Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks*

Jie Jiang† ‡ and Xiaojun Chen§

Abstract. This paper considers a class of nonsmooth nonconvex-nonconcave min-max problems in machine learning and games. We first provide sufficient conditions for the existence of global minimax points and local minimax points. Next, we establish the first-order and second-order optimality conditions for local minimax points by using directional derivatives. These conditions reduce to smooth min-max problems with Fréchet derivatives. We apply our theoretical results to generative adversarial networks (GANs) in which two neural networks contest with each other in a game. Examples are used to illustrate applications of the new theory for training GANs.

Key words. min-max problem, nonsmooth, nonconvex-nonconcave, optimality condition, generative adversarial networks

AMS subject classifications. 90C47, 90C15, 90C33, 65K15

1. Introduction.

Consider the following min-max problem

$$\min_{x\in X}\max_{y\in Y}\ f(x,y), \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ are nonempty, closed and convex sets, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a locally Lipschitz continuous function. Define an envelope function

$$\varphi(x) := \max_{y\in Y} f(x,y).$$

In this paper, we assume that $\varphi(x)$ is finite-valued for any $x \in X$. We say problem (1.1) is nonconvex-nonconcave if for a fixed $x \in X$, $f(x,\cdot)$ is not concave, and for a fixed $y \in Y$, $f(\cdot,y)$ is not convex.

The min-max problem (1.1) has many applications in machine learning and games [20, 30, 35], for instance, the popular generative adversarial networks (GANs) in machine learning [2, 9, 16, 17, 26]. Let $D : \mathbb{R}^m \times \mathbb{R}^{s_1} \to (0,1)$ be a parameterized discriminator, $G : \mathbb{R}^n \times \mathbb{R}^{s_2} \to \mathbb{R}^{s_1}$ be a parameterized generator and ξ_i be a s_i-valued random vector with probability distribution P_i and support $\Xi_i \subseteq \mathbb{R}^{s_i}$ for $i = 1, 2$. Then the plain vanilla GAN model can be formulated as

$$\min_{x\in X}\max_{y\in Y}\ \mathbb{E}_{P_1}\Big[\log\big(D(y,\xi_1)\big)\Big] + \mathbb{E}_{P_2}\Big[\log\Big(1 - D\big(y, G(x,\xi_2)\big)\Big)\Big], \tag{1.2}$$

---

*Revision 1 February 2023

Funding: This work is supported by The Hong Kong Polytechnic University Post-doctoral Fellow Scheme and The Hong Kong Grant Council Grant PolyU15300120.

†College of Mathematics and Statistics, Chongqing University, Chongqing, China (jiangjiecq@163.com).

‡CAS AMSS-PolyU Joint Laboratory of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China.

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (xiaojun.chen@polyu.edu.hk).

where x and y are the parameters to control D and G with ranges X and Y, respectively. Here $\mathbb{E}_{P_i}[\cdot]$ denotes the expected value with probability distribution P_i over Ξ_i for $i = 1, 2$. We assume that the expected values are finite for any fixed $x \in X$ and $y \in Y$. Since the range of D is $(0, 1)$, for any fixed x,

$$\varphi(x) = \max_{y \in Y} \ \mathbb{E}_{P_1}\Big[\log\big(D(y, \xi_1)\big)\Big] + \mathbb{E}_{P_2}\Big[\log\Big(1 - D\big(y, G(x, \xi_2)\big)\Big)\Big]$$

is real-valued. The functions D and G are usually defined by deep neural networks (see section 4 for a specific example). It is noteworthy that unconstrained min-max problems for training GANs are widely used, while constrained min-max problems are also used for improved GANs, Wasserstein GANs and some games. One can refer to [2, 3, 19] for more details.

Since the pioneering work [29] by Von Neumann in 1928, convex-concave min-max problems have been investigated extensively, based on the concept of saddle points (see e.g. [6, 28, 35, 36] and the references therein). In recent years, driving by important applications, nonconvex-nonconcave min-max problems have attracted considerable attention [21, 22, 24, 31]. However, it is well-known that a nonconvex-nonconcave min-max problem may not have a saddle point. How to properly define its local optimal points and optimality conditions has been of great concern. In [1, 12, 25], the concept of local saddle points was studied, but it is pointed out in [21] that the concept of local saddle points is not suitable for most applications of min-max optimization in machine learning. A nonconvex-nonconcave min-max problem may not have a local saddle point (see Example 2.7 in this paper). In [21], the authors argued that a local solution cannot be determined just based on the function value in an arbitrary small neighborhood of a given point. For that reason, they proposed the concept of local minimax points of unconstrained smooth nonconvex-nonconcave min-max problems and studied the first-order and second-order optimality conditions.

Optimality conditions for minimization problems have been extensively studied [7, 32]. Moreover, the study of optimality conditions for simultaneous games has a long history, whose solutions are commonly described as the Nash equilibrium. According to the definition of Nash equilibrium, the optimality conditions are the combination of each player's optimality condition when the rivals' decisions are fixed. Therefore, optimality conditions for simultaneous games can be viewed as an extension of those for minimization problems. For more details, one can refer to [4, 7, 14, 27, 32]. However, optimality and stationarity of nonsmooth nonconvex-nonconcave min-max problems are not well understood. Necessary optimality conditions for unconstrained weakly-convex–concave min-max problems and their application in machine learning were studied in [23, 31]. In [21], from the viewpoint of sequential games, the local minimax points and the first-order and second-order optimality conditions for unconstrained smooth nonconvex-nonconcave min-max problems were defined. Based on the concept of the local minimax points proposed in [21], necessary and sufficient optimality conditions for the local minimax points of constrained smooth min-max problems were studied in [11]. It is worth noting that the min-max problem can be viewed as a specific bi-level optimization problem. The general practice to solve a bi-level optimization problem is to replace the lower level optimization by its first-order optimality conditions, so that the bi-level optimization problem becomes a mathematical programming with equilibrium constraints (MPEC) and its optimality conditions are derived based on the MPEC formulation [13]. However, optimality conditions for global/local minimax points of nonsmooth bi-level problems where the upper level problem is nonconvex and the lower level problem is nonconcave have not been studied yet.

The main contributions of this paper can be summarized as follows.

- We define the first-order and second-order optimality conditions of local minimax points of constrained min-max problem (1.1) by using directional derivatives. Our optimality conditions extend the work [21] for unconstrained smooth min-max problems to constrained nonsmooth min-max problems. These conditions reduce to smooth min-max problems with Fréchet derivatives. Moreover, we rigorously describe the relationships between saddle points, local saddle points, global minimax points, local minimax points and stationary points defined by these first-order and second-order optimality conditions. The relationships among these points is illustrated by interesting examples and summarized in Figure 1.
- We establish new mathematical optimization theory for the GAN model with both smooth and nonsmooth activation functions. In particular, we give new properties of global minimax points, local minimax points and stationary points of problem (1.2) under some specific settings. Examples with the sample average approximation approach show that our results are helpful and efficient for training GANs.

The reminder of the paper is organized as follows. In section 2, we give some notations and preliminaries. In section 3, we study the first-order and second-order optimality conditions of nonsmooth and smooth min-max problems, respectively. In section 4, we apply our results to GANs and use examples to show the effectiveness of our results. Finally, we make some concluding remarks in section 5.

2. Notations and preliminaries.

In this paper, $\mathbb{N}$ denotes the natural numbers. $\mathbb{R}^n_+$ denotes the nonnegative part of $\mathbb{R}^n$. $\|\cdot\|$ denotes the Euclidean norm. $\mathrm{cl}(\Omega)$, $\mathrm{int}(\Omega)$ and $\mathrm{bd}(\Omega)$ denote the closure, the interior and the boundary of set Ω, respectively. $o(|t|)$ denotes the infinitesimal of a higher order than $|t|$ as $t \to 0$. $O(|t|)$ denotes the same order as $|t|$ as $t \to 0$. $\mathbb{B}(x, r)$ denotes the closed ball centred at x with radius $r > 0$. Denote $(\cdot)_+ := \max\{0, \cdot\}$ the ReLU activation function. The indicator function of a set Ω is denoted by δ_Ω, i.e., $\delta_\Omega(x) = 0$ if $x \in \Omega$ and $\delta_\Omega(x) = \infty$ otherwise. The extended-valued functions are functions that are allowed to be extended-real-valued, i.e., to take values in $\mathbb{R} \cup \{\pm\infty\}$.

Let $\Omega \subseteq \mathbb{R}^n$ be a closed and convex set. The tangent cone [32, Definition 6.1] to Ω at $x \in \Omega$, denoted by $\mathcal{T}_\Omega(x)$, is defined as $\mathcal{T}_\Omega(x) = \left\{ w : \exists\, x^k \xrightarrow{\Omega} x, t^k \downarrow 0 \text{ such that } \lim_{k\to\infty} \frac{x^k - x}{t^k} = w \right\}$.

The normal cone [32, Definition 6.3] to Ω at $x \in \Omega$, denoted by $\mathcal{N}_\Omega(x)$, is

$$\mathcal{N}_\Omega(x) := \{y \in \mathbb{R}^n : \langle y, \omega - x\rangle \le 0, \forall \omega \in \Omega\}.$$

It also knows from [32, Proposition 6.5] that $\mathcal{N}_\Omega(x) = \{v : \langle v, \omega\rangle \le 0, \text{ for } \forall \omega \in \mathcal{T}_\Omega(x)\}$.

Definition 2.1. *We say that $(\hat{x}, \hat{y}) \in X \times Y$ is a saddle point of problem (1.1), if*

$$f(\hat{x}, y) \le f(\hat{x}, \hat{y}) \le f(x, \hat{y}) \tag{2.1}$$

holds for any $(x, y) \in X \times Y$.

Definition 2.2. *We say that* $(\hat{x}, \hat{y}) \in X \times Y$ *is a local saddle point of problem* (1.1)*, if there exists a* $\delta > 0$ *such that, for any* $(x, y) \in X \times Y$ *satisfying* $\|x - \hat{x}\| \le \delta$ *and* $\|y - \hat{y}\| \le \delta$*,* (2.1) *holds.*

In the convex-concave setting, saddle points are usually used to describe the optimality of min-max problems. However, one significant drawback of considering (local) saddle points of nonconvex-nonconcave problems is that such points might not exist [21, Proposition 6]. Also, (local) saddle points correspond to simultaneous game, but many applications (such as GANs and adversarial training) correspond to sequential games. In view of this, we consider in what follows global and local minimax points proposed in [21], which are from the viewpoint of sequential games.

Definition 2.3. *We say that* $(\hat{x}, \hat{y}) \in X \times Y$ *is a global minimax point of problem* (1.1)*, if*

$$f(\hat{x}, y) \le f(\hat{x}, \hat{y}) \le \max_{y' \in Y} f(x, y')$$

holds for any $(x, y) \in X \times Y$*.*

Definition 2.4. *We say that* $(\hat{x}, \hat{y}) \in X \times Y$ *is a local minimax point of problem* (1.1)*, if there exist a* $\delta_0 > 0$ *and a function* $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ *satisfying* $\tau(\delta) \to 0$ *as* $\delta \to 0$*, such that for any* $\delta \in (0, \delta_0]$ *and any* $(x, y) \in X \times Y$ *satisfying* $\|x - \hat{x}\| \le \delta$ *and* $\|y - \hat{y}\| \le \delta$*, we have*

$$f(\hat{x}, y) \le f(\hat{x}, \hat{y}) \le \max_{y' \in \{y \in Y : \|y - \hat{y}\| \le \tau(\delta)\}} f(x, y').$$

Remark 2.5. It is noteworthy that the function τ in Definition 2.4 can be further restricted to be monotone or continuous without changing Definition 2.4 [21, Remark 15]. Hereafter, we always assume that τ is monotone and continuous.

Global or local minimax points are motivated by many practical applications and the probable nonconvexity-nonconcavity of the min-max problem. Obviously, a saddle point is a global minimax point and a local saddle point is a local minimax point. However, problem (1.1) may not have a local saddle point. The following proposition gives some sufficient conditions for the existence of global (local) minimax points. Note that the existence of a global (local) minimax point does not imply the existence of a local saddle point.

Proposition 2.6.
(i) If $\Phi_u := \{x \in X : \varphi(x) \le u\}$ *is nonempty and bounded for some scalar* u *and* $\{y \in Y : f(x, y) \ge l_x\}$ *is bounded for every* $x \in \Phi_u$ *and some scalar* l_x*, then problem* (1.1) *has at least a global minimax point.*
(ii) ([21, Lemma 16]) $(x^*, y^*) \in X \times Y$ *is a local minimax point if and only if* y^* *is a local maximum of* $f(x^*, \cdot)$ *and there exists a* $\delta_0 > 0$ *such that* x^* *is a local minimum of* $\varphi_\delta(x) := \max_{y' \in \{y \in Y : \|y - y^*\| \le \delta\}} f(x, y')$ *for any* $\delta \in (0, \delta_0]$*.*

Proof. (i) According to the continuity of $f(x, y)$, φ is lower semicontinuous. We know from [32, Theorem 1.9] that $\arg\min_{x \in X} \varphi(x) \subseteq \Phi_u$ is nonempty and compact. Let $x^* \in \arg\min_{x \in X} \varphi(x)$ and consider the set $\arg\max_{y \in Y} f(x^*, y)$. Since $\{y \in Y : f(x^*, y) \ge l_{x^*}\}$ is bounded, we know from the continuity of $f(x^*, \cdot)$ that the maximum can be achieved. Let $y^* \in \arg\max_{y \in Y} f(x^*, y)$. It is easy to check that (x^*, y^*) is a global minimax point. ∎

Specifically, if both X and Y are bounded, then all conditions in (i) of Proposition 2.6 hold. Thus problem (1.1) has a global minimax point. However, a local minimax point may not exist even X and Y are bounded (see Example 3.24). Also, a global minimax point may not be a local minimax point (see Example 3.24). The following example tells that the global and local minimax points exist but (local) saddle points do not.

Example 2.7 ([21, Figure 1]). Let $n = m = 1$ and $X = Y = [-1, 1]$. Consider $f(x, y) = -x^2 + 5xy - y^2$. Note that

$$\varphi(x) = \max_{y\in[-1,1]} (-x^2 + 5xy - y^2) = \begin{cases} -x^2 - 5x - 1, & x \in \left[-1, -\frac{2}{5}\right]; \\ \frac{21}{4}x^2, & x \in \left[-\frac{2}{5}, \frac{2}{5}\right]; \\ -x^2 + 5x - 1, & x \in \left[\frac{2}{5}, 1\right]. \end{cases}$$

It is not difficult to examine that $\min_{x\in[-1,1]} \varphi(x) = 0$ when $x = 0$. In this case, $y = 0$. Therefore, $(0, 0)$ is a global minimax point. Moreover, let $\delta_0 = \frac{2}{5}$ and $\tau(\delta) = \frac{5}{2}\delta$ in Definition 2.4. Then for any $\delta \le \delta_0$, $(x, y) \in [-1, 1] \times [-1, 1]$ satisfying $|x| \le \delta$ and $|y| \le \delta$, we have

$$\max_{y'\in\{y\in Y:|y|\le\frac{5}{2}\delta\}} f(x, y') = \frac{21}{4}x^2$$

when $y = \frac{5}{2}x$. Thus, we obtain

$$-y^2 = f(0, y) \le f(0, 0) = 0 \le \max_{y'\in\{y\in Y:|y|\le\frac{5}{2}\delta\}} f(x, y') = \frac{21}{4}x^2,$$

which implies that $(0, 0)$ is also a local minimax point.

Note that the solutions of $\max_{y\in[-\delta,\delta]} \min_{x\in[-\delta,\delta]} f(x, y)$ are $(\delta, 0)$ and $(-\delta, 0)$ for any $\delta \in (0, 1]$. Thus, we have

$$\max_{y\in[-\delta,\delta]} \min_{x\in[-\delta,\delta]} f(x, y) = -\delta^2 \ne 0 = \min_{x\in[-\delta,\delta]} \max_{y\in[-\delta,\delta]} f(x, y), \tag{2.2}$$

which implies that $(0, 0)$ is neither a saddle point (i.e., (2.2) holds with $\delta = 1$, see Definition 2.1) nor a local saddle point (i.e., (2.2) holds with a sufficiently small δ, see Definition 2.2).

