# A first-order augmented Lagrangian method for constrained minimax optimization

Zhaosong Lu * Sanyou Mei †

January 5, 2023 (Revised: October 28, 2024)

**Abstract**

In this paper we study a class of constrained minimax problems. In particular, we propose a first-order augmented Lagrangian method for solving them, whose subproblems turn out to be a much simpler structured minimax problem and are suitably solved by a first-order method developed in this paper. Under some suitable assumptions, an *operation complexity* of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$, measured by its fundamental operations, is established for the first-order augmented Lagrangian method for finding an $\varepsilon$-KKT solution of the constrained minimax problems.

**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 constrained minimax problem

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

For notational convenience, throughout this paper we let $\mathcal{X} := \mathrm{dom}\, p$ and $\mathcal{Y} := \mathrm{dom}\, q$, where $\mathrm{dom}\, p$ and $\mathrm{dom}\, q$ 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 $L_{\nabla f}$-smooth on $\mathcal{X}\times\mathcal{Y}$ and $f(x,\cdot)$ is 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 each component $d_i(x,\cdot)$ of $d$ is convex for all $i=1,\ldots,\tilde m$ and $x\in\mathcal{X}$.*

*(iv) The sets $\mathcal{X}$ and $\mathcal{Y}$ (namely, $\mathrm{dom}\, p$ and $\mathrm{dom}\, q$) 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, ONR Award N00014-24-1-2702, and AFOSR Award FA9550-24-1-0343.

[1] The definitions of $L_\phi$-Lipschitz continuity and $L_{\nabla\phi}$-smoothness of a function or mapping $\phi$ are given in Subsection 1.1.

Problem (1) has found applications in machine learning such as perceptual adversarial robustness [28] and robust adversarial classification [21]. Besides, it has potential application to constrained bilevel optimization

$$\min_{x,y} \bar{f}(x,y) + \bar{p}(x) \quad \text{s.t.} \quad y \in \arg\min_z \{\tilde{f}(x,z) + \tilde{p}(z) | \tilde{g}(x,z) \leq 0\}, \tag{2}$$

where $\bar{p}$ and $\tilde{p}$ are proper closed convex functions, $\tilde{g}$, $\nabla \bar{f}$, $\nabla \tilde{f}$ and $\nabla \tilde{g}$ are Lipschitz continuous on $\operatorname{dom} \bar{p} \times \operatorname{dom} \tilde{p}$, and $\tilde{g}_i(x,\cdot)$ is convex for each $x \in \operatorname{dom} \bar{p}$. Specifically, (2) can be tackled by solving a sequence of subproblems in the form of (1). Indeed, observe that (2) is equivalent to

$$\min_{x,y} \bar{f}(x,y) + \bar{p}(x) \quad \text{s.t.} \quad \tilde{g}(x,y) \leq 0, \quad \tilde{f}(x,y) + \tilde{p}(y) - \min_z \{\tilde{f}(x,z) + \tilde{p}(z) | \tilde{g}(x,z) \leq 0\} \leq 0. \tag{3}$$

Notice that any feasible point $(x,y)$ of (3) satisfies $\tilde{f}(x,y) + \tilde{p}(y) - \min_z\{\tilde{f}(x,z) + \tilde{p}(z) | \tilde{g}(x,z) \leq 0\} \geq 0$. As a result, one natural approach to tackling (3) is by solving a sequence of penalty subproblems in the form of

$$\min_{\tilde{g}(x,y)\leq 0} \left\{\bar{f}(x,y) + \bar{p}(x) + \rho\big(\tilde{f}(x,y) + \tilde{p}(y) - \min_z\{\tilde{f}(x,z) + \tilde{p}(z) | \tilde{g}(x,z) \leq 0\}\big)\right\},$$

which turns out to be a special case of (1) given by

$$\min_{\tilde{g}(x,y)\leq 0} \max_{\tilde{g}(x,z)\leq 0} \left\{\bar{f}(x,y) + \rho\big(\tilde{f}(x,y) - \tilde{f}(x,z)\big) + \bar{p}(x) - \rho\tilde{p}(z)\right\}.$$

In the recent years, the minimax problem of a simpler form

$$\min_{x\in X} \max_{y \in Y} f(x;y), \tag{4}$$

where $X$ and $Y$ are closed sets, has received tremendous amount of attention. Indeed, it has found broad applications in many areas, such as adversarial training [18, 35, 47, 53], generative adversarial networks [15, 17, 44], reinforcement learning [9, 13, 37, 40, 48], computational game [1, 42, 49], distributed computing [36, 46], prediction and regression [4, 50, 57, 58], and distributionally robust optimization [14, 45]. Numerous methods have been developed for solving (4) with $X$ and $Y$ being *simple closed convex sets* (e.g., see [7, 20, 22, 29, 30, 32, 34, 39, 55, 59, 60, 63]).

There have also been several studies on some other special cases of problem (1). In particular, two first-order methods, called max-oracle gradient-descent and nested gradient descent/ascent methods, were proposed in [16] for solving (1) with $c(x) \equiv 0$ and $p$ and $q$ being respectively the indicator function of simple compact convex sets $X$ and $Y$, under the assumption that $V(x) = \max_{y\in Y}\{f(x,y) : d(x,y) \leq 0\}$ is convex and moreover an optimal Lagrangian multiplier associated with the constraint $d(x,y) \leq 0$ can be computed for each $x \in X$. An augmented Lagrangian (AL) method was recently proposed in [12] for solving (1) with *only equality constraints*, $p(x) \equiv 0$, $q(y) \equiv 0$ and $c(x) \equiv 0$, under the assumption that a *local min-max point* of the AL subproblem can be found at each iteration. In addition, a multiplier gradient descent method was proposed in [52] for solving (1) with $c(x) \equiv 0$, $d(x,y)$ being an *affine* mapping, and $p$ and $q$ being the indicator function of simple compact convex sets. Also, a proximal gradient multi-step ascent decent method was developed in [10] for (1) with $c(x) \equiv 0$, $d(x,y)$ being an *affine* mapping and $f(x,y) = g(x) + x^T Ay - h(y)$, under the assumption that $f(x,y) - q(y)$ is *strongly concave* in $y$. Besides, primal dual alternating proximal gradient methods were proposed in [62] for (1) with $c(x) \equiv 0$, $d(x,y)$ being an *affine* mapping, and $\{f(x,y)$ being strongly concave in $y$ or $[q(y) \equiv 0$ and $f(x,y)$ being a linear function in $y]\}$. An iteration complexity of the method for finding an approximate stationary point of the aforementioned special minimax problem was established in [10, 16, 62], respectively. Yet, their operation complexity, measured by the number of fundamental operations such as evaluations of gradient of $f$ and proximal operator of $p$ and $q$, was not studied in these works.

There was no algorithmic development for (1) prior to our work, though optimality conditions of (1) were recently studied in [11]. In this paper, we propose a first-order AL method for solving (1). Specifically, given an iterate $(x^k, y^k)$ and a Lagrangian multiplier estimate $(\lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k)$ at the $k$th iteration, the next iterate $(x^{k+1}, y^{k+1})$ is obtained by finding an approximate stationary point of the AL subproblem

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

for some $\rho_k > 0$ through the use of a first-order method proposed in this paper, where $\mathcal{L}$ is the AL function of (1) defined as

$$\mathcal{L}(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}; \rho) = F(x, y) + \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{x}} + \rho c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}\|^2\right) - \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{y}} + \rho d(x,y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right), \tag{5}$$

which is a generalization of the AL function introduced in [12] 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 Section 1.1.

The main contributions of this paper are summarized below.

- We propose a first-order AL method for solving problem (1). To the best of our knowledge, this is the first yet implementable method for solving (1).
- We show that under some suitable assumptions, our first-order AL method enjoys an iteration complexity of $\mathcal{O}(\log \varepsilon^{-1})$ and an operation complexity of $\mathcal{O}(\varepsilon^{-4} \log \varepsilon^{-1})$, measured by the number of evaluations of $\nabla f$, $\nabla c$, $\nabla d$ and proximal operator of $p$ and $q$, for finding an $\varepsilon$-KKT solution of (1).

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 $\delta_S$ denote the indicator function associated with $S$.

A function or mapping $\phi$ is said to be $L_\phi$*-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 $x, x' \in S$. For a closed convex function $p : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the *proximal operator* associated with $p$ is denoted by $\mathrm{prox}_p$, that is,

$$\mathrm{prox}_p(x) = \arg\min_{x' \in \mathbb{R}^n} \left\{ \frac{1}{2}\|x' - x\|^2 + p(x') \right\} \quad \forall x \in \mathbb{R}^n. \tag{6}$$

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 $\operatorname{dom} \phi := \{x | \phi(x) < +\infty\}$. The *upper subderivative* of $\phi$ at $x \in \operatorname{dom} \phi$ 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 $\phi$ at $x \in \operatorname{dom} \phi$ is the set

$$\partial \phi(x) = \{s \in \mathbb{R}^n \big| s^T d \leq \phi'(x; d) \quad \forall d \in \mathbb{R}^n\}.$$

We use $\partial_{x_i} \phi(x)$ to denote the subdifferential with respect to $x_i$. In addition, for an upper semicontinuous function $\phi$, its subdifferential is defined as $\partial \phi = -\partial(-\phi)$. If $\phi$ is locally Lipschitz continuous, the above definition of subdifferential coincides with the Clarke subdifferential. Besides, if $\phi$ is convex, it coincides with the ordinary subdifferential for convex functions. Also, if $\phi$ is continuously differentiable at $x$ , we simply have $\partial \phi(x) = \{\nabla \phi(x)\}$, where $\nabla \phi(x)$ is the gradient of $\phi$ 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 $\phi_1$ is continuously differentiable at $x$ and $\phi_2$ is lower or upper semicontinuous at $x$. See [8, 54] for more details.

Finally, we introduce an (approximate) primal-dual stationary point (e.g., see [10, 11, 26]) for a general minimax problem

$$\min_x \max_y \Psi(x, y), \tag{7}$$

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* (7) *if*

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

*In addition, for any $\epsilon > 0$, a point $(x_\epsilon, y_\epsilon)$ is said to be an $\epsilon$-primal-dual stationary point of the minimax problem* (7) *if*

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

One can see that $(x_\epsilon, y_\epsilon)$ is an $\epsilon$-primal-dual stationary point of (7) if and only if $x_\epsilon$ and $y_\epsilon$ are an $\epsilon$-stationary point of $\min_x \Psi(x, y_\epsilon)$ and $\max_y \Psi(x_\epsilon, y)$, respectively.

## 2 A first-order method for nonconvex-concave minimax problem

In this section, we propose a first-order method for finding an $\epsilon$-primal-dual stationary point of a nonconvex-concave minimax problem introduced in Definition 1, which will be used as a subproblem solver for the first-order AL method proposed in Section 3. In particular, we consider the minimax problem

$$H^* = \min_x \max_y \{H(x, y) := h(x, y) + p(x) - q(y)\}. \tag{8}$$

Assume that problem (8) 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* $\operatorname{dom} p \times \operatorname{dom} q$, *and moreover, $h(x, \cdot)$ is concave for any $x \in$* $\operatorname{dom} p$.

Numerous algorithms have been developed for finding an approximate stationary point of the special case of (8) with $p, q$ being the indicator function of a closed convex set (e.g., see [23, 30, 39, 41, 51, 61]). They are however not applicable to (8) in general. Recently, an accelerated inexact proximal point smoothing (AIPP-S) scheme was proposed in [26] for finding an approximate stationary point of a class of minimax composite nonconvex optimization problems, which includes (8) as a special case. When applied to (8), AIPP-S requires the availability of the oracle including exact evaluation of $\nabla_x h(x,y)$ and

$$\arg\min_x \left\{ p(x) + \frac{1}{2\lambda}\|x - x'\|^2 \right\}, \qquad \arg\max_y \left\{ h(x',y) - q(y) - \frac{1}{2\lambda}\|y - y'\|^2 \right\} \tag{9}$$

for any $\lambda > 0$, $x' \in \mathbb{R}^n$ and $y' \in \mathbb{R}^m$. Notice that $h$ is typically sophisticated and the *exact* solution of the second problem in (9) usually cannot be found. As a result, AIPP-S is generally not implementable for (8), though an operation complexity of $\mathcal{O}(\epsilon^{-5/2})$, measured by the number of evaluations of the aforementioned oracle, was established in [26] for it to find an $\epsilon$-primal-dual stationary point of (8). In addition, a first-order method was proposed in [64] enjoying an operation complexity of $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$, measured by the number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$, for finding an $\epsilon$-primal stationary point $x'$ of (8) satisfying

$$\left\| \lambda^{-1}(x' - \arg\min_x \left\{ \max_y H(x,y) + \frac{1}{2\lambda}\|x - x'\|^2 \right\} \right\| \le \epsilon$$

for some $0 < \lambda < L_{\nabla h}^{-1}$. One can see that such $x'$ is an approximate stationary point of (8) by viewing it as a minimization problem. Consequently, this method does not suit our need since we aim to find an $\epsilon$-primal-dual stationary point of (8) introduced in Definition 1.

In what follows, we first propose a modified optimal first-order method for solving a strongly-convex-strongly-concave minimax problem in Subsection 2.1. Using this method as a subproblem solver for an inexact proximal point scheme, we then propose a first-order method for (8) in Subsection 2.2, which enjoys an operation complexity of $\mathcal{O}(\epsilon^{-5/2}\log\epsilon^{-1})$, measured by the number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$, for finding an $\epsilon$-primal-dual stationary point of (8).

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

In this subsection, we consider the strongly-convex-strongly-concave minimax problem

$$\bar{H}^* = \min_x \max_y \left\{ \bar{H}(x,y) := \bar{h}(x,y) + p(x) - q(y) \right\}, \tag{10}$$

where $p, q$ satisfy Assumption 1 and $\bar{h}$ satisfies the following assumption.

**Assumption 3.** *$\bar{h}(x,y)$ is $\sigma_x$-strongly-convex-$\sigma_y$-strongly-concave and $L_{\nabla\bar{h}}$-smooth on* $\operatorname{dom} p \times \operatorname{dom} q$ *for some $\sigma_x, \sigma_y > 0$.*

Recently, a novel optimal first-order method [27, Algorithm 4] was proposed for solving (10). Though the solution sequence of this method converges to the optimal solution with an optimal rate, it lacks a verifiable termination criterion and also the approximate solution found by it may never be an $\bar{\epsilon}$-primal-dual stationary point of (10) (see Definition 1) for a prescribed tolerance $\bar{\epsilon} > 0$. To tackle these issues, we next propose an optimal first-order method by modifying [27, Algorithm 4] for finding an approximate primal-dual stationary point of (10). Before proceeding, we introduce some notation below, most of which is adopted from [27].

Recall that $\mathcal{X} = \operatorname{dom} p$ and $\mathcal{Y} = \operatorname{dom} q$. Let $(x^*, y^*)$ denote the optimal solution of (10),

$z^* = -\sigma_x x^*$, and

$$D_{\mathbf{x}} := \max\{\|u-v\| \big| u,v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u-v\| \big| u,v \in \mathcal{Y}\}, \tag{11}$$

$$\bar{H}_{\rm low} = \min\left\{\bar{H}(x,y) \middle| \big(x,y) \in \mathcal{X}\times\mathcal{Y}\right\}, \tag{12}$$

$$\hat{h}(x,y) = \bar{h}(x,y) - \sigma_x\|x\|^2/2 + \sigma_y\|y\|^2/2, \tag{13}$$

$$\mathcal{G}(z,y) = \sup_x\{\langle x,z\rangle - p(x) - \hat{h}(x,y) + q(y)\}, \tag{14}$$

$$\mathcal{P}(z,y) = \sigma_x^{-1}\|z\|^2/2 + \sigma_y\|y\|^2/2 + \mathcal{G}(z,y), \tag{15}$$

$$\vartheta_k = \eta_z^{-1}\|z^k - z^*\|^2 + \eta_y^{-1}\|y^k - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(z_f^k, y_f^k) - \mathcal{P}(z^*, y^*)), \tag{16}$$

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

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 1 below. By Assumptions 1 and 3, one can observe that $D_{\mathbf{x}}$, $D_{\mathbf{y}}$ and $\bar{H}_{\rm low}$ are finite.

