K^* y\rangle + \frac{1}{\sigma}\|y\|^2 = \|(x, y)\|_G^2. \tag{7}$$

Since $\tau\sigma L^2 < 1$, H is well defined and positive definite and G defined in (6) is also positive definite. Hence, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, there exist two positive constants β_1 and β_2 such that:

$$\beta_1 \left(\|x\|^2 + \|y\|^2\right) \le \varphi(x, y) \le \beta_2 \left(\|x\|^2 + \|y\|^2\right), \tag{8}$$

where β_1 and β_2 are the smallest and largest eigenvalues of G, respectively. So, we can define a distance function $\mathrm{dist}_G(\cdot, \widehat{\mathcal{U}}) : \mathcal{U} \to \mathcal{R}_+$ such that for any point $u = (x, y) \in \mathcal{U}$, its distance to the set $\widehat{\mathcal{U}}$ is defined as:

$$\mathrm{dist}_G(u, \widehat{\mathcal{U}}) := \min_{(\hat{x},\hat{y})\in\widehat{\mathcal{U}}} \|(x - \hat{x}, y - \hat{y})\|_G = \min_{(\hat{x},\hat{y})\in\widehat{\mathcal{U}}} \sqrt{\varphi(x - \hat{x}, y - \hat{y})}.$$

Now, our I-PDA using a relative error criterion for solving the y-subproblem is given in Algorithm 1.

Algorithm 1 An inexact primal dual algorithm (I-PDA).

0. Initialization: Choose $\eta \in [0, 1)$, $\rho \in (0, 2)$, a sufficient small positive constant ϵ, and $\tau, \sigma > 0$ such that $\tau\sigma L^2 < 1$, and initial points $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$. Set $k = 0$.

1. Solve the x-subproblem: Compute

$$\tilde{x}^k = \mathrm{prox}_{\tau f}(x^k - \tau K^* y^k). \tag{9}$$

2. Solve the y-subproblem inexactly:
Compute $\tilde{y}^k \approx \mathrm{prox}_{\sigma g}(y^k + \sigma K(2\tilde{x}^k - x^k))$ by a method $\mathscr{G}$ such that

$$\|e^k\|_H^2 \le \eta^2 \varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k), \tag{10}$$

$$\text{with} \quad e^k \in \partial g(\tilde{y}^k) - K(2\tilde{x}^k - x^k) + \frac{1}{\sigma}(\tilde{y}^k - y^k). \tag{11}$$

3. Perform correction steps: Compute

$$d_1^k = x^k - \tilde{x}^k + \tau K^* H e^k, \tag{12}$$

$$d_2^k = y^k - \tilde{y}^k + H e^k. \tag{13}$$

If $\varphi(d_1^k, d_2^k) < \epsilon$, **Stop** and output $(\tilde{x}^k, \tilde{y}^k)$; **Otherwise**, compute

$$x^{k+1} = x^k - \rho\alpha_k d_1^k, \tag{14}$$

$$y^{k+1} = y^k - \rho\alpha_k d_2^k, \tag{15}$$

where

$$\alpha_k := \frac{\langle x^k - \tilde{x}^k, \frac{1}{\tau} d_1^k - K^* d_2^k\rangle + \langle y^k - \tilde{y}^k, -K d_1^k + \frac{1}{\sigma} d_2^k\rangle}{\varphi(d_1^k, d_2^k)}, \tag{16}$$

then set $k := k + 1$ and go to step **1**.

For I-PDA, we have the following comments. One observation is that the evaluation of He^k, which involves solving linear system, needs to be calculated at each iteration. When the dimension of $\mathcal{Y}$ is small, one may pre-compute the Cholesky factorization of $I - \tau\sigma KK^*$ and then the evaluation of He^k can be done efficiently by simply performing backward and forward substitution. When the dimension of $\mathcal{X}$ is small, one could pre-compute the Cholesky factorization of $I - \tau\sigma K^*K$ and apply the Sherman-Morrison formula to compute He^k efficiently. On the other hand, when K possesses certain structure, such as the block circulant structure often arising from image processing, the evaluation of He^k could be also done quite efficiently. In the

case of expensive evaluation of He^k, an alternative strategy might be to replace the criterion (10) by:

$$\|e^k\|^2 \leq (\eta^2/\sigma)\lambda_{\min}(I - \sigma\tau KK^*)\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k), \tag{17}$$

where $\lambda_{\min}(\cdot)$ means the minimum eigenvalue of a matrix, and compute d_1^k, d_2^k, and α_k by:

$$d_1^k = \frac{1}{\tau}(x^k - \tilde{x}^k) - K^*(y^k - \tilde{y}^k), \tag{18}$$

$$d_2^k = -K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(y^k - \tilde{y}^k) + e^k, \tag{19}$$

$$\alpha_k = \frac{\langle x^k - \tilde{x}^k, d_1^k\rangle + \langle y^k - \tilde{y}^k, d_2^k\rangle}{\|d_1^k\|^2 + \|d_2^k\|^2}. \tag{20}$$

The criterion (17) is an overestimate of the error and hence, stronger than (10). As a result, similar to Theorems 1 and 2 given in Sections 3 and 4, the global convergence and convergence rates with this modification can be established under 2-norm.

In step 2 of Algorithm 1, the y-subproblem can be solved inexactly by an iterative method $\mathscr{G}$ until criterion (10) is satisfied. Note that the right-hand side of (10) is nonnegative due to the fact $\tau\sigma L^2 < 1$ and (8). We show in the next lemma that the criterion (10) must be satisfied in a finite number of iterations if a method $\mathscr{G}$ satisfying Assumption 1 is applied to solve the y-subproblem in step 2 of Algorithm 1 unless (x^k, y^k) is a solution of (1). The inexact criterion (10) is different from that one used in [20] where an additional variable is involved for collecting the relative error. Also note that two additional correction steps in (14) and (15) are used for the purpose of establishing global convergence of the Algorithm 1. Moreover, if we set $\eta = 0$ and $\rho = 1$, Algorithm 1 would reduce to the classical PDA (2) with $\gamma = 1$. We can also see that Algorithm 1 stops when $\varphi(d_1^k, d_2^k)$ is sufficiently small, that is $\varphi(d_1^k, d_2^k) < \epsilon$ for small positive ϵ. Hence, the stepsize α_k given by (16) is well-defined when the algorithm does not stop. We will show in Corollary 1 that $(\tilde{x}^k, \tilde{y}^k)$ is in fact a solution of (1) if $\varphi(d_1^k, d_2^k) = 0$.

Now, for solving the y-subproblem inexactly in step 2 of Algorithm 1, we have the following lemma.

Lemma 2 *Suppose an iterative method $\mathscr{G}$ satisfying Assumption 1 is applied to solve the y-subproblem in step 2 of Algorithm 1, that is, at the kth iteration of Algorithm 1, $\mathscr{G}$ can generate an infinite sequence $(y^{k,l}, e^{k,l}) \in \mathcal{Y} \times \mathcal{Y}$, $l = 0, 1, 2, \ldots$, satisfying:*

$$\lim_{l\to\infty} e^{k,l} = 0, \quad \text{and} \quad e^{k,l} \in \partial_y \left[g(y) + \frac{1}{2\sigma}\|y - \bar{y}\|^2 \right]_{y=y^{k,l}}, \tag{21}$$

where $\bar{y} = y^k + \sigma K(2\tilde{x}^k - x^k)$. If (x^k, y^k) is not a solution of (1), for sufficiently large l we have:

$$\|e^{k,l}\|_H^2 \leq \eta^2\varphi(x^k - \tilde{x}^k, y^k - y^{k,l}), \tag{22}$$

where η is any constant in $[0, 1)$. Hence, setting $\tilde{y}^k = y^{k,l}$ with l sufficiently large, the criterion (10) will be satisfied.

