A first-order method for nonconvex-strongly-concave constrained minimax optimization*

Zhaosong Lu † Sanyou Mei ‡

May 12, 2024 (Revised: June 17, 2025; October 23, 2025)

Abstract

In this paper we study a nonconvex-strongly-concave constrained minimax problem. Specifically, we propose a first-order augmented Lagrangian method for solving it, whose subproblems are nonconvex-strongly-concave unconstrained minimax problems and suitably solved by a first-order method developed in this paper that leverages the strong concavity structure. Under suitable assumptions, the proposed method achieves an *operation complexity* of $\mathcal{O}(\varepsilon^{-3.5}\log\varepsilon^{-1})$, measured in terms of its fundamental operations, for finding an ε-KKT solution of the constrained minimax problem, which improves the previous best-known operation complexity by a factor of $\varepsilon^{-0.5}$.

Keywords: minimax optimization, augmented Lagrangian method, first-order method, operation complexity

Mathematics Subject Classification: 90C26, 90C30, 90C47, 90C99, 65K05

1 Introduction

In this paper, we consider a nonconvex-strongly-concave constrained minimax problem

$$F^* = \min_{c(x)\le 0}\max_{d(x,y)\le 0}\{F(x,y) := f(x,y) + p(x) - q(y)\}. \tag{1}$$

For notational convenience, throughout this paper we let $\mathcal{X} := \operatorname{dom} p$ and $\mathcal{Y} := \operatorname{dom} q$, where $\operatorname{dom} p$ and $\operatorname{dom} q$ denote the domain of p and q, respectively. Assume that problem (1) has at least one optimal solution and the following additional assumptions hold.

Assumption 1. *(i) f is $L_{\nabla f}$-smooth on $\mathcal{X}\times\mathcal{Y}$, and $f(x,\cdot)$ is σ-strongly-concave for some constant $\sigma > 0$ for any given $x \in \mathcal{X}$.*[1]

(ii) $p : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ and $q : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ are proper closed convex functions, and the proximal operators of p and q can be exactly evaluated.

(iii) $c : \mathbb{R}^n \to \mathbb{R}^{\tilde n}$ is $L_{\nabla c}$-smooth and L_c-Lipschitz continuous on $\mathcal{X}$, $d : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{\tilde m}$ is $L_{\nabla d}$-smooth and L_d-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$, and $d_i(x,\cdot)$ is convex for each $x \in \mathcal{X}$.

(iv) The sets $\mathcal{X}$ and $\mathcal{Y}$ (namely, $\operatorname{dom} p$ and $\operatorname{dom} q$) are compact.

Problem (1) has found application in various areas, such as perceptual adversarial robustness [27], robust adversarial classification [21], adversarial attacks in resource allocation [52], network interdiction problem [14, 48], and power networks [43].

In recent years, the minimax problem of a simpler form has gained significant attention:

$$\min_{x\in X}\max_{y\in Y} f(x;y), \tag{2}$$

*This work was partially supported by the Office of Naval Research under Award N00014-24-1-2702, the Air Force Office of Scientific Research under Award FA9550-24-1-0343, and the National Science Foundation under Award IIS-2211491. It was primarily conducted during the second author's Ph.D. studies at the University of Minnesota.

†Department of Industrial and Systems Engineering, University of Minnesota, USA (email: `zhaosong@umn.edu`).

‡Department of Industrial Engineering and Decision Analytics, the Hong Kong University of Science and Technology, Hong Kong, China (email: `symei@ust.hk`).

[1] The definition of L_F-Lipschitz continuity, $L_{\nabla f}$-smoothness and σ-strongly-concavity is given in Subsection 1.1.

where X and Y are closed sets. This problem has found wide applications in various areas, including adversarial training [18, 35, 47, 53], generative adversarial networks [15, 17, 44], reinforcement learning [8, 12, 37, 40, 49], computational game [1, 41, 50], distributed computing [36, 46], prediction and regression [4, 51, 57, 58], and distributionally robust optimization [13, 45]. Numerous methods have been developed to solve problem (2) when X and Y are *simple closed convex sets* (e.g., see [6, 20, 22, 28, 29, 31, 34, 38, 39, 55, 59, 60, 61, 63]). In addition, first-order methods were developed in [25, 64] for solving problem (1) with $c(x) \equiv 0$ and $d(x, y) \equiv 0$.

There have also been several studies on other special cases of problem (1). Specifically, in [16], two first-order methods called max-oracle gradient-descent and nested gradient descent/ascent methods were proposed for solving (1). These methods assume that $c(x) \equiv 0$ and p and q are the indicator function of simple compact convex sets X and Y, respectively. They also require the convexity of $V(x) = \max_{y \in Y}\{f(x, y) : d(x, y) \leq 0\}$, as well as the ability to compute an optimal Lagrangian multiplier associated with the constraint $d(x, y) \leq 0$ for each $x \in X$. Moreover, in [11], an augmented Lagrangian (AL) method was recently proposed for solving (1) with only equality constraints, $p(x) \equiv 0$, $q(y) \equiv 0$ and $c(x) \equiv 0$. This method assumes that a local min-max point of the AL subproblem can be found at each iteration. Furthermore, [52] introduced a multiplier gradient descent method for solving (1) with $c(x) \equiv 0$, $d(x, y)$ being an affine mapping, and p and q being the indicator function of a simple compact convex set. In addition, [9] developed a proximal gradient multi-step ascent-decent method for problem (1) with $c(x) \equiv 0$, $d(x, y)$ being an affine mapping, and $f(x, y) = g(x) + x^T Ay - h(y)$, assuming that $f(x, y) - q(y)$ is *strongly concave* in y. Furthermore, primal dual alternating proximal gradient methods were proposed in [62] for solving (1) under the conditions of $c(x) \equiv 0$, $d(x, y)$ being an affine mapping, and either $f(x, y)$ being strongly concave in y or $[q(y) \equiv 0$ and $f(x, y)$ being a linear function in $y]\}$. While the aforementioned studies [9, 16, 62] established the iteration complexity of the methods for finding an approximate stationary point of a special minimax problem, the operation complexity, measured by fundamental operations such as gradient evaluations of f and proximal operator evaluations of p and q, was not studied in these works.

Recently, a first-order augmented Lagrangian (AL) method was proposed in [32, Algorithm 3] for solving a nonconvex-concave constrained minimax problem in the form of (1) in which $f(x, \cdot)$ is however merely concave for any given $x \in \mathcal{X}$. Under suitable assumptions, this method achieves an operation complexity of $\mathcal{O}(\varepsilon^{-4} \log \varepsilon^{-1})$, measured by the amount of evaluations of ∇f, ∇c, ∇d and proximal operators of p and q, for finding an ε-KKT solution of the problem. While this method is applicable to problem (1), it does not exploit the strong concavity structure of $f(x, \cdot)$. Consequently, it may not be the most efficient method for solving (1).

In this paper, we propose a first-order AL method for solving problem (1). Our approach follows a similar framework as [32, Algorithm 3], but we enhance it by leveraging the strong concavity of $f(x, \cdot)$. As a result, our method achieves a substantially improved operation complexity compared to [32, Algorithm 3]. Specifically, given an iterate (x^k, y^k) and a Lagrangian multiplier estimate $(\lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k)$ at the kth iteration, the next iterate (x^{k+1}, y^{k+1}) of our method is obtained by finding an approximate stationary point of the AL subproblem

$$\min_x \max_y \mathcal{L}(x, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \tag{3}$$

for some $\rho_k > 0$, where $\mathcal{L}$ is the AL function of (1) defined as

$$\mathcal{L}(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}; \rho) = F(x, y) + \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{x}} + \rho c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}\|^2\right) - \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{y}} + \rho d(x, y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right), \tag{4}$$

which is a generalization of the AL function introduced in [11] for an equality constrained minimax problem. The Lagrangian multiplier estimate is then updated by $\lambda_{\mathbf{x}}^{k+1} = \Pi_{\mathbb{B}_\Lambda^+}(\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1}))$ and $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+$ for some $\Lambda > 0$, where $\Pi_{\mathbb{B}_\Lambda^+}(\cdot)$ and $[\cdot]_+$ are defined in Subsection 1.1. Given that problem (3) is a nonconvex-strongly-concave unconstrained minimax problem, we develop an efficient first-order method for finding an approximate stationary point of it by utilizing its strong concavity structure.

The main contributions of this paper are summarized below.

- We propose a first-order method for solving a nonconvex-strongly-concave unconstrained minimax problem. Under suitable assumptions, we show that this method achieves an operation complexity of $\mathcal{O}(\varepsilon^{-2} \log \varepsilon^{-1})$, measured by its fundamental operations, for finding an ε-primal-dual stationary point of the problem, which improves the previous best-known operation complexity achieved by [32, Algorithm 1] by a factor of $\varepsilon^{-0.5}$.

- We propose a first-order AL method for solving nonconvex-strongly-concave constrained minimax problem (1). Under suitable assumptions, we show that this method achieves an operation complexity of $\mathcal{O}(\varepsilon^{-3.5}\log\varepsilon^{-1})$, measured by its fundamental operations, for finding an ε-KKT solution of (1), which improves the previous best-known operation complexity achieved by [32, Algorithm 3] by a factor of $\varepsilon^{-0.5}$.

The rest of this paper is organized as follows. In Subsection 1.1, we introduce some notation and terminology. In Section 2, we propose a first-order method for solving a nonconvex-concave minimax problem and study its complexity. In Section 3, we propose a first-order AL method for solving problem (1) and present complexity results for it. Finally, we provide the proof of the main results in Section 5.

1.1 Notation and terminology

The following notation will be used throughout this paper. Let $\mathbb{R}^n$ denote the Euclidean space of dimension n and $\mathbb{R}^n_+$ denote the nonnegative orthant in $\mathbb{R}^n$. The standard inner product, l_1-norm and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$, $\|\cdot\|_1$ and $\|\cdot\|$, respectively. For any $\Lambda>0$, let $\mathbb{B}^+_\Lambda=\{x\ge 0:\|x\|\le\Lambda\}$, whose dimension is clear from the context. For any $v\in\mathbb{R}^n$, let v_+ denote the nonnegative part of v, that is, $(v_+)_i=\max\{v_i,0\}$ for all i. Given a point x and a closed set S in $\mathbb{R}^n$, let $\mathrm{dist}(x,S)=\min_{x'\in S}\|x'-x\|$, $\Pi_S(x)$ denote the Euclidean projection of x onto S, and $\mathscr{I}_S$ denote the indicator function associated with S.

A function or mapping ϕ is said to be L_ϕ-*Lipschitz continuous* on a set S if $\|\phi(x)-\phi(x')\|\le L_\phi\|x-x'\|$ for all $x,x'\in S$. In addition, it is said to be $L_{\nabla\phi}$-*smooth* on S if $\|\nabla\phi(x)-\nabla\phi(x')\|\le L_{\nabla\phi}\|x-x'\|$ for all $x,x'\in S$. A function is said to be σ-strongly-convex if it is strongly convex with modulus $\sigma>0$. For a closed convex function $p:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$, the *proximal operator* associated with p is denoted by prox_p, that is,

$$\mathrm{prox}_p(x)=\arg\min_{x'\in\mathbb{R}^n}\left\{\frac{1}{2}\|x'-x\|^2+p(x')\right\}\quad\forall x\in\mathbb{R}^n.$$

Given that evaluation of $\mathrm{prox}_{\gamma p}(x)$ is often as cheap as $\mathrm{prox}_p(x)$, we count the evaluation of $\mathrm{prox}_{\gamma p}(x)$ as one evaluation of proximal operator of p for any $\gamma>0$ and $x\in\mathbb{R}^n$.

For a lower semicontinuous function $\phi:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$, its *domain* is the set $\mathrm{dom}\,\phi:=\{x|\phi(x)<+\infty\}$. The *upper subderivative* of ϕ at $x\in\mathrm{dom}\,\phi$ in a direction $d\in\mathbb{R}^n$ is defined by

$$\phi'(x;d)=\limsup_{x'\xrightarrow{\phi}x,\,t\downarrow 0}\ \inf_{d'\to d}\frac{\phi(x'+td')-\phi(x')}{t},$$

where $t\downarrow 0$ means both $t>0$ and $t\to 0$, and $x'\xrightarrow{\phi}x$ means both $x'\to x$ and $\phi(x')\to\phi(x)$. The *subdifferential* of ϕ at $x\in\mathrm{dom}\,\phi$ is the set

$$\partial\phi(x)=\{s\in\mathbb{R}^n\,|\,s^Td\le\phi'(x;d)\ \ \forall d\in\mathbb{R}^n\}.$$

We use $\partial_{x_i}\phi(x)$ to denote the subdifferential with respect to x_i. In addition, for an upper semicontinuous function ϕ, its subdifferential is defined as $\partial\phi=-\partial(-\phi)$. If ϕ is locally Lipschitz continuous, the above definition of subdifferential coincides with the Clarke subdifferential. Besides, if ϕ is convex, it coincides with the ordinary subdifferential for convex functions. Also, if ϕ is continuously differentiable at x , we simply have $\partial\phi(x)=\{\nabla\phi(x)\}$, where $\nabla\phi(x)$ is the gradient of ϕ at x. In addition, it is not hard to verify that $\partial(\phi_1+\phi_2)(x)=\nabla\phi_1(x)+\partial\phi_2(x)$ if ϕ_1 is continuously differentiable at x and ϕ_2 is lower or upper semicontinuous at x. See [7, 54] for more details.

Finally, we introduce an (approximate) primal-dual stationary point (e.g., see [9, 10, 25]) for a general minimax problem

$$\min_x\max_y\Psi(x,y),\tag{5}$$

where $\Psi(\cdot,y):\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous function, and $\Psi(x,\cdot):\mathbb{R}^m\to\mathbb{R}\cup\{-\infty\}$ is an upper semicontinuous function.

