A. Relation to Mixed Strategy Games

In contrast to pure strategies where each player plays a single action, game theorists have also considered mixed strategies where each player is allowed to play a randomized action sampled from a probability measure \( \mu \in \mathcal{P}(X) \) or \( \nu \in \mathcal{P}(Y) \). Then, the payoff function becomes an expected value \( \mathbb{E}_{x \sim \mu, y \sim \nu} f(x, y) \). For mixed strategy games, even if function is nonconvex-nonconcave, the following minimax theorem still holds.

**Proposition 32 (Glicksberg, 1952).** Assume that the function \( f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \) is continuous and that \( \mathcal{X} \subset \mathbb{R}^{d_1}, \mathcal{Y} \subset \mathbb{R}^{d_2} \) are compact. Then

\[
\min_{\mu \in \mathcal{P}(X)} \max_{\nu \in \mathcal{P}(Y)} \mathbb{E}_{(x, y) \sim \mu \times \nu} f(x, y) = \max_{\nu \in \mathcal{P}(Y)} \min_{\mu \in \mathcal{P}(X)} \mathbb{E}_{(x, y) \sim \mu \times \nu} f(x, y).
\]

This implies that the order which player goes first is no longer important in this setting, and there is no intrinsic difference between simultaneous games and sequential games if mixed strategies are allowed.

Similar to the concept of (pure strategy) Nash equilibrium in Definition 1, we can define mixed strategy Nash equilibrium as

**Definition 33.** A probability measure \((\mu^*, \nu^*)\) is a mixed strategy Nash equilibrium of \( f \), if for any measure \((\mu, \nu)\) in \( \mathcal{P}(X) \times \mathcal{P}(Y) \), we have

\[
\mathbb{E}_{x \sim \mu^*, y \sim \nu^*} f(x, y) \leq \mathbb{E}_{x \sim \mu, y \sim \nu} f(x, y) \leq \mathbb{E}_{x \sim \mu^*, y \sim \nu} f(x, y).
\]

Unlike pure strategy Nash equilibrium, the existence of mixed strategy Nash equilibrium in this setting is always guaranteed by Glicksberg (1952).

One challenge for finding mixed strategy equilibria is that it requires optimizing over a space of probability measures, which is of infinite dimension. However, we can show that finding approximate mixed strategy Nash equilibria for Lipschitz games can be reduced to finding a global solution of a “augmented” pure strategy sequential games, which is a problem of polynomially large dimension.

**Definition 34.** Let \((\mu^*, \nu^*)\) be a mixed strategy Nash equilibrium. A probability measure \((\mu^1, \nu^1)\) is an \( \epsilon \)-approximate mixed strategy Nash equilibrium if:

\[
\forall \nu' \in \mathcal{P}(Y), \quad \mathbb{E}_{(\mu^1, \nu')} f(x, y) \leq \mathbb{E}_{(\mu^*, \nu^*)} f(x, y) + \epsilon \\
\forall \mu' \in \mathcal{P}(X), \quad \mathbb{E}_{(\mu', \nu^1)} f(x, y) \geq \mathbb{E}_{(\mu^*, \nu^*)} f(x, y) - \epsilon.
\]

**Theorem 35.** Assume that function \( f \) is L-Lipschitz, and the diameters of \( \mathcal{X} \) and \( \mathcal{Y} \) are at most \( D \). Let \((\mu^*, \nu^*)\) be a mixed strategy Nash equilibrium. Then there exists an absolute constant \( c \), for any \( \epsilon > 0 \), such that if \( N \geq c \cdot d_2(2LD/\epsilon)^2 \log(\mathbb{D}/\epsilon) \), we have:

\[
\min_{(x_1, \ldots, x_N) \in \mathcal{X}^N} \max_{y \in \mathcal{Y}} \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) \leq \mathbb{E}_{(\mu^*, \nu^*)} f(x, y) + \epsilon.
\]

Intuitively, Theorem 35 holds because function \( f \) is Lipschitz, \( \mathcal{Y} \) is a bounded domain, and thus we can establish uniform convergence of the expectation of \( f(\cdot, y) \) to its average over \( N \) samples for all \( y \in \mathcal{Y} \) simultaneously. A similar argument was made in (Arora et al., 2017).

Theorem 35 implies that in order to find an approximate mixed strategy Nash equilibrium, we can solve a large minimax problem with objective \( F(X, y) := \sum_{i=1}^{N} f(x_i, y)/N \). The global minimax solution \( X^* = (x_1^*, \ldots, x_n^*) \) gives a empirical distribution \( \hat{\mu}^* = \sum_{i=1}^{N} \delta(x - x_i^*)/N \), where \( \delta(\cdot) \) is the Dirac delta function. By symmetry, we can also solve the corresponding maximin problem to find \( \hat{\nu}^* \). It can be shown that \((\hat{\mu}^*, \hat{\nu}^*)\) is an \( \epsilon \)-approximate mixed strategy Nash equilibrium. That is, approximate mixed strategy Nash can be found by finding two global minimax points.

**Proof of Theorem 35.** Note that WLOG, the second player can always play pure strategy. That is,

\[
\min_{\mu \in \mathcal{P}(X)} \max_{\nu \in \mathcal{P}(Y)} \mathbb{E}_{x \sim \mu, y \sim \nu} f(x, y) = \min_{\mu \in \mathcal{P}(X)} \max_{y \in \mathcal{Y}} \mathbb{E}_{x \sim \mu} f(x, y)
\]

Therefore, we only need to solve the problem of RHS. Suppose the minimum over \( \mathcal{P}(X) \) is achieved at \( \mu^* \). First, sample \((x_1, \ldots, x_N)\) i.i.d from \( \mu^* \), and note \( \max_{x_1, x_2 \in \mathcal{X}} |f(x_1, y) - f(x_2, y)| \leq LD \) for any fixed \( y \). Therefore by Hoeffding
This implies:
\[ \Pr \left( \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) - \mathbb{E}_{x \sim \mu^*} f(x, y) \geq t \right) \leq e^{-\frac{Nt^2}{4LD^2}} \]

