What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

Chi Jin [1] **Praneeth Netrapalli** [2] **Michael I. Jordan** [3]

Abstract

Minimax optimization has found extensive applications in modern machine learning, in settings such as generative adversarial networks (GANs), adversarial training and multi-agent reinforcement learning. As most of these applications involve continuous nonconvex-nonconcave formulations, a very basic question arises—"what is a proper definition of local optima?"

Most previous work answers this question using classical notions of equilibria from *simultaneous* games, where the min-player and the max-player act simultaneously. In contrast, most applications in machine learning, including GANs and adversarial training, correspond to *sequential* games, where the order of which player acts first is crucial (since minimax is in general not equal to maximin due to the nonconvex-nonconcave nature of the problems). The main contribution of this paper is to propose a proper mathematical definition of local optimality for this sequential setting—*local minimax*, as well as to present its properties and existence results. Finally, we establish a strong connection to a basic local search algorithm—gradient descent ascent (GDA): under mild conditions, all stable limit points of GDA are exactly local minimax points up to some degenerate points.

1. Introduction

Minimax optimization refers to problems of two agents—one agent tries to minimize the payoff function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ while the other agent tries to maximize it. Such problems arise in a number of fields, including mathematics, biology, social science, and particularly economics (Myerson, 2013). Due to the wide range of applications of these problems and their rich mathematical structure, they have been studied for several decades in the setting of zero-sum games (Morgenstern and Von Neumann, 1953). In the last few years, minimax optimization has also found significant applications in machine learning, in settings such as generative adversarial networks (GAN) (Goodfellow et al., 2014), adversarial training (Madry et al., 2017) and multi-agent reinforcement learning (Omidshafiei et al., 2017). In practice, these minimax problems are often solved using gradient-based algorithms, especially gradient descent ascent (GDA), an algorithm that alternates between a gradient descent step for $\mathbf{x}$ and some number of gradient ascent steps for $\mathbf{y}$.

A well-known notion of optimality in this setting is that of a *Nash equilibrium*—no player can benefit by changing strategy while the other player keeps hers unchanged. That is, a Nash equilibrium is a point $(\mathbf{x}^\star, \mathbf{y}^\star)$ where $\mathbf{x}^\star$ is a global minimum of $f(\cdot, \mathbf{y}^\star)$ and $\mathbf{y}^\star$ is a global maximum of $f(\mathbf{x}^\star, \cdot)$. In the convex-concave setting, it can be shown that an approximate Nash equilibrium can be found efficiently by variants of GDA (Bubeck, 2015; Hazan, 2016). However, most of the minimax problems arising in modern machine learning applications do not have this simple convex-concave structure. Meanwhile, in the general nonconvex-nonconcave setting, one cannot expect to find Nash equilibria efficiently as the special case of nonconvex optimization is already NP-hard. This motivates the quest to find a local surrogate instead of a global optimal point. Most previous work (e.g., Daskalakis and Panageas, 2018; Mazumdar and Ratliff, 2018; Adolphs et al., 2018) studied a notion of *local Nash equilibrium* which replaces all the global minima or maxima in the definition of Nash equilibrium by their local counterparts.

The starting point of this paper is the observation that the notion of local Nash equilibrium is *not* suitable for most machine learning applications of minimax optimization. In fact, the notion of Nash equilibrium (on which local Nash equilibrium is based), was developed in the context of *simultaneous* games, and so it does not reflect the order between the min-player and the max-player. In contrast, most applications in machine learning, including GANs and adversarial training, correspond to *sequential* games, where one player acts first and the other acts second. When f is nonconvex-

[1] Princeton University [2] Microsoft Research, India [3] University of California, Berkeley. Correspondence to: Chi Jin <chij@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

nonconcave, $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ is in general not equal to $\max_{\mathbf{y}} \min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$; the order of which player acts first is crucial for the problem. This motivates the question:

What is a good notion of local optimality in nonconvex-nonconcave minimax optimization?

To answer this question, we start from the optimal solution $(\mathbf{x}^\star, \mathbf{y}^\star)$ for two-player sequential games $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}^\star$ is again the global maximum of $f(\mathbf{x}^\star, \cdot)$, but $\mathbf{x}^\star$ is now the global minimum of $\phi(\cdot)$, where $\phi(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. We call these optimal solutions *global minimax points*. The main contribution of this paper is to propose a proper mathematical definition of local optimality for this sequential setting—*local minimax*—a local surrogate for the global minimax points. This paper also presents existence results for this notion of optimality, and establishes several important properties of local minimax points. These properties naturally reflect the order of which player acts first, and alleviate many of the problematic issues of local Nash equilibria. Finally, the notion of local minimax provides a much stronger characterization of the asymptotic behavior of GDA—under certain idealized parameter settings, all stable limit points of GDA are exactly local minimax points up to some degenerate points. This provides, for the first time, a game-theoretic meaning for all of the stable limit points of GDA.

1.1. Our contributions

To summarize, this paper makes the following contributions.

- We clarify the difference between several notions of global and local optimality in the minimax optimization literature, in terms of definitions, settings, and properties (see Section 2 and Appendix A).
- We propose a new notion of local optimality—*local minimax*—a proper mathematical definition of local optimality for the two-player sequential setting. We also present properties of local minimax points and establish existence results (see Section 3.1 and 3.2).
- We establish a strong connection between local minimax points and the asymptotic behavior of GDA, and provide the first game-theoretic explanation of all stable limit points of GDA, up to some degenerate points (see Section 3.3).
- We provide a general framework and an efficiency guarantee for a special case where the maximization $\max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ can be solved efficiently for any fixed $\mathbf{x}$, or in general when an approximate max-oracle is present (see Appendix 4).