Definition 1. *A point (x,y) is said to be a primal-dual stationary point of the minimax problem* (5) *if*

$$0\in\partial_x\Psi(x,y),\quad 0\in\partial_y\Psi(x,y).$$

In addition, for any $\epsilon>0$, a point (x_ϵ,y_ϵ) is said to be an ϵ-primal-dual stationary point of the minimax problem (5) *if*

$$\mathrm{dist}\left(0,\partial_x\Psi(x_\epsilon,y_\epsilon)\right)\le\epsilon,\quad\mathrm{dist}\left(0,\partial_y\Psi(x_\epsilon,y_\epsilon)\right)\le\epsilon.$$

One can see that (x_ϵ,y_ϵ) is an ϵ-primal-dual stationary point of (5) if and only if x_ϵ and y_ϵ are an ϵ-stationary point of $\min_x\Psi(x,y_\epsilon)$ and $\max_y\Psi(x_\epsilon,y)$, respectively.

2 A first-order method for nonconvex-strongly-concave unconstrained minimax optimization

In this section, we propose a first-order method for finding an ϵ-primal-dual stationary point of a nonconvex-strongly-concave unconstrained minimax problem, which will be used as a subproblem solver for the first-order AL method proposed in Section 3. In particular, we consider a nonconvex-strongly-concave minimax problem

$$H^* = \min_x \max_y \{H(x,y) := h(x,y) + p(x) - q(y)\}. \tag{6}$$

Assume that problem (6) has at least one optimal solution and p, q satisfy Assumption 1. In addition, h satisfies the following assumption.

Assumption 2. *The function h is $L_{\nabla h}$-smooth on $\mathcal{X} \times \mathcal{Y}$, and moreover, $h(x,\cdot)$ is σ_y-strongly-concave for some constant $\sigma_y > 0$ for all $x \in \mathcal{X}$, where $\mathcal{X} := \mathrm{dom}\, p$ and $\mathcal{Y} := \mathrm{dom}\, q$.*

Several first-order methods have been developed for special classes of (6) with p, q being the indicator function of convex compact sets or entire spaces, and they enjoy an operation complexity of $\mathcal{O}(\epsilon^{-2}\log\epsilon^{-1})$, measured by the amount of evaluations of ∇h and proximal operators of p and q, for finding an ϵ-primal-dual stationary point of (6) with such p and q (e.g., see [29, 61]). They are however not applicable to (6) in general.

We now propose a first-order method for problem (6) by solving a sequence of subproblems

$$\min_x \max_y \{H_k(x,y) := h_k(x,y) + p(x) - q(y)\}, \tag{7}$$

which result from applying an inexact proximal point method [24] to the minimization problem $\min_x\{\max_y h(x,y) + p(x) - q(y)\}$, where

$$h_k(x,y) = h(x,y) + L_{\nabla h}\|x - x^k\|^2, \tag{8}$$

and x^k is an approximate x-solution of (7) with k replaced by $k-1$. By Assumption 2, one can observe that (i) h_k is $L_{\nabla h}$-strongly convex in x and σ_y-strongly concave in y on $\mathrm{dom}\, p \times \mathrm{dom}\, q$; (ii) h_k is $3L_{\nabla h}$-smooth on $\mathrm{dom}\, p \times \mathrm{dom}\, q$. Consequently, problem (7) is a special case of (90) and can be suitably solved by Algorithm 3 (see Appendix A). The resulting first-order method for (6) is presented in Algorithm 1.

Algorithm 1 A first-order method for problem (6)

Input: $\epsilon > 0$, $\hat\epsilon_0 \in (0, \epsilon/2]$, $(\hat x^0, \hat y^0) \in \mathrm{dom}\, p \times \mathrm{dom}\, q$, $(x^0, y^0) = (\hat x^0, \hat y^0)$, and $\hat\epsilon_k = \hat\epsilon_0/(k+1)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Call Algorithm 3 (see Appendix A) with $\bar h \leftarrow h_k$, $\bar\epsilon \leftarrow \hat\epsilon_k$, $\sigma_x \leftarrow L_{\nabla h}$, $\sigma_y \leftarrow \sigma_y$, $L_{\nabla \bar h} \leftarrow 3L_{\nabla h}$, $\bar z^0 = z_f^0 \leftarrow -\sigma_x x^k$, $\bar y^0 = y_f^0 \leftarrow y^k$, and denote its output by (x^{k+1}, y^{k+1}), where h_k is given in (8).

3: Terminate the algorithm and output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ if

$$\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h}). \tag{9}$$

4: **end for**

Remark 1. *It is seen from step 2 of Algorithm 1 that (x^{k+1}, y^{k+1}) results from applying Algorithm 3 to the subproblem (7). As will be shown in Lemma 1, (x^{k+1}, y^{k+1}) is an $\hat\epsilon_k$-primal-dual stationary point of (7).*

We next study complexity of Algorithm 1 for finding an ϵ-primal-dual stationary point of problem (6). Before proceeding, we define

$$D_{\mathbf{x}} := \max\{\|u - v\| \,|\, u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| \,|\, u, v \in \mathcal{Y}\}, \tag{10}$$

$$H_{\mathrm{low}} := \min\{H(x,y) | (x,y) \in \mathrm{dom}\, p \times \mathrm{dom}\, q\}. \tag{11}$$

By Assumption 1, one can observe that H_{low} is finite.

The following theorem presents *iteration and operation complexity* of Algorithm 1 for finding an ϵ-primal-dual stationary point of problem (6), whose proof is deferred to Subsection 5.1.

Theorem 1 (Complexity of Algorithm 1). *Suppose that Assumption 2 holds. Let H^*, H $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, and H_{low} be defined in* (6), (10) *and* (11), *$L_{\nabla h}$ be given in Assumption 2, ϵ, $\hat{\epsilon}_0$ and $\hat{x}^0$ be given in Algorithm 1, and*

$$\hat{\alpha} = \min\left\{1, \sqrt{8\sigma_y/L_{\nabla h}}\right\}, \tag{12}$$

$$\hat{\delta} = (2+\hat{\alpha}^{-1})L_{\nabla h}D_{\mathbf{x}}^2 + \max\left\{2\sigma_y, \hat{\alpha}L_{\nabla h}/4\right\}D_{\mathbf{y}}^2, \tag{13}$$

$$\widehat{T} = \left\lceil 16(\max_y H(\hat{x}^0,y) - H^*)L_{\nabla h}\epsilon^{-2} + 32\hat{\epsilon}_0^2(1+\sigma_y^{-2}L_{\nabla h}^2)\epsilon^{-2} - 1\right\rceil_+, \tag{14}$$

$$\begin{aligned}\widehat{N} = {} & 3397\max\left\{2, \sqrt{L_{\nabla h}/(2\sigma_y)}\right\} \\ & \times \left[(\widehat{T}+1)\left(\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\text{low}} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{(9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\} + 3L_{\nabla h})^{-2}\hat{\epsilon}_0^2}\right)_+ \right. \\ & \left. + \widehat{T} + 1 + 2\widehat{T}\log(\widehat{T}+1)\right]. \end{aligned} \tag{15}$$

Then Algorithm 1 terminates and outputs an ϵ-primal-dual stationary point (x_ϵ, y_ϵ) of (6) *in at most $\widehat{T}+1$ outer iterations that satisfies*

$$\max_y H(x_\epsilon, y) \le \max_y H(\hat{x}^0, y) + 2\hat{\epsilon}_0^2\left(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h}\right). \tag{16}$$

Moreover, the total number of evaluations of ∇h and proximal operators of p and q performed in Algorithm 1 is no more than $\widehat{N}$, respectively.

Remark 2. *One can observe from Theorem 1 that $\hat{\alpha} = \mathcal{O}(\kappa^{-1/2})$, $\hat{\delta} = \mathcal{O}(\kappa^{1/2})$, $\widehat{T} = \mathcal{O}(\epsilon^{-2})$, and $\widehat{N} = \mathcal{O}(\kappa^{1/2}\epsilon^{-2}\log\hat{\epsilon}_0^{-1})$, where $\kappa = L_{\nabla h}/\sigma_y$ is the condition number of the maximization part. Consequently, by setting $\hat{\epsilon}_0 = \epsilon/2$, Algorithm 1 achieves an operation complexity of $\mathcal{O}(\kappa^{1/2}\epsilon^{-2}\log\epsilon^{-1})$, measured by the number of evaluations of ∇h and the proximal operators of p and q, for computing an ϵ-primal-dual stationary point of the nonconvex–strongly-concave minimax problem* (6)*. This improves the best-known complexity bound previously obtained by [32, Algorithm 1] by a factor of $\epsilon^{-1/2}$. In addition, an alternating gradient projection (AGP) method was recently proposed in [60] for a subclass of unconstrained minimax problems of the form* (6)*, specifically those where p and q are indicator functions of convex compact sets. A complexity bound is established for AGP in terms of the norm of a gradient mapping, which has slightly better dependence on ϵ (up to a logarithmic factor) than our result. However, it has significantly worse dependence on the condition number κ due to the lack of an acceleration scheme in AGP.*

3 A first-order augmented Lagrangian method for nonconvex-strongly-concave constrained minimax optimization

In this section, we propose a first-order augmented Lagrangian (FAL) method in Algorithm 2 for problem (1), and study its complexity for finding an approximate KKT point of (1). The proposed FAL method follows a similar framework as [32, Algorithm 3]. Specifically, at each iteration, the FAL method finds an approximate primal-dual stationary point of an AL subproblem in the form of

$$\min_x \max_y \mathcal{L}(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}; \rho), \tag{17}$$

where $\mathcal{L}$ is the AL function associated with problem (1) defined in (4), $\lambda_{\mathbf{x}} \in \mathbb{R}_+^{\tilde{n}}$ and $\lambda_{\mathbf{y}} \in \mathbb{R}_+^{\tilde{m}}$ are Lagrangian multiplier estimates, and $\rho > 0$ is a penalty parameter, which are updated by a standard scheme. By Assumption 1, it is not hard to observe that (17) is a special case of nonconvex-strongly-concave unconstrained minimax problem (6). Consequently, our FAL method applies Algorithm 1 to find an approximate primal-dual stationary point of (17).

Before presenting the FAL method for (1), we let

$$\begin{aligned} & \mathcal{L}_{\mathbf{x}}(x, y, \lambda_{\mathbf{x}}; \rho) := F(x,y) + \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{x}} + \rho c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}\|^2\right), \\ & c_{\text{hi}} := \max\{\|c(x)\| \big| x \in \mathcal{X}\}, \quad d_{\text{hi}} := \max\{\|d(x,y)\| \big| (x,y) \in \mathcal{X}\times\mathcal{Y}\}, \end{aligned} \tag{18}$$

where $\mathcal{L}_{\mathbf{x}}(\cdot, y, \lambda_{\mathbf{x}}; \rho)$ can be viewed as the AL function for the minimization part of (1), namely, the problem $\min_x\{F(x,y)|c(x) \le 0\}$ for any $y \in \mathcal{Y}$. Besides, we make one additional assumption below regarding the availability of a nearly feasible point for the minimization part of (1). Given the possible nonconvexity of c_i's, it will be used to specify an initial point for solving the AL subproblems (see step 2 of Algorithm 2) so that the resulting FAL method outputs an approximate KKT point of (1) nearly satisfying the constraint $c(x) \le 0$.

Assumption 3. *For any given $\varepsilon \in (0,1)$, a $\sqrt{\varepsilon}$-nearly feasible point $x_{\mathbf{nf}}$ of problem* (1)*, namely $x_{\mathbf{nf}} \in \mathcal{X}$ satisfying $\|[c(x_{\mathbf{nf}})]_+\| \le \sqrt{\varepsilon}$, can be found.*

Remark 3. *A very similar assumption as Assumption 3 was considered in [5, 19, 32, 33, 56]. In addition, when the error bound condition $\|[c(x)]_+\| = \mathcal{O}(\mathrm{dist}(0, \partial(\|[c(x)]_+\|^2 + \mathscr{I}_{\mathcal{X}}(x)))^\nu)$ holds on a level set of $\|[c(x)]_+\|$ for some $\nu > 0$, Assumption 3 holds for problem* (1) *(e.g., see [30, 42]). In this case, one can find the above $x_{\mathbf{nf}}$ by applying a projected gradient method to the problem $\min_{x\in\mathcal{X}} \|[c(x)]_+\|^2$.*

We are now ready to present the aforementioned FAL method for solving problem (1).

Algorithm 2 A first-order augmented Lagrangian method for problem (1)

Input: $\varepsilon, \tau \in (0,1)$, $\epsilon_k = \tau^k$, $\rho_k = \epsilon_k^{-1}$, $\Lambda > 0$, $\lambda_{\mathbf{x}}^0 \in \mathbb{B}_\Lambda^+$, $\lambda_{\mathbf{y}}^0 \in \mathbb{R}_+^{\tilde{m}}$, $(x^0, y^0) \in \mathrm{dom}\, p \times \mathrm{dom}\, q$, and $x_{\mathbf{nf}} \in \mathrm{dom}\, p$ with $\|[c(x_{\mathbf{nf}})]_+\| \le \sqrt{\varepsilon}$.

