Near-Optimal Methods for Convex-Concave Minimax Optimization

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

August 28, 2024

Abstract

This paper resolves a longstanding open question pertaining to the design of near-optimal first-order algorithms for smooth and strongly-convex-strongly-concave minimax problems. Current state-of-the-art first-order algorithms find an approximate Nash equilibrium using $\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$ [Tseng(1995)] or $\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \sqrt{\kappa_{\mathbf{x}}}\kappa_{\mathbf{y}}\})$ [Alkousa et al.(2019)] gradient evaluations, where $\kappa_{\mathbf{x}}$ and $\kappa_{\mathbf{y}}$ are the condition numbers for the strong-convexity and strong-concavity assumptions. A gap still remains between these results and the best existing lower bound $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ [Ibrahim et al.(2019), Zhang et al.(2019)]. This paper presents the first algorithm with $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ gradient complexity, matching the lower bound up to logarithmic factors. Our algorithm is designed based on an accelerated proximal point method and an accelerated solver for minimax proximal steps. It can be easily extended to the settings of strongly-convex-concave, convex-concave, nonconvex-strongly-concave, and nonconvex-concave functions. This paper also presents algorithms that match or outperform all existing methods in these settings in terms of gradient complexity, up to logarithmic factors.

1 Introduction

Let $\mathbb{R}^m$ and $\mathbb{R}^n$ be finite-dimensional Euclidean spaces and let the function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be smooth. Let $\mathcal{X}$ and $\mathcal{Y}$ are two nonempty closed convex sets in $\mathbb{R}^m$ and $\mathbb{R}^n$. Our problem of interest is the following minimax optimization problem:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \ f(\mathbf{x}, \mathbf{y}) \tag{1}$$

The theoretical study of solutions of problem (1) has been an focus of several decades of research in mathematics, statistics, economics and computer science [Basar and Olsder(1999), Nisan et al.(2007), Von Neumann and Morgenstern(2007), Facchinei and Pang(2007), Berger(2013)]. Recently, this line of research has become increasingly relevant to algorithmic machine learning, with applications including robustness in adversarial learning [Goodfellow et al.(2014), Sinha et al.(2018)], prediction and regression problems [Cesa-Bianchi and Lugosi(2006), Xu et al.(2009)] and distributed computing [Shamma(2008), Mateos et al.(2010)]. Moreover, real-world machine-learning systems are increasingly embedded in multi-agent systems or matching markets and subject to game-theoretic constraints [Jordan(2018)].

Most existing work on minimax optimization focuses on the convex-concave setting, where the function $f(\cdot, \mathbf{y})$ is convex for each $\mathbf{y} \in \mathbb{R}^n$ and the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$. The best known convergence rate in a general convex-concave setting is $O(1/\epsilon)$ in terms of duality gap, which can be achieved by Nemirovski's mirror-prox algorithm [Nemirovski(2004)] (a special case of which is the extragradient algorithm [Korpelevich(1976)]), Nesterov's dual extrapolation algorithm [Nesterov(2007)] or Tseng's accelerated proximal gradient algorithm [Tseng(2008)]. This rate

is known to be optimal for the class of smooth convex-concave problems [Ouyang and Xu(2019)]. Furthermore, optimal algorithms are known for special instances of convex-concave setting; e.g., for the affinely constrained smooth convex problem [Ouyang et al.(2015)] and problems with a composite bilinear objective function, $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$ [Chen et al.(2014)].

Very recently, the lower complexity bound of first-order algorithms have been established for solving general strongly-convex-strongly-concave and strongly-convex-concave minimax optimization problems [Ouyang and Xu(2019), Ibrahim et al.(2019), Zhang et al.(2019)]. For the strongly-convex-strongly-concave setting, in which $\kappa_\mathbf{x}, \kappa_\mathbf{y} > 0$ are the condition numbers for $f(\cdot, \mathbf{y})$ and $f(\mathbf{x}, \cdot)$, respectively, the complexity bound is $\tilde{\Omega}(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}})$ while the best known upper bounds are $\tilde{O}(\kappa_\mathbf{x} + \kappa_\mathbf{y})$ [Tseng(1995), Gidel et al.(2019), Mokhtari et al.(2019b)] and $\tilde{O}(\min\{\kappa_\mathbf{x}\sqrt{\kappa_\mathbf{y}}, \kappa_\mathbf{y}\sqrt{\kappa_\mathbf{x}}\})$ [Alkousa et al.(2019)]. For the strongly-convex-concave setting in which $\kappa_\mathbf{x} > 0$ and $\kappa_\mathbf{y} = 0$, the lower complexity bound is $\tilde{\Omega}(\sqrt{\kappa_\mathbf{x}/\epsilon})$ while the best known upper bound is $O(\kappa_\mathbf{x}/\epsilon)$ [Thekumparampil et al.(2019)]. The existing algorithms that obtain a rate of $O(\sqrt{\kappa_\mathbf{x}/\epsilon})$ in this context are only for special case of strongly-convex-linear, where $\mathbf{x}$ and $\mathbf{y}$ are connected only through a bilinear term $\mathbf{x}^\top A\mathbf{y}$ or $f(\mathbf{x}, \cdot)$ is linear for each $\mathbf{x} \in \mathbb{R}^m$; see, e.g., [Nesterov(2005), Chambolle and Pock(2016), Juditsky and Nemirovski(2011), Hamedani and Aybat(2018)]. Thus, a gap remains between the lower complexity bound and the upper complexity bound for existing algorithms in both the strongly-convex-strongly-concave setting and the strongly-convex-concave setting. Accordingly, we have the following open problem:

Can we design first-order algorithms that achieve the lower bounds in these settings?

This paper presents an affirmative answer by resolving the above open problem up to logarithmic factors. More specifically, our contribution is as follows.

We propose the first near-optimal algorithms for solving the strongly-convex-strongly-concave and strongly-convex-concave minimax optimization problems. In the former setting, our algorithm achieves a gradient complexity of $\tilde{O}(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}})$ which matches the lower complexity bound [Ibrahim et al.(2019), Zhang et al.(2019)] up to logarithmic factors. In the latter setting, our algorithm attains a gradient complexity of $\tilde{O}(\sqrt{\kappa_\mathbf{x}/\epsilon})$ which again matches the lower complexity bound [Ouyang and Xu(2019)] up to logarithmic factors. In addition, our algorithm extends to the general convex-concave setting, achieving a gradient complexity of $\tilde{O}(\epsilon^{-1})$, which matches the lower bound of [Ouyang and Xu(2019)] as well as the best existing upper bounds [Nemirovski(2004), Nesterov(2007), Tseng(2008)] up to logarithmic factors.

Our second contribution is a class of accelerated algorithms for the smooth nonconvex-strongly-concave and nonconvex-concave minimax optimization problems. In the former setting, our algorithm achieves a gradient complexity bound of $\tilde{O}(\sqrt{\kappa_\mathbf{y}}\epsilon^{-2})$ which improves the best known bound $\tilde{O}(\kappa_\mathbf{y}^2\epsilon^{-2})$ [Jin et al.(2019), Rafique et al.(2018), Lin et al.(2019), Lu et al.(2019)]. In the latter setting, our algorithms specialize to a range of different notions of optimality. In particular, expressing our results in terms of stationarity of f, our algorithm achieves a gradient complexity bound of $\tilde{O}(\epsilon^{-2.5})$, which improves the best known bound $\tilde{O}(\epsilon^{-3.5})$ [Nouiehed et al.(2019)]. In terms of stationarity of the function $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$, our algorithm achieves a gradient complexity bound of $\tilde{O}(\epsilon^{-3})$ which matches the current state-of-the-art results [Thekumparampil et al.(2019), Kong and Monteiro(2019)].

We provide a head-to-head comparison between our results and existing results in the literature in Table 1 for convex-concave settings, and Table 2 for nonconvex-concave settings.

Table 1. Comparison of gradient complexities to find an ϵ-saddle point (Definition 4) in the convex-concave setting. This table highlights only the dependency on error tolerance ϵ and the strong-convexity and strong-concavity condition numbers, $\kappa_{\mathbf{x}}, \kappa_{\mathbf{y}}$.

Settings	References	Gradient Complexity
Strongly-Convex-Strongly-Concave	[Tseng(1995)]	$\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$
	[Nesterov and Scrimali(2006)]	
	[Gidel et al.(2019)]	
	[Mokhtari et al.(2019b)]	
	[Alkousa et al.(2019)]	$\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$
	This paper (Theorem 4)	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound [Ibrahim et al.(2019)]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound [Zhang et al.(2019)]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
Strongly-Convex-Linear (special case of strongly-convex-concave)	[Juditsky and Nemirovski(2011)]	$O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	[Hamedani and Aybat(2018)]	
	[Zhao(2019)]	
Strongly-Convex-Concave	[Thekumparampil et al.(2019)]	$\tilde{O}(\kappa_{\mathbf{x}}/\sqrt{\epsilon})$
	This paper (Corollary 1)	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	Lower bound [Ouyang and Xu(2019)]	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
Convex-Concave	[Nemirovski(2004)]	$O(\epsilon^{-1})$
	[Nesterov(2007)]	
	[Tseng(2008)]	
	This paper (Corollary 2)	$\tilde{O}(\epsilon^{-1})$
	Lower bound [Ouyang and Xu(2019)]	$\Omega(\epsilon^{-1})$

1.1 Related work

To the best of our knowledge, the earliest algorithmic schemes for solving the bilinear minimax problem, $\min_{\mathbf{x}\in\Delta^m} \max_{\mathbf{y}\in\Delta^n} \mathbf{x}^\top A\mathbf{y}$, date back to Brown's fictitious play [Brown(1951)] and Dantzig's simplex method [Dantzig(1998)]. This problem can also be solved by Korpelevich's extragradient (EG) algorithm [Korpelevich(1976)], which can be shown to be linearly convergent when A is square and full rank [Tseng(1995)]. There are also several recent papers studying the convergence of EG and its variants; see [Chambolle and Pock(2011), Malitsky(2015), Yadav et al.(2018)] for reflected gradient descent ascent, [Daskalakis et al.(2018), Mokhtari et al.(2019b), Mokhtari et al.(2019a)] for optimistic gradient descent ascent (OGDA) and [Rakhlin and Sridharan(2013a), Rakhlin and Sridharan(2013b), Mertikopoulos et al.(2019), Chavdarova et al.(2019), Hsieh et al.(2019), Mishchenko et al.(2019)] for other variants. In the bilinear setting, [Daskalakis et al.(2018)] established the convergence of the optimistic gradient descent ascent (OGDA) method to a neighborhood of the solution; [Liang and Stokes(2019)] proved the linear convergence of the OGDA algorithm using a dynamical

system approach. Very recently, [Mokhtari et al.(2019b)] have proposed a unified framework for achieving the sharpest convergence rates of both EG and OGDA algorithms.

For the convex-concave minimax problem, [Nemirovski(2004)] proved that his mirror-prox algorithm returns an ϵ-saddle point within the gradient complexity of $O(\epsilon^{-1})$ when $\mathcal{X}$ and $\mathcal{Y}$ are bounded. This algorithm was subsequently generalized by [Auslender and Teboulle(2005)] to a class of distance-generating functions, and the complexity result was extended to unbounded sets and composite objectives [Monteiro and Svaiter(2010), Monteiro and Svaiter(2011)] using the hybrid proximal extragradient algorithm with different error criteria. [Nesterov(2007)] developed a dual extrapolation algorithm which possesses the same complexity bound as in [Nemirovski(2004)]. Later on, [Tseng(2008)] presented a unified treatment of these algorithms and a refined convergence analysis with same complexity result. [Nedić and Ozdaglar(2009)] analyzed the (sub)gradient descent ascent algorithm for convex-concave saddle point problems when the (sub)gradients are bounded over the constraint sets. [Abernethy et al.(2019)] presented a Hamiltonian gradient descent algorithm with last-iterate convergence under a "sufficiently bilinear" condition.

Several papers have studied special cases in the convex-concave setting. For the special case when the objective function is a composite bilinear form, $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$, [Chambolle and Pock(2011)] introduced a primal-dual algorithm that converges to a saddle point with the rate of $O(1/\epsilon)$ when the convex functions g and h are smooth. [Nesterov(2005)] proposed a smoothing technique and proved that the resulting algorithm achieves an improved rate with better dependence on Lipschitz constant of ∇g when h is the convex and smooth function and $\mathcal{X}, \mathcal{Y}$ are both bounded. [He and Monteiro(2016)] and [Kolossoski and Monteiro(2017)] proved that such result also hold when $\mathcal{X}, \mathcal{Y}$ are unbounded or the space is non-Euclidean. [Chen et al.(2014), Chen et al.(2017)] generalized Nesterov's technique to develop optimal algorithms for solving a class of stochastic saddle point problems and stochastic monotone variational inequalities. For a class of certain purely bilinear games where g and h are zero functions, [Azizian et al.(2020)] demonstrated that linear convergence is possible for several algorithms and their new algorithm achieved the tight bound. The second case is the so-called affinely constrained smooth convex problem, i.e., $\min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}), \text{s.t. } A\mathbf{x} = \mathbf{u}$. [Esser et al.(2010)] proposed a $O(\epsilon^{-1})$ primal-dual algorithm while [Lan and Monteiro(2016)] provided a first-order augmented Lagrangian method with the same $O(\epsilon^{-1})$ rate. By exploiting the structure, [Ouyang et al.(2015)] proposed a near-optimal algorithm in this setting.

For the strongly convex-concave minimax problem, [Tseng(1995)] and [Nesterov and Scrimali(2006)] proved that their algorithms find an ϵ-saddle point with a gradient complexity of $\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$ using a variational inequality. Using a different approach, [Gidel et al.(2019)] and [Mokhtari et al.(2019b)] derived the same complexity results for the OGDA algorithm. Very recently, [Alkousa et al.(2019)] proposed an accelerated gradient sliding algorithm with a gradient complexity of $\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$ while [Ibrahim et al.(2019)] and [Zhang et al.(2019)] established a lower complexity bound of $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ among all the first-order algorithms in this setting.

For strongly-convex-concave minimax problems, the best known general lower bound for first-order algorithm is $O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$, as shown by [Ouyang and Xu(2019)]. Several papers have studied strongly-convex-concave minimax problem with additional structures. This includex optimizing a strongly convex function with linear constraints [Goldstein et al.(2014), Xu and Zhang(2018), Xu(2019)], the case when $\mathbf{x}$ and $\mathbf{y}$ are connected only through a bilinear term $\mathbf{x}^\top A\mathbf{y}$ [Nesterov(2005), Chambolle and Pock(2016), Xie and Shi(2019)] and the case when $f(\mathbf{x}, \cdot)$ is linear for each $\mathbf{x} \in \mathbb{R}^m$ [Juditsky and Nemirovski(2011), Hamedani and Aybat(2018), Zhao(2019)]. The algorithms developed in these works were all guaranteed to return an ϵ-saddle point with a gradient complexity of

$\tilde{O}(1/\sqrt{\epsilon})$ and some of them even achieve a near-optimal gradient complexity of $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$ [Nesterov(2005), Chambolle and Pock(2016)]. However, the best known upper complexity bound for general strongly-convex-concave minimax problems is $O(\kappa_{\mathbf{x}}/\sqrt{\epsilon})$ which was shown using the *dual implicit accelerated gradient* algorithm [Thekumparampil et al.(2019)].

For nonconvex-concave minimax problems, a line of recent work [Jin et al.(2019), Rafique et al.(2018), Lin et al.(2019)] has studied various algorithms and proved that they can find an approximate stationary point of $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$. In a deterministic setting, all of these algorithms guarantee a rate of $\tilde{O}(\kappa_{\mathbf{y}}^2 \epsilon^{-2})$ and $\tilde{O}(\epsilon^{-6})$ when $f(\mathbf{x}, \cdot)$ is strongly concave and concave respectively. [Thekumparampil et al.(2019)] consider the same setting as ours and proposed a proximal dual implicit accelerated gradient algorithm and proved that it finds an approximate stationary point of $\Phi(\cdot)$ with the total gradient complexity of $\tilde{O}(\epsilon^{-3})$. [Kong and Monteiro(2019)] consider a general nonconvex minimax optimization model: $\min_{\mathbf{x}} h(\mathbf{x}) + \rho(\mathbf{x})$, where h is a "simple" proper, lower semi-continuous and convex function and $\rho(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ with f satisfying that $-f(\mathbf{x}, \cdot)$ is proper, convex, and lower semi-continuous. They propose to smooth ρ to $\rho_\xi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) - (1/2\xi)\|\mathbf{y} - \mathbf{y}_0\|^2$ and apply an accelerated inexact proximal point method to solve the smoothed problem $\min_{\mathbf{x}} h(\mathbf{x}) + \rho_\xi(\mathbf{x})$. The resulting AIPP-S algorithm attains the iteration complexity of $O(\epsilon^{-3})$ using a slightly different but equivalent notion of stationarity but requires the *exact* gradient of ρ_ξ at each iteration. This amounts to assuming that $\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) - (1/2\xi)\|\mathbf{y} - \mathbf{y}_0\|^2$ can be solved exactly, which is restrictive due to the potentially complicated structure of $f(\mathbf{x}, \cdot)$ or $\mathcal{Y}$. If f is further assumed to be smooth, [Zhao(2020)] developed a variant of AIPP-S algorithm which only requires an inexact gradient of ρ_ξ at each iteration and attains the total gradient complexity of $\tilde{O}(\epsilon^{-3})$. On the other hand, the stationarity of $f(\cdot, \cdot)$ is proposed for quantifying the efficiency in nonconvex-concave minimax optimization [Lu et al.(2019), Nouiehed et al.(2019), Kong and Monteiro(2019), Ostrovskii et al.(2020)]. Using this notion of stationarity, [Kong and Monteiro(2019)] attains the rate of $O(\epsilon^{-2.5})$ but requires the *exact* gradient of ρ_ξ at each iteration. Without this assumption, the current state-of-the-art rate is $\tilde{O}(\epsilon^{-2.5})$ achieved by our Algorithm 5 and the algorithm proposed by a concurrent work [Ostrovskii et al.(2020)]. Both algorithms are based on constructing an auxiliary function $f_{\epsilon, \mathbf{y}}$ and applying an accelerated solver for minimax proximal steps. Finally, several other algorithms have been developed either for specific nonconvex-concave minimax problems or in stochastic setting; see [Namkoong and Duchi(2016), Sinha et al.(2018), Sanjabi et al.(2018), Grnarova et al.(2018)] for the details.

2 Preliminaries

In this section, we clarify the notation used in this paper, review some background and provide formal definitions for the class of functions and optimality measure considered in this paper.

Notation. We use bold lower-case letters to denote vectors, as in $\mathbf{x}, \mathbf{y}, \mathbf{z}$ and calligraphic upper case letters to denote sets, as in $\mathcal{X}$ and $\mathcal{Y}$. For a differentiable function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$, we le4t $\nabla f(\mathbf{z})$ denote the gradient of f at $\mathbf{z}$. For a function $f(\cdot, \cdot) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of two variables, $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ (or $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$) to denote the partial gradient of f with respect to the first variable (or the second variable) at point $(\mathbf{x}, \mathbf{y})$. We also use $\nabla f(\mathbf{x}, \mathbf{y})$ to denote the full gradient at $(\mathbf{x}, \mathbf{y})$ where $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))$. For a vector $\mathbf{x}$, we denote $\|\mathbf{x}\|$ as its ℓ_2-norm. For constraint sets $\mathcal{X}$ and $\mathcal{Y}$, we let $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ denote their diameters, where $D_{\mathbf{x}} = \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$ and $D_{\mathbf{y}} = \max_{\mathbf{y}, \mathbf{y}' \in \mathcal{Y}} \|\mathbf{y} - \mathbf{y}'\|$. We use the notation $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ to denote projections onto the sets $\mathcal{X}$ and $\mathcal{Y}$. Finally, we use the notation $O(\cdot), \Omega(\cdot)$ to hide only absolute constants which do not depend

on any problem parameter, and notation $\tilde{O}(), \tilde{\Omega}()$ to hide only absolute constants and log factors.