1.2. Related work

Minimax optimization: Since the seminal paper of (von Neumann, 1928), notions of equilibria in games and their algorithmic computation have received wide attention. In terms of algorithmic computation, the vast majority of results focus on the convex-concave setting (Korpelevich, 1976; Nemirovski and Yudin, 1978; Nemirovski, 2004). In the context of optimization, these problems have generally been studied in the setting of constrained convex optimization (Bertsekas, 2014). Results beyond convex-concave setting are much more recent. Rafique et al. (2018) and Nouiehed et al. (2019) consider nonconvex-but-concave minimax problems where for any $\mathbf{x}$, $f(\mathbf{x}, \cdot)$ is a concave function. In this case, they propose algorithms combining approximate maximization over $\mathbf{y}$ and a proximal gradient method for $\mathbf{x}$ to show convergence to stationary points. Lin et al. (2018) consider a special case of the nonconvex-nonconcave minimax problem, where the function $f(\cdot, \cdot)$ satisfies a variational inequality. In this setting, they consider a proximal algorithm that requires the solving of certain strong variational inequality problems in each step and show its convergence to stationary points. Hsieh et al. (2018) propose proximal methods that asymptotically converge to a *mixed* Nash equilibrium; i.e., a distribution rather than a point.

The most closely related prior work is that of Evtushenko (1974), who proposed a concept of "local" solution that is similar to the local minimax points proposed in this paper. Note, however, that Evtushenko's "local" notion is not a truly local property (i.e., cannot be determined just based on the function values in a small neighborhood of the given point). As a consequence, Evtushenko's definition does not satisfy the first-order and second-order necessary conditions of local minimax points (Proposition 18 and Proposition 19). We defer detailed comparison to Appendix B. Concurrent to our work, Fiez et al. (2019) also recognizes the important difference between simultaneous games and sequential games in the machine learning context, and proposes a local notion referred to as *Differential Stackelberg Equilibrium*, which implicitly assumes the Hessian for the second player to be nondegenerate,[1] in which case it is equivalent to a *strict* local minimax point (defined in Proposition 20). In contrast, we define a notion of local minimax point in a general setting, including the case in which Hessian matrices are degenerate. Finally, we consider GDA dynamics, which differ from the Stackelberg dynamics considered in Fiez et al. (2019).

[1] The definition in Fiez et al. (2019) implicitly assumes that the best response $r : \mathcal{X} \to \mathcal{Y}$ is well-defined by implicit equation $\nabla_{\mathbf{y}} f(x, r(x)) = 0$, and $g(x) := f(x, r(x))$ is differentiable with respect to x, conditions which do not always hold even if f is infinitely differentiable. These conditions are typically ensured by assuming $\nabla^2_{\mathbf{yy}} f \prec \mathbf{0}$.

GDA dynamics: There have been several lines of work studying GDA dynamics for minimax optimization. Cherukuri et al. (2017) investigate GDA dynamics under some strong conditions and show that the algorithm converges locally to Nash equilibria. Heusel et al. (2017) and Nagarajan and Kolter (2017) similarly impose strong assumptions in the setting of the training of GANs and show that under these conditions Nash equilibria are stable fixed points of GDA. Gidel et al. (2018) investigate the effect of simultaneous versus alternating gradient updates as well as the effect of momentum on the convergence in bilinear games. The analyses most closely related to ours are Mazumdar and Ratliff (2018) and Daskalakis and Panageas (2018). While Daskalakis and Panageas (2018) study minimax optimization (or zero-sum games), Mazumdar and Ratliff (2018) study a much more general setting of non-zero-sum games and multi-player games. Both of these papers show that the stable limit points of GDA are not necessarily Nash equilibria. Adolphs et al. (2018) and Mazumdar et al. (2019) propose Hessian-based algorithms whose stable fixed points are exactly Nash equilibria. We note that all the papers in this setting use Nash equilibrium as the notion of goodness.

Variational inequalities: Variational inequalities are generalizations of minimax optimization problems. The appropriate generalization of convex-concave minimax problems are known as monotone variational inequalities which have applications in the study of differential equations (Kinderlehrer and Stampacchia, 1980). There is a large literature on the design of efficient algorithms for finding solutions to monotone variational inequalities (Bruck, 1977; Nemirovski, 1981; 2004).

2. Preliminaries

In this section, we will first introduce our notation, and then present definitions and basic results for simultaneous games, sequential games, and general game-theoretic dynamics that are relevant to our work. This paper will focus on two-player zero-sum games. For clarity, we restrict our attention to *pure strategy* games in the main paper, that is, each player is restricted to play a single action as her strategy. We will present the discussion on their relations to *mixed strategy* games in Appendix A.

Notation. We use bold upper-case letters $\mathbf{A}, \mathbf{B}$ to denote matrices and bold lower-case letters $\mathbf{x}, \mathbf{y}$ to denote vectors. For vectors we use $\|\cdot\|$ to denote the ℓ_2-norm, and for matrices we use $\|\cdot\|$ and $\rho(\cdot)$ to denote spectral (or operator) norm and spectral radius (largest absolute value of eigenvalues) respectively. Note that these two are in general different for asymmetric matrices. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we use ∇f and $\nabla^2 f$ to denote its gradient and Hessian. For functions of two vector arguments, $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, we use $\nabla_{\mathbf{x}} f$, $\nabla_{\mathbf{y}} f$ and $\nabla^2_{\mathbf{xx}} f$, $\nabla^2_{\mathbf{xy}} f$, $\nabla^2_{\mathbf{yy}} f$ to denote its partial gradient and partial Hessian. We also use $O(\cdot)$ and $o(\cdot)$ notation as follows: $f(\delta) = O(\delta)$ means $\limsup_{\delta\to 0} |f(\delta)/\delta| \le C$ for some large absolute constant C, and $g(\delta) = o(\delta)$ means $\lim_{\delta\to 0} |g(\delta)/\delta| = 0$. For complex numbers, we use $\mathrm{Re}(\cdot)$ to denote its real part, and $|\cdot|$ to denote its modulus. We also use $\mathcal{P}(\cdot)$, operating over a set, to denote the collection of all probability measures over the set.