Example 2.7 gives a nonconvex-nonconcave min-max problem that has global and local minimax points, but does not have a local saddle point. Thus, global and local minimax points defined in Definitions 2.3 and 2.4 respectively are good supplements of (local) saddle points, especially in the nonconvex-nonconcave setting.

3. Optimality and stationarity.

In this section, we first discuss the first-order and second-order optimality conditions when f in problem (1.1) is nonsmooth. The smooth case is considered as a special case of the nonsmooth ones when the directional derivatives can be represented by Fréchet derivatives. Our results extend the study of necessary optimality conditions of unconstrained smooth min-max problems in [21]. In particular, in the nonsmooth case, our results extend [21] from unconstrained smooth ones to constrained nonsmooth ones and in the smooth case, our results extend [21] from unconstrained ones to constrained ones. We also illustrate these theoretical results by three examples.

To proceed further, we give the description of tangents to convex sets.

Lemma 3.1 ([32, Theorem 6.9]). *If $\Omega \subseteq \mathbb{R}^n$ is convex and $\bar{x} \in \Omega$, then*

$$\mathcal{T}_\Omega(\bar{x}) = \mathrm{cl}\{w : \exists \lambda > 0 \text{ with } \bar{x} + \lambda w \in \Omega\},\ \mathrm{int}\,(\mathcal{T}_\Omega(\bar{x})) = \{w : \exists \lambda > 0 \text{ with } \bar{x} + \lambda w \in \mathrm{int}(\Omega)\}.$$

Denote

$$\mathcal{T}^\circ_\Omega(\bar{x}) := \{w : \exists \lambda > 0 \text{ with } \bar{x} + \lambda w \in \Omega\}.$$

It is not difficult to verify that $\mathcal{T}_\Omega(\bar{x})$, $\mathrm{int}(\mathcal{T}_\Omega(\bar{x}))$ and $\mathcal{T}^\circ_\Omega(\bar{x})$ are convex cones if Ω is convex. Moreover, we have the following relationship $\mathrm{int}\,(\mathcal{T}_\Omega(\bar{x})) \subseteq \mathcal{T}^\circ_\Omega(\bar{x}) \subseteq \mathcal{T}_\Omega(\bar{x})$. If Ω is polyhedral, then $\mathcal{T}^\circ_\Omega(\bar{x}) = \mathcal{T}_\Omega(\bar{x})$.

3.1. Nonsmooth case. In this subsection, we consider problem (1.1) when f is not differentiable. For this purpose, we introduce some definitions for nonsmooth analysis.

Let $g : \mathbb{R}^n \to \mathbb{R}$. The (first-order) *subderivative* $\mathrm{d}g(x)(v)$ at $x \in \mathbb{R}^n$ for $v \in \mathbb{R}^n$ is defined as [32, Definition 8.1]

$$\mathrm{d}g(x)(v) := \liminf_{v' \to v, t \downarrow 0} \frac{g(x + tv') - g(x)}{t}.$$

The function g is *semidifferentiable* at x for v [32, Definition 7.20] if the (possibly infinite) limit

$$\lim_{v' \to v, t \downarrow 0} \frac{g(x + tv') - g(x)}{t}$$

exists. Further, if the above limit exists for every $v \in \mathbb{R}^n$, we say that g is semidifferentiable at x. It is easy to see that if g is Lipschitz continuous in a neighborhood of x, then this limit is finite.

There are two types of second-order subderivatives [32, Definition 13.3]. The second-order subderivative at $x \in \mathbb{R}^n$ for w and v is

$$\mathrm{d}^2 g(x|v)(w) := \liminf_{w' \to w, t \downarrow 0} \frac{g(x + tw') - g(x) - t\,\langle v, w'\rangle}{\frac{1}{2}t^2}.$$

The second-order subderivative at $x \in \mathbb{R}^n$ for w (without mention of v) is

$$\mathrm{d}^2 g(x)(w) := \liminf_{w' \to w, t \downarrow 0} \frac{g(x + tw') - g(x) - t\mathrm{d}g(x)(w')}{\frac{1}{2}t^2}.$$

We say that g is *twice semidifferentiable* at x if it is semidifferentiable at x and the (possibly infinite) limit

$$\lim_{w' \to w, t \downarrow 0} \frac{g(x + tw') - g(x) - t\mathrm{d}g(x)(w')}{\frac{1}{2}t^2}$$

exists for any $w \in \mathbb{R}^n$.

The one-side *directional derivative* $g'(x; v)$ at $x \in \mathbb{R}^n$ along the direction $v \in \mathbb{R}^n$ is defined as

$$g'(x; v) := \lim_{t \downarrow 0} \frac{g(x + tv) - g(x)}{t}.$$

The function g is directionally differentiable at x if $g'(x;v)$ exists for all directions $v \in \mathbb{R}^n$. If g is locally Lipschitz continuous near x, then semidifferentiability at x is equivalent to directional differentiability at x.

The *second-order directional derivative* of g at $x \in \mathbb{R}^n$ along the direction $v \in \mathbb{R}^n$ is defined as [32, Chapter 13.B]

$$g^{(2)}(x;v) := \lim_{t\downarrow 0} \frac{g(x+tv) - g(x) - tg'(x;v)}{\frac{1}{2}t^2}.$$

Obviously, if g is semidifferentiable at x, then $\mathrm{d}g(x)(v) = g'(x;v)$; if g is twice semidifferentiable at x, then $\mathrm{d}^2 g(x)(w) = g^{(2)}(x;w)$.

As a generalization of classical directional derivatives, *the (Clarke) generalized directional derivative* of g at $x \in \mathbb{R}^n$ along the direction $v \in \mathbb{R}^n$ is defined as [7, Section 2.1]

$$g^{\circ}(x;v) := \limsup_{x'\to x, t\downarrow 0} \frac{g(x'+tv) - g(x')}{t}.$$

We say that g is *Clarke regular* at x [7, Definition 2.3.4] if $g'(x;v)$ exists and $g^{\circ}(x;v) = g'(x;v)$ for all v. By using the generalized directional derivative, we can define the (Clarke) *generalized subdifferential* as

$$\partial g(x) := \{z \in \mathbb{R}^n : \langle z, v\rangle \le g^{\circ}(x;v)\ \forall v \in \mathbb{R}^n\}.$$

In turn, we know from [7, page 10] that

$$g^{\circ}(x;v) = \max\{\langle \zeta, v\rangle : \zeta \in \partial g(x)\}. \tag{3.1}$$

The *generalized second-order directional derivative* of g at $x \in \mathbb{R}^n$ along the direction $(u,v) \in \mathbb{R}^n \times \mathbb{R}^n$ is defined as ([8, Definition 1.1] and [32, Theorem 13.52])

$$g^{\circ\circ}(x;u,v) := \limsup_{\substack{x'\to x\\ t\downarrow 0,\delta\downarrow 0}} \frac{g(x'+\delta u+tv) - g(x'+\delta u) - g(x'+tv) + g(x')}{\delta t}.$$

Especially, when $u = v$, we write $g^{\circ\circ}(x;v,v)$ as $g^{\circ\circ}(x;v)$ for simplicity.

Remark 3.2. When f is continuously differentiable at $(\hat{x},\hat{y})$, $f_x^{\circ}(\hat{x},\hat{y};v) = \mathrm{d}_x f(\hat{x},\hat{y})(v) = \nabla_x f(\hat{x},\hat{y})^{\top} v$ and $f_y^{\circ}(\hat{x},\hat{y};w) = \mathrm{d}_y f(\hat{x},\hat{y})(w) = \nabla_y f(\hat{x},\hat{y})^{\top} w$ ([32, Exercise 8.20]). Moreover, if f is twice continuously differentiable at $(\hat{x},\hat{y})$, we know from [32, Example 13.8, Proposition 13.56] that $f_x^{\circ\circ}(\hat{x},\hat{y};v) = \mathrm{d}_x^2 f(\hat{x},\hat{y})(v) = v^{\top}\nabla_x^2 f(\hat{x},\hat{y})v$ and $f_y^{\circ\circ}(\hat{x},\hat{y};w) = \mathrm{d}_y^2 f(\hat{x},\hat{y})(w) = w^{\top}\nabla_y^2 f(\hat{x},\hat{y})w$.

Example 3.3. Consider a two-layer neural network with the ReLU activation function as follows:

$$F(W,b) := \rho(W_2(W_1\xi + b_1)_+ + b_2)$$

for a fixed $\xi \in \mathbb{R}^s$, where $W_1 \in \mathbb{R}^{s_1\times s}$, $b_1 \in \mathbb{R}^{s_1}$, $W_2 \in \mathbb{R}^{s_2\times s_1}$, $b_2 \in \mathbb{R}^{s_2}$, $\rho : \mathbb{R}^{s_2} \to \mathbb{R}$ is a continuously differentiable function, $W = (W_1, W_2)$ and $b = (b_1, b_2)$. Obviously, F is locally

Lipschitz continuous. For fixed $\overline{W} = (\overline{W}_1, \overline{W}_2)$ and $\bar{b} = (\bar{b}_1, \bar{b}_2)$, we consider

$$
\begin{aligned}
F'(W, b; \overline{W}, \bar{b}) &= \lim_{t\downarrow 0} \frac{F(W + t\overline{W}, b + t\bar{b}) - F(W, b)}{t} \\
&= \lim_{t\downarrow 0} \frac{\rho((W_2 + t\overline{W}_2)((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ + b_2 + t\bar{b}_2) - \rho(W_2(W_1\xi + b_1)_+ + b_2)}{t}
\end{aligned}
$$

and

$$
\begin{aligned}
&\lim_{t\downarrow 0} \frac{(W_2 + t\overline{W}_2)((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ + b_2 + t\bar{b}_2 - (W_2(W_1\xi + b_1)_+ + b_2)}{t} \\
&= \lim_{t\downarrow 0} \frac{W_2\left(((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ - (W_1\xi + b_1)_+\right) + t\left(\overline{W}_2((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ + \bar{b}_2\right)}{t} \\
&= W_2\left(\lim_{t\downarrow 0} \frac{((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ - (W_1\xi + b_1)_+}{t}\right) + \overline{W}_2(W_1\xi + b_1)_+ + \bar{b}_2.
\end{aligned}
$$

For $i = 1, \cdots, s_1$, denote $\overline{W}_1^i$ and W_1^i the ith row vectors of $\overline{W}_1$ and W_1, and $\bar{b}_1^i$ and b_1^i the ith components of $\bar{b}_1$ and b_1, respectively. Then, for $i = 1, \cdots, s_1$ and sufficiently small $t > 0$, we have

$$
\begin{aligned}
&((W_1^i + t\overline{W}_1^i)^\top\xi + b_1^i + t\bar{b}_1^i)_+ - ((W_1^i)^\top\xi + b_1^i)_+ \\
&\qquad = \begin{cases}
t(\overline{W}_1^i)^\top\xi + t\bar{b}_1^i, & \text{if } (W_1^i)^\top\xi + b_1^i > 0; \\
0, & \text{if } (W_1^i)^\top\xi + b_1^i < 0; \\
t(\overline{W}_1^i)^\top\xi + t\bar{b}_1^i, & \text{if } (W_1^i)^\top\xi + b_1^i = 0 \text{ and } (\overline{W}_1^i)^\top\xi + \bar{b}_1^i > 0; \\
0, & \text{if } (W_1^i)^\top\xi + b_1^i = 0 \text{ and } (\overline{W}_1^i)^\top\xi + \bar{b}_1^i \le 0.
\end{cases}
\end{aligned}
$$

Hence we obtain

$$
\begin{aligned}
&\lim_{t\downarrow 0} \frac{((W_1^i + t\overline{W}_1^i)^\top\xi + b_1^i + t\bar{b}_1^i)_+ - ((W_1^i)^\top\xi + b_1^i)_+}{t} \\
&= \begin{cases}
(\overline{W}_1^i)^\top\xi + \bar{b}_1^i, & \text{if } (W_1^i)^\top\xi + b_1^i > 0 \text{ or } (W_1^i)^\top\xi + b_1^i = 0 \text{ and } (\overline{W}_1^i)^\top\xi + \bar{b}_1^i > 0; \\
0, & \text{if } (W_1^i)^\top\xi + b_1^i < 0 \text{ or } (W_1^i)^\top\xi + b_1^i = 0 \text{ and } (\overline{W}_1^i)^\top\xi + \bar{b}_1^i \le 0.
\end{cases}
\end{aligned}
$$

Thus, we have that the following limit

$$
\Upsilon := W_2\left(\lim_{t\downarrow 0} \frac{((W_1 + t\overline{W}_1)\xi + b_1 + t\bar{b}_1)_+ - (W_1\xi + b_1)_+}{t}\right) + \overline{W}_2(W_1\xi + b_1)_+ + \bar{b}_2
$$

exists. Therefore, we have that F is semidifferentiable based on the locally Lipschitz continuity.

If, moreover, ρ is twice continuously differentiable, we have

$$
\begin{aligned}
\mathrm{d}^2F(W, b)(\overline{W}, \bar{b}) &= \liminf_{\substack{t\downarrow 0 \\ \overline{W}'\to\overline{W}, \bar{b}'\to\bar{b}}} \frac{F(W + t\overline{W}', b + t\bar{b}') - F(W, b) - t\mathrm{d}F(W, b)(\overline{W}', \bar{b}')}{\frac{1}{2}t^2} \\
&= \Upsilon^\top \nabla^2\rho(W_2(W_1\xi + b_1)_+ + b_2)\Upsilon,
\end{aligned}
$$

which implies that F is twice semidifferentiable.

The following lemma tells the necessary optimality conditions for an unconstrained minimization problem by using subderivatives.

Lemma 3.4 ([32, Theorems 10.1 & 13.24]). *Let $g : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper extended-valued function. If $\bar{x}$ is a local minimum of g over $\mathbb{R}^n$, then $\mathrm{d}g(\bar{x})(v) \geq 0$ and $\mathrm{d}^2 g(\bar{x}|0)(v) \geq 0$ for any $v \in \mathbb{R}^n$.*

The following lemma shows that we can replace $\mathrm{d}^2 g(\bar{x}|0)(v) \geq 0$ by $\mathrm{d}^2 g(\bar{x})(v) \geq 0$ under certain mild conditions.

Lemma 3.5. *Let $g : \mathbb{R}^n \to (-\infty, +\infty]$ be twice semidifferentiable at $\bar{x}$. If $\mathrm{d}g(\bar{x})(v) = 0$, then $\mathrm{d}^2 g(\bar{x}|0)(v) = \mathrm{d}^2 g(\bar{x})(v)$.*

Proof. Let $\mathrm{d}g(\bar{x})(v) = 0$. Note that

$$\begin{aligned}\mathrm{d}^2 g(\bar{x})(v) &= \liminf_{v' \to v, t \downarrow 0} \frac{g(\bar{x} + tv') - g(\bar{x}) - t\mathrm{d}g(\bar{x})(v')}{\frac{1}{2}t^2} = \lim_{v' \to v, t \downarrow 0} \frac{g(\bar{x} + tv') - g(\bar{x}) - t\mathrm{d}g(\bar{x})(v')}{\frac{1}{2}t^2} \\ &= \lim_{t \downarrow 0} \frac{g(\bar{x} + tv) - g(\bar{x}) - t\mathrm{d}g(\bar{x})(v)}{\frac{1}{2}t^2} = \lim_{t \downarrow 0} \frac{g(\bar{x} + tv) - g(\bar{x})}{\frac{1}{2}t^2} = \mathrm{d}^2 g(x|0)(v),\end{aligned}$$

where the second equality follows from the twice semidifferentiability of g at $\bar{x}$ and the third equality follows from the existence of the limit. ■

Lemma 3.6 ([32, Theorem 8.2]). *For the indicator function $\delta_{\mathcal{X}}$ of a set $\mathcal{X} \subseteq \mathbb{R}^n$ and any point $x \in \mathcal{X}$, one has $\mathrm{d}\delta_{\mathcal{X}}(x)(v) = \delta_{\mathcal{T}_{\mathcal{X}}(x)}(v)$ for any $v \in \mathbb{R}^n$.*

A function $g : \mathbb{R}^n \to \mathbb{R}$ is called *positively homogeneous of degree $p > 0$* if $g(\lambda w) = \lambda^p g(w)$ for all $\lambda > 0$ and $w \in \mathbb{R}^n$ (see [32, Definition 13.4]).

