

A first-order method for nonconvex–nonconcave minimax problems under a local Kurdyka–Łojasiewicz condition

Zhaosong Lu ^{*} Xiangyuan Wang ^{*}

July 2, 2025 (Revised: May 19, 2026)

Abstract

We study a class of nonconvex–nonconcave minimax problems in which the inner maximization problem satisfies a local Kurdyka–Łojasiewicz (KL) condition that may vary with the outer minimization variable. In contrast to the global KL or Polyak–Łojasiewicz (PL) conditions commonly assumed in the literature—which are significantly stronger and often too restrictive in practice—this local KL condition accommodates a broader range of practical scenarios. However, it also introduces new analytical challenges. In particular, as an optimization algorithm progresses toward a stationary point of the problem, the region over which the KL condition holds may shrink, resulting in a more intricate and potentially ill-conditioned landscape. To address this challenge, we show that the associated maximal function is locally generalized Hölder smooth. Leveraging this key property, we develop an inexact proximal gradient method for solving the minimax problem, where the inexact gradient of the maximal function is computed by applying a proximal gradient method to a KL-structured subproblem. Under mild assumptions, we establish complexity guarantees for computing an approximate stationary point of the minimax problem.

Keywords: nonconvex–nonconcave minimax, local KL condition, local generalized Hölder smoothness, inexact proximal gradient method, first-order oracle complexity

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

1 Introduction

In this paper, we consider a nonconvex–nonconcave minimax problem of the form

$$\min_x \max_y \{f(x, y) + p(x) - q(y)\}, \quad (1)$$

where f is a smooth function that is nonconvex in x and nonconcave in y , and p and q are possibly nonsmooth, closed, and simple convex functions.

Problem (1) arises in a wide range of applications in machine learning and operations research, including generative adversarial networks [1, 15], reinforcement learning [11, 25], adversarial training [22, 30], and distributionally robust optimization [5, 6, 27]. Despite its broad applicability, problem (1) remains computationally challenging due to its inherent nonconvex–nonconcave structure. For instance, computing a global Nash equilibrium—an important special case of (1)—is generally NP-hard (see, e.g., [17]).

In recent years, significant progress has been made under specific structural assumptions on problem (1). For example, several studies focus on the special case where $q = 0$ and impose the global Polyak–Łojasiewicz (PL) condition on the inner maximization problem of (1), which is generally weaker

^{*}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.

than the strong concavity assumption. Under this condition, gradient descent–ascent type methods have been developed, and complexity guarantees have been established for finding approximate stationary points. Remarkably, these guarantees match those obtained under the strong concavity assumption for the inner maximization subproblem of (1) (see, e.g., [16, 24, 32, 33]). In addition, first-order methods have been developed for problem (1) from a variational inequality perspective, typically assuming the existence of a weak Minty variational inequality solution (see, e.g., [7, 9, 21, 26]).

More recently, [20, 34, 35] studied a class of minimax problems of the form (1), where p and q are indicator functions of simple convex compact sets and a global Kurdyka–Łojasiewicz (KL) condition is imposed on the inner maximization problem. They developed gradient descent–ascent-type methods that alternately update the x and y variables using first-order schemes, and established complexity guarantees for finding approximate stationary points. Notably, the class of minimax problems considered in [20, 34, 35] is significantly broader than those studied in [16, 24, 32, 33], since the KL condition generalizes the PL condition (which corresponds to the KL condition with exponent $1/2$) and accommodates nonsmooth objectives.

However, requiring the KL property to hold globally is often too restrictive in practice, which limits the applicability of the proposed methods. To address this limitation, we relax the global KL assumption on the inner maximization problem of (1) that is imposed in the existing literature. Specifically, we assume that for each fixed outer variable $x \in \text{dom } p$, the KL condition holds only on a level set of the inner variable y . Moreover, this level set may depend on x , and its size may vary accordingly (see Assumption 1(iii) for details). This weaker assumption accommodates a broader range of practical scenarios but also introduces new analytical challenges. In particular, as an optimization algorithm progresses toward a stationary point of (1), the region over which the KL condition is valid may shrink, resulting in a more intricate and potentially ill-conditioned landscape. Moreover, we consider more general functions p and q , beyond the indicator functions of simple convex compact sets considered in prior works [20, 34, 35], thereby further broadening the class of minimax problems under consideration.

In this paper, we study problem (1) under the aforementioned local KL condition and other mild assumptions (see Assumption 1 below). In particular, we show that the maximal function, defined as $F^*(x) := \max_y \{f(x, y) - q(y)\}$, is locally generalized Hölder smooth on the set $\{x \in \text{dom } p : 0 \notin \partial \Psi(x)\}$, where $\Psi(x) := F^*(x) + p(x)$ is the value function of problem (1) (see Theorem 1). Leveraging this key property, we develop an inexact proximal gradient method (Algorithm 2) to solve the problem $\min_x \{F^*(x) + p(x)\}$, which is equivalent to the original minimax problem (1). Specifically, given the current iterate x^k , we apply a proximal gradient method (Algorithm 1) to approximately solve the subproblem $\max_y \{f(x^k, y) - q(y)\}$, starting from the previous inner iterate y^{k-1} , and obtain an approximate solution y^k . We then perform an inexact proximal gradient step to compute x^{k+1} , using $-\nabla_x f(x^k, y^k)$ as the forward direction together with a carefully chosen step size. We also establish complexity guarantees for the proposed method for 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, which plays a crucial role in the development of a method for solving problem (1) (see Theorem 1).
- We propose an inexact proximal gradient method for finding approximate stationary points of problem (1). Under mild assumptions, we establish that this method achieves an *iteration complexity* of $\tilde{\mathcal{O}}(\epsilon^{-\max\{(1-\theta)^{-1}, \theta^{-1}\}\sigma})$, and a *first-order oracle complexity* of $\tilde{\mathcal{O}}(\epsilon^{-(1-\theta)^{-1}(2\theta^2-2\theta+1)\max\{(1-\theta)^{-1}, \theta^{-1}\}\sigma})$, measured by the number of gradient evaluations, for finding an $\mathcal{O}(\epsilon)$ -approximate stationary point of (1). Here, θ and σ are the parameters for the local KL condition given in Assumption 1.¹

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 establish a local generalized Hölder

¹ $\tilde{\mathcal{O}}(\cdot)$ represents $\mathcal{O}(\cdot)$ with logarithmic factors hidden.

smoothness property of the maximal function. Section 3 presents a proximal gradient method for minimizing functions that satisfy a local KL property. In Section 4, we propose an inexact proximal gradient method for solving problem (1) and analyze its complexity. Section 5 presents preliminary numerical results illustrating the performance of the proposed method. In Section 6, we provide the proof of the main results. Finally, we make some concluding remarks in Section 7.

1.1 Notation, terminology, and assumptions

The following notation will be used throughout the paper. Let \mathbb{R}^n denote the n -dimensional Euclidean space, and let $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. 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 $\text{dist}(x, S)$ denote the distance from x to S . The closed ball centered at $x \in \mathbb{R}^n$ with radius r is denoted by $\mathcal{B}(x, r)$. In addition, $\text{conv}(\cdot)$, $\text{aff}(\cdot)$, $\text{ri}(\cdot)$, and $\text{int}(\cdot)$ denote the convex hull, affine hull, relative interior, and interior of a set, respectively.

For an extended real-valued function $\phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, its *domain* is denoted by $\text{dom } \phi$, i.e., $\text{dom } \phi = \{x : \phi(x) < \infty\}$. Such ϕ is called *proper* if $\text{dom } \phi \neq \emptyset$, and it is called *closed* (or *lower semicontinuous*) if $\liminf_{z \rightarrow x} \phi(z) \geq \phi(x)$ holds for all $x \in \mathbb{R}^n$. The *limiting subdifferential* (see, e.g., [29, Definition 8.3(b)]) of a proper closed function ϕ at $x \in \text{dom } \phi$ is defined as

$$\partial\phi(x) := \left\{ v \in \mathbb{R}^n : \exists x^k \xrightarrow{\phi} x, v^k \rightarrow v \text{ with } \liminf_{z \rightarrow x^k, z \neq x^k} \frac{\phi(z) - \phi(x^k) - \langle v^k, z - x^k \rangle}{\|z - x^k\|} \geq 0 \ \forall k \right\}.$$

In addition, we use $\partial_{x_i}\phi$ to denote the limiting subdifferential of ϕ with respect to x_i . 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. Moreover, if ϕ is proper and closed, $\partial\phi$ is outer semicontinuous on $\text{dom } \phi$. That is, for any $x \in \text{dom } \phi$, if $x^k \xrightarrow{\phi} x$ and $v^k \in \partial\phi(x^k)$ with $v^k \rightarrow v$, then $v \in \partial\phi(x)$ (see, e.g., [29, Proposition 8.7]).

Suppose that $\phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a locally Lipschitz continuous function on \mathcal{X} . The *Clarke subdifferential* (see, e.g., [10, Definition (1.1)]) of ϕ at $x \in \mathcal{X}$, denoted by $\partial^C\phi(x)$, is defined as

$$\partial^C\phi(x) := \text{conv}\{v : \exists x^k \rightarrow x \text{ such that } \nabla\phi(x^k) \rightarrow v\}. \quad (2)$$

The *restricted Clarke subdifferential* of ϕ with respect to \mathcal{X} , denoted by $\partial_{\mathcal{X}}^C\phi$, is defined as

$$\partial_{\mathcal{X}}^C\phi(x) := \text{conv}\{v : \exists x^k \in \mathcal{X} \rightarrow x \text{ such that } \nabla\phi(x^k) \rightarrow v\} \quad \forall x \in \mathcal{X}. \quad (3)$$

Clearly, $\partial_{\mathcal{X}}^C\phi(x) \subseteq \partial^C\phi(x)$ for all $x \in \mathcal{X}$. When $\partial_{\mathcal{X}}^C\phi(x)$ is a singleton, we denote its unique element by $\nabla_{\mathcal{X}}^C\phi(x)$.

A function $\phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is called L_ϕ -*Lipschitz continuous* on \mathcal{X} if $|\phi(x) - \phi(y)| \leq L_\phi\|x - y\|$ for all $x, y \in \mathcal{X}$, and $L_{\nabla\phi}$ -*smooth* on \mathcal{X} if $\|\nabla\phi(x) - \nabla\phi(y)\| \leq L_{\nabla\phi}\|x - y\|$ for all $x, y \in \mathcal{X}$. In addition, we introduce the following notion of *generalized Hölder smoothness*.

Definition 1 (generalized Hölder smoothness). *Suppose that $\phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a locally Lipschitz continuous function on \mathcal{X} and that $\partial_{\mathcal{X}}^C\phi(x)$ is a singleton for all $x \in \mathcal{X}$. The function ϕ is said to be generalized Hölder smooth on \mathcal{X} if there exist $L_1 \geq 0$, $L_2 \geq 0$, and $\nu \in (0, 1)$ such that*

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

We now introduce an approximate stationary point for the problem $\min_x \phi(x)$, where ϕ is a proper closed function. A similar notion has been used in the context of weakly convex optimization (see, e.g., [12]). As will be shown later, under mild assumptions, the minimax problem (1) can be viewed as a special case of this problem. Consequently, the following definition applies to problem (1) as well.

Definition 2 ((r, ϵ) -stationary point). Suppose ϕ is a proper closed function. For any $\epsilon > 0$ and $r \geq 0$, a point \bar{x} is called an (r, ϵ) -stationary point of the problem $\min_x \phi(x)$ if $\bar{x} \in \text{dom } \phi$ and $\text{dist}(\bar{x}, \mathcal{S}_\epsilon) \leq r$, where $\mathcal{S}_\epsilon = \{x : \text{dist}(0, \partial\phi(x)) \leq \epsilon\}$.

It should be noted that when ϕ is a locally Lipschitz continuous function, any (r, ϵ) -stationary point \bar{x} of ϕ is also an (r, ϵ) -Goldstein stationary point of ϕ , that is, $\text{dist}(0, \partial_r \phi(\bar{x})) \leq \epsilon$, where

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

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

$$\mathcal{X} := \text{dom } p, \quad \mathcal{Y} := \text{dom } q, \quad F(x, y) := f(x, y) - q(y), \quad (4)$$

$$F^*(x) := \max_y F(x, y), \quad Y^*(x) := \{y : F(x, y) = F^*(x)\}, \quad (5)$$

$$\Psi(x) := F^*(x) + p(x), \quad \Psi^* := \min_x \Psi(x). \quad (6)$$

We assume that problem (1) satisfies the following assumption.

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 $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is $L_{\nabla f}$ -smooth on $\mathcal{X} \times \mathcal{Y}$.

(ii) $p : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper closed convex, $q : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper closed convex, and the proximal operators of p and q can be computed exactly. In addition, we assume that $\text{aff}(\mathcal{X}) = \mathbb{R}^n$.²

(iii) For any fixed $x \in \mathcal{X}$, $\max_y F(x, y)$ has a nonempty solution set and a finite optimal value. The function F satisfies the following local Kurdyka-Lojasiewicz (KL) condition in y : there exist constants $C > 0$, $\theta \in [1/2, 1)$, $\gamma > 0$, and $\sigma > 0$ such that for any $x \in \mathcal{X}$,

$$C(F^*(x) - F(x, y))^\theta \leq \text{dist}(0, \partial_y F(x, y)) \quad \forall y \in \mathcal{L}(x), \quad (7)$$

where

$$\mathcal{L}(x) := \{y : 0 < F^*(x) - F(x, y) \leq \gamma \text{dist}(0, \partial\Psi(x))^\sigma\}. \quad (8)$$

Remark 1. We refer to condition (7) as a local KL condition because the associated KL inequality holds only on a level set of the variable y , which may vary with x . This condition is significantly weaker than the global KL condition imposed in the literature [20, 34, 35], where the KL inequality is required to hold for all y . In contrast to the global KL condition, the local KL condition applies to a broader class of minimax problems. However, the local KL condition introduces new analytical challenges. In particular, as an optimization algorithm progresses toward a stationary point of problem (1), the region in which the KL condition holds may shrink, leading to a more intricate and potentially ill-conditioned landscape. As a result, addressing problem (1) under the local KL condition requires substantially different algorithmic design and analysis.

We conclude this subsection with an example of a minimax problem that satisfies the local KL condition in Assumption 1(iii), but not the global KL condition (which requires the KL inequality to hold for all $y \in \mathcal{Y}$). This example demonstrates that the local KL condition applies to a broader class of minimax problems. Moreover, it illustrates that the size of the level set associated with the local KL condition depends on the outer variable x and shrinks as x approaches a stationary point of Ψ .

Example 1. Consider the problem

$$\min_{1 \leq x \leq 2} \max_{-1 \leq y \leq 1} -(1 - y^2)^2 + x(1 - y^2)^3 + \frac{1}{2}(x - 1)^2 - x. \quad (9)$$

²The assumption $\text{aff}(\mathcal{X}) = \mathbb{R}^n$ 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 $\text{aff}(\mathcal{X})$.

Observe that problem (9) is a special case of (1) with $f(x, y) = -(1 - y^2)^2 + x(1 - y^2)^3$, $p(x) = (x - 1)^2/2 - x + \delta_{[1,2]}(x)$, and $q(y) = \delta_{[-1,1]}(y)$. It follows from the definition of F in (4) that $F(x, y) = -(1 - y^2)^2 + x(1 - y^2)^3 - \delta_{[-1,1]}(y)$. By this and the definition of F^* in (5), one has

$$F^*(x) = \max_{-1 \leq y \leq 1} -(1 - y^2)^2 + x(1 - y^2)^3 = \max_{0 \leq t \leq 1} -(1 - t)^2 + x(1 - t)^3 = \max\{x - 1, 0\}. \quad (10)$$

Using this, (6), and the expression of p , we have $\Psi(x) = F^*(x) + p(x) = (x - 1)^2/2 - 1 + \delta_{[1,2]}(x)$, which implies that

$$\text{dist}(0, \partial\Psi(x)) = x - 1 \quad \forall x \in [1, 2]. \quad (11)$$

We first show that the local KL condition fails to hold for the inner maximization problem of (9) on any constant level set. That is, the KL inequality (7) fails to hold for any $C > 0$, $\theta \in (0, 1)$, $\gamma > 0$, and $\sigma = 0$. To this end, suppose $\sigma = 0$ and fix any $\gamma > 0$. It follows from (8) that $\mathcal{L}(x) = \{y : 0 < F^*(x) - F(x, y) \leq \gamma\}$. Let $\bar{x} = \min\{1 + \gamma, 2\}$. Notice that $F^*(\bar{x}) - F(\bar{x}, 1) = \min\{\gamma, 1\}$ and hence $1 \in \mathcal{L}(\bar{x})$. In addition, one can verify that $\text{dist}(0, \partial_y F(\bar{x}, 1)) = 0$. Combining these, we see that the KL inequality (7) fails to hold for the inner maximization problem of (9) at \bar{x} and $y = 1$ for any $C > 0$ and $\theta \in (0, 1)$. Consequently, the global KL condition also fails.

