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 $\mu \in \mathcal{P}(\mathcal{X})$ or $\nu \in \mathcal{P}(\mathcal{Y})$. Then, the payoff function becomes an expected value $\mathbb{E}_{\mathbf{x}\sim\mu,\mathbf{y}\sim\nu} f(\mathbf{x},\mathbf{y})$. 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*

$$\min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\nu\in\mathcal{P}(\mathcal{Y})} \mathbb{E}_{(\mu,\nu)} f(\mathbf{x},\mathbf{y}) = \max_{\nu\in\mathcal{P}(\mathcal{Y})} \min_{\mu\in\mathcal{P}(\mathcal{X})} \mathbb{E}_{(\mu,\nu)} f(\mathbf{x},\mathbf{y}).$$

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 $(\mu^\star, \nu^\star)$ is a **mixed strategy Nash equilibrium** of f, if for any measure (μ, ν) in $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, we have

$$\mathbb{E}_{\mathbf{x}\sim\mu^\star,\mathbf{y}\sim\nu} f(\mathbf{x},\mathbf{y}) \le \mathbb{E}_{\mathbf{x}\sim\mu^\star,\mathbf{y}\sim\nu^\star} f(\mathbf{x},\mathbf{y}) \le \mathbb{E}_{\mathbf{x}\sim\mu,\mathbf{y}\sim\nu^\star} f(\mathbf{x},\mathbf{y}).$$

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 $(\mu^\star, \nu^\star)$ be a mixed strategy Nash equilibrium. A probability measure $(\mu^\dagger, \nu^\dagger)$ is an **ϵ-approximate mixed strategy Nash equilibrium** if:

$$\begin{aligned}
&\forall \nu' \in \mathcal{P}(\mathcal{Y}), \quad \mathbb{E}_{(\mu^\dagger,\nu')} f(\mathbf{x},\mathbf{y}) \le \mathbb{E}_{(\mu^\star,\nu^\star)} f(\mathbf{x},\mathbf{y}) + \epsilon \\
&\forall \mu' \in \mathcal{P}(\mathcal{Y}), \quad \mathbb{E}_{(\mu',\nu^\dagger)} f(\mathbf{x},\mathbf{y}) \ge \mathbb{E}_{(\mu^\star,\nu^\star)} f(\mathbf{x},\mathbf{y}) - \epsilon.
\end{aligned}$$

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

$$\min_{(\mathbf{x}_1,\ldots,\mathbf{x}_N)\in\mathcal{X}^N} \max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) \le \mathbb{E}_{(\mu^\star,\nu^\star)} f(\mathbf{x},\mathbf{y}) + \epsilon.$$

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 $f(\cdot,\mathbf{y})$ to its average over N samples for all $\mathbf{y} \in \mathcal{Y}$ simultaneously. A similar argument was made in (Arora et al., 2017).

Theorem 35 implies that in order to find a approximate mixed strategy Nash equilibrium, we can solve a large minimax problem with objective $F(\mathbf{X},\mathbf{y}) := \sum_{i=1}^N f(\mathbf{x}_i,\mathbf{y})/N$. The global minimax solution $\mathbf{X}^\star = (\mathbf{x}_1^\star,\ldots,\mathbf{x}_n^\star)$ gives a empirical distribution $\hat{\mu}^\star = \sum_{i=1}^N \delta(\mathbf{x}-\mathbf{x}_i^\star)/N$, where $\delta(\cdot)$ is the Dirac delta function. By symmetry, we can also solve the corresponding maximin problem to find $\hat{\nu}^\star$. It can be shown that $(\hat{\mu}^\star, \hat{\nu}^\star)$ is an ϵ-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,

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

Therefore, we only need to solve the problem of RHS. Suppose the minimum over $\mathcal{P}(\mathcal{X})$ is achieved at $\mu^\star$. First, sample $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ i.i.d from $\mu^\star$, and note $\max_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}} |f(\mathbf{x}_1,\mathbf{y}) - f(\mathbf{x}_2,\mathbf{y})| \le LD$ for any fixed $\mathbf{y}$. Therefore by Hoeffding

inequality, for any fixed $\mathbf{y}$:

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) - \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}) \geq t\right) \leq e^{-\frac{Nt^2}{(LD)^2}}$$

Let $\bar{\mathcal{Y}}$ be a minimal $\epsilon/(2L)$-covering over $\mathcal{Y}$. We know the covering number $|\bar{\mathcal{Y}}| \leq (2DL/\epsilon)^d$. Thus by union bound:

$$\mathbb{P}\left(\forall \mathbf{y}\in\bar{\mathcal{Y}},\ \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) - \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}) \geq t\right) \leq e^{d\log\frac{2DL}{\epsilon} - \frac{Nt^2}{(LD)^2}}$$

