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

Zhaosong Lu * Xiangyuan Wang *

July 2, 2025

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 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 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

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

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, 13], reinforcement learning [10, 23], adversarial training [20, 26], and distributionally robust optimization [3, 4, 25]. 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., [15]).

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

*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 Award IIS-2211491.

Polyak-Łojasiewicz (PL) condition on the inner maximization problem of (1), which is generally weaker 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 stronger assumption of concavity in the inner maximization problem of (1) (see, e.g., [14, 22, 28, 29]). 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., [5, 7, 19, 24]).

More recently, [18, 30, 31] 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 [18, 30, 31] is significantly broader than those studied in [14, 22, 28, 29], 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 illustrate this limitation, consider the following simple minimax problem.

Example 1.

$$\min_{1\leq x\leq 2}\ \max_{\pi/4\leq y\leq\pi}\ x(\cos y-1). \tag{2}$$

One can verify that the global KL condition fails to hold for the inner maximization problem of (2) *at any* $x\in[1,2]$*. However, it can be shown that the KL condition holds on the level set* $\mathcal{L}_x=\{y\in[\pi/4,\pi]:0<\phi^*(x)-\phi(x,y)\leq\sqrt{2}x/2\}$ *for all* $x\in[1,2]$*, where* $\phi(x,y)=x(\cos y-1)$ *and* $\phi^*(x)=\max_{\pi/4\leq y\leq\pi}\phi(x,y)$*. Indeed, one can verify that*

$$C(\phi^*(x)-\phi(x,y))^{\theta}\leq \operatorname{dist}\big(0,\nabla_y\phi(x,y)-\mathcal{N}_{[\pi/4,\pi]}(y)\big)\qquad \forall y\in\mathcal{L}_x$$

for all $x\in[1,2]$ *with* $C=2^{1/4}$ *and* $\theta=1/2$*, where* $\operatorname{dist}(\cdot,\cdot)$ *denotes the distance from a point to a set, and* $\mathcal{N}_{[\pi/4,\pi]}(\cdot)$ *is the normal cone to* $[\pi/4,\pi]$ *at the given point.*

The above observation motivates us to 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 $x\in\operatorname{dom}p$, the KL condition holds only on a level set of the inner variable y, where 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 [18, 30, 31], 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 Hölder smooth on the set $\{x:0\notin\partial\Psi(x)\}$, 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 an approximation of $\nabla F^*(x^k)$, along with a carefully chosen step size. We also establish complexity guarantees for the proposed method in computing an approximate stationary point of problem (1).

The main contributions of this paper are summarized below.

- We establish a local Hölder smoothness property for the maximal function F^* under a local KL condition, which plays a crucial role in developing 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 $\widetilde{\mathcal{O}}\big(\epsilon^{-\max\{(1-\theta)^{-1},\theta^{-1}\sigma\}}\big)$, and a *first-order oracle complexity* of $\widetilde{\mathcal{O}}\big(\epsilon^{-(1-\theta)^{-1}(2\theta^2-2\theta+1)\max\{(1-\theta)^{-1},\theta^{-1}\sigma\}}\big)$, measured by the number of gradient evaluations, for finding an $\mathcal{O}(\epsilon)$-approximate stationary point of (1), where θ and σ are the parameters 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 Hölder smoothness property of the maximal function. Section 3 presents a proximal gradient method for minimizing functions that satisfy the 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. Finally, we provide the proof of the main results in Section 6.

1.1 Notation, terminology, and assumptions

The following notation will be used throughout the paper. Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space. The standard inner product, ℓ_1-norm, ℓ_∞-norm, and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$, $\|\cdot\|_1$, $\|\cdot\|_\infty$, and $\|\cdot\|$, respectively. For any two points $u, v \in \mathbb{R}^n$, the notation $[u, v]$ denotes the line segment connecting u and v. Given a point x and a closed set $S \subset \mathbb{R}^n$, let $\mathrm{dist}(x, S)$ denote the distance between x and S. The closed ball centered at $x \in \mathbb{R}^n$ with radius r is denoted by $\mathcal{B}(x, r)$. In addition, $\mathrm{conv}(\cdot)$ denotes the convex hull of the associated set.

A function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is called *L_f-Lipschitz continuous* on Ω if $|f(x)-f(y)| \le L_f\|x-y\|$ for all $x, y \in \Omega$, and *$L_{\nabla f}$-smooth* on Ω if $\|\nabla f(x) - \nabla f(y)\| \le L_{\nabla f}\|x - y\|$ for all $x, y \in \Omega$. More generally, f is said to be *Hölder smooth* on Ω if there exist $L > 0$ and $\nu \in (0, 1]$ such that $\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|^\nu$ for all $x, y \in \Omega$. For an extended real-valued function $\phi : \mathbb{R}^n \to (-\infty, \infty]$, its *domain* is denoted by $\mathrm{dom}\,\phi$, i.e., $\mathrm{dom}\,\phi = \{x : \phi(x) < \infty\}$. For a locally Lipschitz continuous function ϕ, we use $\partial\phi$ to denote its Clarke subdifferential, that is,

$$\partial\phi(x) := \mathrm{conv}\{v : \exists\, x^k \to x \text{ such that } \nabla f(x^k) \to v\} \qquad \forall x \in \mathrm{dom}\,\phi.$$

In addition, we use $\partial_{x_i}\phi$ to denote the Clarke subdifferential with respect to x_i. When ϕ differentiable, then $\partial\phi$ coincides with the gradient $\nabla\phi$. If ϕ is convex, then $\partial\phi$ corresponds to the classical convex subdifferential. It is also well-known that $\partial(\phi_1 + \phi_2)(x) = \nabla\phi_1(x) + \partial\phi_2(x)$ if ϕ_1 is differentiable and ϕ_2 is locally Lipschitz continuous at x (e.g., see [9]).

We now define an approximate stationary point for the problem $\min_x \phi(x)$, where ϕ is a locally Lipschitz continuous function. 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 1. *Suppose ϕ is a locally Lipschitz continuous function. For any $\epsilon > 0$ and $r > 0$, a point x is called an (ϵ, r)-stationary point of the problem $\min_x \phi(x)$ if $\mathrm{dist}(x, \mathcal{S}_\epsilon) \le r$, where $\mathcal{S}_\epsilon = \{x : \mathrm{dist}(0, \partial\phi(x)) \le \epsilon\}$.*

Before ending this subsection, we introduce additional notation and assumptions for problem (1).

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

For convenience, we define

$$\mathcal{X} := \operatorname{dom} p, \quad \mathcal{Y} := \operatorname{dom} q, \quad F(x,y) := f(x,y) - q(y), \tag{3}$$

$$F^*(x) := \max_y F(x,y), \quad Y^*(x) := \{y : F(x,y) = F^*(x)\}, \tag{4}$$

$$\Psi(x) := F^*(x) + p(x), \quad \Psi^* := \min_x \Psi(x). \tag{5}$$

We assume that problem (1) has at least one optimal solution and 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 an open set $\Omega \subset \mathbb{R}^n$ containing $\mathcal{X}$. Moreover, the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is $L_{\nabla f}$-smooth on $\Omega \times \mathcal{Y}$.*

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

(iii) *For any fixed $x \in \Omega$, $\max_y F(x,y)$ 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 $C > 0$, $\theta \in [1/2, 1)$, and $\sigma > 0$ such that for any $x \in \Omega$,*

$$C(F^*(x) - F(x,y))^\theta \leq \operatorname{dist}(0, \partial_y F(x,y)) \qquad \forall y \in \mathcal{L}(x), \tag{6}$$

where

$$\mathcal{L}(x) := \{y : 0 < F^*(x) - F(x,y) \leq \gamma \operatorname{dist}(0, \partial\Psi(x))^\sigma\}.^2 \tag{7}$$

Remark 1. *We refer to condition (6) 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 [18, 30, 31], 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. For example, the minimax problem in Example (2) satisfies the local KL condition but not the global one. 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 over which the KL condition is valid 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.*

2 Local Hölder smoothness of the maximal function

In this section, we establish a local Hölder smoothness property of the maximal function F^*, which will play a crucial role in developing 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:

$$\mathcal{U}_\epsilon := \{x \in \Omega : \operatorname{dist}(0, \partial\Psi(x)) > \epsilon\} \qquad \forall \epsilon > 0. \tag{8}$$

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

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 [22] 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

[2] Under Assumption 1(i), it can be shown that F^* is Lipschitz continuous on Ω (see Lemma 1). Consequently, ∂F^* is well-defined and bounded on Ω. In addition, by convention, we set $\partial p(x) = \emptyset$ for any $x \notin \operatorname{dom} p$. Combining these facts with the definition of Ψ in (5), it follows that $\partial\Psi$ is well-defined on Ω. We also adopt the convention that $\operatorname{dist}(0, \emptyset) = \infty$.

of (1) satisfies only a *local* KL condition as described in Assumption 1(iii). Specifically, it establishes that the maximal function F^* is locally Hölder smooth on $\mathcal{U}_\epsilon$. This property will play a key role in developing 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 $\mathcal{U}_\epsilon$ be defined in (8). Suppose that Assumption 1 holds. Then, F^* is differentiable on $\mathcal{U}_\epsilon$ and*

$$\nabla F^*(x) = \nabla_x f(x, y^*) \qquad \forall x \in \mathcal{U}_\epsilon,\ y^* \in Y^*(x). \tag{9}$$

Moreover, for any $x, x' \in \mathcal{U}_\epsilon$ satisfying $\|x - x'\| \le \gamma\epsilon^\sigma/(2L_f)$, we have

$$\|\nabla F^*(x) - \nabla F^*(x')\| \le L_{\nabla f}\|x - x'\| + (1-\theta)^{-1} C^{-1/\theta} L_{\nabla f}^{1/\theta} \|x - x'\|^{\frac{1-\theta}{\theta}}. \tag{10}$$

Remark 2. *As a consequence of Theorem 1, the maximal function F^* is differentiable on the set $\{x \in \mathcal{X} : 0 \notin \partial\Psi(x)\}$. Moreover, in view of Theorem 1 and the outer semicontinuity of $\partial\Psi$, it is not hard to observe that F^* is locally Hölder smooth on this set.*