We are now ready to present a modified optimal first-order method for solving (10) in Algorithm 1. It is a slight modification of the novel optimal first-order method [27, Algorithm 4] by incorporating a forward-backward splitting scheme and also a verifiable termination criterion (see steps 23-25 in Algorithm 1) in order to find an $\bar{\epsilon}$-primal-dual stationary point of (10) (see Definition 1) for any prescribed tolerance $\bar{\epsilon} > 0$.

**Algorithm 1** A modified optimal first-order method for (10)

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

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

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

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

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

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

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

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

$\quad t = 0$.

$\quad$ **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**

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

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

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

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

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

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

$\quad\quad t \leftarrow t+1$.

$\quad$ **end while**

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

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

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

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

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

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

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

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

$$\|\bar\zeta^{-1}(x^{k+1} - \tilde x^{k+1}, \tilde y^{k+1} - y^{k+1}) - (\nabla\bar h(x^{k+1}, y^{k+1}) - \nabla\bar h(\tilde x^{k+1}, \tilde y^{k+1}))\| \le \bar\epsilon. \tag{17}$$

**end for**

The following theorem presents *iteration and operation complexity* of Algorithm 1 for finding an $\bar\epsilon$-primal-dual stationary point of problem (10), whose proof is deferred to Subsection 4.1.

**Theorem 1 (Complexity of Algorithm 1).** *Suppose that Assumptions 1 and 3 hold. Let $\bar H^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\bar H_{\mathrm{low}}$, and $\vartheta_0$ be defined in (10), (11), (12) and (16), $\sigma_x$, $\sigma_y$ and $L_{\nabla\bar h}$ be given in Assumption 3, $\bar\alpha$, $\eta_y$, $\eta_z$, $\bar\epsilon$, $\bar\zeta$ be given in Algorithm 1, and*

$$\bar\delta = (2+\bar\alpha^{-1})\sigma_x D_{\mathbf{x}}^2 + \max\{2\sigma_y, \bar\alpha\sigma_x/4\} D_{\mathbf{y}}^2, \tag{18}$$

$$\bar K = \left\lceil \max\left\{\frac{2}{\bar\alpha}, \frac{\bar\alpha\sigma_x}{4\sigma_y}\right\} \log\frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\bar\zeta^{-1} + L_{\nabla\bar h})^{-2}\bar\epsilon^2}\right\rceil_+, \tag{19}$$

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

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

$$\times \left( \left\lceil 96\sqrt{2}\left(1 + 8L_{\nabla \bar{h}}\sigma_x^{-1}\right) \right\rceil + 2 \right). \tag{20}$$

*Then Algorithm 1 outputs an $\bar{\epsilon}$-primal-dual stationary point of* (10) *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 1 is no more than $\bar{N}$, respectively.*

**Remark 1.** *It can be observed from Theorem 1 that Algorithm 1 enjoys an operation complexity of $\mathcal{O}(\log(1/\bar{\epsilon}))$, measured by the number of evaluations of $\nabla \bar{h}$ and proximal operator of $p$ and $q$, for finding an $\bar{\epsilon}$-primal-dual stationary point of the strongly-convex-strongly-concave minimax problem* (10).

## 2.2 A first-order method for problem (8)

In this subsection, we propose a first-order method for finding an $\epsilon$-primal-dual stationary point of problem (8) (see Definition 1) for any prescribed tolerance $\epsilon > 0$. In particular, we first add a perturbation to the max part of (8) for obtaining an approximation of (8), which is given as follows:

$$\min_x \max_y \left\{ h(x,y) + p(x) - q(y) - \frac{\epsilon}{4D_{\mathbf{y}}}\|y - \hat{y}^0\|^2 \right\} \tag{21}$$

for some $\hat{y}^0 \in \operatorname{dom} q$, where $D_{\mathbf{y}}$ is given in (11). We then apply an inexact proximal point method [25] to (21), which consists of approximately solving a sequence of subproblems

$$\min_x \max_y \left\{ H_k(x,y) := h_k(x,y) + p(x) - q(y) \right\}, \tag{22}$$

where

$$h_k(x,y) = h(x,y) - \epsilon\|y - \hat{y}^0\|^2/(4D_{\mathbf{y}}) + L_{\nabla h}\|x - x^k\|^2. \tag{23}$$

By Assumption 2, one can observe that (i) $h_k$ is $L_{\nabla h}$-strongly convex in $x$ and $\epsilon/(2D_{\mathbf{y}})$-strongly concave in $y$ on $\operatorname{dom} p \times \operatorname{dom} q$; (ii) $h_k$ is $(3L_{\nabla h} + \epsilon/(2D_{\mathbf{y}}))$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$. Consequently, problem (22) is a special case of (10) and can be suitably solved by Algorithm 1. The resulting first-order method for (8) is presented in Algorithm 2.

---

**Algorithm 2** A first-order method for problem (8)

---

**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 1 with $\bar{h} \leftarrow h_k$, $\bar{\epsilon} \leftarrow \hat{\epsilon}_k$, $\sigma_x \leftarrow L_{\nabla h}$, $\sigma_y \leftarrow \epsilon/(2D_{\mathbf{y}})$, $L_{\nabla \bar{h}} \leftarrow 3L_{\nabla h} + \epsilon/(2D_{\mathbf{y}})$, $\bar{z}^0 = z_f^0 \leftarrow -\sigma_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$, and denote its output by $(x^{k+1}, y^{k+1})$, where $h_k$ is given in (23).

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

$$\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h}). \tag{24}$$

4: **end for**

---

**Remark 2.** *It is seen from step 2 of Algorithm 2 that $(x^{k+1}, y^{k+1})$ results from applying Algorithm 1 to the subproblem* (22). *As will be shown in Lemma 2, $(x^{k+1}, y^{k+1})$ is an $\hat{\epsilon}_k$-primal-dual stationary point of* (22).

We next study complexity of Algorithm 2 for finding an $\epsilon$-primal-dual stationary point of problem (8). Before proceeding, we define

$$H_{\text{low}} := \min\left\{ H(x,y) | (x,y) \in \operatorname{dom} p \times \operatorname{dom} q \right\}. \tag{25}$$

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

The following theorem presents *iteration and operation complexity* of Algorithm 2 for finding an $\epsilon$-primal-dual stationary point of problem (8), whose proof is deferred to Subsection 4.2.

**Theorem 2** (**Complexity of Algorithm 2**)**.** *Suppose that Assumption 2 holds. Let $H^*$, $H$ $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, and $H_{\mathrm{low}}$ be defined in* (8), (11) *and* (25), *$L_{\nabla h}$ be given in Assumption 2, $\epsilon$, $\hat{\epsilon}_0$ and $\hat{x}^0$ be given in Algorithm 2, and*

$$\hat{\alpha} = \min\left\{1, \sqrt{4\epsilon/(D_{\mathbf{y}}L_{\nabla h})}\right\}, \tag{26}$$

$$\hat{\delta} = (2+\hat{\alpha}^{-1})L_{\nabla h}D_{\mathbf{x}}^2 + \max\left\{\epsilon/D_{\mathbf{y}}, \hat{\alpha}L_{\nabla h}/4\right\}D_{\mathbf{y}}^2, \tag{27}$$

$$\widehat{T} = \left\lceil 16(\max_y H(\hat{x}^0,y) - H^* + \epsilon D_{\mathbf{y}}/4)L_{\nabla h}\epsilon^{-2} + 32\hat{\epsilon}_0^2(1+4D_{\mathbf{y}}^2L_{\nabla h}^2\epsilon^{-2})\epsilon^{-2} - 1\right\rceil_+, \tag{28}$$

$$\begin{aligned}\widehat{N} = &\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_{\mathbf{y}})L_{\nabla h}^{-1}\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L_{\nabla h}\epsilon^{-1}}\right\}\\ &\times\left((\widehat{T}+1)\left(\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_{\mathbf{y}}}{\epsilon}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\mathrm{low}}+\epsilon D_{\mathbf{y}}/4+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{[(3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}}))^2/\min\{L_{\nabla h},\epsilon/(2D_{\mathbf{y}})\}+3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}})]^{-2}\,\hat{\epsilon}_0^2}\right)_+\right.\\ &\left.+\widehat{T}+1+2\widehat{T}\log(\widehat{T}+1)\right).\end{aligned} \tag{29}$$

*Then Algorithm 2 terminates and outputs an $\epsilon$-primal-dual stationary point $(x_\epsilon, y_\epsilon)$ of* (8) *in at most $\widehat{T}+1$ outer iterations that satisfies*

$$\max_y H(x_\epsilon, y) \le \max_y H(\hat{x}^0, y) + \epsilon D_{\mathbf{y}}/4 + 2\hat{\epsilon}_0^2\left(L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2L_{\nabla h}\epsilon^{-2}\right). \tag{30}$$

*Moreover, the total number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$ performed in Algorithm 2 is no more than $\widehat{N}$, respectively.*

**Remark 3.** *Since $\hat{\epsilon}_0 \in (0, \epsilon/2]$, one can observe from Theorem 2 that $\hat{\alpha} = \mathcal{O}(\epsilon^{1/2})$, $\hat{\delta} = \mathcal{O}(\epsilon^{-1/2})$, $\widehat{T} = \mathcal{O}(\epsilon^{-2})$, and $\widehat{N} = \mathcal{O}(\epsilon^{-5/2}\log(\hat{\epsilon}_0^{-1}\epsilon^{-1}))$. Consequently, Algorithm 2 enjoys an operation complexity of $\mathcal{O}(\epsilon^{-5/2}\log(\hat{\epsilon}_0^{-1}\epsilon^{-1}))$, measured by the number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$, for finding an $\epsilon$-primal-dual stationary point of the nonconvex-concave minimax problem* (8).

# 3 A first-order augmented Lagrangian method for problem (1)

In this section, we propose a first-order augmented Lagrangian (FAL) method for problem (1), and study its complexity for finding an approximate KKT point of (1).

One standard approach for solving constrained nonlinear program is to solve a sequence of unconstrained nonlinear program problems, which are typically penalty or augmented Lagrangian subproblems (e.g., see [38]). In a similar spirit, we next propose an FAL method in Algorithm 3 for solving (1). In particular, at each iteration, the FAL method finds an approximate primal-dual stationary point of an AL subproblem in the form of

$$\min_x \max_y \mathcal{L}(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}; \rho), \tag{31}$$

where $\mathcal{L}$ is the AL function associated with problem (1) defined in (5), $\lambda_{\mathbf{x}} \in \mathbb{R}_+^{\tilde{n}}$ and $\lambda_{\mathbf{y}} \in \mathbb{R}_+^{\tilde{m}}$ are a Lagrangian multiplier estimate, and $\rho > 0$ is a penalty parameter, which are updated by a standard scheme. In view of Assumption 1, one can observe that $\mathcal{L}$ enjoys the following nice structure.

- For any given $\rho > 0$, $\lambda_{\mathbf{x}} \in \mathbb{R}_+^{\tilde{n}}$ and $\lambda_{\mathbf{y}} \in \mathbb{R}_+^{\tilde{m}}$, $\mathcal{L}$ is the sum of smooth function $f(x,y) + \left(\|[\lambda_{\mathbf{x}}+\rho c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}\|^2\right)/(2\rho) - \left(\|[\lambda_{\mathbf{y}}+\rho d(x,y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right)/(2\rho)$ with Lipschitz continuous gradient and possibly nonsmooth function $p(x) - q(y)$ with exactly computable proximal operator.

- $\mathcal{L}$ is nonconvex in $x$ but concave in $y$.

Thanks to the above nice structure of $\mathcal{L}$, we will use Algorithm 2 as a solver to find an approximate primal-dual stationary point of the AL subproblem (31).

Recall that $\mathcal{X} = \operatorname{dom} p$ and $\mathcal{Y} = \operatorname{dom} q$. Before presenting an FAL method for (1), we let

$$\mathcal{L}_{\mathbf{x}}(x, y, \lambda_{\mathbf{x}}; \rho) := F(x, y) + \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{x}} + \rho c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}\|^2\right), \tag{32}$$

$$c_{\mathrm{hi}} := \max\{\|c(x)\| \,|\, x \in \mathcal{X}\}, \quad d_{\mathrm{hi}} := \max\{\|d(x, y)\| \,|\, (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{33}$$

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 $\min_x\{F(x, y)|c(x) \leq 0\}$ 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). Due to 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 3) so that the resulting FAL method outputs an approximate KKT point of (1) nearly satisfying the constraint $c(x) \leq 0$.

**Assumption 4.** *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 $\|[c(x_{\mathbf{nf}})]_+\| \leq \sqrt{\varepsilon}$, can be found.*

**Remark 4.** *A very similar assumption as Assumption 4 was considered in [6, 19, 33, 56]. In addition, when the error bound condition $\|[c(x)]_+\| = \mathcal{O}(\operatorname{dist}(0, \partial(\|[c(x)]_+\|^2 + \delta_{\mathcal{X}}(x)))^\nu)$ holds on a level set of $\|[c(x)]_+\|$ for some $\nu > 0$, Assumption 4 holds for problem* (1) *(e.g., see [31, 43]). 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 an FAL method for solving problem (1).

---

**Algorithm 3** A first-order augmented Lagrangian method for problem (1)

---

**Input:** $\varepsilon, \tau \in (0, 1)$, $\epsilon_k = \tau^k$, $\rho_k = \epsilon_k^{-1}$, $\Lambda > 0$, $\lambda_{\mathbf{x}}^0 \in \mathbb{B}_\Lambda^+$, $\lambda_{\mathbf{y}}^0 \in \mathbb{R}_+^{\tilde{m}}$, $(x^0, y^0) \in \operatorname{dom} p \times \operatorname{dom} q$, and $x_{\mathbf{nf}} \in \operatorname{dom} p$ with $\|[c(x_{\mathbf{nf}})]_+\| \leq \sqrt{\varepsilon}$ (see Assumption 4).

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

2: Set

$$x_{\mathrm{init}}^k = \begin{cases} x^k, & \text{if } \mathcal{L}_{\mathbf{x}}(x^k, y^k, \lambda_{\mathbf{x}}^k; \rho_k) \leq \mathcal{L}_{\mathbf{x}}(x_{\mathbf{nf}}, y^k, \lambda_{\mathbf{x}}^k; \rho_k), \\ x_{\mathbf{nf}}, & \text{otherwise.} \end{cases} \tag{34}$$

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

$$\min_x \max_y \mathcal{L}(x, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \tag{35}$$

where

$$L_k = L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\mathrm{hi}} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\mathrm{hi}} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d}. \tag{36}$$

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

6: **end for**

---

**Remark 5.** *(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, 24], which has been shown to enjoy many practical and theoretical advantages (see [2] for discussions).*

*(ii) In view of Theorem 2, one can see that an $\epsilon_k$-primal-dual stationary point of* (35) *can be successfully found in step 3 of Algorithm 3 by applying Algorithm 2 to problem* (35). *Consequently, Algorithm 3 is well-defined.*

### 3.1 Complexity results for Algorithm 3

In this subsection we study iteration and operation complexity for Algorithm 3. Recall that $\mathcal{X} = \mathrm{dom}\, p$ and $\mathcal{Y} = \mathrm{dom}\, q$. Before proceeding, we make one additional assumption below 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 10). 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 6 and 7).

**Assumption 5.** *(i) There exist some constants $\delta_c$, $\theta > 0$ 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 $v_x^T \nabla c_i(x) \leq -\delta_c$ for all $i \in \mathcal{A}(x;\theta)$, where $\mathcal{T}_{\mathcal{X}}(x)$ is the tangent cone of $\mathcal{X}$ at $x$, and*

$$\mathcal{F}(\theta) = \{x \in \mathcal{X} \big| \|[c(x)]_+\|_\infty \leq \theta\}, \quad \mathcal{A}(x;\theta) = \{i | c_i(x) \geq -\theta,\ 1 \leq i \leq \tilde{n}\}. \tag{37}$$

*(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 6.** *(i) Assumption 5(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 $\theta$.*

*(ii) The latter part of Assumption 5(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 5(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 5(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 5(iii) is mild.*

In order to characterize the approximate solution found by Algorithm 3, we next introduce a notion called an $\varepsilon$-KKT solution of problem (1).

