Complexity Lower Bounds for Nonconvex-Strongly-Concave Min-Max Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

We provide a first-order oracle complexity lower bound for finding stationary points of min-max optimization problems where the objective function is smooth, nonconvex in the minimization variable, and strongly concave in the maximization variable. We establish a lower bound of $\Omega\left(\sqrt{\kappa}\epsilon^{-2}\right)$ for deterministic oracles, where ϵ defines the level of approximate stationarity and κ is the condition number. Our lower bound matches the best existing upper bound in the ϵ and κ dependence up to logarithmic factors. For stochastic oracles, we provide a lower bound of $\Omega\left(\sqrt{\kappa}\epsilon^{-2} + \kappa^{1/3}\epsilon^{-4}\right)$. It suggests that there is a gap between the best existing upper bound $\mathcal{O}(\kappa^3\epsilon^{-4})$ and our lower bound in the condition number dependence.

1 Introduction

In this paper, we study the oracle complexity lower bound of the following min-max optimization problem:

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

where $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^n$ are nonempty closed convex sets. Such a problem arises in a wide range of applications, e.g., two-player zero-sum games [37], Generative Adversarial Networks [17], robust optimization [4], including defending adversarial attacks [18, 28].

The research in min-max problems (1) has a long history. If f is linear in both $\boldsymbol{x}$ and $\boldsymbol{y}$, the problem is known as bilinear min-max optimization, for which von Neumann's min-max theorem [37] guarantees the existence of a saddle point, a point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ that satisfies strong duality (i.e. min-max equals max-min). Later, Sion [43] generalized the statement to convex-concave functions under certain regularity conditions, which led to the theory of KKT conditions. Aside from studying the existence of strong duality, Nemirovski [32], Nesterov [35] proposed algorithms that can find approximate saddle points for convex-concave objectives with an $\mathcal{O}(1/\epsilon)$ convergence rate. The rate was proven to be worst-case optimal even for min-max problems with a bilinear cross-term [40].

Going beyond the assumption of convex-concavity poses a big challenge. In general nonconvex-nonconcave min-max optimization problems, a saddle-point may not exist [21]. Defining a notion of min-max points that is simultaneously nontrivial and tractable is still an open question. If one uses the standard definition of saddle points as in the strong duality problems, then determining its existence is known to be NP-hard [9], and finding an approximate local saddle point is PPAD-complete [9]. Alternatively, a relaxed definition of min-max point known as Stackelberg equilibrium is guaranteed to exist [21], yet only local asymptotic convergence is known [21, 46].

Due to the lack of a well-formulated suboptimality measure for general nonconvex-nonconcave problems, recent research has considered problems with special structures. For example, Lin et al. [25, 26] studied nonconvex-concave problems, Yang et al. [49] studied problems satisfying the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Polyak-Lojasiewicz inequality, Diakonikolas et al. [11] studied problems with weak Minty variational inequality solutions, Mangoubi and Vishnoi [29] considered an algorithm specific equilibrium.

Our work focuses on the analysis of gradient oracle complexity in the nonconvex-strongly-concave min-max setting. In the standard convex-concave setup, Lin et al. [26] designed a set of algorithms that achieve near-optimal gradient complexity for all the three variants of convex-concave min-max optimization. However, in the nonconvex-concave setting, it is unclear whether their algorithm is optimal given that no matching lower bound exists. In this work, our main contribution is as follows:

- We construct an explicit example on which no first-order zero-respecting algorithm can achieve less than $\Omega(\sqrt{\kappa}\epsilon^{-2})$ oracle complexity for smooth nonconvex-strongly-concave problems, matching the $\tilde{\mathcal{O}}(\sqrt{\kappa}\epsilon^{-2})$ complexity in [26].
- We further extend our results to the stochastic(-oracle) setting, achieving the complexity lower bound $\Omega(\sqrt{\kappa}\epsilon^{-2}+\kappa^{1/3}\epsilon^{-4})$. Compared against the results in [25], our result suggests that the dependency on the condition number may not be optimal.

1.1 Related work

Upper bounds. Since standard gradient descent-ascent diverges for even the simplest case of bilinear zero-sum games, researchers proposed different variations to achieve convergence in the convex-concave setting, with possibly additional structures. One of the variations is known as extra-gradient method [23, 45, 32, 7, 48], which takes two gradient steps per iteration. Another line studies the optimistic gradient method [42, 8, 15, 30, 19]. Mokhtari et al. [31] proposed a unified analysis through the lens of proximal update. Some works [1, 26] focused on accelerating the known rates in terms of the conditional number dependence. Golowich et al. [16] studied last iterate convergence. Lee et al. [24], Daskalakis et al. [10], Wei et al. [47] studied two-player zero-sum stochastic games.

Convergence under the nonconvex-concave setup is less studied. Rafique et al. [41] studied a proximal version of gradient methods. Lin et al. [25] studied gradient descent ascent with different time scales. Thekumparampil et al. [44] proposed an implicit algorithm for the nonconvex-concave setting. Many other works also studied different variations of convergence [38, 22, 27, 39].

Lower bounds. Our work follows the complexity framework introduced by Nemirovski in the 1980s. The most classical lower bounds are due to Nemirovski and Nesterov, and are discussed in their textbooks [34, 36]. Recently, much progress was made [13, 5, 6, 3] by extending the original analysis to the nonconvex setting. Our work builds upon these results. Another line of works studies lower bounds for quadratic problems in min-max setting and utilizes a different framework sometimes known as Stationary Canonical Linear Iterative (SCLI) [2, 40, 50, 20]. It exploits the closed-form updates for linear dynamics and studies convergence properties via the transition matrices.

