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

Zhaosong Lu * Sanyou Mei *

July 12, 2023

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 previously best known operation complexity by a factor of $\sqrt{\varepsilon}$.

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 $\operatorname{dom} q$ are 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_F-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$, f is $L_{\nabla f}$-smooth on $\mathcal{X}\times\mathcal{Y}$, and $f(x,\cdot)$ is σ-strongly-concave 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 operator 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.

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu). This work was partially supported by NSF Award IIS-2211491.

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

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

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

where X and Y are a closed set. where X and Y are closed sets. This problem has found wide applications in various areas, including adversarial training [16, 31, 41, 46], generative adversarial networks [13, 15, 38], reinforcement learning [7, 11, 33, 35, 42], computational game [1, 36, 43], distributed computing [32, 40], prediction and regression [4, 44, 49, 50], and distributionally robust optimization [12, 39]. Numerous methods have been developed to solve problem (2) when X and Y are *simple closed convex sets* (e.g., see [6, 18, 19, 24, 25, 27, 30, 34, 47, 51, 52, 54]).

There have also been several studies on other special cases of problem (1). Specifically, in [14], 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 $V(x) = \max_{y\in Y}\{f(x,y) : d(x,y) \le 0\}$, as well as the ability to compute an optimal Lagrangian multiplier associated with the constraint $d(x,y) \le 0$ for each $x \in X$. Moreover, in [10], 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, [45] 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, [8] 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 [53] 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 [8, 14, 53] 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 [28, 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 operator 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 [28, 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 [28, 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

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

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.
- 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 operation complexity of [28, Algorithm 3] by a factor of $\sqrt{\varepsilon}$.

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 4.

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 $\mathbb{B}^+_\Lambda=\{x\geq 0:\|x\|\leq\Lambda\}$, 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 $\|\phi(x)-\phi(x')\|\leq L_\phi\|x-x'\|$ for all $x,x'\in S$. In addition, it is said to be *$L_{\nabla\phi}$-smooth* on S if $\|\nabla\phi(x)-\nabla\phi(x')\|\leq L_{\nabla\phi}\|x-x'\|$ 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 $x\in\mathbb{R}^n$.

Finally, we introduce an (approximate) primal-dual stationary point (e.g., see [8, 9, 22]) 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

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

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

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

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 $\mathcal{X} := \operatorname{dom} p$ and $\mathcal{Y} := \operatorname{dom} q$.*

Our first-order method applies an inexact proximal point method [21] to (6) and solves a sequence of subproblems

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

where

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

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 $\operatorname{dom} p \times \operatorname{dom} q$; (ii) h_k is $3L_{\nabla h}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$. Consequently, problem (7) is a special case of (83) 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: $\epsilon > 0$, $\hat{\epsilon}_0 \in (0, \epsilon/2]$, $(\hat{x}^0, \hat{y}^0) \in \operatorname{dom} p \times \operatorname{dom} q$, $(x^0, y^0) = (\hat{x}^0, \hat{y}^0)$, 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**

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

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

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

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 4.1.

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*

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

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

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

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

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 operator of p and q performed in Algorithm 1 is no more than $\widehat{N}$, respectively.

Remark 1. *Since $\hat{\epsilon}_0 \in (0, \epsilon/2]$, one can observe from Theorem 1 that $\hat{\alpha} = \mathcal{O}(1)$, $\hat{\delta} = \mathcal{O}(1)$, $\widehat{T} = \mathcal{O}(\epsilon^{-2})$, and $\widehat{N} = \mathcal{O}(\epsilon^{-2}\log\hat{\epsilon}_0^{-1})$. Consequently, if $\hat{\epsilon}_0 = \epsilon/2$, Algorithm 1 enjoys an operation complexity of $\mathcal{O}(\epsilon^{-2}\log\epsilon^{-1})$, measured by the amount of evaluations of ∇h and proximal operator of p and q, for finding an ϵ-primal-dual stationary point of nonconvex-strongly-concave minimax problem (6).*

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 [28, Algorithm 3]. Specifically, at each iteration, the FAL method finds an approximate primal-dual stationary point of an AL subproblem in the form of

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

for some $\rho > 0$, $\lambda_{\mathbf{x}} \in \mathbb{R}_+^{\tilde{n}}$ and $\lambda_{\mathbf{y}} \in \mathbb{R}_+^{\tilde{m}}$, where $\mathcal{L}$ is the AL function associated with problem (1) defined in (4). 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

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

where $\mathcal{X} := \operatorname{dom} p$ and $\mathcal{Y} := \operatorname{dom} q$, and make one additional assumption on problem (1).

Assumption 3. *For any given $\eta \in (0, 1]$, an η-approximately feasible point z_η of problem (1), namely $z_\eta \in \mathcal{X}$ satisfying $\|[c(z_\eta)]_+\| \le \eta$, can be found.*

Remark 2. *A very similar assumption as Assumption 3 was considered in [5, 17, 28, 29, 48]. 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 [26, 37]). In this case, one can find the above* z_η *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: $\varepsilon, \tau \in (0,1)$, $\epsilon_0 \in (\tau\varepsilon, 1]$, $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, $\Lambda > 0$, $\lambda_{\mathbf{x}}^0 \in \mathbb{B}_\Lambda^+$, $\lambda_{\mathbf{y}}^0 \in \mathbb{R}_+^{\tilde{m}}$, $(x^0, y^0) \in \mathrm{dom}\, p \times \mathrm{dom}\, q$, 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 3. *(i) Note that* $x_{\mathbf{nf}}$ *is an* $\sqrt{\varepsilon}$*-approximately feasible point of problem* (1)*, where the subscript "***nf***" stands for "nearly feasible". It follows from Assumption 3 that* $x_{\mathbf{nf}}$ *can be found in advance.*

(ii) $\lambda_{\mathbf{x}}^{k+1}$ *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, 20], which has been shown to enjoy many practical and theoretical advantages (see [2] for discussions).*

(iii) 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 holds for the minimization part of (1) and a uniform Slater's condition holds for the maximization part of (1).

Assumption 4. *(i) There exist some constants* δ_c*,* θ_a*,* $\theta_f > 0$ *such that for each* $x \in \mathcal{F}(\theta_f)$ *there exists some* $v_x \in \mathbb{R}^n$ *satisfying* $\|v_x\| = 1$ *and* $v_x^T \nabla c_i(x) \le -\delta_c$ *for all* $i \in \mathcal{A}(x; \theta_a)$*, where*

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

(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, $\delta_d := \inf\{-d_i(x, \hat{y}_x) | x \in \mathcal{X},\ i = 1, 2, \ldots, \tilde{m}\} > 0$.[2]

In addition, to characterize the approximate solution found by Algorithm 2, we review a terminology called an ε-KKT solution of problem (1), which was introduced in [28, 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*

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

Furthermore, we define

$$f^*(x) := \max\{F(x, y) | d(x, y) \le 0\}, \tag{22}$$

$$f^*_{\text{low}} := \inf\{f^*(x) | x \in \mathcal{X}\}, \tag{23}$$

$$F_{\text{hi}} := \max\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \quad F_{\text{low}} := \min\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{24}$$

$$K := \lceil (\log \varepsilon - \log \epsilon_0)/\log \tau \rceil_+, \quad \mathbb{K} := \{0, 1, \ldots, K + 1\}, \tag{25}$$

where L_F and δ_d are given in Assumptions 1 and 4, and ϵ_0, 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. In addition, it follows from Assumptions 1, 4 and [28, Lemma 4] that f^*_{low} is finite.