One can observe from Lemma 4(iii) in Subsection 4.3 that problem (1) is equivalent to

$$\min_{x, \lambda_{\mathbf{y}}} \big\{ \max_{y} F(x, y) - \langle \lambda_{\mathbf{y}}, d(x, y) \rangle + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}}) \big| c(x) \leq 0 \big\}.$$

By this, one can further see that problem (1) is equivalent to

$$\min_{x, \lambda_{\mathbf{y}}} \max_{\lambda_{\mathbf{x}}} \Big\{ \max_{y} \{F(x, y) - \langle \lambda_{\mathbf{y}}, d(x, y) \rangle + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}})\} + \langle \lambda_{\mathbf{x}}, c(x) \rangle - \delta_{\mathbb{R}_+^{\tilde{n}}}(\lambda_{\mathbf{x}}) \Big\},$$

which is a nonconvex-concave minimax problem

$$\min_{x, \lambda_{\mathbf{y}}} \max_{y, \lambda_{\mathbf{x}}} \Big\{ F(x, y) + \langle \lambda_{\mathbf{x}}, c(x) \rangle - \langle \lambda_{\mathbf{y}}, d(x, y) \rangle - \delta_{\mathbb{R}_+^{\tilde{n}}}(\lambda_{\mathbf{x}}) + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}}) \Big\}. \tag{38}$$

It follows from [11, Theorem 3.1] that if $(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{\tilde{n}}_+ \times \mathbb{R}^{\tilde{m}}_+$ is a local minimax point of problem (38), then it must also be a primal-dual stationary point of (38). This, combined with Definition 1, implies that $(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}})$ is a KKT point of (38) satisfying the conditions:

$$0 \in \partial_x F(x, y) + \nabla c(x)\lambda_{\mathbf{x}} - \nabla_x d(x, y)\lambda_{\mathbf{y}}, \tag{39}$$

$$0 \in \partial_y F(x, y) - \nabla_y d(x, y)\lambda_{\mathbf{y}}, \tag{40}$$

$$c(x) \leq 0, \quad \langle \lambda_{\mathbf{x}}, c(x)\rangle = 0, \tag{41}$$

$$d(x, y) \leq 0, \quad \langle \lambda_{\mathbf{y}}, d(x, y)\rangle = 0. \tag{42}$$

Based on this observation and the equivalence of (1) and (38), we introduce an (approximate) KKT solution for problem (1) below.

**Definition 2.** *The pair $(x, y)$ is said to be a KKT solution of problem* (1) *if there exists $(\lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}) \in \mathbb{R}^{\tilde{n}}_+ \times \mathbb{R}^{\tilde{m}}_+$ such that the conditions* (39)*-*(42) *hold. In addition, for any $\varepsilon > 0$, $(x, y)$ is said to be an $\varepsilon$-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*

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

Recall that $\mathcal{X} = \mathrm{dom}\, p$ and $\mathcal{Y} = \mathrm{dom}\, q$. To study complexity of Algorithm 3, we define

$$f^*(x) := \max\{F(x, y) | d(x, y) \leq 0\}, \tag{43}$$

$$F_{\mathrm{hi}} := \max\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \quad F_{\mathrm{low}} := \min\{F(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{44}$$

$$\Delta := F_{\mathrm{hi}} - F_{\mathrm{low}}, \quad r := 2\delta_d^{-1}\Delta, \tag{45}$$

$$K := \lceil \log\varepsilon / \log\tau \rceil_+, \quad \mathbb{K} := \{0, 1, \ldots, K+1\}, \tag{46}$$

where $\delta_d$ is given in Assumption 5, and $\varepsilon$ and $\tau$ are some input parameters of Algorithm 3. For convenience, we define $\mathbb{K} - 1 = \{k - 1 | k \in \mathbb{K}\}$. One can observe from Assumption 1 that $F_{\mathrm{hi}}$ and $F_{\mathrm{low}}$ are finite. Besides, one can easily observe that

$$f^*(x) \geq F_{\mathrm{low}}, \;\; F(x, y) - f^*(x) \leq \Delta \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{47}$$

We are now ready to present an *iteration and operation complexity* of Algorithm 3 for finding an $\mathcal{O}(\varepsilon)$-KKT solution of problem (1), whose proof is deferred to Section 4.

**Theorem 3.** *Suppose that Assumptions 1, 4 and 5 hold. Let $\{(x^k, y^k, \lambda^k_{\mathbf{x}}, \lambda^k_{\mathbf{y}})\}_{k \in \mathbb{K}}$ be generated by Algorithm 3, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $c_{\mathrm{hi}}$, $d_{\mathrm{hi}}$, $\Delta$ and $K$ be defined in* (11), (33), (45) *and* (46)*, $L_F$, $L_{\nabla f}$, $L_{\nabla d}$, $L_{\nabla c}$, $L_c$, $L_{\nabla d}$, $L_d$, $\delta_c$, $\delta_d$ and $\theta$ be given in Assumptions 1 and 5, $\varepsilon$, $\tau$, $\Lambda$ and $\lambda^0_{\mathbf{y}}$ be given in Algorithm 3, and*

$$L = L_{\nabla f} + L_c^2 + c_{\mathrm{hi}}L_{\nabla c} + \Lambda L_{\nabla c} + L_d^2 + d_{\mathrm{hi}}L_{\nabla d} + L_{\nabla d}\sqrt{\|\lambda^0_{\mathbf{y}}\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}}, \tag{48}$$

$$\alpha = \min\left\{1, \sqrt{4/(D_{\mathbf{y}}L)}\right\}, \quad \delta = (2 + \alpha^{-1})LD_{\mathbf{x}}^2 + \max\{1/D_{\mathbf{y}}, L/4\}D_{\mathbf{y}}^2, \tag{49}$$

$$\begin{aligned}
M = {} & 16\max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}\left[(3L + 1/(2D_{\mathbf{y}}))^2/\min\{L_c^2, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^2\\
& \times \left(\delta + 2\alpha^{-1}\left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda^0_{\mathbf{y}}\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1-\tau} + \rho_k d_{\mathrm{hi}}^2 + \frac{D_{\mathbf{y}}}{4} + LD_{\mathbf{x}}^2\right)\right),
\end{aligned} \tag{50}$$

$$T = \left\lceil 16L\left(2\Delta + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda^0_{\mathbf{y}}\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} + \frac{\Lambda^2}{2} + \frac{D_{\mathbf{y}}}{4}\right) + 8(1 + 4D_{\mathbf{y}}^2L^2)\right\rceil_+, \tag{51}$$

$$\tilde{\lambda}^{K+1}_{\mathbf{x}} = [\lambda^K_{\mathbf{x}} + \tau^{-K}c(x^{K+1})]_+. \tag{52}$$

*Suppose that*

$$\varepsilon^{-1} \geq \max\left\{\theta^{-1}\Lambda, \theta^{-2}\Big\{4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{D_{\mathbf{y}}}{2} + L_c^{-2} + 4D_{\mathbf{y}}^2 L + \Lambda^2\Big\},\right.$$
$$\left.\frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2\tau(1-\tau)}\right\}. \tag{53}$$

*Then the following statements hold.*

*(i) Algorithm 3 terminates after* $K+1$ *outer iterations and outputs an approximate stationary point* $(x^{K+1}, y^{K+1})$ *of* (1) *satisfying*

$$\mathrm{dist}(0, \partial_x F(x^{K+1}, y^{K+1}) + \nabla c(x^{K+1})\tilde{\lambda}_x^{K+1} - \nabla_x d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}) \leq \varepsilon, \tag{54}$$
$$\mathrm{dist}\left(0, \partial_y F(x^{K+1}, y^{K+1}) - \nabla_y d(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\right) \leq \varepsilon, \tag{55}$$
$$\|[c(x^{K+1})]_+\| \leq \varepsilon\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1\right), \tag{56}$$
$$\begin{aligned}|\langle\tilde{\lambda}_{\mathbf{x}}^{K+1}, c(x^{K+1})\rangle| \leq{}& \varepsilon\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1)\\ &\times \max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1), \Lambda\},\end{aligned} \tag{57}$$
$$\|[d(x^{K+1}, y^{K+1})]_+\| \leq 2\varepsilon\delta_d^{-1}(\Delta + D_{\mathbf{y}}), \tag{58}$$
$$|\langle\lambda_{\mathbf{y}}^{K+1}, d(x^{K+1}, y^{K+1})\rangle| \leq 2\varepsilon\delta_d^{-1}(\Delta + D_{\mathbf{y}})\max\{2\delta_d^{-1}(\Delta + D_{\mathbf{y}}), \|\lambda_{\mathbf{y}}^0\|\}. \tag{59}$$

*(ii) The total number of evaluations of* $\nabla f$*,* $\nabla c$*,* $\nabla d$ *and proximal operator of* $p$ *and* $q$ *performed in Algorithm 3 is at most* $N$*, respectively, where*

$$\begin{aligned}N = {}& \left(\left\lceil 96\sqrt{2}\left(1 + (24L + 4/D_{\mathbf{y}})/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\} T(1-\tau^4)^{-1}\\ &\times (\tau\varepsilon)^{-4}\left(28K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right).\end{aligned} \tag{60}$$

**Remark 7.** *(i) The condition* (53) *on* $\varepsilon$ *is to ensure that the final penalty parameter* $\rho_K$ *in Algorithm 3 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 3 that Algorithm 3 enjoys an iteration complexity of* $\mathcal{O}(\log\varepsilon^{-1})$ *and an operation complexity of* $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$*, measured by the number of evaluations of* $\nabla f$*,* $\nabla c$*,* $\nabla d$ *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*

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

*where* $\tilde{\lambda}_{\mathbf{x}}^{K+1} \in \mathbb{R}_+^{\tilde{n}}$ *is defined in* (52) *and* $\lambda_{\mathbf{y}}^{K+1} \in \mathbb{R}_+^{\tilde{m}}$ *is given in Algorithm 3.*

# 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, 2 and 3.

## 4.1 Proof of the main results in Subsection 2.1

In this subsection we prove Theorem 1. Before proceeding, we establish an upper bound on $\vartheta_0$ in terms of the function value gap of (10), where $\vartheta_0$ is given in (16).

**Lemma 1.** *Suppose that Assumptions 2 and 3 hold. Let $\bar{H}^*$, $\bar{H}_{\rm low}$, $\vartheta_0$ and $\bar{\delta}$ be defined in* (10), (12), (16) *and* (18), *and $\bar{\alpha}$ be given in Algorithm 1. Then we have*

$$\vartheta_0 \le \bar{\delta} + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\rm low}\right). \tag{61}$$

*Proof.* By (10), (12), (13) and (14), one has

$$\begin{aligned}
\mathcal{G}(\bar{z}^0, \bar{y}^0) &\overset{(14)}{=} \sup_x \left\{\langle x, \bar{z}^0\rangle - p(x) - \hat{h}(x, \bar{y}^0) + q(\bar{y}^0)\right\} \\
&\overset{(13)}{=} \max_{x\in{\rm dom}\, p} \left\{\langle x, \bar{z}^0\rangle - p(x) - \bar{h}(x, \bar{y}^0) + \frac{\sigma_x}{2}\|x\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 + q(\bar{y}^0)\right\} \\
&\overset{(10)(12)}{\le} \max_{x\in{\rm dom}\, p} \left\{\langle x, \bar{z}^0\rangle + \frac{\sigma_x}{2}\|x\|^2\right\} - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low} \\
&= \max_{x\in{\rm dom}\, p} \frac{\sigma_x}{2}\|x + \sigma_x^{-1}\bar{z}^0\|^2 - \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low} \\
&\le \frac{\sigma_x D_{\mathbf{x}}^2}{2} - \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low},
\end{aligned} \tag{62}$$

where the last inequality follows from (11), $\mathcal{X} = {\rm dom}\, p$, and the fact that $z^0 \in -\sigma_x {\rm dom}\, p$.

Recall that $(x^*, y^*)$ is the optimal solution of (10) and $z^* = -\sigma_x x^*$. It follows from (10), (13) and (14) that

$$\begin{aligned}
\mathcal{G}(z^*, y^*) &\overset{(14)}{=} \sup_x \left\{\langle x, z^*\rangle - p(x) - \hat{h}(x, y^*) + q(y^*)\right\} \ge \langle x^*, z^*\rangle - p(x^*) - \hat{h}(x^*, y^*) + q(y^*) \\
&\overset{(13)}{=} \langle x^*, z^*\rangle + \frac{\sigma_x}{2}\|x^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - p(x^*) - \bar{h}(x^*, y^*) + q(y^*) \\
&= -\frac{\sigma_x^{-1}}{2}\|z^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - \bar{H}^*,
\end{aligned}$$

where the last equality follows from (10), the definition of $(x^*, y^*)$, and $z^* = -\sigma_x x^*$. This together with (15) and (62) implies that

$$\begin{aligned}
\mathcal{P}(\bar{z}^0, \bar{y}^0) - \mathcal{P}(z^*, y^*) &= \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 + \frac{\sigma_y}{2}\|\bar{y}^0\|^2 + \mathcal{G}(\bar{z}^0, \bar{y}^0) - \frac{\sigma_x^{-1}}{2}\|z^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - \mathcal{G}(z^*, y^*) \\
&\le \sigma_x D_{\mathbf{x}}^2/2 - \bar{H}_{\rm low} + \bar{H}^*.
\end{aligned}$$

Notice from Algorithm 1 that $z^0 = z_f^0 = \bar{z}^0 \in -\sigma_x {\rm dom}\, p$ and $y^0 = y_f^0 = \bar{y}^0 \in {\rm dom}\, q$. By these, $z^* = -\sigma_x x^*$, $\mathcal{X} = {\rm dom}\, p$, $\mathcal{Y} = {\rm dom}\, q$, (11), (16), and the above inequality, one has

$$\begin{aligned}
\vartheta_0 &\overset{(16)}{=} \eta_z^{-1}\|\bar{z}^0 - z^*\|^2 + \eta_y^{-1}\|\bar{y}^0 - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(\bar{z}^0, \bar{y}^0) - \mathcal{P}(z^*, y^*)) \\
&\le \eta_z^{-1}\sigma_x^2 D_{\mathbf{x}}^2 + \eta_y^{-1} D_{\mathbf{y}}^2 + 2\bar{\alpha}^{-1}\left(\sigma_x D_{\mathbf{x}}^2/2 - \bar{H}_{\rm low} + \bar{H}^*\right) \\
&= \eta_z^{-1}\sigma_x^2 D_{\mathbf{x}}^2 + \bar{\alpha}^{-1}\sigma_x D_{\mathbf{x}}^2 + \eta_y^{-1} D_{\mathbf{y}}^2 + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\rm low}\right).
\end{aligned}$$

Hence, the conclusion follows from this, (18), $\eta_z = \sigma_x/2$ and $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$. ∎

We are now ready to prove Theorem 1, using Lemma 1, [27, Theorem 3], [27, Lemma 4], and [5, Corollary 2.5].