Concurrently, Zhang et al. [51] provided a similar lower complexity bound in the deterministic nonconvex-strongly-concave setting using a different construction, yet following the same formalism as Nemirovski and Carmon [33, 5]. Whereas Zhang et al. [51] provided novel analyses with extensions to the upper and lower bounds in the finite-sum setting, our proofs and construction are different and we also extend the results to the stochastic setting.

2 Preliminaries

Before introducing our main results, we describe the problem setup and algorithm complexity in this section. First, we introduce the notation. Next, we define the function class and algorithm class. Finally, we provide the formal definition of optimality measure and gradient complexity.

Notation. We use bold lower-case letters to denote vectors and use x_i to denote the i-th coordinate of vector $\boldsymbol{x}$. Let $\mathrm{supp}(\boldsymbol{x}) := \{i : x_i \neq 0\}$ denote the support of $\boldsymbol{x}$. Let $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ denote its ℓ_2 and ℓ_∞ norm respectively. For a matrix A, we use $A_{i,j}$ to denote its (i,j)-th entry and $A^\top$ its transpose. We use $\|A\|_2$ to denote its spectral norm and $\det(A)$ its determinant. We use calligraphic upper-case letters to denote sets, as in $\mathcal{X}$ and $\mathcal{Y}$. For a nonempty closed convex set $\mathcal{X} \in \mathbb{R}^d$, let $\mathsf{P}_{\mathcal{X}}$ denote the Euclidean projection onto $\mathcal{X}$. We also use a semi-comma besides commas to split minimization and maximization variables of a function. For example, when we write $f(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{y})$, it

means $\boldsymbol{x}$ and $\boldsymbol{z}$ are the variables to minimize and $\boldsymbol{y}$ is the variable to maximize. Finally, we use the standard $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ notation, with $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ further hiding log factors.

2.1 Function class

Our analysis focuses on the class of smooth nonconvex-strongly-concave functions defined below.

Definition 1. *Given $L \geq \mu > 0$ and $\Delta > 0$, we use $\mathcal{F}(L, \mu, \Delta)$ to denote the set of all functions $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for some nonempty closed convex sets $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^n$ where $m, n \in \mathbb{N}$, which satisfies the following assumptions:*

1. *f is L-smooth, that is, for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}$,*

$$\|\nabla f(\boldsymbol{x}_1; \boldsymbol{y}_1) - \nabla f(\boldsymbol{x}_2; \boldsymbol{y}_2)\|_2 \leq L \|(\boldsymbol{x}_1, \boldsymbol{y}_1) - (\boldsymbol{x}_2, \boldsymbol{y}_2)\|_2 ;$$

2. *For any fixed $\boldsymbol{x} \in \mathcal{X}$, f is μ-strongly concave in $\boldsymbol{y}$, that is, for any $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}$,*

$$f(\boldsymbol{x}; \boldsymbol{y}_1) \leq f(\boldsymbol{x}; \boldsymbol{y}_2) + \nabla_{\boldsymbol{y}} f(\boldsymbol{x}; \boldsymbol{y}_2) \cdot (\boldsymbol{y}_1 - \boldsymbol{y}_2) - \frac{1}{2}\mu \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2^2 ;$$

3. *$f_m(\mathbf{0}) - \min_{\boldsymbol{x} \in \mathcal{X}} f_m(\boldsymbol{x}) \leq \Delta$, where $f_m(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}; \boldsymbol{y})$.*

For any fixed $\boldsymbol{y} \in \mathcal{Y}$, f is potentially nonconvex in $\boldsymbol{x}$. Note that $\mathcal{F}(L, \mu, \Delta)$ includes functions with domain on $\mathbb{R}^m \times \mathbb{R}^n$ for all $m, n \in \mathbb{N}$, following the framework of dimension-free convergence guarantees [33, 36]. In Section 5, we will construct a hard instance with domain dimensions m, n both growing inversely in the required accuracy ϵ.

2.2 Algorithm class

In this section, we describe the algorithms of interest for the min-max problems with the function class defined above. In particular, we restrict our analysis to first-order algorithms that optimize objectives using first-order oracles defined below.

Definition 2 (Deterministic first-order oracle)**.** *The deterministic first-order oracle of a differentiable function $f : \mathcal{X} \to \mathbb{R}$ is a mapping $O : \boldsymbol{x} \mapsto (f(\boldsymbol{x}), \nabla f(\boldsymbol{x}))$ for $\boldsymbol{x} \in \mathcal{X}$.*

Definition 3 (Stochastic first-order oracle)**.** *A stochastic first-order oracle with bounded variance σ^2 of a differentiable function $f : \mathcal{X} \to \mathbb{R}$ is a mapping $O : \boldsymbol{x} \mapsto (f(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}, \xi))$ for $\boldsymbol{x} \in \mathcal{X}$, where ξ is a random variable satisfying $\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}, \xi)] = \nabla f(\boldsymbol{x})$ and $\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{x}, \xi) - \nabla f(\boldsymbol{x})\|_2^2\right] \leq \sigma^2$.*

We say O is a first-order oracle if it is a deterministic or stochastic first-order oracle.

Furthermore, we consider first-order algorithms satisfying the zero-respecting assumption. Formally, we define first-order algorithms as follows.