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 4.

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}, f^*_{low}, F_{hi}, F_{low} and K be defined in (10), (18), (23), (24) and (25), L_F, $L_{\nabla f}$, $L_{\nabla d}$, $L_{\nabla c}$, L_c, $L_{\nabla d}$, L_d and δ_d be given in Assumption 1, ε, ϵ_0, τ, Λ and $\lambda^0_{\mathbf{y}}$ be given in Algorithm 2, and*

$$L = L_{\nabla f} + L_c^2 + c_{\text{hi}} L_{\nabla c} + \Lambda L_{\nabla c} + L_d^2 + d_{\text{hi}} L_{\nabla d} + L_{\nabla d}\sqrt{\|\lambda^0_{\mathbf{y}}\|^2 + \frac{2(F_{\text{hi}} - f^*_{\text{low}} + D_{\mathbf{y}}\epsilon_0)}{1 - \tau}}, \tag{26}$$

$$\alpha = \min\left\{1, \sqrt{8\sigma/L}\right\}, \quad \delta = (2 + \alpha^{-1})LD^2_{\mathbf{x}} + \max\{2\sigma, L/4\}D^2_{\mathbf{y}}, \tag{27}$$

$$\begin{aligned}
M = {} & 16 \max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}\left[81/\min\{L_c^2, \sigma\} + 3L\right]^2\\
& \times \left(\delta + 2\alpha^{-1}\left(F_{\text{hi}} - F_{\text{low}} + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda^0_{\mathbf{y}}\|^2 + \frac{3(F_{\text{hi}} - f^*_{\text{low}} + D_{\mathbf{y}}\epsilon_0)}{1 - \tau} + \rho_k d^2_{\text{hi}} + LD^2_{\mathbf{x}}\right)\right),
\end{aligned} \tag{28}$$

$$\begin{aligned}
T = \Bigg\lceil & 16\left(L_F D_{\mathbf{y}} + F_{\text{hi}} - f^*_{\text{low}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda^0_{\mathbf{y}}\|^2) + \frac{F_{\text{hi}} - f^*_{\text{low}} + D_{\mathbf{y}}\epsilon_0}{1 - \tau} + \frac{\Lambda^2}{2}\right) L\\
& + 8(1 + \sigma^{-2}L^2)\Bigg\rceil_+,
\end{aligned} \tag{29}$$

$$\tilde{\lambda}^{K+1}_{\mathbf{x}} = [\lambda^K_{\mathbf{x}} + c(x^{K+1})/(\epsilon_0 \tau^K)]_+. \tag{30}$$

[2]The latter part of this assumption 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).

Suppose that

$$\varepsilon^{-1} \geq \max\Bigg\{1, \theta_a^{-1}\Lambda, \theta_f^{-2}\Big\{2L_F D_{\mathbf{y}} + 2F_{\rm hi} - 2f_{\rm low}^* + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} + L_c^{-2} + \sigma^{-2}L + \Lambda^2\Big\}, \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{\delta_d^2\tau(1-\tau)}\Bigg\}. \tag{31}$$

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

$$\begin{aligned}
&\operatorname{dist}(0, \partial_x F(x^{K+1}, y^{K+1}) + \nabla c(x^{K+1})\tilde{\lambda}_x^{K+1} - \nabla_x d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}) \leq \varepsilon, && (32)\\
&\operatorname{dist}\left(0, \partial_y F(x^{K+1}, y^{K+1}) - \nabla_y d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\right) \leq \varepsilon, && (33)\\
&\|[c(x^{K+1})]_+\| \leq \varepsilon\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0\right), && (34)\\
&|\langle\tilde{\lambda}_{\mathbf{x}}^{K+1}, c(x^{K+1})\rangle| \leq \varepsilon\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0)\\
&\qquad\qquad \times \max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0), \Lambda\}, && (35)\\
&\|[d(x^{K+1}, y^{K+1})]_+\| \leq 2\varepsilon\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}}, && (36)\\
&|\langle\lambda_{\mathbf{y}}^{K+1}, d(x^{K+1}, y^{K+1})\rangle| \leq 2\varepsilon\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}}\max\{2\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}}, \|\lambda_{\mathbf{y}}^0\|\}. && (37)
\end{aligned}$$

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

$$\begin{aligned}
N = {} & 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}T(1-\tau^{7/2})^{-1}\\
& \times (\tau\varepsilon)^{-7/2}\left(20K\log(1/\tau) + 20\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right). \qquad (38)
\end{aligned}$$

Remark 4. *(i) 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 operator of p and q, for finding an $\mathcal{O}(\varepsilon)$-KKT solution (x^{K+1}, y^{K+1}) of (1) such that*

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

where $\tilde{\lambda}_{\mathbf{x}}^{K+1} \in \mathbb{R}_+^{\tilde{n}}$ is defined in (30) and $\lambda_{\mathbf{y}}^{K+1} \in \mathbb{R}_+^{\tilde{m}}$ is given in Algorithm 2.

(ii) It shall be mentioned that an $\mathcal{O}(\varepsilon)$-KKT solution of (1) can be found by [28, Algorithm 3] with an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ (see [28, Theorem 3]). As a result, the operation complexity of Algorithm 2 improves that of [28, Algorithm 3] by a factor of $\sqrt{\epsilon}$.

4 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.

4.1 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.