The following lemma shows the expansion of a function via subderivatives.

Lemma 3.7 ([32, Theorem 7.21 & Exercise 13.7]). *Let $g : \mathbb{R}^n \to \mathbb{R}$. Then*

(i) g is semidifferentiable at $\bar{x}$ if and only if

$$g(x) = g(\bar{x}) + \mathrm{d}g(\bar{x})(x - \bar{x}) + o(\|x - \bar{x}\|),$$

where $\mathrm{d}g(\bar{x})(\cdot)$ is a finite, continuous, positively homogeneous function.

(ii) Suppose that g is semidifferentiable at $\bar{x}$. Then g is twice semidifferentiable at $\bar{x}$ if and only if

$$g(x) = g(\bar{x}) + \mathrm{d}g(\bar{x})(x - \bar{x}) + \frac{1}{2}\mathrm{d}^2 g(\bar{x})(x - \bar{x}) + o(\|x - \bar{x}\|^2),$$

where $\mathrm{d}^2 g(\bar{x})(\cdot)$ is a finite, continuous, positively homogeneous of degree 2 function.

The following lemma gives the first-order and second-order optimality conditions for minimizing a semidifferentiable function, which extends a sub-result of [10, Proposition 2.3] from a polyhedral set to a general convex and closed set.

Lemma 3.8. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed and convex set, $g : \mathbb{R}^n \to \mathbb{R}$ be semidifferentiable at $\bar{x} \in \mathcal{X}$, and $\bar{x}$ be a local minimum point of g over $\mathcal{X}$. Then $\mathrm{d}g(\bar{x})(v) \geq 0$ for all $v \in \mathcal{T}_{\mathcal{X}}(\bar{x})$. Moreover, if g is twice semidifferentiable at $\bar{x}$, then $\mathrm{d}^2 g(\bar{x})(v) \geq 0$ for all $v \in \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \cap \{v :$ $\mathrm{d}g(\bar{x})(v) = 0\}$.*

Proof. Since $\bar{x}$ is a local minimum point of g over $\mathcal{X}$, we know from Lemma 3.4 that $\mathrm{d}\bar{g}(\bar{x})(v) \geq 0$ and $\mathrm{d}^2\bar{g}(\bar{x}|0)(v) \geq 0$ for any $v \in \mathbb{R}^n$, where $\bar{g} = g + \delta_{\mathcal{X}}$. From Lemma 3.6, we have for all $v \in \mathcal{T}_{\mathcal{X}}(\bar{x})$ that

$$\begin{aligned} 0 \leq \mathrm{d}\bar{g}(\bar{x})(v) &= \liminf_{v' \to v, t \downarrow 0} \frac{g(\bar{x} + tv') - g(\bar{x}) + \delta_{\mathcal{X}}(\bar{x} + tv') - \delta_{\mathcal{X}}(\bar{x})}{t} \\ &= \liminf_{v' \to v, t \downarrow 0} \frac{g(\bar{x} + tv') - g(\bar{x})}{t} = \mathrm{d}g(\bar{x})(v), \end{aligned}$$

where the second equality follows from the observation that $\delta_{\mathcal{X}}(\bar{x}) = 0$ due to $\bar{x} \in \mathcal{X}$ and v' is selected such that $\delta_{\mathcal{X}}(\bar{x} + tv') = 0$ (see Lemma 3.1) for sufficient small t to achieve the limit inferior.

Based on the above results, for $v \in \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \subseteq \mathcal{T}_{\mathcal{X}}(\bar{x})$, $\mathrm{d}g(\bar{x})(v) = 0$ if and only if $\mathrm{d}\bar{g}(\bar{x})(v) = 0$. Thus, $\mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \cap \{v : \mathrm{d}g(\bar{x})(v) = 0\} = \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \cap \{v : \mathrm{d}\bar{g}(\bar{x})(v) = 0\}$.

We know from Lemma 3.5 that for $v \in \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \cap \{v : \mathrm{d}g(\bar{x})(v) = 0\}$, $\mathrm{d}^2\bar{g}(\bar{x}|0)(v) = \mathrm{d}^2\bar{g}(\bar{x})(v)$. Therefore, for $v \in \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x}) \cap \{v : \mathrm{d}g(\bar{x})(v) = 0\}$, we have

$$\begin{aligned} 0 \leq \mathrm{d}^2\bar{g}(\bar{x})(v) &\overset{(a)}{=} \liminf_{v' \to v, t \downarrow 0} \frac{g(\bar{x} + tv') + \delta_{\mathcal{X}}(\bar{x} + tv') - g(\bar{x}) - \delta_{\mathcal{X}}(\bar{x}) - t\mathrm{d}\bar{g}(\bar{x})(v')}{\frac{1}{2}t^2} \\ &\overset{(b)}{\leq} \liminf_{t \downarrow 0} \frac{g(\bar{x} + tv) + \delta_{\mathcal{X}}(\bar{x} + tv) - g(\bar{x}) - \delta_{\mathcal{X}}(\bar{x}) - t\mathrm{d}\bar{g}(\bar{x})(v)}{\frac{1}{2}t^2} \\ &\overset{(c)}{=} \lim_{t \downarrow 0} \frac{g(\bar{x} + tv) - g(\bar{x}) - t\mathrm{d}g(\bar{x})(v)}{\frac{1}{2}t^2} \overset{(d)}{=} \mathrm{d}^2 g(\bar{x})(v), \end{aligned}$$

where (a) follows from the definition of the second-order subderivative $\mathrm{d}^2\bar{g}(\bar{x})(v)$, (b) follows from the definition of limit inferior (see [32, Definition 1.5]), (c) follows from $\bar{x} \in \mathcal{X}$ and $\bar{x} + tv \in \mathcal{X}$ for sufficiently small t due to $v \in \mathcal{T}^{\circ}_{\mathcal{X}}(\bar{x})$ and (d) follows from the twice semidifferentiability of g at $\bar{x}$. ∎

The following lemma gives a description of the generalized second-order directional derivative by using directional derivatives.

Lemma 3.9 ([8, Proposition 1.3]). *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuous function that admits a directional derivative at every point near x. Then $g^{\circ\circ}(x; u, v)$ is the generalized directional derivative of $g'(\cdot, v)$ at x along direction u, that is*

$$g^{\circ\circ}(x; u, v) = \limsup_{\substack{x' \to x \\ t \downarrow 0}} \frac{g'(x' + tu; v) - g'(x'; v)}{t}.$$

Remark 3.10. Note that

$$g^{\circ\circ}(x; v) \geq \lim_{t \downarrow 0} \frac{g(x + tv + tv) - g(x + tv) - g(x + tv) + g(x)}{t^2} = g^{(2)}(x; v).$$

Recall that $g : \mathbb{R}^n \to \mathbb{R}$ is *twice subregular* at x [8, Definition 3.1] if the limit

$$\lim_{t \downarrow 0, \delta \downarrow 0} \frac{g(x + \delta u + tv) - g(x + \delta u) - g(x + tv) + g(x)}{\delta t}$$

exists and the above limit equals to $g^{\circ\circ}(x;u,v)$. Thus, we know that $g^{\circ\circ}(x;v) = g^{(2)}(x;v)$ if g is twice subregular at x.

Now we are ready to give the main results of this subsection.

Theorem 3.11. *Let the tuple $(\hat{x},\hat{y}) \in X \times Y$ be a local minimax point of problem (1.1).*
(i) If f is semidifferentiable at $(\hat{x},\hat{y})$, then

$$f_x^{\circ}(\hat{x},\hat{y};v) \geq 0 \text{ for all } v \in \mathcal{T}_X(\hat{x}), \tag{3.2a}$$
$$\mathrm{d}_y f(\hat{x},\hat{y})(w) \leq 0 \text{ for all } w \in \mathcal{T}_Y(\hat{y}), \tag{3.2b}$$

where $f_x^{\circ}(\hat{x},\hat{y};v)$ denotes the generalized directional derivative of f with respect to x at $\hat{x}$ along the direction v for fixed $\hat{y}$.
(ii) Assume, further, that f is twice semidifferentiable at $(\hat{x},\hat{y})$ and f is Clarke regular in a neighborhood of $(\hat{x},\hat{y})$. Then

$$f_x^{\circ\circ}(\hat{x},\hat{y};v) \geq 0 \text{ for all } v \in \mathcal{T}_X^{\circ}(\hat{x}) \cap \{v : \exists\delta > 0, \mathrm{d}_x f(\hat{x},y')(v) = 0, \forall y' \in \mathbb{B}(\hat{y},\delta) \cap Y\}, \tag{3.3a}$$
$$\mathrm{d}_y^2 f(\hat{x},\hat{y})(w) \leq 0 \text{ for all } w \in \mathcal{T}_Y^{\circ}(\hat{y}) \cap \{w : \mathrm{d}_y f(\hat{x},\hat{y})(w) = 0\}, \tag{3.3b}$$

where $f_x^{\circ\circ}(\hat{x},\hat{y};v)$ denotes the generalized second-order directional derivative of f with respect to x at $\hat{x}$ along the direction (v,v) for fixed $\hat{y}$.

Proof. (3.2b) and (3.3b) directly follow from Lemma 3.8. Therefore, we only focus on (3.2a) and (3.3a), respectively.

(i) Since $(\hat{x},\hat{y})$ is a local minimax point, there exist a $\delta_0 > 0$ and a function $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\tau(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0,\delta_0]$ and $(x,y) \in X \times Y$ satisfying $\|x - \hat{x}\| \leq \delta$ and $\|y - \hat{y}\| \leq \delta$, we have

$$f(\hat{x},y) \leq f(\hat{x},\hat{y}) \leq \max_{y' \in \{y \in Y : \|y-\hat{y}\| \leq \tau(\delta)\}} f(x,y'). \tag{3.4}$$

For any $v \in \mathcal{T}_X(\hat{x})$, according to the convexity of X, there exist $\{v^k\}_{k\geq 1}$ with $v^k \to v$ as $k \to \infty$ and $\{t_k\}_{k\geq 1}$ with $t_k \downarrow 0$ as $k \to \infty$, such that $x^k := \hat{x} + t_k v^k \in X$ (see Lemma 3.1). Let $\delta_k = \left\|x^k - \hat{x}\right\|$ and $\tilde{y}^k$ be defined by

$$\tilde{y}^k \in \operatorname*{arg\,max}_{y' \in \{y \in Y : \|y-\hat{y}\| \leq \tau(\delta_k)\}} f(x^k,y'). \tag{3.5}$$

Obviously, $\delta_k \to 0$ and $\left\|\tilde{y}^k - \hat{y}\right\| \to 0$ as $k \to \infty$. According to the second inequality of (3.4), we have (for sufficiently large k) that

$$\begin{aligned} 0 \leq f(x^k,\tilde{y}^k) - f(\hat{x},\hat{y}) &= f(x^k,\tilde{y}^k) - f(\hat{x},\tilde{y}^k) + f(\hat{x},\tilde{y}^k) - f(\hat{x},\hat{y}) \\ &\leq f(x^k,\tilde{y}^k) - f(\hat{x},\tilde{y}^k). \end{aligned} \tag{3.6}$$

Note from the mean-value theorem [7, Theorem 2.3.7] that there exists an $\tilde{x}^k$ lying in the segment between x^k and $\hat{x}$ such that

$$f(x^k,\tilde{y}^k) - f(\hat{x},\tilde{y}^k) \in \left\langle \partial f(\tilde{x}^k,\tilde{y}^k), \begin{pmatrix} t_k v^k \\ 0 \end{pmatrix} \right\rangle.$$

It indicates that there exists an element contained in $\left\langle \partial f(\tilde{x}^k, \tilde{y}^k), \begin{pmatrix} t_k v^k \\ 0 \end{pmatrix} \right\rangle$ such that it is not less than 0. Thus, by dividing t_k in both sides and letting $k \to \infty$, due to the upper semicontinuity of $\partial f(\cdot, \cdot)$ (see [7, Proposition 2.1.5]), we obtain

$$0 \le \sup_{\zeta \in \partial f(\hat{x}, \hat{y})} \left\langle \zeta, \begin{pmatrix} v \\ 0 \end{pmatrix} \right\rangle \overset{(a)}{=} f^\circ(\hat{x}, \hat{y}; v, 0) = f_x^\circ(\hat{x}, \hat{y}; v),$$

where (a) follows from (3.1) and $f_x^\circ(\hat{x}, \hat{y}; v)$ denotes the Clarke generalized directional derivative of f with respect to x at $\hat{x}$ along the direction v for fixed $\hat{y}$.

(ii) Let $v \in \mathcal{T}_X^\circ(\hat{x}) \cap \{v : \exists \delta > 0, \mathrm{d}_x f(\hat{x}, y')(v) = 0, \forall y' \in \mathbb{B}(\hat{y}, \delta) \cap Y\}$. Then there exists a sequence $\{t_k\}_{k \ge 1}$ with $t_k \downarrow 0$, such that $x^k := \hat{x} + t_k v \in X$. Let $\delta_k = \|x^k - \hat{x}\|$, and $\tilde{y}^k$ be defined in (3.5).

From the mean-value theorem, there is $\zeta_k \in (0, t_k)$ such that

$$f(\hat{x} + t_k v, \tilde{y}^k) - f(\hat{x}, \tilde{y}^k) \in \partial f(\hat{x} + \zeta_k v, \tilde{y}^k) \begin{pmatrix} t_k v \\ 0 \end{pmatrix}.$$

Similar to (3.6), we have $f(\hat{x} + t_k v, \tilde{y}^k) - f(\hat{x}, \tilde{y}^k) \ge 0$. Thus, we have

$$f^\circ(\hat{x} + \zeta_k v, \tilde{y}^k; v, 0) = \sup_{\theta \in \partial f(\hat{x} + \zeta_k v, \tilde{y}^k)} \left\langle \theta, \begin{pmatrix} v \\ 0 \end{pmatrix} \right\rangle \ge 0. \tag{3.7}$$

Then, according to the Clarke regularity of f near $(\hat{x}, \hat{y})$, we have from (3.7) that

$$\begin{aligned} 0 &\overset{(b)}{\le} \limsup_{k \to \infty} \frac{f^\circ(\hat{x} + \zeta_k v, \tilde{y}^k; v, 0)}{\zeta_k} \overset{(c)}{=} \limsup_{k \to \infty} \frac{f'(\hat{x} + \zeta_k v, \tilde{y}^k; v, 0)}{\zeta_k} \\ &\overset{(d)}{=} \limsup_{k \to \infty} \frac{f'(\hat{x} + \zeta_k v, \tilde{y}^k; v, 0) - f'(\hat{x}, \tilde{y}^k; v, 0)}{\zeta_k} \le \limsup_{\substack{x' \to \hat{x}, y' \to \hat{y} \\ t \downarrow 0}} \frac{f'(x' + tv, y'; v, 0) - f'(\hat{x}, y'; v, 0)}{t} \\ &\overset{(e)}{=} f^{\circ\circ}(\hat{x}, \hat{y}; v, 0) = f_x^{\circ\circ}(\hat{x}, \hat{y}; v), \end{aligned}$$

where (b) follows from (3.7), (c) follows from the Clarke regularity of f near $(\hat{x}, \hat{y})$, (d) follows from $f'(\hat{x}, \tilde{y}^k; v, 0) = 0$ for sufficiently large k, (e) follows from Lemma 3.9 and $f_x^{\circ\circ}(\hat{x}, \hat{y}; v)$ denotes the generalized second-order directional derivative of f with respect to x at $\hat{x}$ along the direction (v, v) for fixed $\hat{y}$. ■

We illustrate Theorem 3.11 by Example A.1 in Appendix A.