3 A proximal gradient method for minimizing KL function

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

$$h^* = \min_z \{h(z) := g(z) + q(z)\}, \tag{11}$$

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

$$C(h(z) - h^*)^\theta \le \operatorname{dist}(0, \partial h(z)) \qquad \forall z \text{ with } h^* < h(z) \le h^* + \delta \tag{12}$$

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

Under a KL condition, general algorithmic frameworks for solving problem (11) and analyzing their convergence properties have been extensively studied in the literature (see, e.g., [2, 12, 17]). Inspired by these works, we now propose a proximal gradient method with backtracking line search for solving problem (11), which 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 (11)

Input: $z^0 \in \{z : h(z) \le h^* + \delta\}$, $\overline{\lambda} > 0$, $\rho \in (0,1)$, and $\tau > 0$.
for $k = 0, 1, 2, \ldots$ **do**
  for $i = 0, 1, 2, \ldots$ **do**
    $\lambda_{k,i} = \overline{\lambda}\rho^i$.
    $z^{k+1,i} = \arg\min_z \left\{\langle \nabla g(z^k), z\rangle + \frac{1}{2\lambda_{k,i}}\|z - z^k\|^2 + q(z)\right\}$.
    if $h(z^{k+1,i}) + \frac{1}{2\lambda_{k,i}}\|z^{k+1,i} - z^k\|^2 \le h(z^k)$ **then**
      $z^{k+1} = z^{k+1,i}$, $\lambda_k = \lambda_{k,i}$.
      break
    end if
  end for
  if $\|z^{k+1} - z^k\| \le \tau$ **then**
    return z^{k+1}.
  end if
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, $\overline{\lambda}, \rho$ be given in Algorithm 1, and*

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

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

$$\min\{\rho/L, \overline{\lambda}\} \leq \lambda_k \leq \overline{\lambda}. \tag{13}$$

The following theorem establishes that Algorithm 1 terminates in a finite number of iterations and yields a desired approximate solution to problem (11). The proof is deferred to Subsection 6.2.

Theorem 3. *Let C, δ, θ be given in* (12)*, $\overline{\lambda}, \rho, \tau$ be given in Algorithm 1, and let*

$$\underline{\lambda} = \min\{\rho/L, \overline{\lambda}\}, \qquad \underline{\beta} = \frac{C^2}{2\overline{\lambda}}\left(L + \underline{\lambda}^{-1}\right)^{-2}, \qquad \overline{\beta} = \frac{C^2}{2\underline{\lambda}}\left(L + \overline{\lambda}^{-1}\right)^{-2}, \tag{14}$$

$$C' = \min\left\{\frac{1}{2}, \frac{(2^{\frac{2\theta-1}{2\theta}} - 1)\delta^{1-2\theta}}{(2\theta-1)\overline{\beta}}\right\}, \qquad \overline{K}_\theta := \begin{cases} \left\lceil \frac{1+\beta}{\underline{\beta}} \log(2\overline{\lambda}\delta\tau^{-2}) \right\rceil_+ + 1 & \text{if } \theta = \frac{1}{2}, \\ \left\lceil \frac{1}{C'(2\theta-1)\underline{\beta}} \left(2\overline{\lambda}\tau^{-2}\right)^{2\theta-1} \right\rceil + 1 & \text{if } \theta \in (\frac{1}{2}, 1). \end{cases} \tag{15}$$

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

$$h(z^{k+1}) - h^* \leq \left(C^{-1}(L + \underline{\lambda}^{-1})\tau\right)^{\frac{1}{\theta}}. \tag{16}$$

4 An inexact proximal gradient method for problem (1)

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

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

$$\mathcal{X}_\epsilon := \{x \in \mathcal{X} : \mathrm{dist}(0, \Psi(x)) \leq \epsilon\}, \qquad \mathcal{X}_\epsilon^{\mathrm{c}} := \{x \in \mathcal{X} : \mathrm{dist}(x, \mathcal{X}_\epsilon) > \gamma\epsilon^\sigma/(4L_f)\}, \tag{17}$$

$$r := \gamma\epsilon^\sigma/(4L_f), \quad M := (1-\theta)^{-1}C^{-1/\theta}L_{\nabla f}^{1/\theta}, \quad \nu := \theta^{-1}(1-\theta), \tag{18}$$

where $C, \theta, \gamma, \sigma, L_f$ are given in Assumption 1.

To propose a method for finding an (ϵ, r)-stationary point of problem (1), we first make some key observations. Suppose $x' \in \mathcal{X}_\epsilon^{\mathrm{c}}$, that is, x' is not an (ϵ, r)-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 $\mathrm{dist}(0, \Psi(z)) > \epsilon$ for all $z \in [x', x]$. In view of these and $\mathcal{X} \subset \Omega$, one can see that $[x', x] \subseteq \mathcal{U}_\epsilon$, where $\mathcal{U}_\epsilon$ is defined in (8). Using this and Theorem 1, we can show that

$$F^*(x) \overset{(10)}{\leq} F^*(x') + \langle \nabla F^*(x'), x - x' \rangle + \frac{1}{2}L_{\nabla f}\|x - x'\|^2 + \frac{M}{1+\nu}\|x - x'\|^{1+\nu} \qquad \forall x \in \mathcal{X} \cap \mathcal{B}(x', r).$$

In addition, notice from $\theta \in [1/2, 1)$ and (18) that $\nu \in (0, 1]$. It then follows from [21, Lemma 2] that

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

Combining the above two inequalities, and using the fact $\Psi(\cdot) = F^*(\cdot) + p(\cdot)$, we obtain that

$$F^*(x) \le F^*(x') + \langle \nabla F^*(x'), x - x' \rangle + \frac{1}{2}\big(L_{\nabla f} + \delta^{\frac{\nu-1}{1+\nu}} M^{\frac{2}{1+\nu}}\big)\|x - x'\|^2 + \frac{\delta}{2} \quad \forall x \in \mathcal{X} \cap \mathcal{B}(x', r), \tag{19}$$

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

As a result, when $x' \in \mathcal{X}$ is not an (ϵ, r)-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 to find an (ϵ, r)-stationary point of problem (1), which generates the sequence $\{x^k\}$ according to

$$x^{k+1} = \mathop{\arg\min}_{x \in \mathcal{B}(x^k, r)} \left\{ \langle \nabla F^*(x^k), x \rangle + \frac{1}{2} L_k \|x - x^k\|^2 + p(x) \right\} \tag{20}$$

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 (ϵ, r)-stationary point of (1) for some $k \ge 0$. However, this method faces a practical limitation: the exact value of $\nabla F^*(x^k)$ 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 $\nabla F^*(x^k)$ in (20) with its approximation $\nabla_x f(x^k, y^k)$, where 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, θ, γ, σ from Assumption 1; $\epsilon > 0$, $\overline{\lambda} > 0$, $\rho \in (0, 1)$, and $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ satisfying $F^*(x^0) - F(x^0, y^0) \le \min\{\gamma\epsilon^\sigma/2, 1\}$ and $\mathrm{dist}(y^0, Y^*(x^0)) \le (C(1-\theta))^{-1} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, 1\}$.

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

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

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

4: Compute

$$x^{k+1} = \mathop{\arg\min}_{x \in \mathcal{B}(x^k, r)} \left\{ \langle \nabla_x f(x^k, y^k), x \rangle + \frac{L_k}{2}\|x - x^k\|^2 + p(x) \right\}. \tag{21}$$

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

6: **end for**

Remark 3. *(i) For the initial point (x^0, y^0), Algorithm 2 requires that y^0 be a nearly optimal solution to the problem $\max_y F(x^0, y)$. Although the inner maximization problem $\max_y F(x, y)$ in (1) is generally nonconcave for arbitrary x, there often exists a particular point $x^0 \in \mathcal{X}$ such that $\max_y F(x^0, y)$ becomes a concave problem in y. In such cases, y^0 can be efficiently computed by solving this concave maximization problem. Moreover, even when $\max_y F(x^0, y)$ is nonconcave, it may still be efficiently solvable for some $x^0 \in \mathcal{X}$, depending on the structure of the problem.*

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

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

Theorem 4. *Let $L_f, L_{\nabla f}, \sigma, \theta$ be given in Assumption 1, M, ν be defined in* (18)*, γ, ϵ be given in Algorithm 2, and*

$$A = (1-\theta)^{-2}C^{-2}L_{\nabla f}^2, \quad \underline{L} = L_{\nabla f} + M^{2/(1+\nu)}, \tag{22}$$

$$a = 8\big(\Psi(x^0) - \Psi^* + 3 + 2A\underline{L}^{-1}\big), \quad b = 8\big(3/2 + A\underline{L}^{-1}\big), \tag{23}$$

$$\widehat{C}_1 = \Big(36(1+\nu)\nu^{-1}b\underline{L}\lceil\log(18(1+\nu)\nu^{-1}b\underline{L})\rceil_+ + 72(1+\nu)\nu^{-1}b\underline{L} + 1\Big)^{\frac{1+\nu}{2\nu}}, \tag{24}$$

$$\widehat{C}_2 = \Big(\frac{4b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}}\Big\lceil\log\Big(\frac{2b(1+\nu)(3M)^{2/\nu}}{M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{8b(1+\nu)(3M)^{2/\nu}}{\nu M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}}, \tag{25}$$

$$\widehat{C}_3 = (36\underline{L}a)^{\frac{1+\nu}{2\nu}} + M^{-1}\big(4a(3M)^{2/\nu}\big)^{\frac{1+\nu}{2}}, \quad \widehat{C}_4 = 72A, \tag{26}$$

$$\widehat{C}_5 = \Big(\frac{144(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}}\Big\lceil\log\Big(\frac{72(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{288(1+\nu)bL_{\nabla f}^2}{M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}}, \tag{27}$$

$$\widehat{C}_6 = (144aL_{\nabla f}^2)^{\frac{1+\nu}{2}}/M, \tag{28}$$

$$\widehat{C}_7 = \Big(\frac{64(1+\nu)bL_f^2}{\gamma^2M^{2/(1+\nu)}}\Big\lceil\log\Big(\frac{32(1+\nu)bL_f^2}{\gamma^2M^{2/(1+\nu)}}\Big)\Big\rceil_+ + \frac{128\sigma(1+\nu)bL_f^2}{\gamma^2M^{2/(1+\nu)}} + 1\Big)^{\frac{1+\nu}{2}}, \tag{29}$$

$$\widehat{C}_8 = (64aL_f^2)^{\frac{1+\nu}{2}}/(\gamma^{1+\nu}M), \tag{30}$$

$$\begin{aligned}\widehat{K}_\epsilon = \Big\lceil &\widehat{C}_1\epsilon^{-\frac{1+\nu}{\nu}}(\log\epsilon^{-1})^{\frac{1+\nu}{2\nu}} + \widehat{C}_2\epsilon^{-\frac{1+\nu}{\nu}}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_3\epsilon^{-\frac{1+\nu}{\nu}} + \widehat{C}_4\epsilon^{-2}\\ &+ \widehat{C}_5\epsilon^{-(1+\nu)}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_6\epsilon^{-(1+\nu)} + \widehat{C}_7\epsilon^{-(1+\nu)\sigma}(\log\epsilon^{-1})^{\frac{1+\nu}{2}} + \widehat{C}_8\epsilon^{-(1+\nu)\sigma}\Big\rceil.\end{aligned} \tag{31}$$