Proof Suppose condition (22) is not satisfied for all l. Then, by (21), we must have:

$$\lim_{l\to\infty} \varphi(x^k - \tilde{x}^k, y^k - y^{k,l}) = 0.$$

Thus, by (8), we have $x^k = \tilde{x}^k$ and $\lim_{l\to\infty} y^{k,l} = y^k$. Hence, it follows from (9) and (21) that:

$$\begin{aligned} -K^* y^k - \frac{1}{\tau}(\tilde{x}^k - x^k) &\in \partial f(\tilde{x}^k), \\ e^{k,l} + K(2\tilde{x}^k - x^k) + \frac{1}{\sigma}(y^{k,l} - y^k) &\in \partial g(y^{k,l}), \end{aligned}$$

which can be simplified as:

$$\begin{aligned} -K^* y^k &\in \partial f(x^k), \\ e^{k,l} + Kx^k + \frac{1}{\sigma}(y^{k,l} - y^k) &\in \partial g(y^{k,l}). \end{aligned}$$

Taking $l \to \infty$ in the above relations and using that the graph of the subdifferential mappings of a proper l.s.c. convex function is closed, we obtain $-K^* y^k \in \partial f(x^k)$ and $Kx^k \in \partial g(y^k)$, which implies (x^k, y^k) is a solution of (1). The proof is complete. □

By the previous Lemma 2, to analyze the global convergence and convergence rate of Algorithm 1, in the following, we assume (x^k, y^k) generated by Algorithm 1 is not a solution of (1) for any k, which implies a $\tilde{y}^k$ satisfying criterion (10) in step 2 of Algorithm 1 can be always computed by a proper method $\mathscr{G}$. Now, we give the key lemma for showing the convergence of I-PDA.

Lemma 3 *Let $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ be the iterates generated by Algorithm 1. Then for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $k \geq 0$, we have:*

$$\begin{aligned} \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k) \leq \frac{1}{\rho}\Big(\varphi(x^k - x, y^k - y) - \varphi(x^{k+1} - x, y^{k+1} - y)\Big) \\ -\frac{1}{4}(1-\eta^2)(2-\rho)\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k). \end{aligned} \tag{23}$$

Proof First, it follows from (9) that:

$$f(x) - f(\tilde{x}^k) + \langle x - \tilde{x}^k, K^* y^k + \frac{1}{\tau}(\tilde{x}^k - x^k)\rangle \geq 0, \quad \forall\, x \in \mathcal{X}. \tag{24}$$

By rearranging terms, we obtain

$$\begin{aligned} &\langle \tilde{x}^k - x, \frac{1}{\tau}(x^k - \tilde{x}^k) - K^*(y^k - \tilde{y}^k)\rangle + \langle \tilde{x}^k - x, K^*(y - \tilde{y}^k)\rangle \\ &\geq f(\tilde{x}^k) - f(x) + \langle \tilde{x}^k - x, K^* y\rangle, \quad \forall\, x \in \mathcal{X}. \end{aligned} \tag{25}$$

Similarly, according to (11), we get:

$$g(y) - g(\tilde{y}^k) + \langle y - \tilde{y}^k, -K\tilde{x}^k - K(\tilde{x}^k - x^k) + \frac{1}{\sigma}(\tilde{y}^k - y^k) - e^k\rangle \geq 0, \quad \forall\, y \in \mathcal{Y}. \tag{26}$$

By rearranging terms, we get:

$$\begin{aligned}&\langle \tilde{y}^k - y, -K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(y^k - \tilde{y}^k) + e^k\rangle - \langle \tilde{x}^k - x, K^*(y - \tilde{y}^k)\rangle \\ \geq\ & g(\tilde{y}^k) - g(y) + \langle \tilde{y}^k - y, -Kx\rangle, \quad \forall\, y \in \mathcal{Y}. \end{aligned} \tag{27}$$

Summing (25) and (27), we can derive:

$$\begin{aligned}&\langle \tilde{x}^k - x, \frac{1}{\tau}(x^k - \tilde{x}^k) - K^*(y^k - \tilde{y}^k)\rangle \\ &\quad + \langle \tilde{y}^k - y, -K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(y^k - \tilde{y}^k) + e^k\rangle \\ &\geq \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k), \quad \forall\, x \in \mathcal{X}, \quad \forall\, y \in \mathcal{Y}. \end{aligned} \tag{28}$$

On the other hand, it follows from (12) and (13) that:

$$\frac{1}{\tau}(x^k - \tilde{x}^k) - K^*(y^k - \tilde{y}^k) = \frac{1}{\tau}d_1^k - K^*d_2^k, \tag{29}$$

$$-K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(y^k - \tilde{y}^k) + e^k = -Kd_1^k + \frac{1}{\sigma}d_2^k. \tag{30}$$

Substituting (29) and (30) into (28), we obtain:

$$\begin{aligned}&\langle x^k - x, \frac{1}{\tau}d_1^k - K^*d_2^k\rangle + \langle y^k - y, -Kd_1^k + \frac{1}{\sigma}d_2^k\rangle \\ \geq\ & \langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^*d_2^k\rangle + \langle y^k - \tilde{y}^k, -Kd_1^k + \frac{1}{\sigma}d_2^k\rangle \\ &+ \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k), \quad \forall\, x \in \mathcal{X}, \quad \forall\, y \in \mathcal{Y}. \end{aligned} \tag{31}$$

By some simple manipulations, we have:

$$\begin{aligned}&\langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^*d_2^k\rangle + \langle y^k - \tilde{y}^k, -Kd_1^k + \frac{1}{\sigma}d_2^k\rangle \\ =\ & \frac{1}{\tau}\langle x^k - \tilde{x}^k, d_1^k\rangle + \frac{1}{\sigma}\langle y^k - \tilde{y}^k, d_2^k\rangle - \langle x^k - \tilde{x}^k, K^*d_2^k\rangle - \langle y^k - \tilde{y}^k, Kd_1^k\rangle \\ =\ & \frac{1}{2\tau}\left(\|x^k - \tilde{x}^k\|^2 + \|d_1^k\|^2 - \|x^k - \tilde{x}^k - d_1^k\|^2\right) \\ &+ \frac{1}{2\sigma}\left(\|y^k - \tilde{y}^k\|^2 + \|d_2^k\|^2 - \|y^k - \tilde{y}^k - d_2^k\|^2\right) - \langle d_1^k, K^*d_2^k\rangle \\ &- \langle x^k - \tilde{x}^k, K^*(y^k - \tilde{y}^k)\rangle + \langle x^k - \tilde{x}^k - d_1^k, K^*(y^k - \tilde{y}^k - d_2^k)\rangle \\ =\ & \frac{1}{2}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \frac{1}{2}\varphi(d_1^k, d_2^k) \\ &- \frac{1}{2}\varphi(x^k - \tilde{x}^k - d_1^k, y^k - \tilde{y}^k - d_2^k). \end{aligned} \tag{32}$$

Then, by the definitions of H and $\varphi(\cdot,\cdot)$ in (6) and (7), (12) and (13), we obtain:

$$\begin{aligned}&\varphi(x^k - \tilde{x}^k - d_1^k, y^k - \tilde{y}^k - d_2^k) = \varphi(-\tau K^*He^k, -He^k) \\ =\ & \tau\|K^*He^k\|^2 - 2\langle \tau K^*He^k, K^*He^k\rangle + \frac{1}{\sigma}\|He^k\|^2 \\ =\ & \|e^k\|_H^2. \end{aligned} \tag{33}$$

Substituting (32) and (33) into (31), and applying the inexact criterion (10), we can further get for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\begin{aligned}
&\langle x^k - x, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - y, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle \\
\geq\ &\langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - \tilde{y}^k, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle \\
&+\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k) \qquad (34)\\
=\ &\frac{1}{2}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \frac{1}{2}\varphi(d_1^k, d_2^k) - \frac{1}{2}\|e^k\|_H^2 + \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k) \\
\geq\ &\frac{1-\eta^2}{2}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \frac{1}{2}\varphi(d_1^k, d_2^k) + \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k). \qquad (35)
\end{aligned}$$

From (34) to (35), a lower bound on the stepsize α_k can be derived as:

$$\begin{aligned}
\alpha_k\ &=\ \frac{\langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - \tilde{y}^k, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle}{\varphi(d_1^k, d_2^k)} \\
&\geq\ \frac{\frac{1-\eta^2}{2}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \frac{1}{2}\varphi(d_1^k, d_2^k)}{\varphi(d_1^k, d_2^k)} \\
&\geq\ \frac{1}{2}. \qquad (36)
\end{aligned}$$

Therefore, we have for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\begin{aligned}
&\varphi(x^k - x, y^k - y) - \varphi(x^{k+1} - x, y^{k+1} - y) \\
=\ &\varphi(x^k - x, y^k - y) - \varphi(x^k - x - \rho\alpha_k d_1^k, y^k - y - \rho\alpha_k d_2^k) \\
=\ &2\rho\alpha_k\left(\langle x^k - x, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - y, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle\right) - \rho^2\alpha_k^2\varphi(d_1^k, d_2^k) \\
\geq\ &2\rho\alpha_k\left(\langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - \tilde{y}^k, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle\right) \\
&-\rho^2\alpha_k^2\varphi(d_1^k, d_2^k) + 2\rho\alpha_k\left(\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)\right) \\
=\ &(2-\rho)\rho\alpha_k\left(\langle x^k - \tilde{x}^k, \frac{1}{\tau}d_1^k - K^* d_2^k\rangle + \langle y^k - \tilde{y}^k, -K d_1^k + \frac{1}{\sigma}d_2^k\rangle\right) \\
&+2\rho\alpha_k\left(\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)\right) \\
\geq\ &\frac{1}{2}(1-\eta^2)(2-\rho)\rho\alpha_k\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + 2\rho\alpha_k\left(\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)\right) \\
\geq\ &\frac{1}{4}(1-\eta^2)(2-\rho)\rho\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \rho\left(\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)\right), \qquad (37)
\end{aligned}$$

where the first inequality follows from (34), the third equality follows from the definition of α_k in (16), the second inequality follows from (8) and (35), and the third inequality follows from (36). This completes the proof. □

Based on the analysis for showing the previous Lemma 3, we can easily have the following corollary.