Let \( \tilde{Y} \) be a minimal \( \epsilon/(2L) \)-covering over \( Y \). We know the covering number \( |\tilde{Y}| \leq (2DL/\epsilon)^d \). Thus by union bound:
\[ \Pr \left( \forall y \in \tilde{Y}, \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) - \mathbb{E}_{x \sim \mu^*} f(x, y) \geq t \right) \leq e^{d \log \frac{2DL}{\epsilon} - \frac{Nt^2}{4LD^2}} \]

Pick \( t = \epsilon/2 \) and \( N \geq c \cdot d(LD/\epsilon)^2 \log(LD/\epsilon) \) for some large absolute constant \( c \), we have:
\[ \Pr \left( \forall y \in \tilde{Y}, \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) - \mathbb{E}_{x \sim \mu^*} f(x, y) \geq \frac{\epsilon}{2} \right) \leq \frac{1}{2} \]

Let \( y^* = \arg\max_y \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) \), by definition of covering, we can always find a \( y' \in \tilde{Y} \) so that \( \|y^* - y'\| \leq \epsilon/(4L) \). Thus, with probability at least 1/2:
\[
\max_y \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) - \max_y \mathbb{E}_{x \sim \mu^*} f(x, y) = \frac{1}{N} \sum_{i=1}^{N} f(x_i, y^*) - \max_y \mathbb{E}_{x \sim \mu^*} f(x, y) \\
\leq \left[ \frac{1}{N} \sum_{i=1}^{N} f(x_i, y^*) - \frac{1}{N} \sum_{i=1}^{N} f(x_i, y') \right] + \left[ \frac{1}{N} \sum_{i=1}^{N} f(x_i, y') - \mathbb{E}_{x \sim \mu^*} f(x, y') \right] \\
+ \|\mathbb{E}_{x \sim \mu^*} f(x, y') - \max_y \mathbb{E}_{x \sim \mu^*} f(x, y)\| \leq \epsilon/2 + \epsilon/2 + 0 \leq \epsilon
\]

That is, with probability at least 1/2:
\[
\max_{y \in Y} \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) \leq \min_{\mu \in \mathcal{P}(X)} \max_{y \in Y} \mathbb{E}_{x \sim \mu} f(x, y) + \epsilon
\]

This implies:
\[
\min_{(x_1, \ldots, x_N) \in X^N} \max_{y \in Y} \frac{1}{N} \sum_{i=1}^{N} f(x_i, y) \leq \min_{\mu \in \mathcal{P}(X)} \max_{y \in Y} \mathbb{E}_{x \sim \mu} f(x, y) + \epsilon
\]

Combine with Proposition 32, we finish the proof. \( \square \)

**B. Proofs for Section 3.1**

In this section, we prove the propositions and theorems presented in Section 3.1.

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

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

**Proof.** Let \( D_{gen} \) be the set of local minimax points according to Definition 14. Let \( D_{mon}, D_{cts} \) be the sets of points if we further restrict function \( h \) in Definition 14 to be monotonic or continuous. We will prove that \( D_{gen} \subseteq D_{mon} \subseteq D_{cts} \subseteq D_{gen} \), so that they are all equivalent.

For simplicity of presentation, we denote \( \mathcal{P}(\delta, \epsilon) \) as the property that \( f(x^*, y^*) \leq g_\epsilon(x) \) for any \( x \) satisfying \( \|x - x^*\| \leq \delta \). By its definition, we know if \( \mathcal{P}(\delta, \epsilon) \) holds, then for any \( \delta' \leq \delta \) and any \( \epsilon' \geq \epsilon \), \( \mathcal{P}(\delta', \epsilon') \) also holds.
What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

\( \mathcal{D}_{\text{gen}} \subset \mathcal{D}_{\text{mon}} \): If \((x^*, y^*) \in \mathcal{D}_{\text{gen}}\), we know there exists \( \delta_0 > 0 \) and function \( h \) with \( h(\delta) \to 0 \) as \( \delta \to 0 \) such that for any \( \delta \in (0, \delta_0] \), \( \mathcal{P}(\delta, h(\delta)) \) holds. We can construct a function \( h'(\delta) = \sup_{\delta \in (0, \delta]} h(\delta) \) for any \( \delta \in (0, \delta_0] \). We can show \( h' \) is a monotonically increasing function, and \( h'(\delta) \to 0 \) as \( \delta \to 0 \). Since \( h'(\delta) \geq h(\delta) \) for all \( \delta \in (0, \delta_0] \), we know \( \mathcal{P}(\delta, h'(\delta)) \) also holds, that is, \((x^*, y^*) \in \mathcal{D}_{\text{mon}}\).

\( \mathcal{D}_{\text{mon}} \subset \mathcal{D}_{\text{cts}} \): For any monotone increasing function \( h : (0, \delta_0] \to \mathbb{R} \) with \( h(\delta) \to 0 \) as \( \delta \to 0 \), by standard argument in analysis, we can show there exists a continuous function \( h' : (0, \delta_0] \to \mathbb{R} \) so that \( h'(\delta) \geq h(\delta) \) for all \( \delta \in (0, \delta_0] \) and \( h'(\delta) \to 0 \) as \( \delta \to 0 \). Then, we can use similar arguments as above to finish the proof of this claim.

\( \mathcal{D}_{\text{cts}} \subset \mathcal{D}_{\text{gen}} \): This is immediate by definitions.