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

$$F^*(x^k) - F(x^k, y^k) \le \min\Big\{\frac{\gamma\epsilon^\sigma}{2}, \frac{1}{k+1}\Big\}, \quad \mathrm{dist}\big(y^k, Y^*(x^k)\big) \le \frac{1}{C(1-\theta)}\min\Big\{\Big(\frac{\gamma}{2}\Big)^{(1-\theta)}\epsilon^{\sigma(1-\theta)}, \frac{1}{\sqrt{k+1}}\Big\}. \tag{32}$$

The next result presents a *first-order oracle complexity* bound for Algorithm 2, measured by the number of evaluations of the gradient ∇f, required to generate an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-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, $\widehat{K}_\epsilon$ be defined in Theorem 4, $L_{\nabla f}, C, \gamma, \theta, \sigma$ be given in Assumption 1, M, ν be defined in* (18)*, $\rho, \overline{\lambda}, \underline{\lambda}$ be given in Algorithm 2, and let*

$$\underline{\beta_f} = \frac{C^2}{2\overline{\lambda}}(L_{\nabla f} + \underline{\lambda}^{-1})^{-2}, \quad \overline{\beta_f} = \frac{C^2}{2\underline{\lambda}}(L_{\nabla f} + \overline{\lambda}^{-1})^{-2},$$

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

$$\overline{K}_{f,\theta} = \begin{cases}\Big\lceil\frac{1+\underline{\beta_f}}{\underline{\beta_f}}\log(2\overline{\lambda}C^{-2}(L_{\nabla f} + \underline{\lambda}^{-1})^2\gamma\epsilon^\sigma\Lambda)\Big\rceil_+ + 1 & \text{if } \theta = \frac{1}{2},\\ \Big\lceil\frac{1}{C_f'(2\theta-1)\underline{\beta_f}}\big(2\overline{\lambda}C^{-2}(L_{\nabla f} + \underline{\lambda}^{-1})^2\Lambda\big)^{2\theta-1}\Big\rceil + 1 & \text{if } \theta \in (\frac{1}{2}, 1),\end{cases} \tag{33}$$

$$\widehat{N}_\epsilon = \widehat{K}_\epsilon\Big(\Big\lceil\frac{\log(2L_{\nabla f}\overline{\lambda})}{\log\rho^{-1}}\Big\rceil_+ + 1\Big)\overline{K}_{f,\theta}. \tag{34}$$

Then the total number of evaluations of the proximal operators of p and q, and the gradient ∇f performed by Algorithm 2 is at most $\widehat{K}_\epsilon$, $\widehat{N}_\epsilon$, and $\widehat{K}_\epsilon + \widehat{N}_\epsilon$, respectively, to generate a pair (x^k, y^k) such that x^k is an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point of of problem (1)*, and y^k satisfies* (32)*.*

Remark 4. *As shown in Theorem 4, Algorithm 2 enjoys an iteration complexity of*

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

to compute an $(\epsilon,\gamma\epsilon^{\sigma}/(4L_f))$*-stationary point of problem* (1)*. Furthermore, as established in Theorem 5, the algorithm requires* $\mathcal{O}\Big(\epsilon^{-\max\left\{\frac{1}{1-\theta},\frac{\sigma}{\theta}\right\}}(\log\epsilon^{-1})^{\frac{1}{2(1-\theta)}}\Big)$ *evaluations of the proximal operator of* p*, and the following number of evaluations of the proximal operator of* q *and the gradient* ∇f *to compute such an approximate stationary point of* (1)*:*

- *If* $\theta=\frac{1}{2}$*,*
$$\mathcal{O}\left(\epsilon^{-2\max\{1,\sigma\}}(\log\epsilon^{-1})^{2}\right).$$
- *If* $\theta\in(\frac{1}{2},1)$*,*
$$\mathcal{O}\left(\epsilon^{-\frac{2\theta^2-2\theta+1}{1-\theta}\max\left\{\frac{1}{1-\theta},\frac{\sigma}{\theta}\right\}}(\log\epsilon^{-1})^{\frac{2\theta^2-2\theta+1}{2(1-\theta)^2}}\right).$$

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:

$$\min_{\|x\|\leq 1}\ \max_{\|y\|_\infty\leq 2}\ \left\{0.01\|x\|_1-\|(y+Ax)\odot(y+Bx)\|^2+0.01\,\|x-c\|^2-0.1\|y\|_1\right\},\tag{35}$$

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 (35) by sampling the entries of A, B, and c independently from the standard normal distribution $\mathcal{N}(0,1)$. Note that problem (35) 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 (35), 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 ith 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

$$\begin{aligned}
&\nabla_x f(x,y)=-2\sum_{i=1}^m w_i(v_ia^i+u_ib^i)+0.02(x-c),\quad \nabla_y f(x,y)=-2\,(u+v)\odot w,\\
&\nabla^2_{xx}f(x,y)=-2\sum_{i=1}^m\left[(v_ia^i+u_ib^i)(v_ia^i+u_ib^i)^T+w_i\big(a^i(b^i)^T+b^i(a^i)^T\big)\right]+0.02I,\\
&\nabla^2_{xy}f(x,y)=-2\big[A^T\operatorname{diag}(v^2+2\,u\odot v)+B^T\operatorname{diag}(u^2+2\,u\odot v)\big],\\
&\nabla^2_{yy}f(x,y)=-2\ \operatorname{diag}((u+v)^2+2\,u\odot v),
\end{aligned}\tag{36}$$

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

$$\begin{aligned}
L_f&=4m(M_aM_b+2M_a+2M_b+4)(M_aM_b+M_a+M_b)+0.02(1+\|c\|),\\
L_{\nabla f}&=4m\big[2(M_aM_b+M_a+M_b)^2+M_aM_b(M_aM_b+2M_a+2M_b+4)\big]\\
&\quad+2\big[\|A\|(M_b+2)(2M_a+M_b+6)+\|B\|(M_a+2)(M_a+2M_b+6)\big]\\
&\quad+2\big[(M_a+M_b+4)^2+2(M_a+2)(M_b+2)\big]+0.02,
\end{aligned}\tag{37}$$

where $\|A\|$ and $\|B\|$ denote the spectral norms of A and B, respectively. In view of (36) and (37), one can verify that $L_f \geq \max_{x\in\mathcal{X},\, y\in\mathcal{Y}} \|\nabla_x f(x,y)\|$, and

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

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 (35) on the randomly generated instances described above. The parameters L_f and $L_{\nabla f}$ are computed using (37), while the remaining parameters are set as follows: $C = 0.2$, $\theta = 0.5$, $\gamma = 0.01$, $\sigma = 0.1$, $\overline{\lambda} = 1$, $\rho = 0.95$, and $\epsilon = 10^{-2}$. The algorithm is initialized at $(x^0, y^0) = (0, 0)$. Note that for this initialization, $y^0 = \arg\max_y f(x^0, y)$, 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 (35), 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 actual final objective value of problem (35), defined as

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

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 approximate final objective value by

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

using the approximate inner solution y_ϵ returned by the algorithm.

The computational results on the random instances are presented in Table 1. The first two columns list the values of m and n. For each pair (m, n), the average initial, actual final, and approximate final objective values over 10 random instances are reported in the remaining columns. From the results, we observe that the approximate solution x_ϵ significantly reduces the objective value compared to the initial point x^0, and that y_ϵ is a good approximate solution to the inner maximization problem $\max_{\|y\|_\infty \leq 2}\{f(x_\epsilon, y) - 0.1\|y\|_1\}$.

Table 1: Numerical results for Algorithm 2

n	m	Initial objective value	Actual final value	Approximate final 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 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.

6.1 Proof of the main results in Section 2

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

The following lemma concerns the Lipschitz continuity of F^* on Ω.

Lemma 1. *Suppose that Assumption 1 holds. Then F^* is L_f-Lipschitz continuous on Ω.*

Proof. Fix any $x, x' \in \Omega$. Recall from Assumption 1 that $f(\cdot, y)$ is L_f-Lipschitz continuous on Ω for any $y \in \mathcal{Y}$. Using this and the expression of F in (3), we have

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

This together with the definition of F^* in (4) implies that

$$F^*(x) \overset{(4)}{=} \max_{y\in\mathcal{Y}} F(x,y) \leq \max_{y\in\mathcal{Y}} F(x',y) + L_f\|x - x'\| = F^*(x') + L_f\|x - x'\|,$$

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 $x, x' \in \Omega$, we conclude that F^* is L_f-Lipschitz continuous on Ω. $\square$

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 (4) that $F^*(x) = F(x,y)$ and $F^*(x+td) \geq F(x+td)$ 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 an open set, then its differentiability extends to the entire set.