Corollary 1 *If $\varphi(d_1^k, d_2^k) = 0$, then $(\tilde{x}^k, \tilde{y}^k)$ is a saddle point solution of (1).*

Proof If $\varphi(d_1^k, d_2^k) = 0$, we have $d_1^k = 0$ and $d_2^k = 0$ because of (8). Then, it follows from (29) to (30) that:

$$\begin{aligned} \frac{1}{\tau}(x^k - \tilde{x}^k) - K^*(y^k - \tilde{y}^k) &= 0 \qquad \text{and} \\ -K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(y^k - \tilde{y}^k) + e^k &= 0. \end{aligned}$$

Substituting the above two equalities into (24) and (26), we obtain:

$$\begin{aligned} f(x) - f(\tilde{x}^k) + \langle x - \tilde{x}^k, K^*\tilde{y}^k\rangle &\geq 0, \quad \forall\, x \in \mathcal{X}, \\ g(y) - g(\tilde{y}^k) + \langle y - \tilde{y}^k, -K\tilde{x}^k\rangle &\geq 0, \quad \forall y \in \mathcal{Y}, \end{aligned}$$

which means $(\tilde{x}^k, \tilde{y}^k)$ is a saddle point solution of (1). □

The following theorem gives the global convergence of the iterates generated by I-PDA as well as its ergodic convergence rate.

Theorem 1 *Let $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ be the iterates generated by Algorithm 1. Then, $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ converge to a same solution of (1). Furthermore, for the ergodic sequence $\{(X^N, Y^N)\}$ given by:*

$$X^N = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{x}^k \quad \textit{and} \quad Y^N = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{y}^k, \tag{38}$$

it holds that:

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{\varphi(x^0 - x, y^0 - y)}{\rho N}, \quad \forall\, x \in \mathcal{X}, \quad \forall\, y \in \mathcal{Y}. \tag{39}$$

Proof Summing the inequality (23) over $k = 0, 1, \ldots, N-1$, we have:

$$\begin{aligned} &\frac{1}{4}(1-\eta^2)(2-\rho)\sum_{k=0}^{N-1}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) + \sum_{k=0}^{N-1}\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k) \\ &\quad + \frac{1}{\rho}\varphi(x^N - x, y^N - y) \leq \frac{1}{\rho}\varphi(x^0 - x, y^0 - y), \quad \forall\, x \in \mathcal{X}, \quad \forall\, y \in \mathcal{Y}. \end{aligned} \tag{40}$$

Setting (x, y) as an arbitrary solution $(\hat{x}, \hat{y})$ of (1) and using (8), (40), and the fact $\mathcal{L}(\tilde{x}^k, \hat{y}) - \mathcal{L}(\hat{x}, \tilde{y}^k) \geq 0$, we conclude that $\{(x^k, y^k)\}$ is bounded and:

$$\beta_1 \sum_{k=0}^{\infty}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) \leq \sum_{k=0}^{\infty}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) < \infty.$$

Hence, we have $\|(x^k, y^k) - (\tilde{x}^k, \tilde{y}^k)\| \to 0$ as $k \to \infty$ and by (10), $e^k \to 0$ as $k \to \infty$. Furthermore, there exists a subsequence $\{(x^{k_j}, y^{k_j})\}$ converging to a limit

point $(x^\infty, y^\infty) \in \mathcal{X} \times \mathcal{Y}$. Hence, substituting k by k_j in (24) and (26) and taking the limits as $j \to \infty$, it follows from the lower semicontinuities of f and g that:

$$\begin{aligned} f(x) - f(x^\infty) + \langle x - x^\infty, K^* y^\infty \rangle &\geq 0, \quad \forall x \in \mathcal{X}, \\ g(y) - g(y^\infty) + \langle y - y^\infty, -K x^\infty \rangle &\geq 0, \quad \forall y \in \mathcal{Y}, \end{aligned}$$

which shows (x^∞, y^∞) is a saddle point solution of (1). Notice that (23) holds for any solution of (1). Hence, we have:

$$\varphi(x^{k+1} - x^\infty, y^{k+1} - y^\infty) \leq \varphi(x^k - x^\infty, y^k - y^\infty), \quad \forall k \geq 0,$$

which implies:

$$\varphi(x^k - x^\infty, y^k - y^\infty) \leq \varphi(x^{k_j} - x^\infty, y^{k_j} - y^\infty), \quad \forall k \geq k_j.$$

Then, it follows from (8) and $\{(x^{k_j}, y^{k_j})\}$ converging to (x^∞, y^∞) that the whole sequence $\{(x^k, y^k)\}$ converges to (x^∞, y^∞). In addition, we also have $\{(\tilde{x}^k, \tilde{y}^k)\}$ converges to (x^∞, y^∞).

Now, it follows from (40) that:

$$\sum_{k=0}^{N-1} \mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k) \leq \frac{1}{\rho}\varphi(x^0 - x, y^0 - y) - \frac{1}{\rho}\varphi(x^N - x, y^N - y).$$

Then, by the convexity of $\mathcal{L}(\cdot, y) - \mathcal{L}(x, \cdot)$ and (8), we have:

$$N\left(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)\right) \leq \frac{1}{\rho}\varphi(x^0 - x, y^0 - y).$$

which gives (39). □

Theorem 1 shows that the iterative sequences generated by I-PDA converge to a solution of (1), which is stronger than that in [20], where it only shows any cluster point of the sequence $\{(x^k, y^k)\}$ is a solution of (1). This stronger result comes from the different inexact criterion (10) and the correction steps used in I-PDA. In addition, exactly similar bounds as (39) are also established in [4, 5] to indicate a worst-case $O(1/N)$ convergence rate at the ergodic iterates. In fact, for any fixed solution $(\hat{x}, \hat{y}) \in \widehat{\mathcal{U}}$, we can consider the functions $\mathcal{L}(\cdot, \hat{y})$ and $\mathcal{L}(\hat{x}, \cdot)$ associated with the saddle point $(\hat{x}, \hat{y})$. Then, by setting $(x, y) = (\hat{x}, \hat{y})$ in (39), we can derive the values of convex function $\mathcal{L}(\cdot, \hat{y})$ at $\{X^N\}$ converge to its minimum value $\mathcal{L}(\hat{x}, \hat{y})$ with the rate of:

$$\mathcal{L}(X^N, \hat{y}) - \mathcal{L}(\hat{x}, \hat{y}) \leq \mathcal{L}(X^N, \hat{y}) - \mathcal{L}(\hat{x}, Y^N) \leq \phi(x^0 - \hat{x}, y^0 - \hat{y})/(\rho N) = \mathcal{O}(1/N).$$

And similarly, we have the values of concave function $\mathcal{L}(\hat{x}, \cdot)$ at $\{Y^N\}$ converges to its maximum value $\mathcal{L}(\hat{x}, \hat{y})$ with the rate of:

$$\mathcal{L}(\hat{x}, \hat{y}) - \mathcal{L}(\hat{x}, Y^N) \leq \mathcal{L}(X^N, \hat{y}) - \mathcal{L}(\hat{x}, Y^N) = \mathcal{O}(1/N).$$

4 Linear convergence

In this section, we establish the Q-linear convergence rate of the distance of the iterate u^k to the solution set $\widehat{\mathcal{U}}$, i.e., $\mathrm{dist}_G(u^k, \widehat{\mathcal{U}})$, which leads to the R-linear convergence rate for the iterates $\{(x^k, y^k)\}$.

The following lemma provides an upper bound for $\|R(\tilde{x}^k, \tilde{y}^k)\|$, where $R(\cdot)$ is defined in (5).