1: **for** $k = 0, 1, \ldots$ **do**

2: Set

$$x_{\mathrm{init}}^k = \begin{cases} x^k, & \text{if } \mathcal{L}_{\mathbf{x}}(x^k, y^k, \lambda_{\mathbf{x}}^k; \rho_k) \le \mathcal{L}_{\mathbf{x}}(x_{\mathbf{nf}}, y^k, \lambda_{\mathbf{x}}^k; \rho_k), \\ x_{\mathbf{nf}}, & \text{otherwise.} \end{cases}$$

3: Call Algorithm 1 with $\epsilon \leftarrow \epsilon_k$, $\hat{\epsilon}_0 \leftarrow \epsilon_k/2$, $(x^0, y^0) \leftarrow (x_{\mathrm{init}}^k, y^k)$, $\sigma_y \leftarrow \sigma$ and $L_{\nabla h} \leftarrow L_k$ to find an ϵ_k-primal-dual stationary point (x^{k+1}, y^{k+1}) of

$$\min_x \max_y \mathcal{L}(x, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \tag{19}$$

where

$$L_k = L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\mathrm{hi}} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\mathrm{hi}} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d}. \tag{20}$$

4: Set $\lambda_{\mathbf{x}}^{k+1} = \Pi_{\mathbb{B}_\Lambda^+}(\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1}))$ and $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+$.

5: If $\epsilon_k \le \varepsilon$, terminate the algorithm and output (x^{k+1}, y^{k+1}).

6: **end for**

Remark 4. *(i) $\lambda_{\mathbf{x}}^{k+1}$ results from projecting onto a nonnegative Euclidean ball the standard Lagrangian multiplier estimate $\tilde{\lambda}_{\mathbf{x}}^{k+1}$ obtained by the classical scheme $\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+$. It is called a safeguarded Lagrangian multiplier in the relevant literature [2, 3, 23], which has been shown to enjoy many practical and theoretical advantages (see [2] for discussions).*

(ii) In view of Theorem 1, one can see that an ϵ_k-primal-dual stationary point of (19) *can be successfully found in step 3 of Algorithm 2 by applying Algorithm 1 to problem* (19)*. Consequently, Algorithm 2 is well-defined.*

In the remainder of this section, we study iteration and operation complexity for Algorithm 2. Recall that $\mathcal{X} = \mathrm{dom}\, p$ and $\mathcal{Y} = \mathrm{dom}\, q$. To proceed, we make one additional assumption that a generalized Mangasarian-Fromowitz constraint qualification (GMFCQ) holds for the minimization part of (1), a uniform Slater's condition holds for the maximization part of (1), and $F(\cdot, y)$ is Lipschitz continuous on $\mathcal{X}$ for any $y \in \mathcal{Y}$. Specifically, GMFCQ and the Lipschitz continuity of $F(\cdot, y)$ will be used to bound the amount of violation on feasibility and complementary slackness by $(x^{k+1}, \tilde{\lambda}_{\mathbf{x}}^{k+1})$ for the minimization part of (1) with $\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+$ (see Lemma 8). Likewise, the uniform Slater's condition will be used to bound the amount of violation on feasibility and complementary slackness by $(x^{k+1}, y^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ for the maximization part of (1) (see Lemmas 4 and 5).

Assumption 4. *(i) There exist some constants δ_c, $\theta > 0$ such that for each $x \in \mathcal{F}(\theta)$ there exists some $v_x \in \mathcal{T}_{\mathcal{X}}(x)$ satisfying $\|v_x\| = 1$ and $v_x^T \nabla c_i(x) \le -\delta_c$ for all $i \in \mathcal{A}(x; \theta)$, where $\mathcal{T}_{\mathcal{X}}(x)$ is the tangent cone of $\mathcal{X}$ at x, and*

$$\mathcal{F}(\theta) = \{x \in \mathcal{X} \big| \|[c(x)]_+\| \le \theta\}, \quad \mathcal{A}(x; \theta) = \{i | c_i(x) \ge -\theta,\ 1 \le i \le \tilde{n}\}. \tag{21}$$

(ii) For each $x \in \mathcal{X}$, there exists some $\hat{y}_x \in \mathcal{Y}$ such that $d_i(x, \hat{y}_x) < 0$ for all $i = 1, 2, \ldots, \tilde{m}$, and moreover, $\delta_d := \inf\{-d_i(x, \hat{y}_x) | x \in \mathcal{X},\ i = 1, 2, \ldots, \tilde{m}\} > 0$.

(iii) $F(\cdot, y)$ is L_F-Lipschitz continuous on $\mathcal{X}$ for any $y \in \mathcal{Y}$.

Remark 5. *(i) Assumption 4(i) can be viewed as a robust counterpart of MFCQ. It implies that MFCQ holds for all the minimization problems, resulting from the minimization part of* (1) *by fixing $y \in \mathcal{Y}$ and perturbing $c_i(x)$ at most by θ.*

(ii) The latter part of Assumption 4(ii) can be weakened to the one that the pointwise Slater's condition holds for the constraint on y in (1)*, that is, there exists $\hat{y}_x \in \mathcal{Y}$ such that $d(x, \hat{y}_x) < 0$ for each $x \in \mathcal{X}$. Indeed, if $\delta_d > 0$, Assumption 4(ii) holds. Otherwise, one can solve the perturbed counterpart of* (1) *with $d(x, y)$ being replaced by $d(x, y) - \epsilon$ for some suitable $\epsilon > 0$ instead, which satisfies Assumption 4(ii).*

(iii) In view of Assumption 1, one can observe that if p is Lipschitz continuous on $\mathcal{X}$, $F(\cdot, y)$ is Lipschitz continuous on $\mathcal{X}$ for any $y \in \mathcal{Y}$. Thus, Assumption 4(iii) is mild.

In addition, to characterize the approximate solution found by Algorithm 2, we review a notion so-called an ε-KKT solution of problem (1), which was introduced in [32, Definition 2].

Definition 2. *For any $\varepsilon > 0$, (x, y) is said to be an ε-KKT point of problem* (1) *if there exists $(\lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}) \in \mathbb{R}^{\tilde{n}}_+ \times \mathbb{R}^{\tilde{m}}_+$ such that*

$$
\begin{aligned}
&\mathrm{dist}(0, \partial_x F(x, y) + \nabla c(x)\lambda_{\mathbf{x}} - \nabla_x d(x, y)\lambda_{\mathbf{y}}) \le \varepsilon, \\
&\mathrm{dist}(0, \partial_y F(x, y) - \nabla_y d(x, y)\lambda_{\mathbf{y}}) \le \varepsilon, \\
&\|[c(x)]_+\| \le \varepsilon, \quad |\langle \lambda_{\mathbf{x}}, c(x)\rangle| \le \varepsilon, \\
&\|[d(x, y)]_+\| \le \varepsilon, \quad |\langle \lambda_{\mathbf{y}}, d(x, y)\rangle| \le \varepsilon.
\end{aligned}
$$

Recall that $\mathcal{X} = \mathrm{dom}\, p$ and $\mathcal{Y} = \mathrm{dom}\, q$. To study complexity of Algorithm 2, we define

$$
\begin{aligned}
&f^*(x) := \max\{F(x, y) | d(x, y) \le 0\}, && (22)\\
&F_{\mathrm{hi}} := \max\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \quad F_{\mathrm{low}} := \min\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, && (23)\\
&\Delta := F_{\mathrm{hi}} - F_{\mathrm{low}}, \quad r := 2\delta_d^{-1}\Delta, && (24)\\
&K := \lceil \log\varepsilon / \log\tau \rceil_+, \quad \mathbb{K} := \{0, 1, \ldots, K+1\}, && (25)
\end{aligned}
$$

where δ_d is given in Assumption 4, and ε and τ are some input parameters of Algorithm 2. For convenience, we define $\mathbb{K} - 1 = \{k - 1 | k \in \mathbb{K}\}$. One can observe from Assumption 1 that F_{hi} and F_{low} are finite. Besides, one can easily observe that

$$
f^*(x) \ge F_{\mathrm{low}}, \ F(x, y) - f^*(x) \le \Delta \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{26}
$$

We are now ready to present an *iteration and operation complexity* of Algorithm 2 for finding an $\mathcal{O}(\varepsilon)$-KKT solution of problem (1), whose proof is deferred to Section 5.

Theorem 2. *Suppose that Assumptions 1, 3 and 4 hold. Let $\{(x^k, y^k, \lambda^k_{\mathbf{x}}, \lambda^k_{\mathbf{y}})\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, c_{hi}, d_{hi}, Δ and K be defined in* (10), (18), (24) *and* (25)*, L_F, $L_{\nabla f}$, $L_{\nabla d}$, $L_{\nabla c}$, L_c, $L_{\nabla d}$, L_d, δ_c, δ_d and θ be given in Assumptions 1 and 4, ε, τ, Λ and $\lambda^0_{\mathbf{y}}$ be given in Algorithm 2, and*

$$
L = L_{\nabla f} + L_c^2 + c_{\mathrm{hi}} L_{\nabla c} + \Lambda L_{\nabla c} + L_d^2 + d_{\mathrm{hi}} L_{\nabla d} + L_{\nabla d}\sqrt{\|\lambda^0_{\mathbf{y}}\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}}, \tag{27}
$$

$$
\alpha = \min\left\{1, \sqrt{8\sigma/L}\right\}, \quad \delta = (2 + \alpha^{-1}) L D_{\mathbf{x}}^2 + \max\{2\sigma, L/4\} D_{\mathbf{y}}^2, \tag{28}
$$

$$
\begin{aligned}
M = {} & 16\max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}\left[81/\min\{L_c^2, \sigma\} + 3L\right]^2 \\
& \times \left(\delta + 2\alpha^{-1}\left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda^0_{\mathbf{y}}\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1-\tau} + \rho_k d_{\mathrm{hi}}^2 + L D_{\mathbf{x}}^2\right)\right),
\end{aligned} \tag{29}
$$

$$
T = \left\lceil 16\left(2\Delta + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda^0_{\mathbf{y}}\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} + \frac{\Lambda^2}{2}\right) L + 8(1 + \sigma^{-2}L^2)\right\rceil_+, \tag{30}
$$

$$
\tilde{\lambda}^{K+1}_{\mathbf{x}} = [\lambda^K_{\mathbf{x}} + c(x^{K+1})/\tau^K]_+. \tag{31}
$$

Suppose that

$$\varepsilon^{-1} \geq \max\Big\{1, \theta^{-1}\Lambda, \theta^{-2}\Big\{4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}$$
$$+ L_c^{-2} + \sigma^{-2}L + \Lambda^2\Big\}, \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2\tau(1-\tau)}\Big\}. \tag{32}$$

Then the following statements hold.

(i) Algorithm 2 terminates after $K+1$ outer iterations and outputs an approximate stationary point (x^{K+1}, y^{K+1}) of (1) *satisfying*

$$\mathrm{dist}(0, \partial_x F(x^{K+1}, y^{K+1}) + \nabla c(x^{K+1})\tilde{\lambda}_x^{K+1} - \nabla_x d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}) \leq \varepsilon, \tag{33}$$
$$\mathrm{dist}\left(0, \partial_y F(x^{K+1}, y^{K+1}) - \nabla_y d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\right) \leq \varepsilon, \tag{34}$$
$$\|[c(x^{K+1})]_+\| \leq \varepsilon\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1\right), \tag{35}$$
$$|\langle\tilde{\lambda}_{\mathbf{x}}^{K+1}, c(x^{K+1})\rangle| \leq \varepsilon\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1)$$
$$\times \max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1), \Lambda\}, \tag{36}$$
$$\|[d(x^{K+1}, y^{K+1})]_+\| \leq 2\varepsilon\delta_d^{-1}(\Delta + D_{\mathbf{y}}), \tag{37}$$
$$|\langle\lambda_{\mathbf{y}}^{K+1}, d(x^{K+1}, y^{K+1})\rangle| \leq 2\varepsilon\delta_d^{-1}(\Delta + D_{\mathbf{y}})\max\{2\delta_d^{-1}(\Delta + D_{\mathbf{y}}), \|\lambda_{\mathbf{y}}^0\|\}. \tag{38}$$

(ii) The total number of evaluations of ∇f, ∇c, ∇d and proximal operators of p and q performed in Algorithm 2 is at most N, respectively, where

$$N = 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}T(1-\tau^{7/2})^{-1}$$
$$\times (\tau\varepsilon)^{-7/2}\left(20K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right). \tag{39}$$

Remark 6. *(i) The condition* (32) *on ε is to ensure that the final penalty parameter ρ_K in Algorithm 2 is large enough so that feasibility and complementarity slackness are nearly satisfied at $(x^{K+1}, y^{K+1}, \tilde{\lambda}_{\mathbf{x}}^{K+1}, \lambda_{\mathbf{y}}^{K+1})$.*

(ii) One can observe from Theorem 2 that Algorithm 2 enjoys an iteration complexity of $\mathcal{O}(\log\varepsilon^{-1})$ and an operation complexity of $\mathcal{O}(\varepsilon^{-3.5}\log\varepsilon^{-1})$, measured by the amount of evaluations of ∇f, ∇c, ∇d and proximal operators of p and q, for finding an $\mathcal{O}(\varepsilon)$-KKT solution (x^{K+1}, y^{K+1}) of (1) *such that*

$$\mathrm{dist}\left(\partial_x F(x^{K+1}, y^{K+1}) + \nabla c(x^{K+1})\tilde{\lambda}_{\mathbf{x}} - \nabla_x d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\right) \leq \varepsilon,$$
$$\mathrm{dist}\left(\partial_y F(x^{K+1}, y^{K+1}) - \nabla_y d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\right) \leq \varepsilon,$$
$$\|[c(x^{K+1})]_+\| = \mathcal{O}(\varepsilon), \quad |\langle\tilde{\lambda}_{\mathbf{x}}^{K+1}, c(x^{K+1})\rangle| = \mathcal{O}(\varepsilon),$$
$$\|[d(x^{K+1}, y^{K+1})]_+\| = \mathcal{O}(\varepsilon), \quad |\langle\lambda_{\mathbf{y}}^{K+1}, d(x^{K+1}, y^{K+1})\rangle| = \mathcal{O}(\varepsilon),$$

where $\tilde{\lambda}_{\mathbf{x}}^{K+1} \in \mathbb{R}_+^{\tilde{n}}$ is defined in (31) *and $\lambda_{\mathbf{y}}^{K+1} \in \mathbb{R}_+^{\tilde{m}}$ is given in Algorithm 2.*

(iii) It shall be mentioned that an $\mathcal{O}(\varepsilon)$-KKT solution of (1) *can be found by [32, Algorithm 3] with an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ (see [32, Theorem 3]). As a result, the operation complexity of Algorithm 2 improves that of [32, Algorithm 3] by a factor of $\epsilon^{-1/2}$.*

4 Numerical results

In this section, we conduct some preliminary experiments to test the performance of our proposed method (namely, Algorithms 1 and 2), and compare them with an alternating gradient projection method (AGP) [60, Algorithm 1] and an augmented Lagrangian method (ALM) [32, Algorithm 3], respectively. All the algorithms are coded in Matlab, and all the computations are performed on a laptop with a 2.30 GHz Intel i9-9880H 8-core processor and 16 GB of RAM.