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

Proof. For simplicity of presentation, we denote \( \mathcal{P}(\delta, \epsilon) \) as the property that \( f(x^*, y^*) \leq g_{\epsilon}(x) \) for any \( x \) satisfying \( \|x - x^*\| \leq \delta \). By its definition, we know if \( \mathcal{P}(\delta, \epsilon) \) holds, then for any \( \delta' \leq \delta \) and any \( \epsilon' \geq \epsilon \), \( \mathcal{P}(\delta', \epsilon') \) also holds. Also, since \( f \) is continuous, if a sequence \( \{\epsilon_i\} \) has a limit, then \( \mathcal{P}(\delta, \epsilon_i) \) holds for all \( i \) implies that \( \mathcal{P}(\delta, \lim_{i \to \infty} \epsilon_i) \) holds.

The “only if” direction: suppose \((x^*, y^*)\) is a local minimax point of \( f \), then there exists \( \delta_0 > 0 \) and a function \( h \) satisfying the properties stated in Definition 14. Let \( \epsilon_0 = \delta_0 \). Since \( h(\delta) \to 0 \) as \( \delta \to 0 \), we know that for any \( \epsilon \in (0, \epsilon_0) \), there exists \( \delta \in (0, \delta_0] \) such that \( h(\delta) \leq \epsilon \). Meanwhile, according to Definition 14, \( \mathcal{P}(\delta, h(\delta)) \) holds, which implies that \( \mathcal{P}(\delta, \epsilon) \) holds. Finally, since \( \epsilon \leq \epsilon_0 = \delta_0 \), we have \( g_{\epsilon}(x^*) = f(x^*, y^*) \), i.e. \( y^* \) achieves the local maximum. Combine this with the fact that \( \mathcal{P}(\delta, \epsilon) \) holds, we finished the proof of this direction.

The “if” direction: since \( y^* \) is a local maximum of \( f(\cdot, x^*) \), there exists \( \delta_0 > 0 \) such that \( g_{\delta_0}(x^*) = f(x^*, y^*) \). Let \( \epsilon_0 = \min\{\epsilon_0, \delta_0\} \). By assumption, there exists a function \( q \) such that for any \( \epsilon \in (0, \epsilon_0] \) we have \( q(\epsilon) > 0 \) and \( \mathcal{P}(q(\epsilon), \epsilon) \) holds. Now, define \( \delta_0 = q(\epsilon_0) > 0 \), and define function \( h \) on \( (0, \delta_0] \) as follows:

\[ h(\delta) = \inf\{\epsilon \in (0, \epsilon_0] \text{ and } q(\epsilon) \geq \delta\} \]

It is easy to verify that when \( \delta \in (0, \delta_0] \), the set on the RHS is always non-empty as \( \epsilon_0 \) is always an element of the set, thus function \( h \) is well-defined on its domain. First, it is clear that \( h \) is a monotonically increasing function. Second, we prove \( h \to 0 \) as \( \delta \to 0 \), this is because for any \( \epsilon' \in (0, \epsilon_0] \), there is a \( \delta' = q(\epsilon') \) so that for any \( \delta \in (0, \delta'] \) we have \( h(\delta) \leq h(\delta') \leq \epsilon' \). Finally, we note for any \( \delta \in (0, \delta_0] \), by definition of \( h(\delta) \), there exists a sequence \( \{\epsilon_i\} \) so that \( \lim_{i \to \infty} \epsilon_i = h(\delta) \), and for any \( i \), we have \( \epsilon_i \in [h(\delta), \epsilon_0] \) and \( q(\epsilon_i) \geq \delta \). By assumptions, we know \( \mathcal{P}(q(\epsilon_i), \epsilon_i) \) holds, thus \( \mathcal{P}(\delta, \epsilon_i) \) holds, which eventually implies \( \mathcal{P}(\delta, h(\delta)) \) since \( \lim_{i \to \infty} \epsilon_i = h(\delta) \). This finishes the proof.

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

Proof. Let \( h \) be the constant function \( h(\delta) = 0 \) for any \( \delta \). Suppose \((x^*, y^*)\) is a local pure strategy Nash equilibrium, by definition it implies the existence of \( \delta_0 \), so that for any \( \delta \leq \delta_0 \), and any \((x, y)\) satisfying \( \|x - x^*\| \leq \delta \) and \( \|y - y^*\| \leq \delta \):

\[ f_2(x^*, y) \leq f_2(x^*, y^*) \leq f(x, y^*) \leq \max_{y : \|y - y^*\| \leq h(\delta)} f_2(x, y'). \]

which finishes the proof.

C. Proofs for Section 3.2

In this section, we prove the propositions and theorems presented in Section 3.2.

Proposition 18 (First-order Necessary Condition). Assuming that \( f \) is continuously differentiable, then any local minimax point \((x, y)\) satisfies \( \nabla_x f(x, y) = 0 \) and \( \nabla_y f(x, y) = 0 \).
Proof. Since $y$ is the local maximum of $f(x, \cdot)$, it implies $\nabla_y f(x, y) = 0$. Denote local optima $\delta_x^*(\delta_x) := \arg\max_{\|\delta_x\| \leq h(\delta)} f(x + \delta_x, y + \delta_y)$. By definition we know, $\|\delta_y(\delta_x)\| \leq h(\delta) \to 0$ as $\delta \to 0$. Thus

\begin{align*}
0 & \leq f(x + \delta_x, y + \delta_y) - f(x, y) \\
& = f(x + \delta_x, y + \delta_y(\delta_x)) - f(x, y + \delta_y(\delta_x)) + f(x, y + \delta_y(\delta_x)) - f(x, y) \\
& \leq f(x + \delta_x, y + \delta_y(\delta_x)) - f(x, y + \delta_y(\delta_x)) \\
& = \nabla_x f(x, y + \delta_y(\delta_x))^\top \delta_x + o(\|\delta_x\|) = \nabla_x f(x, y)^\top \delta_x + o(\|\delta_x\|)
\end{align*}

holds for any small $\delta_x$, which implies $\nabla_x f(x, y) = 0$. \hfill \Box

**Proposition 19** (Second-order Necessary Condition). Assuming that $f$ is twice differentiable, then $(x, y)$ is a local minimax point implies that $\nabla_{xx}^2 f(x, y) \leq 0$, and for any $v$ satisfying $\nabla_{yy}^2 f(x, y) \cdot v \in \text{column\_span}(\nabla_{yy}^2 f(x, y))$ that $v^\top [\nabla_{xx}^2 f - \nabla_{yy}^2 f(\nabla_{yy}^2 f)^\top \nabla_{yy}^2 f](x, y) \cdot v \geq 0$. (Here $^\top$ denotes Moore-Penrose inverse.)