2.1. Simultaneous games

A *two-player zero-sum* game is a game of two players with a common payoff function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The function f maps the actions taken by both players $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ to a real value, which represents the gain of $\mathbf{y}$-player as well as the loss of $\mathbf{x}$-player. We also call $\mathbf{y}$ player *the max-player* who tries to maximize the payoff function f, and $\mathbf{x}$-player *the min-player*. In this paper, we focus on continuous payoff functions f, and assume $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$.

In simultaneous games, both players act simultaneously. That is, each player chooses her action without knowledge of the action chosen by other player. A well-known notion of optimality in game theory is Nash equilibrium, where no player can benefit by changing strategy while the other player keep hers unchanged. If we specialize this concept into this setting, we have:

Definition 1. Point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a **Nash equilibrium** of function f, if for any $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X} \times \mathcal{Y}$:

$$f(\mathbf{x}^\star, \mathbf{y}) \le f(\mathbf{x}^\star, \mathbf{y}^\star) \le f(\mathbf{x}, \mathbf{y}^\star).$$

That is, $\mathbf{x}^\star$ is a global minimum of $f(\cdot, \mathbf{y}^\star)$ which keeps the action of $\mathbf{y}$-player unchanged, and $\mathbf{y}^\star$ is a global maximum of $f(\mathbf{x}^\star, \cdot)$ which keeps the action of $\mathbf{x}$-player unchanged.

Classical works typically focus on finding the Nash equilirbria in the setting where the payoff function f is *convex-concave* (i.e. $f(\cdot, \mathbf{y})$ is convex for all $\mathbf{y} \in \mathcal{Y}$, and $f(\mathbf{x}, \cdot)$ is concave for all $\mathbf{x} \in \mathcal{X}$) (Bubeck, 2015). However, most modern applications in machine learning formulate payoff f as *nonconvex-nonconcave* functions, where the problem of finding global Nash equilibrium is NP hard in general. Therefore, recent work has considered the following local alternative (see, e.g., Mazumdar and Ratliff, 2018; Daskalakis and Panageas, 2018):

Definition 2. Point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a **local Nash equilirium** of f, if there exists $\delta > 0$ such that for any $(\mathbf{x}, \mathbf{y})$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \le \delta$ and $\|\mathbf{y} - \mathbf{y}^\star\| \le \delta$ we have:

$$f(\mathbf{x}^\star, \mathbf{y}) \le f(\mathbf{x}^\star, \mathbf{y}^\star) \le f(\mathbf{x}, \mathbf{y}^\star).$$

This essentially changes the requirements of global optimality in Nash equilibria to local optimality. Local Nash equilibria can be characterized in terms of first-order and second-order conditions.

Proposition 3 (First-order Necessary Condition)**.** *Assuming f is differentiable, any local Nash equilibrium satisfies $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$ and $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$.*

Proposition 4 (Second-order Necessary Condition)**.** *Assuming f is twice-differentiable, any local Nash equilibrium satisfies $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \preceq \mathbf{0}$, and $\nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \succeq \mathbf{0}$.*

Proposition 5 (Second-order Sufficient Condition)**.** *Assuming f is twice-differentiable, any stationary point (i.e., $\nabla f = 0$) satisfying the following condition is a local Nash equilibrium:*

$$\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}, \text{ and } \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \succ \mathbf{0}. \quad (1)$$

*We also call a stationary point satisfying (1) a **strict local Nash equilibrium**.*

One significant drawback of considering local or global Nash equilibria in nonconvex-nonconcave setting is that they may not exist even for simple well-behaved functions.

Proposition 6. *There exists a twice-differentiable function f, where pure strategy Nash equilibria (either local or global) do not exist.*

Proof. Consider a two-dimensional function $f(x, y) = \sin(x + y)$. We have $\nabla f(x, y) = (\cos(x + y), \cos(x + y))$. Assuming (x, y) is a local pure strategy Nash equilibrium, by Proposition 3 it must also be a stationary point; that is, $x + y = (k + 1/2)\pi$ for $k \in \mathbb{Z}$. It is easy to verify, for odd k, $\nabla^2_{xx} f(x, y) = \nabla^2_{yy} f(x, y) = 1 > 0$; for even k, $\nabla^2_{xx} f(x, y) = \nabla^2_{yy} f(x, y) = -1 < 0$. By Proposition 4, none of the stationary points is a local pure strategy Nash equilibrium. □

2.2. Sequential games

In sequential games, there is an intrinsic order that one player chooses her action before the other one chooses hers. Importantly, the second player can observe the action taken by the first player, and adjust her action accordingly. We would like to emphasize that although many recent works have focused on the simultaneous setting, GAN and adversarial training are in fact sequential games in their standard formulations.

Example 7 (Adversarial Training)**.** The target is to train a robust classifier that is robust to adversarial noise. The first player picks a classifier, and the second player then chooses adversarial noise to undermine the performance of the chosen classifier.

Example 8 (Generative Adversarial Network (GAN))**.** The target is to train a generator which can generate samples that are similar to real samples in the world. The first player picks a generator, and the second player then picks a discriminator that is capable of telling the difference between real samples and the samples generated by the chosen generator. Without loss of generality, we assume in this paper that min-player is the first player, and max-player is second. The objective of this game corresponds to following minimax optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \quad (2)$$

where $\min_{\mathbf{x}} \max_{\mathbf{y}}$ already reflects the intrinsic order of the sequential game. While this order does not matter for convex-concave $f(\cdot, \cdot)$ as the minimax theorem (Sion et al., 1958) guarantees that $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y})$, for a general nonconvex-nonconcave $f(\cdot, \cdot)$,

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

This means that the order of which player goes first plays an important role in the solution.