Lemma 1. *Suppose that Assumption 2 holds. Let $\{(x^k, y^k)\}_{k\in\mathbb{T}}$ be generated by Algorithm 1, H^*, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $H_{\rm low}$, $\hat{\alpha}$, $\hat{\delta}$ 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*

$$\hat{N}_k := 3397\left\lceil \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\rm low} + L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h}, \sigma_y\} + 3L_{\nabla h}\right]^{-2}\hat{\epsilon}_k^2}\right\rceil_+. \tag{39}$$

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 operator 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

$$\begin{aligned} H_{k,*} := \min_x \max_y H_k(x,y) &= \min_x \max_y \left\{H(x,y) + L_{\nabla h}\|x - x^k\|^2\right\} \\ &\le \max_y \{H(x^*, y) + L_{\nabla h}\|x^* - x^k\|^2\} \overset{(6)(10)}{\le} H^* + L_{\nabla h}D_{\mathbf{x}}^2. \end{aligned} \tag{40}$$

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

$$H_{k,\rm low} := \min_{(x,y)\in\operatorname{dom} p\times\operatorname{dom} q} H_k(x,y) = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{H(x,y) + L_{\nabla h}\|x - x^k\|^2\right\} \overset{(11)}{\ge} H_{\rm low}. \tag{41}$$

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 that (x^{k+1}, y^{k+1}) results from applying Algorithm 3 to problem (7). The conclusion of this lemma then follows by using (40) and (41) and applying Theorem 3 to (7) with $\bar{\epsilon} = \hat{\epsilon}_k$, $\sigma_x = L_{\nabla h}$, $\sigma_y = \sigma$, $L_{\nabla \bar{h}} = 3L_{\nabla h}$, $\bar{\alpha} = \hat{\alpha}$, $\bar{\delta} = \hat{\delta}$, $\bar{H}_{\rm low} = H_{k,\rm low}$, and $\bar{H}^* = H_{k,*}$. □

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*

$$\min_{0\le k\le K} \|x^{k+1} - x^k\| \le \frac{\max_y H(\hat{x}^0, y) - H^*}{L_{\nabla h}(K+1)} + \frac{2\hat{\epsilon}_0^2(1 + \sigma_y^{-2}L_{\nabla h}^2)}{L_{\nabla h}^2(K+1)}, \tag{42}$$

$$\max_y H(x^{K+1}, y) \le \max_y H(\hat{x}^0, y) + 2\hat{\epsilon}_0^2\left(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h}\right). \tag{43}$$

Proof. For convenience of the proof, let

$$H^*(x) = \max_y H(x,y), \tag{44}$$

$$H_k^*(x) = \max_y H_k(x,y), \quad y_*^{k+1} = \arg\max_y H_k(x^{k+1}, y). \tag{45}$$

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

$$H_k^*(x) = H^*(x) + L_{\nabla h}\|x - x^k\|^2. \tag{46}$$

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 (45), 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

$$\|y^{k+1} - y_*^{k+1}\| \le \sigma_y^{-1} \hat{\epsilon}_k. \tag{47}$$

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

$$u \in \nabla_x h(x^{k+1}, y^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1} - x^k). \tag{48}$$

Also, observe from (7), (8) and (45) 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 (48) 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

$$H_k^*(x^k) \ge H_k^*(x^{k+1}) + \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1}\rangle + L_{\nabla h}\|x^k - x^{k+1}\|^2/2. \tag{49}$$

Using this, (46), (47), (49), $\|u\| \le \hat{\epsilon}_k$, and the Lipschitz continuity of ∇h, we obtain

$$\begin{aligned}
H^*(x^k) - H^*(x^{k+1}) &\overset{(46)}{=} H_k^*(x^k) - H_k^*(x^{k+1}) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(49)}{\ge} \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1}\rangle + 3L_{\nabla h}\|x^k - x^{k+1}\|^2/2 \\
&\ge \big(- \|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|\|x^k - x^{k+1}\| + L_{\nabla h}\|x^k - x^{k+1}\|^2/2\big) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\ge -(2L_{\nabla h})^{-1}\|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\ge -L_{\nabla h}^{-1}\|u\|^2 - L_{\nabla h}^{-1}\|\nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\ge -L_{\nabla h}^{-1}\hat{\epsilon}_k^2 - L_{\nabla h}\|y^{k+1} - y_*^{k+1}\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(47)}{\ge} -(L_{\nabla h}^{-1} + \sigma_y^{-2} L_{\nabla h})\hat{\epsilon}_k^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2,
\end{aligned}$$

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

$$L_{\nabla h}\sum_{k=0}^{K}\|x^k - x^{k+1}\|^2 \le H^*(x^0) - H^*(x^{K+1}) + (L_{\nabla h}^{-1} + \sigma_y^{-2} L_{\nabla h})\sum_{k=0}^{K}\hat{\epsilon}_k^2. \tag{50}$$

In addition, it follows from (6), (10) and (44) that

$$H^*(x^{K+1}) = \max_y H(x^{K+1}, y) \ge \min_x \max_y H(x, y) = H^*, \quad H^*(x^0) = \max_y H(x^0, y). \tag{51}$$

These together with (50) yield

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

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 (42) holds.

Finally, we show that (43) holds. Indeed, it follows from (10), (44), (50), (51), $\hat{\epsilon}_k = \hat{\epsilon}_0(k+1)^{-1}$, and $\sum_{k=0}^{K}(k+1)^{-2} < 2$ that

$$\begin{aligned}\max_y H(x^{K+1}, y) &\overset{(44)}{=} H^*(x^{K+1}) \\ &\overset{(50)}{\leq} H_\epsilon^*(x^0) + (L_{\nabla h}^{-1} + \sigma_y^{-2} L_{\nabla h}) \sum_{k=0}^{K} \hat{\epsilon}_k^2 \\ &\overset{(51)}{\leq} \max_y H(x^0, y) + 2\hat{\epsilon}_0^2 (L_{\nabla h}^{-1} + \sigma_y^{-2} L_{\nabla h}).\end{aligned}$$

It then follows from this and $x^0 = \hat{x}^0$ that (43) holds. □

We are now ready to prove Theorem 1.

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 $0 \leq k \leq \widehat{T}$. On the other hand, by (14) and (42), one has

$$\min_{0\leq k\leq \widehat{T}} \|x^{k+1} - x^k\|^2 \overset{(42)}{\leq} \frac{\max_y H(\hat{x}^0, y) - H^*}{L_{\nabla h}(\widehat{T}+1)} + \frac{2\hat{\epsilon}_0^2(1 + \sigma_y^{-2} L_{\nabla h}^2)}{L_{\nabla h}^2(\widehat{T}+1)} \overset{(14)}{\leq} \frac{\epsilon^2}{16 L_{\nabla h}^2},$$

which implies that there exists some $0 \leq k \leq \widehat{T}$ such that $\|x^{k+1} - x^k\| \leq \epsilon/(4L_{\nabla h})$, and thus (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 $0 \leq k \leq \widehat{T}$, 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 (64). 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 $\mathrm{dist}(0, \partial_x H_k(x^{k+1}, y^{k+1})) \leq \hat{\epsilon}_k$ and $\mathrm{dist}(0, \partial_y H_k(x^{k+1}, y^{k+1})) \leq \hat{\epsilon}_k$. By these, (6), (7) and (8), there exists (u, v) such that

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

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 (10), (9) and $\hat{\epsilon}_k \leq \hat{\epsilon}_0 \leq \epsilon/2$ (see Algorithm 1) imply that

