A first-order method for constrained nonconvex-nonconcave minimax optimization

Zhaosong Lu * Xiangyuan Wang *

October 11, 2025 (Revised: May 21, 2026)

Abstract

We study a class of constrained nonconvex-nonconcave minimax optimization problems in which the inner maximization involves potentially complex constraints. Under the assumption that the inner problem of a novel lifted minimax reformulation satisfies a local Kurdyka-Łojasiewicz (KL) condition, we show that the maximal function of the original problem enjoys a local generalized Hölder smoothness property. We also propose a sequential convex programming (SCP) method for solving constrained optimization problems and establish its convergence rate under a local KL condition. Leveraging these results, we develop an inexact proximal gradient method for the original minimax problem, where the inexact gradient of the maximal function is computed via the SCP method applied to a locally KL-structured subproblem. Finally, we establish complexity guarantees for the proposed method in computing an approximate stationary point of the original minimax problem.

Keywords: constrained nonconvex-nonconcave minimax optimization, local KL condition, local generalized Hölder smoothness, sequential convex programming method, inexact proximal gradient method, first-order oracle complexity

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

1 Introduction

In this paper, we consider a class of constrained nonconvex-nonconcave minimax optimization problems of the form

$$\min_x \max_{c(y)\le 0} \{f(x,y) + p(x) - q(y)\}, \tag{1}$$

where f is a smooth function that is possibly nonconvex in x and nonconcave in y, p and q are simple closed convex functions, and c is a smooth mapping. This problem arises in many applications and also appears as a subproblem when solving more general constrained minimax problems of the form

$$\min_{d(x)\le 0} \max_{c(y)\le 0} \{f(x,y) + p(x) - q(y)\}, \tag{2}$$

where d is a smooth mapping. In fact, by applying a penalty approach, one can naturally convert (2) into a sequence of subproblems

$$\min_x \max_{c(y)\le 0} \left\{f(x,y) + \rho_k \|[d(x)]_+\|^2 + p(x) - q(y)\right\},$$

which are clearly in the form of (1), where $0 < \rho_k \to \infty$ and $u_+ = \max\{u, 0\}$ for any vector u.

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, wan02269@umn.edu). This work was partially supported by the Air Force Office of Scientific Research under Award FA9550-24-1-0343, the Office of Naval Research under Award N00014-24-1-2702, and the National Science Foundation under Awards 2211491 and 2435911.

In recent years, considerable attention has been devoted to unconstrained nonconvex-nonconcave minimax problems of the following form:

$$\min_x \max_y \{f(x,y) + p(x) - q(y)\}. \tag{3}$$

This class of problems arises in a wide range of applications in machine learning and operations research, including generative adversarial networks [1, 12], reinforcement learning [9, 21], adversarial training [18, 26], and distributionally robust optimization [3, 4, 23]. Significant progress has been made in solving (3) under additional structural assumptions. For instance, several works study the special case with $q = 0$ and assume that the inner maximization problem in (3) satisfies a global Polyak–Łojasiewicz (PL) condition, which is generally weaker than strong concavity. Under this assumption, gradient descent–ascent type methods have been developed, and complexity guarantees have been established for obtaining approximate stationary points (see, e.g., [13, 20, 27, 28]). In addition, first-order methods have been proposed for problem (3) from the perspective of variational inequalities, typically assuming the existence of a weak Minty variational inequality solution (see, e.g., [5, 8, 15, 22]).

More recently, [14, 30, 31] studied problem (3) under a global KL condition, where p and q are indicator functions of simple convex compact sets. This setting generalizes that of [13, 20, 27, 28], since the KL condition extends the PL condition (the latter corresponding to the KL condition with exponent $1/2$). However, requiring the KL property to hold globally is often too restrictive in practice. To address this, [17] considered problem (3) with p and q being simple closed convex functions under a local KL condition. Specifically, for each fixed outer variable $x \in \operatorname{dom} p$, the KL property is assumed to hold only on a level set of the inner variable y, where this level set may depend on x and may shrink as x approaches a stationary point of problem (3). Under this weaker assumption, a local generalized Hölder smoothness property of the associated maximal function was established. Leveraging this property, an inexact proximal gradient method was developed, in which the inexact gradient of the maximal function is computed by applying a proximal gradient method to a locally KL-structured subproblem. Complexity guarantees were then established for finding an approximate stationary point of problem (3).

Despite recent advances, existing results primarily focus on nonconvex-nonconcave minimax problems with unconstrained inner maximization. To the best of our knowledge, no algorithmic framework has been developed for the constrained counterpart (1), where the inner maximization problem involves potentially complex constraints. In this paper, we study problem (1) under the assumption that a novel *lifted minimax* problem—equivalent to (1)—satisfies a local KL condition analogous to that considered for problem (3) (see Assumption 1). We establish that the maximal function $F^*(x) := \max_{c(y)\le 0}\{f(x,y) - q(y)\}$ exhibits a local generalized Hölder smoothness property. In addition, for any fixed outer variable $x \in \operatorname{dom} p$, we propose a sequential convex programming (SCP) method to solve the inner maximization subproblem and establish its convergence rate under the local KL condition. Building on these results, we develop an inexact proximal gradient method for solving $\min_x\{F^*(x) + p(x)\}$, which is equivalent to (1). Specifically, given the current iterate (x^k, y^{k-1}), we apply the SCP method to approximately solve $\max_{c(y)\le 0}\{f(x^k,y) - q(y)\}$ initialized at y^{k-1}, and obtain an approximate solution y^k. We then update x^{k+1} via an inexact proximal gradient step, using $-\nabla_x f(x^k, y^k)$ as the forward direction together with a suitably chosen step size. Finally, we establish complexity guarantees for the proposed method in computing an approximate stationary point of problem (1).

The main contributions of this paper are summarized below.

- We establish a local generalized Hölder smoothness property for the maximal function F^* under a local KL condition imposed on a novel lifted minimax problem, which is crucial for developing a method for solving problem (1).
- We propose a sequential convex programming method for solving a constrained optimization

problem and establish its convergence rate under a local KL condition. This method serves as a subroutine for solving problem (1).

- We propose an inexact proximal gradient method for finding approximate stationary points of problem (1), and show that it achieves an *iteration complexity* of $\widetilde{\mathcal{O}}\big(\epsilon^{-\max\{(1-\theta)^{-1},\theta^{-1}\sigma\}}\big)$, and a *first-order oracle complexity* of $\widetilde{\mathcal{O}}\big(\epsilon^{-(1-\theta)^{-1}(2\theta^2-2\theta+1)\max\{(1-\theta)^{-1},\theta^{-1}\sigma\}}\big)$, measured by the number of gradient evaluations, for finding an $\mathcal{O}(\epsilon)$-approximate stationary point of (1), where θ and σ are the parameters of the local KL condition.

The rest of this paper is organized as follows. Subsection 1.1 introduces the notation, terminology, and assumptions used throughout the paper. In Section 2, we study the theoretical properties of problem (1). Section 3 presents a sequential convex programming method for a constrained optimization problem satisfying a local KL property. In Section 4, we propose an inexact proximal gradient method for solving problem (1) and establish its complexity results. Section 5 presents preliminary numerical results illustrating the performance of the proposed method. Finally, we provide the proof of the main results in Section 6.

1.1 Notation, terminology, and assumptions

The following notation will be used throughout the paper. Let $\mathbb{R}^n$ stand for the n-dimensional Euclidean space, and $\overline{\mathbb{R}} = (-\infty, \infty]$. The standard inner product, ℓ_1-norm, ℓ_∞-norm, and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$, $\|\cdot\|_1$, $\|\cdot\|_\infty$, and $\|\cdot\|$, respectively. For any two points $u, v \in \mathbb{R}^n$, the notation $[u, v]$ denotes the line segment connecting u and v. Given a point x and a closed set $S \subset \mathbb{R}^n$, let $\mathrm{dist}(x, S)$ stand for the distance from x to S, and δ_S the indicator function of S. The *regular normal cone* and the *normal cone* (i.e., the *limiting normal cone*) of S at $x \in S$ are denoted by $\widehat{\mathcal{N}}_S(x)$ and $\mathcal{N}_S(x)$, respectively (see [25, Definition 6.3]). The closed ball centered at $x \in \mathbb{R}^n$ with radius r is denoted by $\mathcal{B}(x, r)$. In addition, $\mathrm{conv}(\cdot)$, $\mathrm{aff}(\cdot)$, and $\mathrm{int}(\cdot)$ denote the convex hull, affine hull, and interior of the associated set, respectively.

A function $\phi : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called L_ϕ-*Lipschitz continuous* on $\mathcal{X}$ if $|\phi(x) - \phi(y)| \le L_\phi\|x - y\|$ for all $x, y \in \mathcal{X}$, and $L_{\nabla\phi}$-*smooth* on $\mathcal{X}$ if $\|\nabla\phi(x) - \nabla\phi(y)\| \le L_{\nabla\phi}\|x - y\|$ for all $x, y \in \mathcal{X}$. For a closed convex function $p : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *proximal operator* associated with p is defined as

$$\mathrm{prox}_p(x) := \operatorname*{arg\,min}_{x' \in \mathbb{R}^n} \Big\{ p(x') + \frac{1}{2}\|x - x'\|^2 \Big\}.$$

For a function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, its *domain* is defined as $\mathrm{dom}\,\phi = \{x : \phi(x) < \infty\}$. Such ϕ is called *proper* if $\mathrm{dom}\,\phi \neq \emptyset$, and it is called *closed* or *lower semicontinuous* if $\liminf_{z\to x} \phi(z) \ge \phi(x)$ holds for all $x \in \mathbb{R}^n$. The *regular subdifferential* (see, e.g., [25, Definition 8.3(a)]) of a proper closed function ϕ at $x \in \mathrm{dom}\,\phi$ is defined as

$$\widehat{\partial}\phi(x) := \left\{ v \in \mathbb{R}^n : \liminf_{z \to x,\, z \neq x} \frac{\phi(z) - \phi(x) - \langle v, z - x\rangle}{\|z - x\|} \ge 0 \right\}.$$

Let $z \xrightarrow{\phi} x$ denote $z \to x$ and $\phi(z) \to \phi(x)$. The *limiting subdifferential* (see, e.g., [25, Definition 8.3(b)]) of a proper closed function ϕ at $x \in \mathrm{dom}\,\phi$ is defined as

$$\partial\phi(x) := \left\{ v \in \mathbb{R}^n : \exists\, x^k \xrightarrow{\phi} x,\ v^k \to v \text{ with } v^k \in \widehat{\partial}\phi(x^k) \right\}.$$

We use $\partial_{x_i}\phi$ to denote the limiting subdifferential with respect to x_i. For an upper semicontinuous function ϕ, its limiting subdifferential is defined as $\partial\phi = -\partial(-\phi)$. If ϕ is continuously differentiable, then $\partial\phi$ coincides with the gradient $\nabla\phi$. Besides, if ϕ is convex, then $\partial\phi$ corresponds to the classical convex subdifferential. It is well-known that $\partial(\phi_1 + \phi_2)(x) = \nabla\phi_1(x) + \partial\phi_2(x)$ if ϕ_1 is continuously differentiable at x and ϕ_2 is lower or upper semicontinuous at x (see, e.g., [25, Exercise 8.8(c)]).

Suppose that ϕ is a locally Lipschitz continuous function on $\mathcal{X}$. The *restricted Clarke subdifferential* of ϕ with respect to $\mathcal{X}$, denoted by $\partial^{\mathrm{C}}_{\mathcal{X}}\phi(x)$, is introduced in [17] and defined as

$$\partial^{\mathrm{C}}_{\mathcal{X}}\phi(x) := \mathrm{conv}\{v : \exists\, x^k \in \mathcal{X} \to x \text{ such that } \nabla\phi(x^k) \to v\} \quad \forall x \in \mathcal{X}.$$

When $\partial^{\mathrm{C}}_{\mathcal{X}}\phi(x)$ is a singleton, we denote its unique element by $\nabla^{\mathrm{C}}_{\mathcal{X}}\phi(x)$. Furthermore, suppose additionally that $\partial^{\mathrm{C}}_{\mathcal{X}}\phi(x)$ is a singleton for all $x \in \mathcal{X}$. The function ϕ is said to be *generalized Hölder smooth* on $\mathcal{X}$ (see [17, Definition 1]) if there exist $L_1 \geq 0$, $L_2 \geq 0$, and $\nu \in (0, 1]$ such that

$$\|\nabla^{\mathrm{C}}_{\mathcal{X}}\phi(x) - \nabla^{\mathrm{C}}_{\mathcal{X}}\phi(y)\| \leq L_1\|x - y\| + L_2\|x - y\|^{\nu} \quad \forall x, y \in \mathcal{X}.$$

For a closed function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *slope* of ϕ at $x \in \mathrm{dom}\,\phi$ is defined as

$$|\nabla\phi|(x) := \limsup_{z \to x} \frac{\big(\phi(x) - \phi(z)\big)_+}{\|x - z\|}, \tag{4}$$

where $t_+ = \max\{0, t\}$ for any $t \in \mathbb{R}$. The *limiting slope* of ϕ at $x \in \mathrm{dom}\,\phi$ is defined as

$$\overline{|\nabla\phi|}(x) := \liminf_{z \xrightarrow{\phi} x} |\nabla\phi|(z). \tag{5}$$

If ϕ is differentiable at x, $|\nabla\phi|(x)$ coincides with $\|\nabla\phi(x)\|$. When ϕ is a convex function, $|\nabla\phi|(x)$ reduces to $\mathrm{dist}(0, \partial\phi(x))$ for all $x \in \mathrm{dom}\,\phi$. For more details on the slope and limiting slope, see, for example, [11, Section 2].

We now introduce the notion of an approximate stationary point for the problem $\min_x \phi(x)$, where ϕ is a closed function. Since the minimax problem (1) can be viewed as a special case of this general problem, the following definition applies directly to (1) as well.