4.1 Unconstrained nonconvex-strongly-concave minimax optimization with quadratic objective

In this subsection, we consider the problem

$$\min_x \max_y \; x^T Ax + x^T By + y^T Cy + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) - \mathscr{I}_{[-1,1]^m}(y), \tag{40}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times m}$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}^m$, and $\mathscr{I}_{[-1,1]^n}(\cdot)$ and $\mathscr{I}_{[-1,1]^m}(\cdot)$ are the indicator functions of $[-1,1]^n$ and $[-1,1]^m$ respectively.

For each pair (n,m), we randomly generate 10 instances of problem (40). Specifically, we construct $A = UDU^T$, where $U = \mathrm{orth}(\mathrm{randn}(n))$, and D is a diagonal matrix with entries independently drawn from a normal distribution with mean 0 and standard deviation 0.1. Matrix C is generated in a similar manner, except the diagonal entries of the corresponding matrix are drawn independently from a uniform distribution over $[2,3]$. In addition, we randomly generate vectors c and d with all the entries independently drawn from a normal distribution with mean 0 and standard deviation 0.1.

Notice that (40) is a special case of (6) with $h(x,y) = x^T Ax + x^T By + y^T Cy + c^T x + d^T y$, $p(x) = \mathscr{I}_{[-1,1]^n}(x)$, and $q(y) = \mathscr{I}_{[-1,1]^m}(y)$ and can be suitably solved by Algorithm 1 and AGP [60, Algorithm 1]. In addition, problem (40) is equivalent to as the following minimization problem

$$\min_x \Phi(x), \tag{41}$$

where Φ is the hyper-objective function defined as

$$\Phi(x) = \max_y x^T Ax + x^T By + y^T Cy + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) - \mathscr{I}_{[-1,1]^m}(y).$$

For Algorithm 1, we set the parameters to $(\epsilon, \hat{\epsilon}_0) = (10^{-2}, 5\times 10^{-3})$. For AGP, we use the parameter settings as specified in [60, Subsection 3.1]. Both algorithms are initialized with the all-one vector. Each algorithm is terminated once a 10^{-2}-primal-dual stationary point (x^k, y^k) of (40) is found for some k, and the pair (x^k, y^k) is returned as an approximate solution to (40).

The computational results of the aforementioned algorithms on the randomly generated instances are presented in Table 1. Specifically, the values of n and m are listed in the first two columns. For each pair (n,m), the average initial hyper-objective value $\Phi(x^0)$, the average final hyper-objective value $\Phi(x^k)$, and the average CPU time (in seconds) over 10 random instances are reported in the remaining columns. It can be observed that both Algorithm 1 and AGP [60, Algorithm 1] yield approximate solutions with comparable hyper-objective values, which are significantly lower than the initial value. However, Algorithm 1 consistently achieves significantly lower CPU times, which may be attributed to its more favorable dependence on condition numbers.

		Initial hyper-objective value	Final hyper-objective value		CPU time (seconds)	
n	m		Algorithm 1	AGP	Algorithm 1	AGP
50	50	4.30	−0.30	−0.29	19.3	100.0
100	100	10.34	−1.13	−1.10	82.6	428.6
150	150	22.16	−1.01	−1.09	176.5	910.3
200	200	32.52	−1.43	−1.39	222.6	1141.1
250	250	69.19	−1.80	−1.83	312.7	1219.1
300	300	108.76	−2.11	−2.07	400.2	1245.5
350	350	124.88	−2.06	−2.09	483.0	1366.9
400	400	175.78	−2.17	−2.13	512.9	1443.3

Table 1: Numerical results for problem (40)

4.2 Constrained nonconvex-strongly-concave minimax optimization with quadratic objective and linear constraints

In this subsection, we consider the problem

$$\min_{\widehat{A}x\le\hat{b}} \max_{\widetilde{A}x+\widetilde{B}y\le\tilde{b}} x^T Ax + x^T By + y^T Cy + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) - \mathscr{I}_{[-1,1]^m}(y), \tag{42}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times m}$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}^m$, $\widehat{A} \in \mathbb{R}^{\tilde{n}\times n}$, $\hat{b} \in \mathbb{R}^{\tilde{n}}$, $\widetilde{A} \in \mathbb{R}^{\tilde{m}\times n}$, $\widetilde{B} \in \mathbb{R}^{\tilde{m}\times m}$, $\tilde{b} \in \mathbb{R}^{\tilde{m}}$, and $\mathscr{I}_{[-1,1]^n}(\cdot)$ and $\mathscr{I}_{[-1,1]^m}(\cdot)$ are the indicator functions of $[-1,1]^n$ and $[-1,1]^m$ respectively.

For each tuple $(n, m, \tilde{n}, \tilde{m})$, we randomly generate 10 instances of problem (42). Specifically, we construct $A = UDU^T$, where $U = \mathrm{orth}(\mathrm{randn}(n))$, and D is a diagonal matrix with entries independently drawn from a normal distribution with mean 0 and standard deviation 0.1. Matrix C is generated in a similar manner, except its diagonal entries are independently drawn from a uniform distribution over $[10, 11]$. In addition, we randomly generate matrices B, $\widehat{A}$, $\widetilde{A}$, $\widetilde{B}$, and vectors c, d, $\tilde{b}$ with all the entries independently drawn from a normal distribution with mean 0 and standard deviation 0.1. Finally, we randomly generate $x_{\mathbf{nf}} \in [-1, 1]^n$ by first sampling each entry independently from a normal distribution with mean 0 and standard deviation 0.1, then projecting the resulting vector onto $[-1, 1]^n$. We choose $\hat{b}$ such that $x_{\mathbf{nf}}$ is 0.1-nearly feasible (see Assumption 3) for problem (42).

Notice that (42) is a special case of (1) with

$$\begin{aligned} &f(x, y) = x^T A x + x^T B y + y^T C y + c^T x + d^T y, \quad p(x) = \mathscr{I}_{[-1,1]^n}(x), \\ &q(y) = \mathscr{I}_{[-1,1]^m}(y), \quad c(x) = \widehat{A} x - \hat{b}, \quad d(x, y) = \widetilde{A} x + \widetilde{B} y - \tilde{b}, \end{aligned}$$

and can be suitably solved by Algorithm 2 and ALM [32, Algorithm 3]. In addition, problem (42) is equivalent to the following minimization problem

$$\min_{\widehat{A} x \leq \hat{b}} \Phi(x), \tag{43}$$

where Φ is the hyper-objective function defined as

$$\Phi(x) = \max_{\widetilde{A} x + \widetilde{B} y \leq \tilde{b}} x^T A x + x^T B y + y^T C y + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) - \mathscr{I}_{[-1,1]^m}(y).$$

We choose the parameters as $(\varepsilon, \tau, \Lambda) = (10^{-2}, 0.5, 10)$ for both Algorithm 2 and ALM [32, Algorithm 3], and initialize them at zero. The algorithms are terminated once a 10^{-2}-relative-KKT point[2] (x_k, y_k) of (42) is found for some k, and we output (x_k, y_k) as an approximate solution to (42).

The computational results of the aforementioned algorithms for the instances randomly generated above are presented in Table 2. Specifically, the values of n, m, $\tilde{n}$, and $\tilde{m}$ are listed in the first four columns. For each tuple $(n, m, \tilde{n}, \tilde{m})$, the average initial hyper-objective value $\Phi(x^0)$, the average final hyper-objective value $\Phi(x^k)$, and the average CPU time (in seconds) over 10 random instances are given in the rest of the columns. We observe that both Algorithm 2 and ALM [32, Algorithm 3] produce approximate solutions with comparable hyper-objective values that are significantly lower than the initial ones. Moreover, Algorithm 2 consistently achieves substantially lower CPU times since it effectively exploits the strong concavity structure of the problem.

n	m	$\tilde{n}$	$\tilde{m}$	Initial hyper-objective value	Final hyper-objective value		CPU time (seconds)	
					Algorithm 2	ALM	Algorithm 2	ALM
50	100	5	10	−0.52	−183.09	−183.18	332.8	1111.9
100	200	10	20	−0.40	−625.04	−625.76	2001.9	2996.1
150	300	15	30	−0.45	−895.71	−895.02	4535.1	6396.9
200	400	20	40	−0.34	−1255.49	−1254.74	6252.2	9653.4
250	500	25	50	−0.45	−1631.83	−1632.54	8343.8	13522.1

Table 2: Numerical results for problem (42)

5 Proof of the main result

In this section we provide a proof of our main results presented in Sections 2 and 3, which are particularly Theorems 1 and 2.

5.1 Proof of the main results in Section 2

In this subsection we prove Theorem 1. Before proceeding, let $\{(x^k, y^k)\}_{k \in \mathbb{T}}$ denote all the iterates generated by Algorithm 1, where $\mathbb{T}$ is a subset of consecutive nonnegative integers starting from 0. Also, we define $\mathbb{T} - 1 = \{k - 1 : k \in \mathbb{T}\}$. We first establish two lemmas and then use them to prove Theorem 1 subsequently.

The following lemma shows that an approximate primal-dual stationary point of (7) is found at each iteration of Algorithm 1, and also provides an estimate of operation complexity for finding it.

[2] We say (x, y) is an ϵ-relative-KKT point of (42) if it is an $(|\Phi(x)| + 1)\epsilon$-KKT point of (42).

Lemma 1. *Suppose that Assumption 2 holds. Let $\{(x^k,y^k)\}_{k\in\mathbb{T}}$ be generated by Algorithm 1, H^*, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, H_{low}, $\hat\alpha$, $\hat\delta$ be defined in (6), (10), (11), (12) and (13), $L_{\nabla h}$ be given in Assumption 2, ϵ, $\hat\epsilon_k$ be given in Algorithm 1, and*

$$\hat N_k := 3397\left\lceil \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\}\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat\alpha L_{\nabla h}}\right\}\right\}\left(\hat\delta + 2\hat\alpha^{-1}(H^* - H_{\text{low}} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\} + 3L_{\nabla h}\right]^{-2}\hat\epsilon_k^2}\right\rceil_+ . \tag{44}$$

Then for all $0\le k\in\mathbb{T}-1$, (x^{k+1},y^{k+1}) is an $\hat\epsilon_k$-primal-dual stationary point of (7). Moreover, the total number of evaluations of ∇h and proximal operators of p and q performed at iteration k of Algorithm 1 for generating (x^{k+1},y^{k+1}) is no more than $\hat N_k$, respectively.

Proof. Let (x^*,y^*) be an optimal solution of (6). Recall that H, H_k and h_k are respectively given in (6), (7) and (8), $\mathcal{X} = \operatorname{dom} p$ and $\mathcal{Y} = \operatorname{dom} q$. Notice that $x^*, x^k \in \mathcal{X}$. Then we have

$$\begin{aligned} H_{k,*} &:= \min_x \max_y H_k(x,y) = \min_x \max_y \left\{H(x,y) + L_{\nabla h}\|x-x^k\|^2\right\} \\ &\le \max_y \{H(x^*,y) + L_{\nabla h}\|x^*-x^k\|^2\} \overset{(6)(10)}{\le} H^* + L_{\nabla h}D_{\mathbf{x}}^2. \end{aligned} \tag{45}$$

Moreover, by $\mathcal{X} = \operatorname{dom} p$, $\mathcal{Y} = \operatorname{dom} q$, (10) and (11), one has

$$H_{k,\text{low}} := \min_{(x,y)\in\operatorname{dom} p\times\operatorname{dom} q} H_k(x,y) = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{H(x,y) + L_{\nabla h}\|x-x^k\|^2\right\} \overset{(11)}{\ge} H_{\text{low}}. \tag{46}$$

In addition, by Assumption 2 and the definition of h_k in (8), it is not hard to verify that $h_k(x,y)$ is $L_{\nabla h}$-strongly-convex in x, σ_y-strongly-concave in y, and $3L_{\nabla h}$-smooth on its domain. Also, recall from Remark 1 that (x^{k+1},y^{k+1}) results from applying Algorithm 3 to problem (7). The conclusion of this lemma then follows by using (45) and (46) and applying Theorem 3 to (7) with $\bar\epsilon = \hat\epsilon_k$, $\sigma_x = L_{\nabla h}$, $\sigma_y = \sigma$, $L_{\nabla\bar h} = 3L_{\nabla h}$, $\bar\alpha = \hat\alpha$, $\bar\delta = \hat\delta$, $\bar H_{\text{low}} = H_{k,\text{low}}$, and $\bar H^* = H_{k,*}$. □

The following lemma provides an upper bound on the least progress of the solution sequence of Algorithm 1 and also on the last-iterate objective value of (6).