Lemma 4 *Let $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ be the iterates generated by Algorithm 1. Then, for any $k \geq 0$, there exists a constant $\kappa_1 > 0$ such that:*

$$\|R(\tilde{u}^k)\|^2 \leq \kappa_1 \varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k), \tag{41}$$

where:

$$\kappa_1 := \frac{1}{\beta_1} \max\left\{3L^2 + \frac{2}{\tau^2}, 2L^2 + \frac{3}{\sigma^2}\right\} + \frac{3\eta^2}{\lambda_{\min}(H)},$$

and $\lambda_{\min}(H) > 0$ is the minimum eigenvalue of H.

Proof First, the optimality condition of (9) can be read as:

$$\tilde{x}^k = \mathrm{prox}_f\left[\tilde{x}^k - \left(\frac{1}{\tau}(\tilde{x}^k - x^k) + K^* y^k\right)\right]. \tag{42}$$

Similarly, the optimality condition of (11) can be read as:

$$\tilde{y}^k = \mathrm{prox}_g\left[\tilde{y}^k - (-K(2\tilde{x}^k - x^k) + \frac{1}{\sigma}(\tilde{y}^k - y^k) - e^k)\right]. \tag{43}$$

Then, it follows from (42), (43), and the definition of $R(\cdot)$ in (5) that:

$$\begin{aligned}
\|R(\tilde{u}^k)\|^2 &= \|\tilde{x}^k - \mathrm{prox}_f(\tilde{x}^k - K^*\tilde{y}^k)\|^2 + \|\tilde{y}^k - \mathrm{prox}_g(\tilde{y}^k + K\tilde{x}^k)\|^2 \\
&\leq \| - \frac{1}{\tau}(\tilde{x}^k - x^k) + K^*(\tilde{y}^k - y^k)\|^2 \\
&\quad + \|K(x^k - \tilde{x}^k) + \frac{1}{\sigma}(\tilde{y}^k - y^k) - e^k\|^2 \\
&\leq \frac{2}{\tau^2}\|x^k - \tilde{x}^k\|^2 + 2L^2\|y^k - \tilde{y}^k\|^2 \\
&\quad + 3L^2\|x^k - \tilde{x}^k\|^2 + \frac{3}{\sigma^2}\|y^k - \tilde{y}^k\|^2 + 3\|e^k\|^2 \\
&\leq \left(3L^2 + \frac{2}{\tau^2}\right)\|x^k - \tilde{x}^k\|^2 + \left(2L^2 + \frac{3}{\sigma^2}\right)\|y^k - \tilde{y}^k\|^2 + \frac{3}{\lambda_{\min}(H)}\|e^k\|_H^2 \\
&\leq \left(3L^2 + \frac{2}{\tau^2}\right)\|x^k - \tilde{x}^k\|^2 + \left(2L^2 + \frac{3}{\sigma^2}\right)\|y^k - \tilde{y}^k\|^2 \\
&\quad + \frac{3\eta^2}{\lambda_{\min}(H)}\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k) \\
&\leq \kappa_1 \varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k),
\end{aligned}$$

where the first inequality follows from the 1-Lipschitz continuities of $\text{prox}_f(\cdot)$ and $\text{prox}_g(\cdot)$, the second inequality uses the fact of $\|K\| = L$, the fourth inequality follows from the inexact criterion (10), and the last inequality follows from (8). □

Now, we are ready to establish the linear rate convergence of Algorithm 1 under certain calmness condition on R^{-1}. Note that this calmness condition are also proposed for establishing the linear convergence of alternating direction method of multipliers in [17].

Theorem 2 *Let $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ be the iterates generated by Algorithm 1. The following properties hold.*

(i) there exists a solution $u^\infty := (x^\infty, y^\infty)$ of (1) such that the $\{(x^k, y^k)\}$ and $\{(\tilde{x}^k, \tilde{y}^k)\}$ converge to $u^\infty \in \widehat{\mathcal{U}}$;

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

$$dist(u, \widehat{\mathcal{U}}) \leq \theta \|R(u)\|, \quad \forall\, u \in \{u \in \mathcal{U} \,\big|\, \|u - u^\infty\| \leq r\}, \tag{44}$$

for some $r > 0$, there exists a positive number $\xi \in [\kappa, 1)$ such that:

$$\text{dist}_G(u^{k+1}, \widehat{\mathcal{U}}) \leq \xi \text{dist}_G(u^k, \widehat{\mathcal{U}}), \tag{45}$$

for all $k \geq 0$, where:

$$\kappa := \sqrt{1 - \frac{(1-\eta^2)(2-\rho)\rho}{4(1+\theta\sqrt{\kappa_1\beta_2})^2}} < 1.$$

(iii) The iterates $\{u^k\} := \{(x^k, y^k)\}$ converges R-linearly.

Proof By Theorem 1, we already know that the property (i) holds. Hence, there exists a $\bar{k} \geq 0$ such that for all:

$$\|\tilde{u}^k - u^\infty\| \leq r, \quad \forall\, k \geq \bar{k}.$$

Thus, by using Lemma 4 and (44), we know that for all $k \geq \bar{k}$:

$$\text{dist}(\tilde{u}^k, \widehat{\mathcal{U}}) \leq \theta \|R(\tilde{u}^k)\| \leq \theta\sqrt{\kappa}_1 \sqrt{\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k)}, \tag{46}$$

where κ_1 is given in Lemma 4. Next, by the definition of φ in (7) with G a positive definite operator, it follows from the definition of the distance function $\text{dist}_G(\cdot, \widehat{\mathcal{U}})$ that:

$$\begin{aligned} \text{dist}(\tilde{u}^k, \widehat{\mathcal{U}}) &\geq \frac{1}{\sqrt{\beta_2}} \text{dist}_G(\tilde{u}^k, \widehat{\mathcal{U}}) \\ &\geq \frac{1}{\sqrt{\beta_2}} \left(\text{dist}_G(u^k, \widehat{\mathcal{U}}) - \sqrt{\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k)} \right). \end{aligned} \tag{47}$$

By combining (46) with (47), we obtain for $k \geq \bar{k}$:

$$\text{dist}_G(u^k, \widehat{\mathcal{U}}) \leq (1 + \theta\sqrt{\kappa_1\beta_2}) \sqrt{\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k)}. \tag{48}$$

Note that for any $(\hat{x}, \hat{y}) \in \widehat{\mathcal{U}}$, it follows from (37) in Lemma 3 and $\mathcal{L}(\tilde{x}^k, \hat{y}) - \mathcal{L}(\hat{x}, \tilde{y}^k) \geq 0$ that:

$$\begin{aligned} &\varphi(x^k - \hat{x}, y^k - \hat{y}) - \varphi(x^{k+1} - \hat{x}, y^{k+1} - \hat{y}) \\ \geq\ & \frac{1}{4}(1-\eta^2)(2-\rho)\rho\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k). \end{aligned} \tag{49}$$

Then, by the definition of $\mathrm{dist}_G(\cdot, \widehat{\mathcal{U}})$ with $\widehat{\mathcal{U}}$ being a nonempty closed convex set, for all $k \geq 0$, we have:

$$\mathrm{dist}_G^2(u^k, \widehat{\mathcal{U}}) - \mathrm{dist}_G^2(u^{k+1}, \widehat{\mathcal{U}}) \geq \frac{1}{4}(1-\eta^2)(2-\rho)\rho\varphi(x^k - \tilde{x}^k, y^k - \tilde{y}^k). \tag{50}$$

Hence, by (48), for $k \geq \bar{k}$, we have:

$$\mathrm{dist}_G^2(u^k, \widehat{\mathcal{U}}) - \mathrm{dist}_G^2(u^{k+1}, \widehat{\mathcal{U}}) \geq \frac{(1-\eta^2)(2-\rho)\rho}{4(1+\theta\sqrt{\kappa_1\beta_2})^2}\mathrm{dist}_G^2(u^k, \widehat{\mathcal{U}}). \tag{51}$$

Then, (50), (51), and Lemma 4 imply that the property (ii) holds, i.e., (45) holds for all $k \geq 0$.