***Proof of Theorem 1***. Suppose for contradiction that Algorithm 1 runs for more than $\bar{K}$ outer iterations, where $\bar{K}$ is given in (19). By this and Algorithm 1, one can assert that (17) does not hold for $k = \bar{K} - 1$. On the other hand, by (19) and [27, Theorem 3], one has

$$\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\| \leq (\bar{\zeta}^{-1} + L_{\nabla\bar{h}})^{-1}\bar{\epsilon}/2, \tag{63}$$

where $(x^*, y^*)$ is the optimal solution of problem (10) and $\bar{\zeta}$ is an input of Algorithm 1. Notice from Algorithm 1 that $(\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}})$ results from the forward-backward splitting (FBS) step applied to the strongly monotone inclusion problem $0 \in (\nabla_x \bar{h}(x,y), -\nabla_y \bar{h}(x,y)) + (\partial p(x), \partial q(y))$ at the point $(x^{\bar{K}}, y^{\bar{K}})$. It then follows from this, $\bar{\zeta} = \min\{\sigma_x, \sigma_y\}/L^2_{\nabla\bar{h}}$ (see Algorithm 1), and the contraction property of FBS [5, Corollary 2.5] that $\|(\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}}) - (x^*, y^*)\| \leq \|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\|$. Using this and (63), we have

$$\begin{aligned}
&\|\bar{\zeta}^{-1}(x^{\bar{K}} - \tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}} - y^{\bar{K}}) - (\nabla\bar{h}(x^{\bar{K}}, y^{\bar{K}}) - \nabla\bar{h}(\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}}))\| \\
\leq\ & \bar{\zeta}^{-1}\|(x^{\bar{K}}, y^{\bar{K}}) - (\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}})\| + \|\nabla\bar{h}(x^{\bar{K}}, y^{\bar{K}}) - \nabla\bar{h}(\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}})\| \\
\leq\ & (\bar{\zeta}^{-1} + L_{\nabla\bar{h}})\|(x^{\bar{K}}, y^{\bar{K}}) - (\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}})\| \\
\leq\ & (\bar{\zeta}^{-1} + L_{\nabla\bar{h}})(\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\| + \|(\tilde{x}^{\bar{K}}, \tilde{y}^{\bar{K}}) - (x^*, y^*)\|) \\
\leq\ & 2(\bar{\zeta}^{-1} + L_{\nabla\bar{h}})\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\| \overset{(63)}{\leq} \bar{\epsilon},
\end{aligned}$$

where the second inequality uses the fact that $\bar{h}$ is $L_{\nabla\bar{h}}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$. It follows that (17) holds for $k = \bar{K} - 1$, which contradicts the above assertion. Hence, Algorithm 1 must terminate in at most $\bar{K}$ outer iterations.

We next show that the output of Algorithm 1 is an $\bar{\epsilon}$-primal-dual stationary point of (10). To this end, suppose that Algorithm 1 terminates at some iteration $k$ at which (17) is satisfied. Then by (6) and the definition of $\tilde{x}^{k+1}$ and $\tilde{y}^{k+1}$ (see steps 23 and 24 of Algorithm 1), one has

$$\begin{aligned}
0 &\in \bar{\zeta}\partial p(\tilde{x}^{k+1}) + \tilde{x}^{k+1} - x^{k+1} + \bar{\zeta}\nabla_x \bar{h}(x^{k+1}, y^{k+1}), \\
0 &\in \bar{\zeta}\partial q(\tilde{y}^{k+1}) + \tilde{y}^{k+1} - y^{k+1} - \bar{\zeta}\nabla_y \bar{h}(x^{k+1}, y^{k+1}),
\end{aligned}$$

which yield

$$\bar{\zeta}^{-1}(x^{k+1} - \tilde{x}^{k+1}) - \nabla_x \bar{h}(x^{k+1}, y^{k+1}) \in \partial p(\tilde{x}^{k+1}),\ \bar{\zeta}^{-1}(y^{k+1} - \tilde{y}^{k+1}) + \nabla_y \bar{h}(x^{k+1}, y^{k+1}) \in \partial q(\tilde{y}^{k+1}).$$

These together with the definition of $\bar{H}$ in (10) imply that

$$\begin{aligned}
&\nabla_x \bar{h}(\tilde{x}^{k+1}, \tilde{y}^{k+1}) + \bar{\zeta}^{-1}(x^{k+1} - \tilde{x}^{k+1}) - \nabla_x \bar{h}(x^{k+1}, y^{k+1}) \in \partial_x \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1}), \\
&\nabla_y \bar{h}(\tilde{x}^{k+1}, \tilde{y}^{k+1}) - \bar{\zeta}^{-1}(y^{k+1} - \tilde{y}^{k+1}) - \nabla_y \bar{h}(x^{k+1}, y^{k+1}) \in \partial_y \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1}).
\end{aligned}$$

Using these and (17), we obtain

$$\begin{aligned}
&\operatorname{dist}(0, \partial_x \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1}))^2 + \operatorname{dist}(0, \partial_y \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1}))^2 \\
&\leq \|\bar{\zeta}^{-1}(x^{k+1} - \tilde{x}^{k+1}) + \nabla_x \bar{h}(\tilde{x}^{k+1}, \tilde{y}^{k+1}) - \nabla_x \bar{h}(x^{k+1}, y^{k+1})\|^2 \\
&\quad + \|\bar{\zeta}^{-1}(\tilde{y}^{k+1} - y^{k+1}) + \nabla_y \bar{h}(\tilde{x}^{k+1}, \tilde{y}^{k+1}) - \nabla_y \bar{h}(x^{k+1}, y^{k+1})\|^2 \\
&= \|\bar{\zeta}^{-1}(x^{k+1} - \tilde{x}^{k+1}, \tilde{y}^{k+1} - y^{k+1}) - (\nabla\bar{h}(x^{k+1}, y^{k+1}) - \nabla\bar{h}(\tilde{x}^{k+1}, \tilde{y}^{k+1}))\|^2 \overset{(17)}{\leq} \bar{\epsilon}^2,
\end{aligned}$$

which implies that $\operatorname{dist}(0, \partial_x \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1})) \leq \bar{\epsilon}$ and $\operatorname{dist}(0, \partial_y \bar{H}(\tilde{x}^{k+1}, \tilde{y}^{k+1})) \leq \bar{\epsilon}$. It then follows from these and Definition 1 that the output $(\tilde{x}^{k+1}, \tilde{y}^{k+1})$ of Algorithm 1 is an $\bar{\epsilon}$-primal-dual stationary point of (10).

Finally, we show that the total number of evaluations of $\nabla\bar{h}$ and proximal operator of $p$ and $q$ performed in Algorithm 1 is no more than $\bar{N}$, respectively. Indeed, notice from Algorithm 1 that $\bar{\alpha} = \min\{1, \sqrt{8\sigma_y/\sigma_x}\}$, which implies that $2/\bar{\alpha} = \max\{2, \sqrt{\sigma_x/(2\sigma_y)}\}$ and $\bar{\alpha} \le \sqrt{8\sigma_y/\sigma_x}$. By these, one has

$$\max\left\{\frac{2}{\bar{\alpha}}, \frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\} \le \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}, \sqrt{\frac{8\sigma_y}{\sigma_x}}\frac{\sigma_x}{4\sigma_y}\right\} = \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}. \tag{64}$$

In addition, by [27, Lemma 4], the number of inner iterations performed in each outer iteration of Algorithm 1 is at most

$$\bar{T} := \left\lceil 48\sqrt{2}\left(1 + 8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil - 1.$$

Then one can observe that the number of evaluations of $\nabla\bar{h}$ and proximal operator of $p$ and $q$ performed in Algorithm 1 is at most

$$\begin{aligned}
(2\bar{T}+3)\bar{K} &\le \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil + 2\right)\left\lceil \max\left\{\frac{2}{\bar{\alpha}}, \frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\}\log\frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\bar{\zeta}^{-1}+L_{\nabla\bar{h}})^{-2}\bar{\epsilon}^2}\right\rceil_+ \\
&\overset{(64)}{\le} \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil + 2\right)\left\lceil \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}\log\frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\bar{\zeta}^{-1}+L_{\nabla\bar{h}})^{-2}\bar{\epsilon}^2}\right\rceil_+ \\
&\le \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil + 2\right) \\
&\quad\times \left\lceil \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}\log\frac{4\max\{1/(2\sigma_x), \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}\}\vartheta_0}{(L_{\nabla\bar{h}}^2/\min\{\sigma_x,\sigma_y\}+L_{\nabla\bar{h}})^{-2}\bar{\epsilon}^2}\right\rceil_+ \overset{(20)(61)}{\le} \bar{N},
\end{aligned}$$

where the second last inequality follows from the definition of $\eta_y$, $\eta_z$ and $\bar{\zeta}$ in Algorithm 1. Hence, the conclusion holds as desired. $\square$

### 4.2 Proof of the main results in Subsection 2.2

In this subsection we prove Theorem 2. Before proceeding, let $\{(x^k, y^k)\}_{k\in\mathbb{T}}$ denote all the iterates generated by Algorithm 2, 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 2 subsequently.

The following lemma shows that an approximate primal-dual stationary point of (22) is found at each iteration of Algorithm 2, and also provides an estimate of operation complexity for finding it.

**Lemma 2.** *Suppose that Assumption 2 holds. Let $\{(x^k, y^k)\}_{k\in\mathbb{T}}$ be generated by Algorithm 2, $H^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $H_{\text{low}}$, $\hat{\alpha}$, $\hat{\delta}$ be defined in* (8), (11), (25), (26) *and* (27), *$L_{\nabla h}$ be given in Assumption 2, $\epsilon$, $\hat{\epsilon}_k$ be given in Algorithm 2, and*

$$\begin{aligned}
\hat{N}_k := &\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_{\mathbf{y}})L_{\nabla h}^{-1}\right)\right\rceil + 2\right)\times\left\lceil \max\left\{2, \sqrt{\frac{D_{\mathbf{y}}L_{\nabla h}}{\epsilon}}\right\}\right. \\
&\left.\times\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_{\mathbf{y}}}{\epsilon}, \frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\text{low}}+\epsilon D_{\mathbf{y}}/4+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{[(3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}}))^2/\min\{L_{\nabla h}, \epsilon/(2D_{\mathbf{y}})\}+3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}})]^{-2}\hat{\epsilon}_k^2}\right\rceil_+.
\end{aligned} \tag{65}$$

*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* (22). *Moreover, the total number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$ performed at iteration $k$ of Algorithm 2 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 (8). Recall that $H$, $H_k$ and $h_k$ are respectively given in (8), (22) and (23), $\mathcal{X} = \operatorname{dom} p$ and $\mathcal{Y} = \operatorname{dom} q$. Notice that $x^*, x^k \in \mathcal{X}$. Then we have

$$\begin{aligned}
H_{k,*} := \min_x\max_y H_k(x,y) &= \min_x\max_y\left\{H(x,y) - \frac{\epsilon}{4D_{\mathbf{y}}}\|y-\hat{y}^0\|^2 + L_{\nabla h}\|x-x^k\|^2\right\} \\
&\le \max_y\{H(x^*,y)+L_{\nabla h}\|x^*-x^k\|^2\} \overset{(8)(11)}{\le} H^* + L_{\nabla h}D_{\mathbf{x}}^2.
\end{aligned} \tag{66}$$

Moreover, by $\mathcal{X} = \operatorname{dom} p$, $\mathcal{Y} = \operatorname{dom} q$, (11) and (25), one has

$$H_{k,\text{low}} := \min_{(x,y)\in \operatorname{dom} p\times \operatorname{dom} q} H_k(x,y) = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ H(x,y) - \frac{\epsilon}{4D_{\mathbf{y}}}\|y-\hat{y}^0\|^2 + L_{\nabla h}\|x-x^k\|^2 \right\}$$
$$\overset{(25)}{\geq} H_{\text{low}} - \max_{y\in\mathcal{Y}} \frac{\epsilon}{4D_{\mathbf{y}}}\|y-\hat{y}^0\|^2 \overset{(11)}{\geq} H_{\text{low}} - \epsilon D_{\mathbf{y}}/4. \tag{67}$$

In addition, by Assumption 2 and the definition of $h_k$ in (23), it is not hard to verify that $h_k(x,y)$ is $L_{\nabla h}$-strongly-convex in $x$, $\epsilon/(2D_{\mathbf{y}})$-strongly-concave in $y$, and $(3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}}))$-smooth on its domain. Also, recall from Remark 2 that $(x^{k+1},y^{k+1})$ results from applying Algorithm 1 to problem (22). The conclusion of this lemma then follows by using (66) and (67) and applying Theorem 1 to (22) with $\bar{\epsilon}=\hat{\epsilon}_k$, $\sigma_x = L_{\nabla h}$, $\sigma_y=\epsilon/(2D_{\mathbf{y}})$, $L_{\nabla\bar{h}} = 3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}})$, $\bar{\alpha}=\hat{\alpha}$, $\bar{\delta}=\hat{\delta}$, $\bar{H}_{\text{low}} = H_{k,\text{low}}$, and $\bar{H}^* = H_{k,*}$. $\square$

The following lemma provides an upper bound on the least progress of the solution sequence of Algorithm 2 and also on the last-iterate objective value of (8).

**Lemma 3.** *Suppose that Assumption 2 holds. Let $\{x^k\}_{k\in\mathbb{T}}$ be generated by Algorithm 2, $H$, $H^*$ and $D_{\mathbf{y}}$ be defined in (8) and (11), $L_{\nabla h}$ be given in Assumption 2, and $\epsilon$, $\hat{\epsilon}_0$ and $\hat{x}^0$ be given in Algorithm 2. Then for all $0\le K\in\mathbb{T}-1$, we have*

$$\min_{0\le k\le K}\|x^{k+1}-x^k\| \le \frac{\max_y H(\hat{x}^0,y) - H^* + \epsilon D_{\mathbf{y}}/4}{L_{\nabla h}(K+1)} + \frac{2\hat{\epsilon}_0^2(1+4D_{\mathbf{y}}^2L_{\nabla h}^2\epsilon^{-2})}{L_{\nabla h}^2(K+1)}, \tag{68}$$
$$\max_y H(x^{K+1},y) \le \max_y H(\hat{x}^0,y) + \epsilon D_{\mathbf{y}}/4 + 2\hat{\epsilon}_0^2\left(L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2}\right). \tag{69}$$

*Proof.* For convenience of the proof, let

$$H_\epsilon^*(x) = \max_y \left\{H(x,y) - \epsilon\|y-\hat{y}^0\|^2/(4D_{\mathbf{y}})\right\}, \tag{70}$$
$$H_k^*(x) = \max_y H_k(x,y), \quad y_*^{k+1} = \arg\max_y H_k(x^{k+1},y). \tag{71}$$

One can observe from these, (22) and (23) that

$$H_k^*(x) = H_\epsilon^*(x) + L_{\nabla h}\|x-x^k\|^2. \tag{72}$$

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 2 that $(x^{k+1},y^{k+1})$ is an $\hat{\epsilon}_k$-primal-dual stationary point of problem (22) 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 (71), 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 $\epsilon/(2D_{\mathbf{y}})$-strong concavity of $H_k(x^{k+1},\cdot)$, implies that $\langle -v, y^{k+1}-y_*^{k+1}\rangle \ge \epsilon\|y^{k+1}-y_*^{k+1}\|^2/(2D_{\mathbf{y}})$. This and $\|v\|\le\hat{\epsilon}_k$ yield

$$\|y^{k+1}-y_*^{k+1}\| \le 2\hat{\epsilon}_k D_{\mathbf{y}}/\epsilon. \tag{73}$$

In addition, by $u\in\partial_x H_k(x^{k+1},y^{k+1})$, (22) and (23), one has

$$u \in \nabla_x h(x^{k+1},y^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1}-x^k). \tag{74}$$

Also, observe from (22), (23) and (71) 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 (74) 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

$$H_k^*(x^k) \geq H_k^*(x^{k+1}) + \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1} \rangle + L_{\nabla h}\|x^k - x^{k+1}\|^2/2. \tag{75}$$

Using this, (72), (73), (75), $\|u\| \leq \hat{\epsilon}_k$, and the Lipschitz continuity of $\nabla h$, we obtain

$$\begin{aligned}
H_\epsilon^*(x^k) - H_\epsilon^*(x^{k+1}) &\overset{(72)}{=} H_k^*(x^k) - H_k^*(x^{k+1}) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(75)}{\geq} \langle u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1}), x^k - x^{k+1} \rangle + 3L_{\nabla h}\|x^k - x^{k+1}\|^2/2 \\
&\geq \big( - \|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|\|x^k - x^{k+1}\| + L_{\nabla h}\|x^k - x^{k+1}\|^2/2\big) + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -(2L_{\nabla h})^{-1}\|u + \nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -L_{\nabla h}^{-1}\|u\|^2 - L_{\nabla h}^{-1}\|\nabla_x h(x^{k+1}, y_*^{k+1}) - \nabla_x h(x^{k+1}, y^{k+1})\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\geq -L_{\nabla h}^{-1}\hat{\epsilon}_k^2 - L_{\nabla h}\|y^{k+1} - y_*^{k+1}\|^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2 \\
&\overset{(73)}{\geq} -(L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2})\hat{\epsilon}_k^2 + L_{\nabla h}\|x^k - x^{k+1}\|^2,
\end{aligned}$$

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 $\nabla h$. Summing up the above inequality for $k = 0, 1, \ldots, K$ yields

$$L_{\nabla h}\sum_{k=0}^{K}\|x^k - x^{k+1}\|^2 \leq H_\epsilon^*(x^0) - H_\epsilon^*(x^{K+1}) + (L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2})\sum_{k=0}^{K}\hat{\epsilon}_k^2. \tag{76}$$

In addition, it follows from (8), (11) and (70) that

$$\begin{aligned}
H_\epsilon^*(x^{K+1}) &= \max_y \left\{H(x^{K+1}, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_{\mathbf{y}})\right\} \geq \min_x \max_y H(x, y) - \epsilon D_{\mathbf{y}}/4 = H^* - \epsilon D_{\mathbf{y}}/4, \\
H_\epsilon^*(x^0) &= \max_y \left\{H(x^0, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_{\mathbf{y}})\right\} \leq \max_y H(x^0, y).
\end{aligned} \tag{77}$$

These together with (76) yield

$$\begin{aligned}
L_{\nabla h}(K+1)\min_{0\leq k\leq K}\|x^{k+1} - x^k\|^2 &\leq L_{\nabla h}\sum_{k=0}^{K}\|x^k - x^{k+1}\|^2 \\
&\leq \max_y H(x^0, y) - H^* + \epsilon D_{\mathbf{y}}/4 + (L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2})\sum_{k=0}^{K}\hat{\epsilon}_k^2,
\end{aligned}$$

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 (68) holds.