Definition 1 ((r, ϵ)**-stationary point).** *Suppose ϕ is a closed function. For any $\epsilon > 0$ and $r > 0$, a point $\bar{x}$ is called an (r, ϵ)-stationary point of the problem $\min_x \phi(x)$ if $\bar{x} \in \mathrm{dom}\,\phi$ and $\mathrm{dist}(\bar{x}, \mathcal{X}_\epsilon) \leq r$, where $\mathcal{X}_\epsilon = \{x \in \mathrm{dom}\,\phi : \mathrm{dist}(0, \partial\phi(x)) \leq \epsilon\}$.*

The above definition is closely related to the notion of an ϵ-optimization-stationary point (see, e.g., [14, Definition 3.1(a)]). In particular, suppose that ϕ is a ρ-weakly convex function, i.e., $\phi(\cdot) + \rho\|\cdot\|^2/2$ is convex, and that $\bar{x}$ is an ϵ-optimization-stationary point of the problem $\min_x \phi(x)$, that is, $\bar{x} \in \mathrm{dom}\phi$ and $\|\mathrm{prox}_{\phi/t}(\bar{x}) - \bar{x}\| \leq \epsilon$ for some $t > \rho$. It can be verified that $\bar{x}$ is an $(\epsilon, t\epsilon)$-stationary point of $\min_x \phi(x)$. It should also be noted that when ϕ is locally Lipschitz continuous, any (r, ϵ)-stationary point $\bar{x}$ of ϕ is also an (r, ϵ)-Goldstein stationary point of ϕ, that is, $\mathrm{dist}(0, \partial_r\phi(\bar{x})) \leq \epsilon$, where

$$\partial_r\phi(\bar{x}) := \mathrm{conv}\Big(\bigcup_{x \in \mathcal{B}(\bar{x}, r)} \partial\phi(x)\Big).$$

We next introduce additional notation for problem (1). For convenience, we define

$$\mathcal{X} := \mathrm{dom}\, p, \quad \mathcal{Y} := \mathrm{dom}\, q, \quad \mathcal{S} := \mathcal{Y} \cap \{y : c(y) \leq 0\}, \tag{6}$$

$$D_{\mathcal{Y}} = \max_{u, v \in \mathcal{Y}} \|u - v\|, \quad M_q = \max_{u, v \in \mathcal{Y}} \{q(u) - q(v)\}, \tag{7}$$

$$F(x, y) := f(x, y) - q(y) - \delta_{c(\cdot) \leq 0}(y), \quad F^*(x) = \max_y F(x, y), \quad Y^*(x) = \{y : F(x, y) = F^*(x)\}, \tag{8}$$

$$\Psi(x) := F^*(x) + p(x), \quad \Psi^* := \min_x \Psi(x). \tag{9}$$

To study problem (1), we introduce the following *lifted minimax* problem:

$$\min_x \max_{\bar{c}(y,z) \leq 0} \{f(x, y) + p(x) - q(y) - \delta_{\mathcal{Y}}(z)\}, \tag{10}$$

where

$$\bar{c} = (\bar{c}_1, \ldots, \bar{c}_m) \;\text{ with }\; \bar{c}_j(y,z) := c_j(z) + \langle \nabla c_j(z), y - z\rangle + \frac{L_{c_j}}{2}\|y-z\|^2 \quad j = 1, \ldots, m, \tag{11}$$

and L_{c_j} is the Lipschitz smoothness constant of c_j (see Assumption 1 below). Interestingly, the lifted minimax problem (10) is equivalent to the original minimax problem (1) (see Lemma 3 and Remark 4). For notational convenience, we further define

$$\bar{F}(x,y,z) := f(x,y) - q(y) - \delta_{\bar{c}(\cdot,\cdot)\le 0}(y,z) - \delta_{\mathcal{Y}}(z), \tag{12}$$

$$\bar{F}^*(x) := \max_{y,z} \bar{F}(x,y,z), \quad \overline{Y}^*(x) := \{(y,z) : \bar{F}(x,y,z) = \bar{F}^*(x)\}, \quad \bar{\Psi}(x) := \bar{F}^*(x) + p(x). \tag{13}$$

Consequently, problem (10) can be equivalently written as

$$\min_x \max_{y,z} \{\bar{F}(x,y,z) + p(x)\}. \tag{14}$$

In what follows, we introduce the assumptions for problem (1) and its equivalent problem (14). In particular, we assume that problem (1) has at least one optimal solution and that the following assumption holds.

Assumption 1. *(i) For any fixed $y \in \mathcal{Y}$, the function $f(\cdot, y)$ is L_f-Lipschitz continuous on $\mathcal{X}$. Moreover, the function $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ is $L_{\nabla f}$-smooth on $\mathcal{X} \times \mathcal{Y}$.*

(ii) The functions p, q are proper closed convex, and the proximal operators of p and q can be computed exactly. Moreover, $\operatorname{dom} q$ (i.e., $\mathcal{Y}$) is compact. In addition, we assume that $\operatorname{aff}(\mathcal{X}) = \mathbb{R}^{n_1}$.[1]

(iii) The mapping $c = (c_1, \ldots, c_m)$ satisfies that each component $c_j : \mathbb{R}^{n_2} \to \mathbb{R}$ is L_{c_j}-smooth and twice continuously differentiable on $\mathcal{Y}$.

(iv) The function $\bar{F}$ satisfies the following local Kurdyka-Łojasiewicz (KL) condition with respect to (y,z): there exist constants $C > 0$, $\theta \in [1/2, 1)$, $\gamma > 0$, and $\sigma \ge 0$ such that for any $x \in \mathcal{X}$,

$$C(\bar{F}^*(x) - \bar{F}(x,y,z))^\theta \le \operatorname{dist}(0, \partial_{(y,z)} \bar{F}(x,y,z)) \qquad \forall (y,z) \in \overline{\mathcal{L}}(x), \tag{15}$$

where

$$\overline{\mathcal{L}}(x) := \{(y,z) : 0 < \bar{F}^*(x) - \bar{F}(x,y,z) \le \gamma \operatorname{dist}(0, \partial\bar{\Psi}(x))^\sigma\}. \tag{16}$$

We now make some remarks on Assumption 1.

Remark 1. *(i) The Lipschitz continuity of $f(\cdot, y)$ in Assumption 1(i) is used to establish the Lipschitz continuity of the maximal function F^*. Consequently, by Rademacher's theorem, it follows that F^* is differentiable almost everywhere. This property plays a key role in establishing the local generalized Hölder smoothness property of F^* in Theorem 1.*

(ii) The twice continuous differentiability of the mapping c in Assumption 1(iii) is imposed to ensure that the constraint function $\bar{c}$ in the lifted minimax problem (10) is continuously differentiable. This property will be used to analyze Algorithm 1 for solving $\max_y F(x,y)$.

(iii) Assumption 1(iv) requires that $\bar{F}(x,\cdot,\cdot)$ satisfy the KL property only on the level set $\overline{\mathcal{L}}(x)$, which we refer to as a local KL condition (in contrast to its global counterpart). When $\sigma > 0$, this level set depends on x, in particular on $\operatorname{dist}(0, \partial\bar{\Psi}(x))$. If x is far from a stationary point of $\bar{\Psi}$, then $\overline{\mathcal{L}}(x)$ is relatively large, and hence the KL property holds on a correspondingly large level set of the maximization problem in (10). As x approaches a stationary point, however,

[1]The assumption $\operatorname{aff}(\mathcal{X}) = \mathbb{R}^{n_1}$ is imposed merely for convenience. Without this assumption, the results of the paper remain valid and the analysis proceeds identically, except that the gradients and subdifferentials should be understood as being defined relative to $\operatorname{aff}(\mathcal{X})$.

the level set shrinks, and the KL property is required only on a correspondingly smaller level set. Moreover, this local KL condition is used to analyze Algorithm 1. Since this algorithm is based on the reformulation (14) *of problem* (1)*, it is natural to impose the local KL condition on* $\bar{F}(x,\cdot,\cdot)$.

(iv) For each $x \in \mathcal{X}$*, if* $\bar{F}(x,\cdot,\cdot)$ *is semi-algebraic, subanalytic, or a structured nonsmooth function, then it satisfies a local KL condition with its own KL constant, KL exponent, and corresponding level set (see, e.g., [2, 6, 7]). Consequently, for such* $\bar{F}(x,\cdot,\cdot)$*, if there exist a common KL constant and exponent for all* $x \in \mathcal{X}$*, and if the level set does not shrink faster than* $\mathcal{O}(\mathrm{dist}(0,\partial\bar{\Psi}(x))^\sigma)$ *for some* $\sigma > 0$ *as* x *approaches a stationary point, then Assumption 1(iv) is satisfied. Otherwise, it may be challenging to develop a first-order method with attractive complexity guarantees for finding an approximate stationary point of problem* (1).

We end this subsection with a simple constrained minimax problem that satisfies the local KL condition in Assumption 1(iv) but not the global one (which requires the KL inequality to hold for all $(y,z) \in \mathrm{dom}\,\bar{F}(x,\cdot,\cdot)$), thereby illustrating that the local KL condition applies to a broader class of minimax problems.

Example 1. Consider the problem

$$\min_{1\le x\le 2}\ \max_{c(y)\le 0}\{-x^2\sin y - \delta_{[0,2\pi/3]}(y)\},$$

where $c(y) = y^3/6 + y - \pi$, which is L_c-smooth on $[0,2\pi/3]$ with $L_c = 2\pi/3$. Notice that this problem is a special case of (1) with $f(x,y) = -x^2\sin y$, $p(x) = \delta_{[1,2]}(x)$, and $q(y) = \delta_{[0,2\pi/3]}(y)$. We can show that the function $\bar{F}$ defined in (12) does not satisfy the global KL condition at any $x \in \mathrm{dom}\,p$, since the KL inequality fails when $y = z = \pi/2$. Indeed, one can observe from the definitions of $\bar{c}$ and c that $\bar{c}(\pi/2,\pi/2) = c(\pi/2) = \pi^3/48 - \pi/2 < 0$, which along with (12) yields $\partial_{(y,z)}\bar{F}(x,\pi/2,\pi/2) = (-x^2\cos(\pi/2), 0) = (0,0)$. On the other hand, note that $\bar{F}^*(x) = 0$ for all $x \in \mathrm{dom}\,p$. Using this, (12), (13), and $\bar{c}(\pi/2,\pi/2) \le 0$, one has

$$\bar{F}^*(x) - \bar{F}(x,\pi/2,\pi/2) = 0 - (-x^2\sin(\pi/2)) = x^2 \ge 1 \qquad \forall x \in \mathrm{dom}\,p.$$

Combining these, we see that the KL inequality does not hold for $y = z = \pi/2$ at any $x \in \mathrm{dom}\,p$.

However, it can be shown that the KL condition holds on the level set $\overline{\mathcal{L}}(x)$ defined in (16) with $\gamma = 1/2$ and $\sigma = 0$ for all $x \in \mathrm{dom}\,p$. To this end, it suffices to show that for any $x \in \mathrm{dom}\,p$, the condition holds on a larger level set $\widehat{\mathcal{L}}_x = \{(y,z) : 0 < \bar{F}^*(x) - \bar{F}(x,y,z) \le x^2/2\}$. Indeed, one can show that $\widehat{\mathcal{L}}_x = \{(y,z) : y \in (0,\pi/6],\ z \in [0,2\pi/3],\ \bar{c}(y,z) \le 0\}$ for all $x \in \mathrm{dom}\,p$. Moreover, it can be verified that $\bar{c}(y,z) < 0$ for all $(y,z) \in (0,\pi/6] \times [0,2\pi/3]$, due to $\bar{c}(0,z) < 0$, $\bar{c}(\pi/6,z) < 0$ for all $z \in [0,2\pi/3]$, and the convexity of $\bar{c}(\cdot,z)$. In view of these and (12), we see that $\partial_{(y,z)}\bar{F}(x,y,z) = (-x^2\cos y, -\mathcal{N}_{[0,2\pi/3]}(z))$ for all $x \in \mathrm{dom}\,p$ and $(y,z) \in \widehat{\mathcal{L}}_x$. Then, one can verify that

$$C(\bar{F}^*(x) - \bar{F}(x,y,z))^\theta \le \mathrm{dist}(0,\partial_{(y,z)}\bar{F}(x,y,z)) \qquad \forall (y,z) \in \widehat{\mathcal{L}}_x$$

for all $x \in \mathrm{dom}\,p$ with $C = (3/2)^{1/2}$ and $\theta = 1/2$.

2 Theoretical properties of problem (1)

In this section, we establish several theoretical properties for problem (1), which will be used later for algorithm design and analysis.

The following result shows that F^* possesses a local generalized Hölder smoothness property. Its proof is deferred to Subsection 6.1.

Theorem 1. *Suppose that Assumption 1 holds. Let* $\epsilon > 0$ *be given and*

$$\mathcal{U}_\epsilon := \{x \in \mathcal{X} : \mathrm{dist}(0,\partial\Psi(x)) > \epsilon\}. \tag{17}$$

Then the following statements hold.

(i) $\partial^{\mathrm{C}}_{\mathcal{X}} F^*(x)$ *is a singleton for all* $x \in \mathcal{U}_\epsilon$, *and* F^* *is differentiable on* $\mathcal{U}_\epsilon \cap \mathrm{int}(\mathcal{X})$.

(ii) For any $x, x' \in \mathcal{U}_\epsilon \cap \mathrm{int}(\mathcal{X})$ *satisfying* $\|x - x'\| \le \gamma\epsilon^\sigma/(2L_f)$, *we have*

$$\|\nabla F^*(x) - \nabla F^*(x')\| \le L_{\nabla f}\|x - x'\| + (1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}\|x - x'\|^{\frac{1-\theta}{\theta}}.$$

(iii) For any $x, x' \in \mathcal{U}_\epsilon$ *satisfying* $\|x - x'\| \le \gamma\epsilon^\sigma/(4L_f)$, *we have*

$$\|\nabla^{\mathrm{C}}_{\mathcal{X}} F^*(x) - \nabla^{\mathrm{C}}_{\mathcal{X}} F^*(x')\| \le L_{\nabla f}\|x - x'\| + (1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}\|x - x'\|^{\frac{1-\theta}{\theta}}.$$

(iv) It holds that

$$\nabla^{\mathrm{C}}_{\mathcal{X}} F^*(x) = \nabla_x f(x, y^*) \qquad \forall x \in \mathcal{U}_\epsilon,\ y^* \in Y^*(x).$$

The following result is a consequence of Theorem 1, whose proof is identical to that of [17, Corollary 1] and thus omitted.

Corollary 1. *Let* $\epsilon > 0$ *be given and* $\mathcal{U}_\epsilon$ *be defined in* (17). *Suppose that Assumption 1 holds. Then, for any* x, x' *satisfying* $[x, x'] \subseteq \mathcal{U}_\epsilon$ *and* $\|x - x'\| \le \gamma\epsilon^\sigma/(4L_f)$, *we have*

$$F^*(x) \le F^*(x') + \langle \nabla^{\mathrm{C}}_{\mathcal{X}} F^*(x'), x - x'\rangle + \frac{1}{2}L_{\nabla f}\|x - x'\|^2 + \frac{M}{1+\nu}\|x - x'\|^{1+\nu}, \tag{18}$$

where

$$M := (1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}, \quad \nu := \theta^{-1}(1-\theta). \tag{19}$$

The theorem below establishes a local $(1-\theta)^{-1}$-growth property of $F(x, \cdot)$ for every $x \in \mathcal{X}$, whose proof is deferred to Subsection 6.1.

Theorem 2. *Suppose that Assumption 1 holds. Then it holds that for any* $x \in \mathcal{X}$,

$$F^*(x) - F(x, y) \ge (C(1-\theta))^{\frac{1}{1-\theta}}\,\mathrm{dist}(y, Y^*(x))^{\frac{1}{1-\theta}} \qquad \forall y \in \mathcal{L}(x), \tag{20}$$

where

$$\mathcal{L}(x) := \{y : 0 < F^*(x) - F(x, y) \le \gamma\,\mathrm{dist}(0, \partial\Psi(x))^\sigma\}. \tag{21}$$

We next introduce the Mangasarian–Fromovitz constraint qualification (MFCQ) for problem (1), which will be used frequently in the paper.

Assumption 2 (MFCQ). *For every* $y \in \mathcal{S}$, *there exists some* $\tilde{y} \in \mathcal{Y}$ *such that*

$$\langle \nabla c_j(y), \tilde{y} - y\rangle < 0 \qquad \forall j \in I(y) := \{i : c_i(y) = 0\},$$

where $\mathcal{Y}$ *and* $\mathcal{S}$ *are defined in* (6).

Under the above MFCQ condition, we establish a key property of the mapping $\bar{c}(\cdot, \cdot)$, which will be used subsequently in the paper.

Theorem 3. *Let* $\bar{c}$ *be defined in* (11). *Suppose that Assumptions 1 and 2 hold. Then there exists a constant* $\zeta > 0$ *such that for every* $z \in \mathcal{S}$, *one can find a point* $y \in \mathcal{Y}$ *(possibly depending on* z*) satisfying* $\bar{c}(y, z) \le -\zeta$.