Now, we show property (iii). Select $\hat{u}^k = (\hat{x}^k, \hat{y}^k) \in \widehat{\mathcal{U}}$ such that $\mathrm{dist}_G(u^k, \widehat{\mathcal{U}}) = \|u^k - \hat{u}^k\|_G$ and denote $\delta^k = u^{k+1} - u^k$. Then, it follows from (49) that:

$$\|u^k - \hat{u}^k\|_G^2 - \|u^{k+1} - \hat{u}^k\|_G^2 = \varphi(x^k - \hat{x}^k, y^k - \hat{y}^k) - \varphi(x^{k+1} - \hat{x}^k, y^{k+1} - \hat{y}^k) \geq 0.$$

Hence, by (45), we have:

$$\begin{aligned} \|\delta^k\|_G &= \|u^{k+1} - u^k\|_G \\ &\leq \|u^{k+1} - \hat{u}^k\|_G + \|u^k - \hat{u}^k\|_G \\ &\leq 2\|u^k - \hat{u}^k\|_G = 2\mathrm{dist}_G(u^k, \widehat{\mathcal{U}}) \\ &\leq 2\xi^k \mathrm{dist}_G(u^0, \widehat{\mathcal{U}}). \end{aligned} \tag{52}$$

Then, it follows from $\{u^k\}$ converging to $u^\infty \in \widehat{\mathcal{U}}$ that $u^\infty = u^k + \sum_{j=k}^\infty \delta^j$. So:

$$\begin{aligned} \|u^k - u^\infty\|_G &\leq \sum_{j=k}^\infty \|\delta^j\|_G \leq 2\mathrm{dist}_G(u^0, \widehat{\mathcal{U}}) \sum_{j=k}^\infty \xi^j \\ &= 2\mathrm{dist}_G(u^0, \widehat{\mathcal{U}})\xi^k \sum_{j=0}^\infty \xi^j \\ &= \xi^k \left[2\mathrm{dist}_G(u^0, \widehat{\mathcal{U}})\frac{1}{1-\xi}\right], \end{aligned}$$

which shows $\{u^k\}$ converging to u^∞ R-linearly. □

Under proper calmness condition (44), Theorem 2 shows the Q-linear convergence rate of $\mathrm{dist}_G(u^k, \widehat{\mathcal{U}})$ and the nonergodic R-linear convergence rate for the iterates $\{u^k\}$. Although the constant θ in the calmness condition (44) is not easy to evaluate, our results are more general and stronger than those in [4, 5] which are based on the strong convexity of the objective function.

Corollary 2 *Let* $\{(x^k, y^k)\}$ *and* $\{(\tilde{x}^k, \tilde{y}^k)\}$ *be the iterates generated by Algorithm 1. Assume the mapping* $R : \mathcal{U} \to \mathcal{U}$ *is piecewise polyhedral. Then, the following properties hold.*
(i) There exists a constant $\hat{\theta} > 0$ *such that for all* $k \geq 0$ *we have:*

$$\operatorname{dist}(u, \widehat{\mathcal{U}}) \leq \hat{\theta} \|R(u)\|. \tag{53}$$

(ii) For all $k \geq 0$*, we have:*

$$\operatorname{dist}_G(u^{k+1}, \widehat{\mathcal{U}}) \leq \hat{\kappa} \operatorname{dist}_G(u^k, \widehat{\mathcal{U}}), \tag{54}$$

where:

$$\hat{\kappa} := \sqrt{1 - \frac{(1-\eta^2)(2-\rho)\rho}{4(1+\hat{\theta}\sqrt{\kappa_1\beta_2})^2}} < 1.$$

(iii) The iterates $\{u^k\} := \{(x^k, y^k)\}$ *converges R-linearly.*

Proof Since R^{-1} is piecewise polyhedral if and only if R is piecewise polyhedral [17], it follows from [30] that there exist two constants $\theta > 0$ and $s > 0$ such that:

$$\operatorname{dist}(u, \widehat{\mathcal{U}}) \leq \hat{\theta} \|R(u)\|, \quad \forall\, u \in \{u \in \mathcal{U} \,\big|\, \|R(u)\| \leq s\}. \tag{55}$$

By Theorem 2, we know $\{u^k\}$ converges to $u^\infty \in \widehat{\mathcal{U}}$. Hence, there exists a constant $r > 0$ such that $\|u^k - u^\infty\| \leq r$ for all $k \geq 0$. Note that when $\|R(u^k)\| > s$, we have:

$$\operatorname{dist}(u^k, \widehat{\mathcal{U}}) \leq \|u^k - u^\infty\| \leq r < \frac{r}{s}\|R(u^k)\|. \tag{56}$$

Combining (55) and (56), we have (53) holds with $\hat{\theta} := \max\{\theta, \frac{r}{s}\}$. Using (53), the properties (ii) and (iii) can be similarly proved as the proof in Theorem 2. □

5 Applications to some convex optimization models

In this section, we give some examples arising from practical applications, where the linear convergence results in the previous section will apply. As one can see in Theorem 2, the calmness condition is the key assumption for linear convergence. In order to show the linear convergence rate of I-PDA for solving these problems, it is sufficient to show the KKT mapping (4) of these problems satisfy the calmness condition (44). From the discussions in Section 2, it is sufficient to show the inverse operator of KKT mapping defined in (5) is piecewise polyhedral.

Note that the objective functions f and g involved in the following examples (except the elastic net problem) do not satisfy the strongly convex condition. Hence, the theoretical results given in [4, 5, 27] do not imply the linear convergence rate of PDA. However, from our analysis, these models satisfy the calmness condition and the linear convergence rate can be obtained immediately.

5.1 Matrix games

The matrix games can be applied to model the two-person zero-sum games [5]. Consider the following min-max matrix game [5, 21]:

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\langle Kx, y\rangle, \tag{57}$$

where $K \in \mathcal{R}^{m\times n}$, Δ_n, and Δ_m denote the standard unit simplices in $\mathcal{R}^n$ and $\mathcal{R}^m$, respectively. Note that this problem (57) can be reformulated as:

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m}\delta_{\Delta_n}(x)+\langle Kx, y\rangle-\delta_{\Delta_m}(y). \tag{58}$$

Then, the KKT mapping for this model (58) is:

$$R(u) := \begin{pmatrix} x-\Pi_{\Delta_n}(x-K^*y) \\ y-\Pi_{\Delta_m}(y+Kx)\end{pmatrix}, \quad \forall\, u\in\mathcal{U}.$$

By recalling that Δ_n and Δ_m are polyhedral, Lemma 1 implies that $\Pi_{\Delta_n}(\cdot)$ and $\Pi_{\Delta_m}(\cdot)$ are piecewise polyhedral, and so are R and R^{-1}.

5.2 ℓ_1 regularized least squares

The ℓ_1 regularized least squares model, which includes LASSO model, is widely used in signal processing and sparse optimization. Consider the following ℓ_1 regularized problem [21]:

$$\min_{x\in\mathcal{R}^n}\frac{1}{2}\|Kx-b\|^2+\lambda\|x\|_1, \tag{59}$$

where $K \in \mathcal{R}^{m\times n}$ and $b \in \mathcal{R}^m$. Analogously, we can rewrite (59) as:

$$\min_{x\in\mathcal{R}^n}\max_{y\in\mathcal{R}^m} f(x)+\langle Kx, y\rangle-g(y), \tag{60}$$

where $f(x) = \lambda\|x\|_1$ and $g(y) = \frac{1}{2}\|y\|^2 + b^T y$. Then, the KKT mapping for this model (60) is:

$$R(u) := \begin{pmatrix} x-\mathrm{prox}_f(x-K^*y) \\ y-\mathrm{prox}_g(y+Kx)\end{pmatrix}, \quad \forall\, u\in\mathcal{U}.$$

Since ∂f is piecewise linear, f is piecewise linear-quadratic. In addition, g is quadratic. Consequently, $\mathrm{prox}_f(\cdot)$ and $\mathrm{prox}_g(\cdot)$ are piecewise polyhedral, and so are R and R^{-1}.

5.3 Nonnegative least squares

Consider the following nonnegative least squares problem [21]:

$$\min_{x\in\mathcal{R}^n_+}\frac{1}{2}\|Kx-b\|^2, \tag{61}$$

where $K \in \mathcal{R}^{m\times n}$ and $b \in \mathcal{R}^m$. One saddle point formulation of (61) can be written as:

$$\min_{x\in\mathcal{R}^n} \max_{y\in\mathcal{R}^m} \delta_{\mathcal{R}^n_+}(x) + \langle Kx, y\rangle - g(y), \tag{62}$$

where $g(y) = \frac{1}{2}\|y\|^2 + b^T y$. Then the KKT mapping for this model (62) is:

$$R(u) := \begin{pmatrix} x - \Pi_{\mathcal{R}^n_+}(x - K^* y) \\ y - \text{prox}_g(y + Kx) \end{pmatrix}, \quad \forall\, u \in \mathcal{U}.$$

Since $\mathcal{R}^n_+$ is polyhedral and g is quadratic, $\Pi_{\mathcal{R}^n_+}(\cdot)$ and $\text{prox}_g(\cdot)$ are piecewise polyhedral, and so are R and R^{-1}.