Finally, we show that (69) holds. Indeed, it follows from (11), (70), (76), (77), $\hat{\epsilon}_k = \hat{\epsilon}_0(k+1)^{-1}$, and $\sum_{k=0}^{K}(k+1)^{-2} < 2$ that

$$\begin{aligned}
\max_y H(x^{K+1}, y) &\overset{(11)}{\leq} \max_y \left\{H(x^{K+1}, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_{\mathbf{y}})\right\} + \epsilon D_{\mathbf{y}}/4 \overset{(70)}{=} H_\epsilon^*(x^{K+1}) + \epsilon D_{\mathbf{y}}/4 \\
&\overset{(76)}{\leq} H_\epsilon^*(x^0) + \epsilon D_{\mathbf{y}}/4 + (L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2})\sum_{k=0}^{K}\hat{\epsilon}_k^2 \\
&\overset{(77)}{\leq} \max_y H(x^0, y) + \epsilon D_{\mathbf{y}}/4 + 2\hat{\epsilon}_0^2(L_{\nabla h}^{-1} + 4D_{\mathbf{y}}^2 L_{\nabla h}\epsilon^{-2}).
\end{aligned}$$

It then follows from this and $x^0 = \hat{x}^0$ that (69) holds. □

We are now ready to prove Theorem 2 using Lemmas 2 and 3.

*Proof of Theorem 2.* Suppose for contradiction that Algorithm 2 runs for more than $\widehat{T}+1$ outer iterations, where $\widehat{T}$ is given in (28). By this and Algorithm 2, one can then assert that (24) does not hold for all $0 \le k \le T$. On the other hand, by (28) and (68), one has

$$\min_{0\le k\le \widehat{T}} \|x^{k+1}-x^k\|^2 \overset{(68)}{\le} \frac{\max_y H(\hat{x}^0,y)-H^*+\epsilon D_{\mathbf{y}}/4}{L_{\nabla h}(\widehat{T}+1)} + \frac{2\hat{\epsilon}_0^2(1+4D_{\mathbf{y}}^2 L_{\nabla h}^2\epsilon^{-2})}{L_{\nabla h}^2(\widehat{T}+1)} \overset{(28)}{\le} \frac{\epsilon^2}{16L_{\nabla h}^2},$$

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 thus (24) holds for such $k$, which contradicts the above assertion. Hence, Algorithm 2 must terminate in at most $\widehat{T}+1$ outer iterations.

Suppose that Algorithm 2 terminates at some iteration $0\le k\le \widehat{T}$, namely, (24) holds for such $k$. We next show that its output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ is an $\epsilon$-primal-dual stationary point of (8) and moreover it satisfies (107). Indeed, recall from Lemma 2 that $(x^{k+1}, y^{k+1})$ is an $\hat{\epsilon}_k$-primal-dual stationary point of (22), 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, (8), (22) and (23), there exists $(u,v)$ such that

$$\begin{aligned}
&u\in\partial_x H(x^{k+1},y^{k+1})+2L_{\nabla h}(x^{k+1}-x^k), \quad \|u\|\le\hat{\epsilon}_k,\\
&v\in\partial_y H(x^{k+1},y^{k+1})-\epsilon(y^{k+1}-\hat{y}^0)/(2D_{\mathbf{y}}), \quad \|v\|\le\hat{\epsilon}_k.
\end{aligned}$$

It then follows that $u-2L_{\nabla h}(x^{k+1}-x^k)\in\partial_x H(x^{k+1},y^{k+1})$ and $v+\epsilon(y^{k+1}-\hat{y}^0)/(2D_{\mathbf{y}})\in\partial_y H(x^{k+1},y^{k+1})$. These together with (11), (24) and $\hat{\epsilon}_k\le\hat{\epsilon}_0\le\epsilon/2$ (see Algorithm 2) imply that

$$\mathrm{dist}\left(0,\partial_x H(x^{k+1},y^{k+1})\right)\le\|u-2L_{\nabla h}(x^{k+1}-x^k)\|\le\|u\|+2L_{\nabla h}\|x^{k+1}-x^k\|\overset{(24)}{\le}\hat{\epsilon}_k+\epsilon/2\le\epsilon,$$

$$\mathrm{dist}\left(0,\partial_y H(x^{k+1},y^{k+1})\right)\le\|v+\epsilon(y^{k+1}-\hat{y}^0)/(2D_{\mathbf{y}})\|\le\|v\|+\epsilon\|y^{k+1}-\hat{y}^0\|/(2D_{\mathbf{y}})\overset{(11)}{\le}\hat{\epsilon}_k+\epsilon/2\le\epsilon.$$

Hence, the output $(x^{k+1},y^{k+1})$ of Algorithm 2 is an $\epsilon$-primal-dual stationary point of (8). In addition, (30) holds due to Lemma 3.

Recall from Lemma 2 that the number of evaluations of $\nabla h$ and proximal operator of $p$ and $q$ performed at iteration $k$ of Algorithm 2 is at most $\hat{N}_k$, respectively, where $\hat{N}_k$ is defined in (65). 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 $\nabla h$ and proximal operator of $p$ and $q$ in Algorithm 2 is respectively no more than $\sum_{k=0}^{|\mathbb{T}|-2}\hat{N}_k$. Consequently, to complete the rest of the proof of Theorem 2, it suffices to show that $\sum_{k=0}^{|\mathbb{T}|-2}\hat{N}_k\le\widehat{N}$, where $\widehat{N}$ is given in (29). Indeed, by (29), (65) and $|\mathbb{T}|\le\widehat{T}+2$, one has

$$\begin{aligned}
\sum_{k=0}^{|\mathbb{T}|-2}\hat{N}_k &\overset{(65)}{\le} \sum_{k=0}^{\widehat{T}}\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_{\mathbf{y}})\,L_{\nabla h}^{-1}\right)\right\rceil+2\right)\times\Bigg\lceil\max\left\{2,\sqrt{\frac{D_{\mathbf{y}}L_{\nabla h}}{\epsilon}}\right\}\\
&\quad\times\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}},\min\left\{\frac{D_{\mathbf{y}}}{\epsilon},\frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-H_{\mathrm{low}}+\epsilon D_{\mathbf{y}}/4+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[(3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}}))^2/\min\{L_{\nabla h},\epsilon/(2D_{\mathbf{y}})\}+3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}})\right]^{-2}\hat{\epsilon}_k^2}\Bigg\rceil_+\\
&\le\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_{\mathbf{y}})\,L_{\nabla h}^{-1}\right)\right\rceil+2\right)\max\left\{2,\sqrt{\frac{D_{\mathbf{y}}L_{\nabla h}}{\epsilon}}\right\}\\
&\quad\times\sum_{k=0}^{\widehat{T}}\left(\left(\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}},\min\left\{\frac{D_{\mathbf{y}}}{\epsilon},\frac{4}{\hat{\alpha}L_{\nabla h}}\right\}\right\}\left(\hat{\delta}+2\hat{\alpha}^{-1}(H^*-h_{\mathrm{low}}+\epsilon D_{\mathbf{y}}/4+L_{\nabla h}D_{\mathbf{x}}^2)\right)}{\left[(3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}}))^2/\min\{L_{\nabla h},\epsilon/(2D_{\mathbf{y}})\}+3L_{\nabla h}+\epsilon/(2D_{\mathbf{y}})\right]^{-2}\hat{\epsilon}_k^2}\right)_+ +1\right)\\
&\le\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_{\mathbf{y}})\,L_{\nabla h}^{-1}\right)\right\rceil+2\right)\max\left\{2,\sqrt{\frac{D_{\mathbf{y}}L_{\nabla h}}{\epsilon}}\right\}
\end{aligned}$$

$$\times \left( (\widehat{T}+1) \left( \log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_{\mathbf{y}}}{\epsilon}, \frac{4}{\hat{\alpha} L_{\nabla h}}\right\}\right\} \left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\mathrm{low}} + \epsilon D_{\mathbf{y}}/4 + L_{\nabla h} D_{\mathbf{x}}^2)\right)}{[(3L_{\nabla h} + \epsilon/(2D_{\mathbf{y}}))^2/\min\{L_{\nabla h}, \epsilon/(2D_{\mathbf{y}})\} + 3L_{\nabla h} + \epsilon/(2D_{\mathbf{y}})]^{-2}\,\hat{\epsilon}_0^2} \right)_+ \right.$$

$$\left. + \widehat{T} + 1 + 2\sum_{k=0}^{\widehat{T}} \log(k+1) \right) \overset{(29)}{\leq} \widehat{N},$$

where the last inequality is due to (29) and $\sum_{k=0}^{\widehat{T}} \log(k+1) \leq \widehat{T}\log(\widehat{T}+1)$. This completes the proof of Theorem 2. □

## 4.3 Proof of the main results in Subsection 3.1

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

$$\mathcal{L}_{\mathbf{y}}(x, y, \lambda_{\mathbf{y}}; \rho) = F(x, y) - \frac{1}{2\rho}\left(\|[\lambda_{\mathbf{y}} + \rho d(x, y)]_+\|^2 - \|\lambda_{\mathbf{y}}\|^2\right). \tag{78}$$

In view of (5), (43) and (78), one can observe that

$$f^*(x) \leq \max_y \mathcal{L}_{\mathbf{y}}(x, y, \lambda_{\mathbf{y}}; \rho) \qquad \forall x \in \mathcal{X},\ \lambda_{\mathbf{y}} \in \mathbb{R}_+^{\tilde{m}},\ \rho > 0, \tag{79}$$

which will be frequently used later.

We next establish several lemmas that will be used to prove Theorem 3 subsequently. The following lemma establishes an upper bound on the optimal Lagrangian multipliers of problem (43) and also provides a reformulation of $f^*(x)$.

**Lemma 4.** *Suppose that Assumptions 1 and 5 hold. Let $f^*$, $\Delta$, $r$ and $\delta_d$ be given in* (43), (45) *and Assumption 5, respectively. Then the following statements hold.*

*(i)* $\|\lambda_{\mathbf{y}}^*\| \leq \delta_d^{-1}\Delta$ *and* $\lambda_{\mathbf{y}}^* \in \mathbb{B}_r^+$ *for all* $\lambda_{\mathbf{y}}^* \in \Lambda^*(x)$ *and* $x \in \mathcal{X}$, *where* $\Lambda^*(x)$ *denotes the set of optimal Lagrangian multipliers of problem* (43) *for any* $x \in \mathcal{X}$.

*(ii) It holds that*

$$f^*(x) = \min_{\lambda_{\mathbf{y}}} \max_y F(x, y) - \langle \lambda_{\mathbf{y}}, d(x, y)\rangle + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}}) \qquad \forall x \in \mathcal{X}, \tag{80}$$

*where* $\delta_{\mathbb{R}_+^{\tilde{m}}}(\cdot)$ *is the indicator function associated with* $\mathbb{R}_+^{\tilde{m}}$.

*Proof.* (i) Let $x \in \mathcal{X}$, $\lambda_{\mathbf{y}}^* \in \Lambda^*(x)$ be arbitrarily chosen, and $\hat{y}_x \in \mathcal{Y}$ and $\delta_d > 0$ be given in Assumption 5(ii). It then follows from Assumption 5(ii) that $d_i(x, \hat{y}_x) \leq -\delta_d$ for all $i$. In addition, let $y^* \in \mathcal{Y}$ be such that $(y^*, \lambda_{\mathbf{y}}^*)$ is a pair of primal-dual optimal solutions of (43). Then we have

$$y^* \in \operatorname*{Argmax}_y F(x, y) - \langle \lambda_{\mathbf{y}}^*, d(x, y)\rangle, \quad \langle \lambda_{\mathbf{y}}^*, d(x, y^*)\rangle = 0, \quad d(x, y^*) \leq 0, \quad \lambda_{\mathbf{y}}^* \geq 0.$$

The first relation above yields

$$F(x, y^*) - \langle \lambda_{\mathbf{y}}^*, d(x, y^*)\rangle \geq F(x, \hat{y}_x) - \langle \lambda_{\mathbf{y}}^*, d(x, \hat{y}_x)\rangle.$$

By this and $\langle \lambda_{\mathbf{y}}^*, d(x, y^*)\rangle = 0$, one has

$$\langle \lambda_{\mathbf{y}}^*, -d(x, \hat{y}_x)\rangle \leq F(x, y^*) - F(x, \hat{y}_x),$$

which together with $\lambda_{\mathbf{y}}^* \geq 0$, $d_i(x, \hat{y}_x) \leq -\delta_d$ for all $i$, (44) and (45) implies that

$$\delta_d \|\lambda_{\mathbf{y}}^*\|_1 \leq \langle \lambda_{\mathbf{y}}^*, -d(x, \hat{y}_x)\rangle \leq F(x, y^*) - F(x, \hat{y}_x) \leq \Delta,$$

Hence, we have $\|\lambda_{\mathbf{y}}^*\| \le \|\lambda_{\mathbf{y}}^*\|_1 \le \delta_d^{-1}\Delta$. This and (45) imply that $\lambda_{\mathbf{y}}^* \in \mathbb{B}_r^+$.

(ii) Recall from Assumption 1 that $F(x,\cdot)$ and $d_i(x,\cdot)$, $i=1,\ldots,l$, are convex for any given $x \in \mathcal{X}$. Using this, (43), (45) and the first statement of this lemma, we observe that

$$f^*(x) = \max_y \min_{\lambda\in\mathbb{B}_r^+} F(x,y) - \langle \lambda, d(x,y)\rangle \qquad \forall x \in \mathcal{X}.$$

Also, notice from Assumption 1 that the domain of $F(x,\cdot)$ is compact for all $x \in \mathcal{X}$. By this, the above equality, and the strong duality, one has

$$f^*(x) = \min_{\lambda\in\mathbb{B}_r^+} \max_y F(x,y) - \langle \lambda, d(x,y)\rangle \qquad \forall x \in \mathcal{X}. \tag{81}$$

In addition, one can observe from (43) that for all $x \in \mathcal{X}$,

$$f^*(x) = \max_y \min_{\lambda_{\mathbf{y}}} F(x,y) - \langle \lambda_{\mathbf{y}}, d(x,y)\rangle + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}}) \le \min_{\lambda_{\mathbf{y}}} \max_y F(x,y) - \langle \lambda_{\mathbf{y}}, d(x,y)\rangle + \delta_{\mathbb{R}_+^{\tilde{m}}}(\lambda_{\mathbf{y}}),$$

where the inequality follows from the weak duality. This together with (81) implies that (80) holds. □

The next lemma provides an upper bound for $\{\lambda_{\mathbf{y}}^k\}_{k\in\mathbb{K}}$.