Lemma 2. *Suppose that Assumption 2 holds. Let $\{x^k\}_{k\in\mathbb{T}}$ be generated by Algorithm 1, H, H^* and $D_{\mathbf{y}}$ be defined in (6) and (10), $L_{\nabla h}$ be given in Assumption 2, and ϵ, $\hat\epsilon_0$ and $\hat x^0$ be given in Algorithm 1. Then for all $0\le K\in\mathbb{T}-1$, we have*

$$\min_{0\le k\le K}\|x^{k+1}-x^k\| \le \frac{\max_y H(\hat x^0,y) - H^*}{L_{\nabla h}(K+1)} + \frac{2\hat\epsilon_0^2(1+\sigma_y^{-2}L_{\nabla h}^2)}{L_{\nabla h}^2(K+1)}, \tag{47}$$

$$\max_y H(x^{K+1},y) \le \max_y H(\hat x^0,y) + 2\hat\epsilon_0^2\left(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h}\right). \tag{48}$$

Proof. For convenience of the proof, let

$$H^*(x) = \max_y H(x,y), \tag{49}$$

$$H_k^*(x) = \max_y H_k(x,y), \quad y_*^{k+1} = \arg\max_y H_k(x^{k+1},y). \tag{50}$$

One can observe from these, (7) and (8) that

$$H_k^*(x) = H^*(x) + L_{\nabla h}\|x-x^k\|^2. \tag{51}$$

By this and Assumption 2, one can also see that H_k^* is $L_{\nabla h}$-strongly convex on $\operatorname{dom} p$. In addition, recall from Lemma 1 that (x^{k+1},y^{k+1}) is an $\hat\epsilon_k$-primal-dual stationary point of problem (7) for all $0\le k\in\mathbb{T}-1$. It then follows from Definition 1 that there exist some $u\in\partial_x H_k(x^{k+1},y^{k+1})$ and $v\in\partial_y H_k(x^{k+1},y^{k+1})$ with $\|u\|\le\hat\epsilon_k$ and $\|v\|\le\hat\epsilon_k$. Also, by (50), one has $0\in\partial_y H_k(x^{k+1},y_*^{k+1})$, which, together with $v\in\partial_y H_k(x^{k+1},y^{k+1})$ and σ_y-strong concavity of $H_k(x^{k+1},\cdot)$, implies that $\langle -v, y^{k+1}-y_*^{k+1}\rangle \ge \sigma_y\|y^{k+1}-y_*^{k+1}\|^2$. This and $\|v\|\le\hat\epsilon_k$ yield

$$\|y^{k+1}-y_*^{k+1}\| \le \sigma_y^{-1}\hat\epsilon_k. \tag{52}$$

In addition, by $u \in \partial_x H_k(x^{k+1}, y^{k+1})$, (7) and (8), one has

$$u \in \nabla_x h(x^{k+1}, y^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1} - x^k). \tag{53}$$

Also, observe from (7), (8) and (50) that

$$\partial H_k^*(x^{k+1}) = \nabla_x h(x^{k+1}, y_*^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1} - x^k),$$

which together with (53) yields

$$u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}) \in \partial H_k^*(x^{k+1}).$$

By this and $L_{\nabla h}$-strong convexity of H_k^*, one has

$$H_k^*(x^k) \geq H_k^*(x^{k+1}) + \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1} \rangle + L_{\nabla h}\|x^k - x^{k+1}\|^2/2. \tag{54}$$

Using this, (51), (52), (54), $\|u\| \leq \hat{\epsilon}_k$, and the Lipschitz continuity of ∇h, we obtain

$$\begin{aligned}
H^*(x^k) - H^*(x^{k+1}) &\overset{(51)}{=} H_k^*(x^k) - H_k^*(x^{k+1}) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(54)}{\geq} \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1} \rangle + 3L_{\nabla h}\|x^k - x^{k+1}\|^2/2 \\
&\geq \big(-\|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|\|x^k - x^{k+1}\| + L_{\nabla h}\|x^k - x^{k+1}\|^2/2 \big) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -(2L_{\nabla h})^{-1}\|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -L_{\nabla h}^{-1}\|u\|^2 - L_{\nabla h}^{-1}\|\nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -L_{\nabla h}^{-1}\hat{\epsilon}_k^2 - L_{\nabla h}\|y^{k+1} - y_*^{k+1}\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(52)}{\geq} -(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h})\hat{\epsilon}_k^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2,
\end{aligned}$$

where the second and fourth inequalities follow from Cauchy-Schwartz inequality, and the third inequality is due to Young's inequality, and the fifth inequality follows from $L_{\nabla h}$-Lipschitz continuity of ∇h. Summing up the above inequality for $k = 0, 1, \ldots, K$ yields

$$L_{\nabla h}\sum_{k=0}^{K}\|x^k - x^{k+1}\|^2 \leq H^*(x^0) - H^*(x^{K+1}) + (L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h})\sum_{k=0}^{K}\hat{\epsilon}_k^2. \tag{55}$$

In addition, it follows from (6), (10) and (49) that

$$H^*(x^{K+1}) = \max_y H(x^{K+1}, y) \geq \min_x \max_y H(x, y) = H^*, \quad H^*(x^0) = \max_y H(x^0, y). \tag{56}$$

These together with (55) yield

$$\begin{aligned}
L_{\nabla h}(K+1)\min_{0\leq k\leq K}\|x^{k+1} - x^k\|^2 &\leq L_{\nabla h}\sum_{k=0}^{K}\|x^k - x^{k+1}\|^2 \\
&\leq \max_y H(x^0, y) - H^* + (L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h})\sum_{k=0}^{K}\hat{\epsilon}_k^2,
\end{aligned}$$

which, together with $x^0 = \hat{x}^0$, $\hat{\epsilon}_k = \hat{\epsilon}_0(k+1)^{-1}$ and $\sum_{k=0}^{K}(k+1)^{-2} < 2$, implies that (47) holds.

Finally, we show that (48) holds. Indeed, it follows from (10), (49), (55), (56), $\hat{\epsilon}_k = \hat{\epsilon}_0(k+1)^{-1}$, and $\sum_{k=0}^{K}(k+1)^{-2} < 2$ that

$$\begin{aligned}
\max_y H(x^{K+1}, y) \overset{(49)}{=} H^*(x^{K+1}) &\overset{(55)}{\leq} H_\epsilon^*(x^0) + (L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h})\sum_{k=0}^{K}\hat{\epsilon}_k^2 \\
&\overset{(56)}{\leq} \max_y H(x^0, y) + 2\hat{\epsilon}_0^2(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h}).
\end{aligned}$$

It then follows from this and $x^0 = \hat{x}^0$ that (48) holds. □

We are now ready to prove Theorem 1 using Lemmas 1 and 2.

Proof of Theorem 1. Suppose for contradiction that Algorithm 1 runs for more than $\widehat{T}+1$ outer iterations, where $\widehat{T}$ is given in (14). By this and Algorithm 1, one can then assert that (9) does not hold for all $0 \le k \le \widehat{T}$. On the other hand, by (14) and (47), one has

$$\min_{0\le k\le \widehat{T}} \|x^{k+1}-x^k\|^2 \overset{(47)}{\le} \frac{\max_y H(\hat{x}^0,y)-H^*}{L_{\nabla h}(\widehat{T}+1)} + \frac{2\hat{\epsilon}_0^2(1+\sigma_y^{-2}L_{\nabla h}^2)}{L_{\nabla h}^2(\widehat{T}+1)} \overset{(14)}{\le} \frac{\epsilon^2}{16L_{\nabla h}^2},$$

which implies that there exists some $0 \le k \le \widehat{T}$ such that $\|x^{k+1}-x^k\| \le \epsilon/(4L_{\nabla h})$, and hence (9) holds for such k, which contradicts the above assertion. Hence, Algorithm 1 must terminate in at most $\widehat{T}+1$ outer iterations.

Suppose that Algorithm 1 terminates at some iteration $0 \le k \le \widehat{T}$, namely, (9) holds for such k. We next show that its output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ is an ϵ-primal-dual stationary point of (6) and moreover it satisfies (69). Indeed, recall from Lemma 1 that (x^{k+1}, y^{k+1}) is an $\hat{\epsilon}_k$-primal-dual stationary point of (7), namely, it satisfies $\mathrm{dist}(0, \partial_x H_k(x^{k+1}, y^{k+1})) \le \hat{\epsilon}_k$ and $\mathrm{dist}(0, \partial_y H_k(x^{k+1}, y^{k+1})) \le \hat{\epsilon}_k$. By these, (6), (7) and (8), there exists (u,v) such that

$$u \in \partial_x H(x^{k+1}, y^{k+1}) + 2L_{\nabla h}(x^{k+1}-x^k), \quad \|u\| \le \hat{\epsilon}_k,$$
$$v \in \partial_y H(x^{k+1}, y^{k+1}), \quad \|v\| \le \hat{\epsilon}_k.$$

It then follows that $u - 2L_{\nabla h}(x^{k+1}-x^k) \in \partial_x H(x^{k+1}, y^{k+1})$ and $v \in \partial_y H(x^{k+1}, y^{k+1})$. These together with (9), (10), and $\hat{\epsilon}_k \le \hat{\epsilon}_0 \le \epsilon/2$ (see Algorithm 1) imply that

$$\mathrm{dist}\left(0, \partial_x H(x^{k+1}, y^{k+1})\right) \le \|u - 2L_{\nabla h}(x^{k+1}-x^k)\| \le \|u\| + 2L_{\nabla h}\|x^{k+1}-x^k\| \overset{(9)}{\le} \hat{\epsilon}_k + \epsilon/2 \le \epsilon,$$
$$\mathrm{dist}\left(0, \partial_y H(x^{k+1}, y^{k+1})\right) \le \|v\| \le \hat{\epsilon}_k < \epsilon.$$

Hence, the output (x^{k+1}, y^{k+1}) of Algorithm 1 is an ϵ-primal-dual stationary point of (6). In addition, (16) holds due to Lemma 2.

Recall from Lemma 1 that the number of evaluations of ∇h and proximal operators of p and q performed at iteration k of Algorithm 1 is at most $\hat{N}_k$, respectively, where $\hat{N}_k$ is defined in (44). Also, one can observe from the above proof and the definition of $\mathbb{T}$ that $|\mathbb{T}| \le \widehat{T}+2$. It then follows that the total number of evaluations of ∇h and proximal operators of p and q in Algorithm 1 is respectively no more than $\sum_{k=0}^{|\mathbb{T}|-2} \hat{N}_k$. Consequently, to complete the rest of the proof of Theorem 1, it suffices to show that $\sum_{k=0}^{|\mathbb{T}|-2} \hat{N}_k \le \widehat{N}$, where $\widehat{N}$ is given in (15). Indeed, by (15), (44) and $|\mathbb{T}| \le \widehat{T}+2$, one has

$$\begin{aligned}
\sum_{k=0}^{|\mathbb{T}|-2} \hat{N}_k &\overset{(44)}{\le} \sum_{k=0}^{\widehat{T}} 3397 \times \left\lceil \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \right. \\
&\quad \left. \times \log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\mathrm{low}} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\} + 3L_{\nabla h}\right]^{-2}\hat{\epsilon}_k^2} \right\rceil_+ \\
&\le 3397 \times \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \\
&\quad \times \sum_{k=0}^{\widehat{T}} \left(\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\mathrm{low}} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\} + 3L_{\nabla h}\right]^{-2}\hat{\epsilon}_k^2}\right)_+ + 1\right) \\
&\le 3397 \times \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \\
&\quad \times \left((\widehat{T}+1)\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\mathrm{low}} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\} + 3L_{\nabla h}\right]^{-2}\hat{\epsilon}_0^2}\right)_+ \right. \\
&\quad \left. + \widehat{T} + 1 + 2\sum_{k=0}^{\widehat{T}} \log(k+1)\right) \overset{(15)}{\le} \widehat{N},
\end{aligned}$$

where the last inequality is due to (15) and $\sum_{k=0}^{\widehat{T}} \log(k+1) \le \widehat{T}\log(\widehat{T}+1)$. This completes the proof of Theorem 1. □

5.2 Proof of the main results in Section 3

In this subsection, we provide a proof of our main result presented in Section 3, which is particularly Theorem 2. Before proceeding, let

$$\mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}};\rho) = F(x,y) - \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{y}} + \rho d(x,y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right). \tag{57}$$

In view of (4), (22) and (57), one can observe that

$$f^*(x) \le \max_y \mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}};\rho) \qquad \forall x \in \mathcal{X},\ \lambda_{\mathbf{y}} \in \mathbb{R}^{\tilde{m}}_+,\ \rho > 0, \tag{58}$$

which will be frequently used later.

We next establish several lemmas that will be used to prove Theorem 2 subsequently. The next lemma provides an upper bound for $\{\lambda_{\mathbf{y}}^k\}_{k\in\mathbb{K}}$.

Lemma 3. *Suppose that Assumptions 1 and 4 hold. Let $\{\lambda_{\mathbf{y}}^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, $D_{\mathbf{y}}$ and Δ be defined in (10) and (24), and τ, and ρ_k be given in Algorithm 2. Then we have*

$$\rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 \le \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} \qquad \forall 0 \le k \in \mathbb{K}-1. \tag{59}$$

Proof. Its proof is similar to that of [32, Lemma 5] and thus omitted. □

The following lemma establishes an upper bound on $\|[d(x^{k+1},y^{k+1})]_+\|$ for $0 \le k \in \mathbb{K}-1$.