Lemma 3. *Let $\Gamma \subset \Omega$ be an open set, and $\mathcal{S} = \{x \in \Omega : F^* \text{ is differentiable at } x\}$. Suppose that Assumption 1 holds, and there exist constants $c > 0$, $\alpha > 0$, and $\eta > 0$ such that*

$$\|\nabla F^*(u) - \nabla F^*(v)\| \leq c\|u - v\|^\alpha \qquad \forall\, u, v \in \mathcal{S} \cap \Gamma \text{ with } \|u - v\| \leq \eta. \tag{38}$$

Then F^* is differentiable on Γ.

Proof. Fix any $x \in \Gamma$. Recall from Lemma 1 that F^* is Lipschitz continuous on the open set Ω. It follows from Rademacher's theorem that the set $\mathcal{S}$ differs from Ω only by a set of measure zero. Since $x \in \Omega$, there exists at least one sequence in $\mathcal{S}$ that converges to x.

Let $\{x^k\} \subset \mathcal{S}$ be an arbitrary sequence such that $x^k \to x$. Since Γ is an open set 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 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 $k, k' > K$. By (38), one then has

$$\|\nabla F^*(x^k) - \nabla F^*(x^{k'})\| \leq c\|x^k - x^{k'}\|^\alpha \qquad \forall\, k, k' > K,$$

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 $\{\tilde{x}^k\} \to x$. Interleaving $\{x^k\}$ and $\{\tilde{x}^k\}$, we obtain a sequence $\{z^k\} \subset \mathcal{S}$ such that $z^k \to x$. 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. It follows that the set

$$\mathrm{conv}\left\{\lim_{k\to\infty} \nabla F^*(\hat{x}^k) : \{\hat{x}^k\} \subset \mathcal{S},\ \hat{x}^k \to x\right\}$$

is a singleton. Combined with the Lipschitz continuity of F^* at x (see Lemma 1), this implies that the Clarke subdifferential of F^* at x is a singleton. By this and [8, Proposition (1.13)], we conclude that F^* is differentiable at 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 [11, 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 [16, 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 (6) 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 $x \in \Omega$,*

$$F^*(x) - F(x,y) \geq (C(1-\theta))^{\frac{1}{1-\theta}} \operatorname{dist}(y, Y^*(x))^{\frac{1}{1-\theta}} \qquad \forall y \in \mathcal{L}(x). \tag{39}$$

Proof. Fix any $x \in \Omega$ and $y \in \mathcal{L}(x)$. It then follows from the definition of $\mathcal{L}(x)$ in (7) that $y \notin Y^*(x)$.

Recall from Assumption 1 that f is Lipschitz smooth on $\Omega \times \mathcal{Y}$. 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 $y \in \operatorname{dom} F(x,\cdot)$. By these and [6, Theorem 13], there exists a unique absolutely continuous curve $Y : [0,\infty) \to \mathbb{R}^m$ satisfying

$$Y(0) = y, \quad \dot{Y}(t) \in \partial_y F(x, Y(t)) \quad \text{a.e. } t > 0, \tag{40}$$

$$\frac{d}{dt} F(x, Y(t)) = \|\dot{Y}(t)\|^2 \quad \text{a.e. } (\eta, \infty) \tag{41}$$

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, non-increasing, and continuous on $[0,\infty)$. We next show that

$$\frac{d}{dt} r(t) \leq -C(1-\theta)\|\dot{Y}(t)\| \quad \text{a.e. } (\eta, \infty) \tag{42}$$

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, (6), (40), and (41) that

$$\begin{aligned} \frac{d}{dt} r(t) &\overset{(41)}{=} -(1-\theta)\big(F^*(x) - F(x, Y(t))\big)^{-\theta} \|\dot{Y}(t)\|^2 \overset{(6)}{\leq} -C(1-\theta) \operatorname{dist}\big(0, \partial_y F(x, Y(t))\big)^{-1} \|\dot{Y}(t)\|^2 \\ &\overset{(40)}{\leq} -C(1-\theta)\, \|\dot{Y}(t)\|^{-1} \|\dot{Y}(t)\|^2 = -C(1-\theta)\, \|\dot{Y}(t)\| \quad \text{a.e. } (\eta, \infty), \end{aligned}$$

and hence (42) 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 (41) with η replaced by t_0 that $\|\dot{Y}(t)\| = 0$ almost everywhere on (t_0, ∞). Hence, we obtain

$$\frac{d}{dt} r(t) \leq -C(1-\theta)\|\dot{Y}(t)\| \quad \text{a.e. } (t_0, \infty). \tag{43}$$

It follows that (42) 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 (43), we see that (42) holds in this case as well.

Fix any $T > 0$ and $\delta \in (0, T)$. By (42), the monotonicity of r, and the absolute continuity of $Y(\cdot)$, one has

$$\begin{aligned} r(T) - r(\delta) &\leq \int_\delta^T \frac{d}{dt} r(t)\, dt \overset{(42)}{\leq} -C(1-\theta) \int_\delta^T \|\dot{Y}(t)\|\, dt \leq -C(1-\theta) \left\| \int_\delta^T \dot{Y}(t)\, dt \right\| \\ &= -C(1-\theta)\|Y(T) - Y(\delta)\|, \end{aligned}$$

where the first inequality follows from the monotonicity of r (e.g., see [27, 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 $\delta \to 0$, and using $Y(0) = y$ and the continuity of r and $Y(\cdot)$, we obtain

$$r(T) - r(0) \leq -C(1-\theta)\|Y(T) - y\|. \tag{44}$$

We next show that $\lim_{T\to\infty} r(T) = 0$. 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, (6), (40), (41), and the monotonicity of r, one has that for any $T > 0$ and $\delta \in (0, T)$,

$$\begin{aligned} r(T) - r(\delta) &\leq \int_\delta^T \frac{d}{dt} r(t)\, dt \overset{(41)}{=} -(1-\theta) \int_\delta^T (F^*(x) - F(x, Y(t)))^{-\theta} \|\dot{Y}(t)\|^2\, dt \\ &\overset{(40)}{\leq} -(1-\theta) \int_\delta^T (F^*(x) - F(x, Y(t)))^{-\theta} \operatorname{dist}(0, \partial_y F(x, Y(t)))^2\, dt \\ &\overset{(6)}{\leq} -C^2(1-\theta) \int_\delta^T (F^*(x) - F(x, Y(t)))^{\theta}\, dt = -C^2(1-\theta) \int_\delta^T r(t)^{\frac{\theta}{1-\theta}}\, dt \\ &\leq -C^2(1-\theta)(T-\delta) r(T)^{\frac{\theta}{1-\theta}}, \end{aligned}$$

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 $\delta \to 0$, and using the continuity of r, we obtain that $-r(0) \leq -C^2(1-\theta) T r(T)^{\frac{\theta}{1-\theta}}$, which yields $r(T) \leq [r(0)/(C^2(1-\theta)T)]^{\frac{1-\theta}{\theta}}$. This along with $r(T) > 0$ implies that $r(T) \to 0$ as $T \to \infty$.

By (44) and the nonnegativity of r, one can observe that the range of $Y(\cdot)$ is bounded. In addition, notice from Assumption 1 that $\operatorname{dom} F(x, \cdot)$ is closed. By these facts, there exists a sequence $\{t_k\} \subset (0, \infty)$ such that $t_k \to \infty$ and $\{Y(t_k)\}$ converges to some point $y^* \in \operatorname{dom} F(x, \cdot)$. Recall that $\lim_{t\to\infty} r(t) = 0$, which along with $t_k \to \infty$ implies that $r(t_k) \to 0$. It then follows that $\lim_{k\to\infty} F(x, Y(t_k)) = F^*(x)$. On the other hand, by the upper semicontinuity of $F(x, \cdot)$ and $Y(t_k) \to y^* \in \operatorname{dom} F(x, \cdot)$, one has $\limsup_{k\to\infty} F(x, Y(t_k)) \leq F(x, y^*)$. Combining these relations, we conclude that $y^* \in Y^*(x)$. Finally, letting $T = t_k$ in (44), 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 $k \to \infty$, and using the fact that $r(t_k) \to 0$ and $Y(t_k) \to y^* \in Y^*(x)$, we obtain that

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

which together with the expression of r implies that the conclusion (39) holds. □

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 [11, Theorem 3.7], provides a relationship between $\operatorname{dist}(y, Y^*(x))$ and $\operatorname{dist}(0, \partial_y F(x, y))$.

Lemma 5. *Suppose that Assumption 1 holds. Then it holds that for any $x \in \Omega$,*

$$\operatorname{dist}(y, Y^*(x)) \leq (1-\theta)^{-1} C^{-\frac{1}{\theta}} \operatorname{dist}\left(0, \partial_y F(x,y)\right)^{\frac{1-\theta}{\theta}} \qquad \forall y \in \mathcal{L}(x). \tag{45}$$

Proof. The relation (45) follows from (6) and (39). □

The next result concerns the openness of the set $\mathcal{U}_\epsilon$.