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

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 operator 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 (39). Also, one can observe from the above proof and the definition of $\mathbb{T}$ that $|\mathbb{T}| \leq \widehat{T}+2$. It then follows that the total number of evaluations of ∇h and proximal operator 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 $\sum_{k=0}^{|\mathbb{T}|-2} \hat{N}_k \leq \widehat{N}$, where $\widehat{N}$ is given in (15). Indeed,

by (15), (39) and $|\mathbb{T}| \le \widehat{T}+2$, one has

$$
\begin{aligned}
\sum_{k=0}^{|\mathbb{T}|-2} \hat{N}_k &\overset{(39)}{\le} \sum_{k=0}^{\widehat{T}} 3397 \times \left\lceil \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \right. \\
&\left. \times \log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\rm low}+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\}+3L_{\nabla h}\right]^{-2}\hat{\epsilon}_k^2} \right\rceil_+ \\
&\le 3397 \times \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \\
&\quad \times \sum_{k=0}^{\widehat{T}} \left(\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\rm low}+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\}+3L_{\nabla h}\right]^{-2}\hat{\epsilon}_k^2} \right)_+ + 1 \right) \\
&\le 3397 \times \max\left\{2, \sqrt{\frac{L_{\nabla h}}{2\sigma_y}}\right\} \\
&\quad \times \left((\widehat{T}+1) \left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\rm low}+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h},\sigma_y\}+3L_{\nabla h}\right]^{-2}\hat{\epsilon}_0^2} \right)_+ \right. \\
&\quad \left. + \widehat{T} + 1 + 2\sum_{k=0}^{\widehat{T}} \log(k+1) \right) \overset{(15)}{\le} \widehat{N},
\end{aligned}
$$

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. □

4.2 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

$$
\mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}};\rho) = F(x,y) - \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{y}}+\rho d(x,y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right). \tag{52}
$$

In view of (4), (22) and (52), one can observe that

$$
f^*(x) \le \max_y \mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}};\rho) \qquad \forall x\in\mathcal{X},\ \lambda_{\mathbf{y}}\in\mathbb{R}_+^{\tilde{m}},\ \rho>0, \tag{53}
$$

which will be frequently used later.

We next establish several lemmas that will be used to prove Theorem 2 subsequently.

Lemma 3. *Suppose that Assumptions 1 and 4 hold. Let $\{\lambda_{\mathbf{y}}^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, $D_{\mathbf{y}}$, $f_{\rm low}^*$, and $F_{\rm hi}$ be defined in* (10), (23) *and* (24)*, and ϵ_0, τ, and ρ_k be given in Algorithm 2. Then we have*

$$
\rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 \le \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi}-f_{\rm low}^*+D_{\mathbf{y}}\epsilon_0)}{1-\tau} \qquad \forall 0\le k\in\mathbb{K}-1. \tag{54}
$$

Proof. Its proof is similar to that of [28, Lemma 5] and thus omitted. □

Lemma 4. *Suppose that Assumptions 1 and 4 hold. Let $D_{\mathbf{y}}$, $f_{\rm low}^*$ and $F_{\rm hi}$ be defined in* (10), (23) *and* (24)*, L_F and δ_d be given in Assumptions 1 and 4, and ϵ_0, 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*

$$
\rho_k \ge \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2} + \frac{8(F_{\rm hi}-f_{\rm low}^*+D_{\mathbf{y}}\epsilon_0)}{\delta_d^2(1-\tau)}. \tag{55}
$$

Then we have

$$\|[d(x^{k+1}, y^{k+1})]_+\| \leq \rho_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\| \leq 2\rho_k^{-1}\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}}. \tag{56}$$

Proof. Its proof is similar to that of [28, Lemma 6] and thus omitted. □

Lemma 5. *Suppose that Assumptions 1 and 4 hold. Let $D_{\mathbf{y}}$, f_{low}^* and F_{low} be defined in (10), (23) and (24), L_F and δ_d be given in Assumptions 1 and 4, ϵ_0, 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 $0 \leq k \in \mathbb{K} - 1$ with*

$$\rho_k \geq \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(F_{\text{hi}} - f_{\text{low}}^* + D_{\mathbf{y}}\epsilon_0)}{\delta_d^2\tau(1-\tau)}. \tag{57}$$

Let

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

Then we have

$$\begin{aligned}
&\text{dist}(0, \partial_x F(x^{k+1}, y^{k+1}) + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}) \leq \epsilon_k,\\
&\text{dist}\left(0, \partial_y F(x^{k+1}, y^{k+1}) - \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\right) \leq \epsilon_k,\\
&\|[d(x^{k+1}, y^{k+1})]_+\| \leq 2\rho_k^{-1}\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}},\\
&|\langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1})\rangle| \leq 2\rho_k^{-1}\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} \max\{\|\lambda_{\mathbf{y}}^0\|,\ 2\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}}\}.
\end{aligned}$$

Proof. Its proof is similar to that of [28, Lemma 7] and thus omitted. □

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_{low}^* and F_{hi} be defined in (4), (10), (23) and (24), L_F be given in Assumption 1, and ϵ_0, τ, ρ_k, Λ and x_{init}^k be given in Algorithm 2. Then for all $0 \leq k \in \mathbb{K} - 1$, we have*

$$\max_y \mathcal{L}(x_{\text{init}}^k, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \leq L_F D_{\mathbf{y}} + F_{\text{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{F_{\text{hi}} - f_{\text{low}}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau}. \tag{58}$$

Proof. Its proof is similar to that of [28, Lemma 8] and thus omitted. □

Lemma 7. *Suppose that Assumptions 1, 3 and 4 hold. Let $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, L_k, f_{low}^*, F_{hi} and F_{low} be defined in (10), (20), (23) and (24), L_F be given in Assumption 1, ϵ_0, τ, ϵ_k, ρ_k, Λ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 2, and*

$$\alpha_k = \min\left\{1, \sqrt{8\sigma/L_k}\right\}, \tag{59}$$

$$\delta_k = (2 + \alpha_k^{-1})L_k D_{\mathbf{x}}^2 + \max\{2\sigma, \alpha_k L_k/4\} D_{\mathbf{y}}^2, \tag{60}$$

$$\begin{aligned}
M_k = &\frac{16\max\{1/(2L_k), \min\{1/(2\sigma), 4/(\alpha_k L_k)\}\}\rho_k}{\left[9L_k^2/\min\{L_k, \sigma\} + 3L_k\right]^{-2}\epsilon_k^2} \times \left(\delta_k + 2\alpha_k^{-1}\left(F_{\text{hi}} - F_{\text{low}}\right.\right.\\
&\left.\left. + \frac{\Lambda^2}{2\rho_k} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(F_{\text{hi}} - f_{\text{low}}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} + \rho_k d_{\text{hi}}^2 + L_k D_{\mathbf{x}}^2\right)\right)
\end{aligned} \tag{61}$$

$$\begin{aligned}
T_k = &\left\lceil 16\left(L_F D_{\mathbf{y}} + F_{\text{hi}} - f_{\text{low}}^* + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{F_{\text{hi}} - f_{\text{low}}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau} + \frac{\Lambda^2}{2\rho_k}\right) L_k \epsilon_k^{-2}\right.\\
&\left. + 8(1 + \sigma^{-2}L_k^2)\epsilon_k^2 - 1\right\rceil_+,
\end{aligned} \tag{62}$$

$$N_k = 3397\max\left\{2, \sqrt{L_k/(2\sigma)}\right\} \times \left((T_k + 1)(\log M_k)_+ + T_k + 1 + 2T_k\log(T_k + 1)\right). \tag{63}$$