The global solution for Eq.(2), or *subgame perfect equilibrium* as known in the game theory literature, is for second player to always play the maximizer of $f(\mathbf{x}, \cdot)$ given the action $\mathbf{x}$ taken by the first player, and achieve the maximum value $\phi(\mathbf{x}) := \max_{y \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. Then, the optimal strategy for the first player is to minimize $\phi(\mathbf{x})$, which gives following definition of global optimality.

Definition 9. $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a **global minimax point**, if for any $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X} \times \mathcal{Y}$ we have:

$$f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}').$$

Remark 10. *Equivalently, $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a global minimax point if and only if $\mathbf{y}^\star$ is a global maximum of $f(\mathbf{x}^\star, \cdot)$, and $\mathbf{x}^\star$ is a global minimum of $\phi(\cdot)$ where $\phi(\mathbf{x}) := \max_{y \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$.*

Unlike Nash equilibria, global minimax points always exist even if f is nonconvex-nonconcave, due to the extreme-value theorem.

Proposition 11. *Assume that function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous, and assume that $\mathcal{X} \subset \mathbb{R}^{d_1}$, $\mathcal{Y} \subset \mathbb{R}^{d_2}$ are compact. Then the global minimax point of f always exists.*

Finding global minimax points of nonconvex-nonconcave function is also NP-hard in general. A practical solution is to find a local surrogate. Unfortunately, to the best of our knowledge, there is no formal definition of a local notion of global minimax points in the literature.

Finally, we would like to point out that, there is one easier case where the approximate maximization $\max_{y \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ can be solved efficiently for any $\mathbf{x} \in \mathcal{X}$. Then, Eq.(2) reduces to optimizing $\phi(\cdot)$—a nonsmooth nonconvex function, where efficient guarantees can be obtained (see Section 4).

2.3. Dynamical systems

One of the most popular algorithms for solving minimax problems is Gradient Descent Ascent (GDA). We outline the algorithm in Algorithm 1, with updates written in a general form $\mathbf{z}_{t+1} = \mathbf{w}(\mathbf{z}_t)$, where $\mathbf{w} : \mathbb{R}^d \to \mathbb{R}^d$ is a vector function. One notable distinction from standard gradient descent is that $\mathbf{w}(\cdot)$ may not be a gradient field (i.e., the gradient of a scalar function $\phi(\cdot)$), and so the Jacobian matrix $\mathbf{J} := \partial \mathbf{w}/\partial \mathbf{z}$ may be asymmetric. This results in the possibility of the dynamics $\mathbf{z}_{t+1} = \mathbf{w}(\mathbf{z}_t)$ converging to a limit cycle instead of a single point. Nevertheless, we can still define fixed points and stability for general dynamics.

Definition 12. Point $\mathbf{z}^\star$ is a **fixed point** for dynamical system $\mathbf{w}$ if $\mathbf{z}^\star = \mathbf{w}(\mathbf{z}^\star)$.

Definition 13 (Linear Stability)**.** For a differentiable dynamical system $\mathbf{w}$, a fixed point $\mathbf{z}^\star$ is a **linearly stable point** of $\mathbf{w}$ if its Jacobian matrix $\mathbf{J}(\mathbf{z}^\star) := (\partial \mathbf{w}/\partial \mathbf{z})(\mathbf{z}^\star)$ has spectral radius $\rho(\mathbf{J}(\mathbf{z}^\star)) \le 1$. We also say that a fixed point $\mathbf{z}^\star$ is a **strict linearly stable point** if $\rho(\mathbf{J}(\mathbf{z}^\star)) < 1$ and a **strict linearly unstable point** if $\rho(\mathbf{J}(\mathbf{z}^\star)) > 1$.

Intuitively, linear stability captures whether under the dynamics $\mathbf{z}_{t+1} = \mathbf{w}(\mathbf{z}_t)$ a flow that starts at point that is infinitesimally close to $\mathbf{z}^\star$ will remain in a small neighborhood around $\mathbf{z}^\star$.

3. Main Results

In the previous section, we pointed out that while many modern applications are in fact sequential games, the problem of finding their optima—global minimax points—is NP-hard in general. We now turn to our main results, which provide ways to circumvent this NP-hardness challenge. In Section 3.1, we develop a formal notion of local surrogacy for global minimax points which we refer to as *local minimax points*. In Section 3.2 we study their properties and existence. Finally, in Section 3.3, we establish a close relationship between stable fixed points of GDA and local minimax points.

3.1. Local minimax points

While most previous work (Daskalakis and Panageas, 2018; Mazumdar and Ratliff, 2018) has focused on local Nash equilibria (Definition 2), which are local surrogates for pure strategy Nash equilibria for simultaneous games, we propose a new notion—*local minimax*—as a natural local surrogate for global minimaxity (Definition 9) for sequential games. To the best of our knowledge, this is the first proper mathematical definition of local optimality for the two-player sequential setting.

Definition 14. A point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is said to be a **local minimax point** of f, if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, and any $(\mathbf{x}, \mathbf{y})$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \le \delta$ and $\|\mathbf{y} - \mathbf{y}^\star\| \le \delta$, we have

$$f(\mathbf{x}^\star, \mathbf{y}) \le f(\mathbf{x}^\star, \mathbf{y}^\star) \le \max_{\mathbf{y}' : \|\mathbf{y}' - \mathbf{y}^\star\| \le h(\delta)} f(\mathbf{x}, \mathbf{y}'). \quad (3)$$