2.1 Minimax optimization

We are interested in the ℓ-smooth minimax optimization problems in the form (1). The regularity conditions that we consider for the function f are as follows.

Definition 1. *A function f is L-Lipschitz if for $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$, that $|f(\mathbf{z}) - f(\mathbf{z}')| \leq L\|\mathbf{z} - \mathbf{z}'\|$.*

Definition 2. *A function f is ℓ-smooth if for $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$, that $\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{z}')\| \leq \ell \|\mathbf{z} - \mathbf{z}'\|$.*

Definition 3. *A differentiable function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ-strongly-convex if for any $\mathbf{x}', \mathbf{x} \in \mathbb{R}^d$:*

$$\phi(\mathbf{x}') \geq \phi(\mathbf{x}) + (\mathbf{x}' - \mathbf{x})^\top \nabla \phi(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}' - \mathbf{x}\|^2$$

Furthermore, ϕ is μ-strongly-concave if $-\phi$ is μ-strongly-convex. If we set $\mu = 0$, then we recover the definitions of convexity and concavity for a continuous differentiable function.

Convex-concave setting: we assume that $f(\cdot, \mathbf{y})$ is convex for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$. *Here $\mathcal{X}$ and $\mathcal{Y}$ are both convex and bounded.* Under these conditions, the Sion's minimax theorem [Sion(1958)] guarantees that

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

Furthermore, there exists at least one **saddle point (or Nash equilibrium)** $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{X} \times \mathcal{Y}$ such that the following equality holds true:

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}^\star) \;=\; f(\mathbf{x}^\star, \mathbf{y}^\star) \;=\; \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^\star, \mathbf{y}) \tag{3}$$

Therefore, for any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$, the duality gap $\max_{\mathbf{y} \in \mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}})$ forms the basis for a standard optimality criterion. Formally, we define

Definition 4. *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-saddle point of a convex-concave function $f(\cdot, \cdot)$ if $\max_{\mathbf{y} \in \mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$. If $\epsilon = 0$, then $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a saddle point.*

In the case when $f(\cdot, \mathbf{y})$ is strongly convex for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is strongly concave for each $\mathbf{x} \in \mathcal{X}$, we refer $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ to the strongly-convex or strongly-concave module. If f is further ℓ-smooth, we denote $\kappa_\mathbf{x} = \ell/\mu_\mathbf{x}$ and $\kappa_\mathbf{y} = \ell/\mu_\mathbf{y}$ as the condition numbers of $f(\cdot, \mathbf{y})$ and $f(\mathbf{x}, \cdot)$.

Nonconvex-concave setting: we only assume that $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$. The function $f(\cdot, \mathbf{y})$ can be possibly nonconvex for some $\mathbf{y} \in \mathcal{Y}$. *Here $\mathcal{X}$ is convex but possibly unbounded while $\mathcal{Y}$ is convex and bounded.* In general, finding a global Nash equilibrium of f is intractable since in the special case where $\mathcal{Y}$ has only a single element, this problem reduces to a nonconvex optimization problem in which finding a global minimum is already NP-hard [Murty and Kabadi(1987)]. Similar to the literature in nonconvex constrained optimization, we opt to find local surrogates— stationary points—whose gradient mappings are zero. Formally, we define our optimality criterion as follows.

Algorithm 1 $\text{AGD}(g, \mathcal{X}, \mathbf{x}_0, \ell, \mu, \epsilon)$

1: **Input:** initial point $\mathbf{x}_0 \in \mathcal{X}$, smoothness ℓ, strongly-convex module μ and tolerance $\epsilon > 0$
2: **Initialize:** set $t \leftarrow 0$, $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\eta \leftarrow 1/\ell$, $\kappa \leftarrow \ell/\mu$ and $\theta \leftarrow \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$
3: **repeat**
4:     $t \leftarrow t + 1$
5:     $\mathbf{x}_t \leftarrow \mathcal{P}_{\mathcal{X}}[\tilde{\mathbf{x}}_{t-1} - \eta \nabla g(\tilde{\mathbf{x}}_{t-1})]$
6:     $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$
7: **until** $\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \eta\nabla g(\mathbf{x}_t))\|^2 \leq \frac{\epsilon}{2\kappa^2(\ell-\mu)}$ is satisfied.
8: **Output:** $\mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \eta\nabla g(\mathbf{x}_t))$

Definition 5. *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-**stationary point** of an ℓ-smooth function $f(\cdot, \cdot)$ if*

$$\ell\|\mathcal{P}_{\mathcal{X}}[\hat{\mathbf{x}} - \frac{1}{\ell}\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)] - \hat{\mathbf{x}}\| \leq \epsilon, \quad \ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \leq \epsilon$$

where

$$\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$$

If $\epsilon = 0$, then $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a stationary point.

In the absence of the constraint set $\mathcal{X}$, Definition 5 reduces to the standard condition $\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \leq \epsilon$ and $\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \leq \epsilon$ which is consistent with [Lin et al.(2019), Definition 4.10]. Intuitively, the quantity $\|\mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}})] - \hat{\mathbf{y}}\|$ represents the distance between a point $\hat{\mathbf{y}}$ and a point obtained by performing one-step projected partial gradient ascent at a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ starting from a point $\hat{\mathbf{y}}$. It also refers to the norm of gradient mapping at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$; see [Nesterov(2013)] for the details.

We note that this notion of stationarity of f (Definition 5) is closely related to an optimality notion in terms of stationary points of the function $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ for nonconvex-concave functions. We refer readers to Appendix A.3 for more discussion.

2.2 Nesterov's accelerated gradient descent

Nesterov's Accelerated Gradient Descent (AGD) dates back to the seminal paper [Nesterov(1983)] where it is shown to be optimal among all the first-order algorithms for smooth and convex functions [Nesterov(2018)]. We present a version of AGD in Algorithm 1 which is frequently used to minimize an ℓ-smooth and μ-strongly convex function g over a convex set $\mathcal{X}$. The key steps of the AGD algorithm are Line 5-6, where Lines 5 performs a projected gradient descent step, while Line 6 performs a momentum step, which "overshoots" the iterate in the direction of momentum $(\mathbf{x}_t - \mathbf{x}_{t-1})$. Line 7 is the stopping condition to ensure that the output achieves the desired optimality.

The following theorem provides an upper bound on the gradient complexity of AGD; i.e., the total number of gradient evaluations to find an ϵ-optimal point in terms of function value.

Theorem 1. *Assume that g is ℓ-smooth and μ-strongly convex, the output $\hat{\mathbf{x}} = \text{AGD}(g, \mathbf{x}_0, \ell, \mu, \epsilon)$ satisfies $g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) + \epsilon$ and the total number of gradient evaluations is bounded by*

$$O\left(\sqrt{\kappa}\log\left(\frac{\kappa^3\ell\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right)$$

where $\kappa = \ell/\mu$ is the condition number, and $\mathbf{x}^\star \in \mathcal{X}$ is the unique global minimum of g.

Algorithm 2 INEXACT-APPA$(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$

1: **Input:** initial point $\mathbf{x}_0 \in \mathcal{X}$, proximal parameter ℓ, strongly-convex module μ, tolerance $\epsilon > 0$ and the maximum iteration number $T > 0$
2: **Initialize:** set $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\kappa \leftarrow \frac{\ell}{\mu}$, $\delta \leftarrow \frac{\epsilon}{(10\kappa)^2}$ and $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$
3: **for** $t = 1, \cdots, T$ **do**
4:      find $\mathbf{x}_t$ so that $g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \leq \min_{\mathbf{x} \in \mathcal{X}}\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\} + \delta$
5:      $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$
6: **end for**
7: **Output:** $\mathbf{x}_T$

Compared with the classical result for Gradient Descent (GD), which requires $\tilde{O}(\kappa)$ gradient evaluations in the same setting, AGD improves over GD by a factor of $\sqrt{\kappa}$. AGD will be used as a basic component for acceleration in this paper.

3 Algorithm Components

In this section, we present two main algorithm components. Both of them are crucial for our final algorithms to achieve near-optimal convergence rates.

3.1 Inexact Accelerated Proximal Point Algorithm

Our first component is the Accelerated Proximal Point Algorithm (APPA, Algorithm 2) for minimizing a function $g(\cdot)$. Comparing APPA with classical AGD (Algorithm 1), we note that both of them have momentum steps which yield acceleration. The major difference is in Line 4 of Algorithm 2, where APPA solves a proximal subproblem

$$\mathbf{x}_t \leftarrow \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmin}}\ g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 \tag{4}$$

instead of performing a gradient-descent step as in AGD (Line 5 in Algorithm 1). We refer to the parameter ℓ in (4) as the *proximal parameter*.

We present an inexact version in Algorithm 2 where we tolerate a small error δ in terms of the function value in solving the proximal subproblem (4). That is, the solution $\mathbf{x}_t$ satisfies

$$g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \leq \min_{\mathbf{x} \in \mathcal{X}}\{g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\} + \delta$$

A theoretical guarantee for the inexact APPA algorithm is presented in the following theorem, which claims that as long as δ is sufficiently small, the algorithm finds an ϵ-optimal point of any μ-strongly-convex function g with proximal parameter ℓ in $\tilde{O}(\sqrt{\ell/\mu})$ iterations.

Theorem 2. *Assume that g is μ-strongly convex, $\epsilon \in (0,1)$ and $\ell > \mu$. There exists $T > 0$ such that the output $\hat{\mathbf{x}} = $ INEXACT-APPA$(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$ satisfies $g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) + \epsilon$ and $T > 0$ satisfies the following inequality,*

$$T \geq\ c\sqrt{\kappa}\log\left(\frac{g(\mathbf{x}_0) - g(\mathbf{x}^\star) + (\mu/4)\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)$$

where $\kappa = \ell/\mu$ is an effective condition number, $\mathbf{x}^\star \in \mathcal{X}$ is the unique global minimum of g, and $c > 0$ is an absolute constant.

Algorithm 3 MAXIMIN-AG2$(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{x}, \mu_\mathbf{y}, \epsilon)$

1: **Input:** initial point $\mathbf{x}_0, \mathbf{y}_0$, smoothness ℓ, strongly convex module $\mu_\mathbf{x}, \mu_\mathbf{y}$ and tolerance $\epsilon > 0$
2: **Initialize:** $t \leftarrow 0$, $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\eta \leftarrow \frac{1}{2\kappa_\mathbf{x}\ell}$, $\kappa_\mathbf{x} \leftarrow \frac{\ell}{\mu_\mathbf{x}}$, $\kappa_\mathbf{y} \leftarrow \frac{\ell}{\mu_\mathbf{y}}$, $\theta \leftarrow \frac{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}-1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}+1}$, $\tilde{\epsilon} \leftarrow \frac{\epsilon}{(10\kappa_\mathbf{x}\kappa_\mathbf{y})^7}$
3: **repeat**
4:     $t \leftarrow t + 1$
5:     $\tilde{\mathbf{x}}_{t-1} \leftarrow \text{AGD}(g(\cdot, \tilde{\mathbf{y}}_{t-1}), \mathbf{x}_0, \ell, \mu_\mathbf{x}, \tilde{\epsilon})$
6:     $\mathbf{y}_t \leftarrow \mathcal{P}_\mathcal{Y}[\tilde{\mathbf{y}}_{t-1} + \eta\nabla_\mathbf{y} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})]$
7:     $\tilde{\mathbf{y}}_t \leftarrow \mathbf{y}_t + \theta(\mathbf{y}_t - \mathbf{y}_{t-1})$
8:     $\mathbf{x}_t \leftarrow \text{AGD}(g(\cdot, \mathbf{y}_t), \mathbf{x}_0, \ell, \mu_\mathbf{x}, \tilde{\epsilon})$
9: **until** $\|\mathbf{y}_t - \mathcal{P}_\mathcal{Y}(\mathbf{y}_t + \eta\nabla_\mathbf{y} g(\mathbf{x}_t, \mathbf{y}_t))\|^2 \leq \frac{\epsilon}{(10\kappa_\mathbf{x}\kappa_\mathbf{y})^4\ell}$ is satisfied
10: **Output:** $\mathcal{P}_\mathcal{X}(\mathbf{x}_t - (1/2\kappa_\mathbf{y}\ell)\nabla_\mathbf{x} g(\mathbf{x}_t, \mathbf{y}_t))$

Comparing with Theorem 1, the most important difference here is that Theorem 2 does not require the function g to have any smoothness property. In fact, ℓ is only a proximal parameter in proximal subproblem (4), which does not necessarily relate to the smoothness of g. On the flip side, the proximal subproblem (4) can not be easily solved in general. Theorem 2 guarantees the iteration complexity of Algorithm 1 while the complexity for solving these proximal steps is not discussed.

We conclude that APPA has a unique advantage over AGD in settings where g does not have a smoothness property but the proximal step (4) is easy to solve. These settings include LASSO [Beck and Teboulle(2009)], as well as minimax optimization problems (as we show in later sections).

3.2   Accelerated Solver for Minimax Proximal Steps

In minimax optimization problems of the form (1), we are interested in solving the following proximal subproblem as follows,

$$\mathbf{x}_{t+1} \leftarrow \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}} \; \Phi(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}\|^2, \quad \text{where} \quad \Phi(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \tag{5}$$

which is equivalent to solving the following minimax problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \; \tilde{g}(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}\|^2 \tag{6}$$

For a generic strongly-convex-strongly-concave function $g(\cdot, \cdot)$, solving a minimax problem is equivalent to solving a maximin problem, due to Sion's minimax theorem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} g(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y})$$

A straightforward way of solving the maximin problem is to use a double-loop algorithm which solves the maximization and minimization problems on two different time scales. Specifically, the inner loop performs AGD on function $g(\cdot, \mathbf{y})$ to solve the inner minimization; i.e., to compute $\Psi(\mathbf{y}) := \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y})$ for each $\mathbf{y}$, and the outer loop performs Accelerated Gradient Ascent (AGA) on the function $\Psi(\cdot)$ to solve the outer maximization. Since the algorithm aims to solve a maximin problem we use AGA-AGD, and we name the algorithm MAXIMIN-AG2. See Algorithm 3 for the formal version of this algorithm. We also incorporate Lines 8-9 to check termination

conditions, which ensures that the output achieves the desired optimality. The theoretical guarantee for Algorithm 3 is given in the following theorem.

Theorem 3. *Assume that $g(\cdot, \cdot)$ is ℓ-smooth, $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex for each $\mathbf{y} \in \mathcal{Y}$ and $g(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly concave for each $\mathbf{x} \in \mathcal{X}$. Then $\hat{\mathbf{x}} = \text{MAXIMIN-AG2}(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon)$ satisfies that $\max_{\mathbf{y} \in \mathcal{Y}} g(\hat{\mathbf{x}}, \mathbf{y}) \leq \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y}) + \epsilon$ and the total number of gradient evaluations is bounded by*

$$O\left(\kappa_{\mathbf{x}} \sqrt{\kappa_{\mathbf{y}}} \cdot \log^2 \left(\frac{(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon} \right) \right)$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ are condition numbers, $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \text{argmin}_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$ and $D_{\mathbf{y}} > 0$ is the diameter of the constraint set $\mathcal{Y}$.

Theorem 3 claims that Algorithm 3 finds an ϵ-optimal point in $\tilde{O}(\kappa_{\mathbf{x}} \sqrt{\kappa_{\mathbf{y}}})$ iterations for strongly-convex-strongly-concave functions. This rate does not match the lower bound $\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}} \kappa_{\mathbf{y}}})$ [Ibrahim et al.(2019), Zhang et al.(2019)]. At a high level, it takes AGD $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}})$ steps to solve the inner minimization problem and compute $\Psi(\mathbf{y}) := \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{y})$. Despite the fact that the function g is ℓ-smooth, function Ψ is only guaranteed to be $(\kappa_{\mathbf{x}}\ell)$-smooth in the worst case, which makes the condition number of Ψ be $\kappa_{\mathbf{x}} \kappa_{\mathbf{y}}$. Thus, AGA requires $\tilde{O}(\sqrt{\kappa_{\mathbf{x}} \kappa_{\mathbf{y}}})$ iterations in the outer loop to solve the maximization of Ψ, which gives a total gradient complexity $\tilde{O}(\kappa_{\mathbf{x}} \sqrt{\kappa_{\mathbf{y}}})$.

The key observation here is that although Algorithm 3 is slow for general strongly-convex-strongly-concave functions, the functions $\tilde{g}$ of the form (6) in the proximal steps have a crucial property that $\kappa_{\mathbf{x}} = O(1)$ if the proximal parameter ℓ is chosen to be the smoothness parameter of function f. Therefore, when $f(\mathbf{x}, \cdot)$ is strongly concave, by Theorem 3, it only takes Algorithm 3 $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}})$ gradient evaluations to solve the proximal subproblem (6), which is very efficient. We will see the consequences of this fact in the following section.

4 Accelerating Convex-Concave Optimization

In this section, we present our main results for accelerating convex-concave optimization. We first present our new near-optimal algorithm and its theoretical guarantee for optimizing strongly-convex-strongly-concave functions. Then, we use simple reduction arguments to obtain results for strongly-convex-concave and convex-concave functions.

4.1 Strongly-convex-strongly-concave setting

With the algorithm components from Section 3 in hand, we are now ready to state our near-optimal algorithm. Algorithm 4 is a simple combination of Algorithm 2 and Algorithm 3. Its outer loop performs an inexact APPA to minimize the function $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$, while the inner loop uses Maximin-AG2 to solve the proximal subproblem (5), which is equivalent to solving (6). At the end, after finding a near-optimal $\mathbf{x}_T$, Algorithm 4 performs another AGD on the function $-f(\mathbf{x}_T, \cdot)$ to find a near-optimal $\mathbf{y}_T$. The theoretical guarantee for the algorithm is given in the following theorem.

Theorem 4. *Assume that f is ℓ-smooth and $\mu_{\mathbf{x}}$-strongly-convex-$\mu_{\mathbf{y}}$-strongly-concave. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon, T)$ is an ϵ-saddle*

Algorithm 4 MINIMAX-APPA($f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon, T$)

1: **Input:** initial point $\mathbf{x}_0, \mathbf{y}_0$, proximity ℓ, strongly-convex parameter μ, tolerance δ, iteration T
2: **Initialize:** $\tilde{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\kappa_{\mathbf{x}} \leftarrow \frac{\ell}{\mu_{\mathbf{x}}}$, $\theta \leftarrow \frac{2\sqrt{\kappa_{\mathbf{x}}}-1}{2\sqrt{\kappa_{\mathbf{x}}}+1}$, $\delta \leftarrow \frac{\epsilon}{(10\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^4}$ and $\tilde{\epsilon} \leftarrow \frac{\epsilon}{10^2\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}$
3: **for** $t = 1, \cdots, T$ **do**
4:     denote $g_t(\cdot, \cdot)$ where $g_t(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2$
5:     $\mathbf{x}_t \leftarrow$ MAXIMIN-AG2($g_t, \mathbf{x}_0, \mathbf{y}_0, 3\ell, 2\ell, \mu_{\mathbf{y}}, \delta$)
6:     $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1})$
7: **end for**
8: $\tilde{\mathbf{y}} \leftarrow$ AGD($-f(\mathbf{x}_T, \cdot), \mathbf{y}_0, \ell, \mu_{\mathbf{y}}, \tilde{\epsilon}$)
9: $\mathbf{y}_T \leftarrow \mathcal{P}_{\mathcal{Y}}(\tilde{\mathbf{y}} + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}} f(\mathbf{x}_T, \tilde{\mathbf{y}}))$
10: **Output:** $(\mathbf{x}_T, \mathbf{y}_T)$

---

point, and the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}\log^3\left(\frac{(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})\ell(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ are condition numbers.

    Theorem 4 asserts that Algorithm 4 finds ϵ-saddle points in $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$ gradient evaluations, matching the lower bound [Ibrahim et al.(2019), Zhang et al.(2019)], up to logarithmic factors. At a high level, despite the function Φ having undesirable smoothness properties, APPA minimizes Φ in the outer loop using $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}})$ iterations according to Theorem 2, regardless of the smoothness of Φ. According to the discussion in Section 3.2, Maximin-AG2 solves the proximal step in the inner loop using $\tilde{O}(\sqrt{\kappa_{\mathbf{y}}})$ gradient evaluations, since the condition number of $g_t(\cdot, \mathbf{y})$ for any $\mathbf{y} \in \mathcal{Y}$ is $O(1)$. This gives the total gradient complexity $\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$.