5.4 Elastic net problem

The elastic net problem, which is used for feature selection and sparse coding [5], can be written as:

$$\min_{x\in\mathcal{R}^n} \frac{1}{2}\|Kx - b\|^2 + \lambda_1\|x\|_1 + \lambda_2\|x\|^2, \tag{63}$$

where $K \in \mathcal{R}^{m\times n}$ and $b \in \mathcal{R}^m$. Analogously, we can reformulate (63) as:

$$\min_{x\in\mathcal{R}^n} \max_{y\in\mathcal{R}^m} f(x) + \langle Kx, y\rangle - g(y), \tag{64}$$

where $f(x) = \lambda_1\|x\|_1 + \lambda_2\|x\|^2$ and $g(y) = \frac{1}{2}\|y\|^2 + b^T y$. Then the KKT mapping for this model (64) is:

$$R(u) := \begin{pmatrix} x - \text{prox}_f(x - K^* y) \\ y - \text{prox}_g(y + Kx) \end{pmatrix}, \quad \forall\, u \in \mathcal{U}.$$

Similarly, we can conclude that R^{-1} is piecewise polyhedral.

5.5 Fused LASSO

The fused lasso problem, which was proposed for group variable selection [35], can be written as:

$$\min_{y\in\mathcal{R}^n} F(y) := \|Dy\|_1 + \mu_1\|y\|_1 + \frac{\mu_2}{2}\|Ay - b\|^2, \tag{65}$$

where $A \in \mathcal{R}^{m\times n}$, $b \in \mathcal{R}^m$, and $D \in \mathcal{R}^{(n-1)\times n}$ is given by:

$$D = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \cdots & \cdots \\ & & & -1 \ \ 1 \end{pmatrix}.$$

One min-max reformulation of (65) can be equivalently written as:

$$\min_{x\in\mathcal{R}^m} \max_{y\in\mathcal{R}^n} f(x) + \langle Kx, y\rangle - g(y), \tag{66}$$

where $f(x) = \delta_{\mathcal{B}_\infty}(x)$, $g(y) = \mu_1\|y\|_1 + \frac{\mu_2}{2}\|Ay-b\|^2$ and $K = D^*$. Then the KKT mapping for this model (66) is:

$$R(u) := \begin{pmatrix} x - \Pi_{\mathcal{B}_\infty}(x - K^*y) \\ y - \mathrm{prox}_g(y + Kx) \end{pmatrix}, \quad \forall\, u \in \mathcal{U}.$$

Similarly, we can conclude that R^{-1} is piecewise polyhedral.

5.6 TV-ℓ_2 image restoration

Many image processing problems involve both constraints and regularized terms, such as the tomography reconstruction, where both nonnegative constraints and total variation regularization appear. Consider the following constrained TV-ℓ_2 image restoration problem [16, 20, 23]:

$$\min_{y\in\mathcal{B}} \left\{ \|Dy\|_1 + \frac{1}{2\mu}\|Ay - c\|^2 \right\}, \tag{67}$$

where $c \in \mathcal{R}^n$ is the observed image, A is a blur operator, D is the discrete gradient operator [28], $\|Dy\|_1$ is the discrete TV regularization term, $\mathcal{B} = [0,1]^n$ is the unit box in $\mathcal{R}^n$, and μ is a positive parameter for balancing the data-fidelity and TV regularization. Here, $n = n_1 \times n_2$ is the total number of pixels, where n_1 and n_2 are the numbers of pixels in the horizontal and vertical directions, respectively. Note that the model (67) can be reformulated as the following saddle point problem:

$$\min_{x\in\mathcal{R}^p} \max_{y\in\mathcal{R}^n} \left\{ \delta_{\mathcal{B}_\infty}(x) + \langle D^*x, y\rangle - \delta_{\mathcal{B}}(y) - \frac{1}{2\mu}\|Ay - c\|^2 \right\}. \tag{68}$$

Clearly, (68) is the special case of (1) with $f(x) = \delta_{\mathcal{B}_\infty}(x)$, $g(y) = \delta_{\mathcal{B}}(y) + \frac{1}{2\mu}\|Ay - c\|^2$ and $K = D^*$. Then, the KKT mapping for this model (67) is:

$$R(u) := \begin{pmatrix} x - \Pi_{\mathcal{B}_\infty}(x - K^*y) \\ y - \mathrm{prox}_g(y + Kx) \end{pmatrix}, \quad \forall\, u \in \mathcal{U}.$$

Similarly, we can conclude that R^{-1} is piecewise polyhedral.

6 Numerical experiments

In this section, we would like to demonstrate the linear convergence rate and show the efficiency of I-PDA on several problems mentioned in Section 5. All codes were written by MATLAB R2016a and all the numerical experiments were performed on a laptop ThinkPad X1 Extreme with i7-8750H processor and 16GB memory.

6.1 Matrix games

We first consider a matrix games problem (57), which is generated following the same way given in [5]. The entries of K are generated independently and randomly

with uniformly distribution in the interval $[-1, 1]$. As in [5], for a feasible point pair (x, y), the primal-dual gap can be obtained by:

$$\Theta(x, y) := \max_i (Kx)_i - \min_j (K^* y)_j. \tag{69}$$

In this experiment, we set $m = 100$ and $n = 300$. Note that both the x-subproblem and y-subproblem in I-PDA for solving (57) can be efficiently solved exactly by performing projection onto the unit simplex [12]. So we can simply set $\eta = 0$ in I-PDA. The other parameters are set as $\tau = \sigma = \sqrt{0.99}/L$ and $\rho = 1$. Hence, we have $\tau\sigma L^2 < 1$. The starting point of I-PDA is chosen as $(x^0, y^0) = (\frac{1}{n}(1, \ldots, 1), \frac{1}{m}(1, \ldots, 1))$. By direct calculation, we have:

$$\max_{x \in \Delta_n} \frac{1}{2}\|x - x^0\|^2 = (1 - \frac{1}{n})/2$$

and

$$\max_{y \in \Delta_m} \frac{1}{2}\|y - y^0\|^2 = (1 - \frac{1}{m})/2.$$

Then, it follows from (7), (8) to (39) that:

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \le \frac{1}{N\rho}\left(\frac{1 - \frac{1}{n}}{\tau} + \frac{1 - \frac{1}{m}}{\sigma}\right).$$

To demonstrate the linear convergence rate, we first run I-PDA for sufficiently many iterations to obtain an almost exact solution (x^∞, y^∞) of the problem. Then, Fig. 1 (left) shows the convergence behaviors of:

$$\text{Error} := \|(x^k - x^\infty, y^k - y^\infty)\|_G = \sqrt{\varphi(x^k - x^\infty, y^k - y^\infty)}, \tag{70}$$

and Fig. 1 (right) shows the primal-dual gaps $\Theta(X^N, Y^N)$ on the ergodic iterates (X^N, Y^N) and $\Theta(\tilde{x}^k, \tilde{y}^k)$ on the nonergodic iterates $(\tilde{x}^k, \tilde{y}^k)$.

From Fig. 1 (left), we can see that the Error defined in (70) decreases rapidly at early iterations and then converges to zero in a steady linear rate. Figure 1 (right) shows that the primal-dual gap given in (69) at the nonergodic iterates $(\tilde{x}^k, \tilde{y}^k)$ converges faster than that at the ergodic iterates (X^N, Y^N). Moreover, we can see that

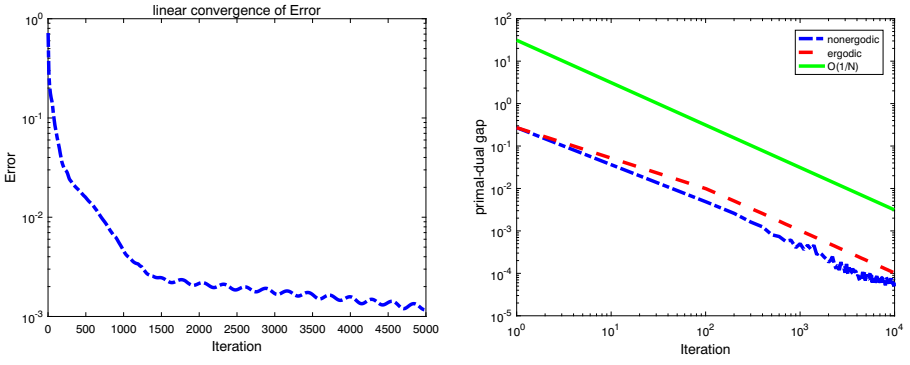

Fig. 1 Convergence plots for matrix games problem (57)

the primal-dual gap at ergodic sequence converges with almost an exact $\mathcal{O}(1/N)$ sublinear rate.