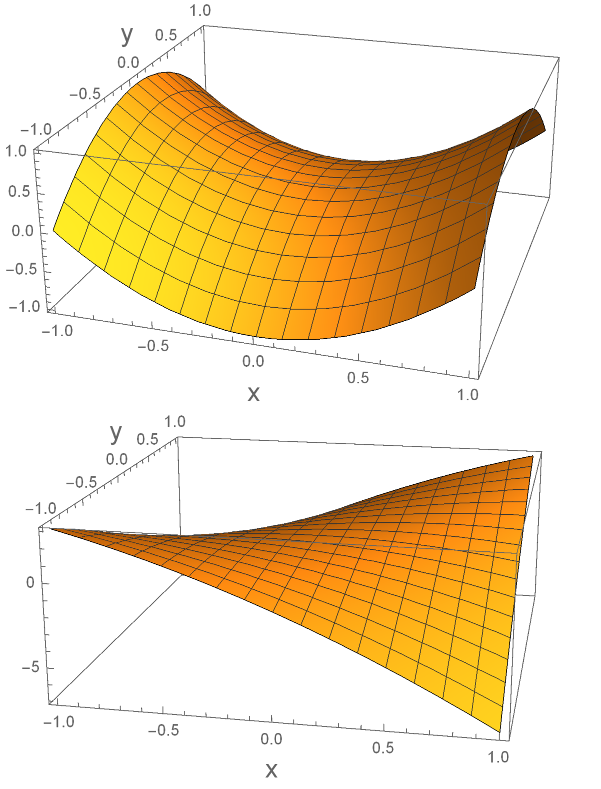

Figure 1. **Left:** $f(x, y) = x^2 - y^2$ where $(0, 0)$ is both local Nash and local minimax. **Right:** $f(x, y) = -x^2 + 5xy - y^2$ where $(0, 0)$ is not local Nash but local minimax with $h(\delta) = \delta$.

Remark 15. *Definition 14 remains equivalent even if we further restrict function h in Definition 14 to be monotonic or continuous. See Appendix C for more details.*

Intuitively, local minimaxity captures the optimal strategies in a two-player sequential game if both players are only allowed to change their strategies locally.

Definition 14 localize the notion of global minimax points (Definition 9) by replacing all global optimality over $\mathbf{x}$ and $\mathbf{y}$ by local optimality. However, since this is a sequential setting, the radius of the local neighborhoods where the maximization or minimization takes over can be different. Definition 14 allows one radius to be δ while the other is $h(\delta)$. The introduction of an arbitrary function h allows the ratio of these two radii to also be arbitrary. The limiting behavior $h(\delta) \to 0$ as $\delta \to 0$ makes this definition a truly local notion. That is, it only depends on the property of function f in an infinitesimal neighborhood around $(\mathbf{x}^\star, \mathbf{y}^\star)$.

Definition 14 is a natural local surrogate for global minimax points. We can alternatively define local minimax points as localized versions of the equivalent characterization of global minimax points as in Remark 10. It turns out that two definitions are equivalent.

Lemma 16. *For a continuous function f, a point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a local minimax point of f if and only if $\mathbf{y}^\star$ is a local maximum of function $f(\cdot, \mathbf{x}^\star)$, and there exists an $\epsilon_0 > 0$ such that $\mathbf{x}^\star$ is a local minimum of function g_ϵ for all $\epsilon \in (0, \epsilon_0]$ where function g_ϵ is defined as $g_\epsilon(\mathbf{x}) := \max_{\mathbf{y}:\|\mathbf{y}-\mathbf{y}^\star\|\le\epsilon} f(\mathbf{x}, \mathbf{y})$.*

Lemma 16 states that local minimaxity can be viewed from a game-theoretic perspective: the second player always plays the action to achieve a local maximum value $g_\epsilon(\mathbf{x}) := \max_{\mathbf{y}:\|\mathbf{y}-\mathbf{y}^\star\|\le\epsilon} f(\mathbf{x}, \mathbf{y})$, for infinitesimal ϵ, given the action $\mathbf{x}$ taken the first player, and the first player minimizes $g_\epsilon(\mathbf{x})$ locally.

Finally, it can be shown that local minimaxity is a weakening of the notion of local Nash equilibrium defined as in Definition 2. It alleviates the non-existence issues of the latter.

Proposition 17. *Any local Nash equilibrium (Definition 2) is a local minimax point.*

3.2. Properties and existence

Local minimax points also enjoy simple first-order and second-order characterizations. Notably, the second-order conditions naturally reflect the order of the sequential game (who plays first).

Proposition 18 (First-order Necessary Condition)**.** *Assuming that f is continuously differentiable, then any local minimax point $(\mathbf{x}, \mathbf{y})$ satisfies $\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) = 0$ and $\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) = 0$.*

Proposition 19 (Second-order Necessary Condition)**.** *Assuming that f is twice differentiable, then $(\mathbf{x}, \mathbf{y})$ is a local minimax point implies that $\nabla^2_{\mathbf{y}\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \preceq \mathbf{0}$. Furthermore, if $\nabla^2_{\mathbf{y}\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}$, then*

$$[\nabla^2_{\mathbf{x}\mathbf{x}} f - \nabla^2_{\mathbf{x}\mathbf{y}} f (\nabla^2_{\mathbf{y}\mathbf{y}} f)^{-1} \nabla^2_{\mathbf{y}\mathbf{x}} f](\mathbf{x}, \mathbf{y}) \succeq \mathbf{0}. \qquad (4)$$

Proposition 20 (Second-order Sufficient Condition)**.** *Assume that f is twice differentiable. Any stationary point $(\mathbf{x}, \mathbf{y})$ satisfying $\nabla^2_{\mathbf{y}\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}$ and*

$$[\nabla^2_{\mathbf{x}\mathbf{x}} f - \nabla^2_{\mathbf{x}\mathbf{y}} f (\nabla^2_{\mathbf{y}\mathbf{y}} f)^{-1} \nabla^2_{\mathbf{y}\mathbf{x}} f](\mathbf{x}, \mathbf{y}) \succ \mathbf{0} \qquad (5)$$

is a local minimax point. We call stationary points satisfying (5) strict local minimax points.

We note that if $\nabla^2_{\mathbf{y}\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ is non-degenerate, then the second-order necessary condition (Proposition 19) becomes $\nabla^2_{\mathbf{y}\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}$ and $[\nabla^2_{\mathbf{x}\mathbf{x}} f - \nabla^2_{\mathbf{x}\mathbf{y}} f (\nabla^2_{\mathbf{y}\mathbf{y}} f)^{-1} \nabla^2_{\mathbf{y}\mathbf{x}} f](\mathbf{x}, \mathbf{y}) \succeq \mathbf{0}$, which is identical to the sufficient condition in Eq. (5) up to an equals sign.