4.2 Strongly-convex-concave setting

Our result in the strongly-convex-strongly-concave setting readily implies a near-optimal result in the strongly-convex-concave setting. Consider the following auxiliary function for an arbitrary $\mathbf{y}_0 \in \mathcal{Y}$ which is defined by

$$f_{\epsilon,\mathbf{y}}(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) - \frac{\epsilon}{4D_{\mathbf{y}}^2}\|\mathbf{y} - \mathbf{y}_0\|^2 \tag{7}$$

By construction, it is clear that the difference between f and $f_{\epsilon,\mathbf{y}}$ is small in terms of function value:

$$\max_{(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}} |f(\mathbf{x}, \mathbf{y}) - f_{\epsilon,\mathbf{y}}(\mathbf{x}, \mathbf{y})| \leq \epsilon/4$$

This implies, according to Definition 4, that any $(\epsilon/2)$-saddle point of function $f_{\epsilon,\mathbf{y}}$ is also a ϵ-saddle point of function f, and thus it is sufficient to only solve the problem $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{x} \in \mathcal{Y}} f_{\epsilon,\mathbf{y}}(\mathbf{x}, \mathbf{y})$. Finally, when f is a $\mu_{\mathbf{x}}$-strongly-convex-concave function, $f_{\epsilon,\mathbf{y}}$ becomes $\mu_{\mathbf{x}}$-strongly-convex-$\epsilon/(2D_{\mathbf{y}}^2)$-strongly-concave, which can be fed into Algorithm 4 to obtain the following result.

Corollary 1. *Assume that f is ℓ-smooth and $\mu_{\mathbf{x}}$-strongly-convex-concave. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$ MINIMAX-APPA($f_{\epsilon,\mathbf{y}}, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \epsilon/(4D_{\mathbf{y}}^2), \epsilon/2, T$) is an ϵ-saddle*

point, and the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\frac{\kappa_{\mathbf{x}}\ell}{\epsilon}}D_{\mathbf{y}}\log^3\left(\frac{\kappa_{\mathbf{x}}\ell(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ is the condition number, and $f_{\epsilon,\mathbf{y}}$ is defined as in (7).

4.3 Convex-concave setting

Similar to the previous subsection, when f is only convex-concave, we can construct following strongly-convex-strongly-concave function f_ϵ:

$$f_\epsilon(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) + \frac{\epsilon}{8D_{\mathbf{x}}^2}\|\mathbf{x} - \mathbf{x}_0\|^2 - \frac{\epsilon}{8D_{\mathbf{y}}^2}\|\mathbf{y} - \mathbf{y}_0\|^2 \tag{8}$$

which can be fed into Algorithm 4 to obtain the following result.

Corollary 2. *Assume function f is ℓ-smooth and convex-concave, then there exists $T > 0$, where the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \frac{\epsilon}{4D_{\mathbf{x}}^2}, \frac{\epsilon}{4D_{\mathbf{y}}^2}, \epsilon/2, T)$ will be an ϵ-saddle point, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell D_{\mathbf{x}} D_{\mathbf{y}}}{\epsilon}\log^3\left(\frac{\ell(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where f_ϵ is defined as in (8).

5 Proofs for Algorithm Components

In this section, we present proofs for our algorithm components.

5.1 Proof of Theorem 1

We divide the proof into three parts. In the first part, we show that the output $\hat{\mathbf{x}}$ satisfies $g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) + \epsilon$. In the second part, we derive the sufficient condition for guaranteeing the stopping criteria in Algorithm 1. In the third part, we derive the gradient complexity of the algorithm using the condition derived in the second part.

Part I. Let $\tilde{\mathbf{x}}_t = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t))$ be defined as the point achieved by one-step projected gradient descent from $\mathbf{x}_t$. Since g is ℓ-smooth and μ-strongly convex, it is straightforward to derive from [Nesterov(2018), Corollary 2.3.2] that

$$g(\mathbf{x}) \geq g(\tilde{\mathbf{x}}_t) + \ell(\mathbf{x}_t - \tilde{\mathbf{x}}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \frac{\ell}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_t\|^2, \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

Using the Young's inequality, we have $(\mathbf{x}_t - \tilde{\mathbf{x}}_t)^\top(\mathbf{x} - \mathbf{x}_t) \geq -(1/2)(\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 + \|\mathbf{x} - \mathbf{x}_t\|^2)$. Putting these pieces together with $\mathbf{x} = \mathbf{x}^\star$ yields that

$$g(\tilde{\mathbf{x}}_t) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) = g(\tilde{\mathbf{x}}_t) - g(\mathbf{x}^\star) \leq \left(\frac{\ell - \mu}{2}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

Without loss of generality, we assume $\ell > \mu$. Indeed, if $\ell = \mu$, then one-step projected gradient descent from any points in $\mathcal{X}$ guarantees that $g(\tilde{\mathbf{x}}_t) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) = 0$. Since $\hat{\mathbf{x}} = \tilde{\mathbf{x}}_t$ in Algorithm 1, it suffices to show that the following statement holds true,

$$\left\| \mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t)) \right\| \leq \sqrt{\frac{\epsilon}{2\kappa^2(\ell - \mu)}} \implies \|\mathbf{x}_t - \mathbf{x}^\star\| \leq \sqrt{\frac{2\epsilon}{\ell - \mu}} \tag{9}$$

Let $\tilde{\mathbf{x}}_t = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t))$ be defined as the point achieved by one-step projected gradient descent from $\mathbf{x}_t$, the ℓ-smoothness of g implies

$$\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\| \leq \|\mathbf{x}_k - \mathbf{x}^\star\| \tag{10}$$

Using the definition of $\tilde{\mathbf{x}}_t$ and $\mathbf{x}^\star$, we have

$$(\mathbf{x}^\star - \tilde{\mathbf{x}}_t)^\top (\ell(\tilde{\mathbf{x}}_t - \mathbf{x}_t) + \nabla g(\mathbf{x}_t)) \geq 0, \qquad (\tilde{\mathbf{x}}_t - \mathbf{x}^\star)^\top \nabla g(\mathbf{x}^\star) \geq 0$$

Summing up the above two inequalities and rearranging yields that

$$(\mathbf{x}^\star - \mathbf{x}_t)^\top (\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}^\star)) \geq \ell(\mathbf{x}^\star - \tilde{\mathbf{x}}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_t) + (\tilde{\mathbf{x}}_t - \mathbf{x}_t)^\top (\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}^\star))$$

Since g is ℓ-smooth and μ-strongly convex, we have

$$-\mu\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \geq -\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|(\|\mathbf{x}^\star - \tilde{\mathbf{x}}_t\| + \|\mathbf{x}^\star - \mathbf{x}_t\|) \overset{(10)}{\geq} -2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|\|\mathbf{x}_t - \mathbf{x}^\star\|$$

Therefore, we conclude that

$$\|\mathbf{x}_t - \mathbf{x}^\star\| \leq 2\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\| = 2\kappa\left\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t))\right\| \overset{(9)}{\leq} \sqrt{\frac{2\epsilon}{\ell - \mu}}$$

Part II. We first show that

$$\|\mathbf{x}_t - \mathbf{x}^\star\| \leq \frac{1}{3\kappa}\sqrt{\frac{\epsilon}{2(\ell - \mu)}} \implies \left\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t))\right\| \leq \sqrt{\frac{\epsilon}{2\kappa^2(\ell - \mu)}}$$

By the definition of $\mathbf{x}^\star$, we have $\mathbf{x}^\star = \mathcal{P}_{\mathcal{X}}(\mathbf{x}^\star - \frac{1}{\ell}\nabla g(\mathbf{x}^\star))$. This equation together with the triangle inequality and the nonexpansiveness of $\mathcal{P}_{\mathcal{X}}$ yields that $\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \frac{1}{\ell}\nabla g(\mathbf{x}_t))\| \leq 3\|\mathbf{x}_t - \mathbf{x}^\star\|$ which implies the desired result. Then we derive a sufficient condition for guaranteeing that $\|\mathbf{x}_t - \mathbf{x}^\star\| \leq (1/(3\kappa))\sqrt{\epsilon/(2(\ell - \mu))}$. Since g is μ-strongly convex and $\mathbf{x}_t \in \mathcal{X}$, [Nesterov(2018), Theorem 2.1.5] together with the fact that $(\mathbf{x}_t - \mathbf{x}^\star)^\top \nabla g(\mathbf{x}^\star) \geq 0$ implies that

$$\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \leq \frac{2}{\mu}\left(g(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x})\right)$$

Putting these pieces together yields the desired sufficient condition as follows,

$$g(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) \leq \frac{\epsilon}{36\kappa^3} \tag{11}$$

13

Part III. We proceed to derive the gradient complexity of the algorithm using the condition in Eq. (11). Since Algorithm 1 is exactly Nesterov's accelerated gradient descent, standard arguments based on estimate sequence [Nesterov(2018)] implies

$$g(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(g(\mathbf{x}_0) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) + \frac{\mu \|\mathbf{x}^\star - \mathbf{x}_0\|^2}{2}\right)$$

Therefore, the gradient complexity of Algorithm 1 to guarantee Eq. (11) is bounded by

$$O\left(1 + \sqrt{\kappa} \log\left(\frac{\kappa^3 \ell \|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right)$$

This completes the proof.

5.2 Proof of Theorem 2

Letting $\hat{\mathbf{x}} = \text{INEXACT-APPA}(g, \mathbf{x}_0, \ell, \mu, \epsilon, T)$. Since $\hat{\mathbf{x}} = \mathbf{x}_T$, it suffices for us to estimate an lower bound for the maximum number of iterations T such that $g(\mathbf{x}_T) \leq \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) + \epsilon$. The following technical lemma is crucial to the subsequent analysis.

Lemma 1. *For any $\mathbf{x} \in \mathcal{X}$ and $\{(\mathbf{x}_t, \tilde{\mathbf{x}}_t)\}_{t \geq 0}$ generated by Algorithm 2, we have*

$$g(\mathbf{x}) \geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \tilde{\mathbf{x}}_{t-1})^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + 2\ell \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + \frac{\mu \|\mathbf{x} - \mathbf{x}_t\|^2}{4} - 7\kappa\delta \qquad (12)$$

Proof. Using the definition of $\mathbf{x}_t$ in Algorithm 2, we have

$$g(\mathbf{x}_t) + \ell \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \leq \min_{\mathbf{x} \in \mathcal{X}} \left\{g(\mathbf{x}) + \ell \|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\right\} + \delta$$

Defining $\mathbf{x}_t^\star = \text{argmin}_{\mathbf{x} \in \mathcal{X}}\{g(\mathbf{x}) + \ell \|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2\}$ and using μ-strongly convexity of g, we have the following for any $\mathbf{x} \in \mathcal{X}$:

$$g(\mathbf{x}) \geq g(\mathbf{x}_t^\star) + \ell \|\mathbf{x}_t^\star - \tilde{\mathbf{x}}_{t-1}\|^2 - \ell \|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 + \left(\ell + \frac{\mu}{2}\right) \|\mathbf{x} - \mathbf{x}_t^\star\|^2$$

Equivalently, we have

$$\begin{aligned} g(\mathbf{x}) &\geq g(\mathbf{x}_t) + \ell \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 - \ell \|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 + \left(\ell + \frac{\mu}{2}\right) \|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \delta \\ &\geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) - \ell \|\mathbf{x} - \mathbf{x}_t\|^2 + \left(\ell + \frac{\mu}{2}\right) \|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \delta \end{aligned}$$

On the other hand, we have

$$\left(\ell + \frac{\mu}{2}\right) \|\mathbf{x} - \mathbf{x}_t^\star\|^2 - \ell \|\mathbf{x} - \mathbf{x}_t\|^2 = \frac{\mu \|\mathbf{x} - \mathbf{x}_t\|^2}{2} + (2\ell + \mu)(\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^\star) + \left(\ell + \frac{\mu}{2}\right) \|\mathbf{x}_t - \mathbf{x}_t^\star\|^2$$

Using Young's inequality yields

$$(\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}_t^\star) \geq -\frac{\mu \|\mathbf{x} - \mathbf{x}_t\|^2}{4(2\ell + \mu)} - (1 + 2\kappa)\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2$$

Putting these pieces together yields that

$$g(\mathbf{x}) \geq g(\mathbf{x}_t) - 2\ell(\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + \frac{\mu \|\mathbf{x} - \mathbf{x}_t\|^2}{4} - (2\ell + \mu)(1 + 2\kappa)\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2 - \delta$$

14

Furthermore, we have

$$(\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) \;=\; (\mathbf{x} - \tilde{\mathbf{x}}_{t-1})^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) - \|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2$$

and

$$\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2 \;\leq\; \frac{2}{\mu + 2\ell} \left(g(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 - \min_{\mathbf{x}\in\mathcal{X}}\left\{ g(\mathbf{x}) + \ell\|\mathbf{x} - \tilde{\mathbf{x}}_{t-1}\|^2 \right\} \right) \;\leq\; \frac{2\delta}{\mu + 2\ell}$$

Putting these pieces together with $\kappa \geq 1$ yields the desired inequality. $\qquad\square$

The remaining proof is based on Lemma 1. Indeed, we have

$$\left(1 - \frac{1}{2\sqrt{\kappa}}\right) g(\mathbf{x}_{t-1}) + \frac{1}{2\sqrt{\kappa}}\left(g(\mathbf{x}^\star) + 14\kappa^{3/2}\delta \right)$$

$$\overset{\text{Eq. (12)}}{\geq}\; \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\left(g(\mathbf{x}_t) - 2\ell(\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + \frac{\mu\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2}{4} - 7\kappa\delta \right)$$

$$+\frac{1}{2\sqrt{\kappa}}\left(g(\mathbf{x}_t) - 2\ell(\mathbf{x}^\star - \tilde{\mathbf{x}}_{t-1})^\top(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + \frac{\mu\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{4} - 7\kappa\delta \right) + 7\kappa\delta$$

$$=\; g(\mathbf{x}_t) - 2\ell\left(\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} + \frac{\mathbf{x}^\star}{2\sqrt{\kappa}} - \tilde{\mathbf{x}}_{t-1} \right)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + 2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 + \frac{\mu\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{8\sqrt{\kappa}}$$

Equivalently, we have

$$g(\mathbf{x}_t) - g(\mathbf{x}^\star) \;\leq\; \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\left(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star) \right) + 2\ell\left(\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} + \frac{\mathbf{x}^\star}{2\sqrt{\kappa}} - \tilde{\mathbf{x}}_{t-1} \right)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1})$$

$$-2\ell\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 - \frac{\mu\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{8\sqrt{\kappa}} + 7\kappa\delta \tag{13}$$

Consider $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}(\mathbf{x}_t - \mathbf{x}_{t-1})$, we let $\mathbf{w}_t = \tilde{\mathbf{x}}_t + 2\sqrt{\kappa}(\tilde{\mathbf{x}}_t - \mathbf{x}_t)$ and obtain that

$$\mathbf{w}_t \;=\; (1 + 2\sqrt{\kappa})\tilde{\mathbf{x}}_t - 2\sqrt{\kappa}\mathbf{x}_t \;=\; 2\sqrt{\kappa}\mathbf{x}_t - (2\sqrt{\kappa} - 1)\mathbf{x}_{t-1} \;=\; \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + 2\sqrt{\kappa}\mathbf{x}_t - \frac{4\kappa - 1}{2\sqrt{\kappa}}\tilde{\mathbf{x}}_{t-1}$$

$$=\; \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + 2\sqrt{\kappa}\,(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}}$$

This implies that

$$\|\mathbf{w}_t - \mathbf{x}^\star\|^2 \;=\; \left\| \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star + 2\sqrt{\kappa}\,(\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}) \right\|^2 \tag{14}$$

$$=\left\| \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star \right\|^2 + 4\sqrt{\kappa}\left(\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{w}_{t-1} + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star \right)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1})$$

$$+4\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2$$

Since $\mathbf{w}_{t-1} = \tilde{\mathbf{x}}_{t-1} + 2\sqrt{\kappa}(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1})$, we have

$$(1 - \frac{1}{2\sqrt{\kappa}})\mathbf{w}_{t-1} + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} \;=\; 2\sqrt{\kappa}\tilde{\mathbf{x}}_{t-1} - (2\sqrt{\kappa} - 1)\mathbf{x}_{t-1} \tag{15}$$

15

Using the Young's inequality, we have

$$\left\| \left(1 - \frac{1}{2\sqrt{\kappa}}\right) \mathbf{w}_{t-1} + \frac{\tilde{\mathbf{x}}_{t-1}}{2\sqrt{\kappa}} - \mathbf{x}^\star \right\|^2 \tag{16}$$

$$\leq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)^2 \left(1 + \frac{5}{8\sqrt{\kappa} - 5}\right) \|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \frac{1}{4\kappa}\left(1 + \frac{8\sqrt{\kappa} - 5}{5}\right) \|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2$$

$$\leq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\left(1 + \frac{1}{8\sqrt{\kappa} - 5}\right) \|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \frac{2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2}{5\sqrt{\kappa}}$$

$$\leq \left(1 - \frac{1}{6\sqrt{\kappa}}\right) \|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \frac{2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2}{5\sqrt{\kappa}}$$

Using the Young's inequality again, we have

$$\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}^\star\|^2 \leq \frac{5\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{4} + 5\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_t\|^2 \tag{17}$$

Putting Eq. (14)-Eq. (17) together with $\kappa \geq 1$, we have

$$\|\mathbf{w}_t - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{1}{6\sqrt{\kappa}}\right) \|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2 + \frac{\|\mathbf{x}^\star - \mathbf{x}_t\|^2}{2\sqrt{\kappa}} + 6\kappa\|\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1}\|^2 \tag{18}$$

$$+ 8\kappa\left(\tilde{\mathbf{x}}_{t-1} - \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\mathbf{x}_{t-1} - \frac{\mathbf{x}^\star}{2\sqrt{\kappa}}\right)^\top (\mathbf{x}_t - \tilde{\mathbf{x}}_{t-1})$$

Combining Eq. (13) and Eq. (18) yields that

$$g(\mathbf{x}_t) - g(\mathbf{x}^\star) + \frac{\mu\|\mathbf{w}_t - \mathbf{x}^\star\|^2}{4} \leq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star)) + \left(1 - \frac{1}{6\sqrt{\kappa}}\right)\frac{\mu\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2}{4} + 7\kappa\delta$$

$$\leq \left(1 - \frac{1}{6\sqrt{\kappa}}\right)\left(g(\mathbf{x}_{t-1}) - g(\mathbf{x}^\star) + \frac{\mu\|\mathbf{w}_{t-1} - \mathbf{x}^\star\|^2}{4}\right) + 7\kappa\delta$$

Repeating the above inequality yields that

$$g(\mathbf{x}_T) - g(\mathbf{x}^\star) + \frac{\mu\|\mathbf{w}_T - \mathbf{x}^\star\|^2}{4} \leq \left(1 - \frac{1}{6\sqrt{\kappa}}\right)^T \left(g(\mathbf{x}_0) - g(\mathbf{x}^\star) + \frac{\mu\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{4}\right) + 42\kappa^{3/2}\delta$$

Therefore, we conclude that

$$g(\mathbf{x}_T) - g(\mathbf{x}^\star) \leq \left(1 - \frac{1}{6\sqrt{\kappa}}\right)^T \left(g(\mathbf{x}_0) - g(\mathbf{x}^\star) + \frac{\mu\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{4}\right) + 42\kappa^{3/2}\delta$$

Since the tolerance $\delta \leq \epsilon\kappa^{-3/2}/84$, we conclude that the iteration complexity of Algorithm 2 to guarantee that $g(\mathbf{x}_T) - \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}) \leq \epsilon$ if there exists an absolute constant $c > 0$ such that

$$T \geq c\sqrt{\kappa}\log\left(\frac{g(\mathbf{x}_0) - g(\mathbf{x}^\star) + (\mu/4)\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)$$

This completes the proof.