Definition 4 (First-order algorithm)**.** *A first-order (zero-respecting) algorithm is one that for any function $f : \mathcal{X} \to \mathbb{R}$ and its associated first-order oracle $O_f : \boldsymbol{x} \mapsto (f(\boldsymbol{x}), \boldsymbol{g})$, the $(t+1)$-th iterate $\boldsymbol{x}^{t+1}$ satisfies*

$$\boldsymbol{x}^{t+1} \in \left\{ \mathsf{P}_{\mathcal{X}}(\boldsymbol{v}) : \operatorname{supp}(\boldsymbol{v}) \subset \bigcup_{0 \leq i \leq t} \left(\operatorname{supp}(\boldsymbol{x}^i) \cup \operatorname{supp}(\boldsymbol{g}^i) \right) \right\}.$$

Definition 4 extends the standard zero-respecting algorithms [5, 3] to the constrained setting. It covers most existing first-order methods used in the literature including (projected) stochastic gradient descent, adaptive methods, and more importantly, the algorithms used in [25, 26] which achieve the upper bounds for nonconvex-strongly-concave min-max optimization in the deterministic and stochastic settings, respectively. There is a standard reduction from a lower bound for zero-respecting algorithms to that for arbitrary deterministic algorithms with deterministic or even stochastic first-order oracles [5, 3]. We defer this extension to future work.

2.3 Optimality via approximate stationarity

We measure the progress of solving the nonconvex-strongly-concave problem via the gradient norm of the maximized function with respect to the minimization variable $\boldsymbol{x}$. Following [14], we define the notion of ϵ-stationary points in presence of constraints as follows.

Definition 5. *Let $\mathcal{X}$ be a nonempty closed convex set. A point $\boldsymbol{x} \in \mathcal{X}$ is said to be an ϵ-stationary point of an L-smooth function $f : \mathcal{X} \to \mathbb{R}$ if*

$$L \left\| \mathsf{P}_{\mathcal{X}} \left[\boldsymbol{x} - (1/L) \nabla f(\boldsymbol{x}) \right] - \boldsymbol{x} \right\|_2 \leq \epsilon.$$

Note that the algorithm studied in [25] also assumes bounded domain and used the same notion of stationarity. This definition of stationary points reduces to $\|\nabla f(\boldsymbol{x})\|_2 \leq \epsilon$ when $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$. For any $f \in \mathcal{F}(L, \mu, \Delta)$, Definition 5 applies to f_m as it is L_m-smooth, where $L_m \leq (\kappa + 1)L$ by [25, Lemma 4.3].

Our goal of solving the nonconvex-strongly-concave min-max optimization problem is to find an ϵ-stationary point of f_m. We show that no deterministic first-order algorithm can achieve less than $\Omega(\sqrt{\kappa}\epsilon^{-2})$ gradient complexity for smooth nonconvex-strongly-concave problems, matching the $\tilde{\mathcal{O}}(\sqrt{\kappa}\epsilon^{-2})$ complexity in [26]. We further extend our result to the stochastic setting, achieving a complexity lower bound $\Omega(\sqrt{\kappa}\epsilon^{-2} + \kappa^{1/3}\epsilon^{-4})$.

3 Main results

In this section, we present our main results on the minimum number of gradient oracle calls required to find an ϵ-stationary point. The results for deterministic and stochastic settings are presented in the following two subsections respectively.

3.1 Lower bound on first-order oracle complexity in the deterministic setting

Nonconvex-strongly-concave min-max optimization subsumes nonconvex optimization, the lower bound $\Omega(L\Delta/\epsilon^2)$ in nonconvex optimization [5] also holds for nonconvex-strongly-concave min-max optimization. However, compared with the $\tilde{\mathcal{O}}(L\Delta\sqrt{\kappa}/\epsilon^2)$ upper bound [26], a $\sqrt{\kappa}$ factor is missing. Our main result below fills this gap, showing that the known rate by [25] is optimal up to log factors.

Theorem 1. *For any $\mu, L, \Delta, \epsilon > 0$ such that $\kappa = L/\mu \geq 1$, there exists a function instance $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ in $\mathcal{F}(L, \mu, \Delta)$ for some $m, n \in \mathbb{N}$ with its deterministic first-order oracle such that for any first-order algorithm, we have $\|\nabla f_m(\boldsymbol{x}^t)\|_2 > \epsilon$, where $f_m(\boldsymbol{x}) := \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, whenever*

$$t \leq \frac{c_0 L \Delta \sqrt{\kappa}}{\epsilon^2},$$

where c_0 is a numerical constant.

Similar to the lower bound in [5], our lower bound applies to dimension-free optimization in nature. That is, for a given ϵ, we construct a hard instance with dimension $d = \mathcal{O}(1/\epsilon^2)$, which can be very large if ϵ is small. The discussion on the proof is deferred until Section 5.

3.2 Lower bound in the stochastic setting

In the stochastic setting, Arjevani et al. [3] provided a lower bound of $\Omega\left(L\Delta\sigma^2/\epsilon^4\right)$ for smooth nonconvex optimization which is a special case of nonconvex-strongly-concave min-max optimization. Our analysis improves this bound by a factor of $\kappa^{1/3}$ in Theorem 2.

Theorem 2. *For any $\mu, L, \Delta, \epsilon, \sigma > 0$ such that $\kappa = L/\mu \geq 1$, there exists a function instance $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in $\mathcal{F}(L, \mu, \Delta)$ and a stochastic first-order oracle O for f with variance σ^2 such that for any first-order algorithm, we have $\mathbb{E}\left[L_m \left\| \mathsf{P}_{\mathcal{X}} \left[\boldsymbol{x}^t - (1/L_m)\nabla f_m(\boldsymbol{x}^t) \right] - \boldsymbol{x}^t \right\|_2\right] > \epsilon$, where $f_m(\boldsymbol{x}) := \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, whenever*

$$t \leq c_0 L \Delta \left(\frac{\sqrt{\kappa}}{\epsilon^2} + \frac{\kappa^{1/3}\sigma^2}{\epsilon^4} \right)$$

where L_m is the smoothness parameter of f_m and c_0 is a numerical constant.