Proof. Denote $A := \nabla_{xx}^2 f(x, y)$, $B := \nabla_{yy}^2 f(x, y)$ and $C := \nabla_{xy}^2 f(x, y)$. Since $y$ is the local maximum of $f(x, \cdot)$, it implies $B \leq 0$. On the other hand,

$$f(x + \delta_x, y + \delta_y) = f(x, y) + \frac{1}{2} \delta_x^\top A \delta_x + \delta_x^\top C \delta_y + \frac{1}{2} \delta_y^\top B \delta_y + o(\|\delta_x\|^2 + \|\delta_y\|^2).$$

Since $(x, y)$ is a local minimax point, by definition, there exists a function $h$ such that Eq.(3) holds. Denote $h'(\delta) = 2\|B^{-1} C^\top \delta_x\|$. We note both $h(\delta)$ and $h'(\delta) \to 0$ as $\delta \to 0$. For any $\delta_x$ satisfying $C^\top \delta_x \in \text{column\_span}(B)$, it is not hard to verify that $\arg\max_{\|\delta_y\| \leq \max(h(\delta), h'(\delta))} f(x + \delta_x, y + \delta_y) = -B^\top C \delta_x + o(\|\delta_x\|)$. Since $(x, y)$ is a local minimax point, we have

$$0 \leq \max_{\|\delta_y\| \leq h(\delta)} f(x + \delta_x, y + \delta_y) - f(x, y) \leq \max_{\|\delta_y\| \leq \max(h(\delta), h'(\delta))} f(x + \delta_x, y + \delta_y) - f(x, y)$$

$$= \frac{1}{2} \delta_x^\top (A - CB^\top C^\top) \delta_x + o(\|\delta_x\|^2).$$

Above equation holds for any $\delta_x$ satisfying $C^\top \delta_x \in \text{column\_span}(B)$, which finishes the proof. \hfill \Box

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

$$\nabla_{xx}^2 f - \nabla_{yy}^2 f(\nabla_{yy}^2 f)^\top \nabla_{yy}^2 f(x, y) > 0$$

is a local minimax point. We call stationary points satisfying (4) strict local minimax points.

Proof. Again denote $A := \nabla_{xx}^2 f(x, y)$, $B := \nabla_{yy}^2 f(x, y)$ and $C := \nabla_{xy}^2 f(x, y)$. Since $(x, y)$ is a stationary point, and $B < 0$, it is clear that $y$ is the local maximum of $f(x, \cdot)$. On the other hand, pick $\delta_y = B^{-1} C^\top \delta_x$ and let $h(\delta) = \|B^{-1} C^\top \delta_x\|$, we know when $\|\delta_x\| \leq \delta, \|\delta_y\| \leq h(\delta)$, thus

$$\max_{\|\delta_y\| \leq h(\delta)} f(x + \delta_x, y + \delta_y) - f(x, y) \geq f(x + \delta_x, y + \delta_y) - f(x, y)$$

$$= \frac{1}{2} \delta_x^\top (A - CB^\top C^\top) \delta_x + o(\|\delta_x\|^2) > 0$$

which finishes the proof. \hfill \Box

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

Proof. Consider function $f(x, y) = 0.2xy - \cos(y)$ in region $[-1, 1] \times [-2\pi, 2\pi]$ as shown in Figure 2. Clearly, the gradient is equal to $(0.2y, 0.2x + \sin(y))$. And for any fixed $x$, there are only two maxima $y^*(x)$ satisfying $0.2x + \sin(y^*) = 0$ where $y_1^*(x) \in (-\pi/2, -\pi/2)$ and $y_2^*(x) \in (\pi/2, 3\pi/2)$. On the other hand, $f(x, y_1^*(x))$ is monotonically decreasing with respect to $x$, while $f(x, y_2^*(x))$ is monotonically increasing, with $f(0, y_1^*(0)) = f(0, y_2^*(0))$ by symmetry. It is not hard to check $y_1^*(0) = -\pi$ and $y_2^*(0) = \pi$. Therefore, $(0, -\pi)$ and $(0, \pi)$ are two global solutions of minimax problem. However, the gradients at both points are not 0, thus they are not stationary points. By Proposition 18 they are also not local minimax points. \hfill \Box
Lemma 22. There exists a twice-differentiable function $f$ and a compact domain, where local minimax points do not exist.

Proof. Consider a two dimensional function $f(x, y) = y^2 - 2xy$ on $[-1, 1] \times [-1, 1]$. Suppose $(x^*, y^*)$ is a local minimax point, then at least $y^*$ is a local maximum of $f(x^*, \cdot)$, which restricts the possible local minimax points within set $[-1, 1] \times \{1\}$ or $(-1, 1] \times \{-1\}$. It is easy to check no points in either set is local minimax. \hfill \Box

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

Proof. Denote $A := \nabla^2_{xx}f(x, y)$, $B := \nabla^2_{xy}f(x, y)$, $C := \nabla^2_{yy}f(x, y)$, $g_x := \nabla_x f(x, y)$ and $g_y := \nabla_y f(x, y)$. Let $(x, y)$ be a global minimax point. Since $y$ is the global argmax of $f(x, \cdot)$ and locally strongly concave, we know $g_y = 0$ and $B < 0$. Let us now consider a second-order Taylor approximation of $f$ around $(x, y)$.