5.3 Proof of Theorem 3

Before presenting the main proof, we define the following important functions:

$$
\begin{aligned}
\Phi_g(\cdot) &= \max_{\mathbf{y}\in\mathcal{Y}} \ g(\cdot,\mathbf{y}), & \mathbf{y}_g^\star(\cdot) &= \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} \ g(\cdot,\mathbf{y}) \\
\Psi_g(\cdot) &= \min_{\mathbf{x}\in\mathcal{X}} \ g(\mathbf{x},\cdot), & \mathbf{x}_g^\star(\cdot) &= \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \ g(\mathbf{x},\cdot)
\end{aligned}
$$

All the above functions are well defined since $g(\cdot,\cdot)$ is strongly convex-concave. We provide their complete characterization in the following structural lemma.

Lemma 2. *Under the assumptions imposed in Theorem 3, we have*

(a) *A function $\mathbf{y}_g^\star(\cdot)$ is $\kappa_{\mathbf{y}}$-Lipschitz.*

(b) *A function $\Phi_g(\cdot)$ is $2\kappa_{\mathbf{y}}\ell$-smooth and $\mu_{\mathbf{x}}$-strongly convex with $\nabla\Phi_g(\cdot) = \nabla_{\mathbf{x}}g(\cdot,\mathbf{y}_g^\star(\cdot))$.*

(c) *A function $\mathbf{x}_g^\star(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz.*

(d) *A function $\Psi_g(\cdot)$ is $2\kappa_{\mathbf{x}}\ell$-smooth and $\mu_{\mathbf{y}}$-strongly concave with $\nabla\Psi_g(\cdot) = \nabla_{\mathbf{y}}g(\mathbf{x}_g^\star(\cdot),\cdot)$.*

where $\kappa_{\mathbf{x}} = \ell/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ are condition numbers.

Proof. We proceed as follows

Part (a): Let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, the points $\mathbf{y}_g^*(\mathbf{x})$ and $\mathbf{y}_g^*(\mathbf{x}')$ satisfy that

$$
(\mathbf{y} - \mathbf{y}_g^*(\mathbf{x}))^\top \nabla_{\mathbf{y}}g(\mathbf{x},\mathbf{y}_g^*(\mathbf{x})) \ \leq \ 0, \qquad \forall \mathbf{y}\in\mathcal{Y} \tag{19}
$$

$$
(\mathbf{y} - \mathbf{y}_g^*(\mathbf{x}'))^\top \nabla_{\mathbf{y}}g(\mathbf{x}',\mathbf{y}_g^*(\mathbf{x}')) \ \leq \ 0, \qquad \forall \mathbf{y}\in\mathcal{Y} \tag{20}
$$

Summing up Eq. (19) with $\mathbf{y} = \mathbf{y}_g^*(\mathbf{x}')$ and Eq. (20) with $\mathbf{y} = \mathbf{y}_g^*(\mathbf{x})$ yields

$$
(\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x}))^\top (\nabla_{\mathbf{y}}g(\mathbf{x},\mathbf{y}_g^*(\mathbf{x})) - \nabla_{\mathbf{y}}g(\mathbf{x}',\mathbf{y}_g^*(\mathbf{x}'))) \ \leq \ 0
$$

Since $g(\mathbf{x},\cdot)$ is $\mu_{\mathbf{y}}$-strongly concave, we have

$$
(\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x}))^\top (\nabla_{\mathbf{y}}g(\mathbf{x},\mathbf{y}_g^*(\mathbf{x}')) - \nabla_{\mathbf{y}}g(\mathbf{x},\mathbf{y}_g^*(\mathbf{x}))) + \mu_{\mathbf{y}}\|\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x})\|^2 \ \leq \ 0
$$

Summing up the above two inequalities yields that

$$
(\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x}))^\top (\nabla_{\mathbf{y}}g(\mathbf{x},\mathbf{y}_g^*(\mathbf{x}')) - \nabla_{\mathbf{y}}g(\mathbf{x}',\mathbf{y}_g^*(\mathbf{x}'))) + \mu_{\mathbf{y}}\|\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x})\|^2 \ \leq \ 0
$$

Since $\nabla_{\mathbf{y}}g$ is ℓ-Lipschitz, we have

$$
\mu_{\mathbf{y}}\|\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x})\|^2 \ \leq \ \ell\|\mathbf{y}_g^*(\mathbf{x}') - \mathbf{y}_g^*(\mathbf{x})\|\|\mathbf{x}' - \mathbf{x}\|
$$

Therefore, we conclude that the function $\mathbf{y}_g^*(\cdot)$ is $\kappa_{\mathbf{y}}$-Lipschitz.

17

Part (b): Since the function $\mathbf{y}_g^*(\cdot)$ is unique, Danskin's theorem [Rockafellar(1970)] implies that Φ_g is differentiable and $\nabla\Phi_g(\cdot) = \nabla_{\mathbf{x}}g(\cdot, \mathbf{y}_g^*(\cdot))$. Let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, we have

$$
\begin{aligned}
\|\nabla\Phi_g(\mathbf{x}) - \nabla\Phi_g(\mathbf{x}')\| &= \|\nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}_g^*(\mathbf{x})) - \nabla_{\mathbf{x}}g(\mathbf{x}', \mathbf{y}_g^*(\mathbf{x}'))\| \leq \ell\|\mathbf{x} - \mathbf{x}'\| + \ell\|\mathbf{y}_g^*(\mathbf{x}) - \mathbf{y}_g^*(\mathbf{x}')\| \\
&\overset{\bar{\kappa}\geq 1}{\leq} \kappa_{\mathbf{y}}\ell\|\mathbf{x} - \mathbf{x}'\| + \ell\|\mathbf{y}_g^*(\mathbf{x}) - \mathbf{y}_g^*(\mathbf{x}')\|
\end{aligned}
$$

Since $\mathbf{y}_g^*(\cdot)$ is $\kappa_{\mathbf{y}}$-Lipschitz, the function Φ_g is $2\kappa_{\mathbf{y}}\ell$-smooth. Furthermore, let $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, we have

$$
\begin{aligned}
\Phi_g(\mathbf{x}') - \Phi_g(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top\nabla\Phi_g(\mathbf{x}) &= g(\mathbf{x}', \mathbf{y}_g^*(\mathbf{x}')) - g(\mathbf{x}, \mathbf{y}_g^*(\mathbf{x})) - (\mathbf{x}' - \mathbf{x})^\top\nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}_g^*(\mathbf{x})) \\
&\geq g(\mathbf{x}', \mathbf{y}_g^*(\mathbf{x})) - g(\mathbf{x}, \mathbf{y}_g^*(\mathbf{x})) - (\mathbf{x}' - \mathbf{x})^\top\nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}_g^*(\mathbf{x}))
\end{aligned}
$$

Since $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex for each $\mathbf{y} \in \mathcal{Y}$, we have

$$
\Phi_g(\mathbf{x}') - \Phi_g(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top\nabla\Phi_g(\mathbf{x}) \geq \frac{\mu_{\mathbf{x}}\|\mathbf{x}' - \mathbf{x}\|^2}{2}
$$

Therefore, the function Φ_g is $\mu_{\mathbf{x}}$-strongly convex.

Part (c): Let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, the points $\mathbf{x}_g^*(\mathbf{y})$ and $\mathbf{x}_g^*(\mathbf{y}')$ satisfy that

$$
(\mathbf{x} - \mathbf{x}_g^*(\mathbf{y}))^\top\nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) \geq 0, \qquad \forall\mathbf{x} \in \mathcal{X} \tag{21}
$$

$$
(\mathbf{x} - \mathbf{x}_g^*(\mathbf{y}'))^\top\nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}'), \mathbf{y}') \geq 0, \qquad \forall\mathbf{x} \in \mathcal{X} \tag{22}
$$

Summing up Eq. (21) with $\mathbf{x} = \mathbf{x}_g^*(\mathbf{y}')$ and Eq. (22) with $\mathbf{x} = \mathbf{x}_g^*(\mathbf{y})$ yields

$$
(\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y}))^\top(\nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}'), \mathbf{y}')) \geq 0
$$

Since $g(\cdot, \mathbf{y})$ is $\mu_{\mathbf{x}}$-strongly convex, we have

$$
(\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y}))^\top(\nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}'), \mathbf{y}') - \nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}')) - \mu_{\mathbf{x}}\|\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y})\|^2 \geq 0
$$

Summing up the above two inequalities yields that

$$
(\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y}))^\top(\nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}')) - \mu_{\mathbf{x}}\|\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y})\|^2 \geq 0
$$

Since $\nabla_{\mathbf{x}}g$ is ℓ-smooth, we have

$$
\mu_{\mathbf{x}}\|\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y})\|^2 \leq \ell\|\mathbf{x}_g^*(\mathbf{y}') - \mathbf{x}_g^*(\mathbf{y})\|\|\mathbf{y}' - \mathbf{y}\|
$$

Therefore, we conclude that the function $\mathbf{x}_g^*$ is $\kappa_{\mathbf{x}}$-Lipschitz.

Part (d): Since the function $\mathbf{x}_g^*(\cdot)$ is unique, Danskin's theorem [Rockafellar(1970)] implies that Ψ_g is differentiable and $\nabla\Psi_g(\cdot) = \nabla_{\mathbf{y}}g(\mathbf{x}_g^*(\cdot), \cdot)$. Let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have

$$
\|\nabla\Psi_g(\mathbf{y}) - \nabla\Psi_g(\mathbf{y}')\| = \|\nabla_{\mathbf{y}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) - \nabla_{\mathbf{y}}g(\mathbf{x}_g^*(\mathbf{y}'), \mathbf{y})\| \leq \ell\|\mathbf{x}_g^*(\mathbf{y}) - \mathbf{x}_g^*(\mathbf{y}')\| + \ell\|\mathbf{y} - \mathbf{y}'\|
$$

Since $\mathbf{x}_g^*(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz, the function Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth. Furthermore, let $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have

$$
\begin{aligned}
\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla\Psi_g(\mathbf{y}) &= g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) - g(\mathbf{x}_g^*(\mathbf{y}'), \mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla_{\mathbf{y}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) \\
&\geq g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}) - g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla_{\mathbf{y}}g(\mathbf{x}_g^*(\mathbf{y}), \mathbf{y})
\end{aligned}
$$

Since $g(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$
\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}') - (\mathbf{y} - \mathbf{y}')^\top\nabla\Psi_g(\mathbf{y}) \geq \frac{\mu_{\mathbf{y}}\|\mathbf{y}' - \mathbf{y}\|^2}{2}
$$

Therefore, the function Ψ_g is $\mu_{\mathbf{y}}$-strongly concave. $\qquad\square$

Now we are ready to prove Theorem 3. We divide the proof into three parts. In the first part, we show that the output $\hat{\mathbf{x}} = \text{MAXIMIN-AG2}(g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{x}}, \mu_{\mathbf{y}}, \epsilon)$ satisfies

$$\max_{\mathbf{y} \in \mathcal{Y}} g(\hat{\mathbf{x}}, \mathbf{y}) \leq \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y}) + \epsilon \tag{23}$$

In the second part, we get the sufficient condition for guaranteeing the stopping criteria in Algorithm 3. In the third part, we estimate an upper bound for the gradient complexity of the algorithm using the condition derived in the second part. For the ease of presentation, we denote $(\mathbf{x}_g^{\star}, \mathbf{y}_g^{\star})$ as the unique solution to the minimax optimization $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$.

Part I. By the definition of Φ_g, the inequality in Eq. (23) can be rewritten as follows,

$$\Phi_g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{X}} \Phi_g(\mathbf{x}) + \epsilon$$

Since $\hat{\mathbf{x}} = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_T - (1/2\kappa_{\mathbf{y}}\ell)\nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_T))$, we have

$$\begin{aligned}
0 &\leq (\mathbf{x} - \hat{\mathbf{x}})^{\top}(2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_T)) \\
&= (\mathbf{x} - \hat{\mathbf{x}})^{\top}(2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla\Phi_g(\mathbf{x}_T)) + (\mathbf{x} - \hat{\mathbf{x}})^{\top}(\nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T))
\end{aligned}$$

Since $\nabla\Phi_g(\mathbf{x}_T) = \nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_g^{\star}(\mathbf{x}_T))$, we have $\|\nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T)\| \leq \ell\|\mathbf{y}_T - \mathbf{y}_g^{\star}(\mathbf{x}_T)\|$. Using the Young's inequality, we have

$$(\mathbf{x} - \hat{\mathbf{x}})^{\top}(\nabla_{\mathbf{x}}g(\mathbf{x}_T, \mathbf{y}_T) - \nabla\Phi_g(\mathbf{x}_T)) \leq \frac{\kappa_{\mathbf{y}}\ell\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2}{2} + \frac{\kappa_{\mathbf{y}}\ell\|\mathbf{x} - \mathbf{x}_T\|^2}{2} + \mu_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^{\star}(\mathbf{x}_T)\|^2$$

Since Φ_g is $2\kappa_{\mathbf{y}}\ell$-smooth and $\mu_{\mathbf{x}}$-strongly convex, we have

$$\begin{aligned}
(\mathbf{x} - \hat{\mathbf{x}})^{\top}(2\kappa_{\mathbf{y}}\ell(\hat{\mathbf{x}} - \mathbf{x}_T) + \nabla\Phi_g(\mathbf{x}_T)) &\leq 2\kappa_{\mathbf{y}}\ell(\mathbf{x} - \mathbf{x}_T)^{\top}(\hat{\mathbf{x}} - \mathbf{x}_T) + \Phi_g(\mathbf{x}) - \Phi_g(\hat{\mathbf{x}}) \\
&\quad -\kappa_{\mathbf{y}}\ell\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2 - \frac{\mu_{\mathbf{x}}\|\mathbf{x} - \mathbf{x}_T\|^2}{2}
\end{aligned}$$

Using the Young's inequality, we have $(\mathbf{x} - \mathbf{x}_T)^{\top}(\hat{\mathbf{x}} - \mathbf{x}_T) \leq \|\mathbf{x} - \mathbf{x}_T\|^2 + (1/4)\|\hat{\mathbf{x}} - \mathbf{x}_T\|^2$. Putting these pieces together yields with $\mathbf{x} = \mathbf{x}_g^{\star}$ yields that

$$\Phi_g(\hat{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}} \Phi_g(\mathbf{x}) \leq 3\kappa_{\mathbf{y}}\ell\|\mathbf{x}_T - \mathbf{x}_g^{\star}\|^2 + \mu_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^{\star}(\mathbf{x}_T)\|^2 \tag{24}$$

In what follows, we prove that $\Phi_g(\hat{\mathbf{x}}) \leq \min_{\mathbf{x} \in \mathcal{X}} \Phi_g(\mathbf{x}) + \epsilon$ if the following stopping conditions hold true,

$$g(\mathbf{x}_T, \mathbf{y}_T) - g(\mathbf{x}_g^{\star}(\mathbf{y}_T), \mathbf{y}_T) \leq \frac{\epsilon}{648\kappa_{\mathbf{x}}^3\kappa_{\mathbf{y}}^3} \tag{25}$$

$$\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T, \mathbf{y}_T))\| \leq \frac{1}{24\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}}\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \tag{26}$$

Indeed, we observe that $\|\mathbf{x}_T - \mathbf{x}_g^{\star}\| \leq \|\mathbf{x}_T - \mathbf{x}_g^{\star}(\mathbf{y}_T)\| + \|\mathbf{x}_g^{\star}(\mathbf{y}_T) - \mathbf{x}_g^{\star}(\mathbf{y}_g^{\star})\| + \|\mathbf{x}_g^{\star}(\mathbf{y}_g^{\star}) - \mathbf{x}_g^{\star}\|$. By definition, we have $\mathbf{x}_g^{\star}(\mathbf{y}_g^{\star}) = \mathbf{x}_g^{\star}$. Also, $\mathbf{x}_g^{\star}(\cdot)$ is $\kappa_{\mathbf{x}}$-Lipschitz. Therefore, we have

$$\|\mathbf{x}_T - \mathbf{x}_g^{\star}\| \leq \|\mathbf{x}_T - \mathbf{x}_g^{\star}(\mathbf{y}_T)\| + \kappa_{\mathbf{x}}\|\mathbf{y}_T - \mathbf{y}_g^{\star}\| \tag{27}$$

19

By the similar argument, we have

$$\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\| \ \leq \ \|\mathbf{y}_T - \mathbf{y}_g^\star\| + \kappa_{\mathbf{y}}\|\mathbf{x}_T - \mathbf{x}_g^\star\| \ \leq \ \kappa_{\mathbf{y}}\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| + \kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\|\mathbf{y}_T - \mathbf{y}_g^\star\| \tag{28}$$

First, we bound the term $\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\|$. Since $g(\cdot, \mathbf{y}_T)$ is $\mu_{\mathbf{x}}$-strongly convex, we have

$$\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| \ \leq \ \sqrt{\frac{2(g(\mathbf{x}_T, \mathbf{y}_T) - g(\mathbf{x}^\star(\mathbf{y}_T), \mathbf{y}_T))}{\mu_{\mathbf{x}}}} \ \leq \ \frac{1}{18\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \tag{29}$$

It remains to bound the term $\|\mathbf{y}_T - \mathbf{y}_g^\star\|$. Indeed, we have $\nabla\Psi_g(\mathbf{y}_T) = \nabla_{\mathbf{y}}g(\mathbf{x}_g^\star(\mathbf{y}_T), \mathbf{y}_T)$ and

$$\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| \ \leq \ \|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T, \mathbf{y}_T))\|$$
$$+\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T, \mathbf{y}_T)) - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\|$$

Since $\mathcal{P}_{\mathcal{Y}}$ is nonexpansive and $\nabla_{\mathbf{y}}g$ is ℓ-Lipschitz, we have

$$\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla_{\mathbf{y}}g(\mathbf{x}_T, \mathbf{y}_T)) - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| \ \leq \ \frac{\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\|}{2\kappa_{\mathbf{x}}}$$

Putting these pieces together with Eq. (26) and Eq. (29) yields that

$$\|\mathbf{y}_T - \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))\| \ \leq \ \frac{1}{18\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}}\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \tag{30}$$

Since $\mathbf{y}_g^\star = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}}\Psi_g(\mathbf{y})$ and $\tilde{\mathbf{y}}_T = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_T + (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T))$ is achieved by one-step projected gradient ascent from $\mathbf{y}_T$, we derive from the $2\kappa_{\mathbf{x}}\ell$-smoothness of Ψ_g, we have

$$\|\tilde{\mathbf{y}}_T - \mathbf{y}_g^\star\| \ \leq \ \|\mathbf{y}_T - \mathbf{y}_g^\star\| \tag{31}$$

Using the definition of $\tilde{\mathbf{y}}_T$ and $\mathbf{y}_g^\star$, we have

$$(\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top(\tilde{\mathbf{y}}_T - \mathbf{y}_T - (1/2\kappa_{\mathbf{x}}\ell)\nabla\Psi_g(\mathbf{y}_T)) \ \geq \ 0, \qquad (\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top\nabla\Psi_g(\mathbf{y}_g^\star) \ \geq \ 0$$

Summing up the above two inequalities and rearranging yields that

$$(\mathbf{y}_g^\star - \mathbf{y}_T)^\top(\nabla\Psi_g(\mathbf{y}_g^\star) - \nabla\Psi_g(\mathbf{y}_T)) \ \geq \ 2\kappa_{\mathbf{x}}\ell(\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T)^\top(\mathbf{y}_T - \tilde{\mathbf{y}}_T) + (\tilde{\mathbf{y}}_T - \mathbf{y}_T)^\top(\nabla\Psi_g(\mathbf{y}_g^\star) - \nabla\Psi_g(\mathbf{y}_T))$$

Since Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth and $\mu_{\mathbf{y}}$-strongly concave, we have

$$-\mu_{\mathbf{y}}\|\mathbf{y}_g^\star - \mathbf{y}_T\|^2 \ \geq \ -2\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}_T - \mathbf{y}_T\|(\|\mathbf{y}_g^\star - \tilde{\mathbf{y}}_T\| + \|\mathbf{y}_g^\star - \mathbf{y}_T\|) \ \overset{(31)}{\geq} \ -4\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}}_T - \mathbf{y}_T\|\|\mathbf{y}_g^\star - \mathbf{y}_T\|$$

This implies that

$$\|\mathbf{y}_g^\star - \mathbf{y}_T\| \ \leq \ 4\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\|\mathbf{y}_T - \tilde{\mathbf{y}}_T\| \ \overset{(30)}{\leq} \ \frac{1}{4\kappa_{\mathbf{x}}}\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \tag{32}$$

Plugging Eq. (29) and Eq. (32) into Eq. (27) yields that

$$\|\mathbf{x}_T - \mathbf{x}_g^\star\| \ \leq \ \left(\frac{1}{18\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}} + \frac{1}{4}\right)\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \ \overset{\kappa_{\mathbf{x}},\kappa_{\mathbf{y}}\geq 1}{\leq} \ \frac{1}{2}\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}}$$

Plugging Eq. (29) and Eq. (32) into Eq. (28) yields that

$$\|\mathbf{y}_T - \mathbf{y}_g^\star(\mathbf{x}_T)\| \ \leq \ \left(\frac{1}{18\kappa_{\mathbf{x}}} + \frac{\kappa_{\mathbf{y}}}{4}\right)\sqrt{\frac{\epsilon}{\kappa_{\mathbf{y}}\ell}} \ \overset{\kappa_{\mathbf{x}},\kappa_{\mathbf{y}}\geq 1}{\leq} \ \frac{1}{2}\sqrt{\frac{\kappa_{\mathbf{y}}\epsilon}{\ell}}$$

Putting these pieces together Eq. (24) yields the desired result.