Remark 3.12. We know from (3.1) that for any v, $f_x^\circ(\hat{x}, \hat{y}; v) = \max_{z \in \partial_x f(\hat{x}, \hat{y})} \langle z, v \rangle$. Thus, (3.2a) can be equivalently reformulated as $\max_{z \in \partial_x f(\hat{x}, \hat{y})} \langle z, v \rangle \ge 0$, $\forall v \in \mathcal{T}_X(\hat{x})$, which, based on the definition of normal cone, is equivalent to $0 \in \partial_x f(\hat{x}, \hat{y}) + \mathcal{N}_X(\hat{x})$.

Generally, (3.2b) implies the Clarke stationary condition $0 \in -\partial_y f(\hat{x}, \hat{y}) + \mathcal{N}_Y(\hat{y})$, but not vice versa. Moreover, by using the (generalized) directional derivatives, we can establish the second-order necessary optimality conditions for the nonsmooth case. Therefore, the (generalized) directional derivatives are employed in Theorem 3.11.

Remark 3.13. It is noteworthy that the necessary optimality conditions (3.2a)-(3.2b) and (3.3a)-(3.3b) with respect to x and y are not symmetric. Generally, (3.2a) and (3.3a) are weaker than

$$\mathrm{d}_x f(\hat{x}, \hat{y}; v) \geq 0 \text{ for all } v \in \mathcal{T}_X(\hat{x}) \tag{3.8}$$

and

$$\mathrm{d}_x^2 f(\hat{x}, \hat{y}; v) \geq 0 \text{ for all } v \in \mathcal{T}_X^\circ(\hat{x}) \cap \{v : \mathrm{d}_x f(\hat{x}, \hat{y})(v) = 0\}, \tag{3.9}$$

respectively, because $f_x^\circ(\hat{x}, \hat{y}; v) \geq \mathrm{d}_x f(\hat{x}, \hat{y}; v)$, $f_x^{\circ\circ}(\hat{x}, \hat{y}; v) \geq \mathrm{d}_x^2 f(\hat{x}, \hat{y}; v)$ (Remark 3.10) and

$$\mathcal{T}_X^\circ(\hat{x}) \cap \{v : \exists \delta > 0, \mathrm{d}_x f(\hat{x}, y')(v) = 0, \forall y' \in \mathbb{B}(\hat{y}, \delta) \cap Y\} \subseteq \mathcal{T}_X^\circ(\hat{x}) \cap \{v : \mathrm{d}_x f(\hat{x}, \hat{y})(v) = 0\}.$$

The main reason is that a local minimax point may not be a local saddle point. If we replace (3.2a) and (3.3a) by (3.8) and (3.9) respectively, the necessary optimality conditions for local saddle points are derived. Indeed, if $(\hat{x}, \hat{y}) \in X \times Y$ is a local saddle point of problem (1.1), then $\hat{x}$ is a local minimum of $\min_{x \in X} f(x, \hat{y})$ and $\hat{y}$ is a local maximum of $\max_{y \in Y} f(\hat{x}, y)$ by Definition 2.2. Hence by Lemma 3.8, we obtain that (3.8) and (3.9) are necessary optimality conditions for local saddle points of problem (1.1).

If, in addition, f is Clarke regular at $(\hat{x}, \hat{y})$, then

$$f_x^\circ(\hat{x}, \hat{y}; v) \overset{(a)}{=} f^\circ(\hat{x}, \hat{y}; v, 0) \overset{(b)}{=} f'(\hat{x}, \hat{y}; v, 0) \overset{(c)}{=} \mathrm{d} f(\hat{x}, \hat{y})(v, 0) \overset{(d)}{=} \mathrm{d}_x f(\hat{x}, \hat{y})(v),$$

where (a) follows from the definition of f_x°, (b) follows from the Clarke regularity, (c) follows from [10, Section 2.1] and (d) follows from the definition of $\mathrm{d}_x f$. Thus, (3.2a) can be replaced by (3.8).

If, in addition, f is twice subregular at $(\hat{x}, \hat{y})$, then

$$f^{\circ\circ}(\hat{x}, \hat{y}; v, 0) \overset{(e)}{=} \mathrm{d}^2 f(\hat{x}, \hat{y})(v, 0) \overset{(f)}{=} \mathrm{d}_x^2 f(\hat{x}, \hat{y})(v),$$

where (e) follows from [10, Section 2.1] and (f) follows from the definition of $\mathrm{d}_x^2 f$. Thus (3.3a) can be replaced by

$$\mathrm{d}_x^2 f(\hat{x}, \hat{y})(v) \geq 0 \text{ for all } v \in \mathcal{T}_X^\circ(\hat{x}) \cap \{v : \exists \delta > 0, \mathrm{d}_x f(\hat{x}, y')(v) = 0, \forall y' \in \mathbb{B}(\hat{y}, \delta) \cap Y\}.$$

Remark 3.14. Suppose that f is twice semidifferentiable, Clarke regular and twice subregular. Then we have $f_x^\circ(\hat{x}, \hat{y}; v) = \mathrm{d}_x f(\hat{x}, \hat{y})(v)$ and $f_x^{\circ\circ}(\hat{x}, \hat{y}; v) = \mathrm{d}_x^2 f(\hat{x}, \hat{y})(v)$. Based on Lemma C.4 and (3.3), we can have

$$\begin{aligned} &f_x^{\circ\circ}(\hat{x}, \hat{y}; v) > 0 \text{ for all } 0 \neq v \in \mathcal{T}_X(\hat{x}) \cap \{v : \mathrm{d}_x f(\hat{x}, \hat{y})(v) = 0\}, \\ &\mathrm{d}_y^2 f(\hat{x}, \hat{y})(w) > 0 \text{ for all } 0 \neq w \in \mathcal{T}_Y(\hat{y}) \cap \{w : \mathrm{d}_y f(\hat{x}, \hat{y})(w) = 0\}, \end{aligned} \tag{3.10}$$

with (3.2) as a second-order sufficient condition for a local saddle point. Since a local saddle point is a local minimax point, (3.10) together with (3.2) is also a sufficient condition for a local minimax point.

Based on Theorem 3.11, we define the first-order and second-order d-stationary points of min-max problems.

Definition 3.15. *We call that $(\hat{x}, \hat{y}) \in X \times Y$ is a first-order d-stationary point of problem* (1.1) *if it satisfies* (3.2a)-(3.2b). *If $(\hat{x}, \hat{y})$ also satisfies* (3.3a)-(3.3b), *we call it a second-order d-stationary point of problem* (1.1).

3.2. Smooth case. In this subsection, we consider the necessary optimality conditions of problem (1.1) when f is (twice) continuously differentiable. For any $(x, y) \in X \times Y$, denote

$$\Gamma_1^\circ(x,y) = \{v \in \mathcal{T}_X^\circ(x) : v \perp \nabla_x f(x,y)\}, \quad \Gamma_1(x,y) = \{v \in \mathcal{T}_X(x) : v \perp \nabla_x f(x,y)\},$$
$$\Gamma_2^\circ(x,y) = \{w \in \mathcal{T}_Y^\circ(y) : w \perp \nabla_y f(x,y)\}, \quad \Gamma_2(x,y) = \{w \in \mathcal{T}_Y(y) : w \perp \nabla_y f(x,y)\}.$$

It is noteworthy that $\mathrm{cl}(\Gamma_1^\circ(x,y)) \neq \Gamma_1(x,y)$ and $\mathrm{cl}(\Gamma_2^\circ(x,y)) \neq \Gamma_2(x,y)$ generally even if we have $\mathrm{cl}(\mathcal{T}_X^\circ(x)) = \mathcal{T}_X(x)$ and $\mathrm{cl}(\mathcal{T}_Y^\circ(y)) = \mathcal{T}_Y(y)$. We summarize their relationships as follows.

Lemma 3.16. *Let $(x,y) \in X \times Y$. Then $\Gamma_1^\circ(x,y)$, $\Gamma_1(x,y)$, $\Gamma_2^\circ(x,y)$ and $\Gamma_2(x,y)$ are convex cones, and we have $\mathrm{cl}\Gamma_1^\circ(x,y) \subseteq \Gamma_1(x,y)$ and $\mathrm{cl}\Gamma_2^\circ(x,y) \subseteq \Gamma_2(x,y)$. Moreover, if X and Y are polyhedral, then $\Gamma_1^\circ(x,y) = \mathrm{cl}\Gamma_1^\circ(x,y) = \Gamma_1(x,y)$ and $\Gamma_2^\circ(x,y) = \mathrm{cl}\Gamma_2^\circ(x,y) = \Gamma_2(x,y)$.*

Proof. Since X and Y are closed and convex, we know from Lemma 3.1 that $\mathcal{T}_X^\circ(x)$, and $\mathcal{T}_Y^\circ(y)$ are convex cones, $\mathcal{T}_X(x)$ and $\mathcal{T}_Y(y)$ are closed convex cones, and

$$\mathrm{cl}\mathcal{T}_X^\circ(\bar{x}) \subseteq \mathcal{T}_X(\bar{x}) \text{ and } \mathrm{cl}\mathcal{T}_Y^\circ(\bar{y}) \subseteq \mathcal{T}_Y(\bar{y}).$$

Thus, we obtain that $\Gamma_1^\circ(x,y)$, $\Gamma_1(x,y)$, $\Gamma_2^\circ(x,y)$ and $\Gamma_2(x,y)$ are convex cones. Moreover, we have

$$\begin{aligned}\mathrm{cl}\Gamma_1^\circ(x,y) &= \mathrm{cl}\{v \in \mathcal{T}_X^\circ(x) : v \perp \nabla_x f(x,y)\} \subseteq \{v \in \mathrm{cl}\mathcal{T}_X^\circ(x) : v \perp \nabla_x f(x,y)\} \\ &\subseteq \{v \in \mathcal{T}_X(x) : v \perp \nabla_x f(x,y)\} = \Gamma_1(x,y).\end{aligned}$$

Similarly, we can verify $\mathrm{cl}\Gamma_2^\circ(x,y) \subseteq \Gamma_2(x,y)$.

If, further, X and Y are polyhedral, we have $\mathcal{T}_X^\circ(\bar{x}) = \mathcal{T}_X(\bar{x})$ and $\mathcal{T}_Y^\circ(\bar{y}) = \mathcal{T}_Y(\bar{y})$. Thus,

$$\begin{aligned}\mathrm{cl}\Gamma_1^\circ(x,y) \subseteq \Gamma_1(x,y) &= \{v \in \mathcal{T}_X(x) : v \perp \nabla_x f(x,y)\} \\ &= \{v \in \mathcal{T}_X^\circ(x) : v \perp \nabla_x f(x,y)\} = \Gamma_1^\circ(x,y),\end{aligned}$$

which implies that $\Gamma_1^\circ(x,y) = \mathrm{cl}\Gamma_1^\circ(x,y) = \Gamma_1(x,y)$. Similarly, we can verify $\Gamma_2^\circ(x,y) = \mathrm{cl}\Gamma_2^\circ(x,y) = \Gamma_2(x,y)$. ■

Theorem 3.17. *Let f be continuously differentiable and the tuple $(\hat{x}, \hat{y}) \in X \times Y$ be a local minimax point of problem (1.1).*

(i) Then it holds that

$$0 \in \nabla_x f(\hat{x}, \hat{y}) + \mathcal{N}_X(\hat{x}), \tag{3.11a}$$
$$0 \in -\nabla_y f(\hat{x}, \hat{y}) + \mathcal{N}_Y(\hat{y}). \tag{3.11b}$$

(ii) Assume, further, that f is twice continuously differentiable. Then

$$\left\langle v, \nabla^2_{xx} f(\hat{x}, \hat{y}) v \right\rangle \geq 0 \text{ for all } v \in \mathrm{cl}\left\{\bar{v} : \exists \delta > 0, \bar{v} \in \Gamma_1^\circ(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\right\}, \tag{3.12a}$$
$$\left\langle w, \nabla^2_{yy} f(\hat{x}, \hat{y}) w \right\rangle \leq 0 \text{ for all } w \in \mathrm{cl}\Gamma_2^\circ(\hat{x}, \hat{y}). \tag{3.12b}$$

Proof. (i) The proof is similar to Theorem 3.11. Here we give a simple proof of (3.11a) and (3.12a) for completeness. For any $x^k \xrightarrow{X} \hat{x}$ as $k \to \infty$, denote $\delta_k = \|x^k - \hat{x}\|$ and $\tilde{y}^k$

is defined in (3.5). Obviously, $\delta_k \to 0$ and $\left\|\tilde{y}^k - \hat{y}\right\| \to 0$ as $k \to \infty$. From the continuous differentiability of f, we have

$$0 \le f(x^k, \tilde{y}^k) - f(\hat{x}, \tilde{y}^k) = \nabla f(\bar{x}^k, \tilde{y}^k)^\top \begin{pmatrix} x^k - \hat{x} \\ \tilde{y}^k - \tilde{y}^k \end{pmatrix} = \nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) + o\left(\left\|x^k - \hat{x}\right\|\right),$$

where $\bar{x}^k$ is some point lying in the segment between $\hat{x}$ and x^k. Thus, we obtain

$$-\nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) \le o\left(\left\|x^k - \hat{x}\right\|\right).$$

We know from [32, Definition 6.3] that $-\nabla_x f(\hat{x}, \hat{y}) \in \mathcal{N}_X(\hat{x})$, which verifies (3.11a).

(ii) We only need to prove that (3.12a) holds with $v \in \Gamma_1^\circ(\hat{x}, y')$ for all $y' \in \mathbb{B}(\hat{y}, \delta)$ and some $\delta > 0$. According to the definition of $\mathcal{T}_X^\circ(\hat{x})$, there exists a sequence $\{t_k\}_{k\ge 1}$ with $t_k \downarrow 0$ as $k \to \infty$, such that $x^k := \hat{x} + t_k v \in X$. Let $\delta_k = t_k \|v\|$, and $\tilde{y}^k$ is denoted in (3.5). Similarly, we have that

$$\begin{aligned} 0 \le f(x^k, \tilde{y}^k) - f(\hat{x}, \tilde{y}^k) &\overset{(a)}{=} \nabla_x f(\hat{x}, \tilde{y}^k)^\top (x^k - \hat{x}) + \frac{1}{2}(x^k - \hat{x})^\top \nabla_{xx}^2 f(\tilde{x}^k, \tilde{y}^k)(x^k - \hat{x}) \\ &\overset{(b)}{=} \nabla_x f(\hat{x}, \tilde{y}^k)^\top (x^k - \hat{x}) + \frac{1}{2}(x^k - \hat{x})^\top \nabla_{xx}^2 f(\hat{x}, \hat{y})(x^k - \hat{x}) + o\left(\left\|x^k - \hat{x}\right\|^2\right), \end{aligned}$$

where (a) follows from Taylor's theorem for multivariate functions with Lagrange's remainder, and $\tilde{x}^k$ is some point lying in the segment between $\hat{x}$ and x^k; (b) follows from the twice continuous differentiability of f and $\tilde{x}^k \to \hat{x}$ as $k \to \infty$. Thus, we obtain

$$t_k \nabla_x f(\hat{x}, \tilde{y}^k)^\top v + t_k^2 \frac{1}{2} v^\top \nabla_{xx}^2 f(\hat{x}, \hat{y}) v + \|v\|^2 o(t_k^2) \ge 0.$$

Since $\nabla_x f(\hat{x}, \tilde{y}^k)^\top v = 0$ for sufficiently large k, dividing by t_k^2 in both sides and letting $k \to \infty$, we complete the proof. ■

Remark 3.18. The asymmetry between (3.12a) and (3.12b) mainly arises from the asymmetry between x and y in a local minimax point. Conversely, if the conditions in (ii) of Theorem 3.17 hold except that $\mathrm{cl}\{w : \exists \delta > 0, w \in \Gamma_1^\circ(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\}$ and $\mathrm{cl}\Gamma_2^\circ(\hat{x}, \hat{y})$ are replaced by $\Gamma_1(\hat{x}, \hat{y})$ and $\Gamma_2(\hat{x}, \hat{y})$, respectively, and the inequality is strict when $v \ne 0$ and $w \ne 0$, then $(\hat{x}, \hat{y})$ is a local saddle point. In that case, (3.12) together with (3.11) are the so-called second-order sufficient condition for a local saddle point. This fact can be easily derived by using the sufficient optimality condition for minimization problems (see [32, Example 13.25]) and the definition of local saddle points (see Definition 2.2). Specifically, by invoking Lemma C.3 (ii), these conditions imply that $\hat{y}$ is a local maximum of $\max_{y \in Y} f(\hat{x}, y)$ for fixed $\hat{x}$, and $\hat{x}$ is a local minimum of $\min_{x \in X} f(x, \hat{y})$ for fixed $\hat{y}$. Hence $(\hat{x}, \hat{y})$ is a local saddle point.