We next show that the local KL condition holds for the inner maximization problem of (9) on a variable level set, particularly, the KL inequality (7) holds with $C = 1/2$, $\theta = 1/2$, $\gamma = 1/2$, and $\sigma = 1$. To this end, fix any $x \in (1, 2]$ and $y \in \mathcal{L}(x)$, where $\mathcal{L}(\cdot)$ is given in (8) with $\gamma = 1/2$ and $\sigma = 1$. For convenience, let $t = 1 - y^2$. It then follows from $y \in \mathcal{L}(x)$, (10), and (11) that

$$x - 1 + t^2 - xt^3 = x - 1 + (1 - y^2)^2 - x(1 - y^2)^3 = F^*(x) - F(x, y) \leq (x - 1)/2,$$

which along with $x \in (1, 2]$ implies that $t^2(xt - 1) = xt^3 - t^2 \geq (x - 1)/2 > 0$, and hence $t > 1/x \geq 1/2$. Using $t > 1/2$ and $t = 1 - y^2$, we have $|y| < \sqrt{2}/2$, which along with the expression of F yields $\partial_y F(x, y) = 2yt(2 - 3xt)$. In addition, since $t > 1/x$, one has $2 - 3xt < -1$. Using these and $t > 1/2$, we obtain that $\text{dist}(0, \partial_y F(x, y)) = 2|y||t||2 - 3xt| \geq |y|$. On the other hand, by (10), $t = 1 - y^2$, $x \leq 2$, $|y| < \sqrt{2}/2$, $t \in (1/2, 1]$, and the expression of F , we have

$$F^*(x) - F(x, y) = x - 1 + t^2 - xt^3 = (1 - t)(x(1 + t + t^2) - 1 - t) = y^2(x(1 + t + t^2) - 1 - t) \leq y^2(1 + t + 2t^2) \leq 4y^2,$$

where the inequalities follow from $x \leq 2$ and $t \in (1/2, 1]$. Hence, we can see that

$$C(F^*(x) - F(x, y))^\theta \leq \text{dist}(0, \partial_y F(x, y))$$

holds with $C = 1/2$ and $\theta = 1/2$.

2 Local generalized Hölder smoothness of the maximal function

In this section, we establish a local generalized Hölder smoothness property of the maximal function F^* , which will play a crucial role in the development of a first-order method for solving problem (1).

As our goal is to develop a first-order method for computing an $\mathcal{O}(\epsilon)$ -stationary point of problem (1), it is important to characterize the behavior of the objective function Ψ over the following subset of nonstationary points:

$$\mathcal{U}_\epsilon := \{x \in \mathcal{X} : \text{dist}(0, \partial\Psi(x)) > \epsilon\} \quad \forall \epsilon > 0. \quad (12)$$

Given that p is a simple component of Ψ , it suffices to study the behavior of the more sophisticated component F^* on \mathcal{U}_ϵ .

For a special case of problem (1), where $q = 0$ and the inner maximization problem of (1) satisfies a *global* PL condition (i.e., a global KL condition with exponent $1/2$), the work [24] shows that the maximal function F^* is globally Lipschitz smooth. The following theorem extends this result to a more general setting, in which q is a possibly nonsmooth convex function and the inner maximization problem of (1) satisfies only a *local* KL condition as described in Assumption 1(iii). Specifically, it establishes

that the maximal function F^* is locally generalized Hölder smooth on \mathcal{U}_ϵ . This property will play a key role in the development of an inexact proximal gradient method for solving problem (1). The proof of this result, which relies on an error bound for $F(x, \cdot)$, is deferred to Subsection 6.1.

Theorem 1. *Let $\epsilon > 0$ be given and \mathcal{U}_ϵ be defined in (12). Suppose that Assumption 1 holds. Then the following statements hold.*

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

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

$$\|\nabla F^*(x) - \nabla F^*(x')\| \leq L_{\nabla f} \|x - x'\| + (1 - \theta)^{-1} C^{-1/\theta} L_{\nabla f}^{1/\theta} \|x - x'\|^{1-\theta}. \quad (13)$$

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

$$\|\nabla_{\mathcal{X}}^{\text{C}} F^*(x) - \nabla_{\mathcal{X}}^{\text{C}} F^*(x')\| \leq L_{\nabla f} \|x - x'\| + (1 - \theta)^{-1} C^{-1/\theta} L_{\nabla f}^{1/\theta} \|x - x'\|^{1-\theta}. \quad (14)$$

(iv) It holds that

$$\nabla_{\mathcal{X}}^{\text{C}} F^*(x) = \nabla_x f(x, y^*) \quad \forall x \in \mathcal{U}_\epsilon, y^* \in Y^*(x). \quad (15)$$

The following result is a consequence of Theorem 1, whose proof is deferred to Subsection 6.1.

Corollary 1. *Let $\epsilon > 0$ be given and \mathcal{U}_ϵ be defined in (12). Suppose that Assumption 1 holds. Then, for any x, x' satisfying $[x, x'] \subseteq \mathcal{U}_\epsilon$ and $\|x - x'\| \leq \gamma\epsilon^\sigma / (4L_f)$, we have*

$$F^*(x) \leq F^*(x') + \langle \nabla_{\mathcal{X}}^{\text{C}} F^*(x'), x - x' \rangle + \frac{1}{2} L_{\nabla f} \|x - x'\|^2 + \frac{M}{1 + \nu} \|x - x'\|^{1+\nu}, \quad (16)$$

where

$$M := (1 - \theta)^{-1} C^{-1/\theta} L_{\nabla f}^{1/\theta}, \quad \nu := \theta^{-1}(1 - \theta). \quad (17)$$

3 A proximal gradient method for minimizing KL function

In this section, we consider a composite optimization problem under a local KL condition:

$$h^* = \min_z \{h(z) := g(z) + q(z)\}, \quad (18)$$

where $q : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is closed and convex, and g is L -smooth on $\text{dom } q$. Additionally, h satisfies the following local KL condition:

$$C(h(z) - h^*)^\theta \leq \text{dist}(0, \partial h(z)) \quad \forall z \text{ with } h^* < h(z) \leq h^* + \delta \quad (19)$$

for some constants $C > 0$, $\theta \in [1/2, 1)$, and $\delta > 0$.

Under a global KL condition (i.e., (19) with $\delta = \infty$), general algorithmic frameworks for solving problem (18) and analyzing their convergence properties have been extensively studied in the literature (see, e.g., [2, 3, 14, 19]). Inspired by these works, we propose a proximal gradient method with backtracking line search for solving problem (18) under the local KL condition (19), and provide a self-contained convergence analysis for completeness and ease of reference. This algorithm will subsequently serve as a subroutine for solving problem (1). Specifically, at each iteration, the method performs multiple proximal gradient steps along with a backtracking line search to ensure sufficient reduction in the objective function h . The method terminates once the change between consecutive iterates becomes sufficiently small. The proposed method is detailed in Algorithm 1.

Algorithm 1 A proximal gradient method for problem (18)

Input: $z^0 \in \{z : h(z) \leq h^* + \delta\}$, $\bar{\lambda} > 0$, $\rho \in (0, 1)$, and $\tau > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: **for** $i = 0, 1, 2, \dots$ **do**
- 3: $\lambda_{k,i} = \bar{\lambda}\rho^i$.
- 4: $z^{k+1,i} = \arg \min_z \left\{ \langle \nabla g(z^k), z \rangle + \frac{1}{2\lambda_{k,i}} \|z - z^k\|^2 + q(z) \right\}$.
- 5: **if** $h(z^{k+1,i}) + \frac{1}{2\lambda_{k,i}} \|z^{k+1,i} - z^k\|^2 \leq h(z^k)$ **then**
- 6: $z^{k+1} = z^{k+1,i}$, $\lambda_k = \lambda_{k,i}$.
- 7: **break**
- 8: **end if**
- 9: **end for**
- 10: **if** $\|z^{k+1} - z^k\| \leq \tau$ **then**
- 11: **return** z^{k+1} .
- 12: **end if**
- 13: **end for**

The following result establishes bounds on λ_k and on the number of inner iterations performed during each outer iteration k . As a consequence, it justifies the well-definedness of Algorithm 1. The proof of this result is deferred to Subsection 6.2.

Theorem 2. *Let L be the Lipschitz smoothness constant of g , $\bar{\lambda}, \rho$ be given in Algorithm 1, and*

$$\bar{i} = \left\lceil \frac{\log(L\bar{\lambda})}{\log \rho^{-1}} \right\rceil_+.$$

Then it holds that the number of inner iterations of Algorithm 1 at each outer iteration k is at most $\bar{i} + 1$. Moreover,

$$\min\{\rho/L, \bar{\lambda}\} \leq \lambda_k \leq \bar{\lambda}. \quad (20)$$

The following theorem establishes that Algorithm 1 terminates in a finite number of iterations and yields a desired approximate solution to problem (18). The proof is deferred to Subsection 6.2.

Theorem 3. *Let C, δ, θ be given in (19), $\bar{\lambda}, \rho, \tau$ be given in Algorithm 1, and let*

$$\underline{\lambda} = \min\{\rho/L, \bar{\lambda}\}, \quad \underline{\beta} = \frac{C^2}{2\bar{\lambda}}(L + \underline{\lambda}^{-1})^{-2}, \quad \bar{\beta} = \frac{C^2}{2\bar{\lambda}}(L + \bar{\lambda}^{-1})^{-2}, \quad (21)$$

$$C' = \min \left\{ \frac{1}{2}, \frac{(2^{\frac{2\theta-1}{2\theta}} - 1)\delta^{1-2\theta}}{(2\theta-1)\bar{\beta}} \right\}, \quad \bar{K}_\theta := \begin{cases} \left\lceil \frac{1+\bar{\beta}}{\bar{\beta}} \log(2\bar{\lambda}\delta\tau^{-2}) \right\rceil_+ + 1 & \text{if } \theta = \frac{1}{2}, \\ \left\lceil \frac{1}{C'(2\theta-1)\bar{\beta}} (2\bar{\lambda}\tau^{-2})^{2\theta-1} \right\rceil_+ + 1 & \text{if } \theta \in (\frac{1}{2}, 1). \end{cases} \quad (22)$$

Then Algorithm 1 terminates in at most \bar{K}_θ iterations, and outputs a point z^{k+1} satisfying $\|z^{k+1} - z^k\| \leq \tau$ for some $k < \bar{K}_\theta$. Moreover, it holds that

$$h(z^{k+1}) - h^* \leq (C^{-1}(L + \underline{\lambda}^{-1})\tau)^{\frac{1}{\theta}}. \quad (23)$$

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 a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of (1) for $\epsilon > 0$.

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

$$\mathcal{X}_\epsilon := \{x \in \mathcal{X} : \text{dist}(0, \partial\Psi(x)) \leq \epsilon\}, \quad \mathcal{X}_\epsilon^c := \{x \in \mathcal{X} : \text{dist}(x, \mathcal{X}_\epsilon) > \gamma\epsilon^\sigma/(4L_f)\}, \quad r := \gamma\epsilon^\sigma/(4L_f), \quad (24)$$

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 some key observations. Suppose $x' \in \mathcal{X}_\epsilon^c$, that is, x' is not an (r, ϵ) -stationary point of (1). Given any $x \in \mathcal{X} \cap \mathcal{B}(x', r)$, we observe that $[x', x] \subseteq \mathcal{X}$ and moreover $\text{dist}(0, \partial\Psi(z)) > \epsilon$ for all $z \in [x', x]$. In view of these, one can see that $[x', x] \subseteq \mathcal{U}_\epsilon$, where \mathcal{U}_ϵ is defined in (12). It follows from Corollary 1 that (16) holds for such x and x' . In addition, notice from $\theta \in [1/2, 1)$ and (17) that $\nu \in (0, 1]$. By this and [23, Lemma 2], one has

$$M(1 + \nu)^{-1} \|x - x'\|^{1+\nu} \leq (\delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}} \|x - x'\|^2 + \delta) / 2 \quad \forall \delta > 0.$$

Combining this inequality with (16), and using the fact $\Psi(\cdot) = F^*(\cdot) + p(\cdot)$, we obtain that

$$\begin{aligned} F^*(x) &\leq F^*(x') + \langle \nabla_{\mathcal{X}}^C F^*(x'), x - x' \rangle + \frac{1}{2} (L_{\nabla f} + \delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}) \|x - x'\|^2 + \frac{\delta}{2} \quad \forall x \in \mathcal{X} \cap \mathcal{B}(x', r), \quad (25) \\ \Psi(x) &\leq F^*(x') + \langle \nabla_{\mathcal{X}}^C F^*(x'), x - x' \rangle + \frac{1}{2} (L_{\nabla f} + \delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}) \|x - x'\|^2 + p(x) + \frac{\delta}{2} \quad \forall x \in \mathcal{X} \cap \mathcal{B}(x', r). \end{aligned}$$

As a result, 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 (PG) type-method for finding an (r, ϵ) -stationary point of problem (1). The method generates the sequence $\{x^k\}$ according to

$$x^{k+1} = \arg \min_{x \in \mathcal{B}(x^k, r)} \left\{ \langle \nabla_{\mathcal{X}}^C F^*(x^k), x \rangle + \frac{1}{2} L_k \|x - x^k\|^2 + p(x) \right\} \quad (26)$$

with $L_k = L_{\nabla f} + \delta_k^{(\nu-1)/(1+\nu)} M^{2/(1+\nu)}$ for a suitable choice of $\delta_k > 0$, and terminates when x^k is an (r, ϵ) -stationary point of (1) for some $k \geq 0$. However, this method faces a practical limitation: the exact value of $\nabla_{\mathcal{X}}^C F^*(x^k)$ is typically unavailable, since F^* is a maximal function.

To address this issue, we propose an inexact PG method for solving problem (1). Specifically, we replace $\nabla_{\mathcal{X}}^C F^*(x^k)$ in (26) with its approximation $\nabla_x f(x^k, y^k)$. Here, y^k is an approximate solution to the subproblem $\max_y \{f(x^k, y) - q(y)\}$, or equivalently, $\min_y \{-f(x^k, y) + q(y)\}$, obtained via Algorithm 1 (see lines 4 and 5 of Algorithm 2).

We now present an inexact PG method for solving problem (1).

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

Input: $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma$ from Assumption 1; $\epsilon > 0, \bar{\lambda} > 0, \rho \in (0, 1)$, and $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ satisfying $F^*(x^0) - F(x^0, y^0) \leq \min\{\gamma\epsilon^\sigma/2, 1\}$.

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

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

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

4: Compute

$$x^{k+1} = \arg \min_{x \in \mathcal{B}(x^k, r)} \left\{ \langle \nabla_x f(x^k, y^k), x \rangle + \frac{L_k}{2} \|x - x^k\|^2 + p(x) \right\}. \quad (27)$$

5: Call Algorithm 1 with $g(\cdot) \leftarrow -f(x^{k+1}, \cdot)$, $q(\cdot) \leftarrow q(\cdot)$, $\bar{\lambda} \leftarrow \bar{\lambda}$, $\rho \leftarrow \rho$, $z^0 \leftarrow y^k$, $\tau \leftarrow \frac{C}{L_{\nabla f} + \underline{\lambda}^{-1}} \min \left\{ (\frac{1}{2}\gamma\epsilon^\sigma)^\theta, \eta_{k+1}^{\frac{\theta}{2(1-\theta)}} \right\}$, and denote its output as y^{k+1} .

6: **end for**

Remark 2. (i) As for the initial point (x^0, y^0) , Algorithm 2 requires that y^0 be a nearly optimal solution to the problem $\max_y F(x^0, y)$. Such a pair (x^0, y^0) can often be found in practice. For example, if $x^0 \in \mathcal{X}$ is such that $F(x^0, y)$ is concave in y , or satisfies a global or local KL condition over a known neighborhood, then y^0 can be obtained by applying a first-order method to the problem $\max_y F(x^0, y)$.