Proof. Suppose for contradiction that the conclusion of this theorem does not hold. Then there exist a positive sequence $\{\zeta_k\}$ and a sequence $\{z^k\} \subset \mathcal{S}$ such that $\lim_{k\to\infty}\zeta_k = 0$ and

$$\bar{c}(y, z^k) \not\le -\zeta_k \quad \forall y \in \mathcal{Y},\ \forall k. \tag{22}$$

Since $\mathcal{Y}$ is compact and c is continuous, it follows that $\mathcal{S}$ is compact. Hence, the sequence $\{z^k\}$ admits a convergent subsequence. Passing to such a subsequence if necessary, we may assume without loss of generality that $z^k \to z$ for some $z \in \mathcal{S}$. By this fact and Assumption 2, there exists some $\tilde{z} \in \mathcal{Y}$ such that

$$\langle \nabla c_j(z), \tilde{z} - z \rangle < 0 \qquad \forall j \in I(z) = \{i : c_i(z) = 0\}.$$

Let $y(t) = z + t(\tilde{z} - z)$ for all t. Then, for each $j \in I(z)$, one has

$$\lim_{t \downarrow 0} \frac{\bar{c}_j(y(t), z)}{t} = \langle \nabla c_j(z), \tilde{z} - z \rangle < 0,$$

which implies that $\bar{c}_j(y(t), z) < 0$ for all sufficiently small $t > 0$. In addition, for each $j \notin I(z)$, we have $\bar{c}_j(y(t), z) < 0$ for all sufficiently small $t > 0$, thanks to $\bar{c}_j(z, z) = c_j(z) < 0$ and the continuity of $\bar{c}_j$. Moreover, it follows from $z, \tilde{z} \in \mathcal{Y}$ and the convexity of $\mathcal{Y}$ that $y(t) \in \mathcal{Y}$ for all $t \in [0, 1]$. Consequently, there exists some $\bar{y} \in \mathcal{Y}$ such that

$$\bar{c}(\bar{y}, z) < 0. \tag{23}$$

On the other hand, since $\bar{y} \in \mathcal{Y}$, it follows from (22) that $\bar{c}(\bar{y}, z^k) \not\leq -\zeta_k$ for all k. Hence, there exists some $\bar{j}$ such that $\bar{c}_{\bar{j}}(\bar{y}, z^k) > -\zeta_k$ holds for infinitely many k. Passing to a subsequence if necessary, we may assume without loss of generality that this inequality holds for all k. Taking limits on both sides of this inequality as $k \to \infty$ and using $\lim_{k\to\infty} \zeta_k = 0$, $\lim_{k\to\infty} z^k = z$, and the continuity of $\bar{c}_{\bar{j}}$, we obtain that $\bar{c}_{\bar{j}}(\bar{y}, z) \geq 0$, which contradicts (23). Hence, the conclusion of this theorem holds. □

3 A sequential convex programming method for constrained KL function minimization

In this section, we consider constrained optimization problems of the form

$$h^* = \min_{c(z) \leq 0} \{h(z) := g(z) + q(z)\}, \tag{24}$$

where q and c are defined in Section 1 that satisfy Assumption 1, and g and h satisfy the following assumption.

Assumption 3. *The function g is L-smooth on $\mathcal{Y}$, and $\bar{h}$ satisfies the following KL condition:*

$$C(\bar{h}(z, w) - \bar{h}^*)^\theta \leq \operatorname{dist}(0, \partial \bar{h}(z, w)) \qquad \forall (z, w) \text{ with } \bar{h}^* < \bar{h}(z, w) \leq \bar{h}^* + \eta \tag{25}$$

for some constants $C, \eta > 0$ and $\theta \in [1/2, 1)$, where $\mathcal{Y} = \operatorname{dom} q$, and

$$\bar{h}(z, w) := h(z) + \delta_{\bar{c}(\cdot,\cdot) \leq 0}(z, w) + \delta_{\mathcal{Y}}(w), \qquad \bar{h}^* := \min_{z,w} \bar{h}(z, w), \tag{26}$$

and $\bar{c}$ is given in (11).

Sequential convex programming (SCP) methods have been studied in the literature (see, e.g., [16, 29]) for solving problem (24). These methods solve a sequence of simple convex optimization problems of the form (27). In particular, [16, 29] apply line search schemes to both the objective and the constraints of (24) to generate a sequence of feasible points that ensures sufficient reduction in the objective function h along the sequence. Motivated by these works, we propose a variant of the SCP method for solving (24), which will subsequently serve as a subroutine for solving problem (1). To suit our purpose, we apply the line search scheme only to the objective, so that the resulting sequence is stronger than those generated by the SCP methods in [16, 29]—in particular, it satisfies the constraints more strictly. Specifically, at each iteration, the method performs multiple constrained

proximal gradient steps using the surrogate constraint mapping $\bar{c}$, along with a backtracking line search to ensure sufficient reduction in h. The method terminates once a practical stopping criterion—designed to guarantee that $\mathrm{dist}(0, \partial\bar{h}(z^{k+1}, z^k))$ is sufficiently small—is met for some z^k and z^{k+1}. The proposed method is described in Algorithm 1, where $\bar{c}$ is defined in (11).

Algorithm 1 A sequential convex programming method for problem (24)

Input: $\{L_{c_j}\}_{j=1}^m$ from Assumption 1; $\underline{L} > 0$, $\rho > 1$, $\beta > 0$, $\tau > 0$, and a point $z^0 \in \{z : h(z) \le h^* + \eta,\ c(z) \le 0\}$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Set $L_{k,i} = \underline{L}\rho^i$.
4: Compute

$$z^{k+1,i} = \arg\min_z \Big\{ \langle \nabla g(z^k), z\rangle + \frac{L_{k,i}}{2}\|z - z^k\|^2 + q(z) \Big\} \quad \text{s.t.} \quad \bar{c}(z, z^k) \le 0, \tag{27}$$

 and its optimal Lagrange multiplier $\lambda^{k,i}$.
5: **if** $h(z^{k+1,i}) \le h(z^k) - \beta\|z^{k+1,i} - z^k\|^2/2$ **then**
6: Set $z^{k+1} = z^{k+1,i}$, $L_k = L_{k,i}$, $\lambda^k = \lambda^{k,i}$.
7: **break**
8: **end if**
9: **end for**
10: Terminate the algorithm and output z^{k+1} if

$$\|\nabla g(z^{k+1}) - \nabla g(z^k) - L_k(z^{k+1} - z^k)\|^2 + 4\Big(\sum_{j=1}^m \lambda_j^k L_{c_j}\Big)^2 \|z^{k+1} - z^k\|^2 \le \tau^2. \tag{28}$$

11: **end for**

We now make some remarks on subproblem (27) in Algorithm 1. By rearranging the terms in the constraint functions of (27), one can see that the constraint is equivalent to

$$z \in \bigcap_{j=1}^m \mathcal{B}\big(s^{k,j}, \sqrt{R_{k,j}}\big),$$

where

$$s^{k,j} = z^k - \nabla c_j(z^k)/L_{c_j}, \quad R_{k,j} = \|\nabla c_j(z^k)\|^2/L_{c_j}^2 - 2c_j(z^k)/L_{c_j}.$$

Thus, the constraint in (27) corresponds to the intersection of Euclidean balls. In addition, since the function q is a simple closed convex function, subproblem (27) can typically be reformulated as a convex conic optimization problem, and a customized primal-dual interior-point method (IPM) can be applied to compute its optimal solution and the associated Lagrange multipliers. Moreover, as the Hessians of the objective and the constraint functions in (27) are multiples of the identity matrix, the Newton system arising in the IPM can be solved cheaply when the epigraph of q is polyhedral (e.g., $q(\cdot) = \|\cdot\|_1$).

The following theorem establishes the well-definedness of Algorithm 1 and several key relations used in the subsequent analysis. Similar results were obtained in [29, Lemma 2.4] for a more sophisticated SCP method. Since our SCP method is simpler, we provide a more concise proof in Subsection 6.2 for reference.

Theorem 4. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{L_{c_j}\}_{j=1}^m, L, \underline{L}, \rho, \beta$ be given in Assumptions 1 and 3, and Algorithm 1, respectively, $\bar{i} = \lceil \log_\rho((\beta + L)/(2\underline{L}))\rceil_+$, and $\{L_k\}$, $\{z^k\}$, and $\{\lambda^k\}$ be generated in Algorithm 1. Then the following statements hold.*

(i) The subproblem (27) *has an optimal solution* $z^{k+1,i}$ *and an optimal Lagrange multiplier* $\lambda^{k,i}$*, and the inner loop terminates in at most* $\bar{i}+1$ *iterations and outputs a point* $z^{k+1} \in \operatorname{dom} q$ *satisfying* $c(z^{k+1}) \leq 0$ *at each outer iteration* k*.*

(ii) For each k*, it holds that*

$$\underline{L} \leq L_k \leq \bar{L} := \max\{\underline{L}, (\beta+L)\rho/2\}, \tag{29}$$

$$\lambda^k \geq 0, \quad \lambda_j^k\Big(c_j(z^k) + \langle \nabla c_j(z^k), z^{k+1}-z^k\rangle + \frac{L_{c_j}}{2}\|z^{k+1}-z^k\|^2\Big) = 0 \quad \forall j \in \{1,...,m\}, \tag{30}$$

$$0 \in \nabla g(z^k) + \Big(L_k + \sum_{j=1}^m \lambda_j^k L_{c_j}\Big)(z^{k+1}-z^k) + \partial q(z^{k+1}) + \sum_{j=1}^m \lambda_j^k \nabla c_j(z^k). \tag{31}$$

To establish the convergence rate of Algorithm 1, we now present a result that provides an upper bound on the sequence of Lagrange multipliers $\{\lambda^k\}$ generated by Algorithm 1. Its proof is deferred to Subsection 6.2.

Lemma 1. *Suppose that Assumptions 1, 2, and 3 hold. Let* λ^k *be generated in the* k*th outer iteration of Algorithm 1 for some* $k \geq 0$*,* ζ *be given in Theorem 3,* $D_{\mathcal{Y}}$*,* M_q*,* $\bar{L}$ *be given in* (7) *and* (29)*, and let* $G = \max_{z\in\mathcal{Y}} \|\nabla g(z)\|$*. Then it holds that*

$$\|\lambda^k\|_1 \leq A := \zeta^{-1}(GD_{\mathcal{Y}} + \bar{L}D_{\mathcal{Y}}^2/2 + M_q). \tag{32}$$

The theorem below shows that Algorithm 1 terminates in a finite number of iterations and outputs a desired approximate solution to problem (24). Its proof is deferred to Subsection 6.2.

Theorem 5. *Suppose that Assumptions 1, 2, and 3 hold. Let* $\{L_{c_j}\}_{j=1}^m, L, C, \theta, \eta, \bar{L}, A, \beta, \tau$ *be given in Assumptions 1 and 3,* (29)*,* (32)*, and Algorithm 1, respectively, and let*

$$\omega = \Big((L+\bar{L})^2 + 4A^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)^{\frac{1}{2}}, \tag{33}$$

$$\alpha = \frac{\beta C^2}{2\omega^2}, \qquad C' = \min\left\{\frac{\alpha}{2}, \frac{(2^{\frac{2\theta-1}{2\theta}}-1)\eta^{1-2\theta}}{2\theta-1}\right\}, \tag{34}$$

$$\overline{K}_\theta := \begin{cases} \Big\lceil \log_{1+\alpha}\big(\frac{2\omega^2\eta}{\beta\tau^2}\big)\Big\rceil_+ + 1 & \text{if } \theta = \frac{1}{2}, \\ \Big\lceil \frac{1}{C'(2\theta-1)}\big(\frac{2\omega^2}{\beta\tau^2}\big)^{2\theta-1}\Big\rceil + 1 & \text{if } \theta \in (\frac{1}{2}, 1). \end{cases} \tag{35}$$

Then Algorithm 1 terminates in at most $\overline{K}_\theta$ *outer iterations and outputs a point* z^{k+1} *satisfying*

$$h(z^{k+1}) - h^* \leq (C^{-1}\tau)^{\frac{1}{\theta}} \tag{36}$$

for some $k < \overline{K}_\theta$*.*

4 An inexact proximal gradient method for problem (1)

In this section, we propose an inexact proximal gradient method for solving problem (1) and analyze its complexity for finding an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$-stationary point of (1) for $\epsilon > 0$.

Before proceeding, we introduce some additional notation below. Given any $\epsilon > 0$, let

$$\mathcal{X}_\epsilon := \{x \in \mathcal{X} : \operatorname{dist}(0, \partial\Psi(x)) \leq \epsilon\}, \quad \mathcal{X}_\epsilon^{\mathrm{c}} := \{x \in \mathcal{X} : \operatorname{dist}(x, \mathcal{X}_\epsilon) > \gamma\epsilon^\sigma/(4L_f)\}, \quad r := \gamma\epsilon^\sigma/(4L_f) \tag{37}$$

where γ, σ, L_f are given in Assumption 1.

To propose a method for finding an (r,ϵ)-stationary point of problem (1), we first make several key observations. Suppose $x' \in \mathcal{X}_\epsilon^{\rm c}$, that is, x' is not an (r,ϵ)-stationary point of (1). For any $x \in \mathcal{X} \cap \mathcal{B}(x',r)$, we observe that $[x',x] \subseteq \mathcal{X}$ and moreover $\mathrm{dist}(0,\partial\Psi(z)) > \epsilon$ for all $z \in [x',x]$. In view of these observations and the definition of $\mathcal{U}_\epsilon$ in (17), it follows that $[x',x] \subseteq \mathcal{U}_\epsilon$. Hence, by Corollary 1, the relation (18) holds for such x and x'. In addition, note from $\theta \in [1/2,1)$ and (19) that $\nu \in (0,1]$. By this relation and [19, (2.15)] with M_ν and t replaced by M and $\|x-x'\|$, respectively, one has

$$M(1+\nu)^{-1}\|x-x'\|^{1+\nu} \le \frac{1}{2}\left[\frac{1-\nu}{1+\nu}\cdot\frac{1}{\delta}\right]^{\frac{1-\nu}{1+\nu}} M^{\frac{2}{1+\nu}}\|x-x'\|^2+\frac{\delta}{2} \le \frac{1}{2}\Big(\delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}\|x-x'\|^2+\delta\Big) \qquad \forall \delta>0,$$

where the first relation follows from Young's inequality (see [19, (2.15)]), and the second relation is due to $\nu \in (0,1]$. Combining this inequality with (18), we obtain that

$$F^*(x) \le F^*(x') + \langle \nabla_{\mathcal{X}}^{\rm C} F^*(x'), x-x'\rangle + \frac{1}{2}\big(L_{\nabla f}+\delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}\big)\|x-x'\|^2 + \frac{\delta}{2} \qquad \forall x \in \mathcal{X}\cap\mathcal{B}(x',r). \tag{38}$$

By this relation and $\Psi(\cdot) = F^*(\cdot)+p(\cdot)$, we further have

$$\Psi(x) \le F^*(x') + \langle \nabla_{\mathcal{X}}^{\rm C} F^*(x'), x-x'\rangle + \frac{1}{2}\big(L_{\nabla f}+\delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}\big)\|x-x'\|^2 + p(x) + \frac{\delta}{2} \qquad \forall x \in \mathcal{X}\cap\mathcal{B}(x',r).$$

Therefore, when $x' \in \mathcal{X}$ is not an (r,ϵ)-stationary point of (1), Ψ is bounded above by a much simpler function that is the sum of a simple quadratic function and $p(\cdot)$ in a neighborhood of x'.