Corollary 3.19. *Let f be twice continuously differentiable. If, further, for local minimax point $(\hat{x}, \hat{y})$, there exists an τ such that $\tau(\delta) = o(\delta)$ as $\delta \downarrow 0$, then (3.12a) can be replaced by*

$$\left\langle v, \nabla_{xx}^2 f(\hat{x}, \hat{y}) v \right\rangle \ge 0 \text{ for all } v \in \mathrm{cl}\Gamma_1^\circ(\hat{x}, \hat{y}).$$

Proof. Let $0 \neq v \in \Gamma_1^\circ(\hat{x}, \hat{y})$. According to the definition of $\mathcal{T}_X^\circ(\hat{x})$, there exists a sequence $\{t_k\}_{k\geq 1}$ with $t_k \downarrow 0$ as $k \to \infty$, such that $x^k := \hat{x} + t_k v \in X$. Let $\delta_k := \left\|x^k - \hat{x}\right\|$, and $\tilde{y}^k$ be denoted in (3.5). Since $\tau(\delta) = o(\delta)$ as $\delta \downarrow 0$, we have $\left\|\tilde{y}^k - \hat{y}\right\| = o\left(\left\|x^k - \hat{x}\right\|\right)$ for sufficiently large k. We know from the twice continuous differentiability of f that

$$\begin{aligned} f(x^k, \tilde{y}^k) &= f(\hat{x}, \hat{y}) + \nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) + \nabla_y f(\hat{x}, \hat{y})^\top (\tilde{y}^k - \hat{y}) \\ &\quad + \frac{1}{2}(x^k - \hat{x})^\top \nabla_{xx}^2 f(\hat{x}, \hat{y})(x^k - \hat{x}) + (x^k - \hat{x})^\top \nabla_{xy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y}) \\ &\quad + \frac{1}{2}(\tilde{y}^k - \hat{y})^\top \nabla_{yy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y}) + o\left(\left\|x^k - \hat{x}\right\|^2 + \left\|\tilde{y}^k - \hat{y}\right\|^2\right), \\ f(\hat{x}, \tilde{y}^k) &= f(\hat{x}, \hat{y}) + \nabla_y f(\hat{x}, \hat{y})^\top (\tilde{y}^k - \hat{y}) + \frac{1}{2}(\tilde{y}^k - \hat{y})^\top \nabla_{yy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y}) \\ &\quad + o\left(\left\|\tilde{y}^k - \hat{y}\right\|^2\right). \end{aligned}$$

Using $t_k \nabla_x f(\hat{x}, \hat{y})^\top v = \nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) = 0$ for $v \in \Gamma_1^\circ(\hat{x}, \hat{y})$, we have

$$\begin{aligned} 0 &\leq f(x^k, \tilde{y}^k) - f(\hat{x}, \tilde{y}^k) \\ &= \frac{1}{2}(x^k - \hat{x})^\top \nabla_{xx}^2 f(\hat{x}, \hat{y})(x^k - \hat{x}) + (x^k - \hat{x})^\top \nabla_{xy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y}) \\ &\quad + o\left(\left\|x^k - \hat{x}\right\|^2 + \left\|\tilde{y}^k - \hat{y}\right\|^2\right) - o\left(\left\|\tilde{y}^k - \hat{y}\right\|^2\right) \\ &\overset{(a)}{=} \frac{1}{2}(x^k - \hat{x})^\top \nabla_{xx}^2 f(\hat{x}, \hat{y})(x^k - \hat{x}) + (x^k - \hat{x})^\top \nabla_{xy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y}) + o\left(\left\|x^k - \hat{x}\right\|^2\right) \\ &\overset{(b)}{=} t_k^2 \frac{1}{2} v^\top \nabla_{xx}^2 f(\hat{x}, \hat{y}) v + o(t_k^2), \end{aligned}$$

where (a) follows from the fact that $\left\|\tilde{y}^k - \hat{y}\right\| = o\left(\left\|x^k - \hat{x}\right\|\right)$ for sufficiently large k and (b) follows from the fact that

$$\left|(x^k - \hat{x})^\top \nabla_{xy}^2 f(\hat{x}, \hat{y})(\tilde{y}^k - \hat{y})\right| \leq \left\|x^k - \hat{x}\right\| \left\|\nabla_{xy}^2 f(\hat{x}, \hat{y})\right\| \left\|\tilde{y}^k - \hat{y}\right\| = o(t_k^2).$$

Finally, dividing by t_k^2 in both sides and letting $t_k \to 0$, we complete the proof. ■

Remark 3.20. In Corollary 3.19, the asymmetry of optimality conditions between on x and on y has been removed. The main reason lies in that we restrict the scope of the local minimax points by requiring $\tau(\delta) = o(\delta)$ as $\delta \downarrow 0$ in Definition 2.4.

The following example illustrates $\operatorname{cl}\{w : w \in \Gamma_1^\circ(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\}$ for some $\delta > 0$.

Example 3.21. Let $n = m = 1$, $X = Y = [-1, 1]$. Consider

$$\min_{x\in[-1,1]} \max_{y\in[-1,1]} \; f(x, y) := -x^4 + 4x^2y^2 - y^4.$$

We have

$$\varphi(x)=\max_{y\in[-1,1]}(-x^4+4x^2y^2-y^4)=\begin{cases}3x^4, & x\in\left[-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right]\ (y^*=\pm\sqrt{2}x);\\ -x^4+4x^2-1, & [-1,1]\backslash\left[-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right]\ (y^*=1),\end{cases}$$

which is not a convex function over $[-1,1]$. Moreover, it can be examined that $(0,0)$ is a global minimax point. In fact, it is also a local minimax point. Let $\tau(\delta)=2\delta^2$ and $\delta_0=\frac{\sqrt{2}}{2}$. Then, for any $\delta\in(0,\delta_0]$ and any $(x,y)\in[-1,1]^2$ satisfying $|x|\le\delta$ and $|y|\le\delta$, we have

$$-y^4=f(0,y)\le f(0,0)\le\max_{y'\in\{y\in Y:|y|\le\tau(\delta)\}}f(x,y')=3x^4.$$

Therefore, for any $\delta\in(0,1]$,

$$\mathrm{cl}\left\{w:w\in\Gamma_1^\circ(0,y'),\forall y'\in\mathbb{B}(0,\delta)\right\}=\mathrm{cl}\left(\bigcap_{y'\in\mathbb{B}(0,\delta)}\{w_1\in\mathcal{T}^\circ_{[-1,1]}(0):w_1\perp\nabla_x f(0,y')\}\right)=\mathbb{R}.$$

Similarly, we have $\mathrm{cl}\Gamma_2^\circ(0,0)=\left\{w_2\in\mathcal{T}^\circ_{[-1,1]}(0):w_2\perp\nabla_y f(0,0)\right\}=\mathbb{R}$.

In this case, the second-order optimality condition (3.12) means $\nabla^2_{xx}f(0,0)\ge 0$ and $\nabla^2_{yy}f(0,0)\le 0$.

In Theorem 3.17, the first-order and second-order optimality necessary conditions are given in a sense of geometry. In particular, for the case that X and Y are polyhedral, we derive the corresponding Karush-Kuhn-Tucker (KKT) systems in Appendix B.

Definition 3.22. *We call that $(\hat{x},\hat{y})\in X\times Y$ is a first-order stationary point of problem* (1.1) *if it satisfies* (3.11a)-(3.11b). *Moreover, if $(\hat{x},\hat{y})$ also satisfies* (3.12a)-(3.12b), *we call it a second-order stationary point of problem* (1.1).

The existence results of the first-order stationary points can be obtained by using existing results in [15, Proposition 2.2.3, Corollary 2.2.5]. Let $F(x,y)=\begin{pmatrix}\nabla_x f(x,y)\\ -\nabla_y f(x,y)\end{pmatrix}$.

(i) If there exist a bounded open set $\mathcal{Z}\subseteq X\times Y$ and a point $(\bar{x},\bar{y})\in(X\times Y)\cap\mathcal{Z}$ such that

$$\left\langle F(x,y),\begin{pmatrix}x-\bar{x}\\ y-\bar{y}\end{pmatrix}\right\rangle\ge 0,\ \forall(x,y)\in(X\times Y)\cap\mathrm{bd}(\mathcal{Z}),$$

then problem (1.1) has at least a first-order stationary point.

(ii) Specially, if X and Y are bounded, the first-order stationary point set of problem (1.1) is nonempty.

We know from [21, Proposition 21] that a global minimax point can be neither a local minimax point nor a stationary point. However, some global minimax points can be the first-order stationary points.

The following proposition claims that under mild conditions a class of global minimax points are first-order stationary points.

Proposition 3.23. *Let f be continuously differentiable over $X \times Y$, and $(\hat{x}, \hat{y})$ be a global minimax point of* (1.1) *satisfying*

$$\hat{y} \in \limsup_{x \to \hat{x}} \left(\arg\max_{y' \in Y} f(x, y') \right),$$

where "$\limsup$" denotes outer limit ([32, Definition 4.1]), then $(\hat{x}, \hat{y})$ is a first-order stationary point.

Proof. Since $(\hat{x}, \hat{y})$ is a global minimax point, we have for any $(x, y) \in X \times Y$ that

$$f(\hat{x}, y) \overset{(a)}{\leq} f(\hat{x}, \hat{y}) \overset{(b)}{\leq} \max_{y' \in Y} f(x, y'). \tag{3.13}$$

The inequality (a) of (3.13) implies (3.11b). In the sequel, we only consider (3.11a) through inequality (b) of (3.13). Since

$$\hat{y} \in \limsup_{x \to \hat{x}} \left(\arg\max_{y' \in Y} f(x, y') \right),$$

without loss of generality, we know from the definition of outer limit that there exist a sequence $\{x^k\}$ and $\tilde{y}^k \in \arg\max_{y' \in Y} f(x^k, y')$ such that $\tilde{y}^k \to \hat{y}$ as $k \to \infty$. By a similar procedure to the proof for (i) of Theorem 3.17, we have

$$\begin{aligned} 0 &\leq \nabla_x f(\hat{x}, \tilde{y}^k)^\top (x^k - \hat{x}) + o\left(\left\|x^k - \hat{x}\right\|\right) \\ &= \nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) + (\nabla_x f(\hat{x}, \tilde{y}^k) - \nabla_x f(\hat{x}, \hat{y}))^\top (x^k - \hat{x}) + o\left(\left\|x^k - \hat{x}\right\|\right) \\ &= \nabla_x f(\hat{x}, \hat{y})^\top (x^k - \hat{x}) + o\left(\left\|x^k - \hat{x}\right\|\right), \end{aligned}$$

which implies that $-\nabla_x f(\hat{x}, \hat{y}) \in \mathcal{N}_X(\hat{x})$. ∎

In general, a global minimax point can be neither a local minimax point nor a stationary point [21, Proposition 21]. Moreover, a first-order stationary point may not be a local minimax point. We use the following example to show this assertion.

Example 3.24 ([21, Figure 2]). Let $n = m = 1$, $X = [-1, 1]$ and $Y = [-5, 5]$. Consider the following minimax problem

$$\min_{x \in [-1,1]} \max_{y \in [-5,5]} \ f(x, y) := xy - \cos(y). \tag{3.14}$$

By direct calculation, we have

$$\varphi(x) = \max_{y \in [-5,5]} (xy - \cos(y)) = \begin{cases} x \cdot (\pi - \arcsin(-x)) - \cos(\pi - \arcsin(-x)), & x \in [0, 1]; \\ x \cdot (-\pi - \arcsin(-x)) - \cos(-\pi - \arcsin(-x)), & x \in [-1, 0], \end{cases}$$

where the optima is achieved when $y = \pi - \arcsin(-x)$ and $y = -\pi - \arcsin(-x)$, respectively. It can observe from the definition of $\varphi(x)$ that $x = 0$ is the minimum. In this case, $(0, -\pi)$

and $(0, \pi)$ are two global minimax points. However, they both fail to satisfy (3.11a)-(3.11b), that is,

$$\begin{cases} 0 \in y + \mathcal{N}_{[-1,1]}(x), \\ 0 \in x + \sin(y) + \mathcal{N}_{[-5,5]}(y), \end{cases}$$

which has a unique solution $(0, 0)$. Thus, neither $(0, -\pi)$ nor $(0, \pi)$ is a first-order stationary point, which implies from Theorem 3.17 that they cannot be local minimax points either. Therefore, a global minimax point can be neither a local minimax point nor a first-order stationary point.

Next, we show that even $(0, 0)$ is not a local minimax point. For any y satisfying $0 < |y| \leq \delta$ with any sufficiently small $\delta > 0$, we have $-\cos(y) = f(0, y) > f(0, 0) = -1$, which, according to the definition of local minimax points in Definition 2.4, concludes that $(0, 0)$ is not a local minimax point. Therefore, problem (3.14) here does not have a local minimax point even both X and Y are bounded.

Sometimes we can find that a global minimax point may be a stationary point (Example 2.7). In the following proposition, we conclude some sufficient conditions such that a global minimax point is a local minimax point.

Proposition 3.25. *Let $(\hat{x}, \hat{y})$ be a global minimax point and f be Lipschitz continuous over $X \times Y$. Assume that for each x in a neighborhood of $\hat{x}$, $\max_{y' \in Y} f(x, y')$ has a unique and uniformly bounded solution. Then $(\hat{x}, \hat{y})$ is a local minimax point.*

Proof. Since $\max_{y' \in Y} f(x, y')$ has a unique solution for all x in a neighborhood of $\hat{x}$, we use $\bar{y}(x)$ to denote this unique solution. Consider

$$\max_{y' \in Y} g(y') := f(\hat{x}, y') \quad \text{and} \quad \max_{y' \in Y} \tilde{g}(y') := f(x, y').$$

Note that $f(\hat{x}, \cdot)$ is continuous and $\bar{y}(x)$ is uniformly bounded for x in a neighborhood of $\hat{x}$. Then, by using Lemma C.1, we know that $\|\bar{y}(x) - \hat{y}\| \to 0$ as $x \to \hat{x}$, which implies that there exists a $\delta_0 > 0$ such that for any $x \in X$ satisfying $\|x - \hat{x}\| \leq \delta \leq \delta_0$, $\tau(\delta) \to 0$ where $\tau(\delta) := \sup_{\{x \in X : \|x - \hat{x}\| \leq \delta\}} \|\bar{y}(x) - \hat{y}\|$. As $(\hat{x}, \hat{y})$ is a global minimax point, we have for any $x \in X$ and $y \in Y$ that $f(\hat{x}, y) \leq f(\hat{x}, \hat{y}) \leq \max_{y' \in Y} f(x, y')$. This indicates that for x satisfying $\|x - \hat{x}\| \leq \delta (\leq \delta_0)$ and y satisfying $\|y - \hat{y}\| \leq \tau(\delta)$, we have

$$f(\hat{x}, y) \leq f(\hat{x}, \hat{y}) \leq \max_{y' \in Y} f(x, y') = f(x, \bar{y}(x)) = \max_{y' \in \{y \in Y : \|y - \hat{y}\| \leq \tau(\delta)\}} f(x, y').$$

Thus, $(\hat{x}, \hat{y})$ is a local minimax point based on Definition 2.4. ∎

Obviously, when $f(x, \cdot)$ is strictly concave for all x in a neighborhood of $\hat{x}$, the condition for the uniqueness of the maximization problem holds.