20

Part II. We first show that $\|\mathbf{y}_T - \mathbf{y}_g^\star\| \leq (1/216\kappa_\mathbf{x}^2\kappa_\mathbf{y})\sqrt{\epsilon/\kappa_\mathbf{y}\ell}$ and Eq. (25) are sufficient to guarantee Eq. (26). Indeed, we have $\mathbf{y}_g^\star = \mathcal{P}_\mathcal{Y}(\mathbf{y}_g^\star + (1/2\kappa_\mathbf{x}\ell)\nabla\Psi_g(\mathbf{y}_g^\star))$. This together with the triangle inequality and the nonexpansiveness of $\mathcal{P}_\mathcal{Y}$ yields

$$\|\mathbf{y}_T - \mathcal{P}_\mathcal{Y}(\mathbf{y}_T + (1/2\kappa_\mathbf{x}\ell)\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T))\| \leq 2\|\mathbf{y}_T - \mathbf{y}_g^\star\| + \frac{\|\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T) - \nabla\Psi_g(\mathbf{y}_g^\star)\|}{2\kappa_\mathbf{x}\ell}$$

Furthermore, $\nabla\Psi_g(\mathbf{y}_T) = \nabla_\mathbf{y}g(\mathbf{x}^\star(\mathbf{y}_T),\mathbf{y}_T)$ and

$$\|\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T) - \nabla\Psi_g(\mathbf{y}_g^\star)\| \leq \|\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T) - \nabla_\mathbf{y}g(\mathbf{x}_g^\star(\mathbf{y}_T),\mathbf{y}_T)\| + \|\nabla\Psi_g(\mathbf{y}_T) - \nabla\Psi_g(\mathbf{y}_g^\star)\|$$

Since g is ℓ-smooth and Ψ_g is $2\kappa_\mathbf{x}\ell$-smooth, we have

$$\|\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T) - \nabla\Psi_g(\mathbf{y}_g^\star)\| \leq \ell\|\mathbf{x}_T - \mathbf{x}_g^\star(\mathbf{y}_T)\| + 2\kappa_\mathbf{x}\ell\|\mathbf{y}_T - \mathbf{y}_g^\star\|$$

Also, Eq. (25) guarantees that Eq. (29) holds true. Then we have

$$\|\mathbf{y}_T - \mathcal{P}_\mathcal{Y}(\mathbf{y}_T + (1/2\kappa_\mathbf{x}\ell)\nabla_\mathbf{y}g(\mathbf{x}_T,\mathbf{y}_T))\| \leq 3\|\mathbf{y}_T - \mathbf{y}_g^\star\| + \frac{1}{36\kappa_\mathbf{x}^2\kappa_\mathbf{y}}\sqrt{\frac{\epsilon}{\kappa_\mathbf{y}\ell}}$$

The above inequality together with $\|\mathbf{y}_T - \mathbf{y}_g^\star\| \leq (1/216\kappa_\mathbf{x}^2\kappa_\mathbf{y})\sqrt{\epsilon/\kappa_\mathbf{y}\ell}$ guarantees Eq. (26). Next we derive a sufficient condition for guaranteeing $\|\mathbf{y}_T - \mathbf{y}_g^\star\| \leq (1/216\kappa_\mathbf{x}^2\kappa_\mathbf{y})\sqrt{\epsilon/\kappa_\mathbf{y}\ell}$. Since Ψ_g is $\mu_\mathbf{y}$-strongly concave, [Nesterov(2018), Theorem 2.1.5] implies that

$$\|\mathbf{y}_T - \mathbf{y}_g^\star\|^2 \leq \frac{2}{\mu_\mathbf{y}}\left(\max_{\mathbf{y}\in\mathcal{Y}}\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}_T)\right)$$

Putting these pieces together yields the desired condition as follows,

$$\max_{\mathbf{y}\in\mathcal{Y}} \Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}_T) \leq \frac{\epsilon}{93312\kappa_\mathbf{x}^4\kappa_\mathbf{y}^4} \tag{33}$$

Part III. We proceed to estimate an upper bound for the gradient complexity of Algorithm 3 using Eq. (33). Note that $\tilde{\epsilon} \leq \epsilon/(4477676(\kappa_\mathbf{x}\kappa_\mathbf{y})^{11/2})$ and we provide a key technical lemma which is crucial to the subsequent analysis.

Lemma 3. *For any $\mathbf{y} \in \mathcal{Y}$ and $\{(\mathbf{y}_t, \tilde{\mathbf{y}}_t)\}_{t\geq 0}$ generated by Algorithm 3, we have*

$$\Psi_g(\mathbf{y}) \leq 2\kappa_\mathbf{x}\ell(\mathbf{y}-\tilde{\mathbf{y}}_{t-1})^\top(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) + \Psi_g(\mathbf{y}_t) - \frac{\kappa_\mathbf{x}\ell\|\mathbf{y}_t-\tilde{\mathbf{y}}_{t-1}\|^2}{2} - \frac{\mu_\mathbf{y}\|\mathbf{y}-\tilde{\mathbf{y}}_{t-1}\|^2}{4} + 3\kappa_\mathbf{x}\kappa_\mathbf{y}\tilde{\epsilon}$$

Proof. For any $\mathbf{y} \in \mathcal{Y}$, the update formula $\mathbf{y}_t \leftarrow \mathcal{P}_\mathcal{Y}(\tilde{\mathbf{y}}_{t-1} + (1/2\kappa_\mathbf{x}\ell)\nabla_\mathbf{y}g(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1}))$ implies that

$$\begin{aligned}0 &\leq (\mathbf{y}-\mathbf{y}_t)^\top(2\kappa_\mathbf{x}\ell(\mathbf{y}_t-\tilde{\mathbf{y}}_{t-1}) - \nabla_\mathbf{y}g(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})) \\ &= (\mathbf{y}-\mathbf{y}_t)^\top(2\kappa_\mathbf{x}\ell(\mathbf{y}_t-\tilde{\mathbf{y}}_{t-1}) - \nabla\Psi_g(\tilde{\mathbf{y}}_{t-1})) + (\mathbf{y}-\mathbf{y}_t)^\top(\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_\mathbf{y}g(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1}))\end{aligned}$$

Since $\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) = \nabla_\mathbf{y}g(\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}),\tilde{\mathbf{y}}_{t-1})$, we have

$$\|\nabla\Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_\mathbf{y}g(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})\| \leq \ell\|\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}) - \tilde{\mathbf{x}}_{t-1}\|$$

Since $g(\cdot,\tilde{\mathbf{y}}_{t-1})$ is $\mu_\mathbf{x}$-strongly convex, we have

$$\|\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}) - \tilde{\mathbf{x}}_{t-1}\| \leq \sqrt{\frac{2(g(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1}) - g(\mathbf{x}_g^\star(\tilde{\mathbf{y}}_{t-1}),\tilde{\mathbf{y}}_{t-1}))}{\mu_\mathbf{x}}} \leq \sqrt{\frac{2\tilde{\epsilon}}{\mu_\mathbf{x}}}$$

21

Using Young's inequality, we have

$$(\mathbf{y} - \mathbf{y}_t)^\top (\nabla \Psi_g(\tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{y}} g(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})) \leq \frac{\kappa_{\mathbf{x}} \ell \|\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}\|^2}{2} + \frac{\mu_{\mathbf{y}} \|\mathbf{y} - \tilde{\mathbf{y}}_{t-1}\|^2}{4} + 3 \kappa_{\mathbf{x}} \kappa_{\mathbf{y}} \tilde{\epsilon}$$

Since Ψ_g is $2\kappa_{\mathbf{x}}\ell$-smooth and $\mu_{\mathbf{y}}$-strongly concave, we have

$$(\mathbf{y} - \mathbf{y}_t)^\top (2\kappa_{\mathbf{x}}\ell(\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) - \nabla \Psi_g(\tilde{\mathbf{y}}_{t-1})) \leq 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}}_{t-1})^\top (\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}) + \Psi_g(\mathbf{y}_t) - \Psi_g(\mathbf{y})$$
$$- \kappa_{\mathbf{x}}\ell \|\mathbf{y}_t - \tilde{\mathbf{y}}_{t-1}\|^2 - \frac{\mu_{\mathbf{y}} \|\mathbf{y} - \tilde{\mathbf{y}}_{t-1}\|^2}{2}$$

Putting these pieces together yields the desired inequality. $\qquad\square$

The remaining proof is based on the modification of Nesterov's techniques [Nesterov(2018), Section 2.2.5]. Indeed, we define the estimate sequence as follows,

$$\Gamma_0(\mathbf{y}) = \Psi_g(\mathbf{y}_0) - \frac{\mu_{\mathbf{y}} \|\mathbf{y} - \mathbf{y}_0\|^2}{2}$$

$$\Gamma_{t+1}(\mathbf{y}) = \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(\Psi_g(\mathbf{y}_{t+1}) + 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}}_t)^\top (\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t) - \frac{\kappa_{\mathbf{x}}\ell \|\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t\|^2}{2} \right.$$
$$\left. - \frac{\mu_{\mathbf{y}} \|\mathbf{y} - \tilde{\mathbf{y}}_t\|^2}{4} - 12(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon} \right) + \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \Gamma_t(\mathbf{y}) \quad \text{for all } t \geq 0$$

We apply the inductive argument to prove,

$$\max_{\mathbf{y} \in \mathbb{R}^n} \Gamma_t(\mathbf{y}) \leq \Psi_g(\mathbf{y}_t) \quad \text{for all } t \geq 0 \tag{34}$$

Eq. (34) holds trivially when $t = 0$. In what follows, we show that Eq. (34) holds true when $t = T$ if Eq. (34) holds true for all $t \leq T - 1$. Let $\mathbf{v}_t = \text{argmax}_{\mathbf{y} \in \mathbb{R}^n} \Gamma_t(\mathbf{y})$ and $\Gamma_t^\star = \max_{\mathbf{y} \in \mathbb{R}^n} \Gamma_t(\mathbf{y})$, we have the canonical form $\Gamma_t(\mathbf{y}) = \Gamma_t^\star - (\mu_{\mathbf{y}}/4)\|\mathbf{y} - \mathbf{v}_t\|^2$. The following recursive rules hold for $\mathbf{v}_t$ and $\Gamma_t^\star$:

$$\mathbf{v}_{t+1} = \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \mathbf{v}_t + \frac{\tilde{\mathbf{y}}_t}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} + \sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}(\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t)$$

$$\Gamma_{t+1}^\star = \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \Gamma_t^\star + \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(\Psi_g(\mathbf{y}_{t+1}) - 12(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon} \right) - \left(\frac{\ell}{8}\sqrt{\frac{\kappa_{\mathbf{x}}}{\kappa_{\mathbf{y}}}} - \frac{\kappa_{\mathbf{x}}\ell}{4} \right) \|\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t\|^2$$
$$- \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \left(\frac{\mu_{\mathbf{y}} \|\tilde{\mathbf{y}}_t - \mathbf{v}_t\|^2}{4} - 2\kappa_{\mathbf{x}}\ell(\mathbf{v}_t - \tilde{\mathbf{y}}_t)^\top (\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t) \right)$$

It follows from the recursive rule for Γ_t and its canonical form that

$$\nabla \Gamma_{t+1}(\mathbf{y}) = -\left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \frac{\mu_{\mathbf{y}}(\mathbf{y} - \mathbf{v}_t)}{2} + \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(2\kappa_{\mathbf{x}}\ell(\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t) - \frac{\mu_{\mathbf{y}}(\mathbf{y} - \tilde{\mathbf{y}}_t)}{2} \right)$$

The recursive rule for $\mathbf{v}_t$ can be achieved by solving $\nabla \Gamma_{t+1}(\mathbf{v}_{t+1}) = 0$. Then we have

$$\Gamma_{t+1}^\star = \Gamma_{t+1}(\mathbf{v}_{t+1})$$
$$= \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \Gamma_t^\star - \left(1 - \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \right) \frac{\mu_{\mathbf{y}} \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2}{4} + \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(\Psi_g(\mathbf{y}_{t+1}) - 12(\kappa_{\mathbf{x}}\kappa_{\mathbf{y}})^{3/2}\tilde{\epsilon} \right.$$
$$\left. - \frac{\kappa_{\mathbf{x}}\ell \|\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t\|^2}{2} \right) + \frac{1}{4\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}}} \left(2\kappa_{\mathbf{x}}\ell(\mathbf{v}_{t+1} - \tilde{\mathbf{y}}_t)^\top (\mathbf{y}_{t+1} - \tilde{\mathbf{y}}_t) - \frac{\mu_{\mathbf{y}} \|\mathbf{v}_{t+1} - \tilde{\mathbf{y}}_t\|^2}{4} \right)$$

Then we conclude the recursive rule for $\Gamma_t^\star$ by plugging the recursive rule for $\mathbf{v}_k$ into the above equality. By the induction, Eq. (34) holds true when $t = T - 1$ which implies

$$
\begin{aligned}
\Gamma_T^\star \;\leq\; & \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\Psi_g(\mathbf{y}_{T-1}) + \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(\Psi_g(\mathbf{y}_T) - 12(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}\right) \\
& - \left(\frac{\ell}{8}\sqrt{\frac{\kappa_\mathbf{x}}{\kappa_\mathbf{y}}} - \frac{\kappa_\mathbf{x}\ell}{4}\right)\|\mathbf{y}_T - \tilde{\mathbf{y}}_{T-1}\|^2 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\left(\frac{\mu_\mathbf{y}\|\tilde{\mathbf{y}}_{T-1} - \mathbf{v}_{T-1}\|^2}{2}\right. \\
& \left. -2\kappa_\mathbf{x}\ell(\mathbf{v}_{T-1} - \tilde{\mathbf{y}}_{T-1})^\top(\mathbf{y}_T - \tilde{\mathbf{y}}_{T-1})\right)
\end{aligned}
$$

Applying Lemma 3 with $t = T$ and $\mathbf{y} = \mathbf{y}_{T-1}$ further implies that

$$
\begin{aligned}
\Psi_g(\mathbf{y}_{T-1}) \;\leq\; & 2\kappa_\mathbf{x}\ell(\mathbf{y}_{T-1} - \tilde{\mathbf{y}}_{T-1})^\top(\mathbf{y}_T - \tilde{\mathbf{y}}_{T-1}) + \Psi(\mathbf{y}_T) - \frac{\kappa_\mathbf{x}\ell\|\mathbf{y}_T - \tilde{\mathbf{y}}_{T-1}\|^2}{2} \\
& - \frac{\mu_\mathbf{y}\|\mathbf{y}_{T-1} - \tilde{\mathbf{y}}_{T-1}\|^2}{2} + 3\kappa_\mathbf{x}\kappa_\mathbf{y}\tilde{\epsilon}
\end{aligned}
$$

Putting these pieces together yields that

$$
\Gamma_T^\star \;\leq\; \Psi_g(\mathbf{y}_T) + \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)2\kappa_\mathbf{x}\ell(\mathbf{y}_T - \tilde{\mathbf{y}}_{T-1})^\top\left[(\mathbf{y}_{T-1} - \tilde{\mathbf{y}}_{T-1}) + \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}(\mathbf{v}_{T-1} - \tilde{\mathbf{y}}_{T-1})\right]
$$

Using the update formula $\tilde{\mathbf{y}}_t = \mathbf{y}_t + \frac{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}-1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}+1}(\mathbf{y}_t - \mathbf{y}_{t-1})$ and the recursive rule for $\mathbf{v}_t$ with the inductive argument, it is straightforward that $(\mathbf{y}_t - \tilde{\mathbf{y}}_t) + \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}(\mathbf{v}_t - \tilde{\mathbf{y}}_t) = 0$ for all $t \geq 0$. This implies that $\Gamma_T^\star \leq \Psi_g(\mathbf{y}_T)$. Therefore, we conclude that Eq. (34) holds true for all $t \geq 0$.

On the other hand, Lemma 3 and the update formula for Γ_t implies that

$$
\Gamma_{t+1}(\mathbf{y}) \;\geq\; \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\left(\Psi_g(\mathbf{y}) - 12(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon} - 3\kappa_\mathbf{x}\kappa_\mathbf{y}\tilde{\epsilon}\right) + \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)\Gamma_t(\mathbf{y})
$$

Since $\kappa_\mathbf{x}, \kappa_\mathbf{y} \geq 1$, we have

$$
\Psi_g(\mathbf{y}) - \Gamma_{t+1}(\mathbf{y}) \;\leq\; \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)(\Psi_g(\mathbf{y}) - \Gamma_t(\mathbf{y})) + 6\kappa_\mathbf{x}\kappa_\mathbf{y}\tilde{\epsilon}
$$

Repeating the above inequality yields that

$$
\Psi_g(\mathbf{y}) - \Gamma_T(\mathbf{y}) \;\leq\; \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)^T(\Psi_g(\mathbf{y}) - \Gamma_0(\mathbf{y})) + 24(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}
$$

Therefore, we conclude that

$$
\max_{\mathbf{y}\in\mathcal{Y}}\Psi_g(\mathbf{y}) - \Psi_g(\mathbf{y}_T) \;\leq\; \left(1 - \frac{1}{4\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}}\right)^T\frac{(2\kappa_\mathbf{x}\ell + \bar{\mu})D_\mathbf{y}^2}{2} + 24(\kappa_\mathbf{x}\kappa_\mathbf{y})^{3/2}\tilde{\epsilon}
$$

Since the tolerance $\tilde{\epsilon} \leq \frac{\epsilon}{4477676(\kappa_\mathbf{x}\kappa_\mathbf{y})^{11/2}}$, we conclude that the iteration complexity Algorithm 3 to guarantee Eq. (33) is bounded by $O(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}\log(\ell D_\mathbf{y}^2/\epsilon))$.