$$f(x + \delta_x, y + \delta_y) = f(x, y) + B^T \delta_x + \frac{1}{2} \delta_x^T A \delta_x + \frac{1}{2} \delta_y^T C \delta_y + \frac{1}{2} \delta_y^T B \delta_y + o(||\delta_x||^2 + ||\delta_y||^2)$$

Since by hypothesis, $B < 0$, we see that when $||\delta_x||$ is sufficiently small, there is a unique $\delta_y^*(\delta_x)$ so that $y + \delta_y^*(\delta_x)$ is a local maximum of $f(x + \delta_x, \cdot)$, where $\delta_y^*(\delta_x) = -B^{-1}C^T \delta_x + o(||\delta_x||)$. It is clear that $||\delta_y^*(\delta_x)|| \leq (||B^{-1}C^T|| + 1)||\delta_x||$ for sufficiently small $||\delta_x||$. Let $h(\delta) = (||B^{-1}C^T|| + 1)\delta$, we know for small enough $\delta$:

$$f(x + \delta_x, y + \delta_y^*(\delta_x)) = \max_{||\delta_y|| \leq h(\delta)} f(x + \delta_x, y + \delta_y)$$

Finally, since by assumption for any $f(x, \cdot)$ all local maxima are global maxima and $x$ is the global min of $\max_y f(x, y)$, we know:

$$f(x, y) \leq \max_{y'} f(x + \delta_x, y') = f(x + \delta_x, y + \delta_y^*(\delta_x)) = \max_{||\delta_y|| \leq h(\delta)} f(x + \delta_x, y + \delta_y)$$

which finishes the proof. \hfill \Box

D. Proofs for Section 3.3

In this section, we provides proofs for propositions and theorems presented in Section 3.3.

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

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

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

Proof. Consider GDA dynamics with step size $\eta$, then the Jacobian matrix of this dynamic system is $I + \eta J_{\gamma}$, whose eigenvalues are $\{1 + \eta \lambda_i\}$. Therefore, $(x, y)$ is a strict linearly stable point if and only if $\rho(I + \eta J_{\gamma}) < 1$, that is $|1 + \eta \lambda_i| < 1$ for all $i$. When taking $\eta \to 0$, this is equivalent to $\text{Re}(\lambda_i) < 0$ for all $i$. \hfill \Box

Proposition 26 ((Daskalakis and Panageas, 2018)). For any fixed $\gamma$, for any twice-differentiable $f$, Local Nash $\subset \gamma - \text{GDA}$, but there exist twice-differentiable $f$ such that $\gamma - \text{GDA} \notin \text{Local Nash}$.

Proof. Daskalakis and Panageas (2018) showed the proposition holds for 1-GDA. For completeness, here we show how similar proof goes through for $\gamma$-GDA for general $\gamma$. Let $\epsilon = 1/\gamma$, and denote $A := \nabla^2_{xx}f(x, y)$, $B := \nabla^2_{yy}f(x, y)$ and $C := \nabla^2_{xy}f(x, y)$. To prove the statement local Nash $\subset \gamma - \text{GDA}$, we note by definition, $(x, y)$ is a strict linear stable point of $1/\epsilon$-GDA if the real part of the eigenvalues of Jacobian matrix

$$J_{\epsilon} := \begin{pmatrix} -\epsilon A & -\epsilon C \\ C^T & B \end{pmatrix}$$
What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?

satisfy that $\text{Re}(\lambda_i) < 0$ for all $1 \leq i \leq d_1 + d_2$. We first note that:

$$\tilde{J}_x := \begin{pmatrix} B & \sqrt{\epsilon}C^T \\ -\sqrt{\epsilon}C & -\epsilon A \end{pmatrix} = UJ_xU^{-1}, \quad \text{where} \quad U = \begin{pmatrix} 0 & \sqrt{\epsilon}I \\ I & 0 \end{pmatrix}$$

Thus, the eigenvalues of $\tilde{J}_e$ and $J_e$ are the same. We can also decompose:

$$\tilde{J}_e = P + Q, \quad \text{where} \quad P := \begin{pmatrix} B & -\epsilon A \end{pmatrix}, \quad Q := \begin{pmatrix} 0 & \sqrt{\epsilon}C^T \\ -\sqrt{\epsilon}C & 0 \end{pmatrix}$$

If $(x, y)$ is a strict local pure strategy Nash equilibrium, then $A > 0, B < 0$, then $P$ is a negative definite symmetric matrix, and $Q$ is anti-symmetric matrix, i.e., $Q = -Q^T$. For any eigenvalue $\lambda$ if $\tilde{J}_e$, assume $w$ is the associated eigenvector. That is, $\tilde{J}_ew = \lambda w$; also let $w = x + iy$ where $x$ and $y$ are real vectors, and $\bar{w}$ be the complex conjugate of vector $w$. Then:

$$\text{Re}(\lambda) = [\bar{w}^T \tilde{J}_x w + w^T \tilde{J}_x \bar{w}] / 2 = [(x - iy)^T \tilde{J}_e (x + iy) + (x + iy)^T \tilde{J}_e (x - iy)] / 2$$

$$= x^T \tilde{J}_e x + y^T \tilde{J}_e y = x^T P x + y^T P y + x^T Q x + y^T Q y$$

Since $P$ is negative definite, that is $x^T P x < 0$. Meanwhile, since $Q$ is antisymmetric $x^T Q x = x^T Q^T x = 0$ and $y^T Q y = y^T Q^T y = 0$. This proves $\text{Re}(\lambda) < 0$, that is $(x, y)$ is a strict linear stable point of $1/\epsilon$-GDA.