Lemma 6. *Let $\mathcal{U}_\epsilon$ be defined in* (8)*. Suppose that Assumption 1 holds. Then $\mathcal{U}_\epsilon$ is an open set for any $\epsilon > 0$.*

Proof. Fix any $\epsilon > 0$. Recall from (8) and Assumption 1 that $\mathcal{U}_\epsilon = \{x \in \Omega : \operatorname{dist}(0, \partial\Psi(x)) > \epsilon\}$, and the set Ω is open. Consequently, to prove the openness of the set $\mathcal{U}_\epsilon$, it suffices to show that the set $\{x \in \mathbb{R}^n : \operatorname{dist}(0, \partial\Psi(x)) > \epsilon\}$ is open, or equivalently, $\mathcal{C} = \{x \in \mathbb{R}^n : \operatorname{dist}(0, \partial\Psi(x)) \leq \epsilon\}$ is closed. To this end, consider any convergent sequence $\{x^k\} \subset \mathcal{C}$ with $x^k \to x$ for some $x \in \mathbb{R}^n$. Clearly, $\mathcal{C} \subseteq \mathcal{X}$ and hence $\{x^k\} \subseteq \mathcal{X}$. It then follows from $x^k \to x$ 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, assume that $s^k \to s$ for some s with $\|s\| \leq \epsilon$. Recall from Lemma 1 that F^* is Lipschitz continuous on an open set containing $\mathcal{X}$, which implies that ∂F^* is bounded on $\mathcal{X}$ and hence $\partial^\infty F^*(z) = \{0\}$ for all $z \in \mathcal{X}$. By this and $\Psi = F^* + p$, 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 $x^k \to x$, $s^k \in \partial\Psi(x^k)$, and $s^k \to s$ 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. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Fix any $\epsilon > 0$ and $x \in \mathcal{U}_\epsilon$. It follows that $x \in \Omega$ and $\operatorname{dist}(0, \partial\Psi(x)) > \epsilon$. This together with (7) implies that $\{y : F^*(x) > F(x,y) \geq F^*(x) - \gamma\epsilon^\sigma\} \subseteq \mathcal{L}(x)$. By this and (45), one has

$$\operatorname{dist}(y, Y^*(x)) \leq (1-\theta)^{-1} C^{-\frac{1}{\theta}} \operatorname{dist}(0, \partial_y F(x,y))^{\frac{1-\theta}{\theta}} \qquad \forall y \ \text{ with } F^*(x) > F(x,y) \geq F^*(x) - \gamma\epsilon^\sigma. \tag{46}$$

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 (46) that

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

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 f on $\Omega \times \mathcal{Y}$, we obtain

$$\partial_y F(x, y^*(x')) = \nabla_y f(x, y^*(x')) - \partial q(y^*(x')), \qquad \partial_y F(x', y^*(x')) = \nabla_y f(x', y^*(x')) - \partial q(y^*(x')). \tag{47}$$

The first relation in (47) 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 (47) 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

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

Combining this with (46) yields

$$\mathrm{dist}(y^*(x'), Y^*(x)) \le (1-\theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1-\theta}{\theta}} \|x'-x\|^{\frac{1-\theta}{\theta}}.$$

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

$$\|y^*(x') - y^*(x)\| \le (1-\theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1-\theta}{\theta}} \|x'-x\|^{\frac{1-\theta}{\theta}}. \tag{48}$$

In addition, notice from Lemma 1 and $\mathcal{U}_\epsilon \subseteq \Omega$ that F^* is Lipschitz continuous on $\mathcal{U}_\epsilon$. Moreover, $\mathcal{U}_\epsilon$ is an open set due to Lemma 6. It then follows that F^* is differentiable almost everywhere on $\mathcal{U}_\epsilon$. Without loss generality, assume 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

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

By these, (48), $\theta \in [1/2, 1)$, $\|x - x'\| \le \gamma\epsilon^\sigma/(2L_f)$, and the Lipschitz smoothness of f, one has

$$\begin{aligned}
&\|\nabla F^*(x') - \nabla F^*(x)\| = \|\nabla_x f(x', y^*(x')) - \nabla_x f(x, y^*(x))\| \\
&\le \|\nabla_x f(x', y^*(x')) - \nabla_x f(x, y^*(x'))\| + \|\nabla_x f(x, y^*(x')) - \nabla_x f(x, y^*(x))\| \\
&\le L_{\nabla f}\|x'-x\| + L_{\nabla f}\|y^*(x') - y^*(x)\| \overset{(48)}{\le} L_{\nabla f}\|x'-x\| + (1-\theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1}{\theta}} \|x'-x\|^{\frac{1-\theta}{\theta}} &(49)\\
&\le L_{\nabla f}\Big(\big(\gamma\epsilon^\sigma/(2L_f)\big)^{\frac{2\theta-1}{\theta}} + (1-\theta)^{-1} C^{-\frac{1}{\theta}} L_{\nabla f}^{\frac{1-\theta}{\theta}}\Big)\|x'-x\|^{\frac{1-\theta}{\theta}}, &(50)
\end{aligned}$$

where the last inequality follows from $\|x - x'\| \le \gamma\epsilon^\sigma/(2L_f)$ and $\theta \in [1/2, 1)$. Hence, inequality (50) holds for all $x, x' \in \mathcal{U}_\epsilon \cap \mathcal{S}$ with $\|x - x'\| \le \gamma\epsilon^\sigma/(2L_f)$, where $\mathcal{S}$ is defined in Lemma 3. Using this, the openness of $\mathcal{U}_\epsilon$, and Lemma 3, we conclude that F^* is differentiable on $\mathcal{U}_\epsilon$. In addition, the relations (9) and (10) directly follow from Lemma 2 and (49), respectively. □

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 kth outer iteration. Then one can observe from Algorithm 1 that

$$h(z^{k+1,\bar{i}}) + \frac{1}{2\lambda_{k,\bar{i}}}\|z^{k+1,\bar{i}} - z^k\|^2 > h(z^k). \tag{51}$$

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

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

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

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

Combining these two inequalities yields

$$h(z^{k+1,\bar{i}}) + \Big(\frac{1}{\lambda_{k,\bar{i}}} - \frac{L}{2}\Big)\|z^{k+1,\bar{i}} - z^k\|^2 \le h(z^k). \tag{52}$$

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 (52) 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 (51). 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 3 holds. □

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 $\mathrm{dist}(0, \partial h(z^{k+1}))$ 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*

$$\mathrm{dist}(0, \partial h(z^{k+1})) \leq \big(L + \lambda_k^{-1}\big)\|z^{k+1} - z^k\|. \tag{53}$$

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

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

and hence the conclusion holds. □

For notational convenience, let

$$a_k = \frac{1}{2\lambda_k}, \qquad b_k = \left(L + \frac{1}{\lambda_k}\right)^{-1}. \tag{54}$$

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

$$h(z^{k+1}) + a_k\|z^{k+1} - z^k\|^2 \leq h(z^k), \tag{55}$$

$$b_k\, \mathrm{dist}(0, \partial h(z^{k+1})) \leq \|z^{k+1} - z^k\|, \tag{56}$$

$$(2\bar{\lambda})^{-1} \leq a_k \leq (2\underline{\lambda})^{-1}, \qquad (L + \underline{\lambda}^{-1})^{-1} \leq b_k \leq (L + \bar{\lambda}^{-1})^{-1}, \tag{57}$$

where $\underline{\lambda}$ is defined in (14). In addition, by (55) 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 [12, Theorem 3.4].

Lemma 8. *Let $\delta, \theta, \underline{\lambda}, \underline{\beta}, \bar{\beta}, C'$ and $\bar{\lambda}$ be given in (12), (14), (15), (54) 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

$$h(z^k) - h^* \leq \delta e^{-\frac{\underline{\beta}}{1+\underline{\beta}}k}. \tag{58}$$

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

$$h(z^k) - h^* \leq \left(\frac{1}{C'(2\theta - 1)\underline{\beta}}\right)^{\frac{1}{2\theta-1}} k^{-\frac{1}{2\theta-1}}. \tag{59}$$

Proof. For notational convenience, let $r_\ell = h(z^\ell) - h^*$ for all ℓ. Since $h(z^0) - h^* \le \delta$ and $\{h(z^\ell)\}$ is nonincreasing, (12) holds with $z = z^\ell$ for all $\ell \ge 0$. By this, (55), and (56), one has

$$r_\ell - r_{\ell+1} \overset{(55)}{\ge} a_\ell \|z^{\ell+1} - z^\ell\|^2 \overset{(56)}{\ge} a_\ell b_\ell^2 \operatorname{dist}\left(0, \partial h(z^{\ell+1})\right)^2 \overset{(12)}{\ge} a_\ell b_\ell^2 C^2 r_{\ell+1}^{2\theta}.$$

Let $\beta_\ell := a_\ell b_\ell^2 C^2$ for all ℓ. Using (57) and (14), we have

$$r_\ell - r_{\ell+1} \ge \beta_\ell r_{\ell+1}^{2\theta}, \qquad \beta_\ell \in [\underline{\beta}, \overline{\beta}], \tag{60}$$

(i) Suppose $\theta = 1/2$. It then follows from (60) that $r_{\ell+1} \le (1+\beta_\ell)^{-1} r_\ell$ for all ℓ. Hence,

$$r_k \le r_0 \prod_{\ell=0}^{k-1} (1+\beta_\ell)^{-1} \qquad \forall k \ge 0. \tag{61}$$

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

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

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

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

which together with (61) and $r_0 \le \delta$ implies that (58) holds.

(ii) Suppose $\theta \in (1/2, 1)$. Fix any $k \ge 1$. If $r_k = 0$, (59) clearly holds. Now we assume that $r_k > 0$. It then follows from the monotonicity of $\{r_\ell\}$ that $r_\ell > 0$ for all $0 \le \ell < k$. Let $\psi(t) = \frac{1}{2\theta-1} t^{1-2\theta}$. Then we have

$$\psi(r_{\ell+1}) - \psi(r_\ell) = \int_{r_\ell}^{r_{\ell+1}} \psi'(t) dt = \int_{r_{\ell+1}}^{r_\ell} t^{-2\theta} dt \ge r_\ell^{-2\theta}(r_\ell - r_{\ell+1}) \qquad \forall 0 \le \ell < k. \tag{62}$$

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

Case a): $r_{\ell+1}^{-2\theta} \le 2 r_\ell^{-2\theta}$. It along with (60) and (62) implies that

$$\psi(r_{\ell+1}) - \psi(r_\ell) \ge \frac{1}{2} r_{\ell+1}^{-2\theta}(r_\ell - r_{\ell+1}) \overset{(60)}{\ge} \frac{1}{2}\beta_\ell.$$

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

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

Combining the above two cases, and using the definition of C' in (14), we obtain that $\psi(r_{\ell+1}) - \psi(r_\ell) \ge C'\beta_\ell$ for all $0 \le \ell < k$. It then follows that

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

This and the expression of ψ lead to

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

and hence (59) holds. □

We are now ready to prove Theorem 3.