Now it suffices to establish the gradient complexity of the two AGD subroutines at each iteration. In particular, we use the gradient complexity of the AGD subroutine to guarantee that $g(\hat{\mathbf{x}}) \leq \min_{\mathcal{X}} g(\mathbf{x}) + \epsilon$ is bounded by

$$
O\left(1 + \sqrt{\kappa}\log\left(\frac{\kappa^3\ell\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\epsilon}\right)\right)
$$

23

where κ is the condition number of g and $\mathbf{x}^\star$ is the global optimum of g over $\mathcal{X}$. Since $\mathcal{Y}$ is a convex and bounded set, $\{\mathbf{y}_t\}_{t\geq 0}$ is a bounded sequence. Hence $\{\tilde{\mathbf{y}}_t\}_{t\geq 0}$ is also a bounded sequence. Since $\mathbf{x}_g^\star(\cdot)$ is $\kappa_\mathbf{x}$-Lipschitz (cf. Lemma 2), the sequences $\{\mathbf{x}_g^\star(\tilde{\mathbf{y}}_t)\}_{t\geq 0}$ and $\{\mathbf{x}_g^\star(\mathbf{y}_t)\}_{t\geq 0}$ are bounded. Thus, we have

$$\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_t)\|^2 = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\tilde{\mathbf{y}}_t)\|^2 = O(\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|^2 + \kappa_\mathbf{x}^2 D_\mathbf{y}^2)$$

Putting these pieces together yields that the gradient complexity of every AGD subroutines at each iteration is bounded by $O(\sqrt{\kappa_\mathbf{x}} \log((\kappa_\mathbf{x}^3 \ell(\|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|^2 + \kappa_\mathbf{x}^2 D_\mathbf{y}^2)/\tilde{\epsilon}))$. Therefore, the gradient complexity of Algorithm 3 to guarantee Eq. (33) is bounded by

$$O\left(\kappa_\mathbf{x}\sqrt{\kappa_\mathbf{y}} \cdot \log^2\left(\frac{(\kappa_\mathbf{x} + \kappa_\mathbf{y})\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right)$$

where $\kappa_\mathbf{x} = \ell/\mu_\mathbf{x}$ and $\kappa_\mathbf{y} = \ell/\mu_\mathbf{y}$ are condition numbers, $\tilde{D}_\mathbf{x} = \|\mathbf{x}_0 - \mathbf{x}_g^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$ and $D_\mathbf{y} > 0$ is the diameter of the constraint set $\mathcal{Y}$.

6   Proofs for Convex-Concave Settings

In this section, we present proofs for all results in Section 4.

6.1   Proof of Theorem 4

We first show that there exists $T > 0$ such that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{Minimax-APPA}(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{x}, \mu_\mathbf{y}, \epsilon, T)$ is an ϵ-saddle point. Then we estimate the total number of gradient evaluations required to output an ϵ-approximate saddle point.

First, we note that Minimax-APPA in Algorithm 4 can be interpreted as an inexact accelerated proximal point algorithm Inexact-APPA with the inner loop solver Maximin-AG2 and AGD. Using Theorem 1 and Theorem 2, the point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \ \leq\ \left(1 - \frac{1}{6\sqrt{\kappa_\mathbf{x}}}\right)^T \left(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}^\star) + \frac{\mu_\mathbf{x}\|\mathbf{x}^\star - \mathbf{x}_0\|^2}{4}\right) + 42\kappa_\mathbf{x}^{3/2}\delta$$

and $\hat{\mathbf{y}} \leftarrow \mathcal{P}_\mathcal{Y}\left(\tilde{\mathbf{y}} + (1/2\kappa_\mathbf{x}\ell)\nabla_\mathbf{y} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\right)$ where $\tilde{\mathbf{y}} \in \mathcal{Y}$ satisfies that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \tilde{\mathbf{y}}) \ \leq\ \tilde{\epsilon}$$

We let $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ and note that Φ is $\mu_\mathbf{x}$-strongly convex function. Since f is $\mu_\mathbf{x}$-strongly-convex-$\mu_\mathbf{y}$-strongly-concave, the Nash equilibrium $(\mathbf{x}^\star, \mathbf{y}^\star)$ is unique and $\mathbf{x}^\star = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$. Therefore, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2 \ \leq\ \frac{2}{\mu_\mathbf{x}}\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})\right)$$

Since $f(\hat{\mathbf{x}}, \cdot)$ is $\mu_\mathbf{y}$-strongly concave, [Nesterov(2018), Theorem 2.1.5] implies that

$$\|\tilde{\mathbf{y}} - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 \ \leq\ \frac{2}{\mu_\mathbf{y}}\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\right) \ \leq\ \frac{2\tilde{\epsilon}}{\mu_\mathbf{y}}$$

24

Since $\mathbf{y}^\star(\cdot) = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is $\kappa_{\mathbf{y}}$-Lipschitz (cf. Lemma 2), $\|\mathbf{y}^\star - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 = \|\mathbf{y}^\star(\mathbf{x}^\star) - \mathbf{y}^\star(\hat{\mathbf{x}})\|^2 \leq \kappa_{\mathbf{y}}^2 \|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2$. Thus, we have

$$\|\tilde{\mathbf{y}} - \mathbf{y}^\star\|^2 \leq 2\kappa_{\mathbf{y}}^2 \|\hat{\mathbf{x}} - \mathbf{x}^\star\|^2 + \frac{4\tilde{\epsilon}}{\mu_{\mathbf{y}}}$$

Let $\Psi(\cdot) = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \cdot)$. By the definition of $\hat{\mathbf{y}}$, the following inequality holds true for any $\mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned}
0 &\leq (\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})) \\
&= (\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla\Psi(\tilde{\mathbf{y}})) + (\mathbf{y} - \hat{\mathbf{y}})^\top (\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}}))
\end{aligned}$$

Since $\nabla\Psi(\tilde{\mathbf{y}}) = \nabla_{\mathbf{y}} f(\mathbf{x}^\star(\tilde{\mathbf{y}}), \tilde{\mathbf{y}})$, we have $\|\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})\| \leq \ell\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \hat{\mathbf{x}}\|$. Using the Young's inequality, we have

$$(\mathbf{y} - \hat{\mathbf{y}})^\top (\nabla\Psi(\tilde{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \tilde{\mathbf{y}})) \leq \frac{\kappa_{\mathbf{x}}\ell\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{2} + \frac{\kappa_{\mathbf{x}}\ell\|\mathbf{y} - \tilde{\mathbf{y}}\|^2}{2} + \mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \hat{\mathbf{x}}\|^2$$

Since Ψ is $\mu_{\mathbf{y}}$-strongly concave and $2\kappa_{\mathbf{x}}\ell$-smooth, we have

$$\begin{aligned}
(\mathbf{y} - \hat{\mathbf{y}})^\top (2\kappa_{\mathbf{x}}\ell(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) - \nabla\Psi(\tilde{\mathbf{y}})) &\leq 2\kappa_{\mathbf{x}}\ell(\mathbf{y} - \tilde{\mathbf{y}})^\top(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) + \Psi(\hat{\mathbf{y}}) - \Psi(\mathbf{y}) \\
&\quad - \kappa_{\mathbf{x}}\ell\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 - \frac{\mu_{\mathbf{x}}\|\mathbf{y} - \tilde{\mathbf{y}}\|^2}{2}
\end{aligned}$$

Using the Young's inequality, we have $(\mathbf{y} - \tilde{\mathbf{y}})^\top(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \leq \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 + (1/4)\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2$. Putting these pieces together with $\mathbf{y} = \mathbf{y}^\star$ yields that

$$\begin{aligned}
\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) &= \Psi(\mathbf{y}^\star) - \Psi(\hat{\mathbf{y}}) \leq 3\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}} - \mathbf{y}^\star\|^2 + \mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \hat{\mathbf{x}}\|^2 \\
&\leq 3\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}} - \mathbf{y}^\star\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star(\tilde{\mathbf{y}}) - \mathbf{x}^\star(\mathbf{y}^\star)\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star - \hat{\mathbf{x}}\|^2 \\
&\leq 5\kappa_{\mathbf{x}}\ell\|\tilde{\mathbf{y}} - \mathbf{y}^\star\|^2 + 2\mu_{\mathbf{x}}\|\mathbf{x}^\star - \hat{\mathbf{x}}\|^2
\end{aligned}$$

Therefore, we conclude that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq 20\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}\tilde{\epsilon} + (20\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2 + 5)\left(\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})\right)$$

Note that $\tilde{\epsilon} \leq \epsilon/80\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}$ and $\delta \leq \epsilon/4200\kappa_{\mathbf{x}}^{7/2}\kappa_{\mathbf{y}}^2$. This together with the above inequality implies that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \frac{3\epsilon}{4} + (20\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2 + 5)\left(1 - \frac{1}{6\sqrt{\kappa_{\mathbf{x}}}}\right)^T \left(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}^\star) + \frac{\mu_{\mathbf{x}}\|\mathbf{x}^\star - \mathbf{x}_0\|^2}{4}\right)$$

To this end, there exists an absolute constant $c > 0$ such that $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$ if the maximum number of iterations $T \geq c\sqrt{\kappa_{\mathbf{x}}} \log(\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2\ell\|\mathbf{x}^\star - \mathbf{x}_0\|^2/\epsilon)$. This implies that the total number of iterations is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{x}}} \log\left(\frac{\kappa_{\mathbf{x}}^2\kappa_{\mathbf{y}}^2\ell\|\mathbf{x}^\star - \mathbf{x}_0\|^2}{\epsilon}\right)\right)$$

Furthermore, we call the solver MAXIMIN-AG2 at each iteration. Using Theorem 3 and $\delta = \epsilon/(10\kappa_\mathbf{x}\kappa_\mathbf{y})^4$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_\mathbf{y}}\log\left(\frac{\kappa_\mathbf{x}^{7/2}\kappa_\mathbf{y}^3\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\log\left(\frac{\kappa_\mathbf{x}^4\kappa_\mathbf{y}^4\ell D_\mathbf{y}^2}{\epsilon}\right)\right)$$

Recalling $D = \max\{D_\mathbf{x}, D_\mathbf{y}\} < +\infty$, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\kappa_\mathbf{x}\kappa_\mathbf{y}}\log^3\left(\frac{\kappa_\mathbf{x}\kappa_\mathbf{y}\ell D^2}{\epsilon}\right)\right)$$

This completes the proof.

6.2 Proof of Corollary 1

We first show that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_{\epsilon,\mathbf{y}}, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{x}, \epsilon/(4D_\mathbf{y}^2), \epsilon/2, T)$ is an ϵ-saddle point. Then we estimate the number of gradient evaluations to output an ϵ-saddle point using Theorem 4. By the definition of f_ϵ, the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}}\left\{f(\hat{\mathbf{x}}, \mathbf{y}) - \frac{\epsilon\|\mathbf{y} - \mathbf{y}_0\|^2}{4D_\mathbf{y}^2}\right\} - \min_{\mathbf{x}\in\mathcal{X}}\left\{f(\mathbf{x}, \hat{\mathbf{y}}) - \frac{\epsilon\|\hat{\mathbf{y}} - \mathbf{y}_0\|^2}{4D_\mathbf{y}^2}\right\} \leq \frac{\epsilon}{2}$$

Since the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$\max_{\mathbf{y}\in\mathcal{Y}}\left\{f(\hat{\mathbf{x}}, \mathbf{y}) - \frac{\epsilon\|\mathbf{y} - \mathbf{y}_0\|^2}{4D_\mathbf{y}^2}\right\} \geq \max_{\mathbf{y}\in\mathcal{Y}}f(\mathbf{x}_{T+1}, \mathbf{y}) - \frac{\epsilon}{4}$$

On the other hand, we have

$$\min_{\mathbf{x}\in\mathcal{X}}\left\{f(\mathbf{x}, \hat{\mathbf{y}}) - \frac{\epsilon\|\hat{\mathbf{y}} - \mathbf{y}_0\|^2}{4D_\mathbf{y}^2}\right\} \leq \min_{\mathbf{x}\in\mathcal{X}}f(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon}{4}$$

Putting these pieces together yields that $\max_{\mathbf{y}\in\mathcal{Y}}f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}}f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$.

Furthermore, letting $\kappa_\mathbf{y} = 2\ell D_\mathbf{y}^2/\epsilon$ in the gradient complexity bound presented in Theorem 4, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\sqrt{\frac{\kappa_\mathbf{x}\ell}{\epsilon}}D_\mathbf{y}\log^3\left(\frac{\kappa_\mathbf{x}\ell D^2}{\epsilon}\right)\right)$$

This completes the proof.

6.3 Proof of Corollary 2

We first show that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-APPA}(f_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \frac{\epsilon}{4D_\mathbf{x}^2}, \frac{\epsilon}{4D_\mathbf{y}^2}, \epsilon/2, T)$ is an ϵ-saddle point. Then we estimate the number of gradient evaluations to output an ϵ-saddle point using Theorem 4. By the definition of f_ϵ, the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies

$$\max_{\mathbf{y}\in\mathcal{Y}}\left\{f(\hat{\mathbf{x}}, \mathbf{y}) + \frac{\epsilon\|\hat{\mathbf{x}} - \mathbf{x}_0\|^2}{8D_\mathbf{x}^2} - \frac{\epsilon\|\mathbf{y} - \mathbf{y}_0\|^2}{8D_\mathbf{y}^2}\right\} - \min_{\mathbf{x}\in\mathcal{X}}\left\{f(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon\|\mathbf{x} - \mathbf{x}_0\|^2}{8D_\mathbf{x}^2} - \frac{\epsilon\|\hat{\mathbf{y}} - \mathbf{y}_0\|^2}{8D_\mathbf{y}^2}\right\} \leq \frac{\epsilon}{2}$$

26

Since the function $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$, we have

$$\max_{\mathbf{y} \in \mathcal{Y}} \left\{ f(\hat{\mathbf{x}}, \mathbf{y}) + \frac{\epsilon \|\hat{\mathbf{x}} - \mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \frac{\epsilon \|\mathbf{y} - \mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} \geq \max_{\mathbf{y} \in \mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \frac{\epsilon}{4}$$

On the other hand,

$$\min_{\mathbf{x} \in \mathcal{X}} \left\{ f(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon \|\mathbf{x} - \mathbf{x}_0\|^2}{8D_{\mathbf{x}}^2} - \frac{\epsilon \|\hat{\mathbf{y}} - \mathbf{y}_0\|^2}{8D_{\mathbf{y}}^2} \right\} \leq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon}{4}$$

Putting these pieces together yields that $\max_{\mathbf{y} \in \mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon$.

Furthermore, letting $\kappa_{\mathbf{x}} = 4\ell D_{\mathbf{x}}^2/\epsilon$ and $\kappa_{\mathbf{y}} = 2\ell D_{\mathbf{y}}^2/\epsilon$ in the gradient complexity bound presented in Theorem 4, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell D_{\mathbf{x}} D_{\mathbf{y}}}{\epsilon} \log^3 \left(\frac{\ell D^2}{\epsilon} \right) \right)$$

This completes the proof.

7   Conclusions

This paper has provided the first set of *near-optimal* algorithms for strongly-convex-(strongly)-concave minimax optimization problems and the state-of-the-art algorithms for nonconvex-(strongly)-concave minimax optimization problems. For the former class of problems, our algorithms match the lower complexity bound for first-order algorithms [Ouyang and Xu(2019), Ibrahim et al.(2019), Zhang et al.(2019)] up to logarithmic factors. For the latter class of problems, our algorithms achieve the best known upper bound. In the future research, one important direction is to investigate the lower complexity bound of first-order algorithms for nonconvex-(strongly)-concave minimax problems. Despite several striking results on lower complexity bounds for nonconvex smooth problems [Carmon et al.(2019a), Carmon et al.(2019b)], this problem remains challenging as solving it requires a new construction of "chain-style" functions and resisting oracles.

References

[Abernethy et al.(2019)] J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization. *ArXiv Preprint: 1906.02027*, 2019.

[Alkousa et al.(2019)] M. Alkousa, D. Dvinskikh, F. Stonyakin, and A. Gasnikov. Accelerated methods for composite non-bilinear saddle point problem. *ArXiv Preprint: 1906.03620*, 2019.

[Auslender and Teboulle(2005)] A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical programming*, 104(1):39–68, 2005.

[Azizian et al.(2020)] W. Azizian, D. Scieur, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. Accelerating smooth games by manipulating spectral shapes. *ArXiv Preprint: 2001.00602*, 2020.

[Basar and Olsder(1999)] T. Basar and G. J. Olsder. *Dynamic Noncooperative Game Theory*, volume 23. SIAM, 1999.

[Beck and Teboulle(2009)] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Science*, 2(1):183–202, 2009.

[Berger(2013)] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013.

[Brown(1951)] G. W. Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951.

[Carmon et al.(2019a)] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, Jun 2019a. ISSN 1436-4646. doi: 10.1007/s10107-019-01406-y. URL https://doi.org/10.1007/s10107-019-01406-y.

[Carmon et al.(2019b)] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, Sep 2019b. ISSN 1436-4646. doi: 10.1007/s10107-019-01431-x. URL https://doi.org/10.1007/s10107-019-01431-x.

[Cesa-Bianchi and Lugosi(2006)] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[Chambolle and Pock(2011)] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[Chambolle and Pock(2016)] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[Chavdarova et al.(2019)] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. *ArXiv Preprint: 1904.08598*, 2019.

[Chen et al.(2014)] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[Chen et al.(2017)] Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[Dantzig(1998)] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1998.

[Daskalakis et al.(2018)] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018.

[Davis and Drusvyatskiy(2019)] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[Esser et al.(2010)] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[Facchinei and Pang(2007)] F. Facchinei and J-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

[Gidel et al.(2019)] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

[Goldstein et al.(2014)] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[Goodfellow et al.(2014)] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

[Grnarova et al.(2018)] P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. In *ICLR*, 2018.

[Hamedani and Aybat(2018)] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *ArXiv Preprint: 1803.01401*, 2018.

[He and Monteiro(2016)] Y. He and R. D. C. Monteiro. An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016.

[Hsieh et al.(2019)] Y-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6936–6946, 2019.

[Ibrahim et al.(2019)] A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Lower bounds and conditioning of differentiable games. *ArXiv Preprint: 1906.07300*, 2019.

[Jin et al.(2019)] C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *ArXiv Preprint: 1902.00618*, 2019.

[Jordan(2018)] M. I. Jordan. Artificial intelligence–the revolution hasn't happened yet. *Medium. Vgl. Ders.(2018): Perspectives and Challenges. Presentation SysML*, 2018.

[Juditsky and Nemirovski(2011)] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9): 149–183, 2011.

[Kolossoski and Monteiro(2017)] O. Kolossoski and R. D. C. Monteiro. An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

[Kong and Monteiro(2019)] W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *ArXiv Preprint: 1905.13433*, 2019.

[Korpelevich(1976)] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Lan and Monteiro(2016)] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

[Liang and Stokes(2019)] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915, 2019.

[Lin et al.(2019)] T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *ArXiv Preprint: 1906.00331*, 2019.

[Lu et al.(2019)] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *ArXiv Preprint: 1902.08294*, 2019.

[Malitsky(2015)] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[Mateos et al.(2010)] G. Mateos, J. A. Bazerque, and G. B. Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010.

[Mertikopoulos et al.(2019)] P. Mertikopoulos, B. Lecouat, H. Zenati, C-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019.

[Mishchenko et al.(2019)] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. *ArXiv Preprint: 1905.11373*, 2019.

[Mokhtari et al.(2019a)] A. Mokhtari, A. Ozdaglar, and S. Pattathil. Proximal point approximations achieving a convergence rate of o(1/k) for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *ArXiv Preprint: 1906.01115*, 2019a.

[Mokhtari et al.(2019b)] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *ArXiv Preprint: 1901.08511*, 2019b.

[Monteiro and Svaiter(2010)] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

[Monteiro and Svaiter(2011)] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of tseng's modified fb splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[Murty and Kabadi(1987)] K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming: Series A and B*, 39(2):117–129, 1987.

[Namkoong and Duchi(2016)] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, pages 2208–2216, 2016.

[Nedić and Ozdaglar(2009)] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

[Nemirovski(2004)] A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[Nesterov(2005)] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[Nesterov(2007)] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[Nesterov(2013)] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Nesterov(2018)] Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

[Nesterov and Scrimali(2006)] Y. Nesterov and L. Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Available at SSRN 970903*, 2006.

[Nesterov(1983)] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o$(1/k^2)$. In *Dokl. Akad. Nauk Sssr*, volume 269, pages 543–547, 1983.