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

$$\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \le L_F D_{\mathbf{y}} + F_{\rm hi} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau} + \frac{1}{2}\left(L_k^{-1} + \sigma^{-2}L_k\right)\epsilon_k^2. \tag{64}$$

Moreover, the total number of evaluations of ∇f, ∇c, ∇d and proximal operator 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), (23), (52), (53) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ (see Algorithm 2), one has

$$\begin{aligned}&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \overset{(4)(52)}{=} \min_x \max_y \left\{\mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}}^k;\rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\right\}\\ &\overset{(53)}{\ge} \min_x \left\{f^*(x) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\right\} \overset{(23)}{\ge} f_{\rm low}^* - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 \ge f_{\rm low}^* - \frac{\Lambda^2}{2\rho_k}.\end{aligned} \tag{65}$$

Let (x^*, y^*) be an optimal solution of (1). It then follows that $c(x^*) \le 0$. Using this, (4), (24) and (54), we obtain that

$$\begin{aligned}&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \le \max_y \mathcal{L}(x^*, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k)\\ &\overset{(4)}{=} \max_y \left\{F(x^*,y) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^*)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right)\right\}\\ &\le \max_y \left\{F(x^*,y) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right)\right\}\\ &\overset{(24)}{\le} F_{\rm hi} + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \overset{(54)}{\le} F_{\rm hi} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau},\end{aligned} \tag{66}$$

where the second inequality is due to $c(x^*) \leq 0$. Moreover, it follows from this, (4), (18), (24), (54), $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$ that

$$\begin{aligned}
&\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \overset{(4)}{\geq} \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\|[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+\|^2 \right\} \\
&\geq \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\left(\|\lambda_{\mathbf{y}}^k\| + \rho_k\|[d(x,y)]_+\|\right)^2 \right\} \\
&\geq \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 - \rho_k\|[d(x,y)]_+\|^2 \right\} \\
&\geq F_{\rm low} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} - \rho_k d_{\rm hi}^2,
\end{aligned} \tag{67}$$

where the second inequality is due to $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and the last inequality is due to (18), (24), (54) and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$.

To complete the rest of the proof, let

$$H(x,y) = \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \quad H^* = \min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \tag{68}$$

$$H_{\rm low} = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k). \tag{69}$$

In view of these, (58), (65), (66), (67), we obtain that

$$\begin{aligned}
&\max_y H(x_{\rm init}^k, y) \overset{(58)}{\leq} L_F D_{\mathbf{y}} + F_{\rm hi} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau}, \\
&f_{\rm low}^* - \frac{\Lambda^2}{2\rho_k} \overset{(65)}{\leq} H^* \overset{(66)}{\leq} F_{\rm hi} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau}, \\
&H_{\rm low} \overset{(67)}{\geq} F_{\rm low} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} - \rho_k d_{\rm hi}^2.
\end{aligned}$$

Using these and Theorem 1 with $x^0 = x_{\rm init}^k$, $\epsilon = \epsilon_k$, $\hat{\epsilon}_0 = \epsilon_k/2$, $L_{\nabla h} = L_k$, $\sigma_y = \sigma$, $\hat{\alpha} = \alpha_k$, $\hat{\delta} = \delta_k$, and H, H^*, $H_{\rm low}$ given in (68) and (69), we can conclude that Algorithm 1 performs at most N_k evaluations of ∇f, ∇c, ∇d and proximal operator of p and q for finding an ϵ_k-primal-dual stationary point of problem (19) satisfying (64). □

Lemma 8. *Suppose that Assumptions 1, 3 and 4 hold. Let $D_{\mathbf{y}}$, $f_{\rm low}^*$, $F_{\rm hi}$ and L be defined in (10), (23), (24) and (26), L_F, L_c, δ_c, θ_f and θ_a be given in Assumptions 1 and 4, and ϵ_0, 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 $0 \leq k \in \mathbb{K} - 1$ with*

$$\begin{aligned}
\rho_k \geq \max\Bigg\{ &\theta_a^{-1}\Lambda, \theta_f^{-2}\Big\{2L_F D_{\mathbf{y}} + 2F_{\rm hi} - 2f_{\rm low}^* + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} \\
&+ L_c^{-2} + \sigma^{-2}L + \Lambda^2\Big\}, \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{\delta_d^2\tau(1-\tau)} \Bigg\}.
\end{aligned} \tag{70}$$

Let

$$\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+. \tag{71}$$

Then we have

$$\|[c(x^{k+1})]_+\| \leq \rho_k^{-1}\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0\right), \tag{72}$$

$$|\langle\tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1})\rangle| \leq \rho_k^{-1}\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0)\max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\epsilon_0 + L_F)D_{\mathbf{y}} + \epsilon_0), \Lambda\}. \tag{73}$$

Proof. One can observe from (4), (23), (52) and (53) that

$$\begin{aligned}\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) = &\max_y \mathcal{L}_{\mathbf{y}}(x^{k+1}, y, \lambda_{\mathbf{y}}^k; \rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\\ &\overset{(53)}{\geq} f^*(x^{k+1}) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\\ &\overset{(23)}{\geq} f_{\rm low}^* + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right).\end{aligned}$$

By this inequality, (64) and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$, one has

$$\begin{aligned}&\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 \leq 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k f_{\rm low}^* + \|\lambda_{\mathbf{x}}^k\|^2\\ &\leq 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k f_{\rm low}^* + \Lambda^2\\ &\overset{(64)}{\leq} 2\rho_k L_F D_{\mathbf{y}} + 2\rho_k F_{\rm hi} + 2\rho_k \Lambda + \rho_k(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{2\rho_k(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau}\\ &\quad + L_k^{-1}\epsilon_k^2 + \sigma^{-2}L_k\epsilon_k^2 - 2\rho_k f_{\rm low}^* + \Lambda^2.\end{aligned}$$

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

$$\begin{aligned}\|[c(x^{k+1})]_+\|^2 \leq \rho_k^{-1}&\left(2L_F D_{\mathbf{y}} + 2F_{\rm hi} - 2f_{\rm low}^* + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau}\right)\\ &+ \rho_k^{-2}\left(L_k^{-1}\epsilon_k^2 + \sigma^{-2}L_k\epsilon_k^2 + \Lambda^2\right).\end{aligned} \tag{74}$$