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

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

Theorem 4. Let $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma$ be given in Assumption 1, M, ν be defined in (17), ϵ be given in Algorithm 2, and

$$A = (1 - \theta)^{-2} C^{-2} L_{\nabla f}^2, \quad \underline{L} = L_{\nabla f} + M^{2/(1+\nu)}, \quad (28)$$

$$a = 8(\Psi(x^0) - \Psi^* + 3 + 2A\underline{L}^{-1}), \quad b = 8(3/2 + A\underline{L}^{-1}), \quad (29)$$

$$\widehat{C}_1 = \left(36(1 + \nu)\nu^{-1}b\underline{L} \left[\log(18(1 + \nu)\nu^{-1}b\underline{L}) \right]_+ + 72(1 + \nu)\nu^{-1}b\underline{L} + 1 \right)^{\frac{1+\nu}{2\nu}}, \quad (30)$$

$$\widehat{C}_2 = \left(\frac{4b(1 + \nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}} \left[\log \left(\frac{2b(1 + \nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}} \right) \right]_+ + \frac{8b(1 + \nu)(3M)^{2/\nu}}{\nu M^{2/(1+\nu)}} + 1 \right)^{\frac{1+\nu}{2}}, \quad (31)$$

$$\widehat{C}_3 = (36\underline{L}a)^{\frac{1+\nu}{2\nu}} + M^{-1}(4a(3M)^{2/\nu})^{\frac{1+\nu}{2}}, \quad \widehat{C}_4 = 72A, \quad (32)$$

$$\widehat{C}_5 = \left(\frac{144(1 + \nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}} \left[\log \left(\frac{72(1 + \nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}} \right) \right]_+ + \frac{288(1 + \nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}} + 1 \right)^{\frac{1+\nu}{2}}, \quad (33)$$

$$\widehat{C}_6 = (144aL_{\nabla f}^2)^{\frac{1+\nu}{2}} / M, \quad (34)$$

$$\widehat{C}_7 = \left(\frac{64(1 + \nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}} \left[\log \left(\frac{32(1 + \nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}} \right) \right]_+ + \frac{128\sigma(1 + \nu)bL_f^2}{\gamma^2 M^{2/(1+\nu)}} + 1 \right)^{\frac{1+\nu}{2}}, \quad (35)$$

$$\widehat{C}_8 = (64aL_f^2)^{\frac{1+\nu}{2}} / (\gamma^{1+\nu} M), \quad (36)$$

$$\begin{aligned} \widehat{K}_\epsilon = & \left[\widehat{C}_1 \epsilon^{-\frac{1+\nu}{2\nu}} (\log \epsilon^{-1})^{\frac{1+\nu}{2\nu}} + \widehat{C}_2 \epsilon^{-\frac{1+\nu}{2\nu}} (\log \epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_3 \epsilon^{-\frac{1+\nu}{2\nu}} + \widehat{C}_4 \epsilon^{-2} \right. \\ & \left. + \widehat{C}_5 \epsilon^{-(1+\nu)} (\log \epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_6 \epsilon^{-(1+\nu)} + \widehat{C}_7 \epsilon^{-(1+\nu)\sigma} (\log \epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_8 \epsilon^{-(1+\nu)\sigma} \right]. \end{aligned} \quad (37)$$

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

$$F^*(x^k) - F(x^k, y^k) \leq \min \left\{ \frac{\gamma\epsilon^\sigma}{2}, \frac{1}{k+1} \right\}, \quad \text{dist}(y^k, Y^*(x^k)) \leq \frac{1}{C(1-\theta)} \min \left\{ \left(\frac{\gamma}{2} \right)^{(1-\theta)} \epsilon^{\sigma(1-\theta)}, \frac{1}{\sqrt{k+1}} \right\}. \quad (38)$$

Remark 3. In view of Theorem 4, Algorithm 2 enjoys an iteration complexity of

$$\widetilde{O} \left(\max \left\{ \epsilon^{-\frac{1}{1-\theta}}, (L_f/\gamma)^{\frac{1}{\theta}} \epsilon^{-\frac{\sigma}{\theta}} \right\} \right)$$

for computing a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1).

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 a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1) for any $\epsilon \in (0, 1/e]$. The proof is deferred to Subsection 6.3.

Theorem 5. Let $\epsilon \in (0, 1/e]$ be given, \widehat{K}_ϵ be defined in Theorem 4, $L_{\nabla f}, C, \theta, \gamma, \sigma$ be given in Assump-

tion 1, M, ν be defined in (17), $\rho, \bar{\lambda}, \underline{\lambda}$ be given in Algorithm 2, and let

$$\begin{aligned}\underline{\beta}_f &= \frac{C^2}{2\bar{\lambda}}(L_{\nabla f} + \underline{\lambda}^{-1})^{-2}, \quad \bar{\beta}_f = \frac{C^2}{2\lambda}(L_{\nabla f} + \bar{\lambda}^{-1})^{-2}, \\ C'_f &= \min \left\{ \frac{1}{2}, \frac{(2^{\frac{2\theta-1}{2\theta}} - 1)(\gamma\epsilon^\sigma)^{1-2\theta}}{(2\theta-1)\bar{\beta}_f} \right\}, \quad \Lambda = \max \left\{ (\frac{1}{2}\gamma\epsilon^\sigma)^{-2\theta}, (\widehat{K}_\epsilon + 1)^{\frac{\theta}{1-\theta}} \right\}, \\ \bar{K}_{f,\theta} &= \begin{cases} \left\lceil \frac{1+\bar{\beta}_f}{\bar{\beta}_f} \log(2\bar{\lambda}C^{-2}(L_{\nabla f} + \underline{\lambda}^{-1})^2\gamma\epsilon^\sigma\Lambda) \right\rceil + 1 & \text{if } \theta = \frac{1}{2}, \\ \left\lceil \frac{1}{C'_f(2\theta-1)\bar{\beta}_f} (2\bar{\lambda}C^{-2}(L_{\nabla f} + \underline{\lambda}^{-1})^2\Lambda)^{2\theta-1} \right\rceil + 1 & \text{if } \theta \in (\frac{1}{2}, 1), \end{cases} \\ \widehat{N}_\epsilon &= \widehat{K}_\epsilon \left(\left\lceil \frac{\log(2L_{\nabla f}\bar{\lambda})}{\log \rho^{-1}} \right\rceil + 1 \right) \bar{K}_{f,\theta}. \end{aligned} \quad (39)$$

$$\widehat{N}_\epsilon = \widehat{K}_\epsilon \left(\left\lceil \frac{\log(2L_{\nabla f}\bar{\lambda})}{\log \rho^{-1}} \right\rceil + 1 \right) \bar{K}_{f,\theta}. \quad (40)$$

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}_ϵ , \widehat{N}_ϵ , and $\widehat{K}_\epsilon + \widehat{N}_\epsilon$, respectively, to generate a pair (x^k, y^k) such that x^k is a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1), and y^k satisfies (38).

Remark 4. From Theorem 5, we observe that, in order to compute a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1), Algorithm 2 requires $\widetilde{\mathcal{O}}(\max\{\epsilon^{-\frac{1}{1-\theta}}, (L_f/\gamma)^{\frac{1}{\theta}}\epsilon^{-\frac{\sigma}{\theta}}\})$ evaluations of the proximal operator of p . Moreover, the numbers of evaluations of the proximal operator of q and the gradient ∇f are given as follows:

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

5 Numerical results

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

Consider the following minimax optimization problem:

$$\min_{\|x\| \leq 1} \max_{\|y\|_\infty \leq 2} \{0.01\|x\|_1 - \|(y + Ax) \odot (y + Bx)\|^2 + 0.01\|x - c\|^2 - 0.1\|y\|_1\}, \quad (41)$$

where $A, B \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, and \odot denotes the Hadamard (elementwise) product.

For each pair (m, n) , we randomly generate 10 instances of problem (41) by sampling the entries of A , B , and c independently from the standard normal distribution $\mathcal{N}(0, 1)$. Note that problem (41) is a special case of problem (1) with $f(x, y) = -\|(y + Ax) \odot (y + Bx)\|^2 + 0.01\|x - c\|^2$, $p(x) = 0.01\|x\|_1 + \mathcal{I}_{\mathcal{B}(0,1)}(x)$, and $q(y) = 0.1\|y\|_1 + \mathcal{I}_{[-2,2]^m}(y)$, where $\mathcal{I}_{\mathcal{B}(0,1)}$ and $\mathcal{I}_{[-2,2]^m}$ denote the indicator functions of the unit Euclidean ball $\mathcal{B}(0, 1)$ and the m -dimensional box $[-2, 2]^m$, respectively.

In order to apply Algorithm 2 to solve problem (41), we need to estimate the Lipschitz constant L_f of $f(\cdot, y)$ and the Lipschitz constant $L_{\nabla f}$ of ∇f over the set $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathcal{B}(0, 1)$ and $\mathcal{Y} = [-2, 2]^m$. To this end, let $(a^i)^T$ and $(b^i)^T$ denote the i th row vectors of A and B , respectively, and define $u = y + Ax$, $v = y + Bx$, and $w = u \odot v$. Then we obtain that $f(x, y) = -\sum_{i=1}^m w_i^2 + 0.01\|x - c\|^2$, and

$$\begin{aligned}\nabla_x f(x, y) &= -2 \sum_{i=1}^m w_i (v_i a^i + u_i b^i) + 0.02(x - c), \quad \nabla_y f(x, y) = -2(u + v) \odot w, \\ \nabla_{xx}^2 f(x, y) &= -2 \sum_{i=1}^m [(v_i a^i + u_i b^i)(v_i a^i + u_i b^i)^T + w_i (a^i (b^i)^T + b^i (a^i)^T)] + 0.02I, \\ \nabla_{xy}^2 f(x, y) &= -2[A^T \text{diag}(v^2 + 2u \odot v) + B^T \text{diag}(u^2 + 2u \odot v)], \\ \nabla_{yy}^2 f(x, y) &= -2 \text{diag}((u + v)^2 + 2u \odot v), \end{aligned} \quad (42)$$

where $z^2 := z \odot z$ for any vector z . Let $M_a = \max_i \|a^i\|$, $M_b = \max_i \|b^i\|$, and

$$\begin{aligned} L_f &= 4m(M_a M_b + 2M_a + 2M_b + 4)(M_a M_b + M_a + M_b) + 0.02(1 + \|c\|), \\ L_{\nabla f} &= 4m[2(M_a M_b + M_a + M_b)^2 + M_a M_b(M_a M_b + 2M_a + 2M_b + 4)] \\ &\quad + 2[\|A\|(M_b + 2)(2M_a + M_b + 6) + \|B\|(M_a + 2)(M_a + 2M_b + 6)] \\ &\quad + 2[(M_a + M_b + 4)^2 + 2(M_a + 2)(M_b + 2)] + 0.02, \end{aligned} \tag{43}$$

where $\|A\|$ and $\|B\|$ denote the spectral norms of A and B , respectively. In view of (42) and (43), one can verify that $L_f \geq \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \|\nabla_x f(x, y)\|$, and

$$L_{\nabla f} \geq \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \{ \|\nabla_{xx}^2 f(x, y)\| + \|\nabla_{xy}^2 f(x, y)\| + \|\nabla_{yy}^2 f(x, y)\| \} \geq \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \|\nabla^2 f(x, y)\|.$$

It then follows that $f(\cdot, y)$ is L_f -Lipschitz continuous on \mathcal{X} for any fixed $y \in \mathcal{Y}$, and ∇f is $L_{\nabla f}$ -Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$.

We now apply Algorithm 2 to solve problem (41) on the randomly generated instances described above. The parameters L_f and $L_{\nabla f}$ are computed using (43), while the remaining parameters are set as follows: $C = 0.2$, $\theta = 0.5$, $\gamma = 0.01$, $\sigma = 0.1$, $\bar{\lambda} = 1$, $\rho = 0.95$, and $\epsilon = 10^{-2}$.³ The algorithm is initialized at $(x^0, y^0) = (0, 0)$. Note that for this initialization, $y^0 = \arg \max_{\|y\|_\infty \leq 2} \{f(x^0, y) - 0.1\|y\|_1\}$, making it a suitable starting point for y . We run the algorithm for 10,000 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 in (41), while y_ϵ is an approximate solution to the inner maximization problem $\max_{\|y\|_\infty \leq 2} \{f(x_\epsilon, y) - 0.1\|y\|_1\}$.

To evaluate the performance of Algorithm 2, we compute the final objective value of problem (41) by

$$\Psi(x_\epsilon) = \max_{\|y\|_\infty \leq 2} \{f(x_\epsilon, y) - 0.1\|y\|_1\} + 0.01\|x_\epsilon\|_1.$$

Thanks to the separable structure of the problem, this maximization problem can be decomposed into m independent scalar subproblems. Each subproblem is solved using the MATLAB subroutine `GlobalSearch`, which is a solver for finding global optima of nonconvex problems.

In addition, we compute an estimate of the 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 m and n . For each pair (m, n) , the remaining columns report the averages of the initial objective value, the final objective value $\Psi(x_\epsilon)$, and the estimated final objective value $\widehat{\Psi}(x_\epsilon)$ over 10 random instances. 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 $\max_{\|y\|_\infty \leq 2} \{f(x_\epsilon, y) - 0.1\|y\|_1\}$.

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 Theorems 1-5.

³Observe that the inner maximization problem in (41) admits a decomposition into m independent one-dimensional minimization subproblems, each of which has an objective function of the form $(z - a)^2(z - b)^2 + 0.1|z| + \delta_{[-2, 2]}(z)$ for some $a, b \in \mathbb{R}$. Numerical evidence suggests that such functions satisfy a local KL condition with exponent $\theta = 1/2$. The other parameters were chosen empirically and yield reasonable performance in our experiments.

Table 1: Numerical results for Algorithm 2

n	m	Initial objective value	Final objective value	Estimated final objective value
100	100	1.03	-224.55	-224.87
100	200	0.98	-228.22	-228.69
100	300	1.05	-260.45	-261.29
200	100	1.95	-808.08	-808.45
200	200	2.03	-816.54	-817.28
200	300	1.95	-837.63	-838.33
300	100	2.95	-1102.26	-1102.51
300	200	2.90	-1082.37	-1082.71
300	300	3.03	-1022.22	-1022.83

6.1 Proof of the main results in Section 2

In this subsection we prove Theorem 1 and Corollary 1. To proceed, we first establish several technical lemmas below.

The following lemma concerns the Lipschitz continuity of F^* on \mathcal{X} .

Lemma 1. *Suppose that Assumption 1 holds. Then F^* is L_f -Lipschitz continuous on \mathcal{X} .*

Proof. Fix any $x, x' \in \mathcal{X}$. Recall from Assumption 1 that $f(\cdot, y)$ is L_f -Lipschitz continuous on \mathcal{X} for any $y \in \mathcal{Y}$. Using this and the expression of F in (4), we have

$$F(x, y) - F(x', y) = f(x, y) - f(x', y) \leq L_f \|x - x'\| \quad \forall y \in \mathcal{Y}.$$

This together with the definition of F^* in (5) implies that

$$F^*(x) \stackrel{(5)}{=} \max_{y \in \mathcal{Y}} F(x, y) \leq \max_{y \in \mathcal{Y}} F(x', y) + L_f \|x - x'\| = F^*(x') + L_f \|x - x'\|,$$

and hence $F^*(x) - F^*(x') \leq L_f \|x - x'\|$. Similarly, one can show that $F^*(x') - F^*(x) \leq L_f \|x' - x\|$. It then follows that $|F^*(x) - F^*(x')| \leq L_f \|x - x'\|$. By this and the arbitrariness of $x, x' \in \mathcal{X}$, we conclude that F^* is L_f -Lipschitz continuous on \mathcal{X} . \square

The next result provides a formula for $\nabla F^*(x)$ at a point x where F^* is differentiable.

Lemma 2. *Suppose that Assumption 1 holds and F^* is differentiable at some $x \in \mathbb{R}^n$. Then $\nabla F^*(x) = \nabla_x f(x, y)$ for all $y \in Y^*(x)$.*

Proof. Fix any $y \in Y^*(x)$ and $d \in \mathbb{R}^n$. Observe from (5) that $F^*(x) = F(x, y)$ and $F^*(x + td) \geq F(x + td, y)$ for any $t \in \mathbb{R}$. By these, the differentiability of F^* at x , and the expression of F , one has

$$\begin{aligned} \langle \nabla F^*(x), d \rangle &= \lim_{t \downarrow 0} \frac{F^*(x + td) - F^*(x)}{t} \geq \lim_{t \downarrow 0} \frac{F(x + td, y) - F(x, y)}{t} \\ &= \lim_{t \downarrow 0} \frac{f(x + td, y) - f(x, y)}{t} = \langle \nabla_x f(x, y), d \rangle. \end{aligned}$$

Using this and the arbitrariness of d , we conclude that $\nabla F^*(x) = \nabla_x f(x, y)$. \square

The following lemma establishes that if F^* is Hölder smooth almost everywhere on a relatively open subset of \mathcal{X} , then its differentiability extends to the entire subset.