Based on the above observation, it is natural to propose a proximal-gradient-type method for finding an (r,ϵ)-stationary point of problem (1). Specifically, the method generates the sequence $\{x^k\}$ according to

$$x^{k+1} = \mathop{\arg\min}_{x\in\mathcal{B}(x^k,r)} \Big\{ \langle \nabla_{\mathcal{X}}^{\rm C} F^*(x^k), x\rangle + \frac{1}{2}\bar{L}_k\|x-x^k\|^2 + p(x)\Big\}$$

with $\bar{L}_k = L_{\nabla f} + \delta_k^{(\nu-1)/(1+\nu)} M^{2/(1+\nu)}$ for a suitable choice of $\delta_k > 0$. However, since F^* is a maximal function, the exact value of $\nabla_{\mathcal{X}}^{\rm C} F^*(x^k)$ is generally unavailable. To overcome this difficulty, we approximate $\nabla_{\mathcal{X}}^{\rm C} F^*(x^k)$ by $\nabla_x f(x^k,y^k)$, where y^k is a suitably chosen approximate solution of the kth subproblem

$$\min_y \{-f(x^k,y)+q(y) : c(y) \le 0\}. \tag{39}$$

Such y^k is obtained using Algorithm 1, initialized from y^{k-1} (see line 5 of Algorithm 2). We show that if y^0 is a suitable approximate solution to the initial subproblem and $\{x^\ell\}_{0\le\ell<k}$ are not $(\gamma\epsilon^\sigma/(4L_f),\epsilon)$-stationary points of (1), then y^k generated in this manner is indeed a desired approximate solution to (39) (see Lemma 10).

We are now ready to present an inexact proximal gradient method for solving problem (1).

Algorithm 2 An inexact proximal gradient method for problem (1)

Input: L_f, $L_{\nabla f}$, $\{L_{c_j}\}_{j=1}^m$, C, θ, γ, σ from Assumption 1; $\epsilon>0$, $\underline{L}>0$, $\rho>1$, $\beta>0$, and initial point $(x^0,y^0) \in \mathcal{X}\times\mathcal{Y}$ satisfying $F^*(x^0)-F(x^0,y^0) \le \min\{\gamma\epsilon^\sigma/2,1\}$.

1: Set $r=\gamma\epsilon^\sigma/(4L_f)$, $M=(1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}$, $\nu=\theta^{-1}(1-\theta)$.

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

3: Set $\delta_k = 1/(k+1)$, $\eta_k = 1/(k+1)$, $\bar{L}_k = L_{\nabla f} + \delta_k^{(\nu-1)/(1+\nu)} M^{2/(1+\nu)}$.

4: Compute

$$x^{k+1} = \mathop{\arg\min}_{x\in\mathcal{B}(x^k,r)} \Big\{ \langle \nabla_x f(x^k,y^k), x\rangle + \frac{\bar{L}_k}{2}\|x-x^k\|^2 + p(x)\Big\}.$$

5: Call Algorithm 1 with $g(\cdot) \leftarrow -f(x^{k+1},\cdot)$, $q(\cdot)\leftarrow q(\cdot)$, $c(\cdot)\leftarrow c(\cdot)$, $z^0 \leftarrow y^k$, $\underline{L}\leftarrow\underline{L}$, $\rho\leftarrow\rho$, $\beta\leftarrow\beta$, $\{L_{c_j}\}_{j=1}^m \leftarrow \{L_{c_j}\}_{j=1}^m$, $\tau \leftarrow C\min\Big\{(\frac{1}{2}\gamma\epsilon^\sigma)^\theta, \eta_{k+1}^{\frac{\theta}{2(1-\theta)}}\Big\}$, and denote its output as y^{k+1}.

6: **end for**

Remark 2. *(i) The required y^0 for Algorithm 2 can be obtained by applying Algorithm 1 to the problem $\max_y F(x^0, y)$, provided that an initial point in a region where the local KL condition holds is readily available. Alternatively, such a y^0 can be efficiently computed when $F(x^0, \cdot)$ is concave. Moreover, even when $F(x^0, \cdot)$ is nonconcave, the problem corresponding to the specific choice of x^0 may still possess a favorable structure, making it computationally tractable to compute such a y^0 by exploiting this structure.*

(ii) Some input parameters required by Algorithm 2 may not be readily available in practice. It would therefore be worthwhile to develop a parameter-free variant of Algorithm 2. Alternatively, in practical implementations, one may run the algorithm with a range of trial parameters and adjust them until its performance stabilizes.

To analyze the complexity of Algorithm 2 for computing an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$-stationary point of problem (1), it is necessary to establish that the sequence of Lagrange multipliers generated by Algorithm 1 for solving subproblem (39) remains uniformly bounded, independent of k.

Lemma 2. *Suppose that Assumption 1 holds. Let $\{\lambda^{k,\ell}\}$ denote the sequence of Lagrange multipliers generated by Algorithm 1 during the kth iteration of Algorithm 2, and let ζ be given in Theorem 3, $\bar{L}$ be given in (29), $D_{\mathcal{Y}}, M_q$ be defined in (7), and*

$$G_f = \max\{\|\nabla_y f(x,y)\| : (x,y) \in \mathcal{X} \times \mathcal{Y}\}.$$

Suppose that $G_f < \infty$. Then it holds that

$$\|\lambda^{k,\ell}\|_1 \le A_f := \zeta^{-1}(G_f D_{\mathcal{Y}} + \bar{L} D_{\mathcal{Y}}^2/2 + M_q) \qquad \forall k, \ell. \tag{40}$$

The following theorem establishes an *iteration complexity* bound for Algorithm 2 to compute an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$-stationary point of problem (1) for any $\epsilon \in (0, 1/e]$. The proof is deferred to Subsection 6.3.

Theorem 6. *Let $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma$ be given in Assumption 1, M, ν be defined in (19), ϵ be given in Algorithm 2, and*

$$\begin{aligned}
&\widehat{A} = (1-\theta)^{-2} C^{-2} L_{\nabla f}^2, \quad \widehat{L} = L_{\nabla f} + M^{2/(1+\nu)},\\
&a = 8\big(\Psi(x^0) - \Psi^* + 3 + 2\widehat{L}^{-1}\widehat{A}\big), \quad b = 8\big(3/2 + \widehat{L}^{-1}\widehat{A}\big),\\
&\widehat{C}_1 = \Big(36(1+\nu)\nu^{-1} b\widehat{L}\lceil \log(18(1+\nu)\nu^{-1} b\widehat{L})\rceil_+ + 72(1+\nu)\nu^{-1} b\widehat{L} + 1\Big)^{\frac{1+\nu}{2\nu}},\\
&\widehat{C}_2 = \Big(\frac{4b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}}\Big\lceil \log\Big(\frac{2b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{8b(1+\nu)(3M)^{2/\nu}}{\nu M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}},\\
&\widehat{C}_3 = \big(36\widehat{L}a\big)^{\frac{1+\nu}{2\nu}} + M^{-1}\big(4a(3M)^{2/\nu}\big)^{\frac{1+\nu}{2}}, \quad \widehat{C}_4 = 72\widehat{A},\\
&\widehat{C}_5 = \Big(\frac{144(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}}\Big\lceil \log\Big(\frac{72(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{288(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}},\\
&\widehat{C}_6 = (144aL_{\nabla f}^2)^{\frac{1+\nu}{2}}/M,\\
&\widehat{C}_7 = \Big(\frac{64(1+\nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}}\Big\lceil \log\Big(\frac{32(1+\nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{128\sigma(1+\nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}},\\
&\widehat{C}_8 = (64aL_f^2)^{\frac{1+\nu}{2}}/(\gamma^{1+\nu}M),\\
&\widehat{K}_\epsilon = \Big\lceil \widehat{C}_1\epsilon^{-\frac{1+\nu}{\nu}}(\log\epsilon^{-1})^{\frac{1+\nu}{2\nu}} + \widehat{C}_2\epsilon^{-\frac{1+\nu}{\nu}}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_3\epsilon^{-\frac{1+\nu}{\nu}} + \widehat{C}_4\epsilon^{-2}\\
&\qquad + \widehat{C}_5\epsilon^{-(1+\nu)}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_6\epsilon^{-(1+\nu)} + \widehat{C}_7\epsilon^{-(1+\nu)\sigma}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_8\epsilon^{-(1+\nu)\sigma}\Big\rceil.
\end{aligned}$$

Suppose that $\epsilon \in (0, 1/e]$ and Assumptions 1 and 2 hold. Then Algorithm 2 generates a pair (x^k, y^k) in at most $\widehat{K}_\epsilon$ iterations such that x^k is an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$-stationary point of problem (1) (or equivalently

the problem $\min_x \Psi(x)$*), and* y^k *satisfies*

$$F^*(x^k)-F(x^k,y^k) \le \min\Big\{\frac{\gamma\epsilon^\sigma}{2}, \frac{1}{k+1}\Big\}, \quad \operatorname{dist}\big(y^k, Y^*(x^k)\big) \le \frac{1}{C(1-\theta)} \min\Big\{\Big(\frac{\gamma}{2}\Big)^{(1-\theta)}\epsilon^{\sigma(1-\theta)}, \frac{1}{\sqrt{k+1}}\Big\}. \tag{41}$$

The next result presents a *first-order oracle complexity* bound for Algorithm 2, measured by the number of evaluations of the gradient ∇f, required to generate an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$-stationary point of problem (1) for any $\epsilon \in (0, 1/e]$. The proof is deferred to Subsection 6.3.

Theorem 7. *Let* $\epsilon \in (0, 1/e]$ *be given,* $\widehat{K}_\epsilon$ *be defined in Theorem 6,* $L_{\nabla f}, C, \theta, \gamma, \sigma, \{L_{c_j}\}_{j=1}^m$ *be given in Assumption 1,* M, ν *be defined in* (19)*,* $\underline{L}, \beta, \rho$ *be given in Algorithm 2,* A_f *be given in* (40)*, and let*

$$\bar{L}_{\nabla f} = \max\Big\{\underline{L}, \frac{(\beta+L_{\nabla f})\rho}{2}\Big\}, \quad \omega_f = \Big((L_{\nabla f}+\bar{L}_{\nabla f})^2+4A_f^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)^{\frac{1}{2}}, \quad \alpha_f = \frac{\beta C^2}{2\omega_f^2},$$

$$C_f' = \min\Big\{\frac{1}{2}\alpha_f, \frac{(2^{\frac{2\theta-1}{2\theta}}-1)(\gamma\epsilon^\sigma)^{1-2\theta}}{2\theta-1}\Big\}, \quad \Lambda = \max\Big\{(\tfrac{1}{2}\gamma\epsilon^\sigma)^{-2\theta}, (\widehat{K}_\epsilon+1)^{\frac{\theta}{1-\theta}}\Big\},$$

$$\overline{K}_{f,\theta} = \begin{cases} \Big\lceil \log_{1+\alpha_f}(2\omega_f^2\beta^{-1}C^{-2}\gamma\epsilon^\sigma\Lambda)\Big\rceil_+ + 1 & \text{if } \theta = \frac{1}{2}, \\ \Big\lceil \frac{1}{C_f'(2\theta-1)}\Big(2\omega_f^2\beta^{-1}C^{-2}\Lambda\Big)^{2\theta-1}\Big\rceil + 1 & \text{if } \theta \in (\frac{1}{2}, 1), \end{cases}$$

$$\widehat{N}_\epsilon = \widehat{K}_\epsilon\Big(\Big\lceil \log_\rho\big(\frac{\beta+L_{\nabla f}}{2\underline{L}}\big)\Big\rceil_+ + 1\Big)\overline{K}_{f,\theta}.$$

Suppose that Assumptions 1 and 2 hold. Then the total number of evaluations of the proximal operators of p *and* q*, and the gradient* ∇f *performed by Algorithm 2 is at most* $\widehat{K}_\epsilon$*,* $\widehat{N}_\epsilon$*, and* $\widehat{K}_\epsilon+\widehat{N}_\epsilon$*, respectively, to generate a pair* (x^k, y^k) *such that* x^k *is an* $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$*-stationary point of problem* (1)*, and* y^k *satisfies* (41)*.*

Remark 3. *As shown in Theorem 6, Algorithm 2 achieves an iteration complexity of*

$$\mathcal{O}\left(\epsilon^{-\max\{\frac{1}{1-\theta}, \frac{\sigma}{\theta}\}}(\log\epsilon^{-1})^{\frac{1}{2(1-\theta)}}\right)$$

to compute an $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$*-stationary point of problem* (1)*. Furthermore, as established in Theorem 7, the algorithm requires* $\mathcal{O}\Big(\epsilon^{-\max\{\frac{1}{1-\theta}, \frac{\sigma}{\theta}\}}(\log\epsilon^{-1})^{\frac{1}{2(1-\theta)}}\Big)$ *evaluations of the proximal operator of* p*, and the following number of evaluations of the proximal operator of* q *and the gradient* ∇f *to compute such an approximate stationary point of* (1)*:*

$$\begin{cases} \mathcal{O}\left(\epsilon^{-2\max\{1,\sigma\}}(\log\epsilon^{-1})^2\right) & \text{if } \theta = \frac{1}{2}, \\ \mathcal{O}\left(\epsilon^{-\frac{2\theta^2-2\theta+1}{1-\theta}\max\{\frac{1}{1-\theta}, \frac{\sigma}{\theta}\}}(\log\epsilon^{-1})^{\frac{2\theta^2-2\theta+1}{2(1-\theta)^2}}\right) & \text{if } \theta \in (\frac{1}{2}, 1). \end{cases}$$

5 Numerical results

In this section, we conduct preliminary experiments to evaluate the performance of our proposed method (Algorithm 2).

Consider the following constrained minimax optimization problem:

$$\min_x \max_{c(y)\le 0} \big\{-\|(y+Ax)\odot(y+Bx)\|^2+0.01\,\|x-u\|^2+0.01\|x\|_1+\delta_{\mathcal{B}(0,2)}(x)-0.1\|y\|_1-\delta_{[-2,2]^{n_2}}(y)\big\}, \tag{42}$$

where $A, B \in \mathbb{R}^{n_2 \times n_1}$, $u \in \mathbb{R}^{n_1}$, and $\odot$ denotes the Hadamard (elementwise) product. The mapping c is defined as follows. Assuming n_2 is a multiple of 10, we set the number of constraints as $m = n_2/10$, and for each $j \in \{1, ..., m\}$, the jth component of c is

$$c_j(y) = e^{y_{10j-9}} + e^{y_{10j-8}} + \cdots + e^{y_{10j}} - 10. \tag{43}$$

For each pair (n_1, n_2), we randomly generate 5 instances of problem (42) by sampling the entries of A, B, and u independently from the standard normal distribution $\mathcal{N}(0, 1)$. Note that problem (42) is a special case of problem (1) with $f(x, y) = -\|(y + Ax) \odot (y + Bx)\|^2 + 0.01\|x - u\|^2$, $p(x) = 0.01\|x\|_1 + \delta_{\mathcal{B}(0,2)}(x)$, $q(y) = 0.1\|y\|_1 + \delta_{[-2,2]^{n_2}}(y)$, and $c = (c_1, \ldots, c_m)$ defined in (43).

We now apply Algorithm 2 to solve problem (42) on the randomly generated instances described above. Notice that problem (42) is similar to the one studied in [17, Section 5], except that the constraint $c(y) \leq 0$ is imposed on the inner maximization problem. Consequently, the Lipschitz constant L_f of $f(\cdot, y)$ and the Lipschitz constant $L_{\nabla f}$ of ∇f over $\mathcal{B}(0, 2) \times [-2, 2]^{n_2}$ are computed as in [17, Section 5]. In addition, one can verify that c_j is L_{c_j}-smooth over $[-2, 2]^{n_2}$ with $L_{c_j} = e^2$ for all $j \in \{1, ..., m\}$. The remaining input parameters for Algorithm 2 are set as $C = 0.1$, $\theta = 0.5$, $\gamma = 0.01$, $\sigma = 0.1$, $\underline{L} = 1$, $\rho = 1.25$, $\beta = 10$, $\epsilon = 10^{-2}$.[2] The algorithm is initialized at $(x^0, y^0) = (0, 0)$. Note that for this choice of (x^0, y^0), y^0 is the maximizer of the problem $\max_{c(y)\leq 0}\{f(x^0, y) - 0.1\|y\|_1 - \delta_{[-2,2]^{n_2}}(y)\}$, making it a suitable starting point for y. We run the algorithm for 2,500 iterations and return the final output denoted by (x_ϵ, y_ϵ). Here, x_ϵ serves as an approximate solution to the outer minimization problem of (42), while y_ϵ is an approximate solution to the inner maximization problem $\max_{c(y)\leq 0}\{f(x_\epsilon, y) - 0.1\|y\|_1 - \delta_{[-2,2]^{n_2}}(y)\}$.