Proof of Theorem 3. Suppose for contradiction that Algorithm 1 runs for more than $\overline{K}_\theta$ outer iterations. Then there exists some $\ell \geq \overline{K}_\theta - 1$ such that $\|z^{\ell+1} - z^\ell\| > \tau$. By (55) and (57), one has

$$\|z^{\ell+1} - z^\ell\| \overset{(55)}{\leq} \sqrt{\frac{r_\ell - r_{\ell+1}}{a_\ell}} \leq a_\ell^{-\frac{1}{2}} r_\ell^{\frac{1}{2}} \overset{(57)}{\leq} (2\overline{\lambda})^{\frac{1}{2}} r_\ell^{\frac{1}{2}}, \tag{63}$$

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

Case (i): $\theta = 1/2$. By this, (15), and $\ell \geq \overline{K}_\theta - 1$, one has $\ell \geq \underline{\beta}^{-1}(1+\underline{\beta}) \log(2\overline{\lambda}\delta\tau^{-2})$. Using this relation and (58), we have $r_\ell \leq \delta e^{-\underline{\beta}(1+\underline{\beta})^{-1}\ell} \leq \tau^2/(2\overline{\lambda})$.

Case (ii): $\theta \in (1/2, 1)$. Using this, (15), and $\ell \geq \overline{K}_\theta - 1$, we obtain that $\ell \geq \frac{1}{C'(2\theta-1)\underline{\beta}} \left(2\overline{\lambda}\tau^{-2}\right)^{2\theta-1}$. By this relation and (59), one has

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

We thus conclude that $r_\ell \leq \tau^2/(2\overline{\lambda})$. This together with (63) implies $\|z^{\ell+1} - z^\ell\| \leq \tau$, which leads to a contradiction. Hence, Algorithm 1 runs at most $\overline{K}_\theta$ outer iterations.

We next show that (16) holds. Notice from (13) and (54) that $\lambda_k \geq \underline{\lambda}$. By this and (53), one has

$$\mathrm{dist}(0, \partial h(z^{k+1})) \leq \left(L + \underline{\lambda}^{-1}\right)\|z^{k+1} - z^k\|. \tag{64}$$

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 (12), we have

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

By this, (64), and $\|z^{k+1} - z^k\| \leq \tau$, one has

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

and hence (16) holds as desired. □

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 $\gamma, \sigma, C, \theta, L_{\nabla f}, \epsilon, \mathcal{X}_\epsilon^{\mathrm{c}}$, and $\{\eta_\ell\}$ be given in (17), 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 $x^\ell \in \mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0 \leq \ell < k$. Then, for all $0 \leq \ell \leq k$, it holds that*

$$F^*(x^\ell) - F(x^\ell, y^\ell) \leq \min\left\{\gamma\epsilon^\sigma/2, \eta_\ell\right\}, \qquad \mathrm{dist}\left(y^\ell, Y^*(x^\ell)\right) \leq \frac{1}{C(1-\theta)} \min\left\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\right\}, \tag{65}$$

$$\|\nabla F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell)\| \leq \frac{L_{\nabla f}}{C(1-\theta)} \min\left\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, \eta_\ell^{1/2}\right\}. \tag{66}$$

Proof. We first show that (65) and (66) hold for $\ell = 0$. One can observe from Algorithm 2 that (65) holds for $\ell = 0$. We now show that (66) also holds for $\ell = 0$. By the assumption in this lemma, we know that $x^0 \in \mathcal{X}_\epsilon^{\mathrm{c}}$ and hence $\mathrm{dist}(0, \partial\Psi(x^0)) > \epsilon$. This together with $x^0 \in \mathcal{X} \subset \Omega$ implies that $x^0 \in \mathcal{U}_\epsilon$. It then follows from Theorem 1 that F^* is differentiable at x^0, and moreover, $\nabla F^*(x^0) =$

$\nabla_x f(x^0, y^*)$, where $y^* \in Y^*(x^0)$ with $\|y^* - y^0\| = \mathrm{dist}(y^0, Y^*(x^0))$. Using this, $\mathrm{dist}(y^0, Y^*(x^0)) \le (C(1-\theta))^{-1} \min\{(\gamma/2)^{1-\theta}\epsilon^{\sigma(1-\theta)}, 1\}$, and the $L_{\nabla f}$-smoothness of f, one has

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

This together with $\eta_0 = 1$ implies that (66) holds for $\ell = 0$ as desired.

We next show that (65) and (66) hold for $0 < \ell \le k$. 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 $\overline{\lambda}, \rho, \tau$ specified in Algorithm 2. Then it follows from $\tau = \frac{C}{L_{\nabla f}+\underline{\lambda}^{-1}} \min\left\{(\frac{1}{2}\gamma\epsilon^\sigma)^\theta, \eta_\ell^{\frac{\theta}{2(1-\theta)}}\right\}$, the definitions of F^* and F, and Theorem 3 with $h(\cdot) = -f(x^\ell, \cdot) + q(\cdot)$ that

$$F^*(x^\ell) - F(x^\ell, y^\ell) \le [C^{-1}(L_{\nabla f} + \underline{\lambda}^{-1})\tau]^{\frac{1}{\theta}} = \min\left\{\frac{1}{2}\gamma\epsilon^\sigma, \eta_\ell^{\frac{1}{2(1-\theta)}}\right\}, \tag{67}$$

which, together with $\eta_\ell \in (0, 1)$ and $\theta \in [1/2, 1)$, implies that the first relation in (65) holds for $\ell > 0$. In addition, notice from the assumption that $x^{\ell-1} \in \mathcal{X}^{\mathrm{c}}_\epsilon$. Also, observe from Algorithm 2 that $r = \gamma\epsilon^\sigma/(4L_f)$ and $\|x^{\ell+1} - x^\ell\| \le r$. It then follows that $\|x^{\ell+1} - x^\ell\| \le \gamma\epsilon^\sigma/(4L_f)$, which together with $x^{\ell-1} \in \mathcal{X}^{\mathrm{c}}_\epsilon$ implies that $\mathrm{dist}(0, \partial\Psi(x^\ell)) > \epsilon$. Using this and (67), we obtain that $F^*(x^\ell) - F(x^\ell, y^\ell) \le \gamma\mathrm{dist}(0, \partial\Psi(x^\ell))^\sigma$ and hence $y^\ell \in \mathcal{L}(x^\ell)$, where $\mathcal{L}(\cdot)$ is defined in (7). By this, (67), $x^\ell \in \mathcal{X} \subset \Omega$, and Lemma 4, one can conclude that the second relation in (65) holds for $\ell > 0$. Lastly, (66) also holds for $\ell > 0$, due to the second relation in (65) and arguments similar to those used in the case $\ell = 0$. □

Lemma 10. *Let $\epsilon > 0$ be given, M, $\mathcal{X}^{\mathrm{c}}_\epsilon$ be defined in* (17) *and* (18)*, $L_f, L_{\nabla f}, C, \theta, \gamma, \sigma, \{\delta_\ell\}, \{\eta_\ell\}, \{L_\ell\}$ be given in Assumption 1 and Algorithm 2, and let*

$$\Delta_k := 8\Big[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^{k}\Big(1 + \tfrac{L_{\nabla f}^2}{(1-\theta)^2C^2L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k}\frac{\delta_\ell}{2}\Big], \tag{68}$$

$$\underline{K}_\epsilon := \max\{k \ge 1 : \Delta_k/(kL_{\lceil k/2\rceil}) \ge \gamma^2\epsilon^{2\sigma}/(16L_f^2)\}, \tag{69}$$

$$\overline{K}_\epsilon := \max\{k \ge 0 : x^k \in \mathcal{X}^{\mathrm{c}}_\epsilon\}, \tag{70}$$

$$\ell(k) := \operatorname*{arg\,min}_{\lceil k/2\rceil \le \ell \le k} L_\ell\|x^{\ell+1} - x^\ell\|^2. \tag{71}$$

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

$$\mathrm{dist}\left(0, \partial\Psi(x^{\ell(k)+1})\right) \le L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}} + \sqrt{\frac{L_k\Delta_k}{k}} + M\Big(\frac{\Delta_k}{L_{\lceil k/2\rceil}k}\Big)^{\frac{\nu}{2}} + (1-\theta)^{-1}C^{-1}L_{\nabla f}\eta_{\lceil k/2\rceil}^{\frac{1}{2}}. \tag{72}$$

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

$$F(x^{\ell+1}, y^{\ell+1}) + p(x^{\ell+1}) \le F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4}\|x^{\ell+1} - x^\ell\|^2 + \Big(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2C^2L_\ell}\Big)\eta_\ell + \frac{\delta_\ell}{2}. \tag{73}$$

To this end, let us fix any $0 \le \ell \le k$. By optimality condition of (21), one has

$$\langle\nabla_x f(x^\ell, y^\ell), x^{\ell+1}\rangle + L_\ell\|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) \le \langle\nabla_x f(x^\ell, y^\ell), x^\ell\rangle + p(x^\ell). \tag{74}$$

Observe from Algorithm 2 that $\|x^{\ell+1} - x^\ell\| \le \gamma\epsilon^\sigma/(4L_f)$. Using this relation, $x^\ell \in \mathcal{X}^{\mathrm{c}}_\epsilon$, and the definition of $\mathcal{X}^{\mathrm{c}}_\epsilon$ in (17), we deduce that $\mathrm{dist}(0, \partial\Psi(x)) > \epsilon$ for any $x \in [x^\ell, x^{\ell+1}]$. In addition, by $x^\ell, x^{\ell+1} \in \mathcal{X}$, the convexity of $\mathcal{X}$, and $\mathcal{X} \subset \Omega$, one can see that $[x^\ell, x^{\ell+1}] \subseteq \Omega$. It follows from these and (8) that $[x^\ell, x^{\ell+1}] \subseteq \mathcal{U}_\epsilon$. In view of this, (19), and the definition of L_ℓ in Algorithm 2, we have

$$F^*(x^{\ell+1}) \le F^*(x^\ell) + \langle \nabla F^*(x^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2}\|x^{\ell+1} - x^\ell\|^2 + \frac{\delta_\ell}{2}. \tag{75}$$