To prove the statement $\gamma - GDA \notin \text{localMinimax}$, since $\epsilon$ is also fixed, we consider function $f(x, y) = x^2 + 2\sqrt{\epsilon}xy + (\epsilon/2)y^2$. It is easy to see $(0, 0)$ is a fixed point of $1/\epsilon$-GDA, and Hessian $A = 2, B = \epsilon, C = 2\sqrt{\epsilon}$. Thus the Jacobian matrix

$$J_e := \begin{pmatrix} -2\epsilon & -2\epsilon^{3/2} \\ 2\epsilon^{1/2} & -\epsilon \end{pmatrix}$$

has two eigenvalues $\epsilon(-1 \pm i\sqrt{\epsilon})/2$. Therefore, $\text{Re}(\lambda_1) = \text{Re}(\lambda_2) < 0$, which implies $(0, 0)$ is a strict linear stable point. However $B = \epsilon > 0$, thus it is not a strict local pure strategy Nash equilibrium.

**Proposition 27.** For any fixed $\gamma$, there exists a twice-differentiable $f$ such that $\text{LocalMinimax} \notin \gamma - GDA$; there also exists a twice-differentiable $f$ such that $\gamma - GDA \notin \text{LocalMinimax} \cup \text{LocalMaximin}$. 

**Proof.** Let $\epsilon = 1/\gamma$, and denote $A := \nabla^2_{xx} f(x, y), B := \nabla^2_{xy} f(x, y)$ and $C := \nabla^2_{yy} f(x, y)$.

To prove the first statement $\text{localminimax} \notin \gamma - GDA$, since $\epsilon$ is also fixed, we consider function $f(x, y) = x^2 + 2\sqrt{\epsilon}x y + (\epsilon/2)y^2$. It is easy to see $(0, 0)$ is a fixed point of $1/\epsilon$-GDA, and Hessian $A = -2, B = -\epsilon, C = 2\sqrt{\epsilon}$. It is easy to verify that $B < 0$ and $A - CB^{-1}C = 2 > 0$, thus $(0, 0)$ is a local minimax point. However, inspect the Jacobian matrix of $1/\epsilon$-GDA:

$$J_e := \begin{pmatrix} 2\epsilon & -2\epsilon^{3/2} \\ -2\epsilon^{1/2} & -\epsilon \end{pmatrix}$$

We know the two eigenvalues are $\epsilon(1 \pm i\sqrt{\epsilon})/2$. Therefore, $\text{Re}(\lambda_1) = \text{Re}(\lambda_2) > 0$, which implies $(0, 0)$ is not a strict linear stable point.

To prove the second statement $\gamma - GDA \notin \text{localminimax} \cup \text{localmaximin}$, since $\epsilon$ is also fixed, we consider function $f(x, y) = x^2 + 2\sqrt{\epsilon}x y_1 + (\epsilon/2)y_1^2 - x_2^2/2 + 2\sqrt{\epsilon}x_2 y_2 - \epsilon y_2^2$. It is easy to see $(0, 0)$ is a fixed point of $1/\epsilon$-GDA, and Hessian $A = \text{diag}(2, -1), B = \text{diag}(\epsilon, -2\epsilon), C = 2\sqrt{\epsilon} \cdot \text{diag}(1, 1)$. Thus the Jacobian matrix

$$J_e := \begin{pmatrix} -2\epsilon & 0 & -2\epsilon^{3/2} & 0 \\ 0 & \epsilon & 0 & -2\epsilon^{3/2} \\ 2\epsilon^{1/2} & 0 & \epsilon & 0 \\ 0 & 2\epsilon^{1/2} & 0 & -2\epsilon \end{pmatrix}$$

has four eigenvalues $\epsilon(-1 \pm i\sqrt{\epsilon})/2$ (each with multiplicity of 2). Therefore, $\text{Re}(\lambda_i) < 0$ for $1 \leq i \leq 4$, which implies $(0, 0)$ is a strict linear stable point. However, $B$ is not negative definite, thus $(0, 0)$ is not a strict local minimax point; similarly, $A$ is also not positive definite, thus $(0, 0)$ is not a strict local maximin point.
Theorem 28 (Asymptotic Behavior of $\infty$-GDA). For any twice-differentiable $f$, $Local_{\text{Minimax}} \subset \infty-GDA \subset \infty-\bar{GDA} \subset Local_{\text{Minimax}} \cup \{(x, y)|(x, y)\text{ is stationary and } \nabla^2_{xy} f(x, y)\text{ is degenerate}\}$.

Proof. For simplicity, denote $A := \nabla^2_{xy} f(x, y)$, $B := \nabla^2_{xy} f(x, y)$ and $C := \nabla^2_{xy} f(x, y)$. Let $\epsilon = 1/\gamma$. Consider sufficiently small $\epsilon$ (i.e. sufficiently large $\gamma$), we know the Jacobian $J$ of $1/\epsilon$-GDA at $(x, y)$ is:

$$J_\epsilon := \begin{pmatrix} -\epsilon A & -\epsilon C \\ C^T & B \end{pmatrix}$$

According to Lemma 36, for sufficient $\epsilon$, $J_\epsilon$ has $d_1 + d_2$ complex eigenvalues $\{\lambda_i\}_{i=1}^{d_1 + d_2}$ with following form for sufficient small $\epsilon$:

$$|\lambda_i + \epsilon \mu_i| = o(\epsilon), \quad 1 \leq i \leq d_1$$

$$|\lambda_i + d_1 - \nu_i| = o(1), \quad 1 \leq i \leq d_2$$

(6)

where $\{\mu_i\}_{i=1}^{d_1}$ and $\{\nu_i\}_{i=1}^{d_2}$ are the eigenvalues of matrices $A - CB^{-1}C^T$ and $B$ respectively. Now we are ready to prove the three inclusion statement in Theorem 28 separately.

First, for $\infty-\bar{GDA} \subset \infty-GDA$ always holds by their definitions.

Second, for $Local_{\text{Minimax}} \subset \infty-GDA$ statement, if $(x, y)$ is strict local minimax point, then by its definition:

$$B \prec 0, \quad \text{and} \quad A - CB^{-1}C^T \succ 0$$

By Eq. (6) the eigenvalue structure of $J_\epsilon$, we know there exists sufficiently small $\epsilon_0$, so that for any $\epsilon < \epsilon_0$, the real part $\Re(\lambda_i) < 0$, i.e. $(x, y)$ is a strict linear stable point of $1/\epsilon$-GDA.