Lemma 3. Let $\Gamma \subseteq \mathcal{X}$ be relatively open in \mathcal{X} , $\mathcal{S} = \{x \in \mathcal{X} : F^* \text{ is differentiable at } x\}$, and $\partial_{\mathcal{X}}^{\mathcal{C}} F^*$ be defined as in (3). Suppose that Assumption 1 holds, and there exist constants $c_1, c_2 > 0$, $\alpha > 0$, and $\eta > 0$ such that

$$\|\nabla F^*(u) - \nabla F^*(v)\| \leq c_1 \|u - v\| + c_2 \|u - v\|^\alpha \quad \forall u, v \in \mathcal{S} \cap \Gamma \text{ with } \|u - v\| \leq \eta. \quad (44)$$

Then $\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x)$ is a singleton for every $x \in \Gamma$. Moreover, F^* is differentiable on $\Gamma \cap \text{int}(\mathcal{X})$.

Proof. Fix any $x \in \Gamma$. We first claim that there exists at least one sequence in \mathcal{S} converging to x . Indeed, since \mathcal{X} is a convex set and $\text{aff}(\mathcal{X}) = \mathbb{R}^n$, one has $\text{int}(\mathcal{X}) \neq \emptyset$. Let us fix any $x' \in \text{int}(\mathcal{X})$. Then, by $x \in \mathcal{X}$ and the convexity of \mathcal{X} , it follows from [4, Proposition 1.3.1] that $(x, x') \subset \text{int}(\mathcal{X})$. Consequently, there exists a sequence $\{\hat{x}^k\} \subset \text{int}(\mathcal{X})$ such that $\hat{x}^k \rightarrow x$. This implies that for each k , there exists some $r_k \in (0, 1/k]$ such that $\mathcal{B}(\hat{x}^k, r_k) \subset \text{int}(\mathcal{X})$. In addition, by Lemma 1, F^* is Lipschitz continuous on the open set $\text{int}(\mathcal{X})$, which implies that $\text{int}(\mathcal{X}) \setminus \mathcal{S}$ has measure zero due to Rademacher's theorem. In view of these, one can see that there exists some $\hat{z}^k \in \mathcal{B}(\hat{x}^k, r_k) \cap \mathcal{S}$ for every k . Hence, we obtain

$$\|\hat{z}^k - x\| \leq \|\hat{x}^k - x\| + \|\hat{z}^k - \hat{x}^k\| \leq \|\hat{x}^k - x\| + r_k \leq \|\hat{x}^k - x\| + 1/k \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

It follows that $\{\hat{z}^k\} \subset \mathcal{S}$ and $\hat{z}^k \rightarrow x$, which proves the claim.

Now let $\{x^k\} \subset \mathcal{S}$ be an arbitrary sequence such that $x^k \rightarrow x$. Since Γ is relatively open in \mathcal{X} and $x \in \Gamma$, it follows that $x^k \in \mathcal{S} \cap \Gamma$ for all sufficiently large k . Hence, without loss of generality, we may assume that $\{x^k\} \subset \mathcal{S} \cap \Gamma$. We now claim that $\{\nabla F^*(x^k)\}$ converges. Indeed, since $\{x^k\}$ converges, it is a Cauchy sequence. Hence, there exists K such that $\|x^k - x^{k'}\| \leq \eta$ for all $k, k' \geq K$. By (44), one then has

$$\|\nabla F^*(x^k) - \nabla F^*(x^{k'})\| \leq c_1 \|x^k - x^{k'}\| + c_2 \|x^k - x^{k'}\|^\alpha \quad \forall k, k' \geq K,$$

which implies that $\{\nabla F^*(x^k)\}$ is also a Cauchy sequence and hence converges as claimed. Next, we show that the limit of $\{\nabla F^*(x^k)\}$ is independent of the choice of sequence. To this end, let $\{\tilde{x}^k\} \subset \mathcal{S}$ be another sequence such that $\{\tilde{x}^k\} \rightarrow x$. Interleaving $\{x^k\}$ and $\{\tilde{x}^k\}$, we obtain a sequence $\{z^k\} \subset \mathcal{S}$ such that $z^k \rightarrow x$. It then follows from the above claim that $\{\nabla F^*(z^k)\}$ converges. Since both $\{\nabla F^*(x^k)\}$ and $\{\nabla F^*(\tilde{x}^k)\}$ are subsequences of $\{\nabla F^*(z^k)\}$, they must share the same limit. Hence, the limit of $\{\nabla F^*(x^k)\}$ is independent of the sequence chosen. In view of these and the definition of $\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x)$ in (3), one can see that $\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x)$ is a singleton for any $x \in \Gamma$.

Next, let us fix any $x \in \Gamma \cap \text{int}(\mathcal{X})$. By this and (3), one can observe that

$$\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x) = \text{conv}\{v : \exists x^k \rightarrow x \text{ such that } \nabla F^*(x^k) \rightarrow v\}. \quad (45)$$

Recall from Lemma 1 that F^* is Lipschitz continuous on \mathcal{X} . Moreover, note that $\Gamma \cap \text{int}(\mathcal{X}) \subseteq \text{int}(\mathcal{X})$. It follows that the Clarke subdifferential of F^* at x is well defined as in (2). Using this and (45), we obtain that $\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x) = \partial^{\mathcal{C}} F^*(x)$. Notice from above that $\partial_{\mathcal{X}}^{\mathcal{C}} F^*(x)$ is a singleton. Hence, $\partial^{\mathcal{C}} F^*(x)$ is also a singleton. By this and [10, Proposition (1.13)], we conclude that F^* is differentiable at any $x \in \Gamma \cap \text{int}(\mathcal{X})$. \square

The following result establishes a local $(1 - \theta)^{-1}$ -growth property of $F(x, \cdot)$ for any $x \in \mathcal{X}$, which was previously derived in the proof of [13, Theorem 3.7]. Here, we provide an alternative and self-contained proof. Our proof generalizes the one used to derive the global quadratic growth result in [18, Appendix G] for the special case where $F(x, \cdot)$ satisfies the global KL condition with exponent $1/2$ —that is, $F(x, \cdot)$ satisfies (7) with $\theta = 1/2$, and $\mathcal{L}(x)$ replaced by \mathcal{Y} for any $x \in \mathcal{X}$.

Lemma 4. Suppose that Assumption 1 holds. Then it holds that for any $x \in \mathcal{X}$,

$$F^*(x) - F(x, y) \geq (C(1 - \theta))^{\frac{1}{1-\theta}} \text{dist}(y, Y^*(x))^{\frac{1}{1-\theta}} \quad \forall y \in \mathcal{L}(x). \quad (46)$$

Proof. Fix any $x \in \mathcal{X}$ and $y \in \mathcal{L}(x)$. It then follows from the definition of $\mathcal{L}(x)$ in (8) that $y \notin Y^*(x)$.

Recall from Assumption 1 that f is Lipschitz smooth on $\mathcal{X} \times \mathcal{Y}$. It together with the convexity of q and the expression of F implies that $F(x, \cdot)$ is weakly concave on \mathcal{Y} . In addition, since $y \in \mathcal{L}(x)$, one has $y \in \text{dom } F(x, \cdot)$. By these and [8, Theorem 13], there exists a unique absolutely continuous curve $Y : [0, \infty) \rightarrow \mathbb{R}^m$ satisfying

$$Y(0) = y, \quad \dot{Y}(t) \in \partial_y F(x, Y(t)) \quad \text{a.e. } t > 0, \quad (47)$$

$$\frac{d}{dt} F(x, Y(t)) = \|\dot{Y}(t)\|^2 \quad \text{a.e. } (\eta, \infty) \quad (48)$$

for any $\eta > 0$, and moreover, $F(x, Y(\cdot))$ is non-decreasing and continuous on $[0, \infty)$. It follows that $Y(t) \in \mathcal{L}(x)$ for any $t \geq 0$.

Let $r(t) = (F^*(x) - F(x, Y(t)))^{1-\theta}$. By $y \notin Y^*(x)$ and the monotonicity and continuity of $F(x, Y(\cdot))$, one can observe that $r(0) > 0$, and r is non-negative, nonincreasing, and continuous on $[0, \infty)$. We next show that

$$\frac{d}{dt} r(t) \leq -C(1-\theta) \|\dot{Y}(t)\| \quad \text{a.e. } (\eta, \infty) \quad (49)$$

for any $\eta > 0$. To this end, let us fix any $\eta > 0$ and consider two separate cases below.

Case 1) $r(t) > 0$ on $[0, \infty)$. It follows from this, (7), (47), and (48) that

$$\begin{aligned} \frac{d}{dt} r(t) &\stackrel{(48)}{=} -(1-\theta)(F^*(x) - F(x, Y(t)))^{-\theta} \|\dot{Y}(t)\|^2 \stackrel{(7)}{\leq} -C(1-\theta) \text{dist}(0, \partial_y F(x, Y(t)))^{-1} \|\dot{Y}(t)\|^2 \\ &\stackrel{(47)}{\leq} -C(1-\theta) \|\dot{Y}(t)\|^{-1} \|\dot{Y}(t)\|^2 = -C(1-\theta) \|\dot{Y}(t)\| \quad \text{a.e. } (\eta, \infty), \end{aligned}$$

and hence (49) holds as desired.

Case 2) $r(t) = 0$ for some $t > 0$. Since r is continuous on $[0, \infty)$, one has that $t_0 := \min\{t > 0 : r(t) = 0\} > 0$. By this and the nonnegativity and monotonicity of r , we have $r(t) = 0$ and hence $F(x, Y(t)) = F^*(x)$ for all $t \geq t_0$. It then follows from (48) with η replaced by t_0 that $\|\dot{Y}(t)\| = 0$ almost everywhere on (t_0, ∞) . Hence, we obtain

$$\frac{d}{dt} r(t) \leq -C(1-\theta) \|\dot{Y}(t)\| \quad \text{a.e. } (t_0, \infty). \quad (50)$$

It follows that (49) holds if $\eta \geq t_0$. We now assume that $\eta < t_0$. Note that $r(t) > 0$ for all $t \in [\eta, t_0)$. By a similar argument as in Case 1), one can conclude that

$$\frac{d}{dt} r(t) \leq -C(1-\theta) \|\dot{Y}(t)\| \quad \text{a.e. } (\eta, t_0).$$

Combining this with (50), we see that (49) holds in this case as well.

Fix any $T > 0$ and $\delta \in (0, T)$. By (49), the monotonicity of r , and the absolute continuity of $Y(\cdot)$, one has

$$\begin{aligned} r(T) - r(\delta) &\leq \int_{\delta}^T \frac{d}{dt} r(t) dt \stackrel{(49)}{\leq} -C(1-\theta) \int_{\delta}^T \|\dot{Y}(t)\| dt \leq -C(1-\theta) \left\| \int_{\delta}^T \dot{Y}(t) dt \right\| \\ &= -C(1-\theta) \|Y(T) - Y(\delta)\|, \end{aligned}$$

where the first inequality follows from the monotonicity of r (e.g., see [31, Chapter 3, Exercise 16]), and the equality uses the absolute continuity of $Y(\cdot)$. Taking the limit on both sides of the above relation as $\delta \rightarrow 0$, and using $Y(0) = y$ and the continuity of r and $Y(\cdot)$, we obtain

$$r(T) - r(0) \leq -C(1-\theta) \|Y(T) - y\|. \quad (51)$$

We next show that $\lim_{T \rightarrow \infty} r(T) = 0$. It clearly holds if there exists some $t > 0$ such that $r(t) = 0$, due to the nonnegativity and monotonicity of r . We now assume that $r(t) > 0$ for all $t \in [0, \infty)$. By this, (7), (47), (48), and the monotonicity of r , one has that for any $T > 0$ and $\delta \in (0, T)$,

$$\begin{aligned} r(T) - r(\delta) &\leq \int_{\delta}^T \frac{d}{dt} r(t) dt \stackrel{(48)}{=} -(1 - \theta) \int_{\delta}^T (F^*(x) - F(x, Y(t)))^{-\theta} \|\dot{Y}(t)\|^2 dt \\ &\stackrel{(47)}{\leq} -(1 - \theta) \int_{\delta}^T (F^*(x) - F(x, Y(t)))^{-\theta} \text{dist}(0, \partial_y F(x, Y(t)))^2 dt \\ &\stackrel{(7)}{\leq} -C^2(1 - \theta) \int_{\delta}^T (F^*(x) - F(x, Y(t)))^{\theta} dt = -C^2(1 - \theta) \int_{\delta}^T r(t)^{\frac{\theta}{1-\theta}} dt \\ &\leq -C^2(1 - \theta)(T - \delta)r(T)^{\frac{\theta}{1-\theta}}, \end{aligned}$$

where the first and last inequalities follow from the monotonicity of r . This relation and $r(T) > 0$ imply that $-r(\delta) \leq -C^2(1 - \theta)(T - \delta)r(T)^{\frac{\theta}{1-\theta}}$. Taking the limit on both sides of this relation as $\delta \rightarrow 0$, and using the continuity of r , we obtain that $-r(0) \leq -C^2(1 - \theta)Tr(T)^{\frac{\theta}{1-\theta}}$, which yields $r(T) \leq [r(0)/(C^2(1 - \theta)T)]^{\frac{1-\theta}{\theta}}$. This along with $r(T) > 0$ implies that $r(T) \rightarrow 0$ as $T \rightarrow \infty$.

By (51) and the nonnegativity of r , one can observe that the range of $Y(\cdot)$ is bounded. In addition, notice from Assumption 1 that $\text{dom } F(x, \cdot)$ is closed. By these facts, there exists a sequence $\{t_k\} \subset (0, \infty)$ such that $t_k \rightarrow \infty$ and $\{Y(t_k)\}$ converges to some point $y^* \in \text{dom } F(x, \cdot)$. Recall that $\lim_{t \rightarrow \infty} r(t) = 0$, which along with $t_k \rightarrow \infty$ implies that $r(t_k) \rightarrow 0$. It then follows that $\lim_{k \rightarrow \infty} F(x, Y(t_k)) = F^*(x)$. On the other hand, by the upper semicontinuity of $F(x, \cdot)$ and $Y(t_k) \rightarrow y^* \in \text{dom } F(x, \cdot)$, one has $\limsup_{k \rightarrow \infty} F(x, Y(t_k)) \leq F(x, y^*)$. Combining these relations, we conclude that $y^* \in Y^*(x)$. Finally, letting $T = t_k$ in (51), one has $r(t_k) - r(0) \leq -C(1 - \theta)\|Y(t_k) - y\|$. Taking the limit on both sides of this inequality as $k \rightarrow \infty$, and using the fact that $r(t_k) \rightarrow 0$ and $Y(t_k) \rightarrow y^* \in Y^*(x)$, we obtain that

$$r(0) \geq C(1 - \theta)\|y^* - y\| \geq C(1 - \theta) \text{dist}(y, Y^*(x)),$$

which together with the expression of r implies that the conclusion (46) holds. \square

As an immediate consequence of Lemma 4 and Assumption 1, the following lemma establishes an error bound for $F(x, \cdot)$. This result, originally derived in [13, Theorem 3.7], provides a relationship between $\text{dist}(y, Y^*(x))$ and $\text{dist}(0, \partial_y F(x, y))$.

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

$$\text{dist}(y, Y^*(x)) \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} \text{dist}(0, \partial_y F(x, y))^{\frac{1-\theta}{\theta}} \quad \forall y \in \mathcal{L}(x). \quad (52)$$

Proof. The relation (52) follows from (7) and (46). \square

The next result concerns the relative openness of the set \mathcal{U}_ϵ in \mathcal{X} .

Lemma 6. *Let \mathcal{U}_ϵ be defined in (12). Suppose that Assumption 1 holds. Then \mathcal{U}_ϵ is relatively open in \mathcal{X} for any $\epsilon > 0$.*

Proof. Fix any $\epsilon > 0$. To prove that \mathcal{U}_ϵ is relatively open in \mathcal{X} , it suffices to show that the set $\{x \in \mathbb{R}^n : \text{dist}(0, \partial\Psi(x)) > \epsilon\}$ is open, or equivalently, $\mathcal{C} = \{x \in \mathbb{R}^n : \text{dist}(0, \partial\Psi(x)) \leq \epsilon\}$ is closed. To this end, consider any convergent sequence $\{x^k\} \subset \mathcal{C}$ with $x^k \rightarrow x$ for some $x \in \mathbb{R}^n$. Clearly, $\mathcal{C} \subseteq \mathcal{X}$ and hence $\{x^k\} \subset \mathcal{X}$. It then follows from $x^k \rightarrow x$ and the closedness of \mathcal{X} that $x \in \mathcal{X}$. Also, since $x^k \in \mathcal{C}$, there exists $s^k \in \partial\Psi(x^k)$ with $\|s^k\| \leq \epsilon$. Without loss of generality, we may assume that $s^k \rightarrow s$ for some s with $\|s\| \leq \epsilon$. Observe from Assumption 1 that $f(\cdot, y)$ is $L_{\nabla f}$ -smooth on \mathcal{X} for all $y \in \mathcal{Y}$, which together with (5) implies that F^* is weakly convex on \mathcal{X} . By this, the convexity of p , $\text{ri}(\text{dom } F^*) \cap \text{ri}(\text{dom } p) = \text{int}(\mathcal{X}) \neq \emptyset$, $\Psi = F^* + p$, and [28, Theorem 28.3], one has $\partial\Psi = \partial F^* + \partial p$ on \mathcal{X} . Additionally, notice that ∂F^* and ∂p are outer semicontinuous on \mathcal{X} . Hence, $\partial\Psi$ is outer semicontinuous on \mathcal{X} , which together with $x^k \rightarrow x$, $s^k \in \partial\Psi(x^k)$, and $s^k \rightarrow s$ implies that $s \in \partial\Psi(x)$. By this and $\|s\| \leq \epsilon$, we conclude that $x \in \mathcal{C}$. Hence, \mathcal{C} is closed as desired. \square

We are now ready to prove Theorem 1.