In addition, notice that $F(x^{\ell+1}, y^{\ell+1}) \le F^*(x^{\ell+1})$. Using this, (65), (74), and (75), we obtain that

$$\begin{aligned}
&F(x^{\ell+1}, y^{\ell+1}) + p(x^{\ell+1}) \le F^*(x^{\ell+1}) + p(x^{\ell+1}) \\
&\overset{(75)}{\le} F^*(x^\ell) + \langle \nabla F^*(x^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2}\|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) + \frac{\delta_\ell}{2} \\
&= F(x^\ell, y^\ell) + \langle \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{L_\ell}{2}\|x^{\ell+1} - x^\ell\|^2 + p(x^{\ell+1}) + F^*(x^\ell) - F(x^\ell, y^\ell) \\
&\quad + \langle \nabla F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{\delta_\ell}{2} \\
&\overset{(65)(74)}{\le} F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{2}\|x^{\ell+1} - x^\ell\|^2 + \eta_\ell + \langle \nabla F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \frac{\delta_\ell}{2} \\
&= F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4}\|x^{\ell+1} - x^\ell\|^2 - \frac{L_\ell}{4}\|x^{\ell+1} - x^\ell\|^2 + \langle \nabla F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell), x^{\ell+1} - x^\ell \rangle + \eta_\ell + \frac{\delta_\ell}{2} \\
&\le F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4}\|x^{\ell+1} - x^\ell\|^2 + \frac{\|\nabla F^*(x^\ell) - \nabla_x f(x^\ell, y^\ell)\|^2}{L_\ell} + \eta_\ell + \frac{\delta_\ell}{2} \\
&\overset{(66)}{\le} F(x^\ell, y^\ell) + p(x^\ell) - \frac{L_\ell}{4}\|x^{\ell+1} - x^\ell\|^2 + \Big(1 + \frac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell}\Big)\eta_\ell + \frac{\delta_\ell}{2},
\end{aligned}$$

where the fourth inequality follows from the Young's inequality $\langle u, v\rangle \le \alpha\|u\|^2/4 + \|v\|^2/\alpha$ for all $\alpha > 0$ and $u, v \in \mathbb{R}^n$. By this and the arbitrariness of ℓ, we see that (73) holds for all $0 \le \ell \le k$.

Summing up (73) over $\ell = 0, \ldots, k$ yields

$$\sum_{\ell=\lceil k/2\rceil}^{k} L_\ell\|x^{\ell+1} - x^\ell\|^2 \le 4\Big[F(x^0, y^0) + p(x^0) - F(x^{k+1}, y^{k+1}) - p(x^{k+1}) + \sum_{\ell=0}^{k}\Big(1 + \tfrac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k}\frac{\delta_\ell}{2}\Big]. \tag{76}$$

For notational convenience, let $\hat{k} = \ell(k)$. By this, (68), and (71), one has

$$\begin{aligned}
L_{\hat{k}}\|x^{\hat{k}+1} - x^{\hat{k}}\|^2 &\overset{(71)}{\le} \frac{1}{\lceil k/2\rceil}\sum_{\ell=\lceil k/2\rceil}^{k} L_\ell\|x^{\ell+1} - x^\ell\|^2 \\
&\overset{(76)}{\le} \frac{4}{\lceil k/2\rceil}\Big[F(x^0, y^0) + p(x^0) - F(x^{k+1}, y^{k+1}) - p(x^{k+1}) + \sum_{\ell=0}^{k}\Big(1 + \tfrac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k}\frac{\delta_\ell}{2}\Big] \\
&\le \frac{8}{k}\Big[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^{k}\Big(1 + \tfrac{L_{\nabla f}^2}{(1-\theta)^2 C^2 L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k}\frac{\delta_\ell}{2}\Big] \overset{(68)}{=} \frac{\Delta_k}{k},
\end{aligned} \tag{77}$$

where the last inequality uses the fact that $F(x^0, y^0) \le F^*(x^0)$, $\Psi(\cdot) = F^*(\cdot) + p(\cdot)$, $\Psi(x^{k+1}) \ge \Psi^*$, and $F^*(x^{k+1}) - F(x^{k+1}, y^{k+1}) \le \eta_{k+1}$ due to (65). Additionally, by $\hat{k} \ge \lceil k/2\rceil$, the monotonicity of $\{\delta_\ell\}$, $\nu \in (0, 1]$, and the definition of L_ℓ in Algorithm 2, one has $L_{\hat{k}} \ge L_{\lceil k/2\rceil}$. It then together with (77) implies that

$$\|x^{\hat{k}+1} - x^{\hat{k}}\| \le \sqrt{\frac{\Delta_k}{L_{\hat{k}}k}} \le \sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}}. \tag{78}$$

In addition, notice from Algorithm 2 that $r = \gamma\epsilon^\sigma/(4L_f)$. By this, $k > \underline{K}_\epsilon$, and (69), one has $\Delta_k/(L_{\lceil k/2\rceil}k) < r^2$. Using this and (78), we have

$$\|x^{\hat{k}+1} - x^{\hat{k}}\| < r. \tag{79}$$

This together with the first-order optimality condition of (21) for $x^{\hat{k}+1}$ implies that

$$0 \in \nabla_x f(x^{\hat{k}}, y^{\hat{k}}) + L_{\hat{k}}(x^{\hat{k}+1} - x^{\hat{k}}) + \partial p(x^{\hat{k}+1}). \tag{80}$$

By $\hat{k} \leq k$ and the assumption that $x^\ell \in \mathcal{X}_\epsilon^{\rm c}$ for all $0 \leq \ell \leq k$, one has $x^{\hat{k}} \in \mathcal{X}_\epsilon^{\rm c}$. Using this and (79), we conclude that $\mathrm{dist}(0, \partial\Psi(x^{\hat{k}})) > \epsilon$ and $\mathrm{dist}(0, \partial\Psi(x^{\hat{k}+1})) > \epsilon$. By these and $x^{\hat{k}}, x^{\hat{k}+1} \in \mathcal{X} \subset \Omega$, one can see that $x^{\hat{k}}, x^{\hat{k}+1} \in \mathcal{U}_\epsilon$. In view of this, (18), (79), and Theorem 1, we obtain that

$$\|\nabla F^*(x^{\hat{k}+1}) - \nabla F^*(x^{\hat{k}})\| \leq L_{\nabla f}\|x^{\hat{k}+1} - x^{\hat{k}}\| + M\|x^{\hat{k}+1} - x^{\hat{k}}\|^\nu. \tag{81}$$

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, (66), (78), (80), (81), and $\partial\Psi(\cdot) = \nabla F^*(\cdot) + \partial p(\cdot)$, we have

$$\begin{aligned}
\mathrm{dist}\left(0, \partial\Psi(x^{\hat{k}+1})\right) &\overset{(80)}{\leq} \|\nabla F^*(x^{\hat{k}+1}) - \nabla_x f(x^{\hat{k}}, y^{\hat{k}}) - L_{\hat{k}}(x^{\hat{k}+1} - x^{\hat{k}})\| \\
&\leq \|\nabla F^*(x^{\hat{k}+1}) - \nabla F^*(x^{\hat{k}})\| + \|\nabla F^*(x^{\hat{k}}) - \nabla_x f(x^{\hat{k}}, y^{\hat{k}})\| + L_{\hat{k}}\|x^{\hat{k}+1} - x^{\hat{k}}\| \\
&\overset{(66)(81)}{\leq} L_{\nabla f}\|x^{\hat{k}+1} - x^{\hat{k}}\| + M\|x^{\hat{k}+1} - x^{\hat{k}}\|^\nu + (1-\theta)^{-1}C^{-1}L_{\nabla f}\eta_{\hat{k}}^{\frac{1}{2}} + L_{\hat{k}}\|x^{\hat{k}+1} - x^{\hat{k}}\| \\
&\overset{(78)}{\leq} L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\hat{k}}k}} + \sqrt{\frac{L_{\hat{k}}\Delta_k}{k}} + M\Big(\frac{\Delta_k}{L_{\hat{k}}k}\Big)^{\frac{\nu}{2}} + (1-\theta)^{-1}C^{-1}L_{\nabla f}\eta_{\hat{k}}^{\frac{1}{2}} \\
&\leq L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}} + \sqrt{\frac{L_k\Delta_k}{k}} + M\Big(\frac{\Delta_k}{L_{\lceil k/2\rceil}k}\Big)^{\frac{\nu}{2}} + (1-\theta)^{-1}C^{-1}L_{\nabla f}\eta_{\lceil k/2\rceil}^{\frac{1}{2}}.
\end{aligned}$$

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 $t \geq \max\Big\{(2a\zeta^{-1})^{1/\omega},\ \Big(\lfloor 4b(\omega\zeta)^{-1}\log(2b/(\omega\zeta))\rfloor_+ + 1\Big)^{1/\omega}\Big\}$, 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

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

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 $k = \widehat{K}_\epsilon$. Suppose for contradiction that an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-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 $0 \le \ell < k$. Since none of $\{x^i\}_{i=0}^\ell$ is an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point of (1), it follows that $x^i \in \mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0 \le i \le \ell$, where $\mathcal{X}_\epsilon^{\mathrm{c}}$ is defined in (17). By this and Lemma 9, one has that $F^*(x^\ell) - F(x^\ell, y^\ell) \le \gamma\epsilon^\sigma/2$. In addition, notice from (21) that $x^{\ell+1}$ is well-defined and thus successfully generated by Algorithm 2, and moreover, $\|x^{\ell+1} - x^\ell\| \le r = \gamma\epsilon^\sigma/(4L_f)$. By these, Lemma 1, and the L_f-Lipschitz continuity of $F(\cdot, y)$ for each $y \in \mathcal{Y}$, we have

$$\begin{aligned} F^*(x^{\ell+1}) - F(x^{\ell+1}, y^\ell) &= F^*(x^{\ell+1}) - F^*(x^\ell) + F^*(x^\ell) - F(x^\ell, y^\ell) + F(x^\ell, y^\ell) - F(x^{\ell+1}, y^\ell) \\ &\le 2L_f\|x^{\ell+1} - x^\ell\| + \tfrac{1}{2}\gamma\epsilon^\sigma \le 2L_f r + \tfrac{1}{2}\gamma\epsilon^\sigma \le \gamma\epsilon^\sigma. \end{aligned} \tag{82}$$