To evaluate the performance of Algorithm 2, we compute the actual final objective value

$$\Psi(x_\epsilon) = \max_{c(y)\leq 0}\{f(x_\epsilon, y) - 0.1\|y\|_1 - \delta_{[-2,2]^{n_2}}(y)\} + 0.01\|x_\epsilon\|_1.$$

Thanks to the block-separable structure of the problem, this maximization problem can be decomposed into m independent subproblems, each involving a single component of c and 10 components of y. These subproblems are solved using the `MATLAB` subroutine `GlobalSearch`, which is a solver for finding global optima of nonconvex problems. In addition, we compute an approximate final objective value by

$$\widehat{\Psi}(x_\epsilon) = f(x_\epsilon, y_\epsilon) - 0.1\|y_\epsilon\|_1 + 0.01\|x_\epsilon\|_1,$$

using the approximate inner solution y_ϵ returned by the algorithm.

The computational results on the random instances are presented in Table 1. The first two columns list the values of n_1 and n_2. For each pair (n_1, n_2), the average initial, actual final, and approximate final objective values over five random instances are reported in the remaining columns. From the results, we observe that the approximate solution x_ϵ significantly reduces the objective value compared to the initial point x^0, and that y_ϵ is a good approximate solution to the inner maximization problem $\max_{c(y)\leq 0}\{f(x_\epsilon, y) - 0.1\|y\|_1 - \delta_{[-2,2]^{n_2}}(y)\}$. In addition, for the five random instances with $(n_1, n_2) = (100, 100)$, we plot the average actual objective value in Figure 1 to illustrate the performance of Algorithm 2. As observed, the average objective value decreases rapidly in the early stage and then stabilizes, which illustrates the convergence behavior of the proposed method.

6 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2, 3, and 4, which are particularly Lemma 1 and Theorems 1, 2, 4, 5, 6, and 7.

[2] In our numerical experiments, we tested several values of $\theta \in [1/2, 1)$ and observed that $\theta = 1/2$ yields the largest reduction in the objective value. The remaining parameters were chosen empirically and perform reasonably well across the tested instances.

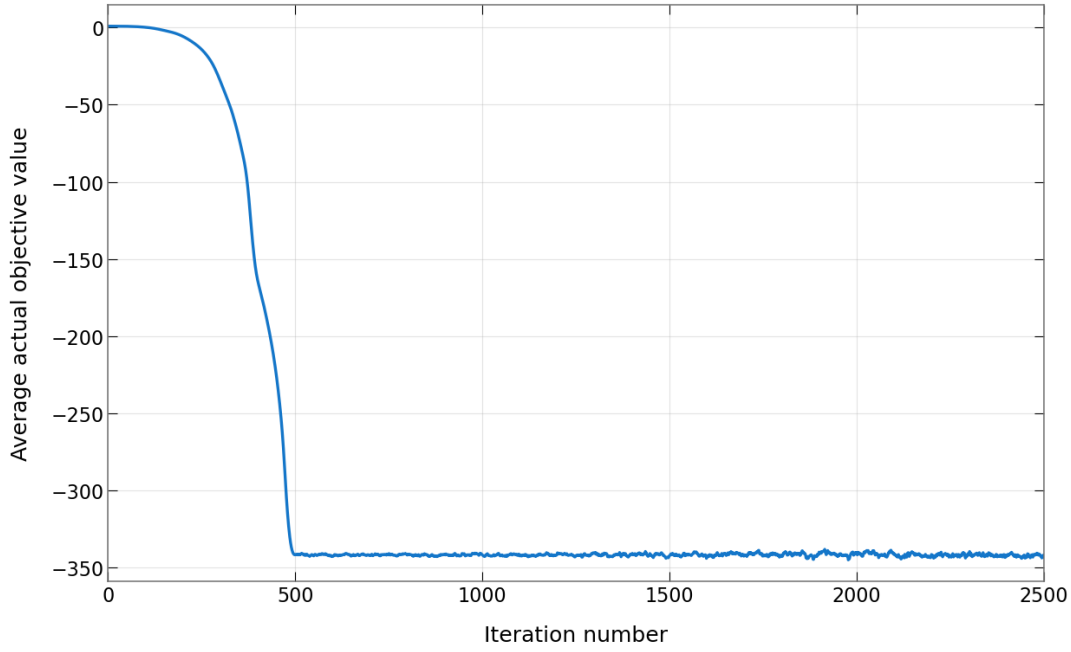

Figure 1: Performance of Algorithm 2 with $(n_1, n_2) = (100, 100)$

Table 1: Numerical results for Algorithm 2

n_1	n_2	Initial objective value	Actual final value	Approximate final value
50	50	0.49	-164.92	-165.09
60	60	0.61	-279.51	-279.53
70	70	0.78	-241.86	-241.91
80	80	0.75	-298.59	-298.76
90	90	0.81	-260.97	-261.38
100	100	1.02	-341.73	-341.88
110	110	1.14	-404.36	-404.43
120	120	1.16	-482.48	-482.50
130	130	1.17	-467.62	-467.75
140	140	1.39	-590.08	-590.23
150	150	1.42	-579.71	-579.88

6.1 Proof of the main results in Section 2

In this subsection, we prove Theorems 1 and 2. To this end, we first present several technical lemmas. The following lemma establishes the equivalence between $\max_y F(x, y)$ and $\max_{y,z} \bar{F}(x, y, z)$ for any $x \in \mathcal{X}$, and between $\min_x \Psi(x)$ and $\min_x \bar{\Psi}(x)$.

Lemma 3. *Let $F^*, Y^*, \Psi, \bar{F}^*, \overline{Y}^*, \bar{\Psi}$ be defined in (8), (9), and (13). Suppose that Assumption 1 holds. Then for any $x \in \mathcal{X}$, the following statements hold.*

(i) If $y^* \in Y^*(x)$, then $(y^*, y^*) \in \overline{Y}^*(x)$.

(ii) $F^*(x) = \bar{F}^*(x)$ and $\Psi(x) = \bar{\Psi}(x)$.

(iii) If $(y^*, z^*) \in \overline{Y}^*(x)$, then $y^* \in Y^*(x)$.

Proof. Fix any $x \in \mathcal{X}$. For notational convenience, let $\tilde{F}(x, y) = f(x, y) - q(y)$. It then follows from the definitions of F and $\bar{F}$ in (8) and (12) that

$$F(x, y) = \tilde{F}(x, y) - \delta_{c(\cdot)\le 0}(y), \qquad \bar{F}(x, y, z) = \tilde{F}(x, y) - \delta_{\bar{c}(\cdot,\cdot)\le 0}(y, z) - \delta_{\mathcal{Y}}(z). \tag{44}$$

We first prove statement (i). Fix any $y^* \in Y^*(x)$. Clearly, $y^* \in \mathcal{Y}$, $c(y^*) \le 0$, and

$$\tilde{F}(x,y) \le \tilde{F}(x,y^*) \quad \forall y \in \mathcal{Y} \text{ with } c(y) \le 0. \tag{45}$$

Morever, by (11) and $c(y^*) \le 0$, we observe that $\bar{c}(y^*,y^*) = c(y^*) \le 0$. Let $(y',z') \in \mathbb{R}^{n_2} \times \mathbb{R}^{n_2}$ be arbitrarily chosen. We claim that $\bar{F}(x,y',z') \le \bar{F}(x,y^*,y^*)$. Indeed, if $\bar{c}(y',z') > 0$ or $(y',z') \notin \mathcal{Y} \times \mathcal{Y}$, then $\bar{F}(x,y',z') = -\infty$, and the claim holds trivially. Now suppose that $\bar{c}(y',z') \le 0$ and $(y',z') \in \mathcal{Y} \times \mathcal{Y}$. By these, (11), and the L_{c_i}-Lipschitz smoothness of each component c_i over $\mathcal{Y}$ (see Assumption 1), we deduce that $c(y') \le \bar{c}(y',z') \le 0$. This along with $y' \in \mathcal{Y}$ and (45) yields $\tilde{F}(x,y') \le \tilde{F}(x,y^*)$. It then follows from (44), $\bar{c}(y',z') \le 0$, $\bar{c}(y^*,y^*) \le 0$, and $y^*, z' \in \mathcal{Y}$ that

$$\bar{F}(x,y',z') = \tilde{F}(x,y') \le \tilde{F}(x,y^*) = \bar{F}(x,y^*,y^*),$$

and the above claim again holds. By $\bar{F}(x,y',z') \le \bar{F}(x,y^*,y^*)$ and the arbitrariness of (y',z'), we conclude that $(y^*,y^*) \in \overline{Y}^*(x)$. Hence, statement (i) holds.

We next prove statement (ii). By Assumption 1, there exists at least one $\hat{y}^* \in Y^*(x)$, and moreover, $F(x,\hat{y}^*)$ is finite. Using this and statement (i), we see that $(\hat{y}^*,\hat{y}^*) \in \overline{Y}^*(x)$. Notice that $\hat{y}^* \in \mathcal{Y}$ and $\bar{c}(\hat{y}^*,\hat{y}^*) = c(\hat{y}^*) \le 0$. By these, (8), (13), and (44), we obtain that

$$F^*(x) = F(x,\hat{y}^*) = \tilde{F}(x,\hat{y}^*) = \bar{F}(x,\hat{y}^*,\hat{y}^*) = \bar{F}^*(x).$$

It follows from this, (9), and (13) that $\Psi(x) = \bar{\Psi}(x)$ holds. This proves statement (ii).

We finally prove statement (iii). Fix any $(y^*,z^*) \in \overline{Y}^*(x)$. By this, (12), (13), and (44), we observe that $(y^*,z^*) \in \mathcal{Y} \times \mathcal{Y}$, $\bar{c}(y^*,z^*) \le 0$, and $\bar{F}^*(x) = \bar{F}(x,y^*,z^*) = \tilde{F}(x,y^*)$. Using these relations, (11), and the L_{c_i}-Lipschitz smoothness of each component c_i over $\mathcal{Y}$, we deduce that $c(y^*) \le \bar{c}(y^*,z^*) \le 0$. By this, (44), and $y^* \in \mathcal{Y}$, one has $F(x,y^*) = \tilde{F}(x,y^*)$, which together with $\bar{F}^*(x) = \tilde{F}(x,y^*)$ and $F^*(x) = \bar{F}^*(x)$ (see statement (ii)) implies that $F(x,y^*) = F^*(x)$. Hence, $y^* \in Y^*(x)$ and statement (iii) holds. □

Remark 4. *In view of* (8), (13), *and Lemma 3(ii), we observe that*

$$\max_y F(x,y) = \max_{y,z} \bar{F}(x,y,z) \qquad \forall x \in \mathcal{X}.$$

Consequently, when interpreted as minimization problems, $\min_x\{\max_y F(x,y) + p(x)\}$ *and* $\min_x\{\max_{y,z} \bar{F}(x,y,z) + p(x)\}$ *have identical objective functions and are thus equivalent. Moreover, from* (8), (12), *and* (13), *these problems correspond to* (1) *and* (14), *respectively. Therefore, the original minimax problem* (1) *is equivalent to the lifted minimax problem* (14).

The next lemma presents properties of the limiting slope and error bound of a proper closed function. Its proof can be found in [10, Proposition 4.6] and [10, Lemma 2.5].

Lemma 4. *Let* $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a proper closed function, and a point* $\bar{u} \in \operatorname{dom}\phi$ *be given. Then the following statements hold.*

(i) $\overline{|\nabla\phi|}(\bar{u}) = \operatorname{dist}(0, \partial\phi(\bar{u}))$.

(ii) Suppose there exist constants $\alpha < \phi(\bar{u})$ *and* $r, K > 0$ *such that*

$$\alpha < \phi(u) \le \phi(\bar{u}) \ \text{ and } \ \|u - \bar{u}\| \le K \quad \Longrightarrow \quad |\nabla\phi|(u) \ge r.$$

If, in addition, $\phi(\bar{u}) - \alpha < Kr$, *then*

$$\operatorname{dist}(\bar{u}, \mathcal{S}_\alpha) \le r^{-1}(\phi(\bar{u}) - \alpha), \quad \text{where } \ \mathcal{S}_\alpha := \{u : \phi(u) \le \alpha\}.$$

The lemma below establishes a local $(1-\theta)^{-1}$-growth property of $\bar{F}(x,\cdot,\cdot)$ for any $x \in \mathcal{X}$, which can also be obtained from [11, Theorem 3.7]. For completeness, we provide a self-contained proof with minimal reliance on the literature.

Lemma 5. *Suppose that Assumption 1 holds. Then it holds that for any $x \in \mathcal{X}$,*

$$\bar{F}^*(x) - \bar{F}(x,y,z) \geq (C(1-\theta))^{\frac{1}{1-\theta}} \mathrm{dist}((y,z), \overline{Y}^*(x))^{\frac{1}{1-\theta}} \qquad \forall (y,z) \in \overline{\mathcal{L}}(x). \tag{46}$$

Proof. Fix any $x \in \mathcal{X}$ and $(\bar{y},\bar{z}) \in \overline{\mathcal{L}}(x)$. If $(\bar{y},\bar{z}) \notin \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$, then $\bar{F}(x,\bar{y},\bar{z}) = -\infty$, and hence relation (46) holds trivially at $(y,z) = (\bar{y},\bar{z})$. We now suppose that $(\bar{y},\bar{z}) \in \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$. For notational convenience, let

$$\phi_x(y,z) = -\bar{F}(x,y,z), \quad \phi_x^* = -\bar{F}^*(x), \tag{47}$$

$$g(y,z) = (\phi_x(y,z) - \phi_x^*)^{1-\theta}, \quad K_x = 1 + (C(1-\theta))^{-1}(\gamma \mathrm{dist}(0, \partial\bar{\Psi}(x))^\sigma)^{1-\theta}. \tag{48}$$

Then, one can see from (15), (16), and $(\bar{y},\bar{z}) \in \overline{\mathcal{L}}(x)$ that

$$0 < C(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^\theta \leq \mathrm{dist}(0, \partial\phi_x(\bar{y},\bar{z})). \tag{49}$$

By the definitions of slope and limiting slope in (4) and (5), one can observe that $|\nabla\phi_x|(\bar{y},\bar{z}) \geq \overline{|\nabla\phi_x|}(\bar{y},\bar{z})$. Also, it follows from Lemma 4(i) that $\overline{|\nabla\phi_x|}(\bar{y},\bar{z}) = \mathrm{dist}(0, \partial\phi_x(\bar{y},\bar{z}))$. Hence, we obtain that

$$|\nabla\phi_x|(\bar{y},\bar{z}) \geq \mathrm{dist}(0, \partial\phi_x(\bar{y},\bar{z})). \tag{50}$$