**Lemma 5.** *Suppose that Assumptions 1 and 5 hold. Let $\{\lambda_{\mathbf{y}}^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 3, $D_{\mathbf{y}}$ and $\Delta$ be defined in* (11) *and* (45)*, and $\tau$ and $\rho_k$ be given in Algorithm 3. Then we have*

$$\rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 \le \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} \qquad \forall 0 \le k \in \mathbb{K}-1. \tag{82}$$

*Proof.* One can observe from (45) and Algorithm 3 that$\Delta \ge 0$ and $\rho_0 \ge 1 > \tau > 0$, which imply that (82) holds for $k=0$. It remains to show that (82) holds for all $1 \le k \in \mathbb{K}-1$.

Since $(x^{t+1}, y^{t+1})$ is an $\epsilon_t$-primal-dual stationary point of (35) for all $0 \le t \in \mathbb{K}-1$, it follows from Definition 1 that there exists some $u \in \partial_y \mathcal{L}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{x}}^t, \lambda_{\mathbf{y}}^t; \rho_t)$ with $\|u\| \le \epsilon_t$. Notice from (5) and (78) that $\partial_y \mathcal{L}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{x}}^t, \lambda_{\mathbf{y}}^t; \rho_t) = \partial_y \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t)$. Hence, $u \in \partial_y \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t)$. Also, observe from (1), (78) and Assumption 1 that $\mathcal{L}_{\mathbf{y}}(x^{t+1}, \cdot, \lambda_{\mathbf{y}}^t; \rho_t)$ is concave. Using this, (11), $u \in \partial_y \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t)$ and $\|u\| \le \epsilon_t$, we obtain

$$\begin{aligned}\mathcal{L}_{\mathbf{y}}(x^{t+1}, y, \lambda_{\mathbf{y}}^t; \rho_t) &\le \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t) + \langle u, y - y^{t+1}\rangle \\ &\le \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t) + D_{\mathbf{y}}\epsilon_t \qquad \forall y \in \mathcal{Y},\end{aligned}$$

which implies that

$$\max_y \mathcal{L}_{\mathbf{y}}(x^{t+1}, y, \lambda_{\mathbf{y}}^t; \rho_t) \le \mathcal{L}_{\mathbf{y}}(x^{t+1}, y^{t+1}, \lambda_{\mathbf{y}}^t; \rho_t) + D_{\mathbf{y}}\epsilon_t. \tag{83}$$

By this, (78) and (79), one has

$$\begin{aligned} f^*(x^{t+1}) &\overset{(79)}{\le} \max_y \mathcal{L}_{\mathbf{y}}(x^{t+1}, y, \lambda_{\mathbf{y}}^t; \rho_t) \\ &\overset{(78)(83)}{\le} F(x^{t+1}, y^{t+1}) - \frac{1}{2\rho_t}\left(\|[\lambda_{\mathbf{y}}^t + \rho_t d(x^{t+1}, y^{t+1})]_+\|^2 - \|\lambda_{\mathbf{y}}^t\|^2\right) + D_{\mathbf{y}}\epsilon_t \\ &= F(x^{t+1}, y^{t+1}) - \frac{1}{2\rho_t}\left(\|\lambda_{\mathbf{y}}^{t+1}\|^2 - \|\lambda_{\mathbf{y}}^t\|^2\right) + D_{\mathbf{y}}\epsilon_t, \end{aligned}$$

where the equality follows from the relation $\lambda_{\mathbf{y}}^{t+1} = [\lambda_{\mathbf{y}}^t + \rho_t d(x^{t+1}, y^{t+1})]_+$ (see Algorithm 3). Using the above inequality,(47) and $\epsilon_t \le 1$ (see Algorithm 3), we have

$$\|\lambda_{\mathbf{y}}^{t+1}\|^2 - \|\lambda_{\mathbf{y}}^t\|^2 \le 2\rho_k\big(F(x^{t+1}, y^{t+1}) - f^*(x^{t+1}) + D_{\mathbf{y}}\epsilon_t\big) \le 2\rho_t(\Delta + D_{\mathbf{y}}).$$

Summing up this inequality for $t = 0, \dots, k-1$ with $1 \le k \in \mathbb{K} - 1$ yields

$$\|\lambda_{\mathbf{y}}^k\|^2 \le \|\lambda_{\mathbf{y}}^0\|^2 + 2(\Delta + D_{\mathbf{y}}) \sum_{t=0}^{k-1} \rho_t. \tag{84}$$

Recall from Algorithm 3 that $\rho_t = \epsilon_t^{-1} = \tau^{-t}$. Then we have $\sum_{t=0}^{k-1} \rho_t \le \rho_{k-1}/(1-\tau)$. Using this, (84) and $\rho_k > \rho_{k-1} \ge 1$ (see Algorithm 3), we obtain that for all $1 \le k \in \mathbb{K} - 1$,

$$\rho_k^{-1} \|\lambda_{\mathbf{y}}^k\|^2 \le \rho_k^{-1} \left( \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})\rho_{k-1}}{1-\tau} \right) \le \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau}.$$

Hence, the conclusion holds as desired. $\square$

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

**Lemma 6.** *Suppose that Assumptions 1 and 5 hold. Let $D_{\mathbf{y}}$ and $\Delta$ be defined in* (11) *and* (45)*, and $\delta_d$ be given in Assumption 5, and $\tau$ and $\rho_k$ be given in Algorithm 3. Suppose that $(x^{k+1}, y^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 3 for some $0 \le k \in \mathbb{K} - 1$ with*

$$\rho_k \ge \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2(1-\tau)}. \tag{85}$$

*Then we have*

$$\|[d(x^{k+1}, y^{k+1})]_+\| \le \rho_k^{-1} \|\lambda_{\mathbf{y}}^{k+1}\| \le 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}}). \tag{86}$$

*Proof.* Suppose that $(x^{k+1}, y^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 3 for some $0 \le k \in \mathbb{K} - 1$ with $\rho_k$ satisfying (85). Since $(x^{k+1}, y^{k+1})$ is an $\epsilon_k$-primal-dual stationary point of (35), it follows from (5) and Definition 1 that

$$\text{dist}\left(0, \partial_y F(x^{k+1}, y^{k+1}) - \nabla_y d(x^{k+1}, y^{k+1})[\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+\right) \le \epsilon_k.$$

Besides, notice from Algorithm 3 that $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+$. Hence, there exists some $u \in \partial_y F(x^{k+1}, y^{k+1})$ such that

$$\|u - \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\| \le \epsilon_k. \tag{87}$$

By Assumption 5(ii), there exists some $\hat{y}^{k+1} \in \mathcal{Y}$ such that $-d_i(x^{k+1}, \hat{y}^{k+1}) \ge \delta_d$ for all $i$. Notice that $\langle \lambda_{\mathbf{y}}^{k+1}, \lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1}) \rangle = \|[\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+\|^2 \ge 0$, which implies that

$$-\langle \lambda_{\mathbf{y}}^{k+1}, \rho_k^{-1}\lambda_{\mathbf{y}}^k \rangle \le \langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1}) \rangle. \tag{88}$$

Using these and (87), we have

$$\begin{aligned}
&F(x^{k+1}, \hat{y}^{k+1}) - F(x^{k+1}, y^{k+1}) + \delta_d \|\lambda_{\mathbf{y}}^{k+1}\|_1 - \rho_k^{-1}\langle \lambda_{\mathbf{y}}^{k+1}, \lambda_{\mathbf{y}}^k \rangle \\
&\le F(x^{k+1}, \hat{y}^{k+1}) - F(x^{k+1}, y^{k+1}) - \langle \lambda_{\mathbf{y}}^{k+1}, \rho_k^{-1}\lambda_{\mathbf{y}}^k + d(x^{k+1}, \hat{y}^{k+1}) \rangle \\
&\overset{(88)}{\le} F(x^{k+1}, \hat{y}^{k+1}) - F(x^{k+1}, y^{k+1}) + \langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1}) - d(x^{k+1}, \hat{y}^{k+1})) \rangle \\
&\le \langle u, \hat{y}^{k+1} - y^{k+1} \rangle + \langle \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, y^{k+1} - \hat{y}^{k+1} \rangle \\
&= \langle u - \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, y^{k+1} - \hat{y}^{k+1} \rangle \le D_{\mathbf{y}}\epsilon_k,
\end{aligned} \tag{89}$$

where the first inequality is due to $\lambda_{\mathbf{y}}^{k+1} \ge 0$ and $-d_i(x^{k+1}, \hat{y}^{k+1}) \ge \delta_d$ for all $i$, the third inequality follows from $u \in \partial_y F(x^{k+1}, y^{k+1})$, $\lambda_{\mathbf{y}}^{k+1} \ge 0$, the concavity of $F(x^{k+1}, \cdot)$ and the convexity of $d_i(x^{k+1}, \cdot)$, and the last inequality is due to (11) and (87).

In view of (44) and (89), one has

$$
\begin{aligned}
D_{\mathbf{y}}\epsilon_k + \Delta &\overset{(44)}{\geq} D_{\mathbf{y}}\epsilon_k - F(x^{k+1}, \hat{y}^{k+1}) + F(x^{k+1}, y^{k+1}) \\
&\overset{(89)}{\geq} \delta_d \|\lambda_{\mathbf{y}}^{k+1}\|_1 - \rho_k^{-1}\langle \lambda_{\mathbf{y}}^{k+1}, \lambda_{\mathbf{y}}^k \rangle \geq (\delta_d - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|)\|\lambda_{\mathbf{y}}^{k+1}\|, \qquad (90)
\end{aligned}
$$

where the last inequality is due to $\|\lambda_{\mathbf{y}}^{k+1}\|_1 \geq \|\lambda_{\mathbf{y}}^{k+1}\|$. In addition, it follows from (82) and (85) that

$$
\delta_d - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\| \overset{(82)}{\geq} \delta_d - \sqrt{\rho_k^{-1}\left(\|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(F_{\mathrm{hi}} - f_{\mathrm{low}}^* + D_{\mathbf{y}})}{1-\tau}\right)} \overset{(85)}{\geq} \frac{1}{2}\delta_d,
$$

which together with (90) yields

$$
\frac{1}{2}\delta_d\|\lambda_{\mathbf{y}}^{k+1}\| \leq (\delta_d - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|)\|\lambda_{\mathbf{y}}^{k+1}\| \overset{(90)}{\leq} D_{\mathbf{y}}\epsilon_k + \Delta.
$$

The conclusion then follows from this, $\epsilon_k \leq 1$, and the relations

$$
\|[d(x^{k+1}, y^{k+1})]_+\| \leq \rho_k^{-1}\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+\| = \rho_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\|.
$$

□

The next lemma provides an upper bound on the amount of violation of the conditions in (39), (40) and (42) at $(x, y, \lambda_{\mathbf{x}}, \lambda_{\mathbf{y}}) = (x^{k+1}, y^{k+1}, \tilde{\lambda}_{\mathbf{x}}^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ for $0 \leq k \in \mathbb{K} - 1$, where $\tilde{\lambda}_{\mathbf{x}}^{k+1}$ is given below.

**Lemma 7.** *Suppose that Assumptions 1 and 5 hold. Let $D_{\mathbf{y}}$ and $\Delta$ be defined in* (11) *and* (45)*, and $\delta_d$ be given in Assumption 5, and $\tau$, $\epsilon_k$, $\rho_k$ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 3. Suppose that $(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 3 for some $0 \leq k \in \mathbb{K} - 1$ with*

$$
\rho_k \geq \frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2 \tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2 \tau(1-\tau)}. \qquad (91)
$$

*Let*

$$
\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+. \qquad (92)
$$

*Then we have*

$$
\begin{aligned}
&\mathrm{dist}(0, \partial_x F(x^{k+1}, y^{k+1}) + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}) \leq \epsilon_k, && (93)\\
&\mathrm{dist}\left(0, \partial_y F(x^{k+1}, y^{k+1}) - \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\right) \leq \epsilon_k, && (94)\\
&\|[d(x^{k+1}, y^{k+1})]_+\| \leq 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}}), && (95)\\
&|\langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1})\rangle| \leq 2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}})\max\{\|\lambda_{\mathbf{y}}^0\|,\ 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})\}. && (96)
\end{aligned}
$$

*Proof.* Suppose that $(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^{k+1}, \lambda_{\mathbf{y}}^{k+1})$ is generated by Algorithm 3 for some $0 \leq k \in \mathbb{K} - 1$ with $\rho_k$ satisfying (91). Since $(x^{k+1}, y^{k+1})$ is an $\epsilon_k$-primal-dual stationary point of (35), it then follows from Definition 1 that

$$
\mathrm{dist}\big(0, \partial_x \mathcal{L}(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k)\big) \leq \epsilon_k,\ \mathrm{dist}\big(0, \partial_y \mathcal{L}(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k)\big) \leq \epsilon_k. \qquad (97)
$$

Observe from Algorithm 3 that $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+$. In view of this, (5) and (92), one has

$$
\begin{aligned}
\partial_x \mathcal{L}(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) &= \partial_x F(x^{k+1}, y^{k+1}) + \nabla c(x^{k+1})[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+ \\
&\quad - \nabla_x d(x^{k+1}, y^{k+1})[\lambda_{\mathbf{y}}^k + \rho_k d(x^{k+1}, y^{k+1})]_+ \\
&= \partial_x F(x^{k+1}, y^{k+1}) + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, \\
\partial_y \mathcal{L}(x^{k+1}, y^{k+1}, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) &= \partial_y F(x^{k+1}, y^{k+1}) - \nabla_y d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}.
\end{aligned}
$$

These relations together with (97) imply that (93) and (94) hold.

Notice from Algorithm 3 that $0 < \tau < 1$, which together with (91) implies that (85) holds for $\rho_k$. It then follows that (86) holds, which immediately yields (95) and

$$\|\lambda_{\mathbf{y}}^{k+1}\| \le 2\delta_d^{-1}(\Delta + D_{\mathbf{y}}). \tag{98}$$

Claim that

$$\|\lambda_{\mathbf{y}}^{k}\| \le \max\{\|\lambda_{\mathbf{y}}^{0}\|,\ 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})\}. \tag{99}$$

Indeed, (99) clearly holds if $k = 0$. We now assume that $k > 0$. Notice from Algorithm 3 that $\rho_{k-1} = \tau\rho_k$, which together with (91) implies that (85) holds with $k$ replaced by $k-1$. By this and Lemma 6 with $k$ replaced by $k-1$, one can conclude that $\|\lambda_{\mathbf{y}}^{k}\| \le 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})$ and hence (99) holds.