Pick $t = \epsilon/2$ and $N \geq c\cdot d(LD/\epsilon)^2\log(LD/\epsilon)$ for some large absolute constant c, we have:

$$\mathbb{P}\left(\forall \mathbf{y}\in\bar{\mathcal{Y}},\ \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) - \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}) \geq \frac{\epsilon}{2}\right) \leq \frac{1}{2}$$

Let $\mathbf{y}^\star = \arg\max_{\mathbf{y}} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y})$, by definition of covering, we can always find a $\mathbf{y}'\in\bar{\mathcal{Y}}$ so that $\|\mathbf{y}^\star - \mathbf{y}'\| \leq \epsilon/(4L)$. Thus, with probability at least $1/2$:

$$\begin{aligned}
&\max_{\mathbf{y}} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) - \max_{\mathbf{y}} \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}^\star) - \max_{\mathbf{y}} \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}) \\
\leq &\left[\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}^\star) - \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}')\right] + \left[\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}') - \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}')\right] \\
&+ \left[\mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y}') - \max_{\mathbf{y}} \mathbb{E}_{\mathbf{x}\sim\mu^\star} f(\mathbf{x},\mathbf{y})\right] \leq \epsilon/2 + \epsilon/2 + 0 \leq \epsilon
\end{aligned}$$

That is, with probability at least $1/2$:

$$\max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) \leq \min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\mathbf{x}\sim\mu} f(\mathbf{x},\mathbf{y}) + \epsilon$$

This implies:

$$\min_{(\mathbf{x}_1,\ldots,\mathbf{x}_N)\in\mathcal{X}^N} \max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i,\mathbf{y}) \leq \min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\mathbf{x}\sim\mu} f(\mathbf{x},\mathbf{y}) + \epsilon$$

Combine with Proposition 32, we finish the proof. □

B. Relation to (Evtushenko, 1974)

In this section, we review a notion similar to local minimax proposed by Evtushenko (1974). To distinguish that notion from our definition (Definition 14), we call it Evtushenko's minimax property. We remark that Evtushenko's minimax is not a truly local property. As a noticable difference, Evtushenko's definition does not satisfy the first-order and second-order necessary conditions of the local minimax notion proposed in this paper as in Proposition 18 and Proposition 19.

Evtushenko's definition of local minimax point can be stated as follows.

Definition 36 ((Evtushenko, 1974))**.** A point $(\mathbf{x}^\star,\mathbf{y}^\star)$ is said to be a **Evtushenko's minimax point** of f, if there exist a local neighborhood $\mathcal{W}$ of $(\mathbf{x}^\star,\mathbf{y}^\star)$ so that $(\mathbf{x}^\star,\mathbf{y}^\star)$ is a global minimax point (Definition 9) within $\mathcal{W}$.

First, we remark that Definition 36 is in fact not a local notion. That is, whether a point $(\mathbf{x}^\star,\mathbf{y}^\star)$ is a Evtushenko's minimax point relies on the property of function f at points which are far away from $(\mathbf{x}^\star,\mathbf{y}^\star)$.

Proposition 37. *There exists a twice differentiable function f, a point $(\mathbf{x},\mathbf{y})$ and its two local neighborhoods $\mathcal{W}_1,\mathcal{W}_2$ satisfying $\mathcal{W}_1\subset\mathcal{W}_2$, so that $(\mathbf{x},\mathbf{y})$ **is not** a Evtushenko's minimax point for $f|_{\mathcal{W}_1}$ but **is** a Evtushenko's minimax point for $f|_{\mathcal{W}_2}$. Here $f|_{\mathcal{W}}$ denotes the function f restriced to the domain $\mathcal{W}$.*

Proof. Consider the function $f(x,y) = 0.2xy - \cos(y)$ in the region $\mathcal{W}_2 = [-1,1] \times [-2\pi, 2\pi]$ as shown in Figure 2. Follow the same analysis as in the proof of Proposition 21, we can show that $(0,-\pi)$ is a global minimax point on $\mathcal{W}_2$, thus a Evtushenko's minimax point of $f|_{\mathcal{W}_2}$.

On the other hand, consider $\mathcal{W}_1 = [-1,1]\times[-2\pi, 0]$. For any fixed x, the global maximium $y^\star(x)$ satisfies $0.2x + \sin(y^\star) = 0$ where $y^\star(x) \in (-3\pi/2, -\pi/2)$. Then for any local neighborhood $\mathcal{W}$ of $(0,-\pi)$ such that $\mathcal{W} \subset \mathcal{W}_1$, there always exists an $\epsilon > 0$ so that $(\epsilon, y^\star(\epsilon)) \in \mathcal{W}$. However, we can verify that

$$f(0,-\pi) \geq f(\epsilon, y^\star(\epsilon)) = \max_{y:(\epsilon,y)\in\mathcal{W}} f(\epsilon, y)$$

That is $(0,-\pi)$ is not a Evtushenko's minimax point on $f|_{\mathcal{W}_1}$.

This concludes that whether $(0,-\pi)$ is a Evtushenko's minimax point depends on the property of function f on set $\mathcal{W}_2 - \mathcal{W}_1$ whose elements are all far away from $(0,-\pi)$. That is, Evtushenko's minimax property is not a local property. □

We further clarify the relation between Evtushenko's minimax point and our definition of local minimax point (Definition 14) as follows.

Proposition 38. *A local minimax point is a Evtushenko's minimax point, but the reverse is not true.*

Proof. If a point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a local minimax point, according to Definition 14, there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, and any $(\mathbf{x}, \mathbf{y})$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \leq \delta$ and $\|\mathbf{y} - \mathbf{y}^\star\| \leq \delta$, we have

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

Therefore, we can choose a $\delta^\dagger \in (0, \delta_0]$ so that $h(\delta^\dagger) \leq \delta_0$. According to the equation above, $(\mathbf{x}^\star, \mathbf{y}^\star)$ is the global minimax point in region $\mathbb{B}_{\mathbf{x}^\star}(\delta^\dagger) \times \mathbb{B}_{\mathbf{y}^\star}(h(\delta^\dagger))$ where $\mathbb{B}_{\mathbf{z}}(r)$ denote the Euclidean ball around $\mathbf{z}$ with radius r. Therefore, $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a Evtushenko's minimax point.

The claim that a Evtushenko's minimax point can be a non local minimax point easily follows from Proposition 21, where a global minimax point (which is always a Evtushenko's minimax point) can be a non local minimax point. □

Finally, as a consequence of Proposition 38, the sufficient conditions for local minimax points are still sufficient conditions for Evtushenko's minimax points. However, the necessary conditions for local minimax points are no longer the necessary conditions for Evtushenko's minimax points.

Proposition 39. *A Evtushenko's minimax point can be a non-stationary point, which does not satisfies Eq.(4) even if* $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}$.

Proof. Consider the function $f(x,y) = -0.03x^2 + 0.2xy - \cos(y)$ in the region $[-1,1] \times [-2\pi, 2\pi]$ as shown in Figure 2. Follow a similar analysis as in the proof of Proposition 21, we can show that $(0,-\pi)$ is a global minimax point, thus a Evtushenko's minimax point. However, we have gradient and Hessian at $(0,-\pi)$:

$$\nabla f = \begin{pmatrix} -0.2\pi \\ 0 \end{pmatrix}, \qquad \nabla^2 f = \begin{pmatrix} -0.06 & 0.2 \\ 0.2 & -1 \end{pmatrix}$$

Therefore, $(0,-\pi)$ is not a stationary point, and despite $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \prec \mathbf{0}$, Eq.(4) does not hold. □

C. Proofs for Section 3.1

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

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

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

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

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

For simplicity of presentation, we denote $\mathcal{P}(\delta, \epsilon)$ as the property that $f(\mathbf{x}^\star, \mathbf{y}^\star) \leq g_\epsilon(\mathbf{x})$ for any $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \leq \delta$. 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.

$\mathcal{D}_{gen} \subset \mathcal{D}_{mon}$**:** If $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{D}_{gen}$, we know there exists $\delta_0 > 0$ and function h with $h(\delta) \to 0$ as $\delta \to 0$ such that for any $\delta \in (0, \delta_0]$, that $\mathcal{P}(\delta, h(\delta))$ 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, $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{D}_{mon}$.

$\mathcal{D}_{mon} \subset \mathcal{D}_{cts}$**:** For any monotonically increasing function $h : (0, \delta_0] \to \mathbb{R}$ with $h(\delta) \to 0$ as $\delta \to 0$, by a 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.

$\mathcal{D}_{cts} \subset \mathcal{D}_{gen}$**:** This is immediate by definitions. □

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

Proof. For simplicity of presentation, we denote $\mathcal{P}(\delta, \epsilon)$ as the property that $f(\mathbf{x}^\star, \mathbf{y}^\star) \leq g_\epsilon(\mathbf{x})$ for any $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \leq \delta$. 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: Supposing $(\mathbf{x}^\star, \mathbf{y}^\star)$ 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 $g_\epsilon(\mathbf{x}^\star) = f(\mathbf{x}^\star, \mathbf{y}^\star)$; i.e., $\mathbf{y}^\star$ achieves the local maximum. Combining this with the fact that $\mathcal{P}(\delta, \epsilon)$ holds, we finish the proof of this direction.

The "if" direction: Since $\mathbf{y}^\star$ is a local maximum of $f(\cdot, \mathbf{x}^\star)$, there exists $\delta_y > 0$ such that $g_{\delta_y}(\mathbf{x}^\star) = f(\mathbf{x}^\star, \mathbf{y}^\star)$. Let $\tilde{\epsilon}_0 = \min\{\epsilon_0, \delta_y\}$. By assumption, there exists a function q such that for any $\epsilon \in (0, \tilde{\epsilon}_0]$ we have $q(\epsilon) > 0$ and $\mathcal{P}(q(\epsilon), \epsilon)$ holds. Now, define $\delta_0 = q(\tilde{\epsilon}_0) > 0$, and define a function h on $(0, \delta_0]$ as follows:

$$h(\delta) = \inf\{\epsilon | \epsilon \in (0, \tilde{\epsilon}] \text{ and } q(\epsilon) \geq \delta\}.$$

It is easy to verify that when $\delta \in (0, \delta_0]$, the set on the RHS is always non-empty as $\tilde{\epsilon}_0$ is always an element of the set, thus the 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 $\epsilon' \in (0, \tilde{\epsilon}]$, 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 $\epsilon_i \in [h(\delta), \tilde{\epsilon}_0]$ and $q(\epsilon_i) \geq \delta$. By assumption, 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 δ. If $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a local pure strategy Nash equilibrium, then by definition this implies the existence of δ_0 such that for any $\delta \leq \delta_0$, and any $(\mathbf{x}, \mathbf{y})$ satisfying $\|\mathbf{x} - \mathbf{x}^\star\| \leq \delta$ and $\|\mathbf{y} - \mathbf{y}^\star\| \leq \delta$:

$$f_2(\mathbf{x}^\star, \mathbf{y}) \leq f_2(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star) \leq \max_{\mathbf{y}':\|\mathbf{y}'-\mathbf{y}^\star\| \leq h(\delta)} f_2(\mathbf{x}, \mathbf{y}').$$

□

D. Proofs for Section 3.2

In this section, we present proofs of 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 $(\mathbf{x}, \mathbf{y})$ satisfies $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$ and $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$.*

Proof. Since $\mathbf{y}$ is the local maximum of $f(\mathbf{x}, \cdot)$, we have $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$. Denote local optima $\delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}}) := \operatorname{argmax}_{\|\delta_{\mathbf{y}}\| \le h(\delta)} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}})$. By definition we know that $\|\delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})\| \le h(\delta) \to 0$ as $\delta \to 0$. Thus

$$\begin{aligned}
0 &\le f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) - f(\mathbf{x}, \mathbf{y}) \\
&= f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) - f(\mathbf{x}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) + f(\mathbf{x}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) - f(\mathbf{x}, \mathbf{y}) \\
&\le f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) - f(\mathbf{x}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) \\
&= \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}}))^\top \delta_{\mathbf{x}} + o(\|\delta_{\mathbf{x}}\|) \\
&= \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})^\top \delta_{\mathbf{x}} + o(\|\delta_{\mathbf{x}}\|)
\end{aligned}$$

holds for any small $\delta_{\mathbf{x}}$, which implies $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$. □

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

$$[\nabla_{\mathbf{xx}}^2 f - \nabla_{\mathbf{xy}}^2 f (\nabla_{\mathbf{yy}}^2 f)^{-1} \nabla_{\mathbf{yx}}^2 f](\mathbf{x}, \mathbf{y}) \succeq \mathbf{0}. \tag{4}$$

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

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

Since $(\mathbf{x}, \mathbf{y})$ is a local minimax point, by definition there exists a function h such that Eq. (3) holds. Denote $h'(\delta) = 2\|\mathbf{B}^{-1}\mathbf{C}^\top\|\delta$. We note both $h(\delta)$ and $h'(\delta) \to 0$ as $\delta \to 0$. In case that $\mathbf{B} \prec 0$, we know $\mathbf{B}$ is invertible, and it is not hard to verify that $\operatorname{argmax}_{\|\delta_{\mathbf{y}}\| \le \max(h(\delta), h'(\delta))} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}) = -\mathbf{B}^{-1}\mathbf{C}^\top \delta_{\mathbf{x}} + o(\|\delta_{\mathbf{x}}\|)$. Since $(\mathbf{x}, \mathbf{y})$ is a local minimax point, we have

$$\begin{aligned}
0 &\le \max_{\|\delta_{\mathbf{y}}\| \le h(\delta)} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}) - f(\mathbf{x}, \mathbf{y}) \le \max_{\|\delta_y\| \le \max(h(\delta), h'(\delta))} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}) - f(\mathbf{x}, \mathbf{y}) \\
&= \frac{1}{2}\delta_{\mathbf{x}}^\top (\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top)\delta_{\mathbf{x}} + o(\|\delta_{\mathbf{x}}\|^2).
\end{aligned}$$

This equation holds for any $\delta_{\mathbf{x}}$, which finishes the proof. □

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

$$[\nabla_{\mathbf{xx}}^2 f - \nabla_{\mathbf{xy}}^2 f (\nabla_{\mathbf{yy}}^2 f)^{-1} \nabla_{\mathbf{yx}}^2 f](\mathbf{x}, \mathbf{y}) \succ \mathbf{0} \tag{5}$$

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

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

$$\begin{aligned}
\max_{\|\delta_{\mathbf{y}}\| \le h(\delta)} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}) - f(\mathbf{x}, \mathbf{y}) &\ge f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\dagger) - f(\mathbf{x}, \mathbf{y}) \\
&= \frac{1}{2}\delta_{\mathbf{x}}^\top (\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top)\delta_{\mathbf{x}} + o(\|\delta_{\mathbf{x}}\|^2) > 0,
\end{aligned}$$

which finishes the proof. □

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

Proof. Consider the function $f(x, y) = 0.2xy - \cos(y)$ in the 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 $y^\star(x)$ satisfying $0.2x + \sin(y^\star) = 0$ where $y_1^\star(x) \in (-3\pi/2, -\pi/2)$ and $y_2^\star(x) \in (\pi/2, 3\pi/2)$. On the other hand, $f(x, y_1^\star(x))$ is monotonically decreasing with respect to x, while $f(x, y_2^\star(x))$ is monotonically increasing, with $f(0, y_1^\star(0)) = f(0, y_2^\star(0))$ by symmetry. It is not hard to check $y_1^\star(0) = -\pi$ and $y_2^\star(0) = \pi$. Therefore, $(0, -\pi)$ and $(0, \pi)$ are two global solutions of the 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. □

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 $(x^\star, y^\star)$ is a local minimax point, then at least $y^\star$ is a local maximum of $f(x^\star, \cdot)$, which restricts the possible local minimax points to be within the set $[-1, 1) \times \{1\}$ or $(-1, 1] \times \{-1\}$. It is easy to check that no point in either set is local minimax. □

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

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

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

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

$$f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) = \max_{\|\delta_{\mathbf{y}}\| \le h(\delta)} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}).$$

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

$$f(\mathbf{x}, \mathbf{y}) \le \max_{\mathbf{y}'} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y}') = f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}^\star(\delta_{\mathbf{x}})) = \max_{\|\delta_{\mathbf{y}}\| \le h(\delta)} f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{y} + \delta_{\mathbf{y}}),$$

which finishes the proof. □

E. Proofs for Section 3.3

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

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

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

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

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

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

Proof. Daskalakis and Panageas (2018) showed the proposition holds for 1-GDA. For completeness, here we show how similar proof goes through for γ-GDA for general γ. Let $\epsilon = 1/\gamma$, and denote $\mathbf{A} := \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y})$, $\mathbf{B} := \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y})$ and $\mathbf{C} := \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y})$.

To prove the statement $\mathit{localNash} \subset \gamma\text{-}\mathcal{GDA}$, we note by definition, $(\mathbf{x}, \mathbf{y})$ is a strict linear stable point of $1/\epsilon$-GDA if the real part of the eigenvalues of Jacobian matrix

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

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

$$\tilde{J}_\epsilon := \begin{pmatrix} \mathbf{B} & \sqrt{\epsilon}\mathbf{C}^\top \\ -\sqrt{\epsilon}\mathbf{C} & -\epsilon\mathbf{A} \end{pmatrix} = \mathbf{U} J_\epsilon \mathbf{U}^{-1}, \text{ where } \mathbf{U} = \begin{pmatrix} 0 & \sqrt{\epsilon}\mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}$$

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

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

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

$$\begin{aligned} \mathrm{Re}(\lambda) =& [\bar{\mathbf{w}}^\top \tilde{J}_\epsilon \mathbf{w} + \mathbf{w}^\top \tilde{J}_\epsilon \bar{\mathbf{w}}]/2 = [(\mathbf{x} - i\mathbf{y})^\top \tilde{J}_\epsilon (\mathbf{x} + i\mathbf{y}) + (\mathbf{x} + i\mathbf{y})^\top \tilde{J}_\epsilon (\mathbf{x} - i\mathbf{y})]/2 \\ =& \mathbf{x}^\top \tilde{J}_\epsilon \mathbf{x} + \mathbf{y}^\top \tilde{J}_\epsilon \mathbf{y} = \mathbf{x}^\top \mathbf{P}\mathbf{x} + \mathbf{y}^\top \mathbf{P}\mathbf{y} + \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{y}^\top \mathbf{Q}\mathbf{y} \end{aligned}$$

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

To prove the statement $\gamma\text{-}\mathcal{GDA} \not\subset \mathit{localNash}$, since ϵ 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

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

has two eigenvalues $\epsilon(-1 \pm i\sqrt{7})/2$. Therefore, $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2) < 0$, 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 γ, there exists a twice-differentiable f such that $\mathit{Local\_Minimax} \not\subset \gamma\text{-}\mathcal{GDA}$; there also exists a twice-differentiable f such that $\gamma\text{-}\mathcal{GDA} \not\subset \mathit{Local\_Minimax} \cup \mathit{Local\_Maximin}$.*

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

To prove the first statement $\mathit{localminimax} \not\subset \gamma\text{-}\mathcal{GDA}$, since ϵ 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 $A = -2, B = -\epsilon, C = 2\sqrt{\epsilon}$. 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:

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

We know the two eigenvalues are $\epsilon(1 \pm i\sqrt{7})/2$. Therefore, $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2) > 0$, which implies $(0, 0)$ is not a strict linear stable point.

To prove the second statement $\gamma\text{-}\mathcal{GDA} \not\subset \mathit{localminimax} \cup \mathit{localmaximin}$, since ϵ is also fixed, we consider function $f(\mathbf{x}, \mathbf{y}) = x_1^2 + 2\sqrt{\epsilon}x_1 y_1 + (\epsilon/2)y_1^2 - x_2^2/2 + 2\sqrt{\epsilon}x_2 y_2 - \epsilon y_2^2$. It is easy to see $(\mathbf{0}, \mathbf{0})$ is a fixed point of $1/\epsilon$-GDA, and Hessian $\mathbf{A} = \mathrm{diag}(2, -1), \mathbf{B} = \mathrm{diag}(\epsilon, -2\epsilon), \mathbf{C} = 2\sqrt{\epsilon} \cdot \mathrm{diag}(1, 1)$. Thus the Jacobian matrix

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

has four eigenvalues $\epsilon(-1 \pm i\sqrt{7})/2$ (each with multiplicity of 2). Therefore, $\mathrm{Re}(\lambda_i) < 0$ for $1 \le i \le 4$, which implies $(\mathbf{0}, \mathbf{0})$ is a strict linear stable point. However, $\mathbf{B}$ is not negative definite, thus $(\mathbf{0}, \mathbf{0})$ is not a strict local minimax point; similarly, $\mathbf{A}$ is also not positive definite, thus $(\mathbf{0}, \mathbf{0})$ is not a strict local maximin point. □

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

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

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

According to Lemma 40, for sufficient ϵ, J_ϵ has $d_1 + d_2$ complex eigenvalues $\{\lambda_i\}_{i=1}^{d_1+d_2}$ with following form for sufficient small ϵ:

$$\begin{aligned} &|\lambda_i + \epsilon\mu_i| = o(\epsilon) && 1 \le i \le d_1 \\ &|\lambda_{i+d_1} - \nu_i| = o(1), && 1 \le i \le d_2 \end{aligned} \tag{7}$$

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

First, for $\underline{\infty\text{-}\mathcal{GDA}} \subset \overline{\infty\text{-}\mathcal{GDA}}$ always holds by their definitions.

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

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

By Eq.(7) the eigenvalue structure of J_ϵ, we know there exists sufficiently small ϵ_0, so that for any $\epsilon < \epsilon_0$, the real part $\mathrm{Re}(\lambda_i) < 0$, i.e. $(\mathbf{x}, \mathbf{y})$ is a strict linear stable point of $1/\epsilon-$GDA.

Finally, for $\overline{\infty\text{-}\mathcal{GDA}} \subset \mathcal{Local\_Minimax} \cup \{(\mathbf{x}, \mathbf{y}) | (\mathbf{x}, \mathbf{y})$ is stationary and $\mathbf{B}$ is degenerate$\}$ statement, if $(\mathbf{x}, \mathbf{y})$ is strict linear stable point of $1/\epsilon-$GDA for a sufficiently small ϵ, then for any i, the real part of eigenvalue of J_ϵ: $\mathrm{Re}(\lambda_i) < 0$. By Eq.(7), if $\mathbf{B}$ is invertible, this implies:

$$\mathbf{B} \prec 0, \quad \text{and} \quad \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top \succeq 0$$

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

Lemma 40. *For any symmetric matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{B} \in \mathbb{R}^{d_2 \times d_2}$, and any rectangular matrix $\mathbf{C} \in \mathbb{R}^{d_1 \times d_2}$, assume $\mathbf{B}$ is nondegenerate. Then, matrix*

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

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

$$\begin{aligned} &|\lambda_i + \epsilon\mu_i| = o(\epsilon) && 1 \le i \le d_1 \\ &|\lambda_{i+d_1} - \nu_i| = o(1), && 1 \le i \le d_2 \end{aligned}$$

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

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

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

We can expand this polynomial as:

$$p_\epsilon(\lambda) = p_0(\lambda) + \sum_{i=1}^{d_1+d_2} \epsilon^i p_i(\lambda), \qquad p_0(\lambda) = \lambda^{d_1} \cdot \det(\lambda \mathbf{I} - \mathbf{B}).$$

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 41, we know the roots of p_ϵ satisfy:

$$\begin{aligned} &|\lambda_i| = o(1) && 1 \le i \le d_1 \\ &|\lambda_{i+d_1} - \nu_i| = o(1), && 1 \le i \le d_2 \end{aligned} \tag{8}$$

Since $\mathbf{B}$ is non-degenerate, we know when ϵ is small enough, $\lambda_1 \dots \lambda_{d_1}$ are very close to 0 while $\lambda_{d_1+1} \dots \lambda_{d_1+d_2}$ 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:

$$p_\epsilon(\epsilon\theta) = \det \begin{pmatrix} \epsilon\theta \mathbf{I} + \epsilon \mathbf{A} & \epsilon \mathbf{C} \\ -\mathbf{C}^\top & \epsilon\theta \mathbf{I} - \mathbf{B} \end{pmatrix} = \epsilon^{d_1} \det \begin{pmatrix} \theta \mathbf{I} + \mathbf{A} & \mathbf{C} \\ -\mathbf{C}^\top & \epsilon\theta \mathbf{I} - \mathbf{B} \end{pmatrix}$$

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

$$q_\epsilon(\theta) = q_0(\theta) + \sum_{i=1}^{d_2} \epsilon^i q_i(\lambda), \qquad q_0(\theta) = \det \begin{pmatrix} \theta \mathbf{I} + \mathbf{A} & \mathbf{C} \\ -\mathbf{C}^\top & -\mathbf{B} \end{pmatrix}$$

It is also clear polynomial q_ϵ and p_ϵ have same roots up to ϵ scaling. Furthermore, we have following factorization:

$$\begin{pmatrix} \theta \mathbf{I} + \mathbf{A} & \mathbf{C} \\ -\mathbf{C}^\top & -\mathbf{B} \end{pmatrix} = \begin{pmatrix} \theta \mathbf{I} + \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{C} \\ 0 & -\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{I} & 0 \\ \mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix}$$

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

$$q_0(\theta) = (-1)^{d_2} \det(\mathbf{B}) \det(\theta \mathbf{I} + \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top)$$

q_0 is d_1-order polynomial having roots $\{\mu_i\}_{i=1}^{d_1}$, which are the eigenvalues of matrices $\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top$. According to Lemma 41, we know q_ϵ has at least d_1 roots so that $|\theta_i + \mu_i| \le o(1)$. This implies d_1 roots of p_ϵ so that:

$$|\lambda_i + \epsilon\mu_i| = o(\epsilon) \qquad 1 \le i \le d_1$$

By Eq.(8), we know p_ϵ has exactly d_1 roots which are of $o(1)$ scaling. This finishes the proof. □

Lemma 41 (Continuity of roots of polynomials (Zedek, 1965))**.** *Given a polynomial $p_n(z) := \sum_{k=0}^n a_k z^k$, $a_n \neq 0$, an integer $m \ge n$ and a number $\epsilon > 0$, there exists a number $\delta > 0$ such that whenever the $m+1$ complex numbers $b_k, 0 \le k \le m$, satisfy the inequlities*

$$|b_k - a_k| < \delta \text{ for } 0 \le k \le n, \text{ and } |b_k| < \delta \text{ for } n+1 \le k \le m$$

then the roots $\beta_k, 1 \le k \le m$ 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 $\alpha_k, 1 \le k \le n$ of $p_n(z)$ the inequalities

$$|\beta_k - \alpha_k| < \epsilon \text{ for } 1 \le k \le n, \text{ and } |\beta_k| > 1/\epsilon \text{ for } n+1 \le k \le m$$

F. Proof for Theorem 31

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

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

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

Proof of Theorem 31. The proof of this theorem mostly follows the proof of Theorem 2.1 from (Davis and Drusvyatskiy, 2018). The only difference is that y_t in Algorithm 2 is only an approximate maximizer and not exact maximizer. However, the proof goes through fairly easily with an additional error term.

We first note an important equation for the gradient of Moreau envelope.

$$\nabla \phi_\lambda(\mathbf{x}) = \lambda^{-1}\left(\mathbf{x} - \operatorname*{argmin}_{\widetilde{\mathbf{x}}}\left(\phi(\widetilde{\mathbf{x}}) + \frac{1}{2\lambda}\|\mathbf{x}-\widetilde{\mathbf{x}}\|^2\right)\right). \tag{9}$$

We also observe that since $f(\cdot)$ is ℓ-smooth and y_t is an approximate maximizer for $\mathbf{x}_t$, we have that any $\mathbf{x}_t$ from Algorithm 2 and $\widetilde{\mathbf{x}}$ satisfy

$$\begin{aligned}\phi(\widetilde{\mathbf{x}}) \geq f(\widetilde{\mathbf{x}}, \mathbf{y}_t) &\geq f(\mathbf{x}_t, \mathbf{y}_t) + \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \widetilde{\mathbf{x}} - \mathbf{x}_t\rangle - \frac{\ell}{2}\|\widetilde{\mathbf{x}} - \mathbf{x}_t\|^2 \\ &\geq \phi(\mathbf{x}_t) - \epsilon + \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \widetilde{\mathbf{x}} - \mathbf{x}_t\rangle - \frac{\ell}{2}\|\widetilde{\mathbf{x}} - \mathbf{x}_t\|^2. \end{aligned} \tag{10}$$

Let $\widehat{\mathbf{x}}_t := \operatorname{argmin}_{\mathbf{x}} \phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2$. We have:

$$\begin{aligned}\phi_{1/2\ell}(\mathbf{x}_{t+1}) &\leq \phi(\widehat{\mathbf{x}}_t) + \ell\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \\ &\leq \phi(\widehat{\mathbf{x}}_t) + \ell\|\mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \widehat{\mathbf{x}}_t\|^2 \\ &\leq \phi(\widehat{\mathbf{x}}_t) + \ell\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + 2\ell\eta\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \widehat{\mathbf{x}}_t - \mathbf{x}_t\rangle + \eta^2\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\leq \phi_{1/2\ell}(\mathbf{x}_t) + 2\eta\ell\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \widehat{\mathbf{x}}_t - \mathbf{x}_t\rangle + \eta^2\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\leq \phi_{1/2\ell}(\mathbf{x}_t) + 2\eta\ell\left(\phi(\widehat{\mathbf{x}}_t) - \phi(\mathbf{x}_t) + \epsilon + \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2\right) + \eta^2\ell L^2, \end{aligned}$$

where the last line follows from (10). Taking a telescopic sum over t, we obtain

$$\phi_{1/2\ell}(\mathbf{x}_T) \leq \phi_{1/2\ell}(\mathbf{x}_0) + 2\eta\ell\sum_{t=0}^{T}\left(\phi(\widehat{\mathbf{x}}_t) - \phi(\mathbf{x}_t) + \epsilon + \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2\right) + \eta^2\ell L^2 T$$

Rearranging this, we obtain

$$\frac{1}{T+1}\sum_{t=0}^{T}\left(\phi(\mathbf{x}_t) - \phi(\widehat{\mathbf{x}}_t) - \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2\right) \leq \epsilon + \frac{\phi_{1/2\ell}(\mathbf{x}_0) - \min_{\mathbf{x}}\phi(\mathbf{x})}{2\eta\ell T} + \frac{\eta L^2}{2}. \tag{11}$$

Since $\phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2$ is ℓ-strongly convex, we have

$$\begin{aligned}&\phi(\mathbf{x}_t) - \phi(\widehat{\mathbf{x}}_t) - \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \\ &\geq \phi(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \mathbf{x}_t\|^2 - \phi(\widehat{\mathbf{x}}_t) - \ell\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \\ &= \left(\phi(\mathbf{x}_t) + \ell\|\mathbf{x}_t - \mathbf{x}_t\|^2 - \min_{\mathbf{x}}\phi(\mathbf{x}) + \ell\|\mathbf{x} - \mathbf{x}_t\|^2\right) + \frac{\ell}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \\ &\geq \ell\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 = \frac{1}{4\ell}\|\nabla\phi_{1/2\ell}(\mathbf{x}_t)\|^2, \end{aligned}$$

where we used (9) in the last step. Plugging this in (11) proves the result. □