In addition, by (48) and the concavity of the function $t^{1-\theta}$ in $[0,\infty)$ due to $\theta \in [1/2,1)$, one has

$$\begin{aligned} g(y,z) &\overset{(48)}{=} (\phi_x(y,z) - \phi_x^*)^{1-\theta} \leq (\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{1-\theta} + (1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta}(\phi_x(y,z) - \phi_x(\bar{y},\bar{z})) \\ &= g(\bar{y},\bar{z}) + (1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta}(\phi_x(y,z) - \phi_x(\bar{y},\bar{z})). \end{aligned}$$

Using this, $\theta \in [1/2,1)$, $\phi_x(\bar{y},\bar{z}) > \phi_x^*$, and the definition of slope in (4), we obtain that

$$\begin{aligned} |\nabla g|(\bar{y},\bar{z}) &\overset{(4)}{=} \limsup_{(y,z)\to(\bar{y},\bar{z})} \frac{\big(g(\bar{y},\bar{z}) - g(y,z)\big)_+}{\|(\bar{y},\bar{z}) - (y,z)\|} \\ &\geq \limsup_{(y,z)\to(\bar{y},\bar{z})} \frac{\big((1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta}(\phi_x(\bar{y},\bar{z}) - \phi_x(y,z))\big)_+}{\|(\bar{y},\bar{z}) - (y,z)\|} \\ &= (1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta} \limsup_{(y,z)\to(\bar{y},\bar{z})} \frac{\big(\phi_x(\bar{y},\bar{z}) - \phi_x(y,z)\big)_+}{\|(\bar{y},\bar{z}) - (y,z)\|} \\ &\overset{(4)}{=} (1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta} |\nabla\phi_x|(\bar{y},\bar{z}) \\ &\overset{(50)}{\geq} (1-\theta)(\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{-\theta} \mathrm{dist}(0, \partial\phi_x(\bar{y},\bar{z})) \overset{(49)}{\geq} C(1-\theta). \end{aligned}$$

By this relation and the arbitrariness of $(\bar{y},\bar{z}) \in \overline{\mathcal{L}}(x) \cap \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$, we conclude that $|\nabla g|(y,z) \geq C(1-\theta)$ holds for all $(y,z) \in \overline{\mathcal{L}}(x) \cap \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$. In addition, by the definitions of g and $\overline{\mathcal{L}}(x)$ along with the fact $(\bar{y},\bar{z}) \in \overline{\mathcal{L}}(x) \cap \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$, one can observe that for any (y,z) with $0 < g(y,z) \leq g(\bar{y},\bar{z})$, we have $(y,z) \in \overline{\mathcal{L}}(x) \cap \mathrm{dom}\, \bar{F}(x,\cdot,\cdot)$, and hence $|\nabla g|(y,z) \geq C(1-\theta)$. Also, notice from (16) and $(\bar{y},\bar{z}) \in \overline{\mathcal{L}}(x)$ that

$$0 < \bar{F}^*(x) - \bar{F}(x,\bar{y},\bar{z}) \leq \gamma\, \mathrm{dist}(0, \partial\bar{\Psi}(x))^\sigma,$$

which along with (47) and (48) implies that

$$\begin{aligned} 0 < g(\bar{y},\bar{z}) &\overset{(48)}{=} (\phi_x(\bar{y},\bar{z}) - \phi_x^*)^{1-\theta} \overset{(47)}{=} (\bar{F}^*(x) - \bar{F}(x,\bar{y},\bar{z}))^{1-\theta} \\ &\leq \big(\gamma\, \mathrm{dist}(0, \partial\bar{\Psi}(x))^\sigma\big)^{1-\theta} \overset{(48)}{<} K_x C(1-\theta). \end{aligned}$$

In view of these, it follows from (47), (48), and Lemma 4 with $\phi = g$, $\bar{u} = (\bar{y},\bar{z})$, $\alpha = 0$, $r = C(1-\theta)$, $K = K_x$, and $\mathcal{S}_\alpha = \overline{Y}^*(x)$ that

$$\mathrm{dist}((\bar{y},\bar{z}), \overline{Y}^*(x)) \overset{\text{Lemma 4}}{\leq} \frac{1}{C(1-\theta)} g(\bar{y},\bar{z}) \overset{(47)(48)}{=} \frac{1}{C(1-\theta)} (\bar{F}^*(x) - \bar{F}(x,\bar{y},\bar{z}))^{1-\theta},$$

which implies that relation (46) holds at $(y,z)=(\bar{y},\bar{z})$. By this and the arbitrariness of $(\bar{y},\bar{z})\in\overline{\mathcal{L}}(x)\cap\operatorname{dom}\bar{F}(x,\cdot,\cdot)$, one can conclude that relation (46) holds for all $(y,z)\in\overline{\mathcal{L}}(x)\cap\operatorname{dom}\bar{F}(x,\cdot,\cdot)$. This completes the proof. □

The following lemma provides a relationship between $\operatorname{dist}\big((y,z),\overline{Y}^*(x)\big)$ and $\operatorname{dist}\big(0,\partial_{(y,z)}\bar{F}(x,y,z)\big)$, following directly from (15) and (46).

Lemma 6. *Suppose that Assumption 1 holds. Then it holds that for any $x\in\mathcal{X}$,*

$$\operatorname{dist}((y,z),\overline{Y}^*(x))\le(1-\theta)^{-1}C^{-\frac{1}{\theta}}\operatorname{dist}\big(0,\partial_{(y,z)}\bar{F}(x,y,z)\big)^{\frac{1-\theta}{\theta}}\qquad\forall(y,z)\in\overline{\mathcal{L}}(x).$$

The lemma below establishes the local generalized Hölder smoothness of $\bar{F}^*$. It is analogous to [17, Theorem 1], whose proof is based on [17, Lemma 4]. Note that [17, Lemma 4] is analogous to Lemma 6. Therefore, the proof of this lemma follows directly from Lemma 6, together with arguments similar to those used in the proof of [17, Theorem 1], and is thus omitted.

Lemma 7. *Let $\epsilon>0$ be given and $\overline{\mathcal{U}}_\epsilon=\{x\in\mathcal{X}:\operatorname{dist}(0,\partial\bar{\Psi}(x))>\epsilon\}$. Suppose that Assumption 1 holds. Then the following statements hold.*

(i) $\partial^{\mathrm{C}}_{\mathcal{X}}\bar{F}^*(x)$ is a singleton for all $x\in\overline{\mathcal{U}}_\epsilon$, and $\bar{F}^*$ is differentiable on $\overline{\mathcal{U}}_\epsilon\cap\operatorname{int}(\mathcal{X})$.

(ii) For any $x,x'\in\overline{\mathcal{U}}_\epsilon\cap\operatorname{int}(\mathcal{X})$ satisfying $\|x-x'\|\le\gamma\epsilon^\sigma/(2L_f)$, we have

$$\|\nabla\bar{F}^*(x)-\nabla\bar{F}^*(x')\|\le L_{\nabla f}\|x-x'\|+(1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}\|x-x'\|^{\frac{1-\theta}{\theta}}.$$

(iii) For any $x,x'\in\overline{\mathcal{U}}_\epsilon$ satisfying $\|x-x'\|\le\gamma\epsilon^\sigma/(4L_f)$, we have

$$\|\nabla^{\mathrm{C}}_{\mathcal{X}}\bar{F}^*(x)-\nabla^{\mathrm{C}}_{\mathcal{X}}\bar{F}^*(x')\|\le L_{\nabla f}\|x-x'\|+(1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}\|x-x'\|^{\frac{1-\theta}{\theta}}.$$

(iv) It holds that

$$\nabla^{\mathrm{C}}_{\mathcal{X}}\bar{F}^*(x)=\nabla_x f(x,y^*)\qquad\forall x\in\overline{\mathcal{U}}_\epsilon,\ y^*\in Y^*(x).$$

We are now ready to prove Theorems 1 and 2.

Proof of Theorem 1. The conclusion of this theorem directly follows from Lemmas 3 and 7. □

Proof of Theorem 2. Fix any $x\in\mathcal{X}$ and $y\in\mathcal{L}(x)$. If $y\notin\operatorname{dom}F(x,\cdot)$, then $F(x,y)=-\infty$, and hence the conclusion holds trivially. We now suppose $y\in\operatorname{dom}F(x,\cdot)$. It follows from this and the definitions of F and $\bar{c}$ in (8) and (11) that $y\in\mathcal{Y}$ and $\bar{c}(y,y)=c(y)\le0$, which implies that $(x,y,y)\in\operatorname{dom}\bar{F}(x,\cdot,\cdot)$ and $F(x,y)=\bar{F}(x,y,y)$. Recall from Lemma 3 that $F^*(x)=\bar{F}^*(x)$. By these, (16), (21), and $y\in\mathcal{L}(x)$, one has $(y,y)\in\overline{\mathcal{L}}(x)$. Using these relations and Lemma 5, we have

$$F^*(x)-F(x,y)=\bar{F}^*(x)-\bar{F}(x,y,y)\ge(C(1-\theta))^{\frac{1}{1-\theta}}\operatorname{dist}((y,y),\overline{Y}^*(x))^{\frac{1}{1-\theta}}.\tag{51}$$

We next show that $\operatorname{dist}(y,Y^*(x))\le\operatorname{dist}((y,y),\overline{Y}^*(x))$. Notice from Assumption 1 and Lemma 3 that $\overline{Y}^*(x)$ is a nonempty closed set. Hence, there exists $(y^*,z^*)\in\overline{Y}^*(x)$ such that $\|(y,y)-(y^*,z^*)\|=\operatorname{dist}((y,y),\overline{Y}^*(x))$. Since $(y^*,z^*)\in\overline{Y}^*(x)$, it follows from Lemma 3 that $y^*\in Y^*(x)$, which implies that $\operatorname{dist}(y,Y^*(x))\le\|y-y^*\|$. In view of these, one has

$$\operatorname{dist}(y,Y^*(x))\le\|y-y^*\|\le\|(y,y)-(y^*,z^*)\|=\operatorname{dist}((y,y),\overline{Y}^*(x)),$$

and hence $\operatorname{dist}(y,Y^*(x))\le\operatorname{dist}((y,y),\overline{Y}^*(x))$ holds as desired. The conclusion (20) directly follows from this and (51). □

6.2 Proof of the main results in Section 3

In this subsection we prove Theorems 4 and 5.

Proof of Theorem 4. We first prove statement (i) of Theorem 4 by induction. Suppose that a point $z^k \in \operatorname{dom} q$ satisfying $c(z^k) \le 0$ is already generated in Algorithm 1 for some $k \ge 0$. Let us fix any $i \ge 0$, and define

$$Q_{k,i}(z) = \langle \nabla g(z^k), z\rangle + \frac{L_{k,i}}{2}\|z - z^k\|^2 + q(z) + \delta_{\bar{c}(\cdot,z^k)\le 0}(z). \tag{52}$$

Since $z^k \in \operatorname{dom} q$ and $\bar{c}(z^k, z^k) = c(z^k) \le 0$, one has $z^k \in \operatorname{dom} Q_{k,i}$. By this and the assumption that q is a closed convex function, one can observe that $Q_{k,i}$ is a proper closed strongly convex function. It follows that the problem $\min_z Q_{k,i}(z)$ has a unique optimal solution. Note that subproblem (27) is equivalent to $z^{k+1,i} = \arg\min_z Q_{k,i}(z)$. Hence, (27) has a unique optimal solution $z^{k+1,i} \in \operatorname{dom} q$ satisfying $\bar{c}(z^{k+1,i}, z^k) \le 0$. By these, $z^k \in \operatorname{dom} q$, and the Lipschitz smoothness of c on $\operatorname{dom} q$, one can conclude that $c(z^{k+1,i}) \le \bar{c}(z^{k+1,i}, z^k) \le 0$. In addition, since $z^k \in \operatorname{dom} q$ and $c(z^k) \le 0$, it follows that $z^k \in \mathcal{S}$, where $\mathcal{S}$ is defined in (6). By this fact and Theorem 3, the Slater's condition holds for (27), that is, there exists a point $\tilde{z} \in \operatorname{dom} q$ such that $\bar{c}(\tilde{z}, z^k) < 0$. Hence, it follows from [24, Corollary 28.2.1, Theorem 28.3] that subproblem (27) has an optimal Lagrange multiplier $\lambda^{k,i}$. We next show that the inner loop terminates in at most $\bar{i}+1$ iterations and outputs a point $z^{k+1} \in \operatorname{dom} q$ with $c(z^{k+1}) \le 0$. To this end, suppose for contradiction that the inner loop runs for more than $\bar{i}+1$ iterations. Then one can observe from Algorithm 1 that

$$h(z^{k+1,\bar{i}}) > h(z^k) - \frac{\beta}{2}\|z^{k+1,\bar{i}} - z^k\|^2. \tag{53}$$

Since $z^{k+1,\bar{i}} = \arg\min_z Q_{k,\bar{i}}(z)$ and $Q_{k,\bar{i}}$ is strongly convex with modulus $L_{k,\bar{i}}$, it follows that $Q_{k,\bar{i}}(z^{k+1,\bar{i}}) \le Q_{k,\bar{i}}(z^k) - L_{k,\bar{i}}\|z^{k+1,\bar{i}} - z^k\|^2/2$, which together with $z^k, z^{k+1,\bar{i}} \in \operatorname{dom} Q_{k,\bar{i}}$ and the definition of $Q_{k,\bar{i}}$ in (52) yields

$$\langle \nabla g(z^k), z^{k+1,\bar{i}}\rangle + \frac{L_{k,\bar{i}}}{2}\|z^{k+1,\bar{i}} - z^k\|^2 + q(z^{k+1,\bar{i}}) \le \langle \nabla g(z^k), z^k\rangle + q(z^k) - \frac{L_{k,\bar{i}}}{2}\|z^{k+1,\bar{i}} - z^k\|^2.$$

By this, (24), and the L-smoothness of g, one has

$$\begin{aligned} h(z^{k+1,\bar{i}}) &\overset{(24)}{=} g(z^{k+1,\bar{i}}) + q(z^{k+1,\bar{i}}) \le g(z^k) + \langle \nabla g(z^k), z^{k+1,\bar{i}} - z^k\rangle + \frac{L}{2}\|z^{k+1,\bar{i}} - z^k\|^2 + q(z^{k+1,\bar{i}}) \\ &\le g(z^k) + q(z^k) - \Big(L_{k,\bar{i}} - \frac{L}{2}\Big)\|z^{k+1,\bar{i}} - z^k\|^2 \overset{(24)}{=} h(z^k) - \Big(L_{k,\bar{i}} - \frac{L}{2}\Big)\|z^{k+1,\bar{i}} - z^k\|^2. \end{aligned} \tag{54}$$

Notice from $L_{k,\bar{i}} = \underline{L}\rho^{\bar{i}}$ and the definition of $\bar{i}$ that $L_{k,\bar{i}} \ge (\beta + L)/2$. This and (54) lead to $h(z^{k+1,\bar{i}}) \le h(z^k) - \beta\|z^{k+1,\bar{i}} - z^k\|^2/2$, which contradicts (53). Hence, the inner loop terminates in at most $\bar{i}+1$ iterations. Moreover, it outputs a point $z^{k+1} \in \operatorname{dom} q$ satisfying $c(z^{k+1}) \le 0$ due to $z^{k+1,i} \in \operatorname{dom} q$ and $c(z^{k+1,i}) \le 0$ for each i. This together with the fact that $z^0 \in \operatorname{dom} q$ and $c(z^0) \le 0$ implies that the induction is complete. Hence, statement (i) holds as desired.