Lin et al. [25] provided an upper bound of $\mathcal{O}\left(\kappa^3\epsilon^{-4}\right)$. Therefore, there is a gap between our lower bound and their upper bound in terms of the dependency on κ.

In summary, our proposed lower bounds improve the known ones by a multiplicative dependence on the condition number. Before proceeding to discuss the concrete techniques, we first summarize the general framework for establishing lower bounds in the next section.

4 Framework for proving lower bound

In this section, we provide an outline for the proof of the lower bound in [36] and [5], which lay the foundation for our construction of the hard instance. Both works utilize the notion of zero-chains [5], which instantiates a class of hard functions for optimization. We first define a zero-chain and then discuss how it is used to establish complexity lower bounds.

Definition 6 (Zero-chain)**.** *We say a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$, is a (first-order) zero-chain if for every $1 \le i \le d$,*

$$\operatorname{supp}(\boldsymbol{x}) := \{i : x_i \neq 0\} \subset \{1, \dots, i-1\} \implies \operatorname{supp}(\nabla f(\boldsymbol{x})) \subset \{1, \dots, i\}.$$

Suppose the domain $\mathcal{X} \subset \mathbb{R}^d$ satisfies $\operatorname{supp}(\mathsf{P}_{\mathcal{X}}(\boldsymbol{x})) = \operatorname{supp}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, i.e., projecting $\boldsymbol{x}$ onto $\mathcal{X}$ does not change its support. For example, this requirement holds when $\mathcal{X}$ is a hypercube or the whole space $\mathbb{R}^d$. If we run a first-order algorithm on a zero-chain initialized at $\boldsymbol{x}^0 = \mathbf{0}$ (which we assume to hold without loss of generality) with a deterministic first-order oracle, then at each iteration, at most one new coordinate of $\boldsymbol{x}$ becomes nonzero ("discovered"). Therefore, $\operatorname{supp}(\boldsymbol{x}^t) \subset \{1, \dots, t\}$. Then we obtain a lower complexity bound of T suppose we can show a good solution exists only if at least T coordinates are discovered.

Therefore, the key to proving a lower bound of first-order algorithms with a deterministic oracle is to find a function f such that:

1. it is a zero-chain that belongs to the function class we are interested in; and that
2. we cannot obtain an ϵ-optimal solution if the t-th coordinate of $\boldsymbol{x}$ is zero for every $t \ge T$.

This is actually a general strategy for proving lower bounds of first-order methods, used in the lower bound construction both by Carmon et al. [5] for smooth nonconvex optimization and by us here for nonconvex-strongly-concave min-max optimization.

In the stochastic setting, we utilize the generalized notion known as the probability-p zero-chain [3] to prove a lower bound.

Definition 7 (Probability-p zero-chain)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ with a stochastic first-order oracle $O : \boldsymbol{x} \mapsto (f(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}, \xi))$ is a probability-p zero-chain if*

$$\operatorname{supp}(\boldsymbol{x}) \subset \{1, \dots, i-1\} \quad \implies \quad \begin{cases} \mathbb{P}\left(\operatorname{supp}(\boldsymbol{g}(\boldsymbol{x}, \xi)) \not\subset \{1, \dots, i-1\}\right) \le p, \\ \mathbb{P}\left(\operatorname{supp}(\boldsymbol{g}(\boldsymbol{x}, \xi)) \subset \{1, \dots, i\}\right) = 1. \end{cases}$$

For a probability-p zero-chain, at each iteration, a new coordinate is discovered with probability at most p if $\operatorname{supp}(\mathsf{P}_{\mathcal{X}}(\boldsymbol{x})) = \operatorname{supp}(\boldsymbol{x})$ for all $\boldsymbol{x}$. Therefore, it takes at least $1/p$ steps in expectation to discover a new coordinate. Formally, the following lemma states that it takes $\mathcal{O}(T/p)$ iterations to reach the end of a probability-p zero-chain with length T.

Lemma 3 ([3, Lemma 1])**.** *Let $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \in \mathbb{R}^T$ satisfying $\operatorname{supp}(\mathsf{P}_{\mathcal{X}}(\boldsymbol{x})) = \operatorname{supp}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^T$, be a probability-$p$ zero-chain with a stochastic first-order oracle. For any first-order algorithm, we have with probability at least $1 - \delta$, $x_T^t = 0$ for all $t \le \frac{T - \log(1/\delta)}{2p}$.*

Therefore, the gradient complexity is enlarged by a factor of $1/p$ compared with the deterministic setting. To obtain a lower bound in the stochastic setting, we first find a zero-chain f satisfying the two requirements in the general strategy we presented above. Then we construct a stochastic first-order oracle which discovers the next coordinate with probability p so that we obtain a probability-p zero-chain. Note that we can not choose an arbitrarily small p since we need to ensure the variance of the stochastic oracle is bounded.

For the ease of exposition of our construction, we now briefly review the lower bound construction for smooth strongly-convex minimization by Nesterov [36] and that for smooth nonconvex optimization by Carmon et al. [5]. As we shall see in the next section, the hard functions in these two cases are the building blocks of our construction.