Finally, for $\infty-GDA \subset Local_{\text{Minimax}} \cup \{(x, y)|(x, y)\text{ is stationary and } B \text{ is degenerate}\}$ statement, if $(x, y)$ is strict linear stable point of $1/\epsilon$-GDA for a sufficiently small $\epsilon$, then for any $i$, the real part of eigenvalue of $J_\epsilon$: $\Re(\lambda_i) < 0$. By Eq. (6), if $B$ is invertible, this implies:

$$B \prec 0, \quad \text{and} \quad A - CB^{-1}C^T \succ 0$$

Finally, suppose matrix $A - CB^{-1}C^T$ has an eigenvalue 0. This means the existence of unit vector $w$ so that $(A - CB^{-1}C^T)w = 0$. It is not hard to verify then $J_\epsilon \cdot (w, -B^{-1}C^T w)^T = 0$. This implies $J_\epsilon$ has a 0 eigen-value, which contradicts the fact that $\Re(\lambda_i) < 0$ for any $i$. Therefore, we can conclude $A - CB^{-1}C^T \succ 0$, and $(x, y)$ is a strict local minimax point.

Lemma 36. For any symmetric matrix $A \in \mathbb{R}^{d_1 \times d_1}$, $B \in \mathbb{R}^{d_2 \times d_2}$, and any rectangular matrix $C \in \mathbb{R}^{d_1 \times d_2}$, assume $B$ is nondegenerate. Then, matrix

$$\begin{pmatrix} -\epsilon A & -\epsilon C \\ C^T & B \end{pmatrix}$$

has $d_1 + d_2$ complex eigenvalues $\{\lambda_i\}_{i=1}^{d_1 + d_2}$ with following form for sufficient small $\epsilon$:

$$|\lambda_i + \epsilon \mu_i| = o(\epsilon), \quad 1 \leq i \leq d_1$$

$$|\lambda_i + d_1 - \nu_i| = o(1), \quad 1 \leq i \leq d_2$$

where $\{\mu_i\}_{i=1}^{d_1}$ and $\{\nu_i\}_{i=1}^{d_2}$ are the eigenvalues of matrices $A - CB^{-1}C^T$ and $B$ respectively.

Proof. By definition of eigenvalues, $\{\lambda_i\}_{i=1}^{d_1 + d_2}$ are the roots of characteristic polynomial:

$$p_\epsilon(\lambda) := \det\begin{pmatrix} \lambda I + \epsilon A & \epsilon C \\ -C^T & \lambda I - B \end{pmatrix}$$

We can expand this polynomial as:

$$p_\epsilon(\lambda) = p_0(\lambda) + \sum_{i=1}^{d_1 + d_2} \epsilon^i p_i(\lambda), \quad p_0(\lambda) = \lambda^{d_1} \cdot \det(\lambda I - B).$$
Here, \( p_i \) are polynomials of order at most \( d_1 + d_2 \). It is clear that the roots of \( p_0 \) are 0 (with multiplicity \( d_1 \)) and \( \{\nu_i\}_{i=1}^{d_2} \). According to Lemma 37, we know the roots of \( p_\epsilon \) satisfy:

\[
\begin{align*}
|\lambda_i| &= o(1) \\
|\lambda_{i+d_1} - \nu_i| &= o(1),
\end{align*}
\]

(7)

Since \( B \) is non-degenerate, we know when \( \epsilon \) is small enough, \( \lambda_1 \ldots \lambda_{d_1} \) are very close to 0 while \( \lambda_{d_1+1} \ldots \lambda_{d_1+d_2} \) have modulus at least \( \Omega(1) \). To provide the sign information of the first \( d_1 \) roots, we proceed to lower order characterization.

On the other hand, reparametrize \( \lambda = \epsilon \theta \), we have:

\[
p_\epsilon(\epsilon \theta) = \det \begin{pmatrix} \epsilon \theta I + \epsilon A & \epsilon C \\ -C^T & \epsilon \theta I - B \end{pmatrix} = \epsilon^{d_1} \det \begin{pmatrix} \theta I + A & C \\ -C^T & \epsilon \theta I - B \end{pmatrix}
\]

Therefore, we know \( q_\epsilon(\theta) := p_\epsilon(\epsilon \theta)/\epsilon^{d_1} \) is still a polynomial, and has polynomial expansion:

\[
q_\epsilon(\theta) = q_0(\theta) + \sum_{i=1}^{d_2} \epsilon^i q_i(\lambda), \quad q_0(\theta) = \det \begin{pmatrix} \theta I + A & C \\ -C^T & -B \end{pmatrix}
\]

It is clear polynomial \( q_\epsilon \) and \( p_\epsilon \) have same roots up to \( \epsilon \) scaling. Furthermore, we have following factorization:

\[
\begin{pmatrix} \theta I + A & C \\ -C^T & -B \end{pmatrix} = \begin{pmatrix} \theta I + A - CB^{-1}C^T & C \\ -C^T & -B \end{pmatrix} \begin{pmatrix} I & 0 \\ B^{-1}C^T & I \end{pmatrix}
\]

Since \( B \) is non-degenerate, we have \( \det(B) \neq 0 \), and

\[
q_0(\theta) = (-1)^{d_2} \det(B) \det(\theta I + A - CB^{-1}C^T)
\]

\( q_0 \) is \( d_1 \)-order polynomial having roots \( \{\mu_i\}_{i=1}^{d_1} \), which are the eigenvalues of matrices \( A - CB^{-1}C^T \). According to Lemma 37, we know \( q_\epsilon \) has at least \( d_1 \) roots so that \( |\theta_i + \mu_i| = o(1) \). This implies \( d_1 \) roots of \( p_\epsilon \) so that:

\[
|\lambda_i + \epsilon \mu_i| = o(\epsilon) \quad 1 \leq i \leq d_1
\]