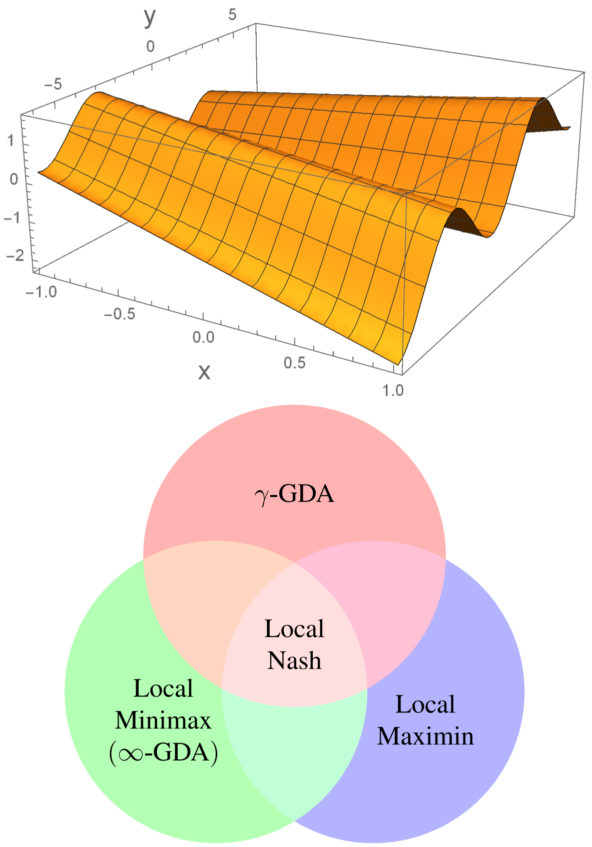

Figure 2. **Left:** $f(x, y) = 0.2xy - cos(y)$, the global minimax points $(0, -\pi)$ and $(0, \pi)$ are not stationary. **Right:** The relations among local Nash equilibria, local minimax points, local maximin points and linearly stable points of γ-GDA, and ∞-GDA (up to degenerate points).

Comparing Eq. (5) to the second-order sufficient condition for local Nash equilibrium in Eq. (1), we see that, instead of requiring $\nabla^2_{\mathbf{x}\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ to be positive definite, local minimaxity requires the Shur complement to be positive definite. Contrary to local Nash equilibria, this characterization of local minimaxity not only takes into account the interaction term $\nabla^2_{\mathbf{x}\mathbf{y}} f$ between $\mathbf{x}$ and $\mathbf{y}$, but also reflects the difference between the first player and the second player.

For existence, we would like to first highlight an interesting fact: in contrast to the well-known fact in nonconvex optimization that global minima are always local minima (thus local minima always exist), global minimax points can be neither local minimax nor even stationary points.

Proposition 21. *The global minimax point can be neither local minimax nor a stationary point.*

Algorithm 1 Gradient Descent Ascent (γ-GDA)

```
Input: $(\mathbf{x}_0, \mathbf{y}_0)$, step size $\eta$, ratio $\gamma$.
  for $t = 0, 1, \ldots,$ do
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - (\eta/\gamma)\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.
    $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)$.
  end for
```

See Figure 2 for an illustration and Appendix C for the proof. The proposition is a natural consequence of the definitions where global minimax points are obtained as a minimum of a *global* maximum function while local minimax points are the minimum of a *local* maximum function. This also illustrates that minimax optimization is a challenging task, and worthy of independent study, beyond nonconvex optimization.

Therefore, although global minimax points always exist as in Proposition 11, it is not necessary for local minimax points to always exist. Unfortunately, similar to local Nash equilibria, local minimax points may not exist in general.

Lemma 22. *There exists a twice-differentiable function f and a compact domain, where local minimax points do not exist.*

Nevertheless, global minimax points can be guaranteed to be local minimax under some further regularity. For instance, this is true when f is strongly-concave in $\mathbf{y}$, or more generally when f satisfies the following properties that have been established in several machine learning problems (Ge et al., 2017; Boumal et al., 2016). There, local minimax points are guaranteed to exist.

Theorem 23. *Assume that f is twice differentiable, and for any fixed $\mathbf{x}$, the function $f(\mathbf{x}, \cdot)$ is strongly concave in the neighborhood of local maxima and satisfies the assumption that all local maxima are global maxima. Then the global minimax point of $f(\cdot, \cdot)$ is also a local minimax point.*

3.3. Relation to the limit points of GDA

In this section, we consider the asymptotic behavior of Gradient Descent Ascent (GDA), and its relation to local minimax points. As shown in the pseudo-code in Algorithm 1, GDA simultaneously performs gradient descent on $\mathbf{x}$ and gradient ascent on $\mathbf{y}$. We consider the general form where the step size for $\mathbf{x}$ can be different from the step size for $\mathbf{y}$ by a ratio γ, and denoted this algorithm by γ-GDA. When the step size η is small, this is essentially equivalent to the algorithm that alternates between one step of gradient descent and γ steps of gradient ascent.

To study the limiting behavior, we primarily focus on linearly stable points of γ-GDA, since with random initialization, γ-GDA will almost surely escape strict linearly unstable points.

Theorem 24 ((Daskalakis and Panageas, 2018))**.** *For any $\gamma > 1$, assuming the function f is ℓ-gradient Lipschitz, and the step size $\eta \le 1/\ell$, then the set of initial points $\mathbf{x}_0$ so that γ-GDA converges to its strict linear unstable point is of Lebesgue measure zero.*

We further simplifiy the problem by considering the limiting case where the step size $\eta \to 0$, which corresponds to γ-GDA flow. We note the asympototic behavior of γ-GDA flow is essentially the same as γ-GDA with very small step size η up to certain error tolerance.