Lemma 4. *Suppose that Assumptions 1 and 4 hold. Let $D_{\mathbf{y}}$ and Δ be defined in (10) and (24), δ_d be given in Assumption 4, and τ, ϵ_k and ρ_k be given in Algorithm 2. Suppose that $(x^{k+1},y^{k+1},\lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 2 for some $0 \le k \in \mathbb{K}-1$ with*

$$\rho_k \ge \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2(1-\tau)}. \tag{60}$$

Then we have

$$\|[d(x^{k+1},y^{k+1})]_+\| \le \rho_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\| \le 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}}). \tag{61}$$

Proof. Its proof is similar to that of [32, Lemma 6] and thus omitted. □

Lemma 5. *Suppose that Assumptions 1 and 4 hold. Let $D_{\mathbf{y}}$ and Δ be defined in (10) and (24), and δ_d be given in Assumption 4, τ, ϵ_k, ρ_k and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 2. Suppose that $(x^{k+1},y^{k+1},\lambda_{\mathbf{x}}^{k+1},\lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 2 for some $0 \le k \in \mathbb{K}-1$ with*

$$\rho_k \ge \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2\tau(1-\tau)}. \tag{62}$$

Let

$$\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+.$$

Then we have

$$\begin{aligned}
&\mathrm{dist}(0, \partial_x F(x^{k+1},y^{k+1}) + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1}) \le \epsilon_k,\\
&\mathrm{dist}\left(0, \partial_y F(x^{k+1},y^{k+1}) - \nabla_y d(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1}\right) \le \epsilon_k,\\
&\|[d(x^{k+1},y^{k+1})]_+\| \le 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}}),\\
&|\langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1},y^{k+1})\rangle| \le 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}})\max\{\|\lambda_{\mathbf{y}}^0\|,\ 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})\}.
\end{aligned}$$

Proof. Its proof is similar to that of [32, Lemma 7] and thus omitted. □

The following lemma provides an upper bound on $\max_y \mathcal{L}(x_{\mathrm{init}}^k,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k)$ for $0 \le k \in \mathbb{K}-1$, which will subsequently be used to derive an upper bound for $\max_y \mathcal{L}(x^{k+1},y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k)$.

Lemma 6. *Suppose that Assumptions 1, 3 and 4 hold. Let $\{(\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k)\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, $\mathcal{L}$, $D_{\mathbf{y}}$, F_{hi} and Δ be defined in (4), (10), (23) and (24), and τ, ρ_k, Λ and x_{init}^k be given in Algorithm 2. Then for all $0 \le k \in \mathbb{K}-1$, we have*

$$\max_y \mathcal{L}(x_{\mathrm{init}}^k,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \le \Delta + F_{\mathrm{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau}. \tag{63}$$

Proof. Its proof is similar to that of [32, Lemma 8] and thus omitted. $\square$

The next lemma shows that an approximate primal-dual stationary point of (19) is found at each iteration of Algorithm 2, and also provides an estimate of operation complexity for finding it.

Lemma 7. *Suppose that Assumptions 1, 3 and 4 hold. Let $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, L_k, $F_{\rm hi}$ and Δ be defined in* (10), (20), (23) *and* (24), *τ, ϵ_k, ρ_k, Λ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 2, and*

$$
\begin{aligned}
\alpha_k &= \min\left\{1, \sqrt{8\sigma/L_k}\right\}, & (64)\\
\delta_k &= (2+\alpha_k^{-1})L_k D_{\mathbf{x}}^2 + \max\{2\sigma, \alpha_k L_k/4\} D_{\mathbf{y}}^2, & (65)\\
M_k &= \frac{16\max\{1/(2L_k), \min\{1/(2\sigma), 4/(\alpha_k L_k)\}\}\rho_k}{[9L_k^2/\min\{L_k,\sigma\}+3L_k]^{-2}\epsilon_k^2} \\
&\quad \times\left(\delta_k + 2\alpha_k^{-1}\left(\Delta + \frac{\Lambda^2}{2\rho_k} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta+D_{\mathbf{y}})}{1-\tau} + \rho_k d_{\rm hi}^2 + L_k D_{\mathbf{x}}^2\right)\right) & (66)\\
T_k &= \left\lceil 16\left(2\Delta + \Lambda + \frac{1}{2}(\tau^{-1}+\|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} + \frac{\Lambda^2}{2\rho_k}\right) L_k \epsilon_k^{-2}\right. \\
&\quad \left. + 8(1+\sigma^{-2}L_k^2)\epsilon_k^2 - 1\right\rceil_+, & (67)\\
N_k &= 3397\max\left\{2, \sqrt{L_k/(2\sigma)}\right\} \times \left((T_k+1)(\log M_k)_+ + T_k + 1 + 2T_k\log(T_k+1)\right). & (68)
\end{aligned}
$$

Then for all $0 \le k \in \mathbb{K}-1$, Algorithm 2 finds an ϵ_k-primal-dual stationary point (x^{k+1}, y^{k+1}) of problem (19) *that satisfies*

$$
\begin{aligned}
\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \le \Delta + F_{\rm hi} + \Lambda + \frac{1}{2}(\tau^{-1}+\|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau}& \\
+ \frac{1}{2}\left(L_k^{-1}+\sigma^{-2}L_k\right)\epsilon_k^2.& \quad (69)
\end{aligned}
$$

Moreover, the total number of evaluations of ∇f, ∇c, ∇d and proximal operators of p and q performed in iteration k of Algorithm 2 is no more than N_k, respectively.

Proof. Observe from (1) and (4) that problem (19) can be viewed as

$$\min_x \max_y \{h(x,y) + p(x) - q(y)\},$$

where

$$h(x,y) = f(x,y) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right).$$

Notice that

$$
\begin{aligned}
\nabla_x h(x,y) &= \nabla_x f(x,y) + \nabla c(x)[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+ + \nabla_x d(x,y)[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+,\\
\nabla_y h(x,y) &= \nabla_y f(x,y) + \nabla_y d(x,y)[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+.
\end{aligned}
$$

It follows from Assumption 1(iii) that

$$\|\nabla c(x)\| \le L_c, \quad \|\nabla d(x,y)\| \le L_d \qquad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}.$$

In view of the above relations, (18) and Assumption 1, one can observe that $\nabla c(x)[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+$ is $(\rho_k L_c^2 + \rho_k c_{\rm hi} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c})$-Lipschitz continuous on $\mathcal{X}$, and $\nabla d(x,y)[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+$ is $(\rho_k L_d^2 + \rho_k d_{\rm hi} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d})$-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$. Using these and the fact that $\nabla f(x,y)$ is $L_{\nabla f}$-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$ and $f(x,\cdot)$ is σ-strongly-concave on $\mathcal{Y}$ for all $x\in\mathcal{X}$, we can see that $h(x,\cdot)$ is σ-strongly-concave on $\mathcal{Y}$, and $h(x,y)$ is L_k-smooth on $\mathcal{X}\times\mathcal{Y}$ for all $0 \le k \in \mathbb{K}-1$, where L_k is given in (20). Consequently, it follows from Theorem 1 that Algorithm 1 can be suitably applied to problem (19) for finding an ϵ_k-primal-dual stationary point (x^{k+1}, y^{k+1}) of it.

In addition, by (4), (26), (57), (58) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ (see Algorithm 2), one has

$$
\begin{aligned}
&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \overset{(4)(57)}{=} \min_x \max_y \left\{ \mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}}^k;\rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) \right\} \\
&\overset{(58)}{\ge} \min_x \left\{ f^*(x) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) \right\} \overset{(26)}{\ge} F_{\mathrm{low}} - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 \ge F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k}.
\end{aligned} \tag{70}
$$

Let (x^*,y^*) be an optimal solution of (1). It then follows that $c(x^*) \le 0$. Using this, (4), (23) and (59), we obtain that

$$
\begin{aligned}
&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \le \max_y \mathcal{L}(x^*,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \\
&\overset{(4)}{=} \max_y \left\{ F(x^*,y) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^*)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right) \right\} \\
&\le \max_y \left\{ F(x^*,y) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right) \right\} \\
&\overset{(23)}{\le} F_{\mathrm{hi}} + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \overset{(59)}{\le} F_{\mathrm{hi}} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{\Delta + D_{\mathbf{y}}}{1-\tau},
\end{aligned} \tag{71}
$$

where the second inequality is due to $c(x^*) \le 0$. Moreover, it follows from this, (4), (18), (23), (59), $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ that

$$
\begin{aligned}
&\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \overset{(4)}{\ge} \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\|[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+\|^2 \right\} \\
&\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\left(\|\lambda_{\mathbf{y}}^k\| + \rho_k\|[d(x,y)]_+\|\right)^2 \right\} \\
&\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 - \rho_k\|[d(x,y)]_+\|^2 \right\} \\
&\ge F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} - \rho_k d_{\mathrm{hi}}^2,
\end{aligned} \tag{72}
$$

where the second inequality is due to $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and the last inequality is due to (18), (23), (59) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$.

To complete the rest of the proof, let

$$
H(x,y) = \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \quad H^* = \min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \tag{73}
$$

$$
H_{\mathrm{low}} = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k). \tag{74}
$$

In view of these, (63), (70), (71), (72), we obtain that

$$
\begin{aligned}
&\max_y H(x_{\mathrm{init}}^k, y) \overset{(63)}{\le} \Delta + F_{\mathrm{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau}, \\
&F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k} \overset{(70)}{\le} H^* \overset{(71)}{\le} F_{\mathrm{hi}} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{\Delta + D_{\mathbf{y}}}{1-\tau}, \\
&H_{\mathrm{low}} \overset{(72)}{\ge} F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} - \rho_k d_{\mathrm{hi}}^2.
\end{aligned}
$$

Using these, (24), and Theorem 1 with $x^0 = x_{\mathrm{init}}^k$, $\epsilon = \epsilon_k$, $\hat{\epsilon}_0 = \epsilon_k/2$, $L_{\nabla h} = L_k$, $\sigma_y = \sigma$, $\hat{\alpha} = \alpha_k$, $\hat{\delta} = \delta_k$, and H, H^*, H_{low} given in (73) and (74), we can conclude that Algorithm 1 performs at most N_k evaluations of ∇f, ∇c, ∇d and proximal operators of p and q for finding an ϵ_k-primal-dual stationary point of problem (19) satisfying (69). □

The following lemma provides an upper bound on $\|[c(x^{k+1})]_+\|$ and $|\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1})\rangle|$ for $0 \le k \in \mathbb{K}-1$, where $\tilde{\lambda}_{\mathbf{x}}^{k+1}$ is given below.

Lemma 8. *Suppose that Assumptions 1, 3 and 4 hold. Let $D_{\mathbf{y}}$, Δ and L be defined in* (10), (24) *and* (27), *L_F, L_c, δ_c and θ be given in Assumption 4, and τ, ρ_k, Λ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 2. Suppose* *that* $(x^{k+1}, \lambda_{\mathbf{x}}^{k+1})$ *is generated by Algorithm 2 for some* $0 \le k \in \mathbb{K} - 1$ *with*

$$\rho_k \ge \max\left\{\theta^{-1}\Lambda, \theta^{-2}\Big\{4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + L_c^{-2} + \sigma^{-2}L + \Lambda^2\Big\}, \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2 \tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2 \tau(1-\tau)}\right\}. \tag{75}$$

Let

$$\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+. \tag{76}$$

Then we have

$$\|[c(x^{k+1})]_+\| \le \rho_k^{-1}\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1\right), \tag{77}$$

$$|\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1})\rangle| \le \rho_k^{-1}\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1)\max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1), \Lambda\}. \tag{78}$$

Proof. One can observe from (4), (26), (57) and (58) that

$$\begin{aligned}\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) = &\ \max_y \mathcal{L}_{\mathbf{y}}(x^{k+1}, y, \lambda_{\mathbf{y}}^k; \rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) \\ \overset{(58)}{\ge} &\ f^*(x^{k+1}) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) \\ \overset{(26)}{\ge} &\ F_{\rm low} + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right).\end{aligned}$$

By this inequality, (69) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$, one has

$$\begin{aligned}&\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 \le 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k F_{\rm low} + \|\lambda_{\mathbf{x}}^k\|^2 \\ &\le 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k F_{\rm low} + \Lambda^2 \\ &\overset{(69)}{\le} 2\rho_k\Delta + 2\rho_k F_{\rm hi} + 2\rho_k\Lambda + \rho_k(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{2\rho_k(Delta + D_{\mathbf{y}})}{1-\tau} \\ &\quad + L_k^{-1}\epsilon_k^2 + \sigma^{-2}L_k\epsilon_k^2 - 2\rho_k F_{\rm low} + \Lambda^2.\end{aligned}$$

This together with (24) and $\rho_k^2\|[c(x^{k+1})]_+\|^2 \le \|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2$ implies that

$$\begin{aligned}\|[c(x^{k+1})]_+\|^2 \le &\ \rho_k^{-1}\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}\right) \\ &+ \rho_k^{-2}\left(L_k^{-1}\epsilon_k^2 + \sigma^{-2}L_k\epsilon_k^2 + \Lambda^2\right).\end{aligned} \tag{79}$$

In addition, we observe from (20), (27), (59), $\rho_k \ge 1$ and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ that for all $0 \le k \le K$,

$$\begin{aligned}&\rho_k L_c^2 \le L_k = L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\rm hi} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\rm hi} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d} \\ &\le L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\rm hi} L_{\nabla c} + \Lambda L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\rm hi} L_{\nabla d} \\ &\quad + L_{\nabla d}\sqrt{\rho_k\left(\|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}\right)} \le \rho_k L.\end{aligned} \tag{80}$$

Using this relation, (75), (79), $\rho_k \ge 1$ and $\epsilon_k \le 1$, we have

$$\begin{aligned}\|[c(x^{k+1})]_+\|^2 \le &\ \rho_k^{-1}\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}\right) \\ &+ \rho_k^{-2}\left((\rho_k L_c^2)^{-1}\epsilon_k^2 + \sigma^{-2}L\epsilon_k^2\rho_k + \Lambda^2\right) \\ \le &\ \rho_k^{-1}\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}\right) \\ &+ \rho_k^{-1}\left(L_c^{-2} + 4\sigma^{-2}L + \Lambda^2\right) \overset{(75)}{\le} \theta^2,\end{aligned}$$

which together with (21) implies that $x^{k+1} \in \mathcal{F}(\theta)$.