By Eq.(7), we know \( p_\epsilon \) has exactly \( d_1 \) roots which are of \( o(1) \) scaling. This finishes the proof. \( \square \)

**Lemma 37** (Continuity of roots of polynomials (Zedek, 1965)). *Given a polynomial \( p_m(z) := \sum_{k=0}^{m} a_k z^k \), an integer \( m \geq n \) and a number \( \epsilon > 0 \), there exists a number \( \delta > 0 \) such that whenever the \( m+1 \) complex numbers \( b_k, 0 \leq k \leq m \), satisfy the inequalities

\[
|b_k - a_k| < \delta \quad \text{for } 0 \leq k \leq n, \quad \text{and} \quad |b_k| < \delta \quad \text{for } n+1 \leq k \leq m
\]

then the roots \( \beta_k, 1 \leq k \leq m \) of the polynomial \( q_m(z) := \sum_{k=0}^{m} b_k z^k \) can be labeled in such a way as to satisfy with respect to the zeros \( \alpha_k, 1 \leq k \leq n \) of \( p_m(z) \) the inequalities

\[
|\beta_k - \alpha_k| < \epsilon \quad \text{for } 1 \leq k \leq n, \quad \text{and} \quad |\beta_k| > 1/\epsilon \quad \text{for } n+1 \leq k \leq m
\]

**E. Proof for Theorem 31**

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

\[
\mathbb{E} \left[ ||\nabla \phi_{1/2\ell}(\bar{x})||^2 \right] 
\leq 2 \cdot \frac{(\phi_{1/2\ell}(x_0) - \min \phi(\bar{x})) + \ell L^2 \gamma^2}{\gamma \sqrt{T+1}} + 4\ell \epsilon,
\]

where \( \phi_{1/2\ell} \) is the Moreau envelope (5) of \( \phi \).
We first note an important equation for the gradient of Moreau envelope.

\[ \nabla \phi_\lambda(x) = \lambda^{-1} \left( x - \arg\min_{\tilde{x}} \left( \phi(\tilde{x}) + \frac{1}{2\lambda} \|x - \tilde{x}\|^2 \right) \right). \]  

(8)

We also observe that since \( f(\cdot) \) is \( \ell \)-smooth and \( y_t \) is an approximate maximizer for \( x_t \), we have that any \( x_t \) from Algorithm 2 and \( \tilde{x} \) satisfy

\[ \phi(\tilde{x}) \geq f(\tilde{x}, y_t) \geq f(x_t, y_t) + \langle \nabla_x f(x_t, y_t), \tilde{x} - x_t \rangle - \frac{\ell}{2} \|\tilde{x} - x_t\|^2 \]

\[ \geq \phi(x_t) - \epsilon + \langle \nabla_x f(x_t, y_t), \tilde{x} - x_t \rangle - \frac{\ell}{2} \|\tilde{x} - x_t\|^2. \]  

(9)

Let \( \tilde{x}_t := \arg\min_x \phi(x) + \ell \|x - x_t\|^2 \). We have:

\[ \phi_{1/2\ell}(x_{t+1}) \leq \phi(\tilde{x}_t) + \ell \|x_{t+1} - \tilde{x}_t\|^2 \]

\[ \leq \phi(\tilde{x}_t) + \ell \|x_t - \eta \nabla_x f(x_t, y_t) - \tilde{x}_t\|^2 \]

\[ \leq \phi(\tilde{x}_t) + \ell \|x_t - \tilde{x}_t\|^2 + 2\ell \eta \|\nabla_x f(x_t, y_t), \tilde{x}_t - x_t\| + \eta^2 \ell \|\nabla_x f(x_t, y_t)\|^2 \]

\[ \leq \phi_{1/2\ell}(x_t) + 2\ell \eta \|\nabla_x f(x_t, y_t), \tilde{x}_t - x_t\| + \eta^2 \ell \|\nabla_x f(x_t, y_t)\|^2 \]

\[ \leq \phi_{1/2\ell}(x_t) + 2\ell \left( \phi(\tilde{x}_t) - \phi(x_t) + \epsilon + \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \right) + \eta^2 \ell L^2, \]

where the last line follows from (9). Taking a telescopic sum over \( t \), we obtain

\[ \phi_{1/2\ell}(x_T) \leq \phi_{1/2\ell}(x_0) + 2\ell \eta \sum_{t=0}^{T} \left( \phi(\tilde{x}_t) - \phi(x_t) + \epsilon + \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \right) + \eta^2 \ell L^2 T \]

Rearranging this, we obtain

\[ \frac{1}{T+1} \sum_{t=0}^{T} \left( \phi(x_t) - \phi(\tilde{x}_t) - \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \right) \leq \epsilon + \frac{\phi_{1/2\ell}(x_0) - \min_x \phi(x)}{2\eta T} + \frac{\eta L^2}{2}. \]  

(10)

Since \( \phi(x) + \ell \|x - x_t\|^2 \) is \( \ell \)-strongly convex, we have

\[ \phi(x_t) - \phi(\tilde{x}_t) - \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \]

\[ \geq \phi(x_t) + \ell \|x_t - x_t\|^2 - \phi(\tilde{x}_t) - \ell \|\tilde{x}_t - x_t\|^2 + \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \]

\[ = \left( \phi(x_t) + \ell \|x_t - x_t\|^2 - \min_x \phi(x) + \ell \|x - x_t\|^2 \right) + \frac{\ell}{2} \|x_t - \tilde{x}_t\|^2 \]

\[ \geq \ell \|x_t - \tilde{x}_t\|^2 = \frac{1}{4\ell} \|\nabla \phi_{1/2\ell}(x_t)\|^2, \]

where we used (8) in the last step. Plugging this in (10) proves the result. \( \square \)