To end this section, we summarize relationships between saddle points, local saddle points, global minimax points, local minimax points and first-order and second-order stationary points in Figure 1.

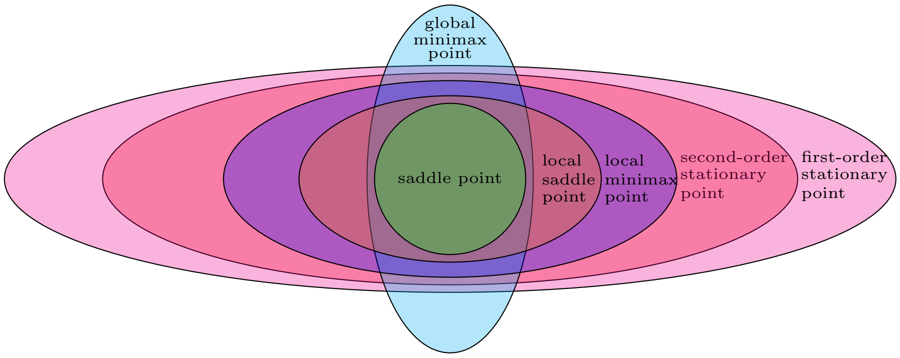

Figure 1. *Venn diagram for saddle points, minimax points and stationary points:*
a saddle point ⇒ a local saddle point (Definitions 2.1 & 2.2),
a global (local) minimax point ⇏ a local saddle point (Example 2.7),
a local saddle point ⇒ a local minimax point (Definitions 2.2 & 2.4),
a local minimax point ⇒ a first-order or second-order stationary point (Theorems 3.11 & 3.17),
a first-order stationary point ⇏ a local minimax point (Example 3.24),
a second-order stationary point ⇒ a first-order stationary point (Definition 3.22).

4. Generative adversarial networks. In this section, we consider the first-order and second -order optimality conditions of the GAN using nonsmooth activation functions, which can be formulated as nonsmooth nonconvex-nonconcave min-max problem (1.1).

The GAN is one of the most popular generative models in machine learning. It is comprised of two ingredients: the generator, which creates samples that are intended to follow the same distribution as the training data, and the discriminator, which examines samples to determine whether they are real or fake. For more motivations and advantages of GANs, one can refer to [17]. Recently, Wang gave a mathematical introduction to GANs in [34].

The plain vanilla GAN model can be formulated as (1.2), where D and G are given by feedforward neural networks with parameters x and y, respectively. The activation function is a function from $\mathbb{R}$ to $\mathbb{R}$ that is used to compute the hidden layer values and introduce the nonlinear property. There are several commonly-used activation functions, such as ReLU $\sigma(z) = \max\{0, z\}$, the logistic sigmoid $\sigma(z) = 1/(1+\exp(-z))$, the softplus activation function $\sigma(z) = \ln(1 + \exp(z))$, etc.

We give an intuition for D and G which are consist of linear models with activation functions in the following example.

Example 4.1. Consider that the discriminator D is a single-layer network with a logistic sigmoid activation function [18] and the generator G is a two-layer network with an activation function σ as follows $G(x, \xi_2) := W_2\boldsymbol{\sigma}(W_1\xi_2 + b_1) + b_2$ and $D(y, \xi_1) := \frac{1}{1+\exp(y^\top \xi_1)}$, where $x = (\mathrm{vec}(W_1)^\top, \mathrm{vec}(W_2)^\top, b_1^\top, b_2^\top)^\top$ and $\mathrm{vec}(\cdot)$ denotes the columnwise vectorization operator of matrices, $W_1 \in \mathbb{R}^{s\times s_2}$, $b_1 \in \mathbb{R}^s$, $W_2 \in \mathbb{R}^{s_1\times s}$, $b_2 \in \mathbb{R}^{s_1}$ and $\boldsymbol{\sigma} : \mathbb{R}^s \to \mathbb{R}^s$. Here, $\boldsymbol{\sigma}$ is a separable vector activation function that aggregates the individual neuron activations.

In this case, the GAN model (1.2) can be explicitly written as

$$\min_{x\in X}\max_{y\in Y}\; f(x,y) = \Bigg(\mathbb{E}_{P_1}\left[\log\left(\frac{1}{1+\exp(y^\top \xi_1)}\right)\right] + \mathbb{E}_{P_2}\left[\log\left(1-\frac{1}{1+\exp(y^\top(W_2\boldsymbol{\sigma}(W_1\xi_2+b_1)+b_2))}\right)\right]\Bigg). \tag{4.1}$$

If X and Y are compact and $\boldsymbol{\sigma}$ is continuous, by Proposition 2.6, problem (4.1) has a global minimax point.

Obviously, if $D(\cdot,\xi_1)$ and $G(\cdot,\xi_2)$ are smooth (i.e. $\boldsymbol{\sigma}$ is smooth), the necessary optimality conditions in Theorem 3.17 hold. Next, we focus on the nonsmooth case with the ReLU activation function.

Proposition 4.2. *Let f be defined in (4.1) with $\boldsymbol{\sigma}(\cdot)=(\cdot)_+$. Assume that support sets Ξ_1 and Ξ_2 are bounded. Then the following statements hold.*

(i) f is locally Lipschitz continuous and twice semidifferentiable in $X\times Y$.

(ii) If, in addition, f is Clarke regular and twice subregular at (x,y), we have

$$f_x^\circ(x,y;v) = \mathbb{E}_{P_2}\left[\nabla\rho_y\left(W_2(W_1\xi_2+b_1)_+ + b_2\right)^\top \Upsilon(v,\xi_2)\right], \tag{4.2a}$$

$$f_x^{\circ\circ}(x,y;v) = \mathbb{E}_{P_2}\left[\Upsilon(v,\xi_2)^\top\nabla^2\rho_y(W_2(W_1\xi_2+b_1)_+ + b_2)\Upsilon(v,\xi_2)\right], \tag{4.2b}$$

where $v=\left(\mathrm{vec}(\overline{W}_1)^\top,\mathrm{vec}(\overline{W}_2)^\top,\bar{b}_1^\top,\bar{b}_2^\top\right)\in\mathbb{R}^n$, $\rho_y(\cdot):=\log\left(1-\frac{1}{1+\exp(y^\top(\cdot))}\right)$ *and*

$$\Upsilon(v,\xi_2) := W_2\left(\lim_{t\downarrow 0}\frac{((W_1+t\overline{W}_1)\xi_2+b_1+t\bar{b}_1)_+ - (W_1\xi+b_1)_+}{t}\right) + \overline{W}_2(W_1\xi_2+b_1)_+ + \bar{b}_2, \tag{4.3}$$

and

$$\mathrm{d}_y f(x,y)(w) = \left(\mathbb{E}_{P_1}\left[\nabla_y\log\left(D(y,\xi_1)\right)\right] + \mathbb{E}_{P_2}\left[\nabla_y\log\left(1-D(y,G(x,\xi_2))\right)\right]\right)^\top w, \tag{4.4a}$$

$$\mathrm{d}_y^2 f(x,y)(w) = w^\top\left(\mathbb{E}_{P_1}\left[\nabla_y^2\log\left(D(y,\xi_1)\right)\right] + \mathbb{E}_{P_2}\left[\nabla_y^2\log\left(1-D(y,G(x,\xi_2))\right)\right]\right)w, \tag{4.4b}$$

where $w\in\mathbb{R}^m$.

Proof. (i) Let $\rho_1(y)=\mathbb{E}_{P_1}\left[\log\left(D(y,\xi_1)\right)\right],\;\; \rho_2(x,y)=\mathbb{E}_{P_2}\left[\log\left(1-D(y,G(x,\xi_2))\right)\right]$. Since for any fixed $\xi_2\in\Xi_2$, $G(x,\xi_2)$ and $\log\left(1-\frac{1}{1+\exp(y^\top G(x,\xi_2))}\right)$ are locally Lipschitz continuous in $X\times Y$, the local Lipschitz continuity of $f(x,y)=\rho_1(y)+\rho_2(x,y)$ follows the continuous differentiability of *log* and *exp* functions. Moreover, the twice semidifferentiability follows directly from Example 3.3.

(ii) Since $\rho_y(\cdot)$ is twice continuously differentiable, we have

$$f_x^\circ(x,y;v) \overset{(a)}{=} f_x'(x,y;v) \overset{(b)}{=} \mathbb{E}_{P_2}\left[\nabla\rho_y\left(W_2(W_1\xi_2+b_1)_+ + b_2\right)^\top\Upsilon(v,\xi_2)\right],$$

where (a) follows from the Clarke regularity, (b) follows from Fatou-Lebesgue theorem and Example 3.3 and $\Upsilon(v,\xi_2)$ is defined in (4.3). Again, by twice subregularity, we have

$$f_x^{\circ\circ}(x,y;v) = f_x^{(2)}(x,y;v) = \mathbb{E}_{P_2}\left[\Upsilon(v,\xi_2)^\top \nabla^2\rho_y(W_2(W_1\xi_2+b_1)_+ + b_2)\Upsilon(v,\xi_2)\right].$$

Note that, for given x, ξ_1 and ξ_2, $D(y,\xi_1)$ and $D(y,G(x,\xi_2))$ are continuously differentiable with respect to y. By Lemma C.2 and the boundedness of Ξ_1 and Ξ_2, we know that $f(x,y)$ is continuously differentiable with respect to y. Moreover, we have (see Remark 3.2)

$$\begin{aligned} \mathrm{d}_y f(x,y)(w) &= \nabla_y f(x,y)^\top w = \left(\nabla_y\rho_1(y) + \nabla_y\rho_2(x,y)\right)^\top w \\ &= \left(\mathbb{E}_{P_1}\left[\nabla_y \log\left(D(y,\xi_1)\right)\right] + \mathbb{E}_{P_2}\left[\nabla_y \log\left(1 - D(y,G(x,\xi_2))\right)\right]\right)^\top w, \end{aligned}$$

where the last equality follows from Lemma C.2. Analogously, by applying Lemma C.2 to $\mathbb{E}_{P_1}\left[\nabla_y \log\left(D(y,\xi_1)\right)\right]$ and $\mathbb{E}_{P_2}\left[\nabla_y \log\left(1 - D(y,G(x,\xi_2))\right)\right]$, we can derive that $f(x,y)$ is twice continuously differentiable with respect to y and (see Remark 3.2)

$$\begin{aligned} \mathrm{d}_y^2 f(x,y)(w) &= w^\top \nabla_y^2 f(x,y) w \\ &= w^\top \left(\mathbb{E}_{P_1}\left[\nabla_y^2 \log\left(D(y,\xi_1)\right)\right] + \mathbb{E}_{P_2}\left[\nabla_y^2 \log\left(1 - D(y,G(x,\xi_2))\right)\right]\right) w. \end{aligned}$$

The proof is complete. ■

By directly using Proposition 4.2, we can apply Theorems 3.11 and 3.17 to problem (4.1).

Proposition 4.3. *Let $(\hat{x},\hat{y})$ be a local minimax point of problem* (4.1).

(i) Suppose the assumptions of Proposition 4.2 hold with $(x,y) = (\hat{x},\hat{y})$. Then the first-order necessary optimality conditions (3.2a)-(3.2b) *hold at $(\hat{x},\hat{y})$ with $f_x^{\circ}(\hat{x},\hat{y};v)$ and $\mathrm{d}_y f(\hat{x},\hat{y})(w)$ being given by* (4.2a) *and* (4.4a). *If, in addition, f is Clarke regular in a neighborhood of $(\hat{x},\hat{y})$, then the second-order necessary optimality conditions* (3.3a)-(3.3b) *hold at $(\hat{x},\hat{y})$ with $f_x^{\circ\circ}(\hat{x},\hat{y};v)$ and $\mathrm{d}_y^2 f(\hat{x},\hat{y})(w)$ being given by* (4.2b) *and* (4.4b).

(ii) If $\boldsymbol{\sigma}(\cdot)$ is twice continuously differentiable, then the first-order and second-order necessary optimality conditions (3.11a)-(3.11b) *and* (3.12a)-(3.12b) *hold at $(\hat{x},\hat{y})$.*

In Appendix D, we discuss the sample average approximation of the first-order and second-order stationary points of problem (4.1).

5. Conclusions.

Many nonconvex-nonconcave min-max problems in dada sciences do not have saddle points. In this paper, we provide sufficient conditions for the existence of global and local minimax points of constrained nonsmooth nonconvex-nonconcave min-max problem (1.1). Moreover, we give the first-order and second-order optimality conditions of local minimax points of problem (1.1), and use these conditions to define the first-order and second-order stationary points of (1.1). The relationships between saddle points, local saddle points, global minimax points, local minimax points, stationary points are summarized in Figure 1. Several examples are employed to illustrate our theoretical results. To demonstrate applications of these optimality conditions, we propose a method to verify the optimality conditions at any given point of generative adversarial network (4.1).

Appendix A. Example.

Example A.1. Let $X = [-1, 1]$ and $Y = [-1, 1]$. We consider

$$\min_{x\in[-1,1]} \max_{y\in[-1,1]} f(x, y) := -|x|^9 + \frac{3}{5}|x|^3 |y|^3 - |y|^5 .$$

Taking $\tau(\delta) = \frac{3}{5}(\sqrt{\delta})^3$, for any $|x| \le \delta$ and $|y| \le \delta$ with sufficiently small $\delta \in (0, 1)$, we have

$$-|y|^5 = f(0, y) \le f(0, 0) \le \max_{y\in[-\tau(\delta),\tau(\delta)]} -|x|^9 + \frac{3}{5}|x|^3 |y|^3 - |y|^5 = -|x|^9 + \frac{2}{5}\left(\frac{3}{5}\right)^4 (\sqrt{|x|})^{15},$$

where $\pm\frac{3}{5}(\sqrt{|x|})^3$ is the maximum of the above maximization problem. This implies that $(0, 0)$ is a local minimax point. Obviously, $f(x, y)$ is not differentiable at $(0, 0)$. In what follows, we examine the necessary optimality conditions in Theorem 3.11. Since $\mathcal{T}_X(0) = \mathbb{R}$ and $\mathcal{T}_Y(0) = \mathbb{R}$, we have for any $v \in \mathcal{T}_X(0)$ that

$$f_x^\circ(0, 0; v) = \limsup_{x'\to 0, t\downarrow 0} \frac{-|x' + tv|^9 + |x'|^9}{t} = 0,$$

which implies that $f_x^\circ(0, 0; v) = f_x'(0, 0; v)$, i.e., the Clarke regularity holds.