[Nisan et al.(2007)] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[Nouiehed et al.(2019)] 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. In *NeurIPS*, pages 14905–14916, 2019.

[Ostrovskii et al.(2020)] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *ArXiv Preprint: 2002.07919*, 2020.

[Ouyang and Xu(2019)] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, Aug 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01420-0. URL https://doi.org/10.1007/s10107-019-01420-0.

[Ouyang et al.(2015)] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

[Rafique et al.(2018)] H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *ArXiv Preprint: 1810.02060*, 2018.

[Rakhlin and Sridharan(2013a)] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *COLT*, pages 993–1019, 2013a.

[Rakhlin and Sridharan(2013b)] S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS*, pages 3066–3074, 2013b.

[Rockafellar(1970)] R. T. Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.

[Sanjabi et al.(2018)] M. Sanjabi, M. Razaviyayn, and J. D. Lee. Solving non-convex non-concave min-max games under polyak-lojasiewicz condition. *ArXiv Preprint: 1812.02878*, 2018.

[Shamma(2008)] J. Shamma. *Cooperative Control of Distributed Multi-agent Systems*. John Wiley & Sons, 2008.

[Sinha et al.(2018)] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018.

[Sion(1958)] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[Thekumparampil et al.(2019)] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pages 12659–12670, 2019.

[Tseng(1995)] P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[Tseng(2008)] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2:3, 2008.

[Von Neumann and Morgenstern(2007)] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition)*. Princeton University Press, 2007.

[Xie and Shi(2019)] Z. Xie and J. Shi. Accelerated primal dual method for a class of saddle point problem with strongly convex component. *ArXiv Preprint: 1906.07691*, 2019.

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

[Xu(2019)] Y. Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *Mathematical Programming*, Aug 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01425-9. URL https://doi.org/10.1007/s10107-019-01425-9.

[Xu and Zhang(2018)] Y. Xu and S. Zhang. Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1): 91–128, 2018.

[Yadav et al.(2018)] A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR*, 2018.

[Zhang et al.(2019)] J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the saddle point problems. *ArXiv Preprint: 1912.07481*, 2019.

[Zhao(2019)] R. Zhao. Optimal algorithms for stochastic three-composite convex-concave saddle point problems. *ArXiv Preprint: 1903.01687*, 2019.

[Zhao(2020)] R. Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *ArXiv Preprint: 2003.04375*, 2020.

A Accelerating Nonconvex-Concave Optimization

In this section, we present methods for accelerating nonconvex-concave optimization. Similar to Section 4, we first present our algorithm and its theoretical guarantee for optimizing nonconvex-strongly-concave functions. We then use a simple reduction argument to obtain results for nonconvex-concave functions. In the first two subsections we present results using the stationarity of the function f (Definition 5) as an optimality measure. In the rest three subsections, we present our results for nonconvex-concave optimization using stationarity of the function $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ (Definition 6 and Definition 9) as the optimality measure.

Algorithm 5 MINIMAX-PPA($g, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{y}, \epsilon, T$)

1: **Input:** initial point $\mathbf{x}_0, \mathbf{y}_0$, proximity ℓ, strongly-convex parameter μ, tolerance δ, iteration T
2: **Initialize:** set $\delta \leftarrow \frac{\epsilon^2}{(10\kappa_\mathbf{y})^4\ell} \cdot (\frac{\epsilon}{\ell D_\mathbf{y}})^2$
3: **for** $t = 1, \cdots, T$ **do**
4:     denote $g_t(\cdot, \cdot)$ where $g_t(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \mathbf{x}_{t-1}\|^2$
5:     $\mathbf{x}_t \leftarrow$ MAXIMIN-AG2($g_t, \mathbf{x}_0, \mathbf{y}_0, 3\ell, \ell, \mu, \delta$)
6: **end for**
7: sample s uniformly from $\{1, 2, \cdots, T\}$
8: $\mathbf{y}_s \leftarrow$ AGD($-f(\mathbf{x}_s, \cdot), \mathbf{y}_0, \ell, \mu, \delta$)
9: **Output:** $(\mathbf{x}_s, \mathbf{y}_s)$

A.1 Nonconvex-strongly-concave setting under Definition 5

Our algorithm for nonconvex-strongly-concave optimization is described in Algorithm 5. Similar to Algorithm 4, we still use our accelerated solver Maximin-AG2 for the same proximal subproblem in the inner loop. The only minor difference is that, in the outer loop, Algorithm 5 only uses the Proximal Point Algorithm (PPA) on function $\Phi(\cdot) := \max_{y \in \mathcal{Y}} f(\cdot, \mathbf{y})$ without acceleration (or momentum steps). This is due to fact that gradient descent is already optimal among all first-order algorithm for finding stationary points of smooth nonconvex functions [Carmon et al.(2019a)]. The standard acceleration technique will not help for smooth nonconvex functions. We presents the theoretical guarantees for Algorithm 5 in the following theorem.

Theorem 5. *Assume that f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is $\mu_\mathbf{y}$-strongly-concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$ MINIMAX-PPA($f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_\mathbf{y}, \epsilon, T$) is an ϵ-stationary point of f with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\kappa_\mathbf{y}} \log^2\left(\frac{\kappa_\mathbf{y}\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right)$$

where $\kappa_\mathbf{y} = \ell/\mu_\mathbf{y}$ is the condition number, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x} \in \mathbb{R}^m} \Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_\mathbf{x} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$.

Theorem 5 claims that Algorithm 5 will find an ϵ-stationary point, with at least constant probability, in $\tilde{O}(\sqrt{\kappa_\mathbf{y}}/\epsilon^2)$ gradient evaluations. Similar to Theorem 4, the inner loop takes $\tilde{O}(\sqrt{\kappa_\mathbf{y}})$ gradient evaluations to solve the proximal step since the condition number of $g_t(\cdot, \mathbf{y})$ is $O(1)$ for any $\mathbf{y} \in \mathcal{Y}$. In the outer loop, regardless of the smoothness of $\Phi(\cdot)$, PPA with proximal parameter ℓ is capable of finding the stationary point in $\tilde{O}(1/\epsilon^2)$ iterations. In total, the gradient complexity is $\tilde{O}(\sqrt{\kappa_\mathbf{y}}/\epsilon^2)$.

A.2 Nonconvex-concave setting under Definition 5

Our result in the nonconvex-strongly-concave setting readily implies a fast result in the nonconvex-concave setting. Consider the following auxiliary function for an arbitrary $\mathbf{y}_0 \in \mathcal{Y}$:

$$\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) - (\epsilon/4D_\mathbf{y})\|\mathbf{y} - \mathbf{y}_0\|^2 \tag{35}$$

By construction, it is clear that the gradient of f and $\tilde{f}_\epsilon$ are close in the sense

$$\max_{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathcal{Y}} \|\nabla f(\mathbf{x}, \mathbf{y}) - \nabla \tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \leq \epsilon/4.$$

This implies that any $(\epsilon/2)$-stationary point of $\tilde{f}_\epsilon$ is also a ϵ-stationary point of f, and thus it is sufficient to solve the problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{x}\in\mathcal{Y}} \tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})$. Finally, the function $\tilde{f}_\epsilon(\mathbf{x}, \cdot)$ is always $\epsilon/(2D_\mathbf{y})$-strongly-concave, which can be fed into Algorithm 5 to obtain the following result.

Corollary 3. *Assume that f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-PPA}(\tilde{f}_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(2D_\mathbf{y}), \epsilon/2, T)$ is an ϵ-stationary point of f with probability at least 2/3, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\frac{\ell D_\mathbf{y}}{\epsilon}} \log^2\left(\frac{\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right)$$

where $D_\mathbf{y} > 0$, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_\mathbf{x} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \text{argmin}_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$.

A.3 Optimality notion based on Moreau envelope

We present another optimality notion based on Moreau envelope for nonconvex-concave setting in which $f(\cdot, \mathbf{y})$ is not necessarily convex for each $\mathbf{y} \in \mathcal{Y}$ but $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$. For simplicity, we let $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y}$ be convex and bounded. In general, finding a global saddle point of f is intractable since solving the special case with a singleton $\mathcal{Y}$ globally is already NP-hard [Murty and Kabadi(1987)] as mentioned in the main text.

One approach, inspired by nonconvex optimization, is to equivalently reformulate problem (1) as the following nonconvex minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \left\{\Phi(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})\right\} \tag{36}$$

and define an optimality notion for the local surrogate of global optimum of Φ. In robust learning, $\mathbf{x}$ is the classifier while $\mathbf{y}$ is the adversarial noise. Practitioners are often only interested in finding a robust classifier $\mathbf{x}$ instead of an adversarial response $\mathbf{y}$ to each data point. Such a stationary point $\mathbf{x}$ precisely corresponds to a robust classifier that is stationary to the robust classification error.

If $f(\mathbf{x}, \cdot)$ is further assumed to be strongly concave for each $\mathbf{x} \in \mathbb{R}^m$, then Φ is smooth and a standard optimality notion is the stationary point.

Definition 6. *We call $\hat{\mathbf{x}}$ an ϵ-stationary point of a **smooth** function Φ if $\|\nabla\Phi(\hat{\mathbf{x}})\| \le \epsilon$. If $\epsilon = 0$, then $\hat{\mathbf{x}}$ is called a stationary point.*

In contrast, when $f(\mathbf{x}, \cdot)$ is merely concave for each $\mathbf{x} \in \mathcal{X}$, Φ is not necessarily smooth and even not differentiable. A weaker sufficient condition for the purpose of our paper is the weak convexity.

Definition 7. *A function $\Phi: \mathbb{R}^d \to \mathbb{R}$ is L-weakly convex if $\Phi(\cdot) + (L/2)\|\cdot\|^2$ is convex.*

First, a function Φ is ℓ-weakly convex if it is ℓ-smooth. Second, the subdifferential of a ℓ-weakly convex function Φ can be uniquely determined by the subdifferential of $\Phi(\cdot) + \frac{\ell}{2}\|\cdot\|^2$. This implies that the optimality notion can be defined by a point $\mathbf{x} \in \mathbb{R}^m$ with at least one small subgradient: $\min_{\xi\in\partial\Phi(\mathbf{x})}\|\xi\| \le \epsilon$. Unfortunately, this notion can be restrictive if Φ is nonsmooth. Considering a one-dimensional function $\Phi(\cdot) = |\cdot|$, a point $\mathbf{x}$ must be 0 if it satisfies the optimality notion with $\epsilon \in [0, 1)$. This means that finding a sufficiently accurate solution under such optimality notion is as difficult as solving the minimization exactly. Another popular optimality notion is based on the *Moreau envelope* of Φ when Φ is weakly convex [Davis and Drusvyatskiy(2019)].

Definition 8. *A function Φ_λ is the Moreau envelope of Φ with $\lambda > 0$ if for $\forall \mathbf{x} \in \mathbb{R}^m$, that*

$$\Phi_\lambda(\mathbf{x}) \;=\; \min_{\mathbf{w} \in \mathbb{R}^m} \Phi(\mathbf{w}) + (1/2\lambda)\|\mathbf{w} - \mathbf{x}\|^2$$

Lemma 4 (Properties of Moreau envelopes)**.** *If the function $\Phi(\cdot)$ is ℓ-weakly convex, its Moreau envelope $\Phi_{1/2\ell}(\cdot)$ is 4ℓ-smooth with the gradient $\nabla\Phi_{1/2\ell}(\cdot) = 2\ell(\cdot - \mathrm{prox}_{\Phi/2\ell}(\cdot))$ in which a point $\mathrm{prox}_{\Phi/2\ell}(\cdot) = \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^m}\{\Phi(\mathbf{w}) + \ell\|\mathbf{w} - \cdot\|^2\}$ is defined.*

Thus, an ϵ-stationary point of an ℓ-weakly convex function Φ can be alternatively defined as a point $\hat{\mathbf{x}}$ satisfying that the gradient norm of Moreau envelope $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|$ is small.

Definition 9. *We call $\hat{\mathbf{x}}$ an ϵ-stationary point of a ℓ-weakly convex function Φ if $\left\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\right\| \leq \epsilon$. If $\epsilon = 0$, then $\hat{\mathbf{x}}$ is called a stationary point.*

Lemma 5 (Properties of ϵ-stationary point)**.** *If $\hat{\mathbf{x}}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ, then there exists $\bar{\mathbf{x}} \in \mathbb{R}^m$ such that $\min_{\xi \in \partial\Phi(\bar{\mathbf{x}})}\|\xi\| \leq \epsilon$ and $\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\| \leq \epsilon/2\ell$.*

Lemma 5 shows that an ϵ-stationary point defined by the Moreau envelope can be interpreted as the relaxation for a point with at least one small subgradient. In particular, if $\hat{\mathbf{x}}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ, then it is close to a point which has small subgradient.

A.4 Nonconvex-strongly-concave setting under Definition 6

In the setting of nonconvex-strongly-concave function, we still use Algorithm 5. Similar to Theorem 5, we can obtain a guarantee, which finds a point $\hat{\mathbf{x}}$ satisfying $\|\nabla\Phi(\hat{\mathbf{x}})\| \leq \epsilon$ in the same number of iterations as in Theorem 5.

Theorem 6. *Assume that f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is $\mu_{\mathbf{y}}$-strongly-concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathrm{MINIMAX\text{-}PPA}(f, \mathbf{x}_0, \mathbf{y}_0, \ell, \mu_{\mathbf{y}}, \epsilon, T)$ satisfies $\|\nabla\Phi(\hat{\mathbf{x}})\| \leq \epsilon$ with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\kappa_{\mathbf{y}}}\log^2\left(\frac{\kappa_{\mathbf{y}}\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $\kappa_{\mathbf{y}} = \ell/\mu_{\mathbf{y}}$ is the condition number, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x} \in \mathbb{R}^m}\Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \mathrm{argmin}_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{y}_0)$.

A.5 Nonconvex-concave setting under Definition 9

We can similarly reduce the problem of optimizing a nonconvex-concave function to the problem of optimizing a nonconvex-strongly-concave function. The only caveat is that, in order to achieve the near-optimal point using Definition 9 as optimality measure, we can only add a $O(\epsilon^2)$ term as follows:

$$\bar{f}_\epsilon(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) - (\epsilon^2/200\ell D_{\mathbf{y}}^2)\|\mathbf{y} - \mathbf{y}_0\|^2 \tag{37}$$

Now $\bar{f}_\epsilon(\mathbf{x}, \cdot)$ is only $\epsilon^2/(100\ell D_{\mathbf{y}}^2)$-concave, by feeding it to Algorithm 5 and through a slightly more complicated reduction argument, we can only obtain gradient complexity bound of $\tilde{O}(\epsilon^{-3})$ instead of $\tilde{O}(\epsilon^{-2.5})$ as in Corollary 3. Formally, we have

Corollary 4. *Assume that f is ℓ-smooth, and $f(\mathbf{x}, \cdot)$ is concave for all $\mathbf{x}$. Then there exists $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \text{MINIMAX-PPA}(\bar{f}_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon^2/(100\ell D_{\mathbf{y}}^2), \epsilon/10, T)$ satisfies $\|\nabla \Phi_{1/2\ell}(\hat{\mathbf{x}})\| \leq \epsilon$ with probability at least $2/3$, and the total number of gradient evaluations is bounded by*

$$\mathcal{O}\left(\frac{\ell^2 D_{\mathbf{y}} \Delta_\Phi}{\epsilon^3} \log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

where $D_{\mathbf{y}} > 0$, $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x} \in \mathbb{R}^m} \Phi(\mathbf{x})$ is the initial function value gap and $\tilde{D}_{\mathbf{x}} = \|\mathbf{x}_0 - \mathbf{x}_{g_1}^\star(\mathbf{y}_0)\|$ is the initial distance where $\mathbf{x}_g^\star(\mathbf{y}_0) = \text{argmin}_{\mathbf{x} \in \mathcal{X}}\, g(\mathbf{x}, \mathbf{y}_0)$.

B Proofs for Nonconvex-Concave Settings

In this section, we present proofs for all results in Section A.

B.1 Proof of Theorem 5

Using the definition of g_t, we have

$$\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}_{t+1}, \mathbf{y}) + \ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \;\leq\; \min_{\mathbf{x} \in \mathcal{X}}\left\{\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right\} + \delta$$

This implies that

$$\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \;\leq\; \min_{\mathbf{x} \in \mathcal{X}}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right\} + \delta \;\leq\; \Phi(\mathbf{x}_t) + \delta$$

Equivalently, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \;\leq\; \frac{\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) + \delta}{\ell} \tag{38}$$

Note that the function $\Phi(\cdot) + \ell\|\cdot - \mathbf{x}_t\|^2$ is ℓ-strongly convex and its minimizer $\mathbf{x}_t^*$ is well defined and unique [Davis and Drusvyatskiy(2019)]. Since the function $\Phi(\cdot) + \ell\|\cdot - \mathbf{x}_t\|^2$ is ℓ-strongly convex, we derive from [Nesterov(2018), Theorem 2.1.5] that

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2 \;\leq\; \frac{2}{\ell}\left(\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \min_{\mathbf{x} \in \mathbb{R}^m}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right\}\right) \;\leq\; \frac{2\delta}{\ell} \tag{39}$$

Since Φ is differentiable, we have

$$\left\|\mathbf{x}_t^* - \mathcal{P}_\mathcal{X}\left(\mathbf{x}_t^* - \frac{\nabla\Phi(\mathbf{x}_t^*) + 2\ell(\mathbf{x}_t^* - \mathbf{x}_t)}{\ell}\right)\right\| \;=\; 0$$

Therefore, we have

$$\left\|\mathbf{x}_{t+1} - \mathcal{P}_\mathcal{X}\left(\mathbf{x}_{t+1} - \frac{\nabla\Phi(\mathbf{x}_{t+1})}{\ell}\right)\right\| \;\leq\; 2\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\|\mathbf{x}_t - \mathbf{x}_t^*\| + \frac{\|\nabla\Phi(\mathbf{x}_t^*) - \nabla\Phi(\mathbf{x}_{t+1})\|}{\ell}$$

Since $\Phi(\cdot)$ is $2\kappa_{\mathbf{y}}\ell$-smooth, we have $\|\nabla\Phi(\mathbf{x}_{t+1}) - \nabla\Phi(\mathbf{x}_t^*)\| \leq 2\kappa_{\mathbf{y}}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|$. Putting these pieces together yields that

$$\left\|\mathbf{x}_{t+1} - \mathcal{P}_\mathcal{X}\left(\mathbf{x}_{t+1} - \frac{\nabla\Phi(\mathbf{x}_{t+1})}{\ell}\right)\right\| \;\leq\; (2\kappa_{\mathbf{y}} + 2)\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\|\mathbf{x}_t - \mathbf{x}_t^*\| \tag{40}$$

$$\overset{\kappa_{\mathbf{y}} \geq 1}{\leq} \; 6\kappa_{\mathbf{y}}\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$$

35

Putting Eq. (38), Eq. (39) and Eq. (40) together with the Cauchy-Schwarz inequality yields

$$(\ell \left\| \mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - \frac{1}{\ell}\nabla\Phi(\mathbf{x}_{t+1})\right)\right\|)^2 \leq 72\kappa_{\mathbf{y}}^2\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t^\star\|^2 + 8\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$

$$\leq 8\ell\left(\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) + \delta\right) + 144\kappa_{\mathbf{y}}^2\ell\delta$$

Summing up the above inequality over $t = 0, 1, \ldots, T-1$ and dividing it by T yields that

$$\frac{1}{T}\left(\sum_{t=0}^{T-1}(\ell\left\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - \frac{1}{\ell}\nabla\Phi(\mathbf{x}_{t+1})\right)\right\|)^2\right) \leq \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 8\ell\delta + 144\kappa_{\mathbf{y}}^2\ell\delta$$

$$\overset{\kappa_{\mathbf{y}} \geq 1}{\leq} \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 152\kappa_{\mathbf{y}}^2\ell\delta$$

Since $\hat{\mathbf{x}} = \mathbf{x}_s$ is uniformly chosen from $\{\mathbf{x}_s\}_{1 \leq s \leq T}$ and $\delta \leq \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, we have

$$\mathbb{E}\left[(\ell\left\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}\left(\hat{\mathbf{x}} - \frac{1}{\ell}\nabla\Phi(\hat{\mathbf{x}})\right)\right\|)^2\right] = \frac{1}{T}\left(\sum_{t=0}^{T-1}(\ell\left\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}}\left(\mathbf{x}_{t+1} - \frac{1}{\ell}\nabla\Phi(\mathbf{x}_{t+1})\right)\right\|)^2\right)$$

$$\leq \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 152\kappa_{\mathbf{y}}^2\ell\delta \leq \frac{8\ell\Delta_\Phi}{T} + \frac{\epsilon^2}{8}$$

Using the Markov inequality, we conclude that there exists $T > c\ell\Delta_\Phi\epsilon^{-2}$, where the output $\hat{\mathbf{x}}$ will satisfy $\ell\left\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}\left(\hat{\mathbf{x}} - \frac{1}{\ell}\nabla\Phi(\hat{\mathbf{x}})\right)\right\| \leq \epsilon/2$ with probability at least $2/3$.