Proof of Theorem 1. (i) Fix any $\epsilon > 0$ and $x \in \mathcal{U}_\epsilon$. It follows that $x \in \mathcal{X}$ and $\text{dist}(0, \partial\Psi(x)) > \epsilon$. This together with (8) implies that $\{y : F^*(x) > F(x, y) \geq F^*(x) - \gamma\epsilon^\sigma\} \subseteq \mathcal{L}(x)$. By this and (52), one has

$$\text{dist}(y, Y^*(x)) \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} \text{dist}(0, \partial_y F(x, y))^{\frac{1-\theta}{\theta}} \quad \forall y \text{ with } F^*(x) > F(x, y) \geq F^*(x) - \gamma\epsilon^\sigma. \quad (53)$$

Clearly, the above relation also holds for any $y \in Y^*(x)$. Now, fix any $x' \in \mathcal{U}_\epsilon$ with $\|x - x'\| \leq \gamma\epsilon^\sigma / (2L_f)$. Observe from Assumption 1 that $Y^*(x') \neq \emptyset$. Let $y^*(x') \in Y^*(x')$ be arbitrarily chosen. Then $F(x', y^*(x')) = F^*(x')$. By these, Assumption 1, and Lemma 1, one has

$$F(x, y^*(x')) - F^*(x) = F(x, y^*(x')) - F(x', y^*(x')) + F^*(x') - F^*(x) \geq -2L_f \|x - x'\| \geq -\gamma\epsilon^\sigma,$$

where the first inequality uses the L_f -Lipschitz continuity of F^* and $F(\cdot, y)$ for each $y \in \mathcal{Y}$ due to Assumption 1 and Lemma 1. Hence, it follows from (53) that

$$\text{dist}(y^*(x'), Y^*(x)) \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} \text{dist}(0, \partial_y F(x, y^*(x')))^{\frac{1-\theta}{\theta}}.$$

Since $y^*(x') \in Y^*(x')$, by the first-order optimality condition, one has $0 \in \partial_y F(x', y^*(x'))$. In addition, by the expression of F and the smoothness of f on $\mathcal{X} \times \mathcal{Y}$, we obtain

$$\partial_y F(x', y^*(x')) = \nabla_y f(x', y^*(x')) - \partial q(y^*(x')), \quad \partial_y F(x, y^*(x')) = \nabla_y f(x, y^*(x')) - \partial q(y^*(x')). \quad (54)$$

The first relation in (54) and $0 \in \partial_y F(x', y^*(x'))$ lead to $\nabla_y f(x', y^*(x')) \in \partial q(y^*(x'))$, which along with the second relation in (54) implies that

$$\nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x')) \in \partial_y F(x, y^*(x')).$$

Using this and the Lipschitz continuity of $\nabla_y f$, we have

$$\text{dist}(0, \partial_y F(x, y^*(x'))) \leq \|\nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x'))\| \leq L_{\nabla_y f} \|x' - x\|.$$

Combining this with (53) yields

$$\text{dist}(y^*(x'), Y^*(x)) \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla_y f}^{\frac{1-\theta}{\theta}} \|x' - x\|^{\frac{1-\theta}{\theta}}.$$

Notice from Assumption 1 that $Y^*(x)$ is a nonempty closed set. Hence, there exists $y^*(x) \in Y^*(x)$ such that $\|y^*(x') - y^*(x)\| = \text{dist}(y^*(x'), Y^*(x))$. By this and the above relation, one has

$$\|y^*(x') - y^*(x)\| \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla_y f}^{\frac{1-\theta}{\theta}} \|x' - x\|^{\frac{1-\theta}{\theta}}. \quad (55)$$

Suppose further that F^* is differentiable at x and x' . Using this, $y^*(x) \in Y^*(x)$, $y^*(x') \in Y^*(x')$, and Lemma 2, we obtain

$$\nabla F^*(x) = \nabla_x f(x, y^*(x)), \quad \nabla F^*(x') = \nabla_x f(x', y^*(x')).$$

By these, (55), $\theta \in [1/2, 1)$, $\|x - x'\| \leq \gamma\epsilon^\sigma / (2L_f)$, and the Lipschitz smoothness of f , one has

$$\begin{aligned} \|\nabla F^*(x') - \nabla F^*(x)\| &= \|\nabla_x f(x', y^*(x')) - \nabla_x f(x, y^*(x))\| \\ &\leq \|\nabla_x f(x', y^*(x')) - \nabla_x f(x, y^*(x'))\| + \|\nabla_x f(x, y^*(x')) - \nabla_x f(x, y^*(x))\| \\ &\leq L_{\nabla_x f} \|x' - x\| + L_{\nabla_x f} \|y^*(x') - y^*(x)\| \stackrel{(55)}{\leq} L_{\nabla_x f} \|x' - x\| + (1 - \theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla_y f}^{\frac{1}{\theta}} \|x' - x\|^{\frac{1-\theta}{\theta}}. \end{aligned} \quad (56)$$

Using this, Lemma 3, and the fact that \mathcal{U}_ϵ is relatively open in \mathcal{X} (see Lemma 6), we conclude that $\partial_{\mathcal{X}}^C F^*(x)$ is a singleton for every $x \in \mathcal{U}_\epsilon$ and F^* is differentiable on $\mathcal{U}_\epsilon \cap \text{int}(\mathcal{X})$. Hence, statement (i) of Theorem 1 holds.

(ii) The relation (13) directly follows from statement (i) of Theorem 1 and (56). This proves statement (ii) of Theorem 1.

(iii) We now prove (14). Let us fix any $x, x' \in \mathcal{U}_\epsilon$ with $\|x - x'\| \leq \gamma\epsilon^\sigma/(4L_f)$. It follows from statement (i) of Theorem 1 that both $\partial_{\mathcal{X}}^{\mathcal{C}}F^*(x)$ and $\partial_{\mathcal{X}}^{\mathcal{C}}F^*(x')$ are a singleton. By this and (3), there exist sequences $\{x^k\} \subset \mathcal{S}$ and $\{\hat{x}^k\} \subset \mathcal{S}$ such that $\nabla_{\mathcal{X}}^{\mathcal{C}}F^*(x) = \lim_{k \rightarrow \infty} \nabla F^*(x^k)$, $\nabla_{\mathcal{X}}^{\mathcal{C}}F^*(x') = \lim_{k \rightarrow \infty} \nabla F^*(\hat{x}^k)$, $x^k \rightarrow x$, and $\hat{x}^k \rightarrow x'$. In view of these and the relative openness of \mathcal{U}_ϵ , we may assume without loss of generality that $x^k, \hat{x}^k \in \mathcal{U}_\epsilon \cap \mathcal{S}$ and $\|x^k - \hat{x}^k\| \leq \gamma\epsilon^\sigma/(2L_f)$ for all k . It then follows from (56) that

$$\|\nabla F^*(x^k) - \nabla F^*(\hat{x}^k)\| \leq L_{\nabla f} \|x^k - \hat{x}^k\| + (1 - \theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1}{\theta}} \|x^k - \hat{x}^k\|^{\frac{1-\theta}{\theta}}.$$

Taking limits on both sides of this inequality as $k \rightarrow \infty$ yields (14). Hence, statement (iii) of Theorem 1 holds.

(iv) We next prove (15). To this end, fix any $x \in \mathcal{U}_\epsilon$ and $y^* \in Y^*(x)$. It follows from statement (i) of Theorem 1 that $\partial_{\mathcal{X}}^{\mathcal{C}}F^*(x)$ is a singleton. By this and (3), there exists a sequence $\{x^k\} \subset \mathcal{S}$ such that $\nabla_{\mathcal{X}}^{\mathcal{C}}F^*(x) = \lim_{k \rightarrow \infty} \nabla F^*(x^k)$ and $x^k \rightarrow x$. Using this and the relative openness of \mathcal{U}_ϵ , we may assume without loss of generality that $x^k \in \mathcal{U}_\epsilon$ and $\|x^k - x\| \leq \gamma\epsilon^\sigma/(2L_f)$ for all k . By these and (55), there exists some $y^k \in Y^*(x^k)$ such that

$$\|y^k - y^*\| \leq (1 - \theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1-\theta}{\theta}} \|x^k - x\|^{\frac{1-\theta}{\theta}}.$$

This along with $x^k \rightarrow x$ implies $y^k \rightarrow y^*$. Also, notice from $x^k \in \mathcal{S}$, $y^k \in Y^*(x^k)$, and Lemma 2 that $\nabla F^*(x^k) = \nabla_x f(x^k, y^k)$ for all k . In view of these and the continuity of $\nabla_x f$, one has $\nabla_{\mathcal{X}}^{\mathcal{C}}F^*(x) = \nabla_x f(x, y^*)$, and hence (15) holds. This proves statement (iv) of Theorem 1. \square

We finally prove Corollary 1.

Proof of Corollary 1. Fix any x, x' satisfying $[x, x'] \subseteq \mathcal{U}_\epsilon$ and $\|x - x'\| \leq \gamma\epsilon^\sigma/(4L_f)$. By $\text{int}(\mathcal{X}) \neq \emptyset$, $x, x' \in \mathcal{X}$, the convexity of \mathcal{X} , and a similar argument as used in the proof of Lemma 3, one can see that there exist sequences $\{x^k\} \subset \text{int}(\mathcal{X})$ and $\{\hat{x}^k\} \subset \text{int}(\mathcal{X})$ such that $x^k \rightarrow x$ and $\hat{x}^k \rightarrow x'$. Then, one can deduce from $[x, x'] \subseteq \mathcal{U}_\epsilon$ and the relative openness of \mathcal{U}_ϵ that $[x^k, \hat{x}^k] \subseteq \mathcal{U}_\epsilon$ for all sufficiently large k . In addition, notice from the convexity of $\text{int}(\mathcal{X})$ that $[x^k, \hat{x}^k] \subseteq \text{int}(\mathcal{X})$ for all k . In view of these, we may assume without loss of generality that $[x^k, \hat{x}^k] \subseteq \mathcal{U}_\epsilon \cap \text{int}(\mathcal{X})$ and $\|x^k - \hat{x}^k\| \leq \gamma\epsilon^\sigma/(2L_f)$ for all k . It then follows from Theorem 1(ii) that for any $z, z' \in [x^k, \hat{x}^k]$,

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

Hence, we obtain that

$$F^*(x^k) \leq F^*(\hat{x}^k) + \langle \nabla F^*(\hat{x}^k), x^k - \hat{x}^k \rangle + \frac{1}{2} L_{\nabla f} \|x^k - \hat{x}^k\|^2 + \frac{M}{1 + \nu} \|x^k - \hat{x}^k\|^{1+\nu}, \quad (57)$$

where M and ν are defined in (17). Note that $\{\hat{x}^k\} \subset \text{int}(\mathcal{X})$ and F^* is Lipschitz continuous on \mathcal{X} (see Lemma 1). This implies that $\{\nabla F^*(\hat{x}^k)\}$ is bounded. By this, $\{\hat{x}^k\} \subset \mathcal{X}$, $\hat{x}^k \rightarrow x'$, and (3), one can observe that any limit point of $\{\nabla F^*(\hat{x}^k)\}$ belongs to $\partial_{\mathcal{X}}^{\mathcal{C}}F^*(x')$. Moreover, $\partial_{\mathcal{X}}^{\mathcal{C}}F^*(x')$ is a singleton due to Theorem 1 and $x' \in \mathcal{U}_\epsilon$. Consequently, the limit of $\{\nabla F^*(\hat{x}^k)\}$ as $k \rightarrow \infty$ exists, and moreover, $\lim_{k \rightarrow \infty} \nabla F^*(\hat{x}^k) = \nabla_{\mathcal{X}}^{\mathcal{C}}F^*(x')$. Using this, $x^k \rightarrow x$, $\hat{x}^k \rightarrow x'$, the continuity of F^* , and taking limits on both sides of (57) as $k \rightarrow \infty$, we conclude that (16) holds, which proves Corollary 1. \square

6.2 Proof of the main results in Section 3

In this subsection, we provide the proofs of Theorems 2 and 3. We begin by proving Theorem 2.

Proof of Theorem 2. Suppose for contradiction that the inner loop runs for more than $\bar{i} + 1$ iterations at the k th outer iteration. Then one can observe from Algorithm 1 that

$$h(z^{k+1, \bar{i}}) + \frac{1}{2\lambda_{k, \bar{i}}} \|z^{k+1, \bar{i}} - z^k\|^2 > h(z^k). \quad (58)$$

By the optimality condition for $z^{k+1, \bar{i}}$, one has

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

In addition, by the L -smoothness of g , we have

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

Combining these two inequalities yields

$$h(z^{k+1, \bar{i}}) + \left(\frac{1}{\lambda_{k, \bar{i}}} - \frac{L}{2} \right) \|z^{k+1, \bar{i}} - z^k\|^2 \leq h(z^k). \quad (59)$$

By the definition of \bar{i} and $\lambda_{k, \bar{i}} = \bar{\lambda} \rho^{\bar{i}}$, one has $L \leq 1/\lambda_{k, \bar{i}}$. This and (59) imply that

$$h(z^{k+1, \bar{i}}) + \frac{1}{2\lambda_{k, \bar{i}}} \|z^{k+1, \bar{i}} - z^k\|^2 \leq h(z^k),$$

which contradicts (58). Hence, the inner loop runs at most $\bar{i} + 1$ iterations. By this and the definition of λ_k , one has $\min\{\rho/L, \bar{\lambda}\} \leq \bar{\lambda} \rho^{\bar{i}} \leq \lambda_k \leq \bar{\lambda}$. Hence, the conclusion of Theorem 2 holds. \square

In the remainder of this subsection, we present the proof of Theorem 3. To this end, we first establish several technical lemmas. The following result provides a bound on $\text{dist}(0, \partial h(z^{k+1}))$ in terms of $\|z^{k+1} - z^k\|$.

Lemma 7. *Let z^k and z^{k+1} be generated by Algorithm 1 for some $k \geq 0$. Then it holds that*

$$\text{dist}(0, \partial h(z^{k+1})) \leq (L + \lambda_k^{-1}) \|z^{k+1} - z^k\|. \quad (60)$$

Proof. By the optimality condition for z^{k+1} , one has

$$0 \in \nabla g(z^k) + \lambda_k^{-1}(z^{k+1} - z^k) + \partial q(z^{k+1}),$$

which implies that

$$\nabla g(z^{k+1}) - \nabla g(z^k) - \lambda_k^{-1}(z^{k+1} - z^k) \in \partial h(z^{k+1}).$$

Using this and the L -smoothness of g , we obtain

$$\text{dist}(0, \partial h(z^{k+1})) \leq \|\nabla g(z^{k+1}) - \nabla g(z^k) - \lambda_k^{-1}(z^{k+1} - z^k)\| \leq (L + \lambda_k^{-1}) \|z^{k+1} - z^k\|,$$

and hence the conclusion holds. \square

For notational convenience, let

$$a_k = \frac{1}{2\lambda_k}, \quad b_k = \left(L + \frac{1}{\lambda_k} \right)^{-1}. \quad (61)$$

In view of these, Algorithm 1, Theorem 2, and Lemma 7, one can observe that the following relations hold:

$$h(z^{k+1}) + a_k \|z^{k+1} - z^k\|^2 \leq h(z^k), \quad (62)$$

$$b_k \operatorname{dist}(0, \partial h(z^{k+1})) \leq \|z^{k+1} - z^k\|, \quad (63)$$

$$(2\bar{\lambda})^{-1} \leq a_k \leq (2\underline{\lambda})^{-1}, \quad (L + \underline{\lambda}^{-1})^{-1} \leq b_k \leq (L + \bar{\lambda}^{-1})^{-1}, \quad (64)$$

where $\underline{\lambda}$ is defined in (21). In addition, by (62) and the choice of z^0 , we can observe that $r_0 \leq \delta$, and $\{r_k\}$ is nonincreasing. Consequently, $r_k \leq \delta$ holds for all k .