6.2 Nonnegative least squares

We then consider the nonnegative least squares problem (61). In this numerical experiment, the entries of $K \in \mathcal{R}^{m\times n}$ are independently randomly generated by the standard Gaussian distribution $\mathcal{N}(0, 1)$. To generate the vector b, we first randomly generate a vector $w \in \mathcal{R}^n$ by the standard Gaussian distribution and then take $b = K\Pi_{\mathcal{R}^n_+}(w)$. Hence, $F^* = 0$ is the optimal objective value of the problem. The problem dimensions are set as $m = 300$ and $n = 1000$. Note that both the x-subproblem and y-subproblem in I-PDA for solving problem (61) also has easily closed-form solution. Hence, we can also set $\eta = 0$ in I-PDA. The other parameters are set as $\tau = 0.4$, $\sigma = \frac{0.99}{\tau L^2}$ and $\rho = 1$. Hence, we have $\tau\sigma L^2 < 1$. We randomly choose a starting point (x^0, y^0) and the typical convergence behaviors of I-PDA are shown in Fig. 2.

Figure 2 (left) again clearly shows the linear convergence of Error defined in (70) to a high precision, where (x^∞, y^∞) is again obtained by running I-PDA sufficiently many iterations. Since F is Lipschitz continuous and x^k converges to x^∞ with a R-linear rate, $F(x^k) - F^*$ also converges to zero with a R-linear rate. We can see from Fig. 2 (right) that the primal function value gap $F(\tilde{x}^k) - F^*$ at nonergodic sequence converges with a faster linear convergence rate, while function value gap $F(X^N) - F^*$ at ergodic sequence only converges at $\mathcal{O}(1/N)$ rate. These convergence behaviors exactly match our theoretical analysis.

6.3 Fused LASSO

We now explore the efficiency of I-PDA for solving the fused LASSO model (65) by solving its subproblems inexactly. We can see that the y-subproblem in I-PDA for solving (65) does not have a closed-form solution. Hence, in our numerical experiments, the y-subproblem in I-PDA is solved inexactly by FISTA [2] to satisfy the

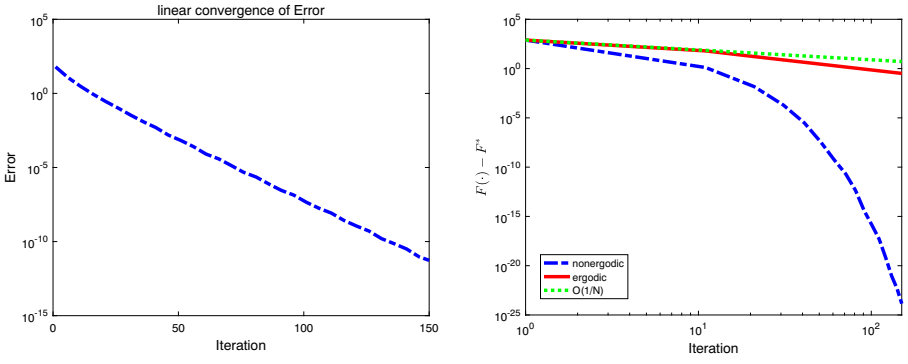

Fig. 2 Convergence plots for nonnegative least squares problem (61)

criterion (10) with $\eta = 0.99$. Note that when the y-subproblem in I-PDA is solved exactly, the I-PDA will reduce to the exact PDA method (2) with $\gamma = 1$. We compare the performance of I-PDA with exact PDA, simply denoted as PDA in the following tables and figures, and two inexact PDAs proposed in [20], that are an inexact PDA with absolute error (I-PDAa) and an inexact PDA with relative error (I-PDAr). For exact PDA, its y-subproblems are solved almost numerically exactly by FISTA until $\|e^k\| \leq 10^{-5}$. To avoid solving linear system at each iteration, we apply the strategy (17) proposed in I-PDA.

We generate the test problems by the same way used in [35]. More precisely, the entries of A are generated by the standard Gaussian distribution $\mathcal{N}(0, 1)$ and b is obtained by $b = Ax + \lambda e$, where e is a standard distributed Gaussian noise and $\lambda = 0.01$. The parameters are set as $\mu_1 = 0.1$ and $\mu_2 = 0.005$. For exact PDA, I-PDAa, and I-PDAr, we set $\tau = 0.8$ and $\sigma = 1/(4\tau)$, which give relatively good numerical results as chosen in [20]. For I-PDA, we set $\tau = 0.56$ and $\sigma = 0.7/(4\tau)$. Note that the $n-1$ eigenvalues of DD^* are $2-2\cos(i\pi/n)$, $i = 1, 2, \cdots, n-1$. and $K = D^*$ in (66). Hence, we have $\tau\sigma L^2 < 1$. We also randomly generate the starting point (x^0, y^0) and the stopping criterion of I-PDA is set as $\varphi(d_1^k, d_2^k) \leq 10^{-3}$.

In this experiment, we generate 5 testing scenarios with different dimensions (m, n) and use 10 different initial points for each scenario. The average performances of exact PDA, I-PDA, I-PDAa, and I-PDAr for each scenario are shown in Table 1, which includes the CPU time in seconds (CPU (s)), the outer iteration number (Iter), and the total inner iteration number (InnerIter) for solving the y-subproblem. From Table 1, we can see that the inner iteration numbers of all inexact PDAs, including I-PDA, I-PDAa, and I-PDAr, are significantly less than that of exact PDA, while the

Table 1 Numerical results for fused LASSO

n	m	PDA			I-PDA		
		CPU (s)	Iter	InnerIter	CPU (s)	Iter	InnerIter
25	500	1.33	380.2	4414.8	0.17	330.6	423.6
40	800	2.11	287.6	4337.2	0.31	300.8	337.3
50	800	2.28	275.1	5085.8	0.44	380.5	502.0
50	1000	3.85	294.9	5110.1	0.72	329.1	470.2
100	2000	87.25	650.8	17653.5	9.95	823.2	1647.1
n	m	I-PDAa			I-PDAr		
		CPU (s)	Iter	InnerIter	CPU (s)	Iter	InnerIter
25	500	0.43	406.0	1338.0	0.91	390.1	2928.4
40	800	0.72	344.0	1183.3	1.80	303.1	2932.5
50	800	0.60	314.7	1043.9	1.27	289.4	2585.3
50	1000	1.37	347.1	1451.4	2.57	317.3	3095.5
100	2000	35.66	651.1	6993.2	57.72	700.0	11518.7

outer iteration numbers of exact PDA are usually less than those of the inexact PDAs, but in a relatively small margin. Hence, we can see from Table 1 that the overall CPU time of the inexact PDAs is much less than that of the exact PDA. On the other hand, compared with I-PDAa and I-PDAr, I-PDA always uses the least CPU time and much less number of inner iterations. So, the relative stopping criterion implemented in I-PDA are more effective than than the inexact subproblem rules used by I-PDAa and I-PDAr. In addition, as pointed in [20], we can also observe that I-PDAa performs better than I-PDAr.

The typical convergence behaviors of these comparison methods against iteration number and CPU time can be illustrated in Figs. 3 and 4 for the case with $n = 50$ and $m = 1000$. In particular, Fig. 3 (left) shows the linear convergence rate of Error (70) for both exact PDA and I-PDA at nonergodic iterates and Fig. 3 (right) illustrates convergence behaviors of Error against CPU time for the four tested methods. Figure 4 (left) shows the sublinear convergence of $F(Y^k) - F^*$ of exact PDA and I-PDA at ergodic iterates, while Fig. 4 (right) demonstrates the linear convergence of $F(y^k) - F^*$ of all comparison methods at nonergodic iterates. Here, the optimal objective value F^* is obtained by running exact PDA for 5000 iterations. From Figs. 3 (right) to 4 (right), we can also observe that I-PDA converges much faster than the other three comparison algorithms. Note again that since F is Lipschitz continuous, the R-linear convergence of y^k to y^∞ implies the R-linear convergence of $F(y^k) - F^*$. These convergence behaviors of I-PDA shown in Figs. 3 and 4 again exactly match our analysis. More importantly, we can observe from Fig. 3 that although the y-subproblem was solved inexactly, I-PDA can still maintain the desired linear convergence rate. Its performance is only slightly worse compared with the exact PDA after the same number of iterations, but much better in terms of CPU time. Hence, for overall efficiency, it is much preferable to solve the subproblems inexactly to a relative accuracy given in I-PDA, when the subproblem is nontrivial to be solved exactly.