For simplicity, we denote $\hat{\mathbf{y}}^+ = \mathcal{P}_{\mathcal{Y}}[\hat{\mathbf{y}} + \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$. Since $\hat{\mathbf{y}}$ is obtained by running AGD on $-f(\hat{\mathbf{x}}, \cdot)$ to optimal with tolerance $\delta \leq \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, and $f(\hat{\mathbf{x}}, \cdot)$ is $\mu_{\mathbf{y}}$-concave function, we know that δ-optimality guarantees:

$$\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \leq \epsilon, \quad \|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\| \leq \frac{\epsilon}{2\ell}$$

Putting these pieces together yields that

$$\ell\left\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}\left(\hat{\mathbf{x}} - \frac{1}{\ell}\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\right)\right\| \leq \ell\left\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}\left(\hat{\mathbf{x}} - \frac{1}{\ell}\nabla\Phi(\hat{\mathbf{x}})\right)\right\| + \|\nabla\Phi(\hat{\mathbf{x}}) - \nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\|$$

$$\leq \ell\left\|\hat{\mathbf{x}} - \mathcal{P}_{\mathcal{X}}\left(\hat{\mathbf{x}} - \frac{1}{\ell}\nabla\Phi(\hat{\mathbf{x}})\right)\right\| + \ell\|\hat{\mathbf{y}}^+ - \mathbf{y}^\star(\hat{\mathbf{x}})\|$$

$$\leq \epsilon$$

This implies that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-stationary point. Furthermore, we call the solver MAXIMIN-AG2 at each iteration. Using Theorem 3 and $\delta \leq \epsilon^2/(10\kappa_{\mathbf{y}})^4\ell$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_{\mathbf{y}}}\log\left(\frac{\kappa_{\mathbf{y}}^5\ell^2(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon^2}\right)\log\left(\frac{\kappa_{\mathbf{y}}^4\ell^2 D_{\mathbf{y}}^2}{\epsilon^2}\right)\right)$$

Therefore, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\kappa_{\mathbf{y}}}\log^2\left(\frac{\kappa_{\mathbf{y}}\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

This completes the proof.

B.2　Proof of Corollary 3

Recall that the function $\tilde{f}_\epsilon$ is defined by

$$\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) \;=\; f(\mathbf{x}, \mathbf{y}) - \frac{\epsilon\|\mathbf{y} - \mathbf{y}_0\|^2}{4D_\mathbf{y}}$$

This implies that the following statement holds for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ that

$$\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{x} \tilde{f}_\epsilon(\mathbf{x}, \mathbf{y}) \;=\; 0$$
$$\|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{y} \tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \;\leq\; \frac{\epsilon}{2}$$

Since $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \textsc{Minimax-PPA}(\tilde{f}_\epsilon, \mathbf{x}_0, \mathbf{y}_0, \ell, \epsilon/(2D_\mathbf{y}), \epsilon/2, T)$, we have

$$\ell\|\mathcal{P}_\mathcal{X}[\hat{\mathbf{x}} - \tfrac{1}{\ell}\nabla_\mathbf{x}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}}_\epsilon^+)] - \hat{\mathbf{x}}\| \leq \frac{\epsilon}{2}, \quad \ell\|\hat{\mathbf{y}}_\epsilon^+ - \hat{\mathbf{y}}\| \leq \frac{\epsilon}{2}, \quad \hat{\mathbf{y}}_\epsilon^+ = \mathcal{P}_\mathcal{Y}[\hat{\mathbf{y}} + \tfrac{1}{\ell}\nabla_\mathbf{y}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$$

Putting these pieces together with $\hat{\mathbf{y}}^+ = \mathcal{P}_\mathcal{Y}[\hat{\mathbf{y}} + \frac{1}{\ell}\nabla_\mathbf{y} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]$ yields that

$$\ell\|\mathcal{P}_\mathcal{X}[\hat{\mathbf{x}} - \tfrac{1}{\ell}\nabla_\mathbf{x} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)] - \hat{\mathbf{x}}\| \;\leq\; \frac{\epsilon}{2} + \|\nabla_\mathbf{x}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}}_\epsilon^+) - \nabla_\mathbf{x} f(\hat{\mathbf{x}}, \hat{\mathbf{y}}^+)\| \;\leq\; \frac{\epsilon}{2} + \|\nabla_\mathbf{y}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}}) - \nabla_\mathbf{y} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| \;\leq\; \epsilon$$

and

$$\ell\|\hat{\mathbf{y}}^+ - \hat{\mathbf{y}}\| \;\leq\; \ell\|\mathcal{P}_\mathcal{Y}[\hat{\mathbf{y}} + \tfrac{1}{\ell}\nabla_\mathbf{y}\tilde{f}_\epsilon(\hat{\mathbf{x}}, \hat{\mathbf{y}})] - \hat{\mathbf{y}}\| + \|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{y}\tilde{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \;\leq\; \epsilon$$

Therefore, we conclude that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-stationary point of f. Furthermore, letting $\kappa_\mathbf{y} = 2\ell D_\mathbf{y}/\epsilon$ in the gradient complexity bound presented in Theorem 5, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2} \cdot \sqrt{\frac{\ell D_\mathbf{y}}{\epsilon}} \log^2\left(\frac{\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right)$$

This completes the proof.

B.3　Proof of Theorem 6

Using the same argument as in Theorem 5, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \;\leq\; \frac{\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) + \delta}{\ell} \tag{41}$$

and

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2 \;\leq\; \frac{2}{\ell}\left(\Phi(\mathbf{x}_{t+1}) + \ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \min_{\mathbf{x} \in \mathbb{R}^m}\left\{\Phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right\}\right) \;\leq\; \frac{2\delta}{\ell} \tag{42}$$

Since Φ is differentiable, we have $\nabla\Phi(\mathbf{x}_t^*) + 2\ell(\mathbf{x}_t^* - \mathbf{x}_t) = 0$ which implies $\|\nabla\Phi(\mathbf{x}_t^*)\| = 2\ell\|\mathbf{x}_t^* - \mathbf{x}_t\|$. Since $\Phi(\cdot)$ is $2\kappa_\mathbf{y}\ell$-smooth, we have $\|\nabla\Phi(\mathbf{x}_{t+1}) - \nabla\Phi(\mathbf{x}_t^*)\| \leq 2\kappa_\mathbf{y}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|$. Putting these pieces together yields that

$$\|\nabla\Phi(\mathbf{x}_{t+1})\| \;\leq\; 2\kappa_\mathbf{y}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\ell\|\mathbf{x}_t^* - \mathbf{x}_t\| \;\leq\; (2\kappa_\mathbf{y}\ell + 2\ell)\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$$
$$\overset{\kappa_\mathbf{y} \geq 1}{\leq} \;4\kappa_\mathbf{y}\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\| + 2\ell\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \tag{43}$$

Putting Eq. (41), Eq. (42) and Eq. (43) together with the Cauchy-Schwarz inequality yields

$$\|\nabla\Phi(\mathbf{x}_{t+1})\|^2 \ \leq \ 32\kappa_\mathbf{y}^2\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2 + 8\ell^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \ \leq \ 8\ell\left(\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) + \delta\right) + 64\kappa_\mathbf{y}^2\ell\delta$$

Summing up the above inequality over $t = 0, 1, \ldots, T-1$ and dividing it by T yields that

$$\frac{1}{T}\left(\sum_{t=0}^{T-1}\|\nabla\Phi(\mathbf{x}_{t+1})\|^2\right) \ \leq \ \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 8\ell\delta + 64\kappa_\mathbf{y}^2\ell\delta \ \overset{\kappa_\mathbf{y}\geq 1}{\leq} \ \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 72\kappa_\mathbf{y}^2\ell\delta$$

Since $\hat{\mathbf{x}} = \mathbf{x}_s$ is uniformly chosen from $\{\mathbf{x}_s\}_{1 \leq s \leq T}$ and $\delta \leq \epsilon^2/144\kappa_\mathbf{y}^2\ell$, we have

$$\mathbb{E}\left[\|\nabla\Phi(\hat{\mathbf{x}})\|^2\right] \ = \ \frac{1}{T}\left(\sum_{t=0}^{T-1}\|\nabla\Phi(\mathbf{x}_{t+1})\|^2\right) \ \leq \ \frac{8\ell(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T))}{T} + 72\kappa_\mathbf{y}^2\ell\delta \ \leq \ \frac{8\ell\Delta_\Phi}{T} + \frac{\epsilon^2}{2}$$

Using the Markov inequality, we conclude that there exists $T > c\ell\Delta_\Phi\epsilon^{-2}$, where the output $\hat{\mathbf{x}}$ will satisfy $\|\nabla\Phi(\hat{\mathbf{x}})\| \leq \epsilon$ with probability at least 2/3. Furthermore, we call the solver MAXIMIN-AG2 at each iteration. Using Theorem 3 and $\delta \leq \epsilon^2/144\kappa_\mathbf{y}^2\ell$, the number of gradient evaluations at each iteration is bounded by

$$O\left(\sqrt{\kappa_\mathbf{y}}\log\left(\frac{\kappa_\mathbf{y}^3\ell^2(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon^2}\right)\log\left(\frac{\kappa_\mathbf{y}^2\ell^2 D_\mathbf{y}^2}{\epsilon^2}\right)\right)$$

Therefore, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell\Delta_\Phi}{\epsilon^2}\cdot\sqrt{\kappa_\mathbf{y}}\log^2\left(\frac{\kappa_\mathbf{y}\ell(\tilde{D}_\mathbf{x}^2 + D_\mathbf{y}^2)}{\epsilon}\right)\right)$$

This completes the proof.

B.4   Proof of Corollary 4

Recall that the function $\bar{f}_\epsilon$ is defined by

$$\bar{f}_\epsilon(\mathbf{x}, \mathbf{y}) \ = \ f(\mathbf{x}, \mathbf{y}) - \frac{\epsilon^2\|\mathbf{y} - \mathbf{y}_0\|^2}{200\ell D_\mathbf{y}^2}$$

This implies that the following statement holds for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ that

$$\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{x}\bar{f}_\epsilon(\mathbf{x}, \mathbf{y}) \ = \ 0$$

$$\|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{y}\bar{f}_\epsilon(\mathbf{x}, \mathbf{y})\| \ \leq \ \frac{\epsilon^2}{100\ell D_\mathbf{y}}$$

Using Theorem 6 and letting $\mathbf{y}_\epsilon^\star(\cdot) = \text{argmin}_{\mathbf{y}\in\mathcal{Y}}\bar{f}_\epsilon(\cdot, \mathbf{y})$, we have

$$\|\nabla_\mathbf{x}\bar{f}_\epsilon(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]\| \ \leq \ \frac{\epsilon}{10}$$

$$\ell\|\mathcal{P}_\mathcal{Y}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + \frac{1}{\ell}\nabla_\mathbf{y}\bar{f}_\epsilon(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))] - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \ = \ 0$$

For simplicity, we define $\mathbf{y}_\epsilon^+ = \mathcal{P}_{\mathcal{Y}}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]$. Then, we have

$$\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\| \le \frac{\epsilon}{10} + \frac{\epsilon^2}{50\ell D_{\mathbf{y}}}, \quad \ell\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \le \frac{\epsilon^2}{50\ell D_{\mathbf{y}}}$$

Now let $\mathbf{x}^\star(\hat{\mathbf{x}}) = \text{argmin}_{\mathbf{x}\in\mathbb{R}^m} \Phi_{1/2\ell}(\mathbf{x}) := \Phi(\mathbf{x}) + \ell\|\mathbf{x} - \hat{\mathbf{x}}\|^2$, we have

$$\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2 = 4\ell^2\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2$$

Since $\Phi(\cdot) + \ell\|\cdot - \hat{\mathbf{x}}\|^2$ is $\ell/2$-strongly-convex, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^*(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^*(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2$$

$$= \Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{x}^*(\hat{\mathbf{x}})) - \ell\|\mathbf{x}^*(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \ge \frac{\ell\|\hat{\mathbf{x}} - \mathbf{x}^*(\hat{\mathbf{x}})\|^2}{4} = \frac{\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2}{16\ell}$$

Furthermore, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^*(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^*(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2$$

$$= \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) + f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^*(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^*(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2$$

$$\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) + \left(f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) - f(\mathbf{x}^*(\hat{\mathbf{x}}), \mathbf{y}_\epsilon^+) - \ell\|\mathbf{x}^*(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2\right)$$

$$\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) + \left(\|\hat{\mathbf{x}} - \mathbf{x}^*(\hat{\mathbf{x}})\|\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\| - \ell\|\hat{\mathbf{x}} - \mathbf{x}^*(\hat{\mathbf{x}})\|^2\right)$$

$$\le \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) + \frac{\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\|^2}{4\ell}$$

Recall that $\mathbf{y}_\epsilon^+ = \mathcal{P}_{\mathcal{Y}}[\mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) + \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))]$, we have

$$(\mathbf{y} - \mathbf{y}_\epsilon^+)^\top \left(\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}) - \frac{1}{\ell}\nabla_{\mathbf{y}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))\right) \ge 0 \text{ for all } \mathbf{y} \in \mathcal{Y}$$

Together with the ℓ-smoothness of the function $f(\hat{\mathbf{x}}, \cdot)$ and the boundedness of $\mathcal{Y}$, we have

$$f(\hat{\mathbf{x}}, \mathbf{y}) - f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+) \le \frac{\ell}{2}(\|\mathbf{y} - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\|^2 - \|\mathbf{y} - \mathbf{y}_\epsilon^+\|^2) \le \ell D_{\mathbf{y}}\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \text{ for all } \mathbf{y} \in \mathcal{Y}$$

Putting these pieces together yields that

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \le \ell D_{\mathbf{y}}\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| + \frac{\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\|^2}{4\ell}$$

Since a point $(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}}))$ satisfies that

$$\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\| \le \frac{\epsilon}{10} + \frac{\epsilon^2}{50\ell D_{\mathbf{y}}}, \quad \ell\|\mathbf{y}_\epsilon^+ - \mathbf{y}_\epsilon^\star(\hat{\mathbf{x}})\| \le \frac{\epsilon^2}{50\ell D_{\mathbf{y}}}$$

we have (assume that $\epsilon \precsim \ell D_{\mathbf{y}}$ without loss of generality)

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \le \frac{\epsilon^2}{50\ell} + \frac{\|\nabla_{\mathbf{x}}f(\hat{\mathbf{x}}, \mathbf{y}_\epsilon^+)\|^2}{4\ell}$$

39

Putting these pieces together yields that $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \leq \epsilon$. Furthermore, letting $\kappa_{\mathbf{y}} = 100\ell^2 D_{\mathbf{y}}^2/\epsilon^2$ in the gradient complexity bound presented in Theorem 6, we conclude that the total number of gradient evaluations is bounded by

$$O\left(\frac{\ell^2 D_{\mathbf{y}}\Delta_\Phi}{\epsilon^3}\log^2\left(\frac{\ell(\tilde{D}_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)}{\epsilon}\right)\right)$$

This completes the proof.

C Proof of Technical Lemmas

In this section, we provide complete proofs for the lemmas in the paper.

C.1 Proof of Lemma 4

We provide a proof for an expanded version of Lemma 4.

Lemma 6. *If Φ is ℓ-weakly convex, we have*

(a) $\Phi_{1/2\ell}(\mathbf{x})$ and $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x}) = \mathrm{argmin}\,\Phi(\mathbf{w}) + \ell\|\mathbf{w} - \mathbf{x}\|^2$ are well defined for any $\mathbf{x} \in \mathbb{R}^m$.

(b) $\Phi(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^m$.

(c) $\Phi_{1/2\ell}$ is 4ℓ-smooth with $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \mathrm{prox}_{\Phi/2\ell}(\mathbf{x}))$.

Proof. Since Φ is ℓ-weakly convex, $\Phi(\cdot) + \frac{\ell}{2}\|\cdot - \mathbf{x}\|^2$ is convex for any $\mathbf{x} \in \mathbb{R}^m$. This implies that $\Phi(\cdot) + \ell\|\cdot - \mathbf{x}\|^2$ is $\frac{\ell}{2}$-strongly convex and $\Phi_{1/2\ell}(\mathbf{x})$ and $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$ are well defined. For any $\mathbf{x} \in \mathbb{R}^m$, the definition of $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$ implies that

$$\Phi(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi_{1/2\ell}(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \leq \Phi(\mathbf{x})$$

By [Davis and Drusvyatskiy(2019), Lemma 2.2], $\Phi_{1/2\ell}$ is differentiable with $\nabla\Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \mathrm{prox}_{\Phi/2\ell}(\mathbf{x}))$. Since $\mathrm{prox}_{\Phi/2\ell}$ is 1-Lipschitz, we $\|\nabla\Phi_{1/2\ell}(\mathbf{x}) - \nabla\Phi_{1/2\ell}(\mathbf{x}')\| \leq 4\ell\|\mathbf{x} - \mathbf{x}'\|$. Therefore, the function $\Phi_{1/2\ell}$ is 4ℓ-smooth. $\square$

C.2 Proof of Lemma 5

Denote $\hat{\mathbf{x}} := \mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$, part (c) in Lemma 4 implies

$$\|\hat{\mathbf{x}} - \mathbf{x}\| = \frac{\|\nabla\Phi_{1/2\ell}(\mathbf{x})\|}{2\ell}$$

Furthermore, we have $2\ell(\mathbf{x} - \hat{\mathbf{x}}) \in \partial\Phi(\hat{\mathbf{x}})$. Putting these pieces together yields the desired result.

Table 2. Comparison of gradient complexities to find an ϵ-stationary point of f (Definition 5) or ϵ-stationary point of $\Phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ (Definition 6, 9) in the nonconvex-concave settings. This table only highlights the dependence on tolerance ϵ and the condition number $\kappa_{\mathbf{y}}$.

Settings	References	Gradient Complexity
Nonconvex-Strongly-Concave (stationarity of f or stationarity of Φ)	[Jin et al.(2019)]	$\tilde{O}(\kappa_{\mathbf{y}}^2 \epsilon^{-2})$
	[Rafique et al.(2018)]	
	[Lin et al.(2019)]	
	[Lu et al.(2019)]	
	This paper (Theorem 5 & 6)	$\tilde{O}(\sqrt{\kappa_{\mathbf{y}}}\epsilon^{-2})$
Nonconvex-Concave (stationarity of f)	[Lu et al.(2019)]	$\tilde{O}(\epsilon^{-4})$
	[Nouiehed et al.(2019)]	$\tilde{O}(\epsilon^{-3.5})$
	[Ostrovskii et al.(2020)]	$\tilde{O}(\epsilon^{-2.5})$
	This paper (Corollary 3)	$\tilde{O}(\epsilon^{-2.5})$
Nonconvex-Concave (stationarity of Φ)	[Jin et al.(2019)]	$\tilde{O}(\epsilon^{-6})$
	[Rafique et al.(2018)]	
	[Lin et al.(2019)]	
	[Thekumparampil et al.(2019)]	$\tilde{O}(\epsilon^{-3})$
	[Zhao(2020)]	
	This paper (Corollary 4)	$\tilde{O}(\epsilon^{-3})$