4.1 Smooth strongly-convex minimization

Nesterov [36] constructed the following hard instance for smooth, strongly-convex functions:

$$f^{\text{sc}}(\boldsymbol{x}) := \frac{\mu(\kappa-1)}{8}\left((x_1-1)^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2\right) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 . \tag{2}$$

Equivalently, $f^{\text{sc}}(\boldsymbol{x}) = \frac{\mu(\kappa-1)}{8}(\boldsymbol{x}^\top A^{\text{sc}}\boldsymbol{x} - 2x_1 + 1) + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ for the tri-diagonal matrix A^{sc} given by

$$A^{\text{sc}} := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3}$$

It is straightforward to verify that A^{sc} is positive semi-definite and $\|A^{\text{sc}}\|_2 \leq 4$. Hence, f^{sc} is L-smooth and μ-strongly-convex for $L := \mu\kappa$. Importantly, if $\operatorname{supp}(\boldsymbol{x}) \subset \{1, \ldots, i-1\}$, we can verify:

$$\operatorname{supp}(\nabla f^{\text{sc}}(\boldsymbol{x})) = \operatorname{supp}(A^{\text{sc}}\boldsymbol{x}) \cup \operatorname{supp}(\boldsymbol{x}) \cup \{1\} \subset \{1, \ldots, i\},$$

where we use the fact that $\operatorname{supp}(A^{\text{sc}}\boldsymbol{x}) \subset \{1, \ldots, i\}$ because A^{sc} is a tri-diagonal matrix. Hence, by Definition 6, f^{sc} is a zero-chain. Based on the general strategy above, it suffices to lower bound $f^{\text{sc}}(\boldsymbol{x}) - f^{\text{sc}}(\boldsymbol{x}^*)$ when fewer than T coordinates are non-zero. Note that the minimizer $\boldsymbol{x}^*$ of f^{sc} is given by $x_i^* = q^i$ where $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Then with fewer than T gradient oracles, for every $t \geq T$, the t-th coordinate of $\boldsymbol{x}$ is still zero. Therefore,

$$f^{\text{sc}}(\boldsymbol{x}) - f^{\text{sc}}(\boldsymbol{x}^*) \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \geq \frac{\mu}{2}\sum_{i=T}^{\infty}(x_i^*)^2 \geq \frac{\mu}{2}q^{2T}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2^2 .$$

Hence, to find a solution satisfying $f^{\text{sc}}(\boldsymbol{x}) - f^{\text{sc}}(\boldsymbol{x}^*) \leq \epsilon$, we need gradient complexity $T \geq \widetilde{\Omega}(\sqrt{\kappa})$.

4.2 Smooth nonconvex minimization

Another key component in our proof is to lower bound the complexity of nonconvex functions. Part of our construction is derived from Carmon et al. [5] who constructed the following unscaled function $\bar{f}^{\text{nc}} : \mathbb{R}^T \to \mathbb{R}$:

$$\bar{f}^{\text{nc}}(\boldsymbol{x}) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T}\left[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right], \tag{4}$$

where the component functions are

$$\Psi(x) := \begin{cases} 0 & x \leq 1/2 \\ \exp\left(1 - \frac{1}{(2x-1)^2}\right) & x > 1/2 \end{cases} \quad \text{and} \quad \Phi(x) := \sqrt{e}\int_{-\infty}^{x} e^{-\frac{1}{2}t^2}dt.$$

The above function also has the zero-chain property defined in 6. Following the general strategy discussed in the previous subsection, by appropriately rescaling $\bar{f}^{\text{nc}}$ so that it meets the requirement of the function class of interest, Carmon et al. [5] derived a lower bound of $T_{\text{nc}} := \Omega\left(1/\epsilon^2\right)$ gradient oracles. More details can be found in Appendix A.

5 Our construction

The hard examples discussed in the previous section show that one can easily construct an additive lower bound by making the objective as the sum of two functions f^{sc} and $\bar{f}^{\text{nc}}$. However, if we would like to improve the lower bound by a multiplicative factor in the condition number, it is far from obvious how one should compose f^{sc} with $\bar{f}^{\text{nc}}$. We describe our approach in the subsections below.

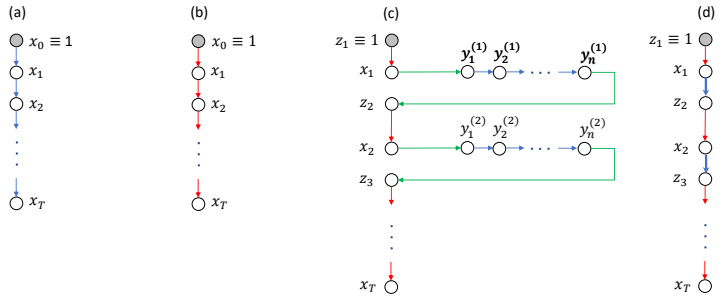

Figure 1: Chains (a, b, c, d) correspond to $f^{\text{sc}}, \bar{f}^{\text{nc}}, \bar{f}^{\text{nc-sc}}, \bar{f}_m^{\text{nc-sc}}$, respectively. Arrows in different colors represent different types of connections between two variables. For example, a blue arrow from u to v corresponds to a term with $(u-v)^2$ in the function. A red arrow corresponds to $\Psi(-u)\Phi(-v) - \Psi(u)\Phi(v)$. Green and bold blue arrows are quadratic connections that need to be carefully designed. The effective length of a bold blue arrow is n and that of all other arrows is 1.