The following lemma establishes a convergence rate for Algorithm 1, following a similar argument as in [14, Theorem 3.4].

Lemma 8. *Let $\delta, \theta, \underline{\lambda}, \underline{\beta}, \bar{\beta}, C'$ and $\bar{\lambda}$ be given in (19), (21), (22), (61) and Algorithm 1, respectively. Suppose that z^k is generated by Algorithm 1 for some $k \geq 1$. Then the following statements hold.*

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

$$h(z^k) - h^* \leq \delta e^{-\frac{\beta}{1+\underline{\beta}}k}. \quad (65)$$

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

$$h(z^k) - h^* \leq \left(\frac{1}{C'(2\theta - 1)\underline{\beta}} \right)^{\frac{1}{2\theta-1}} k^{-\frac{1}{2\theta-1}}. \quad (66)$$

Proof. For notational convenience, let $r_\ell = h(z^\ell) - h^*$ for all ℓ . Since $h(z^0) - h^* \leq \delta$ and $\{h(z^\ell)\}$ is nonincreasing, (19) holds with $z = z^\ell$ for all $\ell \geq 0$. By this, (62), and (63), one has

$$r_\ell - r_{\ell+1} \stackrel{(62)}{\geq} a_\ell \|z^{\ell+1} - z^\ell\|^2 \stackrel{(63)}{\geq} a_\ell b_\ell^2 \operatorname{dist}(0, \partial h(z^{\ell+1}))^2 \stackrel{(19)}{\geq} a_\ell b_\ell^2 C^2 r_{\ell+1}^{2\theta}.$$

Let $\beta_\ell := a_\ell b_\ell^2 C^2$ for all ℓ . Using (21) and (64), we have

$$r_\ell - r_{\ell+1} \geq \beta_\ell r_{\ell+1}^{2\theta}, \quad \beta_\ell \in [\underline{\beta}, \bar{\beta}], \quad (67)$$

(i) Suppose $\theta = 1/2$. It then follows from (67) that $r_{\ell+1} \leq (1 + \beta_\ell)^{-1} r_\ell$ for all ℓ . Hence,

$$r_k \leq r_0 \prod_{\ell=0}^{k-1} (1 + \beta_\ell)^{-1} \quad \forall k \geq 0. \quad (68)$$

By the concavity of $\log(\cdot)$, one has that $\log(1 + t) \leq t$ for all $t > -1$. It follows that

$$\log(1 + \beta_\ell)^{-1} = \log\left(1 - \frac{\beta_\ell}{1 + \beta_\ell}\right) \leq -\frac{\beta_\ell}{1 + \beta_\ell}.$$

Using this and $\beta_\ell \geq \underline{\beta}$ for all ℓ , we obtain

$$\prod_{\ell=0}^{k-1} (1 + \beta_\ell)^{-1} = \exp\left(\sum_{\ell=0}^{k-1} \log(1 + \beta_\ell)^{-1}\right) \leq \exp\left(-\sum_{\ell=0}^{k-1} \frac{\beta_\ell}{1 + \beta_\ell}\right) \leq \exp\left(-\frac{k\underline{\beta}}{1 + \underline{\beta}}\right),$$

which together with (68) and $r_0 \leq \delta$ implies that (65) holds.

(ii) Suppose $\theta \in (1/2, 1)$. Clearly, (66) holds if $r_k = 0$. Now we assume that $r_k > 0$. It then follows from the monotonicity of $\{r_\ell\}$ that $r_\ell > 0$ for all $0 \leq \ell < k$. Let $\psi(t) = \frac{1}{2\theta-1} t^{1-2\theta}$. Then we have

$$\psi(r_{\ell+1}) - \psi(r_\ell) = \int_{r_\ell}^{r_{\ell+1}} \psi'(t) dt = \int_{r_{\ell+1}}^{r_\ell} t^{-2\theta} dt \geq r_\ell^{-2\theta} (r_\ell - r_{\ell+1}) \quad \forall 0 \leq \ell < k. \quad (69)$$

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

Case a): $r_{\ell+1}^{-2\theta} \leq 2r_\ell^{-2\theta}$. It along with (67) and (69) implies that

$$\psi(r_{\ell+1}) - \psi(r_\ell) \geq \frac{1}{2} r_{\ell+1}^{-2\theta} (r_\ell - r_{\ell+1}) \stackrel{(67)}{\geq} \frac{1}{2} \beta_\ell.$$

Case b): $r_{\ell+1}^{-2\theta} > 2r_\ell^{-2\theta}$. It leads to $r_{\ell+1}^{1-2\theta} > 2^{\frac{2\theta-1}{2\theta}} r_\ell^{1-2\theta}$. By this, $r_\ell \leq \delta$, $\beta_\ell \geq \bar{\beta}$, and the expression of ψ , one has

$$\begin{aligned} \psi(r_{\ell+1}) - \psi(r_\ell) &= \frac{1}{2\theta-1} (r_{\ell+1}^{1-2\theta} - r_\ell^{1-2\theta}) > \frac{1}{2\theta-1} \left(2^{\frac{2\theta-1}{2\theta}} - 1 \right) r_\ell^{1-2\theta} \\ &\geq \frac{1}{2\theta-1} \left(2^{\frac{2\theta-1}{2\theta}} - 1 \right) \delta^{1-2\theta} \geq \frac{(2^{\frac{2\theta-1}{2\theta}} - 1) \delta^{1-2\theta}}{(2\theta-1) \bar{\beta}} \beta_\ell. \end{aligned}$$

Combining the above two cases, and using the definition of C' in (21), we obtain that $\psi(r_{\ell+1}) - \psi(r_\ell) \geq C' \beta_\ell$ for all $0 \leq \ell < k$. It then follows that

$$\psi(r_k) \geq \psi(r_0) + C' \sum_{\ell=0}^{k-1} \beta_\ell \geq C' \sum_{\ell=0}^{k-1} \beta_\ell.$$

This and the expression of ψ lead to

$$r_k \leq \left(\frac{1}{C'(2\theta-1)} \right)^{\frac{1}{2\theta-1}} \left(\sum_{\ell=0}^{k-1} \beta_\ell \right)^{-\frac{1}{2\theta-1}} \leq \left(\frac{1}{C'(2\theta-1)\underline{\beta}} \right)^{\frac{1}{2\theta-1}} k^{-\frac{1}{2\theta-1}},$$

and hence (66) holds. \square

We are now ready to prove Theorem 3.

Proof of Theorem 3. Suppose for contradiction that Algorithm 1 runs for more than \bar{K}_θ outer iterations. Then there exists some $\ell \geq \bar{K}_\theta - 1$ such that $\|z^{\ell+1} - z^\ell\| > \tau$. By (62) and (64), one has

$$\|z^{\ell+1} - z^\ell\| \stackrel{(62)}{\leq} \sqrt{\frac{r_\ell - r_{\ell+1}}{a_\ell}} \leq a_\ell^{-\frac{1}{2}} r_\ell^{\frac{1}{2}} \stackrel{(64)}{\leq} (2\bar{\lambda})^{\frac{1}{2}} r_\ell^{\frac{1}{2}}, \quad (70)$$

where $r_\ell = h(z^\ell) - h^*$. We next show that $r_\ell \leq \tau^2/(2\bar{\lambda})$ by considering two separate cases: $\theta = 1/2$ and $\theta \in (1/2, 1)$.

Case (i): $\theta = 1/2$. By this, (22), and $\ell \geq \bar{K}_\theta - 1$, one has $\ell \geq \underline{\beta}^{-1}(1 + \underline{\beta}) \log(2\bar{\lambda}\delta\tau^{-2})$. Using this relation and (65), we have $r_\ell \leq \delta e^{-\underline{\beta}(1+\underline{\beta})^{-1}\ell} \leq \tau^2/(2\bar{\lambda})$.

Case (ii): $\theta \in (1/2, 1)$. Using this, (22), and $\ell \geq \bar{K}_\theta - 1$, we obtain that $\ell \geq \frac{1}{C'(2\theta-1)\underline{\beta}} (2\bar{\lambda}\tau^{-2})^{2\theta-1}$. By this relation and (66), one has

$$r_\ell \leq \left(\frac{1}{C'(2\theta-1)\underline{\beta}} \right)^{\frac{1}{2\theta-1}} \ell^{-\frac{1}{2\theta-1}} \leq \tau^2/(2\bar{\lambda}).$$

We thus conclude that $r_\ell \leq \tau^2/(2\bar{\lambda})$. This together with (70) implies $\|z^{\ell+1} - z^\ell\| \leq \tau$, which leads to a contradiction. Hence, Algorithm 1 runs at most \bar{K}_θ outer iterations.

We next show that (23) holds. Notice from (20) and (61) that $\lambda_k \geq \underline{\lambda}$. By this and (60), one has

$$\text{dist}(0, \partial h(z^{k+1})) \leq (L + \underline{\lambda}^{-1}) \|z^{k+1} - z^k\|. \quad (71)$$

Since $h(z^0) - h^* \leq \delta$ and $\{h(z^\ell)\}$ is nonincreasing, it follows that $h(z^{k+1}) - h^* \leq \delta$. Using this and (19), we have

$$C(h(z^{k+1}) - h^*)^\theta \leq \text{dist}(0, \partial h(z^{k+1})).$$

By this, (71), and $\|z^{k+1} - z^k\| \leq \tau$, one has

$$h(z^{k+1}) - h^* \leq C^{-\frac{1}{\theta}} (\text{dist}(0, \partial h(z^{k+1})))^{\frac{1}{\theta}} \leq (C^{-1}(L + \underline{\lambda}^{-1}))^{\frac{1}{\theta}} \|z^{k+1} - z^k\|^{\frac{1}{\theta}} \leq (C^{-1}(L + \underline{\lambda}^{-1})\tau)^{\frac{1}{\theta}},$$

and hence (23) holds as desired. \square

6.3 Proof of the main results in Section 4

In this subsection we prove Theorems 4 and 5. To proceed, we first establish several technical lemmas below.

Lemma 9. *Let $\gamma, \sigma, C, \theta, L_{\nabla f}, \epsilon, \mathcal{X}_\epsilon^c$, and $\{\eta_\ell\}$ be given in (24), 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 $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell < k$. Then, for all $0 \leq \ell \leq k$, it holds that*

$$F^*(x^\ell) - F(x^\ell, y^\ell) \leq \min\{\gamma\epsilon^\sigma/2, \eta_\ell\}, \quad \text{dist}(y^\ell, Y^*(x^\ell)) \leq \frac{1}{C(1-\theta)} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\}, \quad (72)$$

$$\|\nabla_{\mathcal{X}}^C F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell)\| \leq \frac{L_{\nabla f}}{C(1-\theta)} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\}. \quad (73)$$

Proof. We first show that (72) and (73) hold for $\ell = 0$. One can observe from Algorithm 2 that the first relation in (72) holds for $\ell = 0$. We now show that the second relation in (72) and (73) also hold for $\ell = 0$. By the assumption in this lemma, we know that $x^0 \in \mathcal{X}_\epsilon^c$, and hence $\text{dist}(0, \partial\Psi(x^0)) > \epsilon$. Using this, $F^*(x^0) - F(x^0, y^0) \leq \min\{\gamma\epsilon^\sigma/2, 1\}$, and (8), we see that $y^0 \in \mathcal{L}(x^0)$ or $y^0 \in Y^*(x^0)$. In view of these, $x^0 \in \mathcal{X}$, and $\theta \in [1/2, 1)$, it follows from Lemma 4 that

$$\text{dist}(y^0, Y^*(x^0)) \leq \frac{1}{C(1-\theta)} (F^*(x^0) - F(x^0, y^0))^{1-\theta} \leq \frac{1}{C(1-\theta)} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, 1\}. \quad (74)$$

This together with $\eta_0 = 1$ implies that the second relation in (72) holds for $\ell = 0$. Moreover, one can see from $x^0 \in \mathcal{X}$, $\text{dist}(0, \partial\Psi(x^0)) > \epsilon$ and (12) that $x^0 \in \mathcal{U}_\epsilon$. It then follows from Theorem 1 that $\nabla_{\mathcal{X}}^C F^*(x^0) = \nabla_x f(x^0, y^*)$, where $y^* \in Y^*(x^0)$ with $\|y^* - y^0\| = \text{dist}(y^0, Y^*(x^0))$. Using this, (74), and the $L_{\nabla f}$ -smoothness of f , one has

$$\begin{aligned} \|\nabla_{\mathcal{X}}^C F^*(x^0) - \nabla_x f(x^0, y^0)\| &= \|\nabla_x f(x^0, y^*) - \nabla_x f(x^0, y^0)\| \leq L_{\nabla f} \|y^* - y^0\| \\ &= L_{\nabla f} \text{dist}(y^0, Y^*(x^0)) \leq \frac{L_{\nabla f}}{C(1-\theta)} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, 1\}. \end{aligned}$$

This along with $\eta_0 = 1$ implies that (73) also holds for $\ell = 0$.

We next show that (72) and (73) hold for $0 < \ell \leq k$. Notice from Algorithm 2 that y^ℓ is an approximate solution of the problem $\min_y \{-f(x^\ell, y) + q(y)\}$ obtained by Algorithm 1 with the initial point $y^{\ell-1}$, and the parameters $\bar{\lambda}, \rho, \tau$ specified in Algorithm 2. Then it follows from $\tau = \frac{C}{L_{\nabla f} + \bar{\lambda}^{-1}} \min\left\{(\frac{1}{2}\gamma\epsilon^\sigma)^\theta, \eta_\ell^{\frac{\theta}{2(1-\theta)}}\right\}$, the definitions of F^* and F , and Theorem 3 with $h(\cdot) = -f(x^\ell, \cdot) + q(\cdot)$ that

$$F^*(x^\ell) - F(x^\ell, y^\ell) \leq [C^{-1}(L_{\nabla f} + \bar{\lambda}^{-1})\tau]^{\frac{1}{\theta}} = \min\left\{\frac{1}{2}\gamma\epsilon^\sigma, \eta_\ell^{\frac{1}{2(1-\theta)}}\right\}. \quad (75)$$

This together with $\eta_\ell \in (0, 1)$ and $\theta \in [1/2, 1)$ implies that the first relation in (72) holds for $\ell > 0$. In addition, notice from the assumption that $x^{\ell-1} \in \mathcal{X}_\epsilon^c$. Also, observe from Algorithm 2 that $r = \gamma\epsilon^\sigma/(4L_f)$ and $\|x^\ell - x^{\ell-1}\| \leq r$. It then follows that $\|x^\ell - x^{\ell-1}\| \leq \gamma\epsilon^\sigma/(4L_f)$, which together with $x^{\ell-1} \in \mathcal{X}_\epsilon^c$ implies that $\text{dist}(0, \partial\Psi(x^\ell)) > \epsilon$. Using this and (75), we obtain that $F^*(x^\ell) - F(x^\ell, y^\ell) \leq \gamma \text{dist}(0, \partial\Psi(x^\ell))^\sigma$ and hence $y^\ell \in \mathcal{L}(x^\ell)$, where $\mathcal{L}(\cdot)$ is defined in (8). By this, (75), $x^\ell \in \mathcal{X}$, and Lemma 4, one can conclude that the second relation in (72) holds for $\ell > 0$. Lastly, (73) also holds for $\ell > 0$, due to the second relation in (72) and arguments similar to those used in the case $\ell = 0$. \square

Lemma 10. Let $\epsilon > 0$ be given, $M, \mathcal{X}_\epsilon^c$ be defined in (17) and (24), $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma, \{\delta_\ell\}, \{\eta_\ell\}, \{L_\ell\}$ be given in Assumption 1 and Algorithm 2, and let

$$\Delta_k := 8 \left[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^k \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2} \right], \quad (76)$$

$$\underline{K}_\epsilon := \max\{k \geq 1 : \Delta_k / (k L_{\lceil k/2 \rceil}) \geq \gamma^2 \epsilon^{2\sigma} / (16 L_f^2)\}, \quad (77)$$

$$\overline{K}_\epsilon := \max\{k \geq 0 : x^k \in \mathcal{X}_\epsilon^c\}, \quad (78)$$

$$\ell(k) := \arg \min_{\lceil k/2 \rceil \leq \ell \leq k} L_\ell \|x^{\ell+1} - x^\ell\|^2. \quad (79)$$