We next show that (96) holds. Indeed, by $\lambda_{\mathbf{y}}^{k+1} \ge 0$, (88), (95), (98) and (99), one has

$$\begin{aligned}
\langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1})\rangle &\le \langle \lambda_{\mathbf{y}}^{k+1}, [d(x^{k+1}, y^{k+1})]_+\rangle \le \|\lambda_{\mathbf{y}}^{k+1}\|\|[d(x^{k+1}, y^{k+1})]_+\| \\
&\overset{(95)(98)}{\le} 4\rho_k^{-1}\delta_d^{-2}(\Delta + D_{\mathbf{y}})^2, \\
\langle \lambda_{\mathbf{y}}^{k+1}, d(x^{k+1}, y^{k+1})\rangle &\overset{(88)}{\ge} \langle \lambda_{\mathbf{y}}^{k+1}, -\rho_k^{-1}\lambda_{\mathbf{y}}^{k}\rangle \ge -\rho_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\|\|\lambda_{\mathbf{y}}^{k}\| \\
&\ge -2\rho_k^{-1}\delta_d^{-1}(\Delta + D_{\mathbf{y}})\max\{\|\lambda_{\mathbf{y}}^{0}\|,\ 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})\}.
\end{aligned}$$

These relations imply that (96) holds. □

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 8.** *Suppose that Assumptions 1, 4 and 5 hold. Let $\{(\lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k)\}_{k\in\mathbb{K}}$ be generated by Algorithm 3, $\mathcal{L}$, $D_{\mathbf{y}}$, $F_{\mathrm{hi}}$ and $\Delta$ be defined in (5), (11), (44) and (45), and $\tau$, $\rho_k$, $\Lambda$ and $x_{\mathrm{init}}^k$ be given in Algorithm 3. Then for all $0 \le k \in \mathbb{K}-1$, we have*

$$\max_y \mathcal{L}(x_{\mathrm{init}}^k, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \le \Delta + F_{\mathrm{hi}} + \Lambda + \frac{1}{2}(1 + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau}. \tag{100}$$

*Proof.* In view of (32), (34), (44) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ (see Algorithm 3), one has

$$\begin{aligned}
\mathcal{L}_{\mathbf{x}}(x_{\mathrm{init}}^k, y^k, \lambda_{\mathbf{x}}^k; \rho_k) &\overset{(34)}{\le} \mathcal{L}_{\mathbf{x}}(x_{\mathbf{nf}}, y^k, \lambda_{\mathbf{x}}^k; \rho_k) \overset{(32)}{=} F(x_{\mathbf{nf}}, y^k) + \frac{1}{2\rho_k}\big(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x_{\mathbf{nf}})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\big) \\
&\le F(x_{\mathbf{nf}}, y^k) + \frac{1}{2\rho_k}\big((\|\lambda_{\mathbf{x}}^k\| + \rho_k\|[c(x_{\mathbf{nf}})]_+\|)^2 - \|\lambda_{\mathbf{x}}^k\|^2\big) \\
&= F(x_{\mathbf{nf}}, y^k) + \|\lambda_{\mathbf{x}}^k\|\|[c(x_{\mathbf{nf}})]_+\| + \frac{1}{2}\rho_k\|[c(x_{\mathbf{nf}})]_+\|^2 \\
&\overset{(44)}{\le} F_{\mathrm{hi}} + \Lambda\|[c(x_{\mathbf{nf}})]_+\| + \frac{1}{2}\rho_k\|[c(x_{\mathbf{nf}})]_+\|^2.
\end{aligned} \tag{101}$$

In addition, one can observe from Algorithm 3 that $\epsilon_k > \tau\varepsilon$ for all $0 \le k \in \mathbb{K}-1$. By this and the choice of $\rho_k$ in Algorithm 3, we obtain that $\rho_k = \epsilon_k^{-1} < \tau^{-1}\varepsilon^{-1}$ for all $0 \le k \in \mathbb{K}-1$. It then follows from this, (5), (32), (45), (82), (101), $\|[c(x_{\mathbf{nf}})]_+\| \le \sqrt{\varepsilon} \le 1$, and the Lipschitz continuity of $F$ that

$$
\begin{aligned}
&\max_y \mathcal{L}(x_{\text{init}}^k, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \overset{(5)(32)}{=} \max_y \left\{ \mathcal{L}_{\mathbf{x}}(x_{\text{init}}^k, y, \lambda_{\mathbf{x}}^k; \rho_k) - \frac{1}{2\rho_k}\left( \|[\lambda_{\mathbf{y}}^k + \rho_k d(x_{\text{init}}^k, y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2 \right) \right\} \\
&\le \max_y \left\{ \mathcal{L}_{\mathbf{x}}(x_{\text{init}}^k, y, \lambda_{\mathbf{x}}^k; \rho_k) + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \right\} \\
&\overset{(32)}{=} \max_y \left\{ F(x_{\text{init}}^k, y) - F(x_{\text{init}}^k, y^k) + \mathcal{L}_{\mathbf{x}}(x_{\text{init}}^k, y^k, \lambda_{\mathbf{x}}^k; \rho_k) + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \right\} \\
&\overset{(45)}{\le} \Delta + \mathcal{L}_{\mathbf{x}}(x_{\text{init}}^k, y^k, \lambda_{\mathbf{x}}^k; \rho_k) + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \\
&\le \Delta + F_{\text{hi}} + \Lambda\|[c(x_{\mathbf{nf}})]_+\| + \frac{1}{2}\rho_k\|[c(x_{\mathbf{nf}})]_+\|^2 + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} \\
&\le \Delta + F_{\text{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau},
\end{aligned}
$$

where the third inequality follows from (82) and (101), and the last inequality follows from $\rho_k < \tau^{-1}\varepsilon^{-1}$ and $\|[c(x_{\mathbf{nf}})]_+\| \le \sqrt{\varepsilon} \le 1$. $\square$

The next lemma shows that an approximate primal-dual stationary point of (35) is found at each iteration of Algorithm 3, and also provides an estimate of operation complexity for finding it.

**Lemma 9.** *Suppose that Assumptions 1, 4 and 5 hold. Let $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $L_k$, $F_{\text{hi}}$ and $\Delta$ be defined in* (11), (36), (44) *and* (45), *$\tau$, $\epsilon_k$, $\rho_k$, $\Lambda$ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 3, and*

$$
\begin{aligned}
\alpha_k &= \min\left\{1, \sqrt{4\epsilon_k/(D_{\mathbf{y}} L_k)}\right\}, && (102)\\
\delta_k &= (2+\alpha_k^{-1})L_k D_{\mathbf{x}}^2 + \max\{\epsilon_k/D_{\mathbf{y}}, \alpha_k L_k/4\} D_{\mathbf{y}}^2, && (103)\\
M_k &= \frac{16\max\{1/(2L_k), \min\{D_{\mathbf{y}}/\epsilon_k, 4/(\alpha_k L_k)\}\}\rho_k}{[(3L_k + \epsilon_k/(2D_{\mathbf{y}}))^2/\min\{L_k, \epsilon_k/(2D_{\mathbf{y}})\} + 3L_k + \epsilon_k/(2D_{\mathbf{y}})]^{-2}\epsilon_k^2} \\
&\quad \times \left( \delta_k + 2\alpha_k^{-1}\left( \Delta + \frac{\Lambda^2}{2\rho_k} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1-\tau} + \rho_k d_{\text{hi}}^2 + \frac{\epsilon_k D_{\mathbf{y}}}{4} + L_k D_{\mathbf{x}}^2 \right)\right) && (104)\\
T_k &= \left\lceil 16\left( 2\Delta + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} + \frac{\Lambda^2}{2\rho_k} + \frac{\epsilon_k D_{\mathbf{y}}}{4} \right) L_k \epsilon_k^{-2} \right. \\
&\quad \left. + 8(1 + 4D_{\mathbf{y}}^2 L_k^2 \epsilon_k^{-2})\rho_k^{-1} - 1 \right\rceil_+, && (105)\\
N_k &= \left( \left\lceil 96\sqrt{2}\left(1 + (24L_k + 4\epsilon_k/D_{\mathbf{y}})L_k^{-1}\right)\right\rceil + 2 \right) \max\left\{2, \sqrt{D_{\mathbf{y}} L_k \epsilon_k^{-1}}\right\} \\
&\quad \times \left( (T_k+1)(\log M_k)_+ + T_k + 1 + 2T_k \log(T_k+1) \right). && (106)
\end{aligned}
$$

*Then for all $0 \le k \in \mathbb{K} - 1$, Algorithm 3 finds an $\epsilon_k$-primal-dual stationary point $(x^{k+1}, y^{k+1})$ of problem* (35) *satisfying*

$$
\begin{aligned}
\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) \le \Delta + F_{\text{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} \\
+ \frac{\epsilon_k D_{\mathbf{y}}}{4} + \frac{1}{2\rho_k}\left( L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2 L_k \right). \qquad (107)
\end{aligned}
$$

*Moreover, the total number of evaluations of $\nabla f$, $\nabla c$, $\nabla d$ and proximal operator of $p$ and $q$ performed in iteration $k$ of Algorithm 3 is no more than $N_k$, respectively.*

*Proof.* Observe from (1) and (5) that problem (35) 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, (33) 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 $\mathcal{X} \times \mathcal{Y}$, we can see that $h(x,y)$ 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 (36). Consequently, it follows from Theorem 2 that Algorithm 2 can be suitably applied to problem (35) for finding an $\epsilon_k$-primal-dual stationary point $(x^{k+1}, y^{k+1})$ of it.

In addition, by (5),(47), (78), (79) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ (see Algorithm 3), one has

$$\begin{aligned}
&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \overset{(5)(78)}{=} \min_x \max_y \left\{\mathcal{L}_{\mathbf{y}}(x,y,\lambda_{\mathbf{y}}^k;\rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\right\}\\
&\overset{(79)}{\ge} \min_x \left\{f^*(x) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)\right\} \overset{(47)}{\ge} F_{\rm low} - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 \ge F_{\rm low} - \frac{\Lambda^2}{2\rho_k}.
\end{aligned} \tag{108}$$

Let $(x^*, y^*)$ be an optimal solution of (1). It then follows that $c(x^*) \le 0$. Using this, (5), (44) and (82), we obtain that

$$\begin{aligned}
&\min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \le \max_y \mathcal{L}(x^*,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k)\\
&\overset{(5)}{=} \max_y \left\{F(x^*,y) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^*)]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right)\right\}\\
&\le \max_y \left\{F(x^*,y) - \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{y}}^k + \rho_k d(x^*,y)]_+\|^2 - \|\lambda_{\mathbf{y}}^k\|^2\right)\right\}\\
&\overset{(44)}{\le} F_{\rm hi} + \frac{1}{2\rho_k}\|\lambda_{\mathbf{y}}^k\|^2 \overset{(82)}{\le} F_{\rm hi} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{\Delta + D_{\mathbf{y}}}{1-\tau},
\end{aligned} \tag{109}$$

where the second inequality is due to $c(x^*) \le 0$. Moreover, it follows from this, (5), (33), (44), (82), $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ that

$$\begin{aligned}
&\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k) \overset{(5)}{\ge} \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\|[\lambda_{\mathbf{y}}^k + \rho_k d(x,y)]_+\|^2\right\}\\
&\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \frac{1}{2\rho_k}\left(\|\lambda_{\mathbf{y}}^k\| + \rho_k\|[d(x,y)]_+\|\right)^2\right\}\\
&\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{F(x,y) - \frac{1}{2\rho_k}\|\lambda_{\mathbf{x}}^k\|^2 - \rho_k^{-1}\|\lambda_{\mathbf{y}}^k\|^2 - \rho_k\|[d(x,y)]_+\|^2\right\}\\
&\ge F_{\rm low} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} - \rho_k d_{\rm hi}^2,
\end{aligned} \tag{110}$$

where the second inequality is due to $\lambda_{\mathbf{y}}^k \in \mathbb{R}_+^{\tilde{m}}$ and the last inequality is due to (33), (44), (82) and $\|\lambda_{\mathbf{x}}^k\| \leq \Lambda$.

To complete the rest of the proof, let

$$H(x,y) = \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \quad H^* = \min_x \max_y \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k), \tag{111}$$

$$H_{\mathrm{low}} = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(x,y,\lambda_{\mathbf{x}}^k,\lambda_{\mathbf{y}}^k;\rho_k). \tag{112}$$

In view of these, (100), (108), (109), (110), we obtain that

$$\max_y H(x_{\mathrm{init}}^k, y) \stackrel{(100)}{\leq} \Delta + F_{\mathrm{hi}} + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau},$$
$$F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k} \stackrel{(108)}{\leq} H^* \stackrel{(109)}{\leq} F_{\mathrm{hi}} + \frac{1}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{\Delta + D_{\mathbf{y}}}{1-\tau},$$
$$H_{\mathrm{low}} \stackrel{(110)}{\geq} F_{\mathrm{low}} - \frac{\Lambda^2}{2\rho_k} - \|\lambda_{\mathbf{y}}^0\|^2 - \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} - \rho_k d_{\mathrm{hi}}^2.$$

Using these, (45), and Theorem 2 with $\hat{x}^0 = x_{\mathrm{init}}^k$, $\epsilon = \epsilon_k$, $\hat{\epsilon}_0 = \epsilon_k/(2\sqrt{\rho_k})$, $L_{\nabla h} = L_k$, and $H$, $H^*$, $H_{\mathrm{low}}$ given in (111) and (112), we can conclude that Algorithm 2 performs at most $N_k$ evaluations of $\nabla f$, $\nabla c$, $\nabla d$ and proximal operator of $p$ and $q$ for finding an $\epsilon_k$-primal-dual stationary point of problem (35) satisfying (107). $\square$

The following lemma provides an upper bound on the violation of the conditions in (41) at $(x, \lambda_{\mathbf{x}}) = (x^{k+1}, \tilde{\lambda}_{\mathbf{x}}^{k+1})$ for $0 \leq k \in \mathbb{K} - 1$, where $\tilde{\lambda}_{\mathbf{x}}^{k+1}$ is given below.

**Lemma 10.** *Suppose that Assumptions 1, 4 and 5 hold. Let $D_{\mathbf{y}}$, $\Delta$ and $L$ be defined in* (11), (45) *and* (48), *$L_F$, $L_c$, $\delta_c$ and $\theta$ be given in Assumption 5, and $\tau$, $\rho_k$, $\Lambda$ and $\lambda_{\mathbf{y}}^0$ be given in Algorithm 3. Suppose that $(x^{k+1}, \lambda_{\mathbf{x}}^{k+1})$ is generated by Algorithm 3 for some $0 \leq k \in \mathbb{K} - 1$ with*

$$\rho_k \geq \max\Bigg\{\theta^{-1}\Lambda, \theta^{-2}\Big\{4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{D_{\mathbf{y}}}{2} + L_c^{-2} + 4D_{\mathbf{y}}^2 L + \Lambda^2\Big\},$$
$$\frac{4\|\lambda_{\mathbf{y}}^0\|^2}{\delta_d^2\tau} + \frac{8(\Delta + D_{\mathbf{y}})}{\delta_d^2\tau(1-\tau)}\Bigg\}. \tag{113}$$

*Let*

$$\tilde{\lambda}_{\mathbf{x}}^{k+1} = [\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+. \tag{114}$$

*Then we have*

$$\|[c(x^{k+1})]_+\| \leq \rho_k^{-1}\delta_c^{-1}\left(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1\right), \tag{115}$$
$$|\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1})\rangle| \leq \rho_k^{-1}\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1)\max\{\delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1), \Lambda\}. \tag{116}$$

*Proof.* One can observe from (5),(47), (78) and (79) that

$$\max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) = \max_y \mathcal{L}_{\mathbf{y}}(x^{k+1}, y, \lambda_{\mathbf{y}}^k; \rho_k) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)$$
$$\stackrel{(79)}{\geq} f^*(x^{k+1}) + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right)$$
$$\stackrel{(47)}{\geq} F_{\mathrm{low}} + \frac{1}{2\rho_k}\left(\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 - \|\lambda_{\mathbf{x}}^k\|^2\right).$$

By this inequality, (107) and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$, one has

$$\begin{aligned}
&\|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2 \le 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k F_{\text{low}} + \|\lambda_{\mathbf{x}}^k\|^2 \\
&\le 2\rho_k \max_y \mathcal{L}(x^{k+1}, y, \lambda_{\mathbf{x}}^k, \lambda_{\mathbf{y}}^k; \rho_k) - 2\rho_k F_{\text{low}} + \Lambda^2 \\
&\overset{(107)}{\le} 2\rho_k\Delta + 2\rho_k F_{\text{hi}} + 2\rho_k\Lambda + \rho_k(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{2\rho_k(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{\rho_k\epsilon_k D_{\mathbf{y}}}{2} \\
&\quad + L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2 L_k - 2\rho_k F_{\text{low}} + \Lambda^2.
\end{aligned}$$

This together with (45) and $\rho_k^2\|[c(x^{k+1})]_+\|^2 \le \|[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+\|^2$ implies that

$$\begin{aligned}
\|[c(x^{k+1})]_+\|^2 \le \rho_k^{-1} &\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{\epsilon_k D_{\mathbf{y}}}{2}\right) \\
&+ \rho_k^{-2}\left(L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2 L_k + \Lambda^2\right). \qquad (117)
\end{aligned}$$

In addition, we observe from (36), (48), (82), $\rho_k \ge 1$ and $\|\lambda_{\mathbf{x}}^k\| \le \Lambda$ that for all $0 \le k \le K$,

$$\begin{aligned}
\rho_k L_c^2 \le L_k &= L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\text{hi}} L_{\nabla c} + \|\lambda_{\mathbf{x}}^k\| L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\text{hi}} L_{\nabla d} + \|\lambda_{\mathbf{y}}^k\| L_{\nabla d} \\
&\le L_{\nabla f} + \rho_k L_c^2 + \rho_k c_{\text{hi}} L_{\nabla c} + \Lambda L_{\nabla c} + \rho_k L_d^2 + \rho_k d_{\text{hi}} L_{\nabla d} \\
&\quad + L_{\nabla d}\sqrt{\rho_k\left(\|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_y)}{1-\tau}\right)} \le \rho_k L. \qquad (118)
\end{aligned}$$

Using this relation, (113), (117), $\rho_k \ge 1$ and $\epsilon_k \le 1$, we have

$$\begin{aligned}
\|[c(x^{k+1})]_+\|^2 \le \rho_k^{-1} &\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{\epsilon_k D_{\mathbf{y}}}{2}\right) \\
&+ \rho_k^{-2}\left((\rho_k L_c^2)^{-1}\epsilon_k^2 + 4\rho_k D_{\mathbf{y}}^2 L + \Lambda^2\right) \\
\le \rho_k^{-1} &\left(4\Delta + 2\Lambda + \tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2 + \frac{2(\Delta + D_{\mathbf{y}})}{1-\tau} + \frac{D_{\mathbf{y}}}{2}\right) \\
&+ \rho_k^{-1}\left(L_c^{-2} + 4D_{\mathbf{y}}^2 L + \Lambda^2\right) \overset{(113)}{\le} \theta^2,
\end{aligned}$$

which together with (37) implies that $x^{k+1} \in \mathcal{F}(\theta)$.