5.1 Construction of the hard instance in the deterministic setting

Before discussing the details of our construction to prove Theorem 1, we first highlight the main difficulty and our approach at a high level to provide insight and intuition. First, we need to pinpoint the main difficulty.

The main difficulty. Given the aforementioned lower bound strategy, we need to find a zero-chain $\bar{f}^{\text{nc-sc}} \in \mathcal{F}(L, \mu, \Delta)$ as our hard instance. A natural way to construct $\bar{f}^{\text{nc-sc}}$ is to combine the ideas from the hard instances in [36] and [5]. The main challenge is how to find a good way of combination such that the two components do not interfere with each other's essential properties and that their strengths can be exploited multiplicatively to contribute to the lower bound.

The key idea. The novel combination structure we propose is illustrated in Figure 1. The body of the chain is like $\bar{f}^{\text{nc}}$ (4), which contains the minimization variables $\boldsymbol{x}, \boldsymbol{z}$. Then we insert $T-1$ f^{sc}-style (2) sub-chains with length n, which contain the maximization variable $\boldsymbol{y}$, into the main chain in the way shown in Figure 1(c). Figure 1(d) shows the function $\bar{f}_m^{\text{nc-sc}}(\cdot) := \max_{\boldsymbol{y}} \bar{f}^{\text{nc-sc}}(\cdot; \boldsymbol{y})$ obtained by maximizing $\boldsymbol{y}$. Note that the effective length of chain (d) is $\mathcal{O}(nT) = \mathcal{O}(\sqrt{\kappa}T)$, for the largest allowed choice of $n = \mathcal{O}(\sqrt{\kappa})$. Therefore, suppose we can show that $\bar{f}_m^{\text{nc-sc}}$ shares similar properties to $\bar{f}^{\text{nc}}$ used in [5], we obtain a lower bound of $\Omega\left(\sqrt{\kappa}T_{\text{nc}}\right) = \Omega\left(\sqrt{\kappa}/\epsilon^2\right)$.

However, the requirement that $\bar{f}_m^{\text{nc-sc}}$ behaves similarly to $\bar{f}^{\text{nc}}$ poses another challenge as it basically means that the red arrows in chain (d) dominate the bold blue ones, although their numbers are roughly equal. To overcome this challenge, we carefully design the green arrows directly and the bold blue arrows indirectly. We also need to restrict n such that $n = \mathcal{O}(\sqrt{\kappa})$.

Construction of the hard instance. The formal expression of $\bar{f}^{\text{nc-sc}} : (\mathbb{R}^T \times \mathbb{R}^{T-1}) \times \mathbb{R}^{n(T-1)} \to \mathbb{R}$ is given by

$$\bar{f}^{\text{nc-sc}}(\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{y}}) = -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T} \left[\Psi(-z_i)\Phi(-x_i) - \Psi(z_i)\Phi(x_i)\right] + \sum_{i=1}^{T-1} h(x_i, z_{i+1}; \bar{\boldsymbol{y}}^{(i)}) + \sum_{i=1}^{T-1} \left[c_1 x_i^2 + c_2 z_{i+1}^2\right], \tag{5}$$

where $\boldsymbol{x}, \boldsymbol{z}$ are the minimization variables and $\bar{\boldsymbol{y}} = [\bar{\boldsymbol{y}}^{(1)}, \ldots, \bar{\boldsymbol{y}}^{(T-1)}]$, where each $\bar{\boldsymbol{y}}^{(i)} \in \mathbb{R}^n$, is the maximization variable. The last term of (5) indirectly affects the bold blue arrows in Figure 1(d), where c_1 and c_2 are some parameters (not necessarily nonnegative) we choose later to obtain a much simpler expression to analyze. The function $h : (\mathbb{R} \times \mathbb{R}) \times \mathbb{R}^n \to \mathbb{R}$ is defined as

$$
\begin{aligned}
h(x,z;\boldsymbol{y}) :=& -\frac{1}{2}\sum_{i=1}^{n-1}(y_i-y_{i+1})^2 - \frac{1}{2n^2}\|\boldsymbol{y}\|_2^2 + \sqrt{\frac{C}{n}}\left(xy_1 - \frac{1}{2}zy_n\right) \\
=& -\frac{1}{2}\boldsymbol{y}^\top\left(\frac{1}{n^2}I_n + A\right)\boldsymbol{y} + \sqrt{\frac{C}{n}}\boldsymbol{b}_{x,z}^\top\boldsymbol{y},
\end{aligned} \tag{6}
$$

where $C > 0$ is a large enough numerical constant specified later and $\boldsymbol{b}_{x,z} := x\boldsymbol{e}_1 - \frac{1}{2}z\boldsymbol{e}_n$. Here the term $\sqrt{\frac{C}{n}}\boldsymbol{b}_{x,z}^\top\boldsymbol{y}$ characterizes the green arrows in Figure 1, where the $\sqrt{\frac{C}{n}}$ factor ensures $h_m(x,z) := \max_{\boldsymbol{y}\in\mathbb{R}^n} h(x,z;\boldsymbol{y})$ has an $\mathcal{O}(1)$ dependence on n as shown in Lemma 4. The matrix $A \in \mathbb{R}^{n\times n}$ is defined as

$$
A := \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}, \tag{7}
$$

which is the finite version of (3) in [36]'s instance except that its $(1,1)$-th and (n,n)-th entries are 1. This change is necessary in our example. Note that for fixed scalar variables x and z, $h(x,z;\cdot)$ is $\frac{1}{n^2}$-strongly-concave. Also, $\bar{f}^{\text{nc-sc}}$ is $\Omega(1)$-smooth (Lemma 9). Therefore, $\kappa = \Omega(n^2)$, implying we should choose $n = \mathcal{O}(\sqrt{\kappa})$. We can compute

$$
h_m(x,z) := \max_{\boldsymbol{y}\in\mathbb{R}^n} h(x,z;\boldsymbol{y}) = \frac{C}{2n}\boldsymbol{b}_{x,z}^\top\left(\frac{1}{n^2}I_n + A\right)^{-1}\boldsymbol{b}_{x,z}. \tag{8}
$$

The following lemma shows that h_m actually has a much simpler expression.