In addition, we observe from (20), (26), (54), $\rho_k \geq 1$ and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$ that for all $0 \leq k \leq K$,

$$\begin{aligned}&\rho_k L_c^2 \leq L_k = L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\rm hi} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\rm hi} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d}\\ &\leq L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\rm hi} L_{\nabla c} + \Lambda L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\rm hi} L_{\nabla d}\\ &\quad + L_{\nabla d}\sqrt{\rho_k\left(\|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_y\epsilon_0)}{1-\tau}\right)} \leq \rho_k L.\end{aligned} \tag{75}$$

Using this relation, (70), (74), $\rho_k\epsilon_k = 1$, $\rho_k \geq 1$ and $\epsilon_k \leq \epsilon_0$, we have

$$\begin{aligned}\|[c(x^{k+1})]_+\|^2 &\leq \rho_k^{-1}\left(2L_F D_{\mathbf{y}} + 2F_{\rm hi} - f_{\rm low}^* + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau}\right)\\ &\quad + \rho_k^{-2}\left((\rho_k L_c^2)^{-1}\epsilon_k^2 + \sigma^{-2}L\epsilon_k^2\rho_k + \Lambda^2\right)\\ &\leq \rho_k^{-1}\left(2L_F D_{\mathbf{y}} + 2F_{\rm hi} - f_{\rm low}^* + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} + \frac{\epsilon_0 D_{\mathbf{y}}}{2}\right)\\ &\quad + \rho_k^{-1}\left(L_c^{-2} + 4\sigma^{-2}L + \Lambda^2\right) \overset{(70)}{\leq} \theta_f^2,\end{aligned}$$

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

It follows from $x^{k+1} \in \mathcal{F}(\theta_f)$ and Assumption 4(i) that there exists some v_x such that $\|v_x\| = 1$ and $v_x^T \nabla c_i(x^{k+1}) \leq -\delta_c$ for all $i \in \mathcal{A}(x^{k+1}; \theta_a)$, where $\mathcal{A}(x^{k+1}; \theta_a)$ is defined in (21). Let $\bar{\mathcal{A}}(x^{k+1}; \theta_a) = \{1, 2, \ldots, \tilde{n}\} \backslash \mathcal{A}(x^{k+1}; \theta_a)$. Notice from (21) that $c_i(x^{k+1}) < -\theta_a$ for all $i \in \bar{\mathcal{A}}(x^{k+1}; \theta_a)$. In addition, observe from (70) that $\rho_k \geq \theta_a^{-1}\Lambda$. Using these and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$, we obtain that $(\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1}))_i \leq \Lambda - \rho_k\theta_a \leq 0$ for all $i \in \bar{\mathcal{A}}(x^{k+1}; \theta_a)$. By this and the fact that

$v_x^T\nabla c_i(x^{k+1})\le-\delta_c$ for all $i\in\mathcal{A}(x^{k+1};\theta_a)$, one has

$$v_x^T\nabla c(x^{k+1})\tilde\lambda_{\mathbf{x}}^{k+1}\overset{(71)}{=}v_x^T\nabla c(x^{k+1})[\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+=\sum_{i=1}^{\tilde n}v_x^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+)_i$$
$$=\sum_{i\in\mathcal{A}(x^{k+1};\theta_a)}v_x^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+)_i+\sum_{i\in\bar{\mathcal{A}}(x^{k+1};\theta_a)}v_x^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+)_i$$
$$\le-\delta_c\sum_{i\in\mathcal{A}(x^{k+1};\theta_a)}([\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+)_i=-\delta_c\sum_{i=1}^{\tilde n}([\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+)_i\overset{(71)}{=}-\delta_c\|\tilde\lambda_{\mathbf{x}}^{k+1}\|_1.\qquad(76)$$

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

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

By this, (76) and $\|v_x\|=1$, one has

$$\begin{aligned}\epsilon_k&\ge\|s+\nabla c(x^{k+1})\tilde\lambda_{\mathbf{x}}^{k+1}-\nabla_x d(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1}\|\cdot\|v_x\|\\&\ge\langle s+\nabla c(x^{k+1})\tilde\lambda_{\mathbf{x}}^{k+1}-\nabla_x d(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1},-v_x\rangle\\&=-\langle s-\nabla_x d(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1},v_x\rangle-v_x^T\nabla c(x^{k+1})\tilde\lambda_{\mathbf{x}}^{k+1}\\&\overset{(76)}{\ge}-\Big(\|s\|+\|\nabla_x d(x^{k+1},y^{k+1})\|\|\lambda_{\mathbf{y}}^{k+1}\|\Big)\|v_x\|+\delta_c\|\tilde\lambda_{\mathbf{x}}^{k+1}\|_1.\\&\ge-L_F-L_d\|\lambda_{\mathbf{y}}^{k+1}\|+\delta_c\|\tilde\lambda_{\mathbf{x}}^{k+1}\|_1,\end{aligned}$$

where the last inequality is due to $\|v_x\|=1$ and Assumptions 1(i) and 1(iii). Notice from (70) that (55) holds. It then follows from (56) that $\|\lambda_{\mathbf{y}}^{k+1}\|\le2\delta_d^{-1}(\epsilon_0+L_F)D_{\mathbf{y}}$, which together with the above inequality and $\epsilon_k\le\epsilon_0$ yields

$$\|\tilde\lambda_{\mathbf{x}}^{k+1}\|\le\|\tilde\lambda_{\mathbf{x}}^{k+1}\|_1\le\delta_c^{-1}(L_F+L_d\|\lambda_{\mathbf{y}}^{k+1}\|+\epsilon_k)\le\delta_c^{-1}(L_F+2L_d\delta_d^{-1}(\epsilon_0+L_F)D_{\mathbf{y}}+\epsilon_0).\quad(77)$$

By this and (71), one can observe that

$$\|[c(x^{k+1})]_+\|\le\rho_k^{-1}\|[\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+\|=\rho_k^{-1}\|\tilde\lambda_{\mathbf{x}}^{k+1}\|\le\rho_k^{-1}\delta_c^{-1}(L_F+2L_d\delta_d^{-1}(\epsilon_0+L_F)D_{\mathbf{y}}+\epsilon_0).$$

Hence, (72) holds as desired.