In addition, since $x^\ell \in \mathcal{X}_\epsilon^{\mathrm{c}}$, $\|x^{\ell+1} - x^\ell\| \le \gamma\epsilon^\sigma/(4L_f)$, and $x^{\ell+1} \in \mathcal{X} \subset \Omega$, one can see from (70) that $\mathrm{dist}(0, \partial\Psi(x^{\ell+1})) > \epsilon$. This and (6) imply that

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

Hence, (12) holds for the function $h(\cdot) = -F(x^{\ell+1}, \cdot)$ with $\delta = \gamma\epsilon^\sigma$. It follows from this and (82) 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 (22), $\nu \in (0, 1]$, $\delta_\ell \le 1$, and the definition of L_ℓ in Algorithm 2, we see that $L_\ell \ge \underline{L}$ for all $0 \le \ell \le k$. In addition, observe from (24), (25), (31), $\epsilon \in (0, 1/e]$, and $\nu \in (0, 1]$ that $\widehat{K}_\epsilon \ge 2$, which together with $k = \widehat{K}_\epsilon$ implies $k \ge 2$. In view of these, (68), (22), (23), and the definitions of $\{\delta_\ell\}$ and $\{\eta_\ell\}$, one has

$$\begin{aligned} \Delta_k &\overset{(68)(22)}{=} 8\Big[\Psi(x^0) - \Psi^* + \eta_{k+1} + \sum_{\ell=0}^{k}\Big(1 + \frac{A}{L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k}\frac{\delta_\ell}{2}\Big] \le 8\Big[\Psi(x^0) - \Psi^* + \sum_{\ell=0}^{k+1}\Big(1 + \frac{A}{L_\ell}\Big)\eta_\ell + \sum_{\ell=0}^{k+1}\frac{\delta_\ell}{2}\Big] \\ &\le 8\Big[\Psi(x^0) - \Psi^* + \Big(\frac{3}{2} + \frac{A}{\underline{L}}\Big)\sum_{\ell=0}^{k+1}\frac{1}{\ell+1}\Big] \le 8\Big[\Psi(x^0) - \Psi^* + \Big(\frac{3}{2} + \frac{A}{\underline{L}}\Big)\Big(1 + \int_0^{k+1}\frac{1}{1+t}\,dt\Big)\Big] \\ &= 8\big[\Psi(x^0) - \Psi^* + (3/2 + A\underline{L}^{-1})\big(1 + \log(k+2)\big)\big] \le 8\big[\Psi(x^0) - \Psi^* + (3/2 + A\underline{L}^{-1})(2 + \log k)\big] \\ &= 8\big[\Psi(x^0) - \Psi^* + 3 + 2A\underline{L}^{-1} + (3/2 + A\underline{L}^{-1})\log k\big] \overset{(23)}{=} a + b\log k, \end{aligned} \tag{83}$$

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

$$\frac{\Delta_{k'}}{L_{\lceil k'/2\rceil}k'} \le \frac{a + b\log k'}{(k'/2)^{\frac{1-\nu}{1+\nu}}k'M^{\frac{2}{1+\nu}}} \le \frac{2(a + b\log k')}{k'^{\frac{2}{1+\nu}}M^{\frac{2}{1+\nu}}}. \tag{84}$$

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

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

where the last inequality is due to $\epsilon \in (0, 1/e]$, (29), (30), and (31). It then follows from the arbitrariness of k' and the definition of $\underline{K}_\epsilon$ in (69) that $\underline{K}_\epsilon < k$.

In addition, observe from $k=\widehat{K}_\epsilon$, $\epsilon\in(0,1/e]$, (27), (28), and (31) that

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

It then follows from (84) and Lemma 11 that $\Delta_k/(L_{\lceil k/2\rceil}k)\leq\epsilon^2/(36L_{\nabla f}^2)$, which implies that

$$L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}}\leq\frac{\epsilon}{6}. \tag{85}$$

Similarly, notice from $k=\widehat{K}_\epsilon$, $\epsilon\in(0,1/e]$, (24), (26), and (31) that

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

It then follows from (83) 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

$$\sqrt{\frac{L_k\Delta_k}{k}}\leq\frac{\epsilon}{3}. \tag{86}$$

Also, by $k=\widehat{K}_\epsilon$, $\epsilon\in(0,1/e]$, (25), (26), and (31), we can see that

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

It then follows from (84) and Lemma 11 that $\Delta_k/(L_{\lceil k/2\rceil}k)\leq\epsilon^{2/\nu}/(3M)^{2/\nu}$, which implies that

$$M\Big(\frac{\Delta_k}{L_{\lceil k/2\rceil}k}\Big)^{\frac{\nu}{2}}\leq\frac{\epsilon}{3}. \tag{87}$$

Lastly, using $k=\widehat{K}_\epsilon$, (26) and (31) yields $k\geq\lceil 72A\epsilon^{-2}\rceil$. By this and the definition of η_k, one has

$$A^{\frac12}\eta_{\lceil k/2\rceil}^{\frac12}\leq\left(\frac{2A}{k+2}\right)^{\frac12}\leq\frac{\epsilon}{6}. \tag{88}$$

Recall from the above hypothesis that none of $\{x^\ell\}_{\ell=0}^k$ is an $(\epsilon,\gamma\epsilon^\sigma/(4L_f))$-stationary point of (1). Hence, $x^\ell\in\mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0\leq\ell\leq k$. Moreover, it follows from this and the definition of $\overline{K}_\epsilon$ in (70) that $k\leq\overline{K}_\epsilon$. By $\underline{K}_\epsilon<k\leq\overline{K}_\epsilon$ and $x^\ell\in\mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0\leq\ell\leq k$, it follows from Lemma 10 that (72) holds for such k, which together with the definition of A in (23) leads to

$$\mathrm{dist}\left(0,\partial\Psi(x^{\ell(k)+1})\right)\leq L_{\nabla f}\sqrt{\frac{\Delta_k}{L_{\lceil k/2\rceil}k}}+\sqrt{\frac{L_k\Delta_k}{k}}+M\Big(\frac{\Delta_k}{L_{\lceil k/2\rceil}k}\Big)^{\frac{\nu}{2}}+A^{\frac12}\eta_{\lceil k/2\rceil}^{\frac12}.$$

Combining this with (85), (86), (87), and (88), we obtain that

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

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 $\mathcal{X}_\epsilon^{\mathrm{c}}$ in (17) that $x^{\hat{k}}\notin\mathcal{X}_\epsilon^{\mathrm{c}}$. Since $\hat{k}\leq k$, this contradicts the assumption that $x^\ell\in\mathcal{X}_\epsilon^{\mathrm{c}}$ 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 an $(\epsilon,\gamma\epsilon^\sigma/(4L_f))$-stationary point of problem (1) for some $0\leq k\leq\widehat{K}_\epsilon$. Moreover, it follows from Lemma 9 that y^k satisfies (32). □

We next present the proof of Theorem 5.

Proof of Theorem 5. Let $\mathcal{X}_\epsilon^{\mathrm{c}}$ be defined in (70). The conclusion clearly holds if $x^0 \notin \mathcal{X}_\epsilon^{\mathrm{c}}$. Hence, we assume for the remainder of the proof that $x^0 \in \mathcal{X}_\epsilon^{\mathrm{c}}$. Given this and $\epsilon \in (0, 1/e]$, it follows from Theorem 4 that there exists $0 \leq \overline{k} < \widehat{K}_\epsilon$ such that $x^\ell \in \mathcal{X}_\epsilon^{\mathrm{c}}$ for all $0 \leq \ell \leq \overline{k}$ and $x^{\overline{k}+1} \notin \mathcal{X}_\epsilon^{\mathrm{c}}$. That is, the iterates $\{x^\ell\}_{\ell=0}^{\overline{k}}$ are not, but $x^{\overline{k}+1}$ is, an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-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 $\overline{k} < \widehat{K}_\epsilon$, it follows that the total number of evaluations of p to generate the $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point $x^{\overline{k}+1}$ is $\overline{k} + 1 \leq \widehat{K}_\epsilon$.

We next show that the total number of evaluations of the proximal operator of q performed in Algorithm 2 to generate the $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point $x^{\overline{k}+1}$ 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 $0 \leq \ell' \leq \overline{k}$, through its calls to Algorithm 1. As observed from Algorithm 2 and the proof of Lemma 10, Algorithm 1 is invoked to solve problem (11) with $h(\cdot) = -F(x^{\ell'+1}, \cdot)$, where h satisfies condition (12) with $\delta = \gamma\epsilon^\sigma$. By this, the definitions of τ and $\{\eta_\ell\}$ in Algorithm 2, $\overline{k} < \widehat{K}_\epsilon$, and (15), it follows from Theorem 3 that the number of outer iterations performed in Algorithm 1 at each iteration $0 \leq \ell' \leq \overline{k}$ is at most $\overline{K}_{f,\theta}$, where $\overline{K}_{f,\theta}$ is defined in (33). Using this and Theorem 2, we can see that at each iteration $0 \leq \ell' \leq \overline{k}$, the number of evaluations of q is at most

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

By this bound, the fact that the total number of iterations performed by Algorithm 2 to generate the $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point $x^{\overline{k}+1}$ is at most $\widehat{K}_\epsilon$, and (34), we conclude that the total number of evaluations of the proximal operator of q performed by Algorithm 2 to generate an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point is at most $\widehat{N}_\epsilon$.

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 an $(\epsilon, \gamma\epsilon^\sigma/(4L_f))$-stationary point is at most $\widehat{K}_\epsilon + \widehat{N}_\epsilon$. □

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

[4] J. Blanchet, J. Li, S. Lin, and X. Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

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

[6] 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.

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

[8] F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

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

[10] 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.

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

[12] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165:874–900, 2015.

[13] 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.

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

[15] 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.

[16] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811, 2016.

[17] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] E. M. Stein and R. Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces.* Princeton University Press, 2009.

[28] 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.

[29] 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.

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

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