We next prove statement (ii) of Theorem 4. By the definition of L_k and statement (i), one can see that $L_k = L_{k,i}$ for some $0 \le i \le \bar{i}$, which together with the definition of $\bar{i}$ implies that (29) holds. In addition, notice from Algorithm 1 that (z^{k+1}, λ^k) is a pair of optimal solution and Lagrange multiplier of the problem

$$\min_z \Big\{\langle \nabla g(z^k), z\rangle + \frac{L_k}{2}\|z - z^k\|^2 + q(z) : \bar{c}(z, z^k) \le 0\Big\}.$$

The relations (30) and (31) then follow from the KKT conditions of this problem at (z^{k+1}, λ^k). □

We now turn to the proof of Lemma 1.

Proof of Lemma 1. Let $(z^{k+1}, \lambda^k, L_k)$ be generated in the kth outer iteration of Algorithm 1 for some $k \geq 0$. Observe from Algorithm 1 that

$$z^{k+1} = \arg\min_z \Big\{ \langle \nabla g(z^k), z\rangle + \frac{L_k}{2}\|z - z^k\|^2 + q(z) \Big\} \quad \text{s.t.} \quad \bar{c}(z, z^k) \leq 0,$$

and λ^k is its associated optimal Lagrange multiplier. It follows that

$$\lambda^k \geq 0, \quad \langle \lambda^k, \bar{c}(z^{k+1}, z^k)\rangle = 0, \quad z^{k+1} = \arg\min_z \widetilde{L}(z, \lambda^k), \tag{55}$$

where

$$\widetilde{L}(z, \lambda) = \phi(z) + \langle \lambda, \bar{c}(z, z^k)\rangle \ \text{ with } \ \phi(z) = \langle \nabla g(z^k), z\rangle + \frac{L_k}{2}\|z - z^k\|^2 + q(z).$$

Notice from Theorem 4(i) that $z^k \in \mathcal{S}$. By this and Theorem 3, there exists $\hat{y}^k \in \mathcal{Y}$ such that $\bar{c}(\hat{y}^k, z^k) \leq -\zeta < 0$. Using this and (55), we have

$$\phi(z^{k+1}) = \widetilde{L}(z^{k+1}, \lambda^k) = \min_z \widetilde{L}(z, \lambda^k) \leq \widetilde{L}(\hat{y}^k, \lambda^k) = \phi(\hat{y}^k) + \langle \lambda^k, \bar{c}(\hat{y}^k, z^k)\rangle \leq \phi(\hat{y}^k) - \zeta\|\lambda^k\|_1.$$

It follows from this and $L_k \leq \bar{L}$ (see Theorem 4(ii)) that

$$\begin{aligned}\|\lambda^k\|_1 &\leq \zeta^{-1}\big(\phi(\hat{y}^k) - \phi(z^{k+1})\big) \leq \zeta^{-1}\Big(\langle \nabla g(z^k), \hat{y}^k - z^{k+1}\rangle + \frac{L_k}{2}\|\hat{y}^k - z^{k+1}\|^2 + q(\hat{y}^k) - q(z^{k+1})\Big) \\ &\leq \zeta^{-1}\Big(\|\nabla g(z^k)\|\|\hat{y}^k - z^{k+1}\| + \frac{\bar{L}}{2}\|\hat{y}^k - z^{k+1}\|^2 + q(\hat{y}^k) - q(z^{k+1})\Big).\end{aligned}$$

Using this, $\hat{y}^k, z^k, z^{k+1} \in \mathcal{Y}$, and the definitions of G, $D_{\mathcal{Y}}$ and M_q, we see that (32) holds. □

In the remainder of this subsection, we prove Theorem 5. To this end, we first establish several technical lemmas.

The next lemma provides a bound on $\mathrm{dist}(0, \partial\bar{h}(z^{k+1}, z^k))$ in terms of $\|z^{k+1} - z^k\|$.

Lemma 8. *Suppose that Assumptions 1, 2, and 3 hold. Let $\bar{h}$, $\{L_{c_j}\}_{j=1}^m$, and L be given in* (26) *and Assumptions 1 and 3, respectively. Suppose that z^{k+1}, z^k, L_k, and λ^k are generated by Algorithm 1 for some $k \geq 0$. Then it holds that*

$$\mathrm{dist}(0, \partial\bar{h}(z^{k+1}, z^k)) \leq \Big((L + L_k)^2 + 4\|\lambda^k\|_\infty^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)^{\frac{1}{2}}\|z^{k+1} - z^k\|. \tag{56}$$

Proof. Let us fix any $(z, w) \in \mathrm{dom}\,\bar{h}$. It follows from the definition of $\bar{h}$ in (26) that

$$\begin{aligned}\partial\bar{h}(z, w) &\supseteq \hat{\partial}\bar{h}(z, w) \supseteq \begin{pmatrix} \nabla g(z) + \hat{\partial} q(z) \\ \hat{\partial}\delta_{\mathcal{Y}}(w)\end{pmatrix} + \hat{\partial}\delta_{\bar{c}(\cdot,\cdot)\leq 0}(z, w) \\ &= \begin{pmatrix} \nabla g(z) + \partial q(z) \\ \mathcal{N}_{\mathcal{Y}}(w)\end{pmatrix} + \widehat{\mathcal{N}}_{\bar{c}(\cdot,\cdot)\leq 0}(z, w) \\ &\supseteq \left\{ \begin{pmatrix} \nabla g(z) + \partial q(z) + \sum_{j=1}^m \lambda_j\Big(\nabla c_j(w) + L_{c_j}(z - w)\Big) \\ \mathcal{N}_{\mathcal{Y}}(w) + \sum_{j=1}^m \lambda_j\Big(\nabla^2 c_j(w)(z - w) - L_{c_j}(z - w)\Big)\end{pmatrix} : \lambda \in \mathcal{N}_{-\mathbb{R}^m_+}(\bar{c}(z, w)) \right\},\end{aligned} \tag{57}$$

where the second relation follows from (26) and [25, Exercise 8.8, Proposition 10.5, Corollary 10.9], the third relation uses [25, Proposition 8.12, Exercise 8.14] together with the convexity of q and $\mathcal{Y}$, and the last relation follows from (11), [25, Theorem 6.14], and $\widehat{\mathcal{N}}_{-\mathbb{R}^m_+}(\cdot) = \mathcal{N}_{-\mathbb{R}^m_+}(\cdot)$.

Observe from (26), Theorem 4, and Algorithm 1 that $(z^{k+1}, z^k) \in \operatorname{dom} \bar{h}$. By this and (57), one has that for any $\lambda \in \mathcal{N}_{-\mathbb{R}^m_+}(\bar{c}(z^{k+1}, z^k))$,

$$\partial \bar{h}(z^{k+1}, z^k) \supseteq \begin{pmatrix} \nabla g(z^{k+1}) + \partial q(z^{k+1}) + \sum_{j=1}^m \lambda_j \big(\nabla c_j(z^k) + L_{c_j}(z^{k+1} - z^k)\big) \\ \sum_{j=1}^m \lambda_j \big(\nabla^2 c_j(z^k)(z^{k+1} - z^k) - L_{c_j}(z^{k+1} - z^k)\big) \end{pmatrix}.$$

Notice from (11) and (30) that $\lambda^k \in \mathcal{N}_{-\mathbb{R}^m_+}(\bar{c}(z^{k+1}, z^k))$. In addition, observe from (31) that

$$\nabla g(z^{k+1}) - \nabla g(z^k) - L_k(z^{k+1} - z^k) \in \nabla g(z^{k+1}) + \partial q(z^{k+1}) + \sum_{j=1}^m \lambda_j^k \big(\nabla c_j(z^k) + L_{c_j}(z^{k+1} - z^k)\big).$$

In view of these, one has

$$\begin{pmatrix} \nabla g(z^{k+1}) - \nabla g(z^k) - L_k(z^{k+1} - z^k) \\ \sum_{j=1}^m \lambda_j^k \big(\nabla^2 c_j(z^k)(z^{k+1} - z^k) - L_{c_j}(z^{k+1} - z^k)\big) \end{pmatrix} \in \partial \bar{h}(z^{k+1}, z^k). \tag{58}$$

Notice from the L-smoothness of g that

$$\|\nabla g(z^{k+1}) - \nabla g(z^k) - L_k(z^{k+1} - z^k)\| \le \|\nabla g(z^{k+1}) - \nabla g(z^k)\| + \|L_k(z^{k+1} - z^k)\| \le (L + L_k)\|z^{k+1} - z^k\|. \tag{59}$$

On the other hand, since $\lambda^k \in \mathbb{R}^m_+$, we have

$$\begin{aligned} &\Big\| \sum_{j=1}^m \lambda_j^k \big(\nabla^2 c_j(z^k)(z^{k+1} - z^k) - L_{c_j}(z^{k+1} - z^k)\big) \Big\| \le \sum_{j=1}^m \lambda_j^k \|\nabla^2 c_j(z^k)(z^{k+1} - z^k) - L_{c_j}(z^{k+1} - z^k)\| \\ &\le \sum_{j=1}^m \lambda_j^k \big(\|\nabla^2 c_j(z^k)(z^{k+1} - z^k)\| + \|L_{c_j}(z^{k+1} - z^k)\|\big) \le 2 \sum_{j=1}^m \lambda_j^k L_{c_j} \|z^{k+1} - z^k\| \le 2\|\lambda^k\|_\infty \sum_{j=1}^m L_{c_j} \|z^{k+1} - z^k\|, \end{aligned} \tag{60}$$

where the third inequality follows from $z^k, z^{k+1} \in \operatorname{dom} q$, the convexity of $\operatorname{dom} q$, and L_{c_j}-smoothness of c_j over $\operatorname{dom} q$. Combining (58), (59), and (60) yields (56), and hence the conclusion holds. □

The next lemma establishes the convergence rate of Algorithm 1 under suitable assumptions.

Lemma 9. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{(z^\ell, \lambda^{\ell-1})\}_{\ell=1}^k$ be generated by Algorithm 1 for some $k \ge 1$, and let $\theta, \eta, \omega, \alpha, C'$ be given in (25), (33), and (34), respectively. Then the following statements hold.*

(i) If $\theta = 1/2$, then

$$h(z^k) - h^* \le \eta(1 + \alpha)^{-k}. \tag{61}$$

(ii) If $\theta \in (1/2, 1)$, then

$$h(z^k) - h^* \le \Big(\frac{1}{C'(2\theta - 1)}\Big)^{\frac{1}{2\theta - 1}} k^{-\frac{1}{2\theta - 1}}. \tag{62}$$

Proof. For notational convenience, let $r_\ell := h(z^\ell) - h^*$ for all $0 \le \ell \le k$. If $h(z^k) = h^*$, then relations (61) and (62) clearly hold. For the remainder of the proof, suppose that $h(z^k) > h^*$. Notice from (26) and Algorithm 1 that $(z^{\ell+1}, z^\ell) \in \operatorname{dom} \bar{h}$, which along with (26) implies that $\bar{h}(z^{\ell+1}, z^\ell) = h(z^{\ell+1})$ for all $0 \le \ell < k$. Also, by a similar argument as used in the proof of Lemma 3, one has $h^* = \bar{h}^*$, where $\bar{h}^*$ is defined in (26). These, together with $h(z^0) - h^* \le \eta$, $h(z^k) > h^*$, and the monotonicity of $\{h(z^\ell)\}$, lead to

$$0 < \bar{h}(z^{\ell+1}, z^\ell) - \bar{h}^* = h(z^{\ell+1}) - h^* \le h(z^0) - h^* \le \eta \qquad \forall 0 \le \ell < k.$$

It then follows from (25) that

$$C(\bar{h}(z^{\ell+1}, z^\ell) - \bar{h}^*)^\theta \le \operatorname{dist}(0, \partial \bar{h}(z^{\ell+1}, z^\ell)) \qquad \forall 0 \le \ell < k. \tag{63}$$

In addition, notice from Theorem 4(ii) and Lemma 1 that $L_\ell \le \bar{L}$ and $\|\lambda^\ell\|_\infty \le A$ for all $0 \le \ell < k$, where $\bar{L}$ and A are defined in (29) and (32). Using these, (33), and Lemma 8, we obtain that $\mathrm{dist}(0, \partial \bar{h}(z^{\ell+1}, z^\ell)) \le \omega \|z^{\ell+1} - z^\ell\|$ for all $0 \le \ell < k$. Also, notice from Algorithm 1 that

$$r_\ell - r_{\ell+1} = h(z^\ell) - h(z^{\ell+1}) \ge \frac{\beta}{2}\|z^{\ell+1} - z^\ell\|^2 \qquad \forall 0 \le \ell < k.$$

In view of these, (34), and (63), one has that for all $0 \le \ell < k$,

$$\begin{aligned} r_\ell - r_{\ell+1} &\ge \frac{\beta}{2}\|z^{\ell+1} - z^\ell\|^2 \ge \frac{\beta}{2\omega^2}\mathrm{dist}^2(0, \partial \bar{h}(z^{\ell+1}, z^\ell)) \overset{(63)}{\ge} \frac{\beta C^2}{2\omega^2}\big(\bar{h}(z^{\ell+1}, z^\ell) - \bar{h}^*\big)^{2\theta} \\ &= \frac{\beta C^2}{2\omega^2}(h(z^{\ell+1}) - h^*)^{2\theta} \overset{(34)}{=} \alpha\,(h(z^{\ell+1}) - h^*)^{2\theta} = \alpha\, r_{\ell+1}^{2\theta}. \end{aligned} \tag{64}$$

(i) Suppose $\theta = 1/2$. It then follows from (64) that $r_{\ell+1} \le (1+\alpha)^{-1} r_\ell$ for all $0 \le \ell < k$, which together with $r_0 \le \eta$ implies that $r_k \le r_0(1+\alpha)^{-k} \le \eta(1+\alpha)^{-k}$, and hence (61) holds.

(ii) Suppose $\theta \in (1/2, 1)$. Notice from the above that $r_k > 0$, which together with the monotonicity of $\{r_\ell\}$ implies that $r_\ell > 0$ for all $0 \le \ell \le k$. Letting $\psi(t) = \frac{1}{2\theta-1} t^{1-2\theta}$ and using the monotonicity of $\{r_\ell\}$, we have

$$\psi(r_{\ell+1}) - \psi(r_\ell) = \int_{r_\ell}^{r_{\ell+1}} \psi'(t)dt = \int_{r_{\ell+1}}^{r_\ell} t^{-2\theta} dt \ge r_\ell^{-2\theta}(r_\ell - r_{\ell+1}) \qquad \forall 0 \le \ell < k. \tag{65}$$

For each $0 \le \ell < k$, we consider two separate cases below.

Case a): $r_{\ell+1}^{-2\theta} \le 2 r_\ell^{-2\theta}$. It along with (64) and (65) implies that

$$\psi(r_{\ell+1}) - \psi(r_\ell) \ge \frac{1}{2} r_{\ell+1}^{-2\theta}(r_\ell - r_{\ell+1}) \ge \frac{1}{2}\alpha.$$

Case b): $r_{\ell+1}^{-2\theta} > 2 r_\ell^{-2\theta}$. It leads to $r_{\ell+1}^{1-2\theta} > 2^{\frac{2\theta-1}{2\theta}} r_\ell^{1-2\theta}$. By this, $\theta \in (1/2, 1)$, $r_\ell \le \eta$, and the expression of ψ, one has that

$$\psi(r_{\ell+1}) - \psi(r_\ell) = \frac{1}{2\theta-1}(r_{\ell+1}^{1-2\theta} - r_\ell^{1-2\theta}) > \frac{1}{2\theta-1}\left(2^{\frac{2\theta-1}{2\theta}} - 1\right) r_\ell^{1-2\theta} \ge \frac{1}{2\theta-1}\left(2^{\frac{2\theta-1}{2\theta}} - 1\right)\eta^{1-2\theta}.$$

Combining the above two cases and using the definition of C' in (34), we obtain that $\psi(r_{\ell+1}) - \psi(r_\ell) \ge C'$ for all $0 \le \ell < k$. It follows that $\psi(r_k) \ge \psi(r_0) + kC' \ge kC'$. This together with the expression of ψ yields

$$r_k \le \left(\frac{1}{C'(2\theta-1)}\right)^{\frac{1}{2\theta-1}} k^{-\frac{1}{2\theta-1}}.$$

Relation (62) then follows from this and $r_k = h(z^k) - h^*$. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. Suppose for contradiction that Algorithm 1 runs for more than $\overline{K}_\theta$ outer iterations. It along with (28) implies that there exists some $\ell \ge \overline{K}_\theta - 1$ such that

$$\|\nabla g(z^{\ell+1}) - \nabla g(z^\ell) - L_\ell(z^{\ell+1} - z^\ell)\|^2 + 4\Big(\sum_{j=1}^m \lambda_j^\ell L_{c_j}\Big)^2 \|z^{\ell+1} - z^\ell\|^2 > \tau^2. \tag{66}$$

For notational convenience, let $r_\ell = h(z^\ell) - h^*$ and $r_{\ell+1} = h(z^{\ell+1}) - h^*$. We now show that $r_\ell \le \beta\tau^2/(2\omega^2)$ by considering two separate cases below.