Lemma 4. *Suppose $n \geq 10$, we have*

$$
h_m(x,z) = C\left(\frac{a_1}{2}x^2 - \frac{1}{2}a_2xz + \frac{a_1}{8}z^2\right), \tag{9}
$$

where $a_1, a_2 > 0$ are "almost" numerical constants. That is to say, although a_1, a_2 depends on n, we have $0 < d_1 \leq a_1 \leq f_1$ and $0 < d_2 \leq a_2 \leq f_2$ and d_1, d_2, f_1, f_2 are numerical constants.

Choosing $c_1 = C(a_2 - a_1)/2$, $c_2 = C(a_2 - a_1)/8$ and $C = 12/a_2$, we obtain

$$
\begin{aligned}
\bar{f}_m^{\text{nc-sc}}(\boldsymbol{x},\boldsymbol{z}) :=& \max_{\bar{\boldsymbol{y}}\in\mathbb{R}^{n(T-1)}} \bar{f}^{\text{nc-sc}}(\boldsymbol{x},\boldsymbol{z};\bar{\boldsymbol{y}}) \\
=& -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T}\left[\Psi(-z_i)\Phi(-x_i) - \Psi(z_i)\Phi(x_i)\right] + 6\sum_{i=1}^{T-1}\left(x_i - \frac{1}{2}z_{i+1}\right)^2.
\end{aligned}
$$

We show in the appendix that $\bar{f}_m^{\text{nc-sc}}$ shares similar properties to $\bar{f}^{\text{nc}}$. Then we can prove Theorem 1 by appropriately rescaling $\bar{f}_m^{\text{nc-sc}}$ in the same way as done in [5]. The detailed proof is deferred in Appendix C.

5.2 Construction of the hard instance in the stochastic setting

We start this section by discussing why the techniques in [3] do not directly apply to our construction in the deterministic setting. The following lemma from [3] shows how to construct a probability-p zero-chain by constructing a stochastic first-order oracle over a given zero-chain.

Lemma 5 ([3, Lemma 3])**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a zero-chain on $\mathcal{X} \subset \mathbb{R}^T$. For $\boldsymbol{x} \in \mathcal{X}$, let $i^*(\boldsymbol{x}) := \inf\{i \in [T] : x_i = 0\}$ be the next coordinate to discover. For $p \in (0,1]$, define*

$$
[\boldsymbol{g}(\boldsymbol{x},\xi)]_i := \begin{cases} \frac{\xi}{p}\nabla_i f(\boldsymbol{x}) & \text{if } i = i^*(\boldsymbol{x}) \\ \nabla_i f(\boldsymbol{x}) & \text{otherwise,} \end{cases}
$$

where $\xi \sim \text{Bernoulli}(p)$. Suppose $\|\nabla f(\boldsymbol{x})\|_\infty \leq G$ for all $\boldsymbol{x} \in \mathcal{X}$. Then $O : \boldsymbol{x} \mapsto (f(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x},\xi))$ is a stochastic first-order oracle with bounded variance $\sigma^2 \leq G^2(1-p)/p$. Also, f equipped with O is a probability-p zero-chain.

Therefore, a tempting way to obtain our hard instance in the stochastic setting is to construct a probability-p zero-chain as in Lemma 5 directly over $\bar{f}^{\text{nc-sc}}$. However, as suggested by Lemma 5, to ensure the variance of the oracle is bounded, we require $\left\|\nabla \bar{f}^{\text{nc-sc}}(\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{y}})\right\|_\infty \leq G$ for some bounded $G < \infty$. Actually we need $G = \mathcal{O}(1)$ to obtain a nontrivial lower bound. However, $\bar{f}^{\text{nc-sc}}$ has an unconstrained quadratic component whose gradient is unbounded over the whole space. Therefore, we have to restrict its domain to be a bounded hypercube. However, it turns out to be impossible to find any hypercube such that $\left\|\nabla \bar{f}^{\text{nc-sc}}(\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{y}})\right\|_\infty \leq \mathcal{O}(1)$ without losing the properties of $\bar{f}_m^{\text{nc-sc}}$. To overcome this difficulty, we need to carefully modify the quadratic components in $\bar{f}^{\text{nc-sc}}$, i.e., the blue and green arrows in Figure 1(c).