We next show that (73) holds. Indeed, by $\tilde\lambda_{\mathbf{x}}^{k+1}\ge0$, (72) and (77), one has

$$\begin{aligned}\langle\tilde\lambda_{\mathbf{x}}^{k+1},c(x^{k+1})\rangle&\le\langle\tilde\lambda_{\mathbf{x}}^{k+1},[c(x^{k+1})]_+\rangle\le\|\tilde\lambda_{\mathbf{x}}^{k+1}\|\|[c(x^{k+1})]_+\|\\&\overset{(72)(77)}{\le}\rho_k^{-1}\delta_c^{-2}(L_F+2L_d\delta_d^{-1}(\epsilon_0+L_F)D_{\mathbf{y}}+\epsilon_0)^2.\end{aligned}\qquad(78)$$

Notice that $\langle\lambda_{\mathbf{x}}^{k+1},\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})\rangle=\|[\lambda_{\mathbf{x}}^k+\rho_k c(x^{k+1})]_+\|^2\ge0$. Hence, we have

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

which along with $\|\lambda_{\mathbf{x}}^k\|\le\Lambda$ and (77) yields

$$\langle\tilde\lambda_{\mathbf{x}}^{k+1},c(x^{k+1})\rangle\ge-\rho_k^{-1}\|\tilde\lambda_{\mathbf{x}}^{k+1}\|\|\lambda_{\mathbf{x}}^k\|\ge-\rho_k^{-1}\delta_c^{-1}(L_F+2L_d\delta_d^{-1}(\epsilon_0+L_F)D_{\mathbf{y}}+\epsilon_0)\Lambda.$$

The relation (73) then follows from this and (78). □

We are now ready to prove Theorem 2.

Proof of Theorem 2. (i) Observe from the definition of K in (25) and $\epsilon_k = \epsilon_0\tau^k$ that K is the smallest nonnegative integer such that $\epsilon_K \le \varepsilon$. 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 $\epsilon_K \le \varepsilon$ and $\rho_K \ge \varepsilon^{-1}$. By this and (31), one can see that (57) and (70) holds for $k = K$. It then follows from Lemmas 5 and 8 that (32)-(37) hold.

(ii) Let K and N be given in (25) and (38). Recall from Lemma 7 that the number of evaluations of ∇f, ∇c, ∇d, proximal operator 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 (63). By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f, ∇c, ∇d, proximal operator 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 $\sum_{k=0}^K N_k \le N$. Recall from (75) and Algorithm 2 that $\rho_k L_c^2 \le L_k \le \rho_k L$ and $\rho_k \ge 1 \ge \epsilon_k$. Using these, (27), (28), (29), (59), (60), (61) and (62), we obtain that

$$1 \ge \alpha_k \ge \min\left\{1, \sqrt{8\sigma/(\rho_k L)}\right\} \ge \rho_k^{-1/2}\alpha, \tag{79}$$

$$\delta_k \le (2 + \rho_k^{1/2}\alpha^{-1})\rho_k L D_{\mathbf{x}}^2 + \max\{2\sigma, \rho_k L/4\}D_{\mathbf{y}}^2 \le \rho_k^{3/2}\delta, \tag{80}$$

$$\begin{aligned} M_k &\le \frac{16\max\left\{1/(2\rho_k L_c^2), 4/(\rho_k^{-1/2}\alpha\rho_k L_c^2)\right\}}{\left[9\rho_k^2 L^2/\min\{\rho_k L_c^2, \sigma\} + 3\rho_k L\right]^{-2}\epsilon_k^2} \times \Bigg(\rho_k^{3/2}\delta + 2\rho_k^{1/2}\alpha^{-1} \\ &\quad \times \left(F_{\rm hi} - F_{\rm low} + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} + \rho_k d_{\rm hi}^2 + \rho_k L D_{\mathbf{x}}^2\right)\Bigg) \\ &\le \frac{16\rho_k^{-1/2}\max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}}{\rho_k^{-4}\left[9L^2/\min\{L_c^2, \sigma\} + 3L\right]^{-2}\epsilon_k^2} \times \rho_k^{3/2}\Bigg(\delta + 2\alpha^{-1} \\ &\quad \times \left(F_{\rm hi} - F_{\rm low} + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0)}{1-\tau} + d_{\rm hi}^2 + L D_{\mathbf{x}}^2\right)\Bigg) \le \epsilon_k^{-2}\rho_k^5 M, \end{aligned} \tag{81}$$

$$\begin{aligned} T_k &\le \Bigg\lceil 16\left(L_F D_{\mathbf{y}} + F_{\rm hi} - f_{\rm low}^* + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{F_{\rm hi} - f_{\rm low}^* + D_{\mathbf{y}}\epsilon_0}{1-\tau} + \frac{\Lambda^2}{2}\right)\epsilon_k^{-2}\rho_k L \\ &\quad + 8(1 + \sigma^{-2}\rho_k^2 L^2)\epsilon_k^{-2} - 1\Bigg\rceil_+ \le \epsilon_k^{-2}\rho_k T, \end{aligned}$$

where (81) follows from (27), (28), (29), (79), (80), $\rho_k L_c^2 \le L_k \le \rho_k L$, and $\rho_k \ge 1 \ge \epsilon_k$. By the above inequalities, (63), (75), $T \ge 1$ and $\rho_k \ge 1 \ge \epsilon_k$, one has

$$\begin{aligned} \sum_{k=0}^K N_k &\le \sum_{k=0}^K 3397\max\left\{2, \sqrt{\rho_k L/(2\sigma)}\right\} \\ &\quad \times \left((\epsilon_k^{-2}\rho_k T + 1)(\log(\epsilon_k^{-2}\rho_k^5 M))_+ + \epsilon_k^{-2}\rho_k T + 1 + 2\epsilon_k^{-2}\rho_k T\log(\epsilon_k^{-2}\rho_k T + 1)\right) \\ &\le \sum_{k=0}^K 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\} \times \epsilon_k^{-2}\rho_k^{3/2}\left((T+1)(\log(\epsilon_k^{-2}\rho_k^5 M))_+ + T + 1 + 2T\log(\epsilon_k^{-2}\rho_k T + 1)\right) \\ &\le \sum_{k=0}^K 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-2}\rho_k^{3/2}\left((2\log(\epsilon_k^{-2}\rho_k^5 M))_+ + 2 + 2\log(2\epsilon_k^{-2}\rho_k T)\right) \\ &\le \sum_{k=0}^K 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-2}\rho_k^{3/2}\left(12\log\rho_k - 8\log\epsilon_k + 2(\log M)_+ + 2 + 2\log(2T)\right), \end{aligned} \tag{82}$$

By the definition of K in (25), one has $\tau^K \geq \tau\varepsilon/\epsilon_0$. Also, notice from Algorithm 2 that $\rho_k = \epsilon_k^{-1} = (\epsilon_0\tau^k)^{-1}$. It then follows from these, (38) and (82) that

$$\begin{aligned}
&\sum_{k=0}^{K} N_k \leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_k^{-7/2} \left(20\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&= 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \epsilon_0^{-7/2}\tau^{-7k/2} \left(20k\log(1/\tau) + 20\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \epsilon_0^{-7/2}\tau^{-7k/2} \left(20K\log(1/\tau) + 20\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_0^{-7/2}\tau^{-7/2K}(1-\tau^4)^{-1} \\
&\quad \times \left(20K\log(1/\tau) + 20\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T\epsilon_0^{-7/2}(1-\tau^{7/2})^{-1} \\
&\quad \times (\tau\varepsilon/\epsilon_0)^{-7/2} \left(20K\log(1/\tau) + 20\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \overset{(38)}{=} N,
\end{aligned}$$

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 $\tau^K \geq \tau\varepsilon/\epsilon_0$. Hence, statement (ii) of this theorem holds as desired. □

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] 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.