It follows from $x^{k+1} \in \mathcal{F}(\theta)$ and Assumption 4(i) that there exists some $v \in \mathcal{T}_{\mathcal{X}}(x^{k+1})$ such that $\|v\| = 1$ and $v^T \nabla c_i(x^{k+1}) \le -\delta_c$ for all $i \in \mathcal{A}(x^{k+1};\theta)$, where $\mathcal{A}(x^{k+1};\theta)$ is defined in (21). Let $\bar{\mathcal{A}}(x^{k+1};\theta) = \{1,2,\ldots,\tilde{n}\}\backslash\mathcal{A}(x^{k+1};\theta)$. Notice from (21) that $c_i(x^{k+1}) < -\theta$ for all $i \in \bar{\mathcal{A}}(x^{k+1};\theta)$. In addition, observe from (75) that $\rho_k \ge \theta^{-1}\Lambda$. Using these and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$, we obtain that $(\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1}))_i \le \Lambda - \rho_k\theta \le 0$ for all $i \in \bar{\mathcal{A}}(x^{k+1};\theta)$. By this and the fact that $v^T\nabla c_i(x^{k+1}) \le -\delta_c$ for all $i \in \mathcal{A}(x^{k+1};\theta)$, one has

$$
\begin{aligned}
&v^T\nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} \overset{(76)}{=} v^T\nabla c(x^{k+1})[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+ = \sum_{i=1}^{\tilde{n}} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \\
&= \sum_{i\in\mathcal{A}(x^{k+1};\theta)} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i + \sum_{i\in\bar{\mathcal{A}}(x^{k+1};\theta)} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \\
&\le -\delta_c \sum_{i\in\mathcal{A}(x^{k+1};\theta)} ([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i = -\delta_c\sum_{i=1}^{\tilde{n}}([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \overset{(76)}{=} -\delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1. \qquad (81)
\end{aligned}
$$

Since (x^{k+1}, y^{k+1}) is an ϵ_k-primal-dual stationary point of (19), it follows from (4) and Definition 1 that there exists some $s \in \partial_x F(x^{k+1}, y^{k+1})$ such that

$$
\|s + \nabla c(x^{k+1})[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+ - \nabla_x d(x^{k+1}, y^{k+1})[\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+\| \le \epsilon_k,
$$

which along with (76) and $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_x d(x^{k+1}, y^{k+1})]_+$ implies that

$$
\|s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\| \le \epsilon_k. \qquad (82)
$$

In addition, since $v \in \mathcal{T}_{\mathcal{X}}(x^{k+1})$, there exist $\{z^t\} \subset \mathcal{X}$ and $\{\alpha_t\} \downarrow 0$ such that $z^t = x^{k+1} + \alpha_t v + o(\alpha_t)$ for all t. Also, since $s \in \partial_x F(x^{k+1}, y^{k+1})$, one has $s = \nabla_x f(x^{k+1}, y^{k+1}) + s_p$ for some $s_p \in \partial p(x^{k+1})$. Using these and Assumptions 1 and 4(iii), we have

$$
\begin{aligned}
\langle s, v\rangle &= \langle \nabla_x f(x^{k+1}, y^{k+1}), v\rangle + \lim_{t\to\infty} \alpha_t^{-1}\langle s_p, z^t - x^{k+1}\rangle \\
&= \lim_{t\to\infty} \alpha_t^{-1}(f(z^t, y^{k+1}) - f(x^{k+1}, y^{k+1})) + \lim_{t\to\infty} \alpha_t^{-1}\langle s_p, z^t - x^{k+1}\rangle \\
&\le \lim_{t\to\infty} \alpha_t^{-1}(f(z^t, y^{k+1}) - f(x^{k+1}, y^{k+1})) + \lim_{t\to\infty} \alpha_t^{-1}(p(z^t) - p(x^{k+1})) \\
&= \lim_{t\to\infty} \alpha_t^{-1}(F(z^t, y^{k+1}) - F(x^{k+1}, y^{k+1})) \le L_F \lim_{t\to\infty} \alpha_t^{-1}\|z^t - x^{k+1}\| = L_F, \qquad (83)
\end{aligned}
$$

where the second equality is due to the differentiability of f, the first inequality follows from the convexity of p and $s_p \in \partial p(x^{k+1})$, the second inequality is due to the L_F-Lipschitz continuity of $F(\cdot, y^{k+1})$, and the last equality follows from $\lim_{t\to\infty} \alpha_t^{-1}\|z^t - x^{k+1}\| = \|v\| = 1$.

By (81), (82), (83), and $\|v\| = 1$, one has

$$
\begin{aligned}
\epsilon_k &\ge \|s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\| \cdot \|v\| \\
&\ge \langle s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, -v\rangle \\
&= -\langle s - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, v\rangle - v^T\nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} \\
&\overset{(81)}{\ge} -\langle s, v\rangle - \|\nabla_x d(x^{k+1}, y^{k+1})\|\|\lambda_{\mathbf{y}}^{k+1}\|\|v\| + \delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1 \\
&\ge -L_F - L_d\|\lambda_{\mathbf{y}}^{k+1}\| + \delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1,
\end{aligned}
$$

where the last inequality is due to $\|v\| = 1$ and Assumptions 1(i) and 1(iii). Notice from (75) that (60) holds. It then follows from (61) that $\|\lambda_{\mathbf{y}}^{k+1}\| \le 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})$, which together with the above inequality and $\epsilon_k \le 1$ yields

$$
\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\| \le \|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1 \le \delta_c^{-1}(L_F + L_d\|\lambda_{\mathbf{y}}^{k+1}\| + \epsilon_k) \le \delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1). \qquad (84)
$$

By this and (76), one can observe that

$$
\|[c(x^{k+1})]_+\| \le \rho_k^{-1}\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\| = \rho_k^{-1}\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\| \le \rho_k^{-1}\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1).
$$

Hence, (77) holds as desired.

We next show that (78) holds. Indeed, by $\tilde{\lambda}_{\mathbf{x}}^{k+1} \geq 0$, (77) and (84), one has

$$\begin{aligned} \langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1}) \rangle &\leq \langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, [c(x^{k+1})]_+ \rangle \leq \|\tilde{\lambda}_{\mathbf{x}}^{k+1}\| \|[c(x^{k+1})]_+\| \\ &\overset{(77)(84)}{\leq} \rho_k^{-1} \delta_c^{-2} (L_F + 2L_d \delta_d^{-1} (\Delta + D_{\mathbf{y}}) + 1)^2. \end{aligned} \tag{85}$$

Notice that $\langle \lambda_{\mathbf{x}}^{k+1}, \lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1}) \rangle = \|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 \geq 0$. Hence, we have

$$-\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, \rho_k^{-1} \lambda_{\mathbf{x}}^k \rangle \leq \langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1}) \rangle,$$

which along with $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$ and (84) yields

$$\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1}) \rangle \geq -\rho_k^{-1} \|\tilde{\lambda}_{\mathbf{x}}^{k+1}\| \|\lambda_{\mathbf{x}}^k\| \geq -\rho_k^{-1} \delta_c^{-1} (L_F + 2L_d \delta_d^{-1} (\Delta + D_{\mathbf{y}}) + 1)\Lambda.$$

The relation (78) then follows from this and (85). □

We are now ready to prove Theorem 2 using Lemmas 5, 7 and 8.

Proof of Theorem 2. (i) Observe from the definition of K in (25) and $\epsilon_k = \tau^k$ that K is the smallest nonnegative integer such that $\epsilon_K \leq \varepsilon$. Hence, Algorithm 2 terminates and outputs (x^{K+1}, y^{K+1}) after $K+1$ outer iterations. It follows from these and $\rho_k = \epsilon_k^{-1}$ that $\epsilon_K \leq \varepsilon$ and $\rho_K \geq \varepsilon^{-1}$. By this and (32), one can see that (62) and (75) holds for $k = K$. It then follows from Lemmas 5 and 8 that (33)-(38) hold.

(ii) Let K and N be given in (25) and (39). Recall from Lemma 7 that the number of evaluations of ∇f, ∇c, ∇d, proximal operators of p and q performed by Algorithm 1 at iteration k of Algorithm 2 is at most N_k, where N_k is given in (68). By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f, ∇c, ∇d, proximal operators of p and q performed in Algorithm 2 is no more than $\sum_{k=0}^K N_k$, respectively. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^K N_k \leq N$. Recall from (80) and Algorithm 2 that $\rho_k L_c^2 \leq L_k \leq \rho_k L$ and $\rho_k \geq 1 \geq \epsilon_k$. Using these, (28), (29), (30), (64), (65), (66) and (67), we obtain that

$$1 \geq \alpha_k \geq \min\left\{1, \sqrt{8\sigma/(\rho_k L)}\right\} \geq \rho_k^{-1/2} \alpha, \tag{86}$$

$$\delta_k \leq (2 + \rho_k^{1/2} \alpha^{-1}) \rho_k L D_{\mathbf{x}}^2 + \max\{2\sigma, \rho_k L/4\} D_{\mathbf{y}}^2 \leq \rho_k^{3/2} \delta, \tag{87}$$

$$\begin{aligned} M_k &\leq \frac{16 \max\left\{1/(2\rho_k L_c^2), 4/(\rho_k^{-1/2} \alpha \rho_k L_c^2)\right\}}{[9\rho_k^2 L^2 / \min\{\rho_k L_c^2, \sigma\} + 3\rho_k L]^{-2} \epsilon_k^2} \times \Bigg(\rho_k^{3/2} \delta + 2\rho_k^{1/2} \alpha^{-1} \\ &\quad \times \left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2} \|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1 - \tau} + \rho_k d_{\text{hi}}^2 + \rho_k L D_{\mathbf{x}}^2 \right) \Bigg) \\ &\leq \frac{16 \rho_k^{-1/2} \max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}}{\rho_k^{-4} [9L^2 / \min\{L_c^2, \sigma\} + 3L]^{-2} \epsilon_k^2} \times \rho_k^{3/2} \Bigg(\delta + 2\alpha^{-1} \\ &\quad \times \left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2} \|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1 - \tau} + d_{\text{hi}}^2 + L D_{\mathbf{x}}^2 \right) \Bigg) \leq \epsilon_k^{-2} \rho_k^5 M, \end{aligned} \tag{88}$$

$$\begin{aligned} T_k &\leq \Bigg\lceil 16 \left(2\Delta + \Lambda + \frac{1}{2} (\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1 - \tau} + \frac{\Lambda^2}{2} \right) \epsilon_k^{-2} \rho_k L \\ &\quad + 8(1 + \sigma^{-2} \rho_k^2 L^2) \epsilon_k^{-2} - 1 \Bigg\rceil_+ \leq \epsilon_k^{-2} \rho_k T, \end{aligned}$$

where (88) follows from (28), (29), (30), (86), (87), $\rho_k L_c^2 \leq L_k \leq \rho_k L$, and $\rho_k \geq 1 \geq \epsilon_k$. By the above inequalities, (68), (80), $T \geq 1$ and $\rho_k \geq 1 \geq \epsilon_k$, one has

$$\sum_{k=0}^{K} N_k \leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{\rho_k L/(2\sigma)}\right\}$$
$$\times \left((\epsilon_k^{-2}\rho_k T + 1)(\log(\epsilon_k^{-2}\rho_k^5 M))_+ + \epsilon_k^{-2}\rho_k T + 1 + 2\epsilon_k^{-2}\rho_k T \log(\epsilon_k^{-2}\rho_k T + 1)\right)$$
$$\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} \times \epsilon_k^{-2}\rho_k^{3/2} \left((T+1)(\log(\epsilon_k^{-2}\rho_k^5 M))_+ + T + 1 + 2T\log(\epsilon_k^{-2}\rho_k T + 1)\right)$$
$$\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-2}\rho_k^{3/2} \left((2\log(\epsilon_k^{-2}\rho_k^5 M))_+ + 2 + 2\log(2\epsilon_k^{-2}\rho_k T)\right)$$
$$\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-2}\rho_k^{3/2} \left(12\log\rho_k - 8\log\epsilon_k + 2(\log M)_+ + 2 + 2\log(2T)\right), \tag{89}$$

By the definition of K in (25), one has $\tau^K \geq \tau\varepsilon$. Also, notice from Algorithm 2 that $\rho_k = \tau^{-k}$. It then follows from these, (39) and (89) that

$$\sum_{k=0}^{K} N_k \leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-7/2} \left(20\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right)$$
$$= 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \tau^{-7k/2} \left(20k\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right)$$
$$\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \tau^{-7k/2} \left(20K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right)$$
$$\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\tau^{-7/2K}(1-\tau^4)^{-1}$$
$$\times \left(20K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right)$$
$$\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T(1-\tau^{7/2})^{-1}$$
$$\times (\tau\varepsilon)^{-7/2} \left(20K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right) \overset{(39)}{=} N,$$

where the second last inequality is due to $\sum_{k=0}^{K} \tau^{-7k/2} \leq \tau^{-7K/2}/(1-\tau^{7/2})$, and the last inequality is due to $\tau^K \geq \tau\varepsilon$. Hence, statement (ii) of this theorem holds as desired. □

Declarations

The authors report there are no competing interests to declare.

References

[1] K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *The International Conference on Learning Representations*, 2021.

[2] E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization.* SIAM, 2014.

[3] E. G. Birgin and J. M. Martínez. Complexity and performance of an augmented Lagrangian algorithm. *Optim. Methods and Softw.*, 35(5):885–920, 2020.

[4] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[5] X. Chen, L. Guo, Z. Lu, and J. J. Ye. An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM J. Numer. Anal.*, 55(1):168–193, 2017.

[6] Z. Chen, Y. Zhou, T. Xu, and Y. Liang. Proximal gradient descent-ascent: variable convergence under KŁ geometry. *arXiv preprint arXiv:2102.04653*, 2021.

[7] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[8] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134, 2018.

[9] Y.-H. Dai, J. Wang, and L. Zhang. Optimality conditions and numerical algorithms for a class of linearly constrained minimax optimization problems. *SIAM Journal on Optimization*, 34(3):2883–2916, 2024.