Formally, define the hypercube $\mathcal{C}_R^d \subset \mathbb{R}^d$ as

$$\mathcal{C}_R^d = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_\infty \leq R\}.$$

Our hard instance in the stochastic setting, $\bar{f}^{\text{nc-sc-sg}} : (\mathcal{C}_{R_1}^T \times \mathcal{C}_{R_1}^{T-1}) \times \mathcal{C}_{nR_2}^{n(T-1)} \to \mathbb{R}$, where R_1 and R_2 are postive numerical constants to be set later, is given by

$$\begin{aligned}\bar{f}^{\text{nc-sc-sg}}(\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{y}}) = -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T} \left[\Psi(-z_i)\Phi(-x_i) - \Psi(z_i)\Phi(x_i)\right] \\ + \sum_{i=1}^{T-1} h^{\text{sg}}(x_i, z_{i+1}; \bar{\boldsymbol{y}}^{(i)}) + \sum_{i=1}^{T-1} \left[c_1 x_i^2 + c_2 z_{i+1}^2\right], \qquad (10)\end{aligned}$$

where

$$h^{\text{sg}}(x, z; \boldsymbol{y}) := \frac{C}{n}\left[-\frac{1}{2}\boldsymbol{y}^\top \left(\frac{1}{n^2} I_n + A\right)\boldsymbol{y} + \boldsymbol{b}_{x,z}^\top \boldsymbol{y}\right]. \qquad (11)$$

Here $C, c_1, c_2, \boldsymbol{b}_{x,z}$ are the same as those in (5).

By verifying that $\left\|\nabla \bar{f}^{\text{nc-sc-sg}}(\boldsymbol{x}, \boldsymbol{z}; \bar{\boldsymbol{y}})\right\|_\infty \leq \mathcal{O}(1)$ (Lemma 11.*iv*), we are able to construct a probability-p zero-chain as in Lemma 5 over the rescaled version of $\bar{f}^{\text{nc-sc}}$. According to Lemma 5, to ensure the variance of the rescaled stochastic oracle is bounded by σ^2, p has to be at least $\Omega(\epsilon^2/\sigma^2)$. Then we obtain a lower bound for the stochastic setting of

$$\Omega\left(\frac{nT_{\text{nc}}}{p}\right) = \Omega\left(\frac{\kappa^{1/3}}{\epsilon^4}\right).$$

Note that the deterministic lower bound is $\Omega(\frac{\sqrt{\kappa}}{\epsilon^2})$ which is a special case of the stochastic setting. Therefore we derive a lower bound of

$$\Omega\left(\max\left\{\frac{\sqrt{\kappa}}{\epsilon^2}, \frac{\kappa^{1/3}\sigma^2}{\epsilon^4}\right\}\right) = \Omega\left(\frac{\sqrt{\kappa}}{\epsilon^2} + \frac{\kappa^{1/3}}{\epsilon^4}\right).$$

Detailed analyses are deferred to Appendix D.

6 Conclusion and discussion

In this paper, we proved lower bounds on both the deterministic and the stochastic oracle complexities of nonconvex-strongly-concave min-max optimization for first-order zero-respecting algorithms. Our lower bound in the deterministic setting matches the existing upper bound [26] up to log factors. However, there is still a gap between our lower bound and the upper bound in [25] in the stochastic setting. How to close this gap is an open question. Apart from this, several other questions are worth consideration.

First, one immediate next step is to check if the proposed lower bounds hold for arbitrary, potentially randomized algorithms. We believe that the results are likely to hold but may introduce unexpected complications. Second, so far we have focused on nonconvex-strongly-concave min-max optimization. However, it remains open what the tight lower bound is in the more general nonconvex-concave min-max optimization. Moreover, what about nonconvex-nonconcave min-max optimization? To answer this question, a good measure of the suboptimality is a prerequisite. Last but not least, we only consider first-order oracles. It is also interesting to obtain a lower bound for functions with higher-order smoothness and oracles.

References

[1] Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, Alexander Gasnikov, and Dmitry Kovalev. Accelerated methods for composite non-bilinear saddle point problem. *arXiv preprint arXiv:1906.03620*, 2019.

[2] Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *International Conference on Machine Learning*, pages 908–916. PMLR, 2016.

[3] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

[5] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2019.

[6] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1-2), 2021.

[7] Antonin Chambolle and Thomas 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.

[8] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[9] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. *arXiv preprint arXiv:2009.09623*, 2020.

[10] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *arXiv preprint arXiv:2101.04233*, 2021.

[11] Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

[12] M. El-Mikkawy. On the inverse of a general tridiagonal matrix. *Appl. Math. Comput.*, 150: 669–679, 2004.

[13] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator, 2018.

[14] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, 2013.

[15] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[16] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*, 2019.

[20] Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

[21] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

[22] Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *arXiv preprint arXiv:1905.13433*, 2019.

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

[24] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Linear last-iterate convergence for matrix games and stochastic games. *arXiv preprint arXiv:2006.09517*, 2020.

[25] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[26] Tianyi Lin, Chi Jin, and Michael. I. Jordan. Near-optimal algorithms for minimax optimization, 2020.

[27] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[29] Oren Mangoubi and Nisheeth K. Vishnoi. Greedy adversarial equilibrium: An efficient alternative to nonconvex-nonconcave min-max optimization, 2020.

[30] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[31] Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

[32] Arkadi 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.

[33] Arkadi. S. Nemirovski and David. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[34] A. S. Nemirovsky. Information-based complexity of linear operator equations. *J. Complex.*, 8: 153–175, 1992.

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

[36] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[37] J. v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[38] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

[39] Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

[40] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pages 1–35, 2019.

[41] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[42] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *arXiv preprint arXiv:1311.1869*, 2013.

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

[44] Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *arXiv preprint arXiv:1907.01543*, 2019.

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

[46] Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. *arXiv preprint arXiv:1910.07512*, 2019.

[47] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. *arXiv preprint arXiv:2102.04540*, 2021.

[48] Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

[49] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[50] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems, 2020.

[51] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2103.15888*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work is of theoretical nature. The main contribution is to provide theoretical complexity lower bounds. There are no negative societal impacts from this work, and as a result we did not discuss it.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
(b) Did you include complete proofs of all theoretical results? [Yes] See Section 5 and Appendix

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]