Similarly, we have for any $w \in \mathcal{T}_Y(0)$ that

$$\mathrm{d}_y f(0, 0)(w) = \liminf_{w'\to w, t\downarrow 0} \frac{f(0, tw') - f(0, 0)}{t} = \liminf_{w'\to w, t\downarrow 0} \frac{-|tw'|^5}{t} = 0.$$

Next consider the second-order optimality conditions. Note that $\mathcal{T}_X^\circ(0) = \mathbb{R}$ and for any fixed y', we have

$$\begin{aligned}\mathrm{d}_x f(0, y')(v) &= \liminf_{v'\to v, t\downarrow 0} \frac{f(tv', y') - f(0, y')}{t} \\ &= \liminf_{v'\to v, t\downarrow 0} \frac{-t^9 |v'|^9 + \frac{3}{5}t^3 |v'|^3 |y'|^3 - |y'|^5 + |y'|^5}{t} = 0\end{aligned}$$

for any v, which implies that $\{v : \mathrm{d}_x f(0, y')(v) = 0\} = \mathbb{R}$. Thus, for any $\delta > 0$

$$\mathcal{T}_X^\circ(0) \cap \{v : \mathrm{d}_x f(0, y')(v) = 0, \forall y' \in \mathbb{B}(0, \delta) \cap Y\} = \mathbb{R}.$$

Notice that

$$\begin{aligned}f_x^{\circ\circ}(0, 0; v) &= \limsup_{\substack{x'\to 0\\ t\downarrow 0, \delta\downarrow 0}} \frac{f(x' + \delta v + tv, 0) - f(x' + \delta v, 0) - f(x' + tv, 0) + f(x', 0)}{\delta t} \\ &= \limsup_{\substack{x'\to 0\\ t\downarrow 0, \delta\downarrow 0}} \frac{-|x' + \delta v + tv|^9 + |x' + \delta v|^9 + |x' + tv|^9 - |x'|^9}{\delta t} \ge 0\end{aligned}$$

for any $v \in \mathbb{R}$. Similarly, we have $\mathcal{T}_Y^\circ(0) \cap \{w : \mathrm{d}_y f(0, 0)(w) = 0\} = \mathbb{R}$ and

$$\mathrm{d}_y^2 f(0, 0)(w) = \liminf_{w'\to w, t\downarrow 0} \frac{f(0, tw') - f(0, 0) - t\mathrm{d}_y f(0, 0)(w')}{\frac{1}{2}t^2} = \liminf_{w'\to w, t\downarrow 0} \frac{-|tw'|^5}{\frac{1}{2}t^2} = 0$$

This manuscript is for review purposes only.

for any $w \in \mathbb{R}$.

Appendix B. The polyhedral case.

If both X and Y are polyhedral, we can replace $\mathrm{cl}\{w : \exists\delta > 0, w \in \Gamma_1^\circ(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\}$ and $\mathrm{cl}\Gamma_2^\circ(\hat{x}, \hat{y})$ in Theorem 3.17 by $\mathrm{cl}\{w : \exists\delta > 0, w \in \Gamma_1(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\}$ and $\Gamma_2(\hat{x}, \hat{y})$, respectively (see Lemma 3.16). Specially, we consider that X and Y are defined as follows:

$$X = \{x \in \mathbb{R}^n : Ax \le b\} \text{ and } Y = \{y \in \mathbb{R}^m : Cy \le d\}, \tag{B.1}$$

where $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{q\times m}$ and $d \in \mathbb{R}^q$.

The following proposition establishes the relationship between tangent/normal cones and algebra systems when X and Y are defined in (B.1).

Proposition B.1 ([15]). *Let X and Y be defined in (B.1). Then we have*

$$\mathcal{T}_X(x) = \left\{\lambda \in \mathbb{R}^n : -A_i^\top \lambda \ge 0,\ \forall i \in \mathcal{A}_X(x)\right\}, \mathcal{T}_Y(y) = \left\{\mu \in \mathbb{R}^m : -C_j^\top \mu \ge 0,\ \forall j \in \mathcal{A}_Y(y)\right\}$$
$$\mathcal{N}_X(x) = \left\{-\sum_{i=1}^{p} \alpha_i A_i : \alpha \in \mathcal{N}_{\mathbb{R}_+^p}(b - Ax)\right\},\ \mathcal{N}_Y(y) = \left\{-\sum_{j=1}^{q} \beta_j C_j : \beta \in \mathcal{N}_{\mathbb{R}_+^q}(d - Cy)\right\},$$

where A_i is the ith row vector of matrix A and C_j is the jth row vector of matrix C respectively for $i = 1, \cdots, p$ and $j = 1, \cdots, q$, and $\mathcal{A}_X(x)$ and $\mathcal{A}_Y(y)$ are active sets of X at x and Y at y, respectively.

Theorem B.2. *Let the tuple $(\hat{x}, \hat{y}) \in X \times Y$ be a local minimax point of problem (1.1) with X and Y being defined in (B.1). Then there exist multipliers $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ such that*

$$\begin{cases} \nabla_x f(\hat{x}, \hat{y}) - \sum_{i=1}^p \alpha_i A_i = 0, \quad -\nabla_y f(\hat{x}, \hat{y}) - \sum_{j=1}^q \beta_j C_j = 0, \\ \alpha \in \mathcal{N}_{\mathbb{R}_+^p}(b - A\hat{x}), \quad \beta \in \mathcal{N}_{\mathbb{R}_+^q}(d - C\hat{y}). \end{cases} \tag{B.2}$$

If, moreover, f is twice continuously differentiable, we have, for any $\delta > 0$, that

$$\begin{cases} \left\langle v, \nabla^2_{xx} f(\hat{x}, \hat{y}) v\right\rangle \ge 0 \textit{ for all} \\ \qquad v \in \left\{\lambda \in \mathcal{T}_X(\hat{x}) : \exists\delta > 0,\ \lambda^\top \nabla_x f(\hat{x}, y') = 0 \textit{ for } y' \in \mathbb{B}(\hat{y}, \delta)\right\}, \\ \left\langle w, \nabla^2_{yy} f(\hat{x}, \hat{y}) w\right\rangle \le 0 \textit{ for all } w \in \left\{\mu \in \mathcal{T}_Y(\hat{y}) : \mu^\top \nabla_y f(\hat{x}, \hat{y}) = 0\right\}. \end{cases} \tag{B.3}$$

Proof. We know from (3.11) of Theorem 3.17 that the following first-order optimality necessary condition holds: $0 \in \nabla_x f(\hat{x}, \hat{y}) + \mathcal{N}_X(\hat{x})$ and $0 \in -\nabla_y f(\hat{x}, \hat{y}) + \mathcal{N}_Y(\hat{y})$. This together with the specific reformulations of $\mathcal{N}_X(x)$ and $\mathcal{N}_Y(y)$ in Proposition B.1, we obtain (B.2) directly.

Next, we focus on (B.3). Analogously, we know from (3.12) of Theorem 3.17 that

$$\begin{cases} \left\langle v, \nabla^2_{xx} f(\hat{x}, \hat{y}) v\right\rangle \ge 0 \text{ for all } v \in \mathrm{cl}\left\{\bar{v} : \exists\delta > 0, \bar{v} \in \Gamma_1^\circ(\hat{x}, y'), \forall y' \in \mathbb{B}(\hat{y}, \delta)\right\}, \\ \left\langle w, \nabla^2_{yy} f(\hat{x}, \hat{y}) w\right\rangle \le 0 \text{ for all } w \in \mathrm{cl}\Gamma_2^\circ(\hat{x}, \hat{y}) \end{cases} \tag{B.4}$$

holds. Since X and Y are polyhedral, we know from Lemma 3.16 that $\Gamma_1^\circ(x,y) = \mathrm{cl}\Gamma_1^\circ(x,y) = \Gamma_1(x,y)$ and $\Gamma_2^\circ(x,y) = \mathrm{cl}\Gamma_2^\circ(x,y) = \Gamma_2(x,y)$. Thus, (B.4) can be equivalently rewritten as

$$\text{(B.5)} \qquad \begin{cases} \left\langle v, \nabla_{xx}^2 f(\hat{x},\hat{y})v \right\rangle \geq 0 \text{ for all } v \in \mathrm{cl}\left\{\bar{v} : \exists\delta > 0, \bar{v} \in \Gamma_1(\hat{x},y'), \forall y' \in \mathbb{B}(\hat{y},\delta)\right\}, \\ \left\langle w, \nabla_{yy}^2 f(\hat{x},\hat{y})w \right\rangle \leq 0 \text{ for all } w \in \Gamma_2(\hat{x},\hat{y}). \end{cases}$$

Note that $\Gamma_1(x,y) = \{v \in \mathcal{T}_X(x) : v \perp \nabla_x f(x,y)\}$ and $\Gamma_2(x,y) = \{w \in \mathcal{T}_Y(y) : w \perp \nabla_y f(x,y)\}$. This, together with (B.5) and the reformulations of $\mathcal{T}_X(x)$ and $\mathcal{T}_Y(y)$ in Proposition B.1, verifies (B.3). ∎

We call (B.2) the first-order KKT system of problem (1.1) and (B.2)-(B.3) the second-order KKT system of problem (1.1).

Appendix C. Four lemmas.

Consider the minimization problem

$$\text{(C.1)} \qquad \min_{x \in \mathcal{X}} \; g(x),$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a compact and convex set and $g : \mathcal{X} \to \mathbb{R}$ is continuous, and its a sequence of perturbation problems

$$\text{(C.2)} \qquad \min_{x \in \mathcal{X}} \; \tilde{g}_k(x),$$

where $\tilde{g}_k : \mathcal{X} \to \mathbb{R}$ are continuous for $k \in \mathbb{N}$.

Lemma C.1. *Let v^*, $\mathcal{S}^*$ and v_k^*, $\mathcal{S}_k^*$ denote the optimal values and the optimal solution sets of problems (C.1) and (C.2), respectively. Assume $\sup_{x\in\mathcal{X}} |\tilde{g}_k(x) - g(x)| \to 0$ as $k \to \infty$. Then (i) v^*, v_k^* are finite and $\mathcal{S}^*$, $\mathcal{S}_k^*$ are nonempty; (ii) $\sup_{x\in\mathcal{S}_k^*} \mathrm{d}(x, \mathcal{S}^*) \to 0$ as $k \to \infty$.*

Proof. (i) It follows from that $\mathcal{X}$ is a compact and convex set and $g, \tilde{g}_k$ are continuous. (ii) We give the proof by contradiction. Assume that there exists an $\epsilon_0 > 0$ such that $\sup_{x\in\mathcal{S}_{k_l}^*} \mathrm{d}(x,\mathcal{S}^*) \geq \epsilon_0$, where $\{\mathcal{S}_{k_l}^*\}_{l\geq 1}$ is a subsequence of $\{\mathcal{S}_k^*\}_{k\geq 1}$. Thus, we can select a sequence $\{x_{k_l}\}_{l\geq 1}$ with $x_{k_l} \in \mathcal{S}_{k_l}^*$ such that $\mathrm{d}(x_{k_l}, \mathcal{S}^*) \geq \frac{\epsilon_0}{2}$, $\forall l \in \mathbb{N}$. Due to the boundedness of feasible set $\mathcal{X}$, we know that the sequence $\{x_{k_l}\}_{l\geq 1}$ is bounded, and without loss of generality, we assume that $x_{k_l} \to \bar{x}$ as $l \to \infty$.

$$v_{k_l}^* - g(\bar{x}) = \tilde{g}_{k_l}(x_{k_l}) - g(\bar{x}) = \tilde{g}_{k_l}(x_{k_l}) - g(x_{k_l}) + g(x_{k_l}) - g(\bar{x}).$$

Since $\limsup_{l\to\infty} v_{k_l}^* = \lim_{l\to\infty} v_{k_l}^* = v^*$, we have

$$v^* - g(\bar{x}) = \limsup_{l\to\infty} \left(v_{k_l}^* - g(\bar{x})\right) \geq \liminf_{l\to\infty} \left(\tilde{g}_{k_l}(x_{k_l}) - g(x_{k_l})\right) + \liminf_{l\to\infty} \left(g(x_{k_l}) - g(\bar{x})\right).$$

Note that

$$\left|\liminf_{l\to\infty} \left(\tilde{g}_{k_l}(x_{k_l}) - g(x_{k_l})\right)\right| \leq \sup_{x\in X} |\tilde{g}_{k_l}(x) - g(x)| \to 0 \quad \text{and} \quad \liminf_{l\to\infty} g(x_{k_l}) - g(\bar{x}) \geq 0,$$

which implies that $v^* - g(\bar{x}) \geq 0$ and thus $\bar{x} \in \mathcal{S}^*$. This contradicts with $\frac{\epsilon_0}{2} \leq \mathrm{d}(x_{k_l}, \mathcal{S}^*) \to \mathrm{d}(\bar{x}, \mathcal{S}^*) = 0$. Therefore, $\sup_{x\in\mathcal{S}_k^*} \mathrm{d}(x,\mathcal{S}^*) \to 0$ as $k \to \infty$. ∎

Lemma C.2 ([33, Theorem 7.57]). *Let $U \subseteq \mathbb{R}^n$ be an open set, X be a nonempty compact subset of U and $F : U \times \Xi \to \mathbb{R}$ be a random function. Suppose that: (i) $\{F(x,\xi)\}_{x\in X}$ is dominated by an integrable function; (ii) there exists an integrable function $C(\xi)$ such that $|F(x',\xi) - F(x,\xi)| \le C(\xi)\,\|x' - x\|$ for all $x', x \in U$ and a.e. $\xi \in \Xi$; (iii) for every $x \in X$ the function $F(\cdot,\xi)$ is continuously differentiable at x w.p.1. Then (a) the expectation function $f(x)$ is finite valued and continuously differentiable on X, and (b) for all $x \in X$ the corresponding derivatives can be taken inside the integral, i.e., $\nabla f(x) = \mathbb{E}[\nabla_x F(x,\xi)]$.*

Lemma C.3. *Suppose that g is twice differentiable at $\bar{x} \in \mathcal{X}$. Let $\Gamma^\circ(\bar{x}) := \{w \in \mathcal{T}^\circ_{\mathcal{X}}(\bar{x}) : w \perp \nabla g(\bar{x})\}$ and $\Gamma(\bar{x}) := \{w \in \mathcal{T}_{\mathcal{X}}(\bar{x}) : w \perp \nabla g(\bar{x})\}$. Then $\Gamma^\circ(\bar{x})$ and $\Gamma(\bar{x})$ are convex cones and (i) If $\bar{x}$ is a local minimum point of (C.1), then*

$$0 \in \nabla g(\bar{x}) + \mathcal{N}_{\mathcal{X}}(\bar{x}) \quad \text{and} \quad \langle w, \nabla^2 g(\bar{x}) w\rangle \ge 0 \text{ for all } w \in \text{cl}\Gamma^\circ(\bar{x}). \tag{C.3}$$

(ii) If the conditions in (C.3) hold by replacing $\text{cl}\Gamma^\circ(\bar{x})$ by $\Gamma(\bar{x})$ and "$\ge$" by "$>$" for $w \ne 0$, then $\bar{x}$ is a local minimum point of (C.1).

Proof. (i) For any $w \in \Gamma^\circ(\bar{x})$ with $\|w\| = 1$, there exists a sequence $\{t_k\}_{k\ge1}$ with $t_k \downarrow 0$ as $k \to \infty$ such that $0 \le g(\bar{x}+t_k w) - g(\bar{x}) = t_k \nabla g(\bar{x})^\top w + \frac{t_k^2}{2} w^\top \nabla^2 g(\bar{x}) w + t_k^2 \|w\|^2 o(1)$. Dividing t_k^2 in both sides gives $w^\top \nabla^2 g(\bar{x}) w \ge 0$, since $\nabla g(\bar{x})^\top w = 0$. Hence (C.3) holds.

(ii) We assume by contradiction that $\bar{x}$ is not a local minimum point. Then there exists a sequence $\{x^k\}_{k\ge1} \subseteq \mathcal{X}$ with $x^k \to \bar{x}$ as $k \to \infty$ such that $g(x^k) < g(\bar{x})$. Let $t_l = \|x^{k_l} - \bar{x}\|$ and $w_l = \frac{x^{k_l}-\bar{x}}{\|x^{k_l}-\bar{x}\|} \in \mathcal{T}^\circ_{\mathcal{X}}(\bar{x})$. Then $g(x^k) = g(\bar{x}) + t_l \nabla g(\bar{x})^\top w_l + \frac{t_l^2}{2} w_l^\top \nabla^2 g(\bar{x}) w_l + t_l^2 \|w_l\|^2 o(1)$. Without loss of generality, we assume that $w_l \to \bar{w}$ as $l \to \infty$. Then $\bar{w} \in \text{cl}\Gamma^\circ(\bar{x}) \subseteq \Gamma(\bar{x})$.