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 $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$. Then we have

$$\text{dist}(0, \partial \Psi(x^{\ell(k)+1})) \leq L_{\nabla f} \sqrt{\frac{\Delta_k}{L_{\lceil k/2 \rceil} k}} + \sqrt{\frac{L_k \Delta_k}{k}} + M \left(\frac{\Delta_k}{L_{\lceil k/2 \rceil} k} \right)^{\frac{\nu}{2}} + (1-\theta)^{-1} C^{-1} L_{\nabla f} \eta_{\lceil k/2 \rceil}^{\frac{1}{2}}. \quad (80)$$

Proof. Notice from the above assumption that $\underline{K}_\epsilon < k \leq \overline{K}_\epsilon$ and $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$. We first show that for all $0 \leq \ell \leq k$, it holds that

$$F(x^{\ell+1}, y^{\ell+1}) + p(x^{\ell+1}) \leq F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4} \|x^{\ell+1} - x^\ell\|^2 + \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \frac{\delta_\ell}{2}. \quad (81)$$

To this end, let us fix any $0 \leq \ell \leq k$. By optimality condition of (27), one has

$$\langle \nabla_x f(x^\ell, y^\ell), x^{\ell+1} \rangle + L_\ell \|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) \leq \langle \nabla_x f(x^\ell, y^\ell), x^\ell \rangle + p(x^\ell). \quad (82)$$

Observe from Algorithm 2 that $\|x^{\ell+1} - x^\ell\| \leq \gamma \epsilon^\sigma / (4 L_f)$. Using this relation, $x^\ell \in \mathcal{X}_\epsilon^c$, and the definition of \mathcal{X}_ϵ^c in (24), we deduce that $\text{dist}(0, \partial \Psi(x)) > \epsilon$ for any $x \in [x^\ell, x^{\ell+1}]$. In addition, by $x^\ell, x^{\ell+1} \in \mathcal{X}$, and the convexity of \mathcal{X} , one can see that $[x^\ell, x^{\ell+1}] \subseteq \mathcal{X}$. It follows from these and (12) that $[x^\ell, x^{\ell+1}] \subseteq \mathcal{U}_\epsilon$. In view of this, (25), and the definition of L_ℓ in Algorithm 2, we have

$$F^*(x^{\ell+1}) \leq F^*(x^\ell) + \langle \nabla_{\mathcal{X}}^C F^*(x^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2} \|x^{\ell+1} - x^\ell\|^2 + \frac{\delta_\ell}{2}. \quad (83)$$

In addition, notice that $F(x^{\ell+1}, y^{\ell+1}) \leq F^*(x^{\ell+1})$. Using this, (72), (73), (82), and (83), we obtain that

$$\begin{aligned} F(x^{\ell+1}, y^{\ell+1}) + p(x^{\ell+1}) &\leq F^*(x^{\ell+1}) + p(x^{\ell+1}) \\ &\stackrel{(83)}{\leq} F^*(x^\ell) + \langle \nabla_{\mathcal{X}}^C F^*(x^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2} \|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) + \frac{\delta_\ell}{2} \\ &= F(x^\ell, y^\ell) + \langle \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2} \|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) + F^*(x^\ell) - F(x^\ell, y^\ell) \\ &\quad + \langle \nabla_{\mathcal{X}}^C F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{\delta_\ell}{2} \\ &\stackrel{(72)(82)}{\leq} F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{2} \|x^{\ell+1} - x^\ell\|^2 + \eta_\ell + \langle \nabla_{\mathcal{X}}^C F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{\delta_\ell}{2} \\ &= F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4} \|x^{\ell+1} - x^\ell\|^2 - \frac{L_\ell}{4} \|x^{\ell+1} - x^\ell\|^2 + \langle \nabla_{\mathcal{X}}^C F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \eta_\ell + \frac{\delta_\ell}{2} \\ &\leq F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4} \|x^{\ell+1} - x^\ell\|^2 + \frac{\|\nabla_{\mathcal{X}}^C F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell)\|^2}{L_\ell} + \eta_\ell + \frac{\delta_\ell}{2} \\ &\stackrel{(73)}{\leq} F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4} \|x^{\ell+1} - x^\ell\|^2 + \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \frac{\delta_\ell}{2}, \end{aligned}$$

where the fourth inequality follows from the Young's inequality $\langle u, v \rangle \leq \alpha \|u\|^2/4 + \|v\|^2/\alpha$ for all $\alpha > 0$ and $u, v \in \mathbb{R}^n$. By this and the arbitrariness of ℓ , we see that (81) holds for all $0 \leq \ell \leq k$.

Summing up (81) over $\ell = 0, \dots, k$ yields

$$\sum_{\ell=\lceil k/2 \rceil}^k L_\ell \|x^{\ell+1} - x^\ell\|^2 \leq 4 \left[F(x^0, y^0) + p(x^0) - F(x^{k+1}, y^{k+1}) - p(x^{k+1}) + \sum_{\ell=0}^k \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2} \right]. \quad (84)$$

For notational convenience, let $\hat{k} = \ell(k)$. By this, (76), and (79), one has

$$\begin{aligned} L_{\hat{k}} \|x^{\hat{k}+1} - x^{\hat{k}}\|^2 &\stackrel{(79)}{\leq} \frac{1}{\lceil k/2 \rceil} \sum_{\ell=\lceil k/2 \rceil}^k L_\ell \|x^{\ell+1} - x^\ell\|^2 \\ &\stackrel{(84)}{\leq} \frac{4}{\lceil k/2 \rceil} \left[F(x^0, y^0) + p(x^0) - F(x^{k+1}, y^{k+1}) - p(x^{k+1}) + \sum_{\ell=0}^k \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2} \right] \\ &\leq \frac{8}{k} \left[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^k \left(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell} \right) \eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2} \right] \stackrel{(76)}{=} \frac{\Delta_k}{k}, \end{aligned} \quad (85)$$

where the last inequality uses the fact that $F(x^0, y^0) \leq F^*(x^0)$, $\Psi(\cdot) = F^*(\cdot) + p(\cdot)$, $\Psi(x^{k+1}) \geq \Psi^*$, and $F^*(x^{k+1}) - F(x^{k+1}, y^{k+1}) \leq \eta_{k+1}$ due to (72). Additionally, by $\hat{k} \geq \lceil k/2 \rceil$, the monotonicity of $\{\delta_\ell\}$, $\nu \in (0, 1]$, and the definition of L_ℓ in Algorithm 2, one has $L_{\hat{k}} \geq L_{\lceil k/2 \rceil}$. It then together with (85) implies that

$$\|x^{\hat{k}+1} - x^{\hat{k}}\| \leq \sqrt{\frac{\Delta_k}{L_{\hat{k}} \hat{k}}} \leq \sqrt{\frac{\Delta_k}{L_{\lceil k/2 \rceil} k}}. \quad (86)$$

In addition, notice from Algorithm 2 that $r = \gamma \epsilon^\sigma / (4L_f)$. By this, $k > \underline{K}_\epsilon$, and (77), one has $\Delta_k / (L_{\lceil k/2 \rceil} k) < r^2$. Using this and (86), we have

$$\|x^{\hat{k}+1} - x^{\hat{k}}\| < r. \quad (87)$$

This together with the first-order optimality condition of (27) for $x^{\hat{k}+1}$ implies that

$$0 \in \nabla_x f(x^{\hat{k}}, y^{\hat{k}}) + L_{\hat{k}}(x^{\hat{k}+1} - x^{\hat{k}}) + \partial p(x^{\hat{k}+1}). \quad (88)$$

By $\hat{k} \leq k$ and the assumption that $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$, one has $x^{\hat{k}} \in \mathcal{X}_\epsilon^c$. Using this and (87), we conclude that $\text{dist}(0, \partial \Psi(x^{\hat{k}})) > \epsilon$ and $\text{dist}(0, \partial \Psi(x^{\hat{k}+1})) > \epsilon$. By these and $x^{\hat{k}}, x^{\hat{k}+1} \in \mathcal{X}$, one can see that $x^{\hat{k}}, x^{\hat{k}+1} \in \mathcal{U}_\epsilon$. In view of this, (17), (87), and Theorem 1, we obtain that

$$\|\nabla_{\mathcal{X}}^C F^*(x^{\hat{k}+1}) - \nabla_{\mathcal{X}}^C F^*(x^{\hat{k}})\| \leq L_{\nabla f} \|x^{\hat{k}+1} - x^{\hat{k}}\| + M \|x^{\hat{k}+1} - x^{\hat{k}}\|^\nu. \quad (89)$$

In addition, by $\lceil k/2 \rceil \leq \hat{k} \leq k$, the monotonicity of $\{\eta_\ell\}$ and $\{\delta_\ell\}$, and the definition of L_ℓ , one has $\eta_{\hat{k}} \leq \eta_{\lceil k/2 \rceil}$ and $L_{\lceil k/2 \rceil} \leq L_{\hat{k}} \leq L_k$. Using these, (73), (86), (88), (89), $\partial \Psi = \partial F^* + \partial p$, and $\nabla_{\mathcal{X}}^C F^*(\cdot) \in \partial F^*(\cdot)$, we have

$$\begin{aligned} \text{dist}(0, \partial \Psi(x^{\hat{k}+1})) &\stackrel{(88)}{\leq} \|\nabla_{\mathcal{X}}^C F^*(x^{\hat{k}+1}) - \nabla_x f(x^{\hat{k}}, y^{\hat{k}}) - L_{\hat{k}}(x^{\hat{k}+1} - x^{\hat{k}})\| \\ &\leq \|\nabla_{\mathcal{X}}^C F^*(x^{\hat{k}+1}) - \nabla_{\mathcal{X}}^C F^*(x^{\hat{k}})\| + \|\nabla_{\mathcal{X}}^C F^*(x^{\hat{k}}) - \nabla_x f(x^{\hat{k}}, y^{\hat{k}})\| + L_{\hat{k}} \|x^{\hat{k}+1} - x^{\hat{k}}\| \\ &\stackrel{(73)(89)}{\leq} L_{\nabla f} \|x^{\hat{k}+1} - x^{\hat{k}}\| + M \|x^{\hat{k}+1} - x^{\hat{k}}\|^\nu + (1-\theta)^{-1} C^{-1} L_{\nabla f} \eta_{\hat{k}}^{\frac{1}{2}} + L_{\hat{k}} \|x^{\hat{k}+1} - x^{\hat{k}}\| \\ &\stackrel{(86)}{\leq} L_{\nabla f} \sqrt{\frac{\Delta_k}{L_{\hat{k}} \hat{k}}} + \sqrt{\frac{L_{\hat{k}} \Delta_k}{k}} + M \left(\frac{\Delta_k}{L_{\hat{k}} \hat{k}} \right)^{\frac{\nu}{2}} + (1-\theta)^{-1} C^{-1} L_{\nabla f} \eta_{\hat{k}}^{\frac{1}{2}} \\ &\leq L_{\nabla f} \sqrt{\frac{\Delta_k}{L_{\lceil k/2 \rceil} k}} + \sqrt{\frac{L_k \Delta_k}{k}} + M \left(\frac{\Delta_k}{L_{\lceil k/2 \rceil} k} \right)^{\frac{\nu}{2}} + (1-\theta)^{-1} C^{-1} L_{\nabla f} \eta_{\lceil k/2 \rceil}^{\frac{1}{2}}. \end{aligned}$$

This together with $\hat{k} = \ell(k)$ implies that the conclusion holds. \square

The following lemma will be used to prove Theorem 4 subsequently.

Lemma 11. *Let $\zeta, a, b, \omega > 0$ be given. Then the following statements hold.*

(i) *If $t \geq \lfloor 2\zeta^{-1} \log(1/\zeta) \rfloor_+ + 1$, then $t^{-1} \log t < \zeta$.*

(ii) *If $t \geq \max \left\{ (2a\zeta^{-1})^{1/\omega}, \left(\lfloor 4b(\omega\zeta)^{-1} \log(2b/(\omega\zeta)) \rfloor_+ + 1 \right)^{1/\omega} \right\}$, then $t^{-\omega}(a + b \log t) < \zeta$.*

Proof. We first prove statement (i). Fix any $t \geq \lfloor 2\zeta^{-1} \log(1/\zeta) \rfloor_+ + 1$. Let $\phi(s) = s^{-1} \log s$. It can be verified that ϕ is strictly decreasing on $[e, \infty)$ and $\phi(s) \leq \phi(e) = 1/e$ for all $s > 0$. The latter relation and $t > 0$ imply that $t^{-1} \log t = \phi(t) < \zeta$ holds if $\zeta > 1/e$. We now assume $\zeta \leq 1/e$. It then follows that $t > 2\zeta^{-1} \log(1/\zeta) \geq 2e$, which along with the strict monotonicity of ϕ on $[e, \infty)$ implies that

$$t^{-1} \log t = \phi(t) < \phi(2\zeta^{-1} \log(1/\zeta)) = \frac{\zeta \log((2/\zeta) \log(1/\zeta))}{2 \log(1/\zeta)} = \frac{\zeta}{2} \left(1 + \frac{\log(2 \log(1/\zeta))}{\log(1/\zeta)} \right).$$

In addition, notice that $\zeta \log(1/\zeta) = \phi(1/\zeta) \leq 1/e < 1/2$, which implies that $\log(2 \log(1/\zeta)) \leq \log(1/\zeta)$. By this and the above inequality, one can conclude that statement (i) also holds if $\zeta \leq 1/e$.

We next prove statement (ii). Fix any $t \geq \max\{(2a\zeta^{-1})^{1/\omega}, (\lfloor 4b(\omega\zeta)^{-1} \log(2b/(\omega\zeta)) \rfloor_+ + 1)^{1/\omega}\}$. Since $t \geq (2a\zeta^{-1})^{1/\omega}$, we have $t^{-\omega}a \leq \zeta/2$. In addition, notice that $t^\omega \geq \lfloor 4b(\omega\zeta)^{-1} \log(2b/(\omega\zeta)) \rfloor_+ + 1$, which together with statement (i) implies that $t^{-\omega} \log(t^\omega) < \omega\zeta/(2b)$. It then follows that $bt^{-\omega} \log t = b\omega^{-1}t^{-\omega} \log(t^\omega) < \zeta/2$. By this and $t^{-\omega}a \leq \zeta/2$, one has $t^{-\omega}(a + b \log t) < \zeta$, and hence statement (ii) holds as desired. \square

We are now ready to prove Theorems 4 and 5.

Proof of Theorem 4. For notational convenience, let $k = \hat{K}_\epsilon$. Suppose for contradiction that a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1) is not generated by Algorithm 2 in k iterations.

We first prove by induction that $\{(x^\ell, y^\ell)\}_{\ell=0}^k$ are successfully generated by Algorithm 2. Indeed, since (x^0, y^0) is the initial point, it is generated by Algorithm 2. Now suppose $\{(x^i, y^i)\}_{i=0}^\ell$ are generated by Algorithm 2 for some $0 \leq \ell < k$. Since none of $\{x^i\}_{i=0}^\ell$ is a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of (1), it follows that $x^i \in \mathcal{X}_\epsilon^c$ for all $0 \leq i \leq \ell$, where \mathcal{X}_ϵ^c is defined in (24). By this and Lemma 9, one has that $F^*(x^\ell) - F(x^\ell, y^\ell) \leq \gamma\epsilon^\sigma/2$. In addition, notice from (27) that $x^{\ell+1}$ is well-defined and thus successfully generated by Algorithm 2, and moreover, $\|x^{\ell+1} - x^\ell\| \leq r = \gamma\epsilon^\sigma/(4L_f)$. By these, Lemma 1, and the L_f -Lipschitz continuity of $F(\cdot, y)$ for each $y \in \mathcal{Y}$, we have

$$\begin{aligned} F^*(x^{\ell+1}) - F(x^{\ell+1}, y^\ell) &= F^*(x^{\ell+1}) - F^*(x^\ell) + F^*(x^\ell) - F(x^\ell, y^\ell) + F(x^\ell, y^\ell) - F(x^{\ell+1}, y^\ell) \\ &\leq 2L_f \|x^{\ell+1} - x^\ell\| + \frac{1}{2}\gamma\epsilon^\sigma \leq 2L_f r + \frac{1}{2}\gamma\epsilon^\sigma = \gamma\epsilon^\sigma. \end{aligned} \quad (90)$$

In addition, since $x^\ell \in \mathcal{X}_\epsilon^c$, $\|x^{\ell+1} - x^\ell\| \leq \gamma\epsilon^\sigma/(4L_f)$, and $x^{\ell+1} \in \mathcal{X}$, one can see from (78) that $\text{dist}(0, \partial\Psi(x^{\ell+1})) > \epsilon$. This and (7) imply that

$$C(F^*(x^{\ell+1}) - F(x^{\ell+1}, y))^\theta \leq \text{dist}(0, \partial_y F(x^{\ell+1}, y)) \quad \forall y \text{ with } F^*(x^{\ell+1}) > F(x^{\ell+1}, y) \geq F^*(x^{\ell+1}) - \gamma\epsilon^\sigma.$$

Hence, (19) holds for the function $h(\cdot) = -F(x^{\ell+1}, \cdot)$ with $\delta = \gamma\epsilon^\sigma$. It follows from this and (90) that y^ℓ serves as a suitable initial point for applying Algorithm 1 to solve the problem $\min_y -F(x^{\ell+1}, y)$, or equivalently, $\min_y \{-f(x^{\ell+1}, y) + q(y)\}$. In view of Theorem 3, $y^{\ell+1}$ is then successfully generated by Algorithm 2 via applying Algorithm 1 to this problem. Hence, the induction is completed.