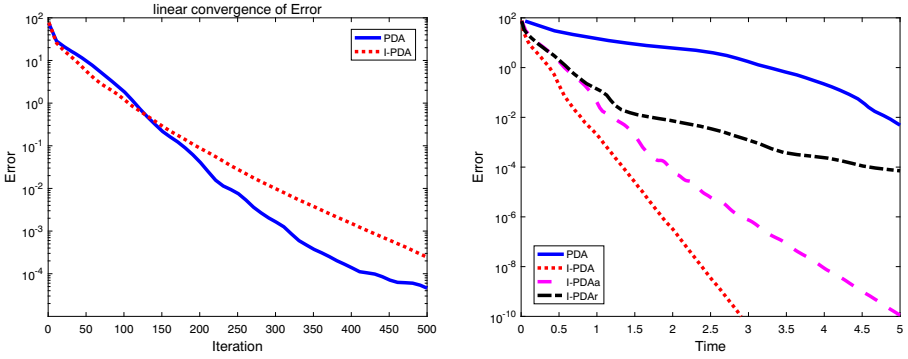

Fig. 3 Linear convergence plots for fused LASSO (65)

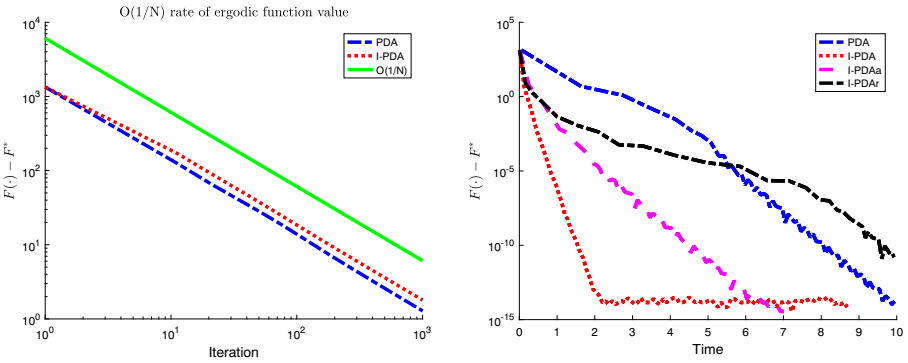

Fig. 4 Convergence rates of primal function value gap for fused LASSO (65)

7 Conclusions

The main contribution of this paper is to provide a road map for analyzing global convergence and the linear convergence rate of an inexact primal-dual algorithm (I-PDA) for solving a class of convex-concave saddle point problems. This I-PDA solves one of the subproblems inexactly to an accuracy relative to the overall optimality error at current iterate. We first analyze the global convergence and convergence rate of I-PDA under the standard condition. Then, with an additional mild calmness condition for the KKT mapping, which naturally holds for many convex models in practical applications, we have established the Q-linear convergence of the distance between the current iterate and the solution set, and the R-linear convergence of the primal-dual gap on the nonergodic iterates generated by I-PDA. These theoretical analyses show that although one subproblem is solved inexactly, the theoretical global convergence and linear convergence rate of exact PDA can still be maintained by I-PDA. Our numerical experiments clearly demonstrate the convergence rates obtained from the theoretical analysis and show that the I-PDA could be much more efficient than exact PDA as well as other compared inexact PDAs when the subproblems do not have closed-from solutions.

Funding This research was partially supported by the National Natural Science Foundation of China under grants 11571178, 11871279 and 12001286, by the China Scholarship Council, by the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX20_1163, and by the USA National Science Foundation under grant 1819161.

References

1. Arrow, K.J., Hurwicz, L., Uzawa, H.: With Contributions by H.B. chenery, S.M. Johnson, S. Karlin, T. Marschak, and R.M. Solow. Studies in Linear and Non-Linear Programming, volume II of Stanford Mathematical Studies in the Social Science. Stanford Unversity Press, Stanford (1958)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)

3. Cai, X.J., Han, D.R., Xu, L.L.: An improved first-order primal-dual algorithm with a new correction step. J. Glob. Optim. **57**(4), 1419–1428 (2013)
4. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
5. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**(1-2), 253–287 (2016)
6. Chambolle, A., Ehrhardt, M.J., Richtárik, P., Schonlieb, C.B.: Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. SIAM J. Optim. **28**(4), 2783–2808 (2018)
7. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. Inverse Probl. **29**(2), 025011 (2013)
8. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions. Fixed Point Theory A **2016**(1), 54 (2016)
9. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
10. Davis, D., Yin, W.T.: A three-operator splitting scheme and its optimization applications. Set.-valued Var. Anal. **25**(4), 829–858 (2017)
11. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings. Springer Monographs in Mathematics, p. 208. Springer, Berlin (2009)
12. Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T.: Efficient projections onto the ℓ_1-ball for learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning, pp. 272–279. ACM (2008)
13. Eckstein, J., Yao, W.: Approximate ADMM algorithms derived from Lagrangian splitting. Comput. Optim. Appl. **68**(2), 363–405 (2017)
14. Eckstein, J., Yao, W.: Relative-error approximate versions of Douglas-Rachford splitting and special cases of the ADMM. Math. Program. **170**(2), 417–444 (2018)
15. Esser, E., Zhang, X.Q., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imag. Sci. **3**(4), 1015–1046 (2010)
16. Han, D.R., He, H.J., Yang, H., Yuan, X.M.: A customized Douglas-Rachford splitting algorithm for separable convex minimization with linear constraints. Numer. Math. **127**(1), 167–200 (2014)
17. Han, D.R., Sun, D.F., Zhang, L.W.: Linear rate convergence of the alternating direction method of multipliers for convex composite programming. Math. Oper. Res. **43**(2), 622–637 (2017)
18. He, B.S., Yuan, X.M.: Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imag. Sci. **5**(1), 119–149 (2012)
19. He, B.S., Ma, F., Yuan, X.M.: An algorithmic framework of generalized primal-dual hybrid gradient methods for saddle point problems. J. Math. Imag. Vis. **58**(2), 279–293 (2017)
20. Jiang, F., Cai, X.J., Wu, Z.M., Han D.R.: Approximate rst-order primal-dual algorithms for saddle point problems. Math. Comput. (2021). https://doi.org/10.1090/mcom/3610
21. Malitsky, Y., Pock, T.: A first-order primal-dual algorithm with linesearch. SIAM J. Optim. **28**(1), 411–432 (2018)
22. Möllenhoff, T., Strekalovskiy, E., Moeller, M., Daniel, C.: The primal-dual hybrid gradient method for semiconvex splittings. SIAM J. Imag. Sci. **8**(2), 827–857 (2015)
23. Morini, S., Porcelli, M., Chan, R.H.: A reduced Newton method for constrained linear least squares problems. J. Comput. Appl. Math. **233**(9), 2200–2212 (2010)
24. Nam, A.S., Davies, M.E., Elad, M., Gribonval, R.: The cosparse analysis model and algorithms. Appl. Comput. Harmon. Anal. **34**(1), 30–56 (2013)
25. Parikh, N., Boyd, S.: Proximal algorithms. Found Trends® Optim. **1**(3), 127–239 (2014)
26. Pedregosa, F., Gidel, G.: Adaptive three operator splitting. arXiv:1804.02339 (2018)
27. Rasch, J., Chambolle, A.: Inexact first-order primal-dual algorithms. Comput. Optim. Appl. **76**, 381–430 (2020). https://doi.org/10.1007/s10589-020-00186-y
28. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1-4), 227–238 (1992)
29. Robinson, S.M.: An implicit-function theorem for generalized variational inequalities. Technical Summary Report 1672, Mathematics Research Center University of Wisconsin-Madison; available from National Technical Information Service under Accession ADA031952 (1976)
30. Robinson, S.M.: Some Continuity Properties of Polyhedral Multifunctions. Mathematical Programming at Oberwolfach, pp. 206–214. Springer, Berlin (1981)

31. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer Science & Business Media, Berlin (2009)
32. Rockafellar, R.T.: Convex analysis. Princeton University Press, Princeton (2015)
33. Sun, T., Barrio, R., Cheng, L., Jiang, H.: Precompact convergence of the nonconvex primal-dual hybrid gradient algorithm. J. Comput. Appl. Math. **330**, 15–27 (2018)
34. Xie, J.X.: On inexact ADMMs with relative error criteria. Comput. Optim. Appl. **71**(3), 743–765 (2018)
35. Yan, M.: A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)
36. Zhao, T., Eldar, Y.C., Beck, A., Nehorai, A.: Smoothing and decomposition for analysis sparse recovery. IEEE Trans. Signal Process. **62**(7), 1762–1774 (2014)
37. Zhu, M.Q., Chan, T.F.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report (2008)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.