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\frac{1}{\gamma}\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \qquad \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}).$$

The strict linearly stable points of the γ-GDA flow have a very simple second-order characterization.

Proposition 25. *Point $(\mathbf{x}, \mathbf{y})$ is a strict linearly stable point of γ-GDA if and only if for all the eigenvalues $\{\lambda_i\}$ of following Jacobian matrix,*

$$\mathbf{J}_\gamma = \begin{pmatrix} -(1/\gamma)\nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) & -(1/\gamma)\nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}), \end{pmatrix}$$

their real part $\mathrm{Re}(\lambda_i) < 0$ for any i.

In the remainder of this section, we assume that f is a twice-differentiable function, and we use $\mathcal{Local\_Nash}$ to represent the set of strict local Nash equilibria, $\mathcal{Local\_Minimax}$ for the set of strict local minimax points, $\mathcal{Local\_Maximin}$ for the set of strict local maximin points, and $\gamma-\mathcal{GDA}$ for the set of strict linearly stable points of the γ-GDA flow. Our goal is to understand the relationship between these sets. Daskalakis and Panageas (2018) and Mazumdar and Ratliff (2018) provided a relation between $\mathcal{Local\_Nash}$ and $1-\mathcal{GDA}$ which can be generalized to $\gamma-\mathcal{GDA}$ as follows.

Proposition 26 ((Daskalakis and Panageas, 2018))**.** *For any fixed γ, for any twice-differentiable f, $\mathcal{Local\_Nash} \subset \gamma-\mathcal{GDA}$, but there exist twice-differentiable f such that $\gamma-\mathcal{GDA} \not\subset \mathcal{Local\_Nash}$.*

That is, if γ-GDA converges, it may converge to points not in $\mathcal{Local\_Nash}$. This raises a basic question as to what those additional stable limit points of γ-GDA are. Are they meaningful? This paper answers this question through the lens of $\mathcal{Local\_Minimax}$. Although for fixed γ, the set $\gamma-\mathcal{GDA}$ does not have a simple relation with $\mathcal{Local\_Minimax}$, it turns out that an important relationship arises when γ goes to ∞. To describe the limit behavior of the set $\gamma-\mathcal{GDA}$ when $\gamma \to \infty$ we define two set-theoretic limits:

$$\overline{\infty-\mathcal{GDA}} := \limsup_{\gamma\to\infty} \gamma-\mathcal{GDA} = \cap_{\gamma_0>0} \cup_{\gamma>\gamma_0} \gamma-\mathcal{GDA}$$

$$\underline{\infty-\mathcal{GDA}} := \liminf_{\gamma\to\infty} \gamma-\mathcal{GDA} = \cup_{\gamma_0>0} \cap_{\gamma>\gamma_0} \gamma-\mathcal{GDA}.$$

The relations between $\gamma-\mathcal{GDA}$ and $\mathcal{Local\_Minimax}$ are given as follows:

Algorithm 2 Gradient Descent with Max-oracle

```
Input: $\mathbf{x}_0$, step size $\eta$.
  for $t = 0, 1, \ldots, T$ do
    find $\mathbf{y}_t$ so that $f(\mathbf{x}_t, \mathbf{y}_t) \geq \max_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}) - \epsilon$.
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$.
  end for
  Pick $t$ uniformly at random from $\{0, \cdots, T\}$.
  return $\bar{\mathbf{x}} \leftarrow \mathbf{x}_t$.
```

Proposition 27. *For any fixed γ, there exists a twice-differentiable f such that $\mathcal{L}ocal\_\mathcal{M}inimax \not\subset \gamma-\mathcal{GDA}$; there also exists a twice-differentiable f such that $\gamma-\mathcal{GDA} \not\subset \mathcal{L}ocal\_\mathcal{M}inimax \cup \mathcal{L}ocal\_\mathcal{M}aximin$.*

Theorem 28 (Asymptotic Behavior of ∞-GDA)**.** *For any twice-differentiable f, $\mathcal{L}ocal\_\mathcal{M}inimax \subset \underline{\infty-\mathcal{GDA}} \subset \overline{\infty-\mathcal{GDA}} \subset \mathcal{L}ocal\_\mathcal{M}inimax \cup \{(\mathbf{x}, \mathbf{y}) | (\mathbf{x}, \mathbf{y})$ is stationary and $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y})$ is degenerate$\}$.*

That is, $\infty-\mathcal{GDA} = \mathcal{L}ocal\_\mathcal{M}inimax$ up to some degenerate points. Intuitively, when γ is large, γ-GDA can move a long distance in $\mathbf{y}$ while only making very small changes in $\mathbf{x}$. As $\gamma \to \infty$, γ-GDA can find the "approximate local maximum" of $f(\mathbf{x} + \delta_{\mathbf{x}}, \cdot)$, subject to any small change in $\delta_{\mathbf{x}}$; therefore, stable limit points are indeed local minimax.

Algorithmically, one can view $\infty-\mathcal{GDA}$ as a set that describes the strict linear stable limit points for GDA with γ very slowly increasing with respect to t, and eventually going to ∞. To the best of our knowledge, this is the first result showing that all stable limit points of GDA are meaningful and locally optimal up to some degenerate points.

4. Gradient Descent with Max-Oracle

In this section, we consider solving the minimax problem Eq.(2) when we have access to an oracle for approximate inner maximization; i.e., for any $\mathbf{x}$, we have access to an oracle that outputs a $\widehat{\mathbf{y}}$ such that $f(\mathbf{x}, \widehat{\mathbf{y}}) \geq \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \epsilon$. A natural algorithm to consider in this setting is to alternate between gradient descent on $\mathbf{x}$ and a (approximate) maximization step on $\mathbf{y}$. The pseudocode is presented in Algorithm 2.