[10] Y.-H. Dai and L. Zhang. Optimality conditions for constrained minimax optimization. *arXiv preprint arXiv:2004.09730*, 2020.

[11] Y.-H. Dai and L.-W. Zhang. The rate of convergence of augmented lagrangian method for minimax optimization problems with equality constraints. *Journal of the Operations Research Society of China*, pages 1–33, 2022.

[12] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058, 2017.

[13] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[14] X. Fu and E. Modiano. Network interdiction using adversarial traffic flows. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1765–1773. IEEE, 2019.

[15] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[16] D. Goktas and A. Greenwald. Convex-concave min-max stackelberg games. *Advances in Neural Information Processing Systems*, 34:2991–3003, 2021.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[19] G. N. Grapiglia and Y. Yuan. On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA J. Numer. Anal.*, 41(2):1508–1530, 2021.

[20] Z. Guo, Y. Yan, Z. Yuan, and T. Yang. Fast objective & duality gap convergence for non-convex strongly-concave min-max problems with pl condition. *J. Mach. Learn. Res.*, 24:148–1, 2023.

[21] N. Ho-Nguyen and S. J. Wright. Adversarial classification via distributional robustness with wasserstein ambiguity. *Mathematical Programming*, 198(2):1411–1447, 2023.

[22] F. Huang, S. Gao, J. Pei, and H. Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 3, 2020.

[23] C. Kanzow and D. Steck. An example comparing the standard and safeguarded augmented Lagrangian methods. *Oper. Res. Lett.*, 45(6):598–603, 2017.

[24] A. Kaplan and R. Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

[25] W. Kong and R. D. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

[26] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[27] C. Laidlaw, S. Singla, and S. Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

[28] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093, 2020.

[29] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779, 2020.

[30] S. Lu. A single-loop gradient descent and perturbed ascent algorithm for nonconvex functional constrained optimization. In *International Conference on Machine Learning*, pages 14315–14357, 2022.

[31] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided nonconvex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[32] Z. Lu and S. Mei. A first-order augmented Lagrangian method for constrained minimax optimization. *Mathematical Programming*, pages 1–42, 2024.

[33] Z. Lu and Y. Zhang. An augmented Lagrangian approach for sparse principal component analysis. *Math. Program.*, 135(1-2):149–193, 2012.

[34] L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[36] G. Mateos, J. A. Bazerque, and G. B. Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58:5262–5276, 2010.

[37] O. Nachum, Y. Chow, B. Dai, and L. Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pages 2315–2325, 2019.

[38] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

[40] S. Qiu, Z. Yang, X. Wei, J. Ye, and Z. Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *arXiv preprint arXiv:2008.10103*, 2020.

[41] A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[42] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] J. Salmeron, K. Wood, and R. Baldick. Analysis of electric grid security under terrorist threat. *IEEE Transactions on power systems*, 19(2):905–912, 2004.

[44] M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

[45] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, page 1576–1584, 2015.

[46] J. Shamma. *Cooperative Control of Distributed Multi-Agent Systems.* Wiley-Interscience, 2008.

[47] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[48] J. C. Smith, M. Prince, and J. Geunes. Modern network interdiction problems and algorithms. In *Handbook of combinatorial optimization*, pages 1949–1987. Springer New York, 2013.

[49] J. Song, H. Ren, D. Sadigh, and S. Ermon. Multi-agent generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.

[50] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, page 2989–2997, 2015.

[51] B. Taskar, S. Lacoste-Julien, and M. Jordan. Structured prediction via the extragradient method. In *Advances in Neural Information Processing Systems*, page 1345–1352, 2006.

[52] I. Tsaknakis, M. Hong, and S. Zhang. Minimax problems with coupled linear constraints: Computational complexity and duality. *SIAM Journal on Optimization*, 33(4):2675–2702, 2023.

[53] J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li. Adversarial attack generation empowered by min-max optimization. In *Advances in Neural Information Processing Systems*, 2021.

[54] D. Ward and J. M. Borwein. Nonsmooth calculus in finite dimensions. *SIAM Journal on control and optimization*, 25(5):1312–1340, 1987.

[55] W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021.

[56] Y. Xie and S. J. Wright. Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *J. Sci. Comput.*, 86(3):1–30, 2021.

[57] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

[58] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems*, page 1537–1544, 2005.

[59] T. Xu, Z. Wang, Y. Liang, and H. V. Poor. Gradient free minimax optimization: Variance reduction and faster convergence. *arXiv preprint arXiv:2006.09361*, 2020.

[60] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201(1):635–706, 2023.

[61] J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems*, pages 5667–5678, 2020.

[62] H. Zhang, J. Wang, Z. Xu, and Y.-H. Dai. Primal dual alternating proximal gradient algorithms for nonsmooth nonconvex minimax problems with coupled linear constraints. *arXiv preprint arXiv:2212.04672*, 2022.

[63] J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.

[64] R. Zhao. A primal-dual smoothing framework for max-structured non-convex optimization. *Mathematics of operations research*, 49(3):1535–1565, 2024.

A A modified optimal first-order method for strongly-convex-strongly-concave minimax problem

In this part, we present a modified optimal first-order method [32, Algorithm 1] in Algorithm 3 below for finding an approximate primal-dual stationary point of strongly-convex-strongly-concave minimax problem

$$\bar{H}^* = \min_x \max_y \left\{ \bar{H}(x,y) := \bar{h}(x,y) + p(x) - q(y) \right\}, \tag{90}$$

which satisfies the following assumptions.

Assumption 5. *(i) $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $q : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ are proper convex functions and continuous on* $\operatorname{dom} p$ *and* $\operatorname{dom} q$*, respectively, and moreover,* $\operatorname{dom} p$ *and* $\operatorname{dom} q$ *are compact.*

(ii) The proximal operators associated with p and q can be exactly evaluated.

(iii) $\bar{h}(x, y)$ is σ_x-strongly-convex-σ_y-strongly-concave and $L_{\nabla \bar{h}}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$ *for some $\sigma_x, \sigma_y > 0$.*

For convenience of presentation, we introduce some notation below, most of which is adopted from [26]. Let $\mathcal{X} = \operatorname{dom} p$, $\mathcal{Y} = \operatorname{dom} q$, (x^*, y^*) denote the optimal solution of (90), $z^* = -\sigma_x x^*$, and

$$
\begin{aligned}
& D_{\mathbf{x}} := \max\{\|u - v\| \big| u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| \big| u, v \in \mathcal{Y}\}, && (91)\\
& \bar{H}_{\text{low}} = \min\big\{\bar{H}(x, y) | \big(x, y) \in \mathcal{X} \times \mathcal{Y}\big\}, && (92)\\
& \hat{h}(x, y) = \bar{h}(x, y) - \sigma_x \|x\|^2/2 + \sigma_y \|y\|^2/2, \\
& \mathcal{G}(z, y) = \sup_x \{\langle x, z\rangle - p(x) - \hat{h}(x, y) + q(y)\}, \\
& \mathcal{P}(z, y) = \sigma_x^{-1}\|z\|^2/2 + \sigma_y \|y\|^2/2 + \mathcal{G}(z, y), \\
& \vartheta_k = \eta_z^{-1}\|z^k - z^*\|^2 + \eta_y^{-1}\|y^k - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(z_f^k, y_f^k) - \mathcal{P}(z^*, y^*)), && (93)\\
& a_x^k(x, y) = \nabla_x \hat{h}(x, y) + \sigma_x(x - \sigma_x^{-1} z_g^k)/2, \quad a_y^k(x, y) = -\nabla_y \hat{h}(x, y) + \sigma_y y + \sigma_x(y - y_g^k)/8,
\end{aligned}
$$

where $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$, $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$, and y^k, y_f^k, y_g^k, z^k, z_f^k and z_g^k are generated at iteration k of Algorithm 3 below. By Assumption 5, one can observe that $D_{\mathbf{x}}$, $D_{\mathbf{y}}$ and $\bar{H}_{\text{low}}$ are finite.

We are now ready to review a modified optimal first-order method [32, Algorithm 1] for solving (90) in Algorithm 3. It is a slight modification of an optimal first-order method [26, Algorithm 4] by incorporating a forward-backward splitting scheme and a verifiable termination criterion (see steps 23-25 in Algorithm 3) in order to find an $\bar{\epsilon}$-primal-dual stationary point of problem (90) for any prescribed tolerance $\bar{\epsilon} > 0$.

Algorithm 3 A modified optimal first-order method for problem (90)

Input: $\bar\epsilon > 0$, $\bar z^0 = z_f^0 \in -\sigma_x \mathrm{dom}\, p$,[3] $\bar y^0 = y_f^0 \in \mathrm{dom}\, q$, $(z^0, y^0) = (\bar z^0, \bar y^0)$, $\bar\alpha = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$, $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y), 4/(\bar\alpha\sigma_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1+8L_{\nabla\bar h}/\sigma_x)\right)^{-1}$, $\gamma_x = \gamma_y = 8\sigma_x^{-1}$, and $\bar\zeta = \min\{\sigma_x, \sigma_y\}/L_{\nabla\bar h}^2$.

for $k = 0, 1, 2, \ldots$ **do**

  $(z_g^k, y_g^k) = \bar\alpha(z^k, y^k) + (1-\bar\alpha)(z_f^k, y_f^k)$.

  $(x^{k,-1}, y^{k,-1}) = (-\sigma_x^{-1} z_g^k, y_g^k)$.

  $x^{k,0} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$.

  $y^{k,0} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$.

  $b_x^{k,0} = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$.

  $b_y^{k,0} = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$.

  $t = 0$.

  while

  $\gamma_x\|a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t}\|^2 + \gamma_y\|a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t}\|^2 > \gamma_x^{-1}\|x^{k,t} - x^{k,-1}\|^2 + \gamma_y^{-1}\|y^{k,t} - y^{k,-1}\|^2$

  do

    $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t})$.

    $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t})$.

    $x^{k,t+1} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$.

    $y^{k,t+1} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$.

    $b_x^{k,t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$.

    $b_y^{k,t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$.

    $t \leftarrow t+1$.

  end while

  $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$.

  $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x \hat h(x_f^{k+1}, y_f^{k+1}) + b_x^{k,t}, -\nabla_y \hat h(x_f^{k+1}, y_f^{k+1}) + b_y^{k,t})$.

  $z^{k+1} = z^k + \eta_z\sigma_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \sigma_x^{-1} z_f^{k+1})$.

  $y^{k+1} = y^k + \eta_y\sigma_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \sigma_y y_f^{k+1})$.

  $x^{k+1} = -\sigma_x^{-1} z^{k+1}$.

  $\tilde x^{k+1} = \mathrm{prox}_{\bar\zeta p}(x^{k+1} - \bar\zeta\nabla_x \bar h(x^{k+1}, y^{k+1}))$.

  $\tilde y^{k+1} = \mathrm{prox}_{\bar\zeta q}(y^{k+1} + \bar\zeta\nabla_y \bar h(x^{k+1}, y^{k+1}))$.

  Terminate the algorithm and output $(\tilde x^{k+1}, \tilde y^{k+1})$ if

$$\|\bar\zeta^{-1}(x^{k+1} - \tilde x^{k+1}, \tilde y^{k+1} - y^{k+1}) - (\nabla \bar h(x^{k+1}, y^{k+1}) - \nabla \bar h(\tilde x^{k+1}, \tilde y^{k+1}))\| \le \bar\epsilon.$$

end for

The following theorem presents *iteration and operation complexity* of Algorithm 3 for finding an $\bar\epsilon$-primal-dual stationary point of problem (90), whose proof can be found in [32, Section 4.1].

Theorem 3 (Complexity of Algorithm 3). *Suppose that Assumption 5 hold. Let $\bar H^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\bar H_{\mathrm{low}}$, and ϑ_0 be defined in* (90), (91), (92) *and* (93), *σ_x, σ_y and $L_{\nabla\bar h}$ be given in Assumption 5, $\bar\alpha$, η_y, η_z, $\bar\epsilon$, $\bar\zeta$ be given in Algorithm 3, and*

$$\begin{aligned}
\bar\delta &= (2+\bar\alpha^{-1})\sigma_x D_{\mathbf{x}}^2 + \max\{2\sigma_y, \bar\alpha\sigma_x/4\} D_{\mathbf{y}}^2,\\
\bar K &= \left\lceil \max\left\{\frac{2}{\bar\alpha}, \frac{\bar\alpha\sigma_x}{4\sigma_y}\right\} \log \frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\bar\zeta^{-1} + L_{\nabla\bar h})^{-2}\bar\epsilon^2} \right\rceil_+,\\
\bar N &= \left\lceil \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\} \log \frac{4\max\{1/(2\sigma_x), \min\{1/(2\sigma_y), 4/(\bar\alpha\sigma_x)\}\}\left(\bar\delta + 2\bar\alpha^{-1}\left(\bar H^* - \bar H_{\mathrm{low}}\right)\right)}{(L_{\nabla\bar h}^2/\min\{\sigma_x,\sigma_y\} + L_{\nabla\bar h})^{-2}\bar\epsilon^2} \right\rceil_+\\
&\quad \times \left(\left\lceil 96\sqrt{2}\left(1 + 8L_{\nabla\bar h}\sigma_x^{-1}\right)\right\rceil + 2\right).
\end{aligned}$$

Then Algorithm 3 outputs an $\bar\epsilon$-primal-dual stationary point of (90) *in at most $\bar K$ iterations. Moreover, the total number of evaluations of $\nabla\bar h$ and proximal operators of p and q performed in Algorithm 3 is no more than $\bar N$, respectively.*

[3] For convenience, $-\sigma_x \mathrm{dom}\, p$ stands for the set $\{-\sigma_x u | u \in \mathrm{dom}\, p\}$.