It follows from $x^{k+1} \in \mathcal{F}(\theta)$ and Assumption 5(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 (37). Let $\bar{\mathcal{A}}(x^{k+1};\theta) = \{1, 2, \ldots, \tilde{n}\}\backslash\mathcal{A}(x^{k+1};\theta)$. Notice from (37) that $c_i(x^{k+1}) < -\theta$ for all $i \in \bar{\mathcal{A}}(x^{k+1};\theta)$. In addition, observe from (113) 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

$$\begin{aligned}
&v^T\nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} \overset{(114)}{=} v^T\nabla c(x^{k+1})[\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+ = \sum_{i=1}^{\tilde{n}} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \\
&= \sum_{i\in\mathcal{A}(x^{k+1};\theta)} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i + \sum_{i\in\bar{\mathcal{A}}(x^{k+1};\theta)} v^T\nabla c_i(x^{k+1})([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \\
&\le -\delta_c \sum_{i\in\mathcal{A}(x^{k+1};\theta)} ([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i = -\delta_c \sum_{i=1}^{\tilde{n}}([\lambda_{\mathbf{x}}^k + \rho_k c(x^{k+1})]_+)_i \overset{(114)}{=} -\delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1. \qquad (119)
\end{aligned}$$

Since $(x^{k+1}, y^{k+1})$ is an $\epsilon_k$-primal-dual stationary point of (35), it follows from (5) and (97) 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 (114) and $\lambda_{\mathbf{y}}^{k+1} = [\lambda_{\mathbf{y}}^k + \rho_x d(x^{k+1}, y^{k+1})]_+$ implies that

$$\|s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\| \leq \epsilon_k. \tag{120}$$

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 5(iii), we have

$$\begin{aligned}
\langle s, v\rangle &= \langle \nabla_x f(x^{k+1}, y^{k+1}), v\rangle + \lim_{t\to\infty} \alpha_t^{-1}\langle s_p, z^t - x^{k+1}\rangle \\
&= \lim_{t\to\infty} \alpha_t^{-1}(f(z^t, y^{k+1}) - f(x^{k+1}, y^{k+1})) + \lim_{t\to\infty} \alpha_t^{-1}\langle s_p, z^t - x^{k+1}\rangle \\
&\leq \lim_{t\to\infty} \alpha_t^{-1}(f(z^t, y^{k+1}) - f(x^{k+1}, y^{k+1})) + \lim_{t\to\infty} \alpha_t^{-1}(p(z^t) - p(x^{k+1})) \\
&= \lim_{t\to\infty} \alpha_t^{-1}(F(z^t, y^{k+1}) - F(x^{k+1}, y^{k+1})) \leq L_F \lim_{t\to\infty} \alpha_t^{-1}\|z^t - x^{k+1}\| = L_F,
\end{aligned} \tag{121}$$

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 (119), (120), (121), and $\|v\| = 1$, one has

$$\begin{aligned}
\epsilon_k &\geq \|s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\| \cdot \|v\| \\
&\geq \langle s + \nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, -v\rangle \\
&= -\langle s - \nabla_x d(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}, v\rangle - v^T\nabla c(x^{k+1})\tilde{\lambda}_{\mathbf{x}}^{k+1} \\
&\overset{(119)}{\geq} -\langle s, v\rangle - \|\nabla_x d(x^{k+1}, y^{k+1})\|\|\lambda_{\mathbf{y}}^{k+1}\|\|v\| + \delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1 \\
&\geq -L_F - L_d\|\lambda_{\mathbf{y}}^{k+1}\| + \delta_c\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1,
\end{aligned}$$

where the last inequality is due to (121), $\|v\| = 1$ and Assumption 1(iii). Notice from (113) that (85) holds. It then follows from (86) that $\|\lambda_{\mathbf{y}}^{k+1}\| \leq 2\delta_d^{-1}(\Delta + D_{\mathbf{y}})$, which together with the above inequality and $\epsilon_k \leq 1$ yields

$$\|\tilde{\lambda}_{\mathbf{x}}^{k+1}\| \leq \|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|_1 \leq \delta_c^{-1}(L_F + L_d\|\lambda_{\mathbf{y}}^{k+1}\| + \epsilon_k) \leq \delta_c^{-1}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1). \tag{122}$$

By this and (114), one can observe that

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

Hence, (115) holds as desired.

We next show that (116) holds. Indeed, by $\tilde{\lambda}_{\mathbf{x}}^{k+1} \geq 0$, (115) and (122), one has

$$\begin{aligned}
\langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, c(x^{k+1})\rangle &\leq \langle \tilde{\lambda}_{\mathbf{x}}^{k+1}, [c(x^{k+1})]_+\rangle \leq \|\tilde{\lambda}_{\mathbf{x}}^{k+1}\|\|[c(x^{k+1})]_+\| \\
&\overset{(115)(122)}{\leq} \rho_k^{-1}\delta_c^{-2}(L_F + 2L_d\delta_d^{-1}(\Delta + D_{\mathbf{y}}) + 1)^2.
\end{aligned} \tag{123}$$

Using a similar argument as for the proof of (88), 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 (122) 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 (116) then follows from this and (123). □

We are now ready to prove Theorem 3 using Lemmas 7, 9 and 10.

***Proof of Theorem 3.*** (i) Observe from the definition of $K$ in (46) and $\epsilon_k = \tau^k$ that $K$ is the smallest nonnegative integer such that $\epsilon_K \le \varepsilon$. Hence, Algorithm 3 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 (53), one can see that (91) and (113) holds for $k = K$. It then follows from Lemmas 7 and 10 that (54)-(59) hold.

(ii) Let $K$ and $N$ be given in (46) and (60). Recall from Lemma 9 that the number of evaluations of $\nabla f$, $\nabla c$, $\nabla d$, proximal operator of $p$ and $q$ performed by Algorithm 2 at iteration $k$ of Algorithm 3 is at most $N_k$, where $N_k$ is given in (106). By this and statement (i) of this theorem, one can observe that the total number of evaluations of $\nabla f$, $\nabla c$, $\nabla d$, proximal operator of $p$ and $q$ performed in Algorithm 3 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 (118) and Algorithm 3 that $\rho_k L_c^2 \le L_k \le \rho_k L$ and $\rho_k \ge 1 \ge \epsilon_k$. Using these, (49), (50), (51), (102), (103), (104) and (105), we obtain that

$$
\begin{aligned}
&1 \ge \alpha_k \ge \min\left\{1, \sqrt{4\epsilon_k/(\rho_k D_{\mathbf{y}} L)}\right\} \ge \epsilon_k^{1/2}\rho_k^{-1/2}\alpha, && (124)\\
&\delta_k \le (2 + \epsilon_k^{-1/2}\rho_k^{1/2}\alpha^{-1})\rho_k L D_{\mathbf{x}}^2 + \max\{1/D_{\mathbf{y}}, \rho_k L/4\} D_{\mathbf{y}}^2 \le \epsilon_k^{-1/2}\rho_k^{3/2}\delta, && (125)\\
&M_k \le \frac{16\max\left\{1/(2\rho_k L_c^2), 4/(\epsilon_k^{1/2}\rho_k^{-1/2}\alpha\rho_k L_c^2)\right\}\rho_k}{\left[(3\rho_k L + 1/(2D_{\mathbf{y}}))^2/\min\{\rho_k L_c^2, \epsilon_k/(2D_{\mathbf{y}})\} + 3\rho_k L + 1/(2D_{\mathbf{y}})\right]^{-2}\epsilon_k^2} \times \Bigg(\epsilon_k^{-1/2}\rho_k^{3/2}\delta \\
&\quad + 2\epsilon_k^{-1/2}\rho_k^{1/2}\alpha^{-1}\left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1-\tau} + \rho_k d_{\mathrm{hi}}^2 + \frac{D_{\mathbf{y}}}{4} + \rho_k L D_{\mathbf{x}}^2\right)\Bigg) && (126)\\
&\le \frac{16\epsilon_k^{-1/2}\rho_k^{-1/2}\max\left\{1/(2L_c^2), 4/(\alpha L_c^2)\right\}\rho_k}{\epsilon_k^2\rho_k^{-4}\left[(3L + 1/(2D_{\mathbf{y}}))^2/\min\{L_c^2, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^{-2}\epsilon_k^2} \times (\epsilon_k^{-1/2}\rho_k^{3/2}) \\
&\quad \times \left(\delta + 2\alpha^{-1}\left(\Delta + \frac{\Lambda^2}{2} + \frac{3}{2}\|\lambda_{\mathbf{y}}^0\|^2 + \frac{3(\Delta + D_{\mathbf{y}})}{1-\tau} + d_{\mathrm{hi}}^2 + \frac{D_{\mathbf{y}}}{4} + L D_{\mathbf{x}}^2\right)\right) \le \epsilon_k^{-5}\rho_k^6 M,\\
&T_k \le \Bigg\lceil 16\left(2\Delta + \Lambda + \frac{1}{2}(\tau^{-1} + \|\lambda_{\mathbf{y}}^0\|^2) + \frac{\Delta + D_{\mathbf{y}}}{1-\tau} + \frac{\Lambda^2}{2} + \frac{D_{\mathbf{y}}}{4}\right)\epsilon_k^{-2}\rho_k L \\
&\quad + 8(1 + 4D_{\mathbf{y}}^2\rho_k^2 L^2\epsilon_k^{-2})\rho_k^{-1} - 1\Bigg\rceil_+ \le \epsilon_k^{-2}\rho_k T,
\end{aligned}
$$

where (126) follows from (49), (50), (51), (124), (125), $\rho_k L_c^2 \le L_k \le \rho_k L$, and $\rho_k \ge 1 \ge \epsilon_k$. By the above inequalities, (106), (118), $T \ge 1$ and $\rho_k \ge 1 \ge \epsilon_k$, one has

$$
\begin{aligned}
\sum_{k=0}^K N_k &\le \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (24\rho_k L + 4/D_{\mathbf{y}})/(\rho_k L_c^2)\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}\rho_k L\epsilon_k^{-1}}\right\} \\
&\qquad \times \left((\epsilon_k^{-2}\rho_k T + 1)(\log(\epsilon_k^{-5}\rho_k^6 M))_+ + \epsilon_k^{-2}\rho_k T + 1 + 2\epsilon_k^{-2}\rho_k T\log(\epsilon_k^{-2}\rho_k T + 1)\right) \\
&\le \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (24L + 4/D_{\mathbf{y}})/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}} L}\right\}\epsilon_k^{-1/2}\rho_k^{1/2} \\
&\qquad \times \epsilon_k^{-2}\rho_k\left((T+1)(\log(\epsilon_k^{-5}\rho_k^6 M))_+ + T + 1 + 2T\log(\epsilon_k^{-2}\rho_k T + 1)\right) \\
&\le \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (24L + 4/D_{\mathbf{y}})/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}} L}\right\} \\
&\qquad \times \epsilon_k^{-5/2}\rho_k^{3/2} T\left(2(\log(\epsilon_k^{-5}\rho_k^6 M))_+ + 2 + 2\log(2\epsilon_k^{-2}\rho_k T)\right) \\
&\le \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (24L + 4/D_{\mathbf{y}})/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}} L}\right\} T
\end{aligned}
$$

$$\times\, \epsilon_k^{-5/2}\rho_k^{3/2}\left(14\log\rho_k - 14\log\epsilon_k + 2(\log M)_+ + 2 + 2\log(2T)\right), \tag{127}$$

By the definition of $K$ in (46), one has $\tau^K \geq \tau\varepsilon$. Also, notice from Algorithm 3 that $\rho_k = \tau^{-k}$. It then follows from these, (60) and (127) that

$$\begin{aligned}
\sum_{k=0}^{K} N_k \leq& \sum_{k=0}^{K}\left(\left\lceil 96\sqrt{2}\left(1+(24L+4/D_y)/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\}T \\
&\times \epsilon_k^{-4}\left(28\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
=& \left(\left\lceil 96\sqrt{2}\left(1+(24L+4/D_y)/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\}T \\
&\times \sum_{k=0}^{K}\tau^{-4k}\left(28k\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
\leq& \left(\left\lceil 96\sqrt{2}\left(1+(24L+4/D_y)/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\}T \\
&\times \sum_{k=0}^{K}\tau^{-4k}\left(28K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
\leq& \left(\left\lceil 96\sqrt{2}\left(1+(24L+4/D_y)/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\}T \\
&\times \tau^{-4K}(1-\tau^4)^{-1}\left(28K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
\leq& \left(\left\lceil 96\sqrt{2}\left(1+(24L+4/D_y)/L_c^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\}T(1-\tau^4)^{-1} \\
&\times \tau^{-4}\varepsilon^{-4}\left(28K\log(1/\tau) + 2(\log M)_+ + 2 + 2\log(2T)\right) \stackrel{(60)}{=} N,
\end{aligned}$$

where the second last inequality is due to $\sum_{k=0}^{K}\tau^{-4k} \leq \tau^{-4K}/(1-\tau^4)$, and the last inequality is due to $\tau^K \geq \tau\varepsilon$. 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] G. H. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] Y.-H. Dai and L. 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.

[13] 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.

[14] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(1):2450–2504, 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. *Journal of Machine Learning Research*, 24:1–63, 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 and first-order momentum methods from mini to minimax optimization. *The Journal of Machine Learning Research*, 23(1):1616–1685, 2022.

[23] C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889, 2020.

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

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

[26] 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.

[27] 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.

[28] C. Laidlaw, S. Singla, and S. Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

[29] 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.

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

[31] 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.

[32] 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.

[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] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, 1999.

[39] 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.

[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] H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

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

[43] 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.

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

[49] 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.

[50] 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.

[51] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[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*, pages 1–72, 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.