If there exists a subsequence $\{k_l\}_{l\ge1}$ such that $\nabla g(\bar{x})^\top w_l = 0$, then $\frac{1}{2} w_l^\top \nabla^2 g(\bar{x}) w_l > 0$ and $\bar{w}^\top \nabla^2 g(\bar{x}) \bar{w} > 0$, which implies $g(x^k) \ge g(\bar{x})$. This leads to a contradiction.

If there exists a subsequence $\{k_l\}_{l\ge1}$ such that $\nabla g(\bar{x})^\top w_l > 0$, then we have $g(x^k) \ge g(\bar{x})$ if $\nabla g(\bar{x})^\top \bar{w} > 0$, and $\bar{w}^\top \nabla^2 g(\bar{x}) \bar{w} > 0$ if $\nabla g(\bar{x})^\top \bar{w} = 0$ (i.e., $\bar{w} \in \Gamma(\bar{x})$), which implies $g(x^k) \ge g(\bar{x})$. This also leads to a contradiction. ∎

Lemma C.4. *Suppose that g is twice semidifferentiable at $\bar{x} \in \mathcal{X}$ and $\mathcal{X}$ is a nonempty, closed and convex set. If $\mathrm{d}g(\bar{x})(v) \ge 0$ for all $v \in \mathcal{T}_{\mathcal{X}}(\bar{x})$ and $0 \ne v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \{w : \mathrm{d}g(\bar{x})(w) = 0\}$ implies that $\mathrm{d}^2 g(\bar{x})(v) > 0$, then $\bar{x}$ is a local minimum point of problem (C.1).*

Proof. Let $\bar{g} := g + \delta_{\mathcal{X}}$. Consider the unconstrained minimization problem $\min_{x\in\mathbb{R}^n} \bar{g}(x)$, which is equivalent to constrained minimization problem (C.1). By applying [32, Theorem 13.24] to the unconstrained minimization problem, we complete the proof. ∎

Appendix D. The sample average approximation. We discuss the sample average approximation (SAA) of a first-order and a second-order stationary points of problem (4.1).

To this end, we assume that $\boldsymbol{\sigma}(\cdot)$ is twice continuously differentiable. Let $X = [a, b]$ and $Y = [c, d]$ where $a, b \in \mathbb{R}^n, c, d \in \mathbb{R}^m$, $a < b$, and $c < d$ with $n = (s+1)(s_1 + s_2)$ and $m = s_1$.

Denote $\{\xi_1^j\}_{j=1}^N$ and $\{\xi_2^j\}_{j=1}^N$ the independent identically distributed (iid) samples of ξ_1

and ξ_2, respectively. We consider the following min-max problem

$$\begin{aligned}\min_{x\in X}\max_{y\in Y}\ \hat{f}_N(x,y) :=&\frac{1}{N}\sum_{i=1}^{N}\left(\log\left(\frac{1}{1+\exp(y^\top \xi_1^i)}\right)\right.\\ &\left.+\log\left(1-\frac{1}{1+\exp(y^\top(W_2\boldsymbol{\sigma}(W_1\xi_2^i+b_1)+b_2))}\right)\right).\end{aligned} \tag{D.1}$$

Use the existing automatic differentiation technique, such as back-propagation, we can compute $\nabla_x \hat{f}_N(x,y)$, $\nabla_y \hat{f}_N(x,y)$, $\nabla^2_{xx}\hat{f}_N(x,y)$, $\nabla^2_{yy}\hat{f}_N(x,y)$. Moreover, we have

$$\mathcal{T}_X(x)=\mathcal{T}_X^\circ(x)=\left\{v\in\mathbb{R}^n\ :\ v_i\in\begin{cases}[0,\infty), & \text{if } x_i=a_i\\ (-\infty,\infty), & \text{if } a_i<x_i<b_i\\ (-\infty,0], & \text{if } x_i=b_i\end{cases}\right\},$$

$$\mathcal{T}_Y(y)=\mathcal{T}_Y^\circ(y)=\left\{w\in\mathbb{R}^m\ :\ w_j\in\begin{cases}[0,\infty), & \text{if } y_j=c_j\\ (-\infty,\infty), & \text{if } c_j<y_j<d_j\\ (-\infty,0], & \text{if } y_j=d_j\end{cases}\right\}$$

and

$$\begin{aligned}\Gamma_1^\circ(x,y)=\Gamma_1(x,y)=\{v\in\mathcal{T}_X(x): v\perp\nabla_x\hat{f}_N(x,y)\},\\ \Gamma_2^\circ(x,y)=\Gamma_2(x,y)=\{w\in\mathcal{T}_Y(y): w\perp\nabla_y\hat{f}_N(x,y)\}.\end{aligned}$$

By Theorem 3.17, if $(\hat{x},\hat{y})$ is a local minimax point of problem (D.1), then $(\hat{x},\hat{y})$ must satisfy the first-order and second-order optimality conditions:

$$\begin{cases}(\nabla_x\hat{f}_N(\hat{x},\hat{y}))_i\geq 0, & \text{if } x_i=a_i;\\ (\nabla_x\hat{f}_N(\hat{x},\hat{y}))_i=0, & \text{if } a_i<x_i<b_i;\\ (\nabla_x\hat{f}_N(\hat{x},\hat{y}))_i\leq 0, & \text{if } x_i=b_i\end{cases}\ \text{and}\ \begin{cases}(\nabla_y\hat{f}_N(\hat{x},\hat{y}))_j\leq 0, & \text{if } y_j=c_j;\\ (\nabla_y\hat{f}_N(\hat{x},\hat{y}))_j=0, & \text{if } c_j<y_j<d_j;\\ (\nabla_y\hat{f}_N(\hat{x},\hat{y}))_j\geq 0, & \text{if } y_j=d_j\end{cases}$$

for $i=1,\dots,n$, $j=1,\dots,m$, and

$$\begin{aligned}&\left\langle v,\nabla^2_{xx}\hat{f}_N(\hat{x},\hat{y})v\right\rangle\geq 0 \text{ for all } v\in\left\{\bar{v}:\exists\delta>0,\bar{v}\in\Gamma_1(\hat{x},y'),\forall y'\in\mathbb{B}(\hat{y},\delta)\right\},\\ &\left\langle w,\nabla^2_{yy}\hat{f}_N(\hat{x},\hat{y})w\right\rangle\leq 0 \text{ for all } w\in\Gamma_2(\hat{x},\hat{y}).\end{aligned}$$

The following proposition tells that the above procedures can ensure an exponential rate of convergence with respect to sample size N.

Proposition D.1. *Let $\boldsymbol{\sigma}(\cdot)$ be twice continuously differentiable. If (x_N,y_N) is a first-order (second-order) stationary point of problem (D.1) with iid samples $\{\xi_1^j\}_{j=1}^N$ and $\{\xi_2^j\}_{j=1}^N$ of ξ_1 and ξ_2 respectively, then (x_N,y_N) converges to a first-order (second-order) stationary point of problem (4.1) exponentially with respect to N.*

This manuscript is for review purposes only.

Proof. Denote

$$h(z) = \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}, \qquad H(z) = \begin{pmatrix} \sup_{v\in\mathcal{V}(x,y)} \langle v, -\nabla^2_{xx} f(x,y) v\rangle \\ \sup_{w\in\mathcal{W}(x,y)} \langle w, \nabla^2_{yy} f(x,y) w\rangle \end{pmatrix},$$
$$\hat{h}_N(z) = \begin{pmatrix} \nabla_x \hat{f}_N(x,y) \\ -\nabla_y \hat{f}_N(x,y) \end{pmatrix}, \qquad \widehat{H}_N(z) = \begin{pmatrix} \sup_{v\in\mathcal{V}(x,y)} \left\langle v, -\nabla^2_{xx} \hat{f}_N(x,y) v\right\rangle \\ \sup_{w\in\mathcal{W}(x,y)} \left\langle w, \nabla^2_{yy} \hat{f}_N(x,y) w\right\rangle \end{pmatrix},$$

where $z = (x^\top, y^\top)^\top$, $\mathcal{V}(x,y) := \mathbb{B}(0,1) \cap \cup_{\delta>0}\mathrm{cl}\,\{\bar{v} : \exists\delta > 0, \bar{v} \in \Gamma_1^\circ(x,y'), \forall y' \in \mathbb{B}(y,\delta)\}$ and $\mathcal{W}(x,y) := \mathbb{B}(0,1) \cap \mathrm{cl}\Gamma_2^\circ(x,y)$.

According to the twice continuous differentiability of f (see Proposition 4.2) and the boundedness of Ξ_1 and Ξ_2, we have $\hat{h}_N(z) \to h(z)$ and $\widehat{H}_N(z) \to H(z)$ exponentially fast uniformly in any compact subset of $\mathcal{Z} \subseteq Z := X \times Y$ ([33, Theorem 7.73]). That is, for any given $\epsilon > 0$, there exist $C = C(\epsilon)$ and $\beta = \beta(\epsilon)$, such that

$$\mathrm{Prob}\left\{\sup_{z\in\mathcal{Z}} \left\|\hat{h}_N(z) - h(z)\right\| \geq \epsilon\right\} \leq Ce^{-N\beta} \text{ and } \mathrm{Prob}\left\{\sup_{z\in\mathcal{Z}} \left|\widehat{H}_N(z) - H(z)\right| \geq \epsilon\right\} \leq Ce^{-N\beta}.$$

Without loss of generality, we assume that $z_N = (x_N^\top, y_N^\top)^\top \in \mathcal{Z}$. Denote the following general growth functions:

$$\psi_1(\tau) := \inf\{\mathrm{d}(0, h(z) + \mathcal{N}_Z(z)) : z \in \mathcal{Z},\ \mathrm{d}(z, \mathcal{S}_1) \geq \tau\},$$
$$\psi_2(\tau) := \inf\{\|(H(z))_+\| : z \in \mathcal{Z},\ \mathrm{d}(z, \mathcal{S}_2) \geq \tau\},$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are the sets satisfying (3.11a)-(3.11b) and (3.12a)-(3.12b), respectively, and "d" denotes the distance from a point to a set. Let the related functions $\Psi_1(t) := \psi_1^{-1}(t) + t$ and $\Psi_2(t) := \psi_2^{-1}(t) + t$, where $\psi_i^{-1}(t) := \sup\{\tau : \psi_i(\tau) \leq \eta\}$ for $i = 1, 2$, which satisfy $\Psi_i(t) \to 0$ as $t \downarrow 0$ for $i = 1, 2$.

Then, by a conventional discussion (see e.g. [5]), we have

$$\mathrm{d}(z_N, \mathcal{S}_1) \leq \Psi_1\left(\sup_{z\in Z} \left\|\hat{h}_N(z) - h(z)\right\|\right) \text{ and } \mathrm{d}(z_N, \mathcal{S}_2) \leq \Psi_2\left(\sup_{z\in Z} \left|\widehat{H}_N(z) - H(z)\right|\right).$$

Thus, we have $\mathrm{Prob}\{\mathrm{d}(z_N, \mathcal{S}_1) \geq \Psi_1(\epsilon)\} \leq Ce^{-N\beta}$ and $\mathrm{Prob}\{\mathrm{d}(z_N, \mathcal{S}_2) \geq \Psi_2(\epsilon)\} \leq Ce^{-N\beta}$, which shows that z_N converges to a first-order stationary point in $\mathcal{S}_1$ (or a first-order stationary point in $\mathcal{S}_2$) exponentially with respect to N. ∎

Acknowledgments. The authors would like to thank editors and two anonymous referees for their insightful and detailed comments.

REFERENCES

[1] L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann, *Local saddle point optimization: A curvature exploitation approach*, in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 486–495.

[2] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[3] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, *Generalization and equilibrium in generative adversarial nets (GANs)*, in International Conference on Machine Learning, PMLR, 2017, pp. 224–232.

[4] Q. Ba and J.-S. Pang, *Exact penalization of generalized Nash equilibrium problems*, Oper. Res., 70 (2022), pp. 1448–1464.

[5] X. Chen, A. Shapiro, and H. Sun, *Convergence analysis of sample average approximation of two-stage stochastic generalized equations*, SIAM J. Optim., 29 (2019), pp. 135–161.

[6] Y. Chen, G. Lan, and Y. Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

[8] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, *Generative adversarial networks: An overview*, IEEE Signal Process. Mag., 35 (2018), pp. 53–65.

[10] Y. Cui, T.-H. Chang, M. Hong, and J.-S. Pang, *A study of piecewise linear-quadratic programs*, J. Optim. Theory Appl., 186 (2020), pp. 523–553.

[11] Y.-H. Dai and L. Zhang, *Optimality conditions for constrained minimax optimization*, CSIAM Transactions on Applied Mathematics, 1 (2020), pp. 296–315.

[12] C. Daskalakis and I. Panageas, *The limit points of (optimistic) gradient descent in min-max optimization*, in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 9236–9246.

[13] S. Dempe and A. Zemkoho, *Bilevel Optimization*, vol. 161, Springer Cham, Switzerland, 2020.

[14] F. Facchinei and C. Kanzow, *Generalized Nash equilibrium problems*, 4OR, 5 (2007), pp. 173–210.

[15] F. Facchinei and J.-S. Pang, *Finite-dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2007.

[16] I. Goodfellow, *NIPS 2016 tutorial: Generative adversarial networks*. arXiv:1701.00160, 2016.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 2672–2680.

[18] B. Grimmer, H. Lu, P. Worah, and V. Mirrokni, *Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems*, in International Conference on Algorithmic Learning Theory, PMLR, 2022, pp. 465–487.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, *Improved training of Wasserstein GANs*, in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5767–5777.

[20] J. Jiang and X. Chen, *Pure characteristics demand models and distributionally robust mathematical programs with stochastic complementarity constraints*, Math. Program., 198 (2023), pp. 1449–1484.

[21] C. Jin, P. Netrapalli, and M. Jordan, *What is local optimality in nonconvex-nonconcave minimax optimization?*, in International Conference on Machine Learning, PMLR, 2020, pp. 4880–4889.

[22] Q. Lin, M. Liu, H. Rafique, and T. Yang, *Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality*. arXiv:1810.10207, 2018.

[23] M. Liu, H. Rafique, Q. Lin, and T. Yang, *First-order convergence theory for weakly-convex-weakly-concave min-max problems.*, J. Machine Learn. Res., 22 (2021), pp. 1–34.

[24] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen, *Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications*, IEEE Trans. Signal Process., 68 (2020), pp. 3676–3691.

[25] E. Mazumdar, L. J. Ratliff, and S. S. Sastry, *On gradient-based learning in continuous games*, SIAM J. Math. Data Sci., 2 (2020), pp. 103–131.

[26] M. Meitz, *Statistical inference for generative adversarial networks*. arXiv:2104.10601, 2021.

[27] J. Nash Jr, *Non-cooperative games*, in Essays on Game Theory, Edward Elgar Publishing, 1996, pp. 22–33.

[28] A. Nemirovski, *Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.

[29] J. V. Neumann, *Zur theorie der gesellschaftsspiele*, Mathematische Annalen, 100 (1928), pp. 295–320.

[30] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, *Robust optimization over multiple domains*, Proceedings of the AAAI Conference on Artificial Intelligence, 33 (2019), pp. 4739–4746.
[31] H. Rafique, M. Liu, Q. Lin, and T. Yang, *Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning*, Optim. Methods Softw., 37 (2022), pp. 1087–1121.
[32] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, New York, 2009.
[33] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2014.
[34] Y. Wang, *A mathematical introduction to generative adversarial nets (GAN)*. arXiv:2009.00169, 2020.
[35] J. Zhang, M. Hong, M. Wang, and S. Zhang, *Generalization bounds for stochastic saddle point problems*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 568–576.
[36] J. Zhang, M. Hong, and S. Zhang, *On lower iteration complexity bounds for the convex concave saddle point problems*, Math. Program., 194 (2022), pp. 901–935.