[8] Y.-H. Dai, J. Wang, and L. Zhang. Optimality conditions and numerical algorithms for a class of linearly constrained minimax optimization problems. *arXiv preprint arXiv:2204.09185*, 2022.

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

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] Z. Guo, Z. Yuan, Y. Yan, and T. Yang. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

[19] 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.

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

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

[22] 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.

[23] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *arXiv preprint arXiv:2205.05653*, 2022.

[24] 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.

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

[26] 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.

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

[28] Z. Lu and S. Mei. A first-order augmented Lagrangian method for constrained minimax optimization. *arXiv preprint arXiv:2301.02060*, 2023.

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

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

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

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

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

[43] 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.

[44] 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.

[45] I. Tsaknakis, M. Hong, and S. Zhang. Minimax problems with coupled linear constraints: computational complexity, duality and solution methods. *arXiv preprint arXiv:2110.11210*, 2021.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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*, pages 1–72, 2023.

[53] 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.

[54] 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.

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 [28, Algorithm 1] in Algorithm 3 below for finding an approximate primal-dual stationary point of strongly-convex-strongly-concave minimax problem

$$\bar{H}^* = \min_x \max_y \left\{ \bar{H}(x,y) \coloneqq \bar{h}(x,y) + p(x) - q(y) \right\}, \tag{83}$$

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 operator 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 [23]. Let $\mathcal{X} = \operatorname{dom} p$, $\mathcal{Y} = \operatorname{dom} q$, (x^*, y^*) denote the optimal solution of (83), $z^* = -\sigma_x x^*$,

and

$$D_{\mathbf{x}} \coloneqq \max\{\|u-v\| \big| u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} \coloneqq \max\{\|u-v\| \big| u, v \in \mathcal{Y}\}, \tag{84}$$

$$\bar{H}_{\text{low}} = \min\left\{\bar{H}(x,y) \big| \big(x,y\big) \in \mathcal{X} \times \mathcal{Y}\right\}, \tag{85}$$

$$\hat{h}(x,y) = \bar{h}(x,y) - \sigma_x\|x\|^2/2 + \sigma_y\|y\|^2/2,$$

$$\mathcal{G}(z,y) = \sup_x\{\langle x, z\rangle - p(x) - \hat{h}(x,y) + q(y)\},$$

$$\mathcal{P}(z,y) = \sigma_x^{-1}\|z\|^2/2 + \sigma_y\|y\|^2/2 + \mathcal{G}(z,y),$$

$$\vartheta_k = \eta_z^{-1}\|z^k - z^*\|^2 + \eta_y^{-1}\|y^k - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(z_f^k, y_f^k) - \mathcal{P}(z^*, y^*)), \tag{86}$$

$$a_x^k(x,y) = \nabla_x\hat{h}(x,y) + \sigma_x(x - \sigma_x^{-1}z_g^k)/2, \quad a_y^k(x,y) = -\nabla_y\hat{h}(x,y) + \sigma_y y + \sigma_x(y - y_g^k)/8,$$

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.

Algorithm 3 A modified optimal first-order method for problem (83)

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 $\bar{\zeta} = \min\{\sigma_x, \sigma_y\}/L_{\nabla\bar{h}}^2$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $(z_g^k, y_g^k) = \bar{\alpha}(z^k, y^k) + (1-\bar{\alpha})(z_f^k, y_f^k)$.
3: $(x^{k,-1}, y^{k,-1}) = (-\sigma_x^{-1} z_g^k, y_g^k)$.
4: $x^{k,0} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$.
5: $y^{k,0} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$.
6: $b_x^{k,0} = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$.
7: $b_y^{k,0} = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$.
8: $t = 0$.
9: **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**
10: $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t})$.
11: $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t})$.
12: $x^{k,t+1} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
13: $y^{k,t+1} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
14: $b_x^{k,t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$.
15: $b_y^{k,t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$.
16: $t \leftarrow t + 1$.
17: **end while**
18: $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$.
19: $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_x^{k,t}, -\nabla_y\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_y^{k,t})$.
20: $z^{k+1} = z^k + \eta_z\sigma_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \sigma_x^{-1} z_f^{k+1})$.
21: $y^{k+1} = y^k + \eta_y\sigma_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \sigma_y y_f^{k+1})$.
22: $x^{k+1} = -\sigma_x^{-1} z^{k+1}$.
23: $\tilde{x}^{k+1} = \mathrm{prox}_{\bar{\zeta}p}(x^{k+1} - \bar{\zeta}\nabla_x\bar{h}(x^{k+1}, y^{k+1}))$.
24: $\tilde{y}^{k+1} = \mathrm{prox}_{\bar{\zeta}q}(y^{k+1} + \bar{\zeta}\nabla_y\bar{h}(x^{k+1}, y^{k+1}))$.
25: Terminate the algorithm and output $(\tilde{x}^{k+1}, \tilde{y}^{k+1})$ if

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

26: **end for**

The following theorem presents *iteration and operation complexity* of Algorithm 3 for finding an $\bar{\epsilon}$-primal-dual stationary point of problem (83), whose proof can be found in [28, 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* (83), (84), (85) *and* (86), *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*

$$\bar{\delta} = (2 + \bar{\alpha}^{-1})\sigma_x D_{\mathbf{x}}^2 + \max\{2\sigma_y, \bar{\alpha}\sigma_x/4\} D_{\mathbf{y}}^2,$$

$$\bar{K} = \left\lceil \max\left\{\frac{2}{\bar{\alpha}}, \frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\} \log \frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\bar{\zeta}^{-1} + L_{\nabla\bar{h}})^{-2}\bar{\epsilon}^2} \right\rceil_+,$$

$$\bar{N} = \left\lceil \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\} \log \frac{4\max\{1/(2\sigma_x), \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}\}\left(\bar{\delta} + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\mathrm{low}}\right)\right)}{(L_{\nabla\bar{h}}^2/\min\{\sigma_x, \sigma_y\} + L_{\nabla\bar{h}})^{-2}\bar{\epsilon}^2} \right\rceil_+$$

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

$$\times \left(\left\lceil 96\sqrt{2}\left(1 + 8L_{\nabla \bar{h}}\sigma_x^{-1}\right) \right\rceil + 2 \right).$$

Then Algorithm 3 outputs an $\bar{\epsilon}$-primal-dual stationary point of (83) *in at most $\bar{K}$ iterations. Moreover, the total number of evaluations of $\nabla \bar{h}$ and proximal operator of p and q performed in Algorithm 3 is no more than $\bar{N}$, respectively.*