It can be shown that Algorithm 2 indeed converges (in contrast with GDA which can converge to limit cycles). Moreover, the limit points of Algorithm 2 satisfy a nice property—they turn out to be approximately stationary points of $\phi(\mathbf{x}) := \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. For a smooth function, "approximately stationary point" means that the norm of gradient is small. However, even when $f(\cdot, \cdot)$ is smooth (up to whatever order), $\phi(\cdot)$ as defined above need not be differentiable. The norm of subgradient can be a discontinuous function which is an undesirable measure for closeness to stationarity. Fortunately, however, we have following structure.

Fact 29. *If function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is ℓ-gradient Lipschitz, then function $\phi(\cdot) := \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ is ℓ-**weakly convex** ; that is, $\phi(\mathbf{x}) + (\ell/2)\|\mathbf{x}\|^2$ is convex function over $\mathbf{x}$.*

The above fact has been also presented in Rafique et al. (2018). In such settings, the approximate stationarity of $\phi(\cdot)$ can be measured by the norm of gradient of its Moreau envelope $\phi_\lambda(\cdot)$.

$$\phi_\lambda(\mathbf{x}) := \min_{\mathbf{x}'} \phi(\mathbf{x}') + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{x}'\|^2. \qquad (6)$$

Here $\lambda < 1/\ell$ is the parameter. The Moreau envelope has the following important property that connects it to the original function ϕ.

Lemma 30 ((Rockafellar, 2015))**.** *Assume function ϕ is ℓ-weakly convex. Let $\lambda < 1/\ell$, and denote $\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}'} \phi(\mathbf{x}') + (1/2\lambda)\|\mathbf{x} - \mathbf{x}'\|^2$. Then $\|\nabla \phi_\lambda(\mathbf{x})\| \leq \epsilon$ implies:*

$$\|\hat{\mathbf{x}} - \mathbf{x}\| = \lambda\epsilon, \quad \text{and} \quad \min_{\mathbf{g} \in \partial\phi(\hat{\mathbf{x}})} \|\mathbf{g}\| \leq \epsilon.$$

where ∂ denotes the subdifferential of a weakly convex function.

Lemma 30 says, $\|\nabla \phi_\lambda(\mathbf{x})\|$ being small means that $\mathbf{x}$ is close to a point $\hat{\mathbf{x}}$ that is a approximately stationary point of original function ϕ. We now present the convergence guarantee for Algorithm 2.

Theorem 31. *Suppose f is ℓ-smooth and L-Lipschitz and define $\phi(\cdot) := \max_{\mathbf{y}} f(\cdot, \mathbf{y})$. Then the output $\bar{\mathbf{x}}$ of GD with Max-oracle (Algorithm 2) with step size $\eta = \gamma/\sqrt{T+1}$ will satisfy*

$$\begin{aligned}&\mathbb{E}\left[\|\nabla \phi_{1/2\ell}(\bar{\mathbf{x}})\|^2\right] \\ &\quad \leq 2 \cdot \frac{\left(\phi_{1/2\ell}(\mathbf{x}_0) - \min \phi(\mathbf{x})\right) + \ell L^2 \gamma^2}{\gamma\sqrt{T+1}} + 4\ell\epsilon,\end{aligned}$$

where $\phi_{1/2\ell}$ is the Moreau envelope (6) of ϕ.

The proof of Theorem 31 is similar to the convergence analysis for nonsmooth weakly-convex functions (Davis and Drusvyatskiy, 2018), except here the max-oracle has error ϵ. Theorem 31 claims, other than an additive error $4\ell\epsilon$ as a result of the oracle solving the maximum approximately, that the remaining term decreases at a rate of $1/\sqrt{T}$. We present the full proof of Theorem in Appendix F.

5. Conclusion

In this paper, we consider general nonconvex-nonconcave minimax optimization problems. Since most these problems arising in modern machine learning correspond to sequential games, we propose a new notion of local optimality—*local minimax*—the first proper mathematical definition of local optimality for the two-player sequential setting. We present favorable results on their properties and existence. We also establish a strong connection to GDA: up to some degenerate points, local minimax points are exactly equal to the stable limit points of GDA.

Acknowledgements

The authors would like to thank Guojun Zhang and Yaoliang Yu for raising a technical issue with Proposition 19 in an earlier version of this paper, Oleg Burdakov for pointing out the related work by Yu. G. Evtushenko, and Lillian J. Ratliff for helpful discussions. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. *arXiv preprint arXiv:1805.05751*, 2018.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANS). *arXiv preprint arXiv:1703.00573*, 2017.

Dimitri P Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 2014.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

Ronald E Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Ashish Cherukuri, Bahman Gharesifard, and Jorge Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9256–9266, 2018.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{\frac{-1}{4}})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

Yurii Gavrilovich Evtushenko. Some local properties of minimax problems. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 14(3):669–679, 1974.

Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in Stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Remi Lepriol, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018.

Irving L Glicksberg. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002*, 2018.

David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and their Applications*, volume 31. SIAM, 1980.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

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.

Eric Mazumdar and Lillian J Ratliff. On the convergence of gradient-based learning in continuous games. *arXiv preprint arXiv:1804.05464*, 2018.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Oskar Morgenstern and John Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

Roger B Myerson. *Game Theory*. Harvard University Press, 2013.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5585–5595, 2017.

Arkadi Nemirovski. Efficient methods for solving variational inequalities. *Ekonomika i Matem. Metody*, 17: 344–359, 1981.

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.

Arkadi Nemirovski and D. Yudin. Cesari convergence of the gradient method for approximation saddle points of convex-concave functions. *Doklady AN SSSRv*, 239:1056–1059, 1978.

Maher Nouiehed, Maziar Sanjabi, 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.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *arXiv preprint arXiv:1703.06182*, 2017.

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.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

Maurice Sion et al. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

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

Mishael Zedek. Continuity and location of zeros of linear combinations of polynomials. *Proceedings of the American Mathematical Society*, 16(1):78–84, 1965.