We next derive a contradiction to the above hypothesis. By the definition of \underline{L} in (28), $\nu \in (0, 1]$, $\delta_\ell \leq 1$, and the definition of L_ℓ in Algorithm 2, we see that $L_\ell \geq \underline{L}$ for all $0 \leq \ell \leq k$. In addition,

observe from (30), (31), (37), $\epsilon \in (0, 1/e]$, and $\nu \in (0, 1]$ that $\widehat{K}_\epsilon \geq 2$, which together with $k = \widehat{K}_\epsilon$ implies $k \geq 2$. In view of these, (28), (29), (76), and the definitions of $\{\delta_\ell\}$ and $\{\eta_\ell\}$, one has

$$\begin{aligned}
\Delta_k &\stackrel{(76)(28)}{=} 8 \left[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^k \left(1 + \frac{A}{L_\ell}\right) \eta_\ell + \sum_{\ell=0}^k \frac{\delta_\ell}{2} \right] \leq 8 \left[\Psi(x^0) - \Psi^* + \sum_{\ell=0}^{k+1} \left(1 + \frac{A}{L_\ell}\right) \eta_\ell + \sum_{\ell=0}^{k+1} \frac{\delta_\ell}{2} \right] \\
&\leq 8 \left[\Psi(x^0) - \Psi^* + \left(\frac{3}{2} + \frac{A}{\underline{L}}\right) \sum_{\ell=0}^{k+1} \frac{1}{\ell+1} \right] \leq 8 \left[\Psi(x^0) - \Psi^* + \left(\frac{3}{2} + \frac{A}{\underline{L}}\right) \left(1 + \int_0^{k+1} \frac{1}{1+t} dt\right) \right] \\
&= 8 \left[\Psi(x^0) - \Psi^* + (3/2 + A\underline{L}^{-1})(1 + \log(k+2)) \right] \leq 8 \left[\Psi(x^0) - \Psi^* + (3/2 + A\underline{L}^{-1})(2 + \log k) \right] \\
&= 8 \left[\Psi(x^0) - \Psi^* + 3 + 2A\underline{L}^{-1} + (3/2 + A\underline{L}^{-1}) \log k \right] \stackrel{(29)}{=} a + b \log k, \tag{91}
\end{aligned}$$

where the last inequality follows from $\log(k+2) \leq \log k + 1$ due to $k \geq 2$. Similarly, one can show that $\Delta_{k'} \leq a + b \log k'$ for all $k' \geq k$. Let us fix any $k' \geq k$. Notice from $\delta_\ell = 1/(\ell+1)$ that $\delta_{\lceil k'/2 \rceil} \leq 2/k'$, which along with the definition of L_ℓ and $\nu \in (0, 1]$ implies that $L_{\lceil k'/2 \rceil} \geq (k'/2)^{(1-\nu)/(1+\nu)} M^{2/(1+\nu)}$. In view of these relations and $\nu \in (0, 1]$, we can see that

$$\frac{\Delta_{k'}}{L_{\lceil k'/2 \rceil} k'} \leq \frac{a + b \log k'}{(k'/2)^{\frac{1-\nu}{1+\nu}} k' M^{\frac{2}{1+\nu}}} \leq \frac{2(a + b \log k')}{k'^{\frac{2}{1+\nu}} M^{\frac{2}{1+\nu}}}. \tag{92}$$

Using this and Lemma 11, we observe that $\Delta_{k'}/(L_{\lceil k'/2 \rceil} k') < \gamma^2 \epsilon^{2\sigma}/(16L_f^2)$, since

$$k' \geq k = \widehat{K}_\epsilon \geq \max \left\{ \left(\frac{64aL_f^2}{\gamma^2 \epsilon^{2\sigma} M^{2/(1+\nu)}} \right)^{\frac{1+\nu}{2}}, \left(\left\lfloor \frac{64(1+\nu)bL_f^2}{\gamma^2 \epsilon^{2\sigma} M^{2/(1+\nu)}} \log \left(\frac{32(1+\nu)bL_f^2}{\gamma^2 \epsilon^{2\sigma} M^{2/(1+\nu)}} \right) \right\rfloor_+ + 1 \right)^{\frac{1+\nu}{2}} \right\},$$

where the last inequality is due to $\epsilon \in (0, 1/e]$, (35), (36), and (37). It then follows from the arbitrariness of k' and the definition of \widehat{K}_ϵ in (77) that $\widehat{K}_\epsilon < k$.

In addition, observe from $k = \widehat{K}_\epsilon$, $\epsilon \in (0, 1/e]$, (33), (34), and (37) that

$$k \geq \max \left\{ \left(\frac{144aL_{\nabla f}^2}{\epsilon^2 M^{2/(1+\nu)}} \right)^{\frac{1+\nu}{2}}, \left(\left\lfloor \frac{144(1+\nu)bL_{\nabla f}^2}{\epsilon^2 M^{2/(1+\nu)}} \log \left(\frac{72(1+\nu)bL_{\nabla f}^2}{\epsilon^2 M^{2/(1+\nu)}} \right) \right\rfloor_+ + 1 \right)^{\frac{1+\nu}{2}} \right\}.$$

It then follows from (92) and Lemma 11 that $\Delta_k/(L_{\lceil k/2 \rceil} k) \leq \epsilon^2/(36L_{\nabla f}^2)$, which implies that

$$L_{\nabla f} \sqrt{\frac{\Delta_k}{L_{\lceil k/2 \rceil} k}} \leq \frac{\epsilon}{6}. \tag{93}$$

Similarly, notice from $k = \widehat{K}_\epsilon$, $\epsilon \in (0, 1/e]$, (30), (32), and (37) that

$$k \geq \max \left\{ \left(\frac{36\underline{L}a}{\epsilon^2} \right)^{\frac{1+\nu}{2\nu}}, \left(\left\lfloor \frac{36(1+\nu)b\underline{L}}{\nu\epsilon^2} \log \left(\frac{18(1+\nu)b\underline{L}}{\nu\epsilon^2} \right) \right\rfloor_+ + 1 \right)^{\frac{1+\nu}{2\nu}} \right\}.$$

It then follows from (91) and Lemma 11 that $\Delta_k/k^{2\nu/(1+\nu)} \leq \epsilon^2/(18\underline{L})$. By this, the definitions of L_ℓ and $\{\delta_\ell\}$, $\underline{L} = L_{\nabla f} + M^{2/(1+\nu)}$, $\nu \in (0, 1]$, and $\delta_\ell \leq 1$, we have

$$\frac{L_k \Delta_k}{k} = \frac{(L_{\nabla f} + \delta_k^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}) \Delta_k}{k} \leq \frac{(L_{\nabla f} + M^{\frac{2}{1+\nu}}) \delta_k^{\frac{\nu-1}{1+\nu}} \Delta_k}{k} = \frac{\underline{L}(k+1)^{\frac{1-\nu}{1+\nu}} \Delta_k}{k} \leq \frac{2\underline{L} \Delta_k}{k^{\frac{2\nu}{1+\nu}}} \leq \frac{\epsilon^2}{9},$$

where the second inequality follows from $(k+1)^{\frac{1-\nu}{1+\nu}} \leq 2k^{\frac{1-\nu}{1+\nu}}$ due to $k \geq 2$. Hence, we obtain

$$\sqrt{\frac{L_k \Delta_k}{k}} \leq \frac{\epsilon}{3}. \tag{94}$$

Also, by $k = \widehat{K}_\epsilon$, $\epsilon \in (0, 1/e]$, (31), (32), and (37), we can see that

$$k \geq \max \left\{ \left(\frac{4a(3M)^{2/\nu}}{M^{2/(1+\nu)}\epsilon^{2/\nu}} \right)^{\frac{1+\nu}{2}}, \left(\left\lfloor \frac{4b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}\epsilon^{2/\nu}} \log \left(\frac{2b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}\epsilon^{2/\nu}} \right) \right\rfloor_+ + 1 \right)^{\frac{1+\nu}{2}} \right\}.$$

It then follows from (92) and Lemma 11 that $\Delta_k/(L_{\lceil k/2 \rceil}k) \leq \epsilon^{2/\nu}/(3M)^{2/\nu}$, which implies that

$$M \left(\frac{\Delta_k}{L_{\lceil k/2 \rceil}k} \right)^{\frac{\nu}{2}} \leq \frac{\epsilon}{3}. \quad (95)$$

Lastly, using $k = \widehat{K}_\epsilon$, (32) and (37) yields $k \geq \lceil 72A\epsilon^{-2} \rceil$. By this and the definition of η_k , one has

$$A^{\frac{1}{2}}\eta_{\lceil k/2 \rceil}^{\frac{1}{2}} \leq \left(\frac{2A}{k+2} \right)^{\frac{1}{2}} \leq \frac{\epsilon}{6}. \quad (96)$$

Recall from the above hypothesis that none of $\{x^\ell\}_{\ell=0}^k$ is a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of (1). Hence, $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$. Moreover, it follows from this and the definition of \overline{K}_ϵ in (78) that $k \leq \overline{K}_\epsilon$. By $\underline{K}_\epsilon < k \leq \overline{K}_\epsilon$ and $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$, it follows from Lemma 10 that (80) holds for such k , which together with the definition of A in (29) leads to

$$\text{dist}(0, \partial\Psi(x^{\ell(k)+1})) \leq L_{\nabla f} \sqrt{\frac{\Delta_k}{L_{\lceil k/2 \rceil}k}} + \sqrt{\frac{L_k\Delta_k}{k}} + M \left(\frac{\Delta_k}{L_{\lceil k/2 \rceil}k} \right)^{\frac{\nu}{2}} + A^{\frac{1}{2}}\eta_{\lceil k/2 \rceil}^{\frac{1}{2}}.$$

Combining this with (93), (94), (95), and (96), we obtain that

$$\text{dist}(0, \partial\Psi(x^{\hat{k}+1})) \leq \frac{\epsilon}{6} + \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{6} = \epsilon.$$

In addition, notice from Algorithm 2 that $\|x^{\hat{k}+1} - x^{\hat{k}}\| \leq \gamma\epsilon^\sigma/(4L_f)$. It follows from these and the definition of \mathcal{X}_ϵ^c in (24) that $x^{\hat{k}} \notin \mathcal{X}_\epsilon^c$. Since $\hat{k} \leq k$, this contradicts the assumption that $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq k$, which is implied by the hypothesis. Hence, Algorithm 2 generates a pair (x^k, y^k) such that x^k is a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1) for some $0 \leq k \leq \widehat{K}_\epsilon$. Moreover, it follows from Lemma 9 that y^k satisfies (38). \square

We next present the proof of Theorem 5.

Proof of Theorem 5. Let \mathcal{X}_ϵ^c be defined in (78). The conclusion clearly holds if $x^0 \notin \mathcal{X}_\epsilon^c$. Hence, we assume for the remainder of the proof that $x^0 \in \mathcal{X}_\epsilon^c$. Given this and $\epsilon \in (0, 1/e]$, it follows from Theorem 4 that there exists $0 \leq \bar{k} < \widehat{K}_\epsilon$ such that $x^\ell \in \mathcal{X}_\epsilon^c$ for all $0 \leq \ell \leq \bar{k}$ and $x^{\bar{k}+1} \notin \mathcal{X}_\epsilon^c$. That is, the iterates $\{x^\ell\}_{\ell=0}^{\bar{k}}$ are not, but $x^{\bar{k}+1}$ is, a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point of problem (1).

We first observe from Algorithm 2 that the number of evaluations of the proximal operator p equals the number of iterations. By this and $\bar{k} < \widehat{K}_\epsilon$, it follows that the total number of evaluations of p to generate the $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point $x^{\bar{k}+1}$ is $\bar{k} + 1 \leq \widehat{K}_\epsilon$.

We next show that the total number of evaluations of the proximal operator of q performed in Algorithm 2 to generate the $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point $x^{\bar{k}+1}$ is at most \widehat{N}_2 . To this end, we analyze the number of evaluations of the proximal operator of q conducted at each iteration $0 \leq \ell' \leq \bar{k}$, through its calls to Algorithm 1. As observed from Algorithm 2 and the proof of Lemma 10, Algorithm 1 is invoked to solve problem (18) with $h(\cdot) = -F(x^{\ell'+1}, \cdot)$, where h satisfies condition (19) with $\delta = \gamma\epsilon^\sigma$. By this, the definitions of τ and $\{\eta_\ell\}$ in Algorithm 2, $\bar{k} < \widehat{K}_\epsilon$, and (22), it follows from Theorem 3 that the number of outer iterations performed in Algorithm 1 at each iteration $0 \leq \ell' \leq \bar{k}$ is at most $\overline{K}_{f,\theta}$, where $\overline{K}_{f,\theta}$ is defined in (39). Using this and Theorem 2, we can see that at each iteration $0 \leq \ell' \leq \bar{k}$, the number of evaluations of q is at most

$$\left(\left\lceil \frac{\log(2L_{\nabla f}\bar{\lambda})}{\log \rho^{-1}} \right\rceil_+ + 1 \right) \overline{K}_{f,\theta}.$$

By this bound, the fact that the total number of iterations performed by Algorithm 2 to generate the $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point $x^{\bar{k}+1}$ is at most \widehat{K}_ϵ , and (40), we conclude that the total number of evaluations of the proximal operator of q performed by Algorithm 2 to generate a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point is at most \widehat{N}_ϵ .

Lastly, notice that the total number of evaluations of ∇f is no more than the sum of the total number of evaluations of the proximal operators of p and q . It then follows that the total number of evaluations of ∇f performed in Algorithm 2 to generate a $(\gamma\epsilon^\sigma/(4L_f), \epsilon)$ -stationary point is at most $\widehat{K}_\epsilon + \widehat{N}_\epsilon$. \square

7 Concluding remarks

In this paper, we consider a class of nonconvex–nonconcave minimax problems for which the inner maximization problem satisfies the local KL condition (7) with parameters $\theta \in [1/2, 1)$, $\gamma > 0$, and $\sigma > 0$. Although minimax problems in which the inner maximization problem satisfies a global KL or PL condition form a special case of this class, the general complexity results developed here may become weak when applied to such problems. This is not surprising, as our complexity analysis targets worst-case instances in this broader class, where the inner maximization problem satisfies only a local KL condition over a region that shrinks as the outer variable approaches a stationary point.

Nevertheless, for the global KL or PL settings, as well as for other scenarios such as (i) $\theta \in (0, 1/2)$, $\gamma > 0$, and $\sigma > 0$, and (ii) $\theta \in (0, 1)$, $\gamma > 0$, and $\sigma = 0$, one can design tailored algorithms together with refined analyses that may yield stronger stationarity guarantees and improved complexity bounds.

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, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical programming*, 137(1):91–129, 2013.
- [3] G. Bento, B. Mordukhovich, T. Mota, and Y. Nesterov. Convergence of descent optimization algorithms under Polyak–Lojasiewicz–Kurdyka conditions. *Journal of Optimization Theory and Applications*, 207(3):41, 2025.
- [4] D. Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.
- [5] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.
- [6] J. Blanchet, J. Li, S. Lin, and X. Zhang. Distributionally robust optimization and robust statistics. *Statistical Science*, 40(3):351–377, 2025.
- [7] A. Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. *arXiv preprint arXiv:2201.12247*, 2022.
- [8] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
- [9] Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. *arXiv preprint arXiv:2210.03096*, 2022.

- [10] F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.
- [11] 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.
- [12] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
- [13] 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.
- [14] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165:874–900, 2015.
- [15] 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.
- [16] F. Huang. Enhanced adaptive gradient algorithms for nonconvex-PL minimax optimization. *arXiv preprint arXiv:2303.03984*, 2023.
- [17] C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.
- [18] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811, 2016.
- [19] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Lojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.
- [24] 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.
- [25] S. Omidshafiei, J. Papis, 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.
- [26] 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.

- [27] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.
- [28] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.
- [29] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [30] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.
- [31] E. M. Stein and R. Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.