Case a): $\theta = 1/2$. By this, (35), and $\ell \ge \overline{K}_\theta - 1$, one has $\ell \ge \log_{1+\alpha}\big(\frac{2\omega^2\eta}{\beta\tau^2}\big)$. It then follows from (61) that $r_\ell \le \eta(1+\alpha)^{-\ell} \le \beta\tau^2/(2\omega^2)$.

Case b): $\theta \in (1/2, 1)$. Using this, (35), and $\ell \geq \overline{K}_\theta - 1$, we have $\ell \geq \frac{1}{C'(2\theta-1)}\big(\frac{2\omega^2}{\beta\tau^2}\big)^{2\theta-1}$. It then follows from (62) that

$$r_\ell \leq \Big(\frac{1}{C'(2\theta-1)}\Big)^{\frac{1}{2\theta-1}} \ell^{-\frac{1}{2\theta-1}} \leq \frac{\beta\tau^2}{2\omega^2}.$$

Combining these two cases, we conclude that $r_\ell \leq \beta\tau^2/(2\omega^2)$. In addition, notice from Algorithm 1 that

$$r_\ell - r_{\ell+1} = h(z^\ell) - h(z^{\ell+1}) \geq \beta\|z^{\ell+1} - z^\ell\|^2/2.$$

By these relations and $r_{\ell+1} \geq 0$, one has

$$\|z^{\ell+1} - z^\ell\| \leq \sqrt{\frac{2(r_\ell - r_{\ell+1})}{\beta}} \leq (2/\beta)^{\frac{1}{2}} r_\ell^{\frac{1}{2}} \leq \tau/\omega. \tag{67}$$

Also, using Theorem 4(ii), we have $0 < L_\ell \leq \bar{L}$, where $\bar{L}$ is defined in (29). This together with (59) yields

$$\|\nabla g(z^{\ell+1}) - \nabla g(z^\ell) - L_\ell(z^{\ell+1} - z^\ell)\|^2 \overset{(59)}{\leq} (L + L_\ell)^2\|z^{\ell+1} - z^\ell\|^2 \leq (L + \bar{L})^2\|z^{\ell+1} - z^\ell\|^2. \tag{68}$$

In addition, notice from Lemma 1 that $\|\lambda^\ell\|_\infty \leq A$, where A is defined in (32). It follows from this and $\lambda^\ell \in \mathbb{R}^m_+$ that

$$\big(\sum_{j=1}^m \lambda_j^\ell L_{c_j}\big)^2 \leq \|\lambda^\ell\|_\infty^2\big(\sum_{j=1}^m L_{c_j}\big)^2 \leq A^2\big(\sum_{j=1}^m L_{c_j}\big)^2. \tag{69}$$

Using this, (67), (68), and the definition of ω in (33), we obtain that

$$\begin{aligned}
&\|\nabla g(z^{\ell+1}) - \nabla g(z^\ell) - L_\ell(z^{\ell+1} - z^\ell)\|^2 + 4\big(\sum_{j=1}^m \lambda_j^\ell L_{c_j}\big)^2\|z^{\ell+1} - z^\ell\|^2 \\
&\overset{(68)(69)}{\leq} \Big((L + \bar{L})^2 + 4A^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)\|z^{\ell+1} - z^\ell\|^2 \overset{(33)}{=} \omega^2\|z^{\ell+1} - z^\ell\|^2 \overset{(67)}{\leq} \tau^2,
\end{aligned}$$

which contradicts (66). Hence, Algorithm 1 terminates in at most $\overline{K}_\theta$ outer iterations.

We next show that (36) holds. If $h(z^{k+1}) = h^*$, (36) clearly holds. For the remainder of the proof, suppose that $h(z^{k+1}) > h^*$. By similar arguments as above, one has that $0 < L_k \leq \bar{L}$, $\|\lambda^k\|_\infty \leq A$, and $\|z^{k+1} - z^k\| \leq \tau/\omega$. It follows from these, (33), and (56) that

$$\begin{aligned}
\mathrm{dist}(0, \partial\bar{h}(z^{k+1}, z^k)) &\overset{(56)}{\leq} \Big((L + L_k)^2 + 4\|\lambda^k\|_\infty^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)^{\frac{1}{2}}\|z^{k+1} - z^k\| \\
&\leq \Big((L + \bar{L})^2 + 4A^2\big(\sum_{j=1}^m L_{c_j}\big)^2\Big)^{\frac{1}{2}}\|z^{k+1} - z^k\| \overset{(33)}{=} \omega\|z^{k+1} - z^k\| \leq \tau.
\end{aligned} \tag{70}$$

In addition, notice that $0 < h(z^{k+1}) - h^* \leq h(z^0) - h^* \leq \eta$, $\bar{h}(z^{k+1}, z^k) = h(z^{k+1})$, and $\bar{h}^* = h^*$. It then follows that (25) holds with $z = z^{k+1}$ and $w = z^k$. In view of these and (70), one has

$$h(z^{k+1}) - h^* = \bar{h}(z^{k+1}, z^k) - \bar{h}^* \leq \big(C^{-1}\mathrm{dist}(0, \partial\bar{h}(z^{k+1}, z^k))\big)^{\frac{1}{\theta}} \leq (C^{-1}\tau)^{\frac{1}{\theta}}.$$

Hence, (36) holds as desired. □

6.3 Proof of the main results in Section 4

In this subsection we prove Lemma 2 and Theorems 6 and 7.

Proof of Lemma 2. Fix any $\bar{k} \geq 0$. Observe that $\{\lambda^{\bar{k},\ell}\}$ reduces to the sequence of Lagrange multipliers generated by Algorithm 1 for solving the subproblem $\min\{-f(x^{\bar{k}}, y) + q(y) : c(y) \leq 0\}$. Also, notice from Algorithm 2 that $x^{\bar{k}} \in \mathcal{X}$, which together with the definition of G_f implies that

$$\max_{y \in \mathcal{Y}} \big\|\nabla_y f(x^{\bar{k}}, y)\big\| \leq G_f.$$

Using this relation, the definition of A_f, and Lemma 1 with G replaced by G_f, we obtain that $\|\lambda^{\bar{k},\ell}\|_1 \leq A_f$ holds for all ℓ. By this and the arbitrariness of $\bar{k}$, one can see that the conclusion holds. □

We now turn to the proofs of Theorems 6 and 7. Notice that Algorithm 2 shares key similarities with [17, Algorithm 2], which is designed to solve the unconstrained nonconvex-nonconcave problem $\min_x \max_y\{f(x,y) + p(x) - q(y)\}$. In particular, Algorithm 2 applies an inexact proximal gradient (IPG) method to solve $\min_x\{F^*(x) + p(x)\}$, while [17, Algorithm 2] applies an IPG method to solve $\min_x\{\tilde{F}^*(x) + p(x)\}$, where $\tilde{F}^*(x) = \max_y\{f(x,y) - q(y)\}$. The two algorithms follow almost identical steps, differing only in how they approximate $\nabla^{\rm C}_{\mathcal{X}} F^*(x^k)$ and $\nabla^{\rm C}_{\mathcal{X}} \tilde{F}^*(x^k)$ at a given iterate x^k. Specifically, Algorithm 2 computes an approximation to $\nabla^{\rm C}_{\mathcal{X}} F^*(x^k)$ by calling the SCP method (Algorithm 1) for the subproblem $\min_y\{-f(x^k, y) + q(y) : c(y) \leq 0\}$, whereas [17, Algorithm 2] computes an approximation to $\nabla^{\rm C}_{\mathcal{X}} \tilde{F}^*(x^k)$ using a proximal gradient method for the subproblem $\min_y\{-f(x^k, y) + q(y)\}$. Thanks to these close similarities, the proofs of Theorems 6 and 7 largely parallel those of [17, Theorems 4 and 5]. We therefore provide only a sketch of the proofs. To this end, we first present two technical lemmas.

Lemma 10. *Let $\mathcal{X}^{\rm c}_\epsilon, L_{\nabla f}, C, \theta, \gamma, \sigma, \epsilon, \{\eta_\ell\}$ be given in (37), Assumption 1, and Algorithm 2, respectively. Suppose that $\{(x^\ell, y^\ell)\}_{\ell=0}^k$ are generated by Algorithm 2 for some $k \geq 1$ such that $x^\ell \in \mathcal{X}^{\rm c}_\epsilon$ for all $0 \leq \ell < k$. Then, for all $0 \leq \ell \leq k$, it holds that*

$$F^*(x^\ell) - F(x^\ell, y^\ell) \leq \min\big\{\gamma\epsilon^\sigma/2, \eta_\ell\big\}, \qquad \mathrm{dist}\big(y^\ell, Y^*(x^\ell)\big) \leq \frac{1}{C(1-\theta)} \min\big\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\big\}, \tag{71}$$

$$\|\nabla^{\rm C}_{\mathcal{X}} F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell)\| \leq \frac{L_{\nabla f}}{C(1-\theta)} \min\big\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\big\}. \tag{72}$$

Proof. This lemma is parallel to [17, Lemma 9], whose proof is based on [17, Theorems 1 and 3] and [17, Lemma 4]. Note that Theorems 1, 5, and 2 are parallel to [17, Theorems 1 and 3] and [17, Lemma 4], respectively. Hence, the conclusion follows from Theorems 1, 2, and 5, together with arguments similar to those used in the proof of [17, Lemma 9]. □

Lemma 11. *Let $\epsilon > 0$ be given, M, $\mathcal{X}^{\rm c}_\epsilon$ be defined in (37) and (19), $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma, \{\delta_\ell\}, \{\eta_\ell\}, \{L_\ell\}$ be given in Assumption 1 and Algorithm 2, and let*

$$\begin{aligned}
&\Delta_k := 8\Big[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^k \Big(1 + \tfrac{L^2_{\nabla f}}{(1-\theta)^2 C^2 L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2}\Big],\\
&\underline{K}_\epsilon := \max\{k \geq 1 : \Delta_k/(kL_{\lceil k/2\rceil}) \geq \gamma^2\epsilon^{2\sigma}/(16L_f^2)\},\\
&\overline{K}_\epsilon := \max\{k \geq 0 : x^k \in \mathcal{X}^{\rm c}_\epsilon\},\\
&\ell(k) := \operatorname*{arg\,min}_{\lceil k/2\rceil \leq \ell \leq k} L_\ell\|x^{\ell+1} - x^\ell\|^2.
\end{aligned}$$

Let $\underline{K}_\epsilon < k \leq \overline{K}_\epsilon$ be given. Suppose that $\{(x^\ell, y^\ell)\}_{\ell=0}^k$ are generated by Algorithm 2 such that $x^\ell \in \mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0 \leq \ell \leq k$. Then we have

$$\operatorname{dist}\big(0, \partial\Psi(x^{\ell(k)+1})\big) \leq L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}} + \sqrt{\frac{L_k\Delta_k}{k}} + M\Big(\frac{\Delta_k}{L_{\lceil k/2\rceil}k}\Big)^{\frac{\nu}{2}} + (1-\theta)^{-1}C^{-1}L_{\nabla f}\eta_{\lceil k/2\rceil}^{\frac{1}{2}}.$$

Proof. This lemma is parallel to [17, Lemma 10], whose proof is based on [17, Eqs. (25), (72), and (73)]. Note that (38), (71), and (72) are parallel to [17, Eqs. (25), (72), and (73)], respectively. Hence, the conclusion follows from (38), (71), and (72), together with arguments similar to those used in the proof of [17, Lemma 10]. □

We are now ready to provide a sketch of the proofs of Theorems 6 and 7.

Proof of Theorem 6. Theorem 6 is parallel to [17, Theorem 4], whose proof is based on [17, Lemmas 9 and 10]. Note that [17, Lemmas 9 and 10] are parallel to Lemmas 10 and 11, respectively. Hence, the conclusion follows from Lemmas 10 and 11, together with arguments similar to those used in the proof of [17, Theorem 4]. □

Proof of Theorem 7. Theorem 7 is parallel to [17, Theorem 5], whose proof is based on [17, Theorems 3 and 4]. Note that [17, Theorems 3 and 4] are parallel to Theorems 5 and 6, respectively. Hence, the conclusion follows from Theorems 5 and 6, together with arguments similar to those used in the proof of [17, Theorem 5]. □

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[3] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[4] J. Blanchet, J. Li, S. Lin, and X. Zhang. Distributionally robust optimization and robust statistics. *Statistical Science*, 40(3):351–377, 2025.

[5] A. Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. *arXiv preprint arXiv:2201.12247*, 2022.

[6] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[7] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[8] Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. *arXiv preprint arXiv:2210.03096*, 2022.

[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] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

[11] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. *Mathematical Programming*, 185:357–383, 2021.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] F. Huang. Enhanced adaptive gradient algorithms for nonconvex-PL minimax optimization. *arXiv preprint arXiv:2303.03984*, 2023.

[14] J. Li, L. Zhu, and A. M.-C. So. Nonsmooth nonconvex–nonconcave minimax optimization: Primal–dual balancing and iteration complexity analysis. *Mathematical Programming*, pages 1–51, 2025.

[15] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

[16] Z. Lu. Sequential convex programming methods for a class of structured nonlinear programming. *arXiv preprint arXiv:1210.3039*, 2012.

[17] Z. Lu and X. Wang. A first-order method for nonconvex-nonconcave minimax problems under a local Kurdyka-Łojasiewicz condition. *arXiv preprint arXiv:2507.01932*, 2025. Accepted by *SIAM Journal on Optimization*.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[19] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[20] 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.

[21] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690, 2017.

[22] T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2302.09831*, 2023.

[23] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[24] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1997.

[25] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[26] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[27] Z. Xu, Z.-Q. Wang, J.-L. Wang, and Y.-H. Dai. Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *Journal of Machine Learning Research*, 24(313):1–25, 2023.

[28] J. Yang, A. Orvieto, A. Lucchi, and N. He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517, 2022.

[29] P. Yu, T. K. Pong, and Z. Lu. Convergence rate analysis of a sequential convex programming method with line search for a class of constrained difference-of-convex optimization problems. *SIAM Journal on Optimization*, 31(3):2024–2054, 2021.

[30] T. Zheng, A. M.-C. So, and J. Li. Doubly smoothed optimistic gradients: A universal approach for smooth minimax problems. *arXiv preprint arXiv:2506.07397*, 2025.

[31] T. Zheng, L. Zhu, A. M.-C. So, J. Blanchet, and J. Li. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. *Advances in Neural Information Processing Systems*, 36:54075–54110, 2023.