Global Convergence to Local Minmax Equilibrium in Classes of Nonconvex Zero-Sum Games

Tanner Fiez
University of Washington
tannerfiez@uw.edu

Lillian J. Ratliff
University of Washington
ratliff1@uw.edu

Eric Mazumdar
University of California, Berkeley
mazumdar@berkeley.edu

Evan Faulkner
University of Washington
evanf@uw.edu

Adhyyan Narang
University of Washington
adhyyan@uw.edu

Abstract

We study gradient descent-ascent learning dynamics with timescale separation in unconstrained continuous action zero-sum games where the minimizing player faces a nonconvex optimization problem and the maximizing player optimizes a Polyak-Łojasiewicz (PŁ) or strongly-concave (SC) objective. In contrast to past work, we assess convergence in relation to game-theoretic equilibria instead of only notions of stationarity. In pursuit of this goal, we prove that the only locally stable points of the continuous-time limiting system correspond to strict local minmax equilibria in each class of games. For the class of nonconvex-PŁ games, we exploit timescale separation to construct a potential function that when combined with the stability characterization and an asymptotic saddle avoidance result gives a global asymptotic almost-sure convergence guarantee to the set of strict local minmax equilibria. For the class of nonconvex-SC games, we show the surprising property that the function of the game can be made a potential with timescale separation. Combining this insight with the stability characterization allows us to generalize methods for efficiently escaping saddle points in nonconvex optimization to obtain a global finite-time convergence guarantee to local minmax equilibria.

1 Introduction

We study continuous action zero-sum games of the form

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

where $f \in C^2(X \times Y, \mathbb{R})$ and $X = \mathbb{R}^{d_1}$ and $Y = \mathbb{R}^{d_2}$ denote the individual action spaces and $d = d_1 + d_2$. In particular, we focus on zero-sum games in which $f(\cdot, y)$ is potentially nonconvex in $x \in X$ and $f(x, \cdot)$ satisfies the Polyak-Łojasiewicz (PŁ) condition or is strongly-concave (SC) in $y \in Y$. We refer to these classes of games as nonconvex-PŁ and nonconvex-SC zero-sum games.

This general formulation has a broad spectrum of applications such as fair classification [Nouiehed et al., 2019], distributionally robust optimization [Namkoong and Duchi, 2016; Rafique et al., 2021], and adversarial learning [Madry et al., 2018]. Consequently, there has been a surge of interest in recent years toward developing methods for solving these problems efficiently. So far, existing work on gradient-based learning in nonconvex-PŁ/SC zero-sum games has exclusively focused on providing global convergence rates to approximate stationary points with no attention given to the characterization in terms of game-theoretic equilibrium concepts [Lin et al., 2020a,b; Lu et al., 2020; Nouiehed et al., 2019; Rafique et al., 2021; Yang et al., 2020].

In contrast, a common theme in the study of general nonconvex-nonconcave zero-sum games is to assess the types of stationary points an algorithm locally converges toward in terms of their higher order structure [Daskalakis and Panageas, 2018; Fiez and Ratliff, 2021; Fiez et al., 2020; Jin et al., 2020; Mazumdar et al., 2021].
The purpose of this analysis is generally to determine whether commonly deployed algorithms can guarantee local convergence to only game-theoretic equilibria or to design algorithms that achieve this objective.

The goal of this paper is to close the gap between the two problem classes and determine whether gradient-based learning algorithms in nonconvex-PL/SC zero-sum games can be shown to globally converge to only game-theoretically meaningful equilibria (local minmax or Nash equilibria). We focus our attention on the canonical gradient descent-ascent learning dynamics with timescale separations between players and a stochastically perturbed variant. In these algorithms, timescale separation is manifested in the different (yet constant) learning rates of the maximizing and minimizing players. The descriptions of these systems, which we refer to as $\tau$-GDA and $\tau$-PGDA where $\tau > 0$ parameterizes the timescale separation between players, are provided in Algorithm 1 and Algorithm 2 respectively. Simply put, $\tau$-GDA (Algorithm 1) corresponds to each player following their individual gradient in a noiseless setting, while $\tau$-PGDA (Algorithm 2) describes each player following their individual (potentially stochastic) gradient with artificial noise injections.

1.1 Contributions

We show that $\tau$-GDA and $\tau$-PGDA have global convergence guarantees in nonconvex-PL/SC zero-sum games to the natural game-theoretic solution concept for this problem class of strict local minmax equilibria. The contributions of this work are summarized as follows.

1) In Theorem 1, we prove the only critical points that are locally stable with respect to the $\tau$-GDA continuous-time limiting system are strict local minmax in nonconvex-PL/SC zero-sum games.

2) In Theorem 2, we combine Theorem 1 with a potential function construction and an asymptotic saddle avoidance result to prove that $\tau$-GDA (Algorithm 1) globally asymptotically converges to strict local minmax equilibria almost surely in nonconvex-PL/SC zero-sum games. In Corollary 1, we show a corresponding local convergence rate of $O(\varepsilon^{-3})$ to an $\varepsilon$-strict local minmax equilibria.

3) In Theorem 3, for the class of nonconvex-SC zero-sum games, we provide high-probability finite-time rates showing that $\tau$-PGDA (Algorithm 2) globally converges to $\varepsilon$-local minmax equilibria with a complexities of $O(\varepsilon^{-4})$ and $O(\varepsilon^{-2})$ in stochastic and deterministic problems, respectively.

To our knowledge, our results provide the broadest existing global convergence guarantee for gradient-based algorithms to game-theoretically meaningful equilibria in zero-sum continuous games.

1.2 Practical Motivation

The study of nonconvex-PL/SC zero-sum games has often been motivated by machine learning problems formulated as games. We remark that given the problem formulations, it is natural to seek notions of minmax equilibria as solutions. Indeed, notions of stationarity are not guaranteed to reflect a meaningful solution since they could even correspond to maxmin solutions. Consequently, it is critical to give convergence guarantees to minmax equilibrium as we pursue in this work. We now provide examples of machine learning applications that belong to the class of games we consider.

Example 1. In the fair classification problem [Nouiehed et al., 2019], the objective is to minimize the maximum loss over multiple categories. An example formulation is

$$\min_W \max_{i=1,\ldots,m} \{L_i(W)\}$$

where $L_i$ represents the loss on category $i$ and $W$ denotes parameter weights of a neural network.

Example 2. To train a robust neural network against adversarial attacks, a common approach is to formulate training as a robust min-max optimization problem of the form

$$\min_w \sum_{i=1}^N \max_{j=1,\ldots,m} \{\ell(f(x_{ij}(w)), y_i)\}$$

1 Strict local minmax equilibria are also known as differential Stackelberg equilibria in the literature [Fiez and Ratliff, 2021; Fiez et al., 2020] and we use the terms interchangeably in this work.
Table 1: The gradient complexity of gradient descent (GD), gradient descent-ascent with timescale separation ($\tau$-GDA), and alternating (including multi-step) gradient descent-ascent (AGDA) or perturbed variants (PGD, $\tau$-PGDA) in deterministic and stochastic nonconvex optimization, nonconvex-PŁ zero-sum games, and nonconvex-SC zero-sum games. We state the complexity in terms of the $\varepsilon$ tolerance of the guarantee and the dimension $d$ with the notation $\tilde{O}(\cdot)$ hiding logarithmic factors in $\varepsilon$ and $d$.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm &amp; Reference</th>
<th>Complexity</th>
<th>Guarantee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td></td>
<td>$\tilde{O}(\varepsilon^{-2})$</td>
<td>$\varepsilon$–Stationarity</td>
</tr>
<tr>
<td>Stochastic</td>
<td></td>
<td>$\tilde{O}(\varepsilon^{-4})$</td>
<td>$\varepsilon$–Stationarity</td>
</tr>
<tr>
<td>Nonconvex Optimization</td>
<td>GD (Jin et al., 2021)</td>
<td>$\tilde{O}(\varepsilon^{-2})$</td>
<td>$\varepsilon$–Stationarity</td>
</tr>
<tr>
<td></td>
<td>PGDA (Jin et al., 2021)</td>
<td>$\tilde{O}(\varepsilon^{-4})$</td>
<td>$\varepsilon$–Local Min</td>
</tr>
<tr>
<td>Nonconvex SC Zero-Sum</td>
<td>$\tau$-GDA, $\tau$-AGDA (Lin et al., 2020a)</td>
<td>$\tilde{O}(\varepsilon^{-2})$</td>
<td>$\varepsilon$–Stationarity</td>
</tr>
<tr>
<td></td>
<td>Theorem 3 ($\tau$-PGDA)</td>
<td>$\tilde{O}(\varepsilon^{-4})$</td>
<td>$\varepsilon$–DSE</td>
</tr>
<tr>
<td>Nonconvex PL Zero-Sum</td>
<td>AGDA (Nouiehed et al., 2019)</td>
<td>$\tilde{O}(\varepsilon^{-2})$</td>
<td>$\varepsilon$–Stationarity</td>
</tr>
<tr>
<td></td>
<td>Theorem 2 ($\tau$-GDA)</td>
<td>Asymptotic</td>
<td>$\varepsilon$–DSE</td>
</tr>
</tbody>
</table>

where $w$ is the parameter vector of the neural network and each $\hat{x}_{ij}(w)$ is the result of a targeted attack on the sample $x_i$, seeking to change the output of the network for label $j$ [Madry et al., 2018; Nouiehed et al., 2019].

**Example 3.** Distributionally robust optimization and robust learning from multiple distributions is commonly formulated as a minmax optimization problem of the form

$$
\min_{x \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} \sum_{i=1}^{n} y_i f(x) - r(y)
$$

where $f_i(x)$ is the loss of a model $x$ on the $i$-th data point, $\mathcal{Y}$ is the simplex in $\mathbb{R}^n$, and $r(y)$ is carefully selected regularizer [Madry et al., 2018; Namkoong and Duchi, 2016; Rafique et al., 2021].

A common approach in each problem is to transform the inner maximization problem through relaxations and regularization methods to give an unconstrained problem in the parameter space.

## 2 Related Work

We now cover the most relevant related work with further discussion provided in Appendix A.

**Nonconvex-Nonconcave Zero-Sum Games.** A common theme in analyzing gradient descent-ascent with or without timescale separation in nonconvex-nonconcave zero-sum games has been to assess the local stability around critical points of the continuous-time limiting system and draw connections to the differential Nash and Stackelberg equilibrium notions (see Definition 4) [Daskalakis and Panageas, 2018; Fiez and Ratliff, 2021; Fiez et al., 2020; Jin et al., 2020; Mazumdar et al., 2020; Mescheder et al., 2018; Nagarajan and Kolter, 2017; Zhang et al., 2020a]. Importantly, unless the timescale separation is chosen very carefully, the stable critical points of gradient descent-ascent may not be game-theoretically meaningful [Fiez and Ratliff, 2021; Jin et al., 2020]. We obtain stronger stability characterizations in our analysis of the continuous-time system using the structure of nonconvex-PŁ/SC zero-sum games. Further discussion of this topic is given in Section 4.

**Nonconvex-PL and Nonconvex-SC Zero-Sum Games.** Table I provides a comprehensive comparison between our results and existing results for gradient descent variants in nonconvex optimization and nonconvex-PL/SC zero-sum games. We leave discussion of research on these classes of games with other algorithmic methods to Appendix A. The key distinction between this paper and past work is that instead of assessing convergence in terms of only reaching an approximate stationary point of the dynamics or a surrogate function, we obtain convergence guarantees in regards to differential Stackelberg equilibria (equivalently strict local minmax equilibria). Despite this being a much stricter and meaningful notion of solving the problem, we
show that global asymptotic convergence guarantees remain obtainable in nonconvex-PL zero-sum games (Theorem 2) and that the local convergence rate (Corollary 1) is comparable to existing results. Moreover, for the subclass of nonconvex-SC zero-sum games, we provide novel global finite-time convergence guarantees to differential Stackelberg equilibria with rates comparable to existing results for finding a local minimum in nonconvex optimization or any stationary point in this class of games (Theorem 3).

**Escaping Saddle Points in Nonconvex Optimization.** A key contribution of this work is in regards to escaping saddles of the dynamics in classes of nonconvex zero-sum games. We build on methods from nonconvex optimization to obtain analogous results. In general, saddle avoidance results for variants of gradient descent in nonconvex optimization are asymptotic or finite-time. The former states that almost surely the algorithm does not converge to saddles points (Lee et al., 2016, 2019), while the latter gives rates of escape to conclude convergence to approximate local minimum (Ge et al., 2015; Jin et al., 2017, 2021). A key assumption in the aforementioned works is what is known as the strict saddle property, which ensures directions of escape exists from a saddle point. We make an analogous assumption refined for games. Further discussion of how we extend the methods from nonconvex optimization to the zero-sum game problem along with the challenges is included in Section 6.

3 Preliminaries

In the zero-sum games we study, we refer to the minimizing player controlling $x$ as player 1 and the maximizing player controlling $y$ as player 2. We consider objective functions $f \in C^2(Z, \mathbb{R})$ where the joint strategy space is denoted by $Z = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ denote the individual action spaces and $d = d_1 + d_2$. We often denote a joint strategy by $z = (x, y) \in Z$.

**Notation.** We denote $\nabla f$ as the total derivative of $f$, $\nabla_i f$ as the derivative of $f$ with respect to the choice variable of player $i$, $\nabla_{ij} f$ as the partial derivative of $\nabla_i f$ with respect to the choice variable of player $j$, and $\nabla_i^2 f$ as the partial derivative of $\nabla_i f$ with respect to the choice variable of player $i$. We let $\| \cdot \|$ denote the 2-norm of vectors unless otherwise specified, $\text{spec}(\cdot)$ denote the set of eigenvalues of a matrix, $\text{Re}(\cdot)$ denote the real part of a complex number, and $\mathbb{C}_+$ and $\mathbb{C}_-$ denote the open left-half and right-half complex plane, respectively. Let $\lambda_{\text{min}}(A)$ denote the eigenvalue of $A$ with the minimum real part, and $\lambda_{\text{max}}(A)$ the eigenvalue of $A$ with the maximum real part.

**Classes of Games.** We study and analyze both nonconvex-PL and nonconvex-SC zero-sum games. Throughout, we make the following assumptions on the class of functions that define the games.

**Assumption 1.** Given a zero-sum game $(f, -f)$ defined by $f \in C^2(Z, \mathbb{R})$, $\nabla_1 f(x, y)$ and $\nabla_2 f(x, y)$ are $L_1$ and $L_2$ Lipschitz, respectively—i.e.,

$$
\|\nabla_1 f(x, y) - \nabla_1 f(x', y')\| \leq L_1(\|x - x'\| + \|y - y'\|),
$$

$$
\|\nabla_2 f(x, y) - \nabla_2 f(x', y')\| \leq L_2(\|x - x'\| + \|y - y'\|).
$$

This assumption immediately implies that the vector of individual gradients given by

$$
g(x, y) = (\nabla_1 f(x, y), -\nabla_2 f(x, y))
$$

is also Lipschitz with parameter $L \leq L_1 + L_2$.

**Assumption 2.** Given a zero-sum game $(f, -f)$ defined by $f \in C^2(Z, \mathbb{R})$, $\nabla^2 f(\cdot, \cdot)$ is $\beta$-Lipschitz with respect to the induced 2-norm. That is, for all $z, z' \in Z$,

$$
\|\nabla^2 f(z) - \nabla^2 f(z')\| \leq \beta \|z - z'\|.
$$

We now define a nonconvex-PL zero-sum game. This class of games allows for the objective function to be nonconvex in $x \in \mathcal{X}$, but it needs to satisfy the Polyak-Łojasiewicz (PL) condition in $y \in \mathcal{Y}$.
The typical solution concept in game theory when an implicit or explicit order of play is present in the structure of the game is the (local) Stackelberg (equivalently minmax in zero-sum games) equilibrium concept. Informally, in nonconvex-PL/SC zero-sum games, a local minmax equilibrium corresponds to a strategy pair \((x^*, y^*) \in \mathcal{Z}\) such that \(x^*\) is a local minimum

Definition 1 (Nonconvex-PL Game). Consider a zero-sum game \((f, -f)\) defined by \(f \in C^2(\mathcal{Z}, \mathbb{R})\). The game is called nonconvex-PL if \(f(x, \cdot)\) is \(\mu\)-PL with respect to the argument \(y \in \mathcal{Y}\). That is, for \(\mu > 0\) and for all \((x, y) \in \mathcal{Z}\),

\[
\|\nabla_x f(x, y)\|^2 \geq 2\mu (\max_{y' \in \mathcal{Y}} f(x, y') - f(x, y)).
\]

Definition 2 (Nonconvex-SC Game). Consider a zero-sum game \((f, -f)\) defined by \(f \in C^2(\mathcal{Z}, \mathbb{R})\). The game is nonconvex-SC if \(f(x, \cdot)\) is \(\mu\)-strongly-concave with respect to the argument \(y \in \mathcal{Y}\). That is, given any \(x \in \mathcal{X}\) and for all \(y, y' \in \mathcal{Y}\),

\[
f(x, y') \leq f(x, y) + \langle \nabla_y f(x, y), y' - y \rangle - \frac{\mu}{2} \|y - y'\|^2.
\]

Learning Dynamics. We study gradient descent-ascent with timescale separation and a perturbed variant. Let \(\gamma\) be the learning rate of player one and \(\tau > 0\) be the timescale parameterization. The deterministic \(\tau\)-GDA dynamics we study are presented in Algorithm 1 and the potentially stochastic variant with injected noise perturbations called \(\tau\)-PGDA is given in Algorithm 2 (see Section 6 for the relevant notation). The key distinction between this study of gradient descent-ascent with timescale separation and past work is how we assess convergence as we now begin to formalize.

Stationarity Notions. A critical point corresponds to a joint strategy profile at which the individual gradient of each player is equal to zero.

Definition 3 (Critical Point). A point \((x, y) \in \mathcal{Z}\) is a critical point if \(\nabla_x f(x, y) = 0\) and \(\nabla_y f(x, y) = 0\).

Critical points correspond to stationary points of the \(\tau\)-GDA dynamics. The nonconvex-PL/SC zero-sum game literature has generally assessed convergence in terms of the complexity of finding an approximate stationary point (see, e.g., (Nouiehed et al. 2019; Yang et al. 2020)). That is, joint strategies \((x, y) \in \mathcal{Z}\) such that \(\|\nabla_x f(x, y)\| \leq \varepsilon\) and \(\|\nabla_y f(x, y)\| \leq \varepsilon\). A related and common notion of convergence in this body of work (see e.g., (Lin et al. 2020a)) is that of finding a stationary point of the function \(\max_{y \in \mathcal{Y}} f(\cdot, y)\). This criterion amounts to seeking to achieve the condition \(\|\nabla_x \max_{y \in \mathcal{Y}} f(x, y)\| \leq \varepsilon\).

In contrast, we assess convergence with connections to the equilibrium notions that are commonly studied in the nonconvex-nonconcave zero-sum game literature. Since either stationarity notion may lack any game-theoretic meaning, we consider a strictly harder notion of solving a game.

Equilibrium Notions. The typical solution concept in game theory when an implicit or explicit order of play is present in the structure of the game is the (local) Stackelberg (equivalently minmax in zero-sum games) equilibrium concept. Informally, in nonconvex-PL/SC zero-sum games, a local minmax equilibrium corresponds to a strategy pair \((x^*, y^*) \in \mathcal{Z}\) such that \(x^*\) is a local minimum.

---

Algorithm 1 \(\tau\)-GDA

| Input: \(x_0 \in \mathbb{R}^d\), \(y_0 \in \mathbb{R}^d\) |
| for \(k = 0, 1, \ldots\) do |
| \(x_{k+1} \leftarrow x_k - \gamma \nabla_1 f(x_k, y_k)\) |
| \(y_{k+1} \leftarrow y_k + \gamma \tau \nabla_2 f(x_k, y_k)\) |
| end for |

Algorithm 2 \(\tau\)-PGDA

| Input: \(x_0 \in \mathbb{R}^d\), \(y_0 \in \mathbb{R}^d\) |
| for \(k = 0, 1, \ldots\) do |
| Sample \(\theta_{i,k} \sim \mathcal{D}_i\), \(i = 1, 2; (\xi_{1,k}, \xi_{2,k}) \sim \mathcal{N}(0, (r^2/d)I)\) |
| \(x_{k+1} \leftarrow x_k - \gamma (g_1(x_k, y_k; \theta_{1,k}) + \xi_{1,k})\), |
| \(y_{k+1} \leftarrow y_k + \gamma \tau (g_2(x_k, y_k; \theta_{2,k}) + \tau^{-1} \xi_{2,k})\) |
| end for |

---

\(^2\)By implicit order of play, we mean the min and max order are not interchangeable.
of the function \( f(x, y_*(x)) \) where \( y_*(x) \in \arg \max_{y \in Y} f(x, y) \) and \( y^* \) is a local maximum of the function \( f(x^*, y) \). When the function \( f \) is bounded or when \( f(\cdot, y) \) is bounded and \( f(x, \cdot) \) is strongly concave, a minmax equilibrium is guaranteed to exist.

We characterize the local minmax (Stackelberg) equilibrium notion in terms of sufficient conditions on player costs as is typical in learning in games (see, e.g., [Fiez and Ratliff, 2021; Fiez et al., 2020; Jin et al., 2020; Mazumdar et al., 2020]). Toward presenting this definition, we denote by \( J(x, y) \) the Jacobian of the vector of individual gradients \( g(x, y) \) that is given by

\[
J(x, y) = \begin{bmatrix}
\nabla_1^2 f(x, y) & \nabla_1 f(x, y) \\
-\nabla_2^2 f(x, y) & -\nabla_2 f(x, y)
\end{bmatrix}.
\]

Let \( S_1(\cdot) \) denote the Schur complement of \( (\cdot) \) with respect to the \( d_2 \times d_2 \) block in \((\cdot)\). The following definition is characterized by sufficient conditions for a local minmax equilibrium in zero-sum games.

**Definition 4** (Differential Stackelberg/Strict Local Minmax Equilibrium (Fiez et al., 2020)). The joint strategy \((x^*, y^*) \in Z\) is a differential Stackelberg equilibrium if the first-order conditions \( \nabla f(x^*, y^*) = \nabla_1 f(x^*, y^*) - \nabla_2 f(x^*, y^*) \left[ \nabla_2^2 f(x^*, y^*) \right]^{-1} \nabla_2 f(x^*, y^*) = 0 \), \( \nabla_2 f(x^*, y^*) = 0 \) hold and the second-order conditions \( S_1(J(x^*, y^*)) > 0 \) and \( \nabla_2^2 f(x^*, y^*) < 0 \) also hold.

A differential Stackelberg equilibrium corresponds to a joint strategy at which the minimizing player is at a local optimum with respect to its choice variable along the best response curve of the maximizing player and the maximizing player is at a local optimum with respect to its choice variable. In the next section, we also provide a definition for an \( \varepsilon \)-differential Stackelberg equilibrium.

### 4 Local Stability Analysis

To characterize the convergence of \( \tau\text{-GDA} \), we begin by studying its continuous-time limiting system

\[
\dot{z} = -\Lambda_\tau g(z)
\]

where \( \dot{z} = (\dot{x}, \dot{y}) \), \( \Lambda_\tau = \text{blockdiag}(I_{d_1}, \tau I_{d_2}) \), and \( g(z) \) is the vector of individual gradients. The Jacobian of this system is given by

\[
J_\tau(z) = \Lambda_\tau J(z).
\]

We analyze the stability of the continuous-time system around critical points \( z^* \) as a function of the timescale separation \( \tau \) using the Jacobian \( J_\tau(z^*) \) in this section toward drawing conclusions about the stability and convergence of the discrete time system \( \tau\text{-GDA} \). A critical point is said to be locally (exponentially) stable when the spectrum of \(-J_\tau(z^*)\) is in the open left-half complex plane \( \mathbb{C}^2_- \) (cf. Theorem 4, Appendix B.2). Simply put, a critical point \( z^* \) is locally exponentially stable if and only if the real parts of the eigenvalues of \(-J_\tau(z^*)\) are strictly negative. Throughout, we use the broader term “stable” to mean the following.

**Definition 5** (Stability). A critical point \( z^* = (x^*, y^*) \in Z \) is locally exponentially stable for \( \dot{z} = -\Lambda_\tau g(z) \) if and only if \( \text{spec}(\Lambda_\tau(z^*)) \subset \mathbb{C}^2_- \). (\( \equiv \text{ spec}(\Lambda_\tau(z^*)) \subset \mathbb{C}^2_+ \).)

Stability with respect to the continuous-time \( \tau\text{-GDA} \) dynamics guarantees that the system asymptotically converges at an exponential rate to the critical point in a local neighborhood. Moreover, given a suitable choice of learning rates, equivalent insights hold for the discrete-time dynamics (Chasnov et al., 2020).

**Stability in Nonconvex-Nonconcave Zero-Sum Games.** Given the implications regarding convergence, a number of papers in the past several years study the stability of \( \tau\text{-GDA} \) around critical points and the connections to differential Nash and Stackelberg equilibria in zero-sum games (Daskalakis and Panageas, 2018; Fiez and Ratliff, 2021; Fiez et al., 2020; Jin et al., 2020; Mazumdar et al., 2020). However, this body of research focuses on general nonconvex-nonconcave zero-sum games. In general across the spectrum of nonconvex-nonconcave zero-sum games, the stable critical points of \( \tau\text{-GDA} \) coincide with the set of differential Stackelberg equilibria only when the timescale separation \( \tau \rightarrow \infty \) (Jin et al., 2020). Given that such a choice of timescale separation requires the learning rate \( \gamma \rightarrow 0 \) in order to retain stability of the discrete-time system, it is not clear how to derive a practical algorithm from this insight.
Toward remedying this problem, Fiez and Ratliff [2021] provide stability results in terms of the timescale separation concerning a given critical point, rather than across the space of nonconvex-nonconcave zero-sum games. Indeed, they prove a stability and instability result as a function of the timescale separation in the τ-GDA dynamics. The stability results say that given a differential Stackelberg equilibrium \( z^* \), there exists a finite \( \tau^* \in (0, \infty) \) that can be constructed such that \( z^* \) is stable for all \( \tau \in (\tau^*, \infty) \). On the other hand, the instability results says that given a critical point which is not a differential Stackelberg equilibrium, there exists a finite \( \tau_0 \in (0, \infty) \) that can be constructed such that \( z^* \) is not stable for all \( \tau \in (\tau_0, \infty) \).

**Stability in Nonconvex-PL/SC Zero-Sum Games.** To our knowledge, the connection between the stability (and instability) of critical points with respect to τ-GDA dynamics and game-theoretic equilibrium notions in the semi-structured problems of nonconvex-PL and nonconvex-SC has not been fully characterized. We show in the following result that when a nonconvex-nonconcave game is specialized to a nonconvex-PL game as from Definition [1] significantly more general stability characterizations can be obtained. Notably, any critical point \( z^* \) that is not a differential Stackelberg equilibrium is unstable for all \( \tau \geq 0 \). Meaning that τ-GDA does not admit spurious stable points in nonconvex-PL game. This is in stark contrast to the known fact that 1-GDA admits spurious stable points in the more general class of nonconvex-nonconcave games as has been shown in previous literature (Daskalakis and Panageas 2018; Mazumdar et al. 2020). Moreover, if \( z^* \) is a differential Stackelberg equilibrium, then \( z^* \) is stable for all \( \tau \) larger than the minimum \( \tau_* \), for which \( z^* \) is stable and such a finite \( \tau_* \) is guaranteed to exist. This result implies that in practice, one can select a finite value of \( \tau \) to run τ-GDA with, and all stable critical points (if they exist) will be differential Stackelberg equilibria and if \( \tau \) is scaled up and the set of stable points grows, then only differential Stackelberg equilibria can be introduced.

**Theorem 1.** Consider a nonconvex-PL zero-sum game \((f, -f)\) where \( f \in C^2(\mathcal{Z}, \mathbb{R}) \). Then, the following hold: 1) Any critical point \( z^* \) that is not a differential Stackelberg equilibrium is unstable for all \( \tau \in (0, \infty) \); 2) If \( z^* \) is a differential Stackelberg equilibrium, then \( \text{spec}(-J_\tau(z^*)) \subset C^- \) for all \( \tau \in (\tau_*, \infty) \) where \( \tau_* \) is the minimum \( \tau \in (0, \infty) \) such that \( \text{spec}(-J_\tau(z^*)) \subset C^- \) and a finite \( \tau_* \) is guaranteed to exist.

In comparison to the stability results for nonconvex-nonconcave zero-sum games of Fiez and Ratliff [2021], we obtain the stronger results in nonconvex-PL zero-sum games that (i) \( \tau_0 = 0 \) for any critical point that is not a differential Stackelberg equilibria and (ii) a differential Stackelberg equilibria is never unstable after it becomes stable as a function of the timescale separation \( \tau \).

The stability characterization also allows us to define a natural approximate differential Stackelberg equilibrium notion. The first-order conditions in Definition [4] can equivalently be reformulated as \( \nabla_1 f(x^*, y^*) = 0 \) and \( \nabla_2^2 f(x^*, y^*) = 0 \). We presented the differential Stackelberg equilibrium using the total derivative for the minimizing player in the first-order condition to mirror the presentation of proper approximate notions. In particular, let \( \hat{g}(x,y) = (\nabla f(x,y), -\nabla_2^2 f(x,y)) \) denote the vector containing the total derivative for the minimizing player and the individual derivative for the maximizing player. We have the following \( \varepsilon \)-differential Stackelberg equilibrium.

**Definition 6.** A point \( z^* = (x^*, y^*) \in \mathcal{X} \times \mathcal{Y} \) is an \( \varepsilon \)-differential Stackelberg equilibrium (ε-local minmax) for a nonconvex–PL zero-sum game \( f \in C^2(\mathcal{Z}, \mathbb{R}) \) satisfying Assumptions [4] and [2] if \( \|\hat{g}(z^*)\| \leq \varepsilon \) and \( \text{Re}(\lambda_{\text{min}}(J_\tau(z^*))) \geq -\sqrt{\varepsilon} \).

### 5 Global Asymptotic Convergence Analysis

Theorem [4] in the previous section completely characterizes the behavior of the continuous-time τ-GDA dynamics. However, ultimately we are concerned with the convergence properties of the discrete-time τ-GDA dynamics. We begin our study of this system by providing a global asymptotic analysis in this section of the deterministic τ-GDA dynamics presented in Algorithm [1].

We prove in Theorem [2] of this section that the deterministic τ-GDA (Algorithm [1]) almost surely converges to a differential Stackelberg equilibrium in nonconvex-PL zero-sum games, a class of games that subsumes nonconvex-SC zero-sum games. Despite this result being asymptotic in nature, to our knowledge it is the most general class of zero-sum games in which a global convergence guarantee to established game-theoretic equilibria has been given.
To begin, observe that the $\tau$-GDA dynamics do not necessarily correspond to a gradient flow a function, which can be observed from the fact that the Jacobian $J_\tau$ is not guaranteed to be symmetric. As a consequence, there may exist complex limiting behavior beyond convergence to a critical point such as non-trivial limit cycles or periodic orbits. To rule out this phenomenon, we prove that there exists potential function that decreases along the iterates of $\tau$-GDA. The best response map is well-defined by the implicit mapping theorem since the PL condition on the maximizing player’s problem implies quadratic growth \cite{karimi2016iteration} which in turn implies that at critical points not only is $\nabla^2 f$ non-degenerate, but also it is negative definite. In order to construct a potential function for nonconvex, $\mu$-PL zero-sum games, we need the best-response map of the maximizing player, $y_\tau(x) \in \text{argmax}_y f(x,y)$, defined implicitly by $\nabla_y f(x,y) = 0$ to be $L_3$-Lipschitz; we show this in Lemma \[7\] in the appendix.\footnote{This mapping is $L(\kappa + 1)$-Lipschitz in the case where $f(x, \cdot)$ is strongly concave \cite[Lemma 4.3]{lin2020cycling}.}

**Lemma 1.** Consider a non-convex, $\mu$-PL zero sum game defined by $f \in C^2(\mathbb{Z}, \mathbb{R})$ which satisfies Assumptions \[1\] and \[2\] and has condition number $\kappa = L_2 / \mu$. Suppose that $\tau > 7\kappa^2$ and $\gamma < \min\{1/(\kappa + \tau), \frac{1}{2L_1 + L_3L_2}\}$ then, for any $\Gamma \in (0, 1/7]$, $\Phi(x,y) = f(x,y(x)) - \Gamma f(x,y)$ is a potential function for $\tau$-GDA.

The function $f(x,y_\tau(x))$ can be seen as the function the $x$ player would minimize if the $y$ player was playing a best-response $y_\tau(x)$. The potential function $\Phi(x,y)$ essentially captures that along trajectories of $\tau$-GDA, the function $f(x,y_\tau(x))$ should either decrease the value of $f(x,y_\tau(x))$, or decrease the value of $f(x,y_\tau(x)) - f(x,y)$ since the $y$ player converges at a fast rate to $y_\tau(x)$ given the time-scale separation. Indeed, this potential function implicitly guarantees that the maximizing player tracks the best response set, and that the minimizing player essentially ends up minimizing $f(x,y_\tau(x))$ as desired. The choice of $\tau$ and the learning rate $\gamma$ allow us to guarantee that this occurs.

Given the potential function in Lemma \[1\] we can conclude that Algorithm \[1\] converges to critical points. The critical points may correspond to stable points of the dynamics (Definition \[5\]) or saddle points of the dynamics, which are defined as follows.\footnote{The terminology of saddle point in this paper is with respect to the dynamics and should not be conflated with the terminology of a saddle point of a function.}

**Definition 7** (Saddle Point). The critical point $z^* = (x^*, y^*) \in \mathbb{Z}$ is a saddle point of the dynamics $\dot{z} = -\Lambda_\tau g(z)$ if $\Re(\lambda_{\max}(-J_\tau(z^*))) = 0$ and a strict saddle if $\Re(\lambda_{\max}(-J_\tau(z^*))) > 0$.

Recall that Theorem \[4\] indicates that the only stable points of the continuous-time $\tau$-GDA dynamics are differential Stackelberg equilibria. Thus, if we can show that the discrete-time $\tau$-GDA dynamics avoid saddle points of the continuous-time $\tau$-GDA dynamics, then we can conclude that Algorithm \[4\] converges to only differential Stackelberg equilibria. The following result of Mazumdar et al. \cite{mazumdar2020} states that the discrete-time $\tau$-GDA dynamics do indeed avoid saddle points of the continuous-time system.\footnote{The saddle avoidance result of Mazumdar et al. \cite{mazumdar2020} holds for both general-sum $n$-player nonconvex games. For simplicity, we only state it in the context of this work.}

**Lemma 2** (Theorem 4.1 \cite{mazumdar2020}). Consider a zero sum game defined by the non-convex, non-concave function $f \in C^2(\mathbb{Z}, \mathbb{R})$. The set of initial conditions $z \in \mathbb{Z}$ from which $\tau$-GDA converges to strict saddle points is of measure zero.
If we fix a $\tau$ satisfying the assumptions of the above theorem, then if we consider the set of stable critical points for $\dot{z} = -\Lambda_{\tau}(z)$, we know that by Theorem 1 this set only contains differential Stackelberg (and hence, strict local minmax) points. It is precisely this set of strict local minmax to which $\tau$-GDA converges almost surely. Moreover, by Theorem 1 as we increase $\tau$, no new spurious non-minmax points are introduced to the set of stable critical points; only additional strict local minmax points can be added to this set.

We complement Theorem 2 with the following characterization of the convergence rate.

**Corollary 1.** Consider a nonconvex-PL zero-sum game $(f, -f)$ defined by $f \in C^2(Z, \mathbb{R})$ that satisfies Assumptions 1–3. Given an initialization in the region of attraction of a strict local minmax, then after around a saddle (with high probability) is achieved by analyzing the local linearization around the regions $\nabla f(z)$. For each $i$, $\nabla f_i(x, y; \theta_i)$ satisfies
\[
\nabla f_i(x, y; \theta_i) = \mathbb{E}_{g_i \sim D_i}[g_i(x, y; \theta_i)] \quad \forall t \in \mathbb{R} \text{ and each } i = 1, 2.
\]

**Assumption 4.** For any $(x, y) \in Z$, the stochastic gradient vector $g(x, y; \theta) = (g_1(x, y; \theta_1), g_2(x, y; \theta_2))$ satisfies
\[
\mathbb{P}(|g_i(x, y; \theta) - \nabla f(x, y)| \geq t) \leq 2 \exp(-\frac{t^2}{2\sigma_i^2}).
\]

**Assumption 5.** For each $i = 1, 2$ and any $\theta_i \in \text{supp}(D_i)$, $g_i(\cdot, \cdot; \theta_i)$ is $L_i$-Lipschitz: i.e., for all $(x_1, y_1), (x_2, y_2) \in Z$, $||g_i(x_1, y_1; \theta_i) - g_i(x_2, y_2; \theta_i)|| \leq L_i ||(x_1, y_1) - (x_2, y_2)||$.

**Assumption 6.** The eigenvalue of $J_{\tau}(z)$ with minimum real part for any strict saddle $z$ is simple. That is, the algebraic multiplicity $m_A(\eta) = 1$ and geometric multiplicity of $m_G(\eta) = 1$ where $-\eta = \text{Re}(\lambda_{\min}(J_{\tau}(z)))$.

Let $\ell = L_1 + L_2$ if $\kappa \leq 1/2$, and $\ell = L_1 + 2\kappa L_2$ otherwise. In either case, $\lambda_{\max}(\nabla^2 f(x, y)) \leq \ell$. In the remainder, we work in the case for which $\ell \geq \sqrt{2\lambda}$; otherwise the problem of finding $\varepsilon$-local minmax is straightforward since all $\varepsilon$-stationary points will be $\varepsilon$-local minmax points.

The steps in showing that $\tau$-PGDA (Algorithm 2) escapes saddles with high probability largely follow those in Daneshmand et al. (2018); Ge et al. (2015); Jin et al. (2017, 2021), with several modifications due to the fact that we consider a zero-sum game as opposed to a single player optimization problem. The challenge in the minmax (or zero-sum game) setting as compared to the single player optimization setting is that the gradient descent-ascent dynamics are not a gradient flow, and hence, the eigenvalues of the local linearization may be complex-valued. In the key step of many of the proofs on saddle avoidance, escaping from the stuck regions around a saddle (with high probability) is achieved by analyzing the local linearization around the saddle point, which crucially, is symmetric. Hence, the problem reduces to analyzing the escape time along the eigendirection of the minimum eigenvalue of the Hessian which relies heavily on the orthogonality of eigenvectors of symmetric matrices. In zero-sum games, the linearization is no longer symmetric and these key results have to be substantially modified. We highlight these distinctions more clearly in Appendix F.

One of the other key distinctions is in constructing a potential function. In the existing literature on convergence in the nonconvex-SC literature (Liu et al. 2020; Wang et al. 2020b; Yang et al. 2020), the
The proposed potential function has a structure closely related to the potential function in Lemma 1 in particular, it takes the form \( f(x, y) + r(x, y) \) where \( r(x, y) \) is a tracking term of the form \( \| y - y_*(x) \|^2 \) or \( f(x, y_*(x)) \). Yet, perhaps surprisingly, despite the fact that \( f \) depends on to sequences generated by two separate gradient updates—one from minimizing \( f \) and one from maximizing it—the descent lemma below (Lemma 3) shows that the function \( f \) can be decomposed into a component that is decreasing along trajectories of \( \tau \)-PGDA (Algorithm 2) and a component that exhibits a possible increase due to randomness in the stochastic gradients and injected noise. The primary reason for the decrease is large gradients in addition to the strongly concave structure of the maximizing player’s problem and the timescale separation which we exploit in the proof to ensure sufficient decrease of \( f \) along trajectories of \( \tau \)-PGDA.

**Lemma 3 (Descent Lemma).** Consider a non-convex, \( \mu \)-SC zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \). Under Assumptions 1–6 if \( \frac{1}{2w} < \gamma < 1/\ell \), then with probability at least \( 1 - \delta \) for some \( \delta > 0 \), \( f(x_k, y_k) \) is a potential function for \( \tau \)-PGDA in both the stochastic and deterministic settings.

The descent lemma enables us to argue that with high probability, either the function value decreases a sufficient amount or the iterates remain in a small region. Using this property of \( \tau \)-PGDA together with the descent lemma, we then show that if two sequences that are coupled (i.e., two sequences starting from the same initial condition and having equivalent noise except along the direction of escape) and are in the region around a saddle point, then with high probability, at least one of them escapes. The full proof is provided in Appendix E.

Putting these results together, we have the following finite time guarantee on convergence to \( \varepsilon \)-differential Stackelberg equilibria (\( \varepsilon \)-local minmax points). Let \( f^* = \min_{x, y} f(x, y) \) be the minimum value of the function \( f \), which we assume is finite. These results provide novel convergence guarantees to not just \( \varepsilon \)-stationary points but those that are game theoretically meaningful.

**Theorem 3.** Consider a non-convex, \( \mu \)-strongly concave zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) and suppose that Assumptions 1–6 hold. For any \( \varepsilon, \delta > 0 \), there exists \( \gamma \) and \( \tau \) such that, with probability \( 1 - \delta \) for some \( \delta > 0 \), starting from any \( z_0 = (x_0, y_0) \), at least half the iterates of \( \tau \)-PGDA will be \( \varepsilon \)-differential Stackelberg equilibria after \( O(\varepsilon^{-4}) \) iterations and \( O(\varepsilon^{-2}) \) in the stochastic and deterministic settings, respectively.

### 7 Discussion

To the best of our knowledge, our results are the first to guarantee global convergence of gradient-based algorithms to game theoretically meaningful equilibria in such a general class of games. We believe that a more detailed analysis could make use of the potential function defined for nonconvex-PL in Section 5 to prove similar finite-time results for \( \tau \)-PGDA, though the proof appears to be tedious and hence, we leave this for future work. We also believe future work can be done to obtain global guarantees in this class of games to game-theoretically meaningful equilibria without gradient information and also in the setting with constraints extending recent saddle avoidance results for the single player nonsmooth, constrained setting.

### Acknowledgements

### References


Appendix

Below we provide a table of contents as a guide to the appendix.

Table of Contents

A Related Work 14
Appendix A contains an extensive discussion of related work on nonconvex-nonconcave zero-sum games, nonconvex-PL/SC zero-sum games, and escaping saddles in nonconvex optimization.

B Preliminaries 17
Appendix B contains preliminaries on game theory (Appendix B.1), dynamical systems (Appendix B.2), linear algebra (Appendix B.4), PL functions and nonconvex-PL zero-sum games (Appendix B.3) and concentration of measure (Appendix B.5). Appendix B.1 includes a formal local Stackelberg equilibrium definition that results in the differential Stackelberg (strict local minmax) equilibrium notion we consider. Appendix B.2 provides background on determining the stability of dynamical systems using the local linearization. Appendix B.3 contains useful properties of PL functions and nonconvex-PL zero-sum games that help in both the stability analysis in Appendix C and for proving the results from Section 5. Appendix B.5 includes concentration inequalities for norm-subGaussian random vectors needed in the analysis of escaping saddle points efficiently given in Appendix E.

C Stability Analysis: Proof of Theorem 1 23
Appendix C contains the proof of Theorem 1 regarding the stability of critical points in nonconvex-PL zero-sum games.

D Global Asymptotic Convergence Analysis 25
Appendix D contains the proofs for the results from Section 5 on the convergence of τ-GDA in nonconvex-PL zero-sum games. In particular, this Appendix D includes the proofs of Lemma 1 (potential function), Theorem 2 (global asymptotic equilibrium convergence guarantee) and Corollary 1 (local convergence rate to ε-differential Stackelberg equilibrium).

E Finite-Time Convergence Analysis 30
Appendix E contains the proof of Theorem 3 regarding the convergence of the τ-PGDA algorithm. This proof is split into subsections, where in Appendix E.1 we prove the stochastic descent lemma (Lemma 3) and a result bounding the iterates of the τ-PGDA algorithm (Lemma 20). Then, in Appendix E.2 we prove the key results on escaping saddle points efficiently in Lemma 21 which follows from several intermediate results. Finally, in Appendix E.3 we formally prove Theorem 3 using the aforementioned results.

A Related Work

We now cover related work in further detail. We begin by discussing past work on nonconvex-nonconcave zero-sum games with a focus on papers that analyze the local stability of critical points using the continuous-time limiting system as we do in Section 4 for nonconvex-PL/SC zero-sum games. Then, we compare our convergence results in nonconvex-PL/SC zero-sum games presented in Section 5 and Section 6 to existing guarantees in the literature for various gradient-based learning algorithms. We remark that there are a number of papers analyzing gradient-based learning in strongly-convex-strongly-concave, strongly-convex-linear, strongly-convex-concave, convex-concave, and nonconvex-concave zero-sum games. Since each of the aforementioned problems are fundamentally different in terms of the structure compared to nonconvex-PL/SC zero-sum games, we do not discuss each body of work in detail and instead refer the reader to the detailed discussion of work on such problems provided in the recent paper of Lin et al. (2020b). Finally, we review techniques and results for escaping saddle points in nonconvex optimization relevant to our analogous methods and guarantees in classes of nonconvex games.
Nonconvex-Nonconcave Zero-Sum Games. In general nonconvex-nonconcave zero-sum games, global convergence guarantees to traditional game-theoretic equilibrium notions are effectively unobtainable as a result of computational hardness results (Daskalakis et al., 2020) and the emergence of non-trivial limit cycles and periodic orbits (Hsieh et al., 2020; Letcher, 2021). Consequently, the analysis of gradient-based learning algorithms in nonconvex-nonconcave zero-sum games has commonly focused on determining the local stability of the continuous-time limiting system around critical points. Typically, the goal is to determine the relationship between the set of stable critical points and the critical points that correspond to game-theoretic equilibrium.

The prototypical equilibrium notions in game theory are the (local) Nash and Stackelberg equilibrium concepts ( Başar and Olsder, 1998). The aforementioned equilibrium notions have been characterized in terms of gradient-based sufficient conditions and critical points satisfying the conditions have sometimes been termed differential Nash (strict local Nash) and differential Stackelberg (strict local Stackelberg) equilibria. The differential characterization of local Nash equilibria was reported by Ratliff et al. (2013), Ratliff et al. (2016), while the differential characterization of local Stackelberg equilibria was given concurrently by Fiez et al. (2020) and Jin et al. (2020), where the conditions presented by the former extend to general-sum games. Moreover, Fiez et al. (2020) and Jin et al. (2020) show that differential Nash equilibria are a subset of differential Stackelberg equilibria. A series of works analyze the properties of the equilibrium notions defined in terms of gradient-based sufficient conditions. In particular, Ratliff et al. (2014) and Mazumdar and Ratliff (2019) prove the genericity and structural stability of differential Nash equilibria in general-sum and zero-sum games, respectively. Analogously, Fiez et al. (2020) prove the genericity and structural stability of differential Stackelberg equilibria in zero-sum games.

Several past works study the local stability of $\tau$-GDA without timescale separation (equal learning rates, $\tau = 1$) around critical points (Daskalakis and Panageas, 2018; Mazumdar et al., 2020). In this direction, each of the aforementioned works prove there can exist stable critical points of $\tau$-GDA with $\tau = 1$ that are not differential Nash equilibria. Moreover, they each show that differential Nash equilibria are stable critical points of $\tau$-GDA with $\tau = 1$.

Building on this line of work, the local stability of $\tau$-GDA with timescale separation around critical points has been analyzed (Fiez and Ratliff, 2021; Jin et al., 2020). We begin by commenting on the results of Jin et al. (2020). It is shown by the authors that differential Nash equilibria are stable critical points of $\tau$-GDA with any timescale separation parameter $\tau > 0$. Moreover, they prove that given any fixed timescale separation $\tau > 0$, there exists a game with a differential Stackelberg equilibrium that is not stable with respect to the $\tau$-GDA system. Finally, they show that as $\tau \to \infty$, the stable critical points of $\tau$-GDA coincide with the set of differential Stackelberg equilibria. Fiez and Ratliff (2021) provide complimentary results by focusing on the stability properties of any given critical point as a function of the timescale separation $\tau$, rather than considering stability properties across the space of nonconvex-nonconcave zero-sum games. In particular, they show a stability and instability result. The stability result says that given a differential Stackelberg equilibrium $(x^*, y^*)$, there exists a finite $\tau^* \in (0, \infty)$ that can be constructed such that $(x^*, y^*)$ is stable for all $\tau \in (\tau^*, \infty)$ with respect to the $\tau$-GDA dynamics. The instability results says that given a critical point $(x^*, y^*)$ which is not a differential Stackelberg equilibrium, there exists a finite $\tau_0 \in (0, \infty)$ that can be constructed such that $(x^*, y^*)$ is not stable for all $\tau \in (\tau_0, \infty)$ with respect to the $\tau$-GDA dynamics. It is important to observe that the constructions of a $\tau^*$ or $\tau_0$ depend on the properties of a given critical point.

The results we provide regarding the local stability of $\tau$-GDA around critical points in nonconvex-PL/SC zero-sum games in Section 4 strengthen the stability characterizations of Fiez and Ratliff (2021) by exploiting the structure of these classes of games. In comparison, we obtain the stronger results in nonconvex-PL zero-sum games that (i) $\tau_0 = 0$ for any critical point that is not a differential Stackelberg equilibria and (ii) a differential Stackelberg equilibria is never unstable after it becomes stable as a function of $\tau$. That is, we obtain insights that apply to the class of games rather than specific critical points, which are easily quantifiable instead of depending on properties of a given critical point.

We remark that another line of work studies the local stability of $\tau$-GDA without timescale separation ($\tau = 1$) in generative adversarial networks under certain assumptions (Mescheder et al., 2018; Nagarajan and Kolter, 2017). The results are generalized to the study of $\tau$-GDA with timescale separation (Fiez and Ratliff, 2021) to show that equivalent conclusions hold for any choice of timescale separation $\tau > 0$.

Finally, there is a line of work developing gradient-based learning algorithms using second order information in nonconvex-nonconcave zero-sum games such that the only locally stable critical points correspond to
game-theoretic equilibrium notions. In particular, Adolphs et al. (2019) and Mazumdar et al. (2019) develop algorithms under which a critical point is locally stable if and only if it is a differential Nash equilibrium. Moreover, Fiez et al. (2020), Wang et al. (2020a), and Zhang et al. (2020b) construct algorithms under which a critical point is locally stable if and only if it is a differential Stackelberg equilibria.

**Nonconvex-PL and Nonconvex-SC Zero-Sum Games.** We now expand our discussion of existing work on nonconvex-PL and nonconvex-SC zero-sum games. As discussed in Section 2 in comparison to past works in this area, the key distinction of this work is that we give global convergence guarantees to game-theoretic equilibria, whereas existing work only considers convergence to notions of stationarity. Despite this, we are able to obtain analogous results for the stricter notion of solving the problem.

In the class of nonconvex-PL zero-sum games, the closest works to this paper are Nouiehed et al. (2019) and Yang et al. (2020). Each of the aforementioned works show that an \( \epsilon \)-approximate stationary point of the game function \( f(\cdot, \cdot) \) can be found in \( O(\epsilon^{-2}) \) gradient calls by variants of alternating gradient descent-ascent with timescale separation (\( \tau \)-AGDA) in the deterministic setting. In comparison, we show in Theorem 2 that simultaneous gradient descent-ascent with timescale separation (\( \tau \)-GDA) globally asymptotically converges to a differential Stackelberg equilibrium (local minmax equilibrium) almost surely. Moreover, we provide a local convergence guarantee to an \( \epsilon \)-differential Stackelberg equilibrium in \( O(\epsilon^{-2}) \) gradient calls in the deterministic setting. Thus, locally the convergence rate we provide is analogous to past works, but the guarantee is to a game-theoretic equilibrium instead of only a stationary point. We note that the reason the guarantee is only local is that even with asymptotic saddle avoidance, there is no guarantee on the rate at which \( \tau \)-GDA escapes saddles of the dynamics. We remark that the primary focus of Yang et al. (2020) is on two-sided PL games, which is a class with a unique equilibrium and the convergence guarantees given for this class are to the equilibrium.

In the class of nonconvex-SC zero-sum games, there is an extensive amount of recent work. In terms of existing results on variants of gradient descent-ascent, Lin et al. (2020a) shows for both simultaneous and alternating gradient descent-ascent (\( \tau \)-GDA and \( \tau \)-AGDA) that with \( O(\epsilon^{-2}) \) and \( O(\epsilon^{-4}) \) gradient calls an \( \epsilon \)-approximate stationary point of the function \( f(\cdot, \cdot) \) or the function \( \phi(\cdot) = \max_{y \in \mathcal{Y}} f(\cdot, y) \) can be obtained in the deterministic and stochastic settings, respectively. For this class of games, we show (Theorem 3) the stronger result that with equivalent complexity in terms of \( \epsilon \) and the dimension up to polylogarithmic factors, a perturbed variant of \( \tau \)-GDA called \( \tau \)-PGDA can reach an \( \epsilon \)-approximate differential Stackelberg equilibrium (strict local minmax equilibrium). Obtaining this result requires both fully characterizing the local stability around critical points (Theorem 1) and generalizing methods for efficiently escaping saddle points from nonconvex optimization to zero-sum games.

Before moving on, we remark that there is a number of works that develop algorithms that improve the complexity of finding stationary points in terms of the dependence on the condition number and polylogarithmic dependencies. This focus is separate from our work, but we refer the reader to Lin et al. (2020b), Lu et al. 2020, Rafique et al. 2021 and the references therein. We believe an interesting direction of future work is strengthening the results along this direction to the stronger notions of solving the problem considered in this work. Furthermore, there is also recent work developing lower bounds for this class of games (Zhang et al. 2021). Finally, there are several works in this class of games with zeroth order feedback (Lin et al. 2020, Wang et al. 2020b, Xu et al. 2020).

**Escaping Saddle Points in Nonconvex Optimization.** In this paper we present asymptotic and finite-time convergence results to differential Stackelberg equilibrium in nonconvex-PL and nonconvex-SC zero-sum games. To obtain such results it is necessary to escape saddle points either asymptotically or in finite-time. We build on methods from nonconvex optimization to prove that the algorithms we analyze do this.

A common assumption in the body of work on escaping saddle points in nonconvex optimization is what is known as the strict saddle property (see, e.g., Ge et al. 2015, Jin et al. 2017, Lee et al. 2016). Informally, if a function class satisfies the strict saddle property, every saddle point of the gradient descent dynamics has a strictly negative eigenvalue in the Jacobian evaluated at the critical point. This assumption ensures that there is a direction of escape from every saddle point. We make an analogous assumption for the classes of nonconvex zero-sum games we analyze (Assumption 3). For nonconvex-PL/SC zero-sum games, if the game satisfies the strict saddle property, every saddle point of the \( \tau \)-GDA dynamics has an eigenvalue with strictly
negative real part in the Jacobian evaluated at the critical point for the given choice of timescale separation $\tau$.

There exists both asymptotic and finite-time saddle avoidance results in nonconvex optimization. The asymptotic saddle avoidance results in nonconvex optimization state that gradient descent dynamics almost surely avoid strict saddle points \cite{Lee16,Lee19}. This result has been extended to show that gradient descent dynamics almost surely avoid saddle points even for functions with non-isolated critical points \cite{Panageas17} in nonconvex optimization. The drawback of this style of result is that it fails to preclude that gradient descent could spend an arbitrarily long time stuck in the neighborhood of a saddle point. In fact, it has been shown that gradient descent can take exponential time to escape from saddle points \cite{Du17}.

A series of works present asymptotic results on escaping saddle points in continuous action games analogous to that from nonconvex optimization. In particular, \cite{Mazumdar20} prove that in $N$-player general-sum games, if each player follows the gradient descent learning rule then the dynamics avoid strict saddles of the dynamics almost surely. This result allows for players to employ distinct learning rates. Translated to nonconvex-nonconcave zero-sum games, the result ensures that $\tau$-GDA dynamics avoid strict saddles of the dynamics almost surely for any timescale separation $\tau > 0$ as is presented in Lemma 2. The aforementioned result is key to the asymptotic convergence guarantee to differential Stackelberg equilibria we provide in Theorem 3 for $\tau$-GDA for nonconvex-PL/SC zero-sum games. In a related line of work, \cite{Daskalakis18} show that in nonconvex-nonconcave zero-sum games the $\tau$-GDA dynamics with $\tau = 1$ (without timescale separation) avoid strict saddle points including when critical points are not isolated. We remark that the result demonstrating that gradient descent can take exponential time to escape from saddle points \cite{Du17} immediately carries over to gradient descent-ascent in zero-sum games since the game could be completely decoupled an correspond to separate nonconvex-optimization problems.

The study of finite-time saddle avoidance results in nonconvex optimization generally focuses on designing variants of gradient descent that not only avoid saddle points of the dynamics almost surely, but escape from them efficiently. The methods of greatest relevance to this paper are the existing works analyzing the rates at which ‘simple’ variants of gradient descent escape saddle points in nonconvex optimization. This line of research dates back to the work of \cite{Ge15}, who proved that stochastic gradient descent with injected noise perturbations finds an approximate local minimum in a number of gradient class that depends polynomially on the dimension. Since the initial work on escaping saddle points efficiently, there have been many follow-up works. The most relevant to our results and techniques are the works of \cite{Jin17} and \cite{Jin21}. Indeed, \cite{Jin17} prove that a perturbed variant of gradient descent with deterministic gradients finds approximate local minima under the strict saddle assumption with only a logarithmic dimension dependence, meaning the result is almost dimension free and near equivalent to the complexity needed to find a critical point using gradient descent. Moreover, \cite{Jin21} generalize such proof techniques to show that variants of gradient and stochastic gradient descent with injected noise achieve an analogous guarantee with a unified analysis. We generalize this analysis method in our study of nonconvex-SC zero-sum games in Section 6 and Appendix E to obtain Theorem 3.

We conclude this section by mentioning that there is an extensive literature on methods for escaping saddle points efficiently in nonconvex optimization beyond the works that have been mentioned thus far. This includes analysis of dynamics using normalized gradients \cite{Levy16}, stochastic gradient descent without artificial noise injections \cite{Daneshmand18}, negative curvature search methods \cite{Agarwal17,AI18,Xu18}, variance reduced methods \cite{Fang18,Zhou18}, cubic regularization \cite{Tripraneni18}, and adaptive acceleration methods \cite{Carmon18,Jin18,Staib19,Wang19}. Finally, there is also a recent, sharper analysis of stochastic gradient descent \cite{Fang19}.

\section{Preliminaries}

We now review preliminaries on game-theory concepts, dynamical systems theory, PL functions and nonconvex-PL zero-sum games, linear algebra, and concentration inequalities. The sections on PL functions and nonconvex-PL zero-sum games, linear algebra, and concentration inequalities include a number of technical lemmas that are needed throughout the rest of the appendix.
B.1 Game Theory

The focus of this work is on developing convergence guarantees for gradient descent-ascent with timescale separation ($\tau$-GDA and $\tau$-PGDA) to game-theoretically meaningful equilibrium. In the class of games under consideration, the natural solution concept from game theory is a minmax or equivalently Stackelberg equilibrium. Given the nonconvex nature of the minimizing players objective, we consider a local refinement of this historically standard equilibrium concept. In the nonconvex-nonconcave zero-sum game literature, recent work [Fiez et al. 2020; Jin et al. 2020] develops sufficient conditions for local Stackelberg equilibrium and points satisfying the conditions correspond to strict local Stackelberg equilibrium. We refer to strict local Stackelberg equilibrium as differential Stackelberg equilibrium following the terminology of [Fiez et al. 2020]. This is the the basis for the equilibrium concept we consider that is given in Definition 4. The fact we focus our attention on strict local equilibrium is standard (Fiez et al. 2020; Jin et al. 2020; Mazumdar et al. 2020; Wang et al. 2020a).

For completeness, we now present a formal local Stackelberg equilibrium definition characterized in terms of the costs. This definition has appeared in past work (Fiez and Ratliff 2021; Fiez et al. 2020) and it is a simple refinement to a local concept from the standard Stackelberg equilibrium definition in continuous action space games (Başar and Olsder 1998). We emphasis that the equilibrium we consider in Definition 4 is characterized by sufficient conditions for the following local Stackelberg equilibrium definition.

**Definition 8 (Local Minmax/Stackelberg Equilibrium).** Consider $U_x \subset X$ and $U_y \subset Y$ where, without loss of generality, player 1 (controlling $x \in X$) is minimizing player and player 2 (controlling $y \in Y$) is the maximizing player. The strategy $x^* \in U_x$ is a local Stackelberg solution for the minimizing player if, $\forall x \in U_x$,

$$
\sup_{y \in r_{U_y}(x^*)} f(x^*, y) \leq \sup_{y \in r_{U_y}(x)} f(x, y),
$$

where $r_{U_y}(x) = \{ y' \in U_y | f(x, y') \geq f(x, y), \forall y \in U_y \}$ is the reaction curve. Moreover, for any $y^* \in r_{U_y}(x^*)$, the joint strategy profile $(x^*, y^*) \in U_x \times U_y$ is a local Stackelberg equilibrium on $U_x \times U_y$.

B.2 Dynamical Systems Theory

In this section, we provide relevant background as it pertains to the local stability analysis we presented in Section 4.

We begin by formally describing what is meant by local stability of a continuous-time dynamical system as used in this paper. In particular, we provide a formal definition for the stability definition that was presented in Definition 5. The following well-known result provides equivalent characterizations of local stability for a critical point (stationary point) of a continuous-time dynamical system of the form $\dot{z} = -g(z)$ in terms of the Jacobian matrix $J(z) = Dg(z)$.

**Theorem 4 ([Khalil 2002, Theorem 4.15]).** Consider a critical point $z^*$ of $g(z)$. The following are equivalent:

(a) $z^*$ is a locally exponentially stable equilibrium of $\dot{z} = -g(z)$; (b) $\text{spec}(-J(z^*)) \subset \mathbb{C}_{\geq 0}$; (c) there exists a symmetric positive-definite matrix $P = P^T > 0$ such that $PJ(z^*) + J(z^*)^TP > 0$.

Before moving on, we provide a brief discussion of the implications of determining stability using the local linearization (Jacobian of the dynamical system) around critical points. The Hartman-Grobman theorem [Sastry 1999, Theorem 7.3]; [Teschl 2000, Theorem 9.9] asserts that it is possible to continuously deform all trajectories of a nonlinear system onto trajectories of the linearization at a fixed point of the nonlinear system. Informally, the theorem states that the qualitative properties of the nonlinear system $\dot{z} = -g(z)$ in the vicinity (which is determined by the neighborhood $U$) of an isolated equilibrium $z^*$ are determined by its linearization if the linearization has no eigenvalues on the imaginary axes in the complex plane. We also remark that Hartman-Grobman can also be applied to discrete time maps (cf. Sastry 1999, Thm. 2.18) with the same qualitative outcome.

In the context of this work, this means that by determining stability using the local linearization around critical points, the behavior of the nonlinear system in a neighborhood of the critical point can be inferred. In particular, given that a critical point is determined to be stable by the local linearization, then there is a neighborhood on which the dynamics converge to the critical point. This observation also applies to the discrete-time system with proper learning rates.
B.3 Polyak-Łojasiewicz Functions and Nonconvex-Polyak-Łojasiewicz Zero-Sum Games

In this section, we state properties of PL functions in the context of nonconvex-PL zero-sum games. Specifically, we characterize the curvature around critical points of PL functions and also state Lipschitz and smoothness properties that follow from our assumptions in the paper. These properties will be used in both the proofs for the local stability and the global convergence in nonconvex-PL zero-sum games.

We begin by stating a known property (Karimi et al., 2016) that μ-PL functions satisfy a quadratic growth condition also with parameter μ. For clarity of presentation, we present this condition in the context of the nonconvex-PL zero-sum games we study.

Lemma 4 (Karimi et al., 2016). Consider a non-convex, μ-PL zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \). For all \( x \in \mathcal{X} \), the function \( f(x, \cdot) \) satisfies the following quadratic growth condition:

\[
\max_{y' \in \mathcal{Y}} f(x, y') - f(x, y) \geq \frac{\mu}{2} \| y' - y \|^2, \quad \forall \, y \in \mathcal{Y}
\]

where \( y_p \) is the projection onto the set \( \arg\max_{y \in \mathcal{Y}} f(x, y) \).

We now show that the quadratic growth property of PL functions implies that \( \nabla_2^2 f(x^*, y^*) \) is negative definite and eigenvalues bounded above by \(-\mu/2\).

Proof. Let us consider any critical point \( (x^*, y^*) \) of the game so that \( \nabla_1 f(x^*, y^*) = 0 \) and \( \nabla_2 f(x^*, y^*) = 0 \). Taking a Taylor expansion of \( f(x^*, \cdot) \) about the point \( y^* \), we and get that

\[
f(x^*, y) \geq f(x^*, y^*) + \nabla_2 f(x^*, y^*)^\top (y - y^*) + \frac{1}{2} (y - y^*)^\top \nabla_2^2 f(x^*, y^*)(y - y^*)
\]

Hence, from the quadratic growth condition of Lemma 4, we have that

\[
\frac{\mu}{2} \| y^* - y \|^2 \leq f(x^*, y^*) - f(x^*, y) \leq -(y - y^*)^\top \nabla_2^2 f(x^*, y^*)(y - y^*)
\]

and consequently

\[
\frac{\mu}{2} \| y^* - y \|^2 I \leq -\nabla_2^2 f(x^*, y^*)\| y^* - y \|^2 \implies \nabla_2^2 f(x^*, y^*) \leq -\frac{\mu}{2}.
\]

Since this holds for any critical point, the conclusion follows. \(\square\)

We now state several more properties that will be used in most of the proofs in Appendix D regarding the results for nonconvex-PL zero-sum games from Section 5.

Danskin’s theorem in optimization provides conditions under which \( \nabla f(x, y_\star(x)) \) where \( y_\star(x) = \arg\max_{y \in \mathcal{Y}} f(x, y) \) is equivalent to \( \nabla_1 f(x, y_\star) \). That is, it gives conditions when the gradient of the function \( f(x, y_\star(x)) \) is equal to the gradient of \( f(x, y_\star) \) evaluated directly at the optimum. Typically this requires the maximizer to be unique. However, it has been shown that for nonconvex-PL zero-sum games, this property carries over even without a unique solution.

Lemma 6 (Danskin-Type Property for PL functions (Nouiehed et al., 2019, Lemma A.5)). Consider a non-convex, μ-PL zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) satisfying Assumption 7. Then,

\[
\nabla f(x, y_\star(x)) = \nabla_1 f(x, y_\star(x)) \quad \text{where} \quad y_\star(x) = \arg\max_{y \in \mathcal{Y}} f(x, y).
\]
We now show that the mapping \( y^*(x) \), defined implicitly by \( \nabla f^2(x,y) = 0 \) is Lipschitz in nonconvex-PL zero-sum games.

**Lemma 7.** Consider a non-convex, \( \mu \)-PL zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) satisfying Assumptions 1 and 2. The best-response map \( y^*(x) \in \text{argmax}_y f(x,y) \), defined implicitly by \( \nabla f^2(x,y) = 0 \) is \( L_3 = \beta \mu \)-Lipschitz. That is, for all \( x, x' \in X \),

\[
\|y^*(x) - y^*(x')\| \leq L_3 \|x - x'\|.
\]

**Proof.** For all \( x, x' \in X \), we have that

\[
\|y^*(x) - y^*(x')\| \leq \max_\nu \|Dy^*(\nu)\| \|x - x'\|
\leq \max_\nu \|\left(\nabla^2 f(\nu, y^*(\nu))\right)^{-1} \|\nabla y^* f(\nu, y^*(\nu))\| \|x - x'\|
\leq \frac{\beta}{\mu} \|x - x'\| = L_3 \|x - x'\|.
\]

Finally, we assume that total derivative \( \nabla f(x,y) \) is Lipschitz in nonconvex-PL zero-sum games.

**Assumption 7.** The total derivative \( \nabla f(x,y) = \nabla f_1(x,y) - \nabla f_2(x,y) \) is \( L_5 \)-Lipschitz. That is, for any \( x \in X \) and all \( y, y' \in Y \),

\[
\|\nabla f(x,y) - \nabla f(x,y')\| \leq L_5 \|y - y'\|.
\]

Note that if \( f \) is Lipschitz in \( y \), then the above assumption follows directly from the proceeding assumptions on the smoothness of \( f \).

### B.4 Linear Algebra

In this section, we state linear algebra properties that are needed for the proofs presented in later sections. Specifically, in Appendix B.4.1 we state a property regarding matrix inertia that is important for the proof of Theorem 1 in Appendix C. Moreover, in Appendix B.4.2 we recall properties of matrix norms and state properties of a particular matrix norm that is important for the several immediate results toward proving Theorem 3 in Appendix E.2.

#### B.4.1 Matrix Inertia

The following result from [Lancaster and Tismenetsky][1985] (Theorem 2, Chapter 13.1) is needed for the proof of Theorem 1 given in Appendix C. We include it here for ease of reference. For a given matrix \( A \in \mathbb{R}^{n \times n} \), \( v_+(A) \), \( v_-(A) \), and \( \zeta(A) \) are the number of eigenvalues of the argument that have positive, negative and zero real parts, respectively.

**Lemma 8** ([Lancaster and Tismenetsky][1985], Theorem 2, Chapter 13.1). Consider a matrix \( A \in \mathbb{R}^{n \times n} \).

(a) If \( P \) is a symmetric matrix such that \( AP + PA^T = Q \) where \( Q = Q^T > 0 \), then \( P \) is nonsingular and \( P \) and \( A \) have the same inertia, meaning that

\[
v_+(A) = v_+(P), \quad v_-(A) = v_-(P), \quad \zeta(A) = \zeta(P).
\]  

(b) On the other hand, if \( \zeta(A) = 0 \), then there exists a matrix \( P = P^T \) and a matrix \( Q = Q^T > 0 \) such that \( AP + PA^T = Q \) and \( P \) and \( A \) have the same inertia so that (7) holds.
We now provide a series of properties for a particular matrix norm that will be key to the proof of Lemma 26 in Appendix E.2. The majority of the following material can be found in [Horn and Johnson (2012)]. In particular, the following result is an exercise in [Horn and Johnson (2012)]. We now show that given a matrix for which every eigenvalue of maximum modulus is semisimple, there exists an induced matrix norm that is equal to the spectral radius. Let \( \rho(\cdot) \) denote the spectral radius of its argument.

**Lemma 9** (Horn and Johnson (2012)). Consider a matrix \( A \in \mathbb{C}^{n \times n} \) and suppose that every eigenvalue of \( A \) of maximum modulus is semisimple. There exists a matrix norm \( \| \cdot \| \) such that \( \rho(A) = \| A \| \). In particular, the following result is an exercise in [Horn and Johnson (2012)]. We now show that given a matrix for which every eigenvalue of maximum modulus is semisimple, there exists an induced matrix norm that is equal to the spectral radius. Let \( \rho(\cdot) \) denote the spectral radius of its argument.

**Proof.** Let \( \lambda_1 \) denote the eigenvalue of \( A \) with maximum modulus and \( \{ \lambda_2, \ldots, \lambda_k \} \) denote all other eigenvalues. Note that \( \lambda_1 \) is semisimple so that its geometric and algebraic multiplicities are the same. Let \( m_1 \) denote the algebraic (and geometric) multiplicity of \( \lambda_1 \); in particular, \( m_1 \) is the number of Jordan blocks associated with \( \lambda_1 \). Let \( J \) be the Jordan normal form of \( A \); that is,

\[
J = \begin{bmatrix}
\Lambda & 0 & \cdots & 0 \\
0 & J_2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & J_k
\end{bmatrix} \in \mathbb{C}^{n \times n},
\]

\[
J_i = \begin{bmatrix}
\lambda_i & 0 & \cdots & 0 \\
0 & \lambda_i & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_i
\end{bmatrix} \in \mathbb{C}^{n_i \times n_i},
\]

with \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_1) \in \mathbb{C}^{m_1 \times m_1} \) and \( n_i \) denoting the size of the Jordan block associated with \( \lambda_i \). The Jordan normal form of the matrix \( A \) is obtained by a similarity transform

\[ A = SJS^{-1} \]

where \( S \in \mathbb{C}^{n \times n} \) is an invertible matrix.

Now, for an arbitrary \( \epsilon > 0 \), let

\[
W(\epsilon) = \begin{bmatrix}
W_{m_1}(\epsilon) & 0 & \cdots & 0 \\
0 & W_{n_2}(\epsilon) & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & W_{n_k}(\epsilon)
\end{bmatrix} \in \mathbb{R}^{n \times n},
\]

where

\[
W_{n_i}(\epsilon) = \begin{bmatrix}
\epsilon & 0 & \cdots & 0 \\
0 & \epsilon^2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \epsilon^{n_i}
\end{bmatrix} \in \mathbb{R}^{n_i \times n_i},
\]

and \( W_{m_1}(\epsilon) = \text{diag}(\epsilon, \ldots, \epsilon) \in \mathbb{R}^{m_1 \times m_1} \). Observe that

\[
W(\epsilon)^{-1}S^{-1}ASW(\epsilon) = \begin{bmatrix}
\Lambda & 0 & \cdots & 0 \\
\Lambda & 0 & \cdots & 0 \\
0 & B_{n_2}(\lambda_2,\epsilon) & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & B_{n_k}(\lambda_k,\epsilon)
\end{bmatrix},
\]

where \( B_{n_i}(\lambda_i,\epsilon) = \begin{bmatrix}
\lambda_i & \epsilon & 0 & \cdots & 0 \\
0 & \lambda_i & \epsilon & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \lambda_i & \epsilon
\end{bmatrix} \).
Since the eigenvalues of $A$ are on the diagonal of $W(\varepsilon)^{-1}S^{-1}ASW(\varepsilon)$ and the superdiagonal of each of the $B_{n_i}(\lambda_i,\varepsilon)$ contains $\varepsilon$, there exists a sufficiently small $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0,\bar{\varepsilon}]$, $\rho(A) = \|W(\varepsilon)^{-1}S^{-1}ASW(\varepsilon)\|$ by the semisimple assumption where $\rho(\cdot)$ denotes the spectral radius and $\| \cdot \|$ denotes the induced 2-norm as noted above. Define $M = W(\varepsilon)^{-1}S^{-1}$ for the largest $\varepsilon \in (0,\bar{\varepsilon})$ for which this holds. We can thus define the matrix norm
\[ \|A\|_* = \|MAM^{-1}\| \]
with the property that $\|A\|_* = \rho(A)$. This is a matrix norm since $\|\cdot\|$ is a matrix norm and $M$ is non-singular since the product of non-singular matrices is non-singular if and only if each matrix in the product is non-singular and these properties are sufficient to guarantee $\|\cdot\|_*$ is a matrix norm \cite{Horn2012}. Moreover, since $\|\cdot\|$ is induced by the vector norm $\|\cdot\|$ on $\mathbb{C}^n$, the matrix norm $\|\cdot\|_*$ is induced by the vector norm $\|\cdot\|_*$ on $\mathbb{C}^n$. That is, for $x \in \mathbb{C}^n$, $\|x\|_* = \|Mx\|$. \qed

We now show properties of the matrix norm constructed in Lemma 9 that are immediate for induced matrix norms.

**Lemma 10.** The induced matrix norm constructed in Lemma 9 denoted by $\|\cdot\|_*$ is subordinate to the vector norm $\|\cdot\|$ that induces it. That is, for any $A \in \mathbb{C}^{n \times n}$, $x \in \mathbb{C}^n$, $\|Ax\|_* \leq \|A\|_* \|x\|_*$.

**Lemma 11.** The induced matrix norm constructed in Lemma 9 denoted by $\|\cdot\|_*$ is submultiplicative. That is, for any $A,B \in \mathbb{C}^{n \times n}$, $\|AB\|_* \leq \|A\|_* \|B\|_*$.

Finally, we show a simple property of the 2-norm that is useful for converting back and forth between the norm construction in the previous results.

**Lemma 12.** Consider a matrix $A \in \mathbb{C}^{n \times n}$ and a vector $x \in \mathbb{C}^n$. Let $\|\cdot\|$ denote the $\ell_2$ norm and denote the minimum and maximum singular values of $A$ by $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$, respectively. Then, $\sigma_{\min}(A) \leq \|Ax\|$ and $\|Ax\| \leq \sigma_{\max}(A)\|x\|$.

### B.5 Concentration Inequalities

In this section, we present concentration inequalities for norm-subGaussian random vectors. Each of the following technical lemmas are from \cite{Jin2021} and we reproduce them here for clarity of presentation and easy reference. These concentration inequalities will be used throughout the finite-time analysis of nonconvex-SC zero-sum games presented in Appendix E to obtain the results from Section 6.

The following defines a norm-subGaussian random vector.

**Definition 9.** A random vector $x \in \mathbb{R}^d$ is norm-subGaussian if there exists $\sigma$ so that:
\[ P(\|x - E[x]\| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}} \quad \forall t \in \mathbb{R}. \]

The next result shows that a bounded random vector and a subGaussian random vector are special cases of a norm-subGaussian random vector.

**Lemma 13.** There exists an absolute constant $c$ so that the following random vectors are $\sigma$-norm-subGaussian:

1. A bounded random vector $x \in \mathbb{R}^d$ such that $\|x\| \leq \sigma$.
2. A random vector $x \in \mathbb{R}^d$, where $x = \psi e_1$ and the random variable $\psi \in \mathbb{R}$ is $\sigma$-subGaussian.
3. A random vector $x \in \mathbb{R}^d$ that is $(\sigma/\sqrt{3})$-subGaussian.

The following says that if a random vector $x$ is norm-subGaussian, then its norm square is subExponential and its component along a single direction is subGaussian.
Lemma 14. There is an absolute constant $c$ such that if the random vector $x \in \mathbb{R}^d$ is zero-mean $\sigma$ norm-subGaussian, then $\|x\|^2$ is $c\sigma^2$-subExponential and for any fixed unit vector $v \in \mathbb{S}^{d-1}$ the quantity $\langle v, x \rangle$ is $c\sigma$-subGaussian.

We now define the properties of norm-subGaussian martingale difference sequences.

Condition 1. Consider random vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ and the corresponding filtrations $\mathcal{F}_i = \sigma(x_1, \ldots, x_i)$ for $i \in [n]$ such that $x_i | \mathcal{F}_{i-1}$ is zero-mean $\sigma_i$-norm-subGaussian with $\sigma_i \in \mathcal{F}_{i-1}$. That is:

$$E[X_i | \mathcal{F}_{i-1}] = 0, \quad P(\|x_i\| \geq t | \mathcal{F}_{i-1}) \leq 2e^{-\frac{t^2}{\sigma_i^2}}, \quad \forall t \in \mathbb{R}, \forall i \in [n].$$

The following result gives a Hoeffding-type inequality for norm-subGaussian random vectors.

Lemma 15. Given random vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ that satisfy condition 1 with fixed $\{\sigma_i\}$, then for any $\iota > 0$ there exists an absolute constant $c$ such that with probability at least $1 - 2de^{-\iota}$:

$$\| \sum_{i=1}^n x_i \| \leq c \sqrt{\sum_{i=1}^n \sigma_i^2}. $$

When the sequence $\{\sigma_i\}$ is also random, then the following holds.

Lemma 16. Given random vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ that satisfy condition 1, then for any $\iota > 0$ and $B > b > 0$ there exists an absolute constant $c$ such that with probability at least $1 - 2d\log(B/b)e^{-\iota}$:

$$\sum_{i=1}^n \sigma_i^2 \geq B \text{ or } \| \sum_{i=1}^n x_i \| \leq c \max \left\{ \sum_{i=1}^n \sigma_i^2, b \right\} \iota. $$

The next results give concentration inequalities for the sum of norm squares of norm-subGaussian random vectors and the sum of inner products of norm-subGaussian random vectors with another set of random vectors.

Lemma 17. Given random vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ that satisfy condition 1 with fixed $\sigma_1 = \cdots = \sigma_n = \sigma$, then for any $\iota > 0$ there exists an absolute constant $c$ such that with probability at least $1 - e^{-\iota}$:

$$\| \sum_{i=1}^n x_i \|^2 \leq c\sigma^2(n + \iota). $$

Lemma 18. Given random vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ that satisfy condition 1 and random vectors $\{u_i\}$ that satisfy $u_i \in \mathcal{F}_{i-1}$ for all $i \in [n]$, then for any $\iota > 0$ and $\lambda > 0$ there exists an absolute constant $c$ such that with probability at least $1 - e^{-\iota}$:

$$\sum_{i=1}^n \langle u_i, x_i \rangle \leq c\lambda \sum_{i=1}^n \|u_i\|^2 \sigma_i^2 + \frac{1}{\lambda} \iota. $$

C Stability Analysis: Proof of Theorem 1

This appendix is devoted to providing the proof of Theorem 1. For ease of reference, we restate the result now before providing the proof.

Theorem 1. Consider a nonconvex-PL zero-sum game $(f, -f)$ where $f \in C^2(\mathbb{Z}, \mathbb{R})$. Then, the following hold: 1) Any critical point $z^*$ that is not a differential Stackelberg equilibrium is unstable for all $\tau \in (0, \infty)$; 2) If $z^*$ is a differential Stackelberg equilibrium, then $\spec(-J_{\tau}(z^*)) \subset C_\tau^\circ$ for all $\tau \in [\tau_*, \infty)$ where $\tau_*$ is the minimum $\tau \in (0, \infty)$ such that $\spec(-J_{\tau}(z^*)) \subset C_\tau^\circ$ and a finite $\tau_*$ is guaranteed to exist.

Proof of Theorem 1. We begin by proving the first claim of the theorem statement.
Proof of 1. Let us first consider the case that a given critical point $z^*$ could be such that $S_1(J_1(z^*))$ or $-\nabla^2_{2} f(z^*)$ are singular. By Lemma 4 we know that for any critical point $z^*$, $-\nabla^2_{2} f(z^*) > 0$ so that $-\nabla^2_{2} f(z^*)$ is non-singular. Observe that at any critical point $z^*$, since $-\tau \nabla^2_{2} f(z^*)$ is positive definite for all $\tau \in (0, \infty)$, the following identity holds for any $\tau \in (0, \infty)$:
\[
\det(J_\tau(z)) = \det(S_1(J(z^*))) \det(-\tau \nabla^2_{2} f(z^*)).
\]
From the fact that $\det(-\tau \nabla^2_{2} f(z^*))$, it then easily follows that $\det(J_\tau(z^*)) = 0$ if and only if $\det(S_1(J(z^*))) = 0$. Note that $\det(J_\tau(z^*)) = 0$ if and only if $0 \in \text{spec}(J_\tau(z^*))$ since eigenvalues of a real square matrix are either purely real or come in complex conjugate pairs. Hence, given any critical point such that $\det(S_1(J(z^*))) = 0$, then $0 \in \text{spec}(J_\tau(z^*))$ and $\text{spec}(J_\tau(z^*)) \not\subseteq \mathbb{C}^\infty$ for all $\tau \in (0, \infty)$. Hence, any critical point such that $S_1(J_1(z^*))$ is singular is unstable for all $\tau \in (0, \infty)$ and such a point is not a differential Stackelberg equilibrium.

Now, suppose that $z^*$ is a critical point such that $S_1(J_1(z^*))$ and $-\nabla^2_{2} f(z^*)$ are non-singular. Let $\text{spec}(-J_{\tau_0}(z^*)) \subset \mathbb{C}^\infty$ for some $\tau_0 \in (0, \infty)$. We know that $-\nabla^2_{2} f(z^*) > 0$ by Lemma 5. We argue by contradiction that $S_1(J_1(z^*)) > 0$. Towards this end, suppose not.

Since $\det(S_1(J_1(z^*))) \neq 0$ and $\det(-\nabla^2_{2} f(z^*)) \neq 0$, by Lemma 8b, there exists non-singular Hermitian matrices $P_1, P_2$ and positive definite Hermitian matrices $Q_1, Q_2$ such that $-S_1(J_1(z^*))P_1 - P_1 S_1(J_1(z^*)) = Q_1$ and $-\nabla^2_{2} f(z^*)P_2 + P_2 -\nabla^2_{2} f(z^*) = Q_2$.

Furthermore, $-S_1(J_1(z^*))$ and $P_1$ have the same inertia, meaning
\[
v_+(-S_1(J_1(z^*))) = v_+(P_1), \quad v_-(-S_1(J_1(z^*))) = v_-(P_1), \quad \zeta(-S_1(J_1(z^*))) = \zeta(P_1)
\]
where for a given matrix $A$, $v_+(A)$, $v_-(A)$, and $\zeta(A)$ are the number of eigenvalues of the argument that have positive, negative and zero real parts, respectively. Similarly, $-\nabla^2_{2} f(z^*)$ and $P_2$ have the same inertia:
\[
v_+(-\nabla^2_{2} f(z^*)) = v_+(P_2), \quad v_-(-\nabla^2_{2} f(z^*)) = v_-(P_2), \quad \zeta(-\nabla^2_{2} f(z^*)) = \zeta(P_2).
\]

Since $-S_1(J_1(z^*))$ has at least one strictly positive eigenvalue, $v_+(P_1) = v_+(-S_1(J_1(z^*)) \geq 1$.

Define
\[
P = \begin{bmatrix} I & L_0^T \\ 0 & I \end{bmatrix} \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ L_0 & I \end{bmatrix}
\]
with
\[
Q_{\tau_0} = \begin{bmatrix} I & L_0^T \\ 0 & I \end{bmatrix} B_{\tau_0} \begin{bmatrix} I & 0 \\ L_0 & I \end{bmatrix}
\]
and
\[
B_{\tau_0} = \begin{bmatrix} Q_1 \\ (P_1 \nabla_{12} f(z^*) - S_1(J_1(z^*))L_0^T P_2) \nabla_{12} f(z^*) + (P_2 L_0 \nabla_{12} f(z^*))^T + \tau_0 Q_2 \end{bmatrix}
\]
which can be verified by straightforward calculations. The matrix $B_{\tau_0}$ is a symmetric matrix, and it is positive definite. Indeed, first observe that $Q_1 > 0$ and $Q_2 > 0$. Then showing $B_{\tau_0} > 0$ reduces to showing $P_2 L_0 \nabla_{12} f(z^*) + (P_2 L_0 \nabla_{12} f(z^*))^T \geq 0$, which is the case because $\nabla^2_{2} f(z^*) < 0$ implies that $P_2 < 0$ so that
\[
P_2(\nabla^2_{2} f(z^*))^{-1} \nabla_{12} f(z^*) \nabla_{12} f(z^*) + (P_2(\nabla^2_{2} f(z^*))^{-1} \nabla_{12} f(z^*) \nabla_{12} f(z^*))^T \geq 0
\]
since the produce of negative definite matrices is positive definite and so is the product of positive definite matrices. Now, since $B_{\tau_0} > 0$ so is $Q_{\tau_0}$ since they are congruent. Since $\text{spec}(-J_{\tau}(z^*)) \subset \mathbb{C}^\infty$, $Q_{\tau_0} > 0$ implies that $P = P^T < 0$ (by Lyapunov’s theorem). Hence, $P_1$ and $P_2$ must be negative definite since $P$ is congruent to $\text{diag}(P_1, P_2)$, but this gives us a contradiction with the fact that $P_1$ has the same inertia as $-S_1(J_1(z^*))$ which we assumed to have at least one positive eigenvalue. Hence, if $\text{spec}(-J_{\tau_0}(z^*)) \subset \mathbb{C}^\infty$, then it must be the case that $S_1(J_1(z^*)) > 0$ which means $z^*$ is a differential Stackelberg equilibrium since we also have $-\nabla^2_{2} f(z^*) > 0$ by Lemma 5.
Thus, we can finish the proof of part 1 as follows. Consider any critical point \( \hat{z} \) that is not a differential Stackelberg equilibrium and is unstable for the nominal \( \tau_0 \). Then, we claim that \( \text{spec}(-J_{\tau}(\hat{z})) \not\subset \mathbb{C}_{\omega} \) for all \( \tau \geq \tau_0 \). Suppose not. That is, there is some \( \tau_1 \geq \tau_0 \) such that \( \text{spec}(-J_{\tau_1}(\hat{z})) \subset \mathbb{C}_{\omega} \). But by our argument above, since \( -\nabla_{2}^2 f(\hat{z}) > 0 \), this implies that \( S_1(J_{\tau_1}(\hat{z})) > 0 \) which contradicts that \( \hat{z} \) is not a differential Stackelberg equilibrium. Hence, any critical point \( z^* \) that is not a differential Stackelberg equilibrium is unstable for all \( \tau \in (0, \infty) \).

**Proof of 2.** We note that the fact there exists a finite \( \tau^* \in (0, \infty) \) such that a differential Stackelberg is stable is known (Fiez and Ratliff [2021]). Let \( \tau^* \) denote the minimum \( \tau^* \) such that a differential Stackelberg equilibrium \( z^* \) is stable. To see that \( \text{spec}(-J_{\tau}(z^*)) \subset \mathbb{C}_{\omega} \) for all \( \tau \geq \tau^* \) given that \( \text{spec}(-J_{\tau^*}(z^*)) \subset \mathbb{C}_{\omega} \), we can again examine the Lyapunov equation under the congruent transformation. We define the matrix \( P \) as above in equation (10) where \( P_1 < 0 \) and \( P_2 < 0 \) so that \( -S(J_1(z^*)) < 0 \) and \( \nabla_{2}^2 f(z^*) < 0 \) with \( Q_1, Q_2 > 0 \). With 

\[
Q_\tau = \begin{bmatrix} I & L_0^\top \\ 0 & I \end{bmatrix} B_\tau \begin{bmatrix} I & 0 \\ L_0 & I \end{bmatrix}
\]

and 

\[
B_\tau = \begin{bmatrix} Q_1 & (P_1 \nabla_{12} f(z^*) - S(J_1(z^*)))L_0^\top P_2 \\ (P_1 \nabla_{12} f(z^*) - S(J_1(z^*)))L_0^\top P_2 & P_2 L_0 \nabla_{12} f(z^*) + (P_2 L_0 \nabla_{12} f(z^*))^\top + \tau Q_2 \end{bmatrix}
\]

we again can see that \( Q_\tau > 0 \) for the same reason as above for any \( \tau \geq \tau^* \). This, in turn, implies that \( \text{spec}(-J_{\tau}(z^*)) \subset \mathbb{C}_{\omega} \) for all \( \tau \geq \tau^* \) since we constructed a Lyapunov function for \( z^* \). Thus, we conclude that if \( z^* \) is a differential Stackelberg equilibrium, then \( \text{spec}(-J_{\tau}(z^*)) \subset \mathbb{C}_{\omega} \) for all \( \tau \in [\tau_*, \infty) \) where \( \tau_* \) is the minimum \( \tau \in (0, \infty) \) such that \( \text{spec}(-J_{\tau}(z^*)) \subset \mathbb{C}_{\omega} \) and a finite \( \tau_* \) is guaranteed to exist.

\[ \square \]

### D Global Asymptotic Convergence Analysis

We now provide the proofs pertaining to the results presented in Section 5.

**D.1 Proof of Lemma 1**

This appendix is devoted to proving Lemma 1. To be clear, we restate the result before proving it.

**Lemma 1.** Consider a non-convex, \( \mu \)-PL zero sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) which satisfies Assumptions 1 and 3 and has condition number \( \kappa = L_2/\mu \). Suppose that \( \tau > 7\kappa^2 \), and \( \gamma < \min\{\frac{1}{3\kappa^2}, \frac{1}{2(1+\gamma^2 L_2)}\} \) then, for any \( \Gamma \in (0, 1/7] \), \( \Phi(x, y) = f(x, y_*(x)) - \Gamma f(x, y) \) is a potential function for \( \tau \)-GDA.

**Proof of Lemma 1** Let the best response be denoted by 

\[
y_*(x) \in \arg\max_{y \in \mathcal{Y}} f(x, y).
\]

Since the function \( f(x, \cdot) \) is PL, there set maximizers may not be a singleton. Hence, \( y_*(x) \) is an element of the set of maximizers.

We claim that for any \( \Gamma \in (0, 1/7] \),

\[
\Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_k) = \Gamma \left( f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \right) + f(x_{k+1}, y_*(x_{k+1})) - f(x_k, y_*(x_k)) < 0
\]

To show this, we need to bound each of the two terms (i) and (ii).
Bounding term (i): \( f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \). To begin, we add and subtract \( f(x_k, y_{k+1}) \) to (i) and get

\[
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) = f(x_k, y_k) - f(x_k, y_{k+1}) + f(x_k, y_{k+1}) - f(x_{k+1}, y_{k+1}).
\]  

(12)

From a Taylor expansion of \(-f(x_k, y_{k+1})\) with respect to \(y_{k+1}\) we obtain

\[
\begin{align*}
- f(x_k, y_{k+1}) &\leq - f(x_k, y_k) - \langle \nabla_2 f(x_k, y_k), y_{k+1} - y_k \rangle + \frac{L_2}{2} \| y_{k+1} - y_k \|_2^2 \\
&\quad - f(x_k, y_k) - \tau \gamma \langle \nabla_2 f(x_k, y_k), \nabla_2 f(x_k, y_k) \rangle + \frac{\tau^2 \gamma^2 L_2}{2} \| \nabla_2 f(x_k, y_k) \|_2^2 \\
&= - f(x_k, y_k) - \tau \gamma \langle \nabla_2 f(x_k, y_k), \nabla_2 f(x_k, y_k) \rangle + \frac{\tau^2 \gamma^2 L_2}{2} \| \nabla_2 f(x_k, y_k) \|_2^2
\end{align*}
\]

Hence, rearranging this bound and combining with (12), we get

\[
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \leq - \left( \tau \gamma - \frac{\tau^2 \gamma^2 L_2}{2} \right) \| \nabla_2 f(x_k, y_k) \|_2^2 + f(x_k, y_{k+1}) - f(x_{k+1}, y_{k+1}).
\]  

(13)

Now, from a Taylor expansion of \(-f(x_{k+1}, y_{k+1})\) with respect to \(x_{k+1}\), we get

\[
\begin{align*}
- f(x_{k+1}, y_{k+1}) &\leq - f(x_k, y_k) - \langle \nabla_1 f(x_k, y_{k+1}), x_{k+1} - x_k \rangle + \frac{L_1}{2} \| x_{k+1} - x_k \|_2^2 \\
&\quad - f(x_k, y_k) + \gamma \langle \nabla_1 f(x_k, y_{k+1}), \nabla_1 f(x_k, y_k) \rangle + \frac{L_1 \gamma^2}{2} \| \nabla_1 f(x_k, y_k) \|_2^2.
\end{align*}
\]

We now add and subtract \(\gamma \| \nabla_1 f(x_k, y_k) \|_2^2\) to obtain

\[
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \leq \gamma \nabla_1 f(x_k, y_{k+1})^\top \nabla_1 f(x_k, y_k) + \frac{L_1 \gamma^2}{2} \| \nabla_1 f(x_k, y_k) \|_2^2 \\
&\quad + \gamma \nabla_1 f(x_k, y_k)^\top \nabla_1 f(x_k, y_k) - \gamma \| \nabla_1 f(x_k, y_k) \|_2^2 \\
&= \gamma \langle \nabla_1 f(x_k, y_{k+1}) - \nabla_1 f(x_k, y_k), \nabla_1 f(x_k, y_k) \rangle + \frac{L_1 \gamma^2 + 2 \gamma}{2} \| \nabla_1 f(x_k, y_k) \|_2^2.
\]  

(14)

Next, we combine this with (13) to get

\[
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \leq - \left( \tau \gamma - \frac{\tau^2 \gamma^2 L_2}{2} \right) \| \nabla_2 f(x_k, y_k) \|_2^2 \\
&\quad + \gamma \langle \nabla_1 f(x_k, y_{k+1}) - \nabla_1 f(x_k, y_k), \nabla_1 f(x_k, y_k) \rangle + \frac{L_1 \gamma^2 + 2 \gamma}{2} \| \nabla_1 f(x_k, y_k) \|_2^2 \\
&\leq - \left( \tau \gamma - \frac{\tau^2 \gamma^2 L_2}{2} \right) \| \nabla_2 f(x_k, y_k) \|_2^2 + \frac{L_1 \gamma^2 + 2 \gamma + \gamma}{2} \| \nabla_1 f(x_k, y_k) \|_2^2 \\
&\quad + \gamma \| \nabla_1 f(x_k, y_{k+1}) - \nabla_1 f(x_k, y_k) \| \| \nabla_1 f(x_k, y_k) \|
\]

where we used Cauchy-Schwartz in the last inequality. Applying Young’s inequality on the last term, we have that

\[
f(x_k, y_k) - f(x_{k+1}, y_{k+1}) \leq - \left( \tau \gamma - \frac{\tau^2 \gamma^2 L_2}{2} \right) \| \nabla_2 f(x_k, y_k) \|_2^2 + \frac{L_1 \gamma^2 + 2 \gamma + \gamma}{2} \| \nabla_1 f(x_k, y_k) \|_2^2 \\
&\quad + \frac{\gamma}{2} \| \nabla_1 f(x_k, y_{k+1}) - \nabla_1 f(x_k, y_k) \|^2 \\
&\leq - \left( \tau \gamma - \frac{\tau^2 \gamma^2 L_2}{2} \right) \| \nabla_2 f(x_k, y_k) \|_2^2 + \frac{L_1 \gamma^2 + 2 \gamma + \gamma}{2} \| \nabla_1 f(x_k, y_k) \|_2^2.
\]  

(15)
Bounding (ii): $f(x_{k+1}, y_*(x_{k+1})) - f(x_k, y_*(x_k))$. To bound (ii), we take a Taylor expansion of $f(x_{k+1}, y_*(x_{k+1}))$ to get that

$$f(x_{k+1}, y_*(x_{k+1})) - f(x_k, y_*(x_k)) \leq \langle \nabla f(x_k, y_*(x_k)), x_{k+1} - x_k \rangle + \frac{L_4 \gamma^2}{2} \| \nabla_1 f(x_k, y_k) \|^2$$

$$= -\gamma \langle \nabla f(x_k, y_*(x_k)), \nabla_1 f(x_k, y_k) \rangle + \frac{L_4 \gamma^2}{2} \| \nabla_1 f(x_k, y_k) \|^2$$

where $L_4 \leq L_1 + L_2 L_3$ is the Lipschitz bound on the total derivative of $f(x, y_*(x))$.

Combining bounds. We now combine the bounds on (i) and (ii). Combining the (15) and (17), we have

$$\Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_k) \leq -\gamma \langle \nabla f(x_k, y_*(x_k)), \nabla_1 f(x_k, y_k) \rangle + \frac{\gamma}{2} \left( \frac{L_4 \gamma + \Gamma L_1 \gamma + 3 \Gamma}{2} \right) \| \nabla_1 f(x_k, y_k) \|^2$$

$$- \Gamma \left( \frac{\tau \gamma}{2} - \frac{\gamma^2 \gamma^2 \gamma}{2} - \frac{\gamma (\tau \gamma)^2 L_2^2}{2} \right) \| \nabla_2 f(x_k, y_k) \|^2.$$

To further bound the above expression, we start by bounding the first two terms. Towards this end, define

$$V = -\gamma \langle \nabla f(x_k, y_*(x_k)), \nabla_1 f(x_k, y_k) \rangle + \frac{\gamma}{2} \left( L_4 \gamma + \Gamma L_1 \gamma + 3 \Gamma \right) \| \nabla_1 f(x_k, y_k) \|^2.$$  

Recall that $\gamma < 1/(2L_4)$ and $\gamma < 1/(2\max\{L_1, L_2\})$. Since $\Gamma \leq 1/7$,

$$\gamma (L_4 + \Gamma L_1) + \Gamma 3 \leq \frac{1}{2} + \frac{\Gamma}{2} \leq 1.$$

Completing the square in (19), we have that

$$V \leq -\gamma \langle \nabla f(x_k, y_*(x_k)), \nabla_1 f(x_k, y_k) \rangle + \frac{\gamma}{2} \| \nabla_1 f(x_k, y_k) \|^2$$

$$\leq -\gamma \| \nabla f(x_k, y_k) \|^2 + \frac{\gamma}{2} \| \nabla f(x_k, y_*(x_k)) - \nabla_1 f(x_k, y_k) \|^2$$

$$\leq -\gamma \| \nabla f(x_k, y_*(x_k)) \|^2 + \frac{\gamma L_2^2}{2} \| y_k - y_*(x_k) \|^2$$

$$\leq -\gamma \| \nabla f(x_k, y_*(x_k)) \|^2 + \frac{\gamma L_2^2}{2} \| \nabla_2 f(x_k, y_k) \|^2$$

where in (20) we used the fact that $\nabla_1 f(x, y)$ is Lipschitz in $y$ and $\nabla f(x_k, y_*(x_k)) = \nabla_1 f(x, y)|_{y=y_*(x)}$ by Lemma [6]. Further, in (21), we used the quadratic growth property of PL functions.

Now, by combining the bound on $V$ with the remaining terms in (18), we have

$$\Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_k) \leq -\gamma \| \nabla f(x_k, y_*(x_k)) \|^2 + \tau \gamma \left( \frac{\kappa^2}{2\tau} - \Gamma + \frac{\Gamma \tau \gamma L_2}{2} + \frac{\Gamma \tau \gamma L_2^2}{2} \right) \| \nabla_2 f(x_k, y_k) \|^2.$$  

(22)

Let

$$C = -\Gamma + \frac{\Gamma \tau \gamma L_2}{2} + \frac{\Gamma \tau \gamma L_2^2}{2} + \frac{\kappa^2}{2\tau}.$$

As long as $C < 0$, as claimed, $\Phi(\cdot)$ will be a potential function. To see this, we upper bound $C$ as follows:

$$C = \frac{\kappa^2}{2\tau} - \Gamma \left( 1 - \frac{\tau \gamma}{2} \left( L_2 + \gamma L_2^2 \right) \right) \leq \frac{\kappa^2}{2\tau} - \Gamma \left( 1 - \frac{\tau \gamma}{2} \frac{3}{4} L_2 \right) \leq \Gamma \left( \frac{\kappa^2}{2\tau} - 2 + \frac{3}{2} \tau \gamma L_2 \right).$$


---


---

27
since \( \gamma < 1/(2 \max \{ L_1, L_2 \}) \). Since \( \tau > \Gamma^{-1} \kappa^2 \) and \( \gamma \leq 1/(3L_2 \tau) \), we have
\[
-C \leq \frac{\Gamma}{2} \left( \frac{\kappa^2}{\Gamma^2} - 2 + \frac{3}{2} \tau \gamma L_2 \right) \leq \frac{\Gamma}{2} \left( \frac{3}{2} \tau \gamma L_2 - 1 \right) < -\frac{\Gamma}{4}.
\]
Hence,
\[
\Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_k) \leq -\frac{\gamma}{2} \| \nabla f(x_k, y_*(x_k)) \|^2 - \frac{\tau \gamma \Gamma}{2} \| \nabla^2 f(x_k, y_k) \|^2
\]
which completes the proof of the claim that \( \Phi \) is a potential function since it is decreasing along trajectories.

\[ \square \]

### D.2 Proof of Theorem 2

In this appendix, we prove Theorem 2. To be clear, we restate the result before proving it.

**Theorem 2.** Consider a nonconvex-PL zero-sum game \((f, -f)\) defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) that satisfies Assumptions 3. Then, \( \tau \)-GDA with \( \tau > 7 \kappa^2 \), and stepsize \( \gamma < \min \{ \frac{1}{3L_2 \tau}, \frac{1}{2(L_1 + L_2 \kappa^2)} \} \) asymptotically converges to the set of strict local minmax that are stable for \( \hat{z} = -\Lambda_\tau g(z) \) almost surely. That is, for almost all initial conditions, \( \tau \)-GDA will converge to a strict local minmax point.

**Proof of Theorem 2.** This result follows nearly immediately from Theorem 1, Lemma 1, and Lemma 2. In particular, the potential function \( \Phi \) from Lemma 1 guarantees that
\[
\Phi(x_{k+1}, y_{k+1}) - \Phi(x_k, y_k) \leq -\frac{\gamma}{2} \| \nabla f(x_k, y_*(x_k)) \|^2 - \frac{\tau \gamma \Gamma}{2} \| \nabla^2 f(x_k, y_k) \|^2.
\]
Thus, the potential function only stops decreasing when we have both
\[
\| \nabla f(x_k, y_*(x_k)) \|^2 = 0 \quad \text{and} \quad \| \nabla^2 f(x_k, y_k) \|^2 = 0.
\]
By the quadratic growth of \( \mu \)-PL functions (see Lemma 3) and the PL property in nonconvex-PL zero-sum games, we have that
\[
\| y_k - y_*(x_k) \|^2 \leq \frac{2}{\mu} (\max_y f(x_k, y) - f(x_k, y_k)) \leq \frac{2}{\mu} \| \nabla^2 f(x_k, y_k) \|^2
\]
Hence, if \( \| \nabla^2 f(x_k, y_k) \|^2 \to 0 \) then \( y_k \to y_*(x_k) \). In particular, when \( \| \nabla^2 f(x_k, y_k) \|^2 = 0 \), we have that \( y_k = y_*(x_k) \) so that \( \| \nabla f(x_k, y_*(x_k)) \|^2 = 0 \) if and only if \( \| \nabla^2 f(x_k, y_k) \|^2 = 0 \). This implies that the potential function only stops decreasing at critical points and this is the only place that a limit cycle could exist. Since the only stable points of \( \hat{z} = -\Lambda_\tau (z) \) are differential Stackelberg equilibrium by Theorem 1 and \( \tau \)-GDA avoids strict saddle points of \( \hat{z} = -\Lambda_\tau (z) \) almost surely by Lemma 2 and the assumption that all saddle points are strict saddles, we can conclude that \( \tau \)-GDA almost surely only reaches a critical point if it is a strict local minmax equilibrium. Thus, these facts ensure that the \( \tau \)-GDA dynamics almost surely converge to a strict local minmax (differential Stackelberg) equilibrium.

\[ \square \]

### D.3 Proof of Corollary 1

In this appendix, we prove Corollary 1. We restate the result and then provide the proof.

**Corollary 1.** Consider a nonconvex-PL zero-sum game \((f, -f)\) defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}) \) that satisfies Assumptions 3. Given an initialization in the region of attraction of a strict local minmax, then after \( \tilde{O}(\varepsilon^{-2}) \) iterations at least one iterate is an \( \varepsilon \)-differential Stackelberg equilibrium.

**Proof of Corollary 1.** For this proof, consider any \( \varepsilon > 0 \). Our approach will be to show that for
\[
T \geq \frac{2(\Phi(x_0, y_0) - \Phi(x_T, y_T))}{\varepsilon^2 \gamma \min \{ 1, \tau \Gamma \}},
\]

...
we have that
\[ \min_{0 \leq k \leq T - 1} \max \left\{ \| \nabla f(x_k, y_*(x_k)) \|, \| \nabla_2 f(x_k, y_k) \| \right\} = \max \left\{ \| \nabla f(x_s, y_*(x_s)) \|, \| \nabla_2 f(x_s, y_s) \| \right\} \leq \varepsilon. \]

Then, we prove that given this fact\(^7\)
\[ \| \nabla f(x_s, y_s) \| \leq \left( 1 + \frac{L_5}{\mu} \right) \varepsilon. \]

This will then allow us to conclude for
\[ T \geq \frac{2 \left( 1 + \frac{L_5}{\mu} \right)^2 (\Phi(x_0, y_0) - \Phi(x_T, y_T))}{\varepsilon^2 \gamma \min\{1, \tau \Gamma\}}, \tag{23} \]
we have both
\[ \| \nabla f(x_s, y_s) \| \leq \varepsilon \quad \text{and} \quad \| \nabla_2 f(x_s, y_s) \| \leq \varepsilon. \]

Then, by selecting the parameters to minimize the right-hand side of (23), we are able to conclude that at least one iterate of the \( \tau \)-GDA dynamics are an \( \varepsilon \)-differential Stackelberg equilibrium after
\[ T \geq \frac{2 \left( 1 + \frac{L_5}{\mu} \right)^2 (\Phi(x_0, y_0) - \Phi(x_T, y_T))}{\varepsilon^2 \gamma \min\{1, \tau \Gamma\}} \]
iterations.

We now formally prove this. Summing the bound on the potential function from Lemma \(^1\) we get the following that is justified below:
\[ \Phi(x_0, y_0) - \Phi(x_T, y_T) = \sum_{k=0}^{T-1} \left( \Phi(x_k, y_k) - \Phi(x_{k+1}, y_{k+1}) \right) \tag{24} \]
\[ \geq \frac{\gamma}{2} \sum_{k=0}^{T-1} \| \nabla f(x_k, y_*(x_k)) \|^2 + \frac{\gamma \Gamma}{2} \sum_{k=0}^{T-1} \| \nabla_2 f(x_k, y_k) \|^2 \tag{25} \]
\[ \geq \frac{\gamma}{2} \min \{1, \tau \Gamma\} \sum_{k=0}^{T-1} \left( \| \nabla f(x_k, y_*(x_k)) \|^2 + \| \nabla_2 f(x_k, y_k) \|^2 \right) \tag{26} \]
\[ \geq \frac{\gamma}{2} \min \{1, \tau \Gamma\} \sum_{k=0}^{T-1} \max \left\{ \| \nabla f(x_k, y_*(x_k)) \|^2, \| \nabla_2 f(x_k, y_k) \|^2 \right\} \tag{27} \]
\[ \geq \frac{\gamma T}{2} \min \{1, \tau \Gamma\} \min_{0 \leq k \leq T - 1} \max \left\{ \| \nabla f(x_k, y_*(x_k)) \|^2, \| \nabla_2 f(x_k, y_k) \|^2 \right\}. \tag{28} \]

Observe that \(24\) follows from telescoping of the sum, \(25\) is a result of applying the bound on the potential function, \(26\) holds since it is replacing a coefficient of a positive number with something smaller, \(27\) holds since the sum of positive numbers is greater than the max, and \(28\) is obtained from the fact that the sum of \( T \) positive numbers is greater than \( T \) times the minimum number.

From the previous steps, and also rearranging terms and then taking the square root, we have
\[ \sqrt{\frac{2 (\Phi(x_0, y_0) - \Phi(x_T, y_T))}{T \gamma \min \{1, \tau \Gamma\}}} \geq \min_{0 \leq k \leq T - 1} \max \left\{ \| \nabla f(x_k, y_*(x_k)) \|, \| \nabla_2 f(x_k, y_k) \| \right\}. \]

We now want to find \( T \) such that
\[ \min_{0 \leq k \leq T - 1} \max \left\{ \| \nabla f(x_k, y_*(x_k)) \|, \| \nabla_2 f(x_k, y_k) \| \right\} \leq \sqrt{\frac{2 (\Phi(x_0, y_0) - \Phi(x_T, y_T))}{T \gamma \min \{1, \tau \Gamma\}}} \leq \varepsilon. \tag{29} \]
\(^7\)See Assumption \(7\) for the derivation of \( L_5 \).
By moving terms around, we find that the inequality in (29) holds for any $T$ such that
\[ T \geq T^* := \frac{2(\Phi(x_0, y_0) - \Phi(x_T, y_T))}{\varepsilon^2 \gamma \min\{1, \tau T\}}. \]

This proves that there exists some iterate $0 \leq s \leq T - 1$ such that for $T \geq T^*$, we have both
\[ \|\nabla f(x_s, y_s(x_s))\| \leq \varepsilon \quad \text{and} \quad \|\nabla^2 f(x_s, y_s)\| \leq \varepsilon. \]

We now show that this implies a bound on $\|\nabla f(x_s, y_s)\|$. In particular, using the fact that $f(x, \cdot)$ is $\mu$-PL, we have that
\[ \|y_s - y_s(x_s)\|^2 \leq \frac{2}{\mu} \left( \max_{y \in \mathcal{Y}} f(x_s, y) - f(x_s, y_s) \right) = \frac{2}{\mu} \left( f(x_s, y_s(x_s)) - f(x_s, y_s) \right) \leq \frac{2}{\mu^2} \|\nabla^2 f(x_s, y_s)\|^2. \]

Since $\|\nabla^2 f(x_s, y_s)\| \leq \varepsilon$, we know that $\|y_s - y_s(x_s)\| \leq \frac{\sqrt{2\varepsilon}}{\mu}$. Then, observe that we have
\[
\|\nabla f(x_s, y_s)\| = \|\nabla f(x_s, y_s) - \nabla f(x_s, y_s(x_s)) + \nabla f(x_s, y_s(x_s))\|
\leq \|\nabla f(x_s, y_s) - \nabla f(x_s, y_s(x_s))\| + \|\nabla f(x_s, y_s(x_s))\|
\leq L_5 \|y_s - y_s(x_s)\| + \|\nabla f(x_s, y_s(x_s))\|
\leq \frac{\sqrt{2L_5\varepsilon}}{\mu} + \varepsilon = \left(1 + \frac{\sqrt{2L_5}}{\mu}\right)\varepsilon,
\]

where we obtain the first inequality using the triangle inequality, the second inequality using the Lipschitz bound, and the final inequality using that $\|\nabla^2 f(x, y_s)\| \leq \varepsilon$ and $\|y_s - y_s(x_s)\| \leq \frac{\sqrt{2\varepsilon}}{\mu}$.

Thus, in order to determine the iteration complexity $T$ needed to get that $\|\nabla f(x_s, y_s)\| \leq \varepsilon$ we redefine the given $\varepsilon$ to be $\varepsilon = \varepsilon'(1 + \frac{\sqrt{2L_5}}{\mu})$ and get that for
\[ T \geq T^* := \frac{2 \left(1 + \frac{\sqrt{2L_5}}{\mu}\right)\left(\Phi(x_0, y_0) - \Phi(x_T, y_T)\right)}{\varepsilon^2 \gamma \min\{1, \tau T\}}, \]

we have both
\[ \|\nabla f(x_s, y_s)\| \leq \varepsilon \quad \text{and} \quad \|\nabla^2 f(x_s, y_s)\| \leq \varepsilon. \]

To finish, we select $\tau = 8\varepsilon^2$, $\Gamma = 1/8$, and $\gamma = \frac{1}{2} \min\{\frac{1}{3L_2^2}, \frac{1}{2(L_1 + L_2)}\}$ to get an iteration complexity of.
\[
T^* = \frac{4 \left(1 + \frac{\sqrt{2L_5}}{\mu}\right)^2 \max\{24L_2k^2, 2(L_1 + L_3L_2)\}(\Phi(x_0, y_0) - \Phi(x_T, y_T))}{\varepsilon^2}. \]

Thus the iteration complexity is $\tilde{O}(\varepsilon^{-2})$ as claimed to reach an $\varepsilon$-differentiable Stackelberg equilibrium. We note that the assumption of initializing in the region of attraction of a strict local minmax is based on the the stability result from Theorem 1 that guarantees there exists a local neighborhood on which $\tau$-GDA converges toward the equilibrium that is stable.

\section{Finite-Time Convergence Analysis}

This appendix contains the analysis for the results from Section 6 on finite-time global convergence guarantees to $\varepsilon$-differentiable Stackelberg equilibria by efficiently escaping saddle points and leveraging the stability result from Theorem 1. We present proofs for the stochastic descent in Appendix E.1 escaping saddle points in Appendix E.2 and convergence in Appendix E.3.
\( \tau \text{-PGDA Dynamics.} \) Recall from Algorithm 2 that the combined update at each time \( k \) of the \( \tau \text{-PGDA} \) dynamics is given by

\[
x_{k+1} = x_{k} - \gamma (g_{1}(x_{k}, y_{k}; \theta_{1,k}) + \xi_{1,k}) = x_{k} - \gamma (\nabla_{1} f(x_{k}, y_{k}) + \zeta_{1,k} + \xi_{1,k}) \\
y_{k+1} = y_{k} + \gamma \tau (g_{2}(x_{k}, y_{k}; \theta_{2,k}) + \xi_{2,k}) = y_{k} + \gamma \tau (\nabla_{2} f(x_{k}, y_{k}) + \zeta_{2,k} + \tau^{-1} \xi_{2,k})
\]

where for each player \( i = \{1, 2\} \), \( g_{i}(x_{k}, y_{k}; \theta_{i,k}) \) denotes the stochastic gradient with \( \theta_{i,k} \) a random variable drawn from a distribution \( D_{i} \), \( \xi_{i,k} = g_{i}(x_{k}, y_{k}; \theta_{i,k}) - \nabla_{i} f(x_{k}, y_{k}) \) denotes the noise in the stochastic gradient, and \( \tilde{\zeta}_{i,k} = \zeta_{i,k} + \xi_{i,k} \) the summation of the noise from the stochastic gradient and the random perturbation for each player \( i = \{1, 2\} \). Furthermore, we state again that by Assumption 4 for each player \( i = \{1, 2\} \), the stochastic gradient \( g_{i}(x_{k}, y_{k}; \theta_{i,k}) \) is \( \sigma_{i} \)-norm-subGaussian (see Definition 9). We also denote by \( \tilde{\sigma}_{i}^{2} = \sigma_{i}^{2} + r^{2} \) for each player \( i = \{1, 2\} \).

\[ \text{Differences from the single player optimization setting.} \] The proof of the finite-time convergence result appears analogous to that of single player setting which as been studied extensively in recent literature (Adolphs et al., 2019; Daneshmand et al., 2018; Fang et al., 2019; Jin et al., 2017, 2018, 2021; Lee et al., 2019; Levy, 2016; Staib et al., 2019; Vlatakis-Gkaragkounis et al., 2019; Xu et al., 2018). However, there are several key differences that make zero-sum game setting significantly more challenging. The fist of which is that the Jacobians of \( g(z) \) and \( \tilde{g}(z) \) are both not symmetric, which means they potentially have complex eigenvalues. This requires a whole new set of analysis techniques for handling saddle avoidance as can be seen in the proof of Lemma which shows that if we define two "coupled sequences" (see Definition 10) that at least one of the two sequences initialized from the same point is guaranteed to escape with high probability. This analysis requires special treatment of the spectral norm of the local linearization of the gradient dynamics as well as the sub-Gaussian random vectors. Similarly, in the proof of Lemma 25 which more precisely bounds the spectral radius of the local linearization, the stepsize size has to be carefully chosen given the fact that the minimum real-part eigenvalue may be complex and hence the imaginary component can contribute significantly to the spectral radius. Moreover, in Lemma 26 unlike analogous results in the single player case, due to the non-symmetric Jacobian, we need to employ novel matrix norm bounds to obtain a convergence rate in the 2-norm which does not pick-up any extra dimensional factors. These are only some of the technical novelties that make the proof in the multiplayer setting depart from the traditional analysis in the single player setting.

\[ \text{E.1 \ Stochastic Descent Lemma.} \] We use the notation \( \lambda_{\min}() \) and \( \lambda_{\max}() \) to denote the minimum and maximum real part of the eigenvalues of the argument, respectively. In addition, let \( \rho() \) denote the spectral radius of the argument. We say a point \( (x^{\ast}, y^{\ast}) \) is an \( \varepsilon \)-local minimax point (or equivalently, an \( \varepsilon \)-differential Stackelberg equilibrium, terms we use interchangeably) for the minmax problem defined by a \( (L_{1} + L_{2}) \)-gradient Lipschitz and \( \beta \)-Hessian Lipschitz function \( f \in \mathcal{C}(X \times Y, \mathbb{R}) \) if

\[
\|\tilde{g}(x^{\ast}, y^{\ast})\| \leq \varepsilon \quad \text{and} \quad \lambda_{\min}(J_{\tau}(x^{\ast}, y^{\ast})) \geq -\sqrt{\beta \varepsilon}.
\]

where \( \tilde{g}(x, y) = (\nabla f(x, y), -\nabla_{2} f(x, y)) \). See also Definition 6. Note that the spectral radius is always greater than the minimum real part eigenvalue:

\[
\rho(J_{\tau}(x^{\ast}, y^{\ast})) \geq \lambda_{\min}(J_{\tau}(x^{\ast}, y^{\ast})) \geq -\sqrt{\beta \varepsilon}
\]

Define

\[
\ell = \begin{cases} 
L_{1} + L_{2}, & \text{if } 2L_{2} \leq \mu \\
L_{1} + 2\kappa L_{2}, & \text{if } \mu < 2L_{2}
\end{cases}
\]

It will be clear in the proof of the descent lemma below precisely why this conditional value of \( \ell \) is chosen. Note that when \( \ell = L_{1} + L_{2} \), since \( f \) is gradient Lipschitz in \( x \) and \( y \), \( \lambda_{\max}(\nabla f(x, y)) \leq \ell \), and when \( \ell = L_{1} + 2\kappa L_{2} \), observe that since \( \frac{1}{\kappa} \mu < \frac{1}{2} L_{2} \), we have that \( \lambda_{\max}(\nabla f(x, y)) \leq L_{1} + L_{2} \leq \ell \). The purpose of defining \( \ell \) as
above is so that we can ensure that we can assume \( \ell \geq \sqrt{\beta}. \) Otherwise, if \( \ell < \sqrt{\beta} \) then finding \( \varepsilon \)-local minmax points is straightforward since all \( \varepsilon \)-stationary points will be \( \varepsilon \)-local minmax points.

The following lemma is Lemma 3 from the main text. We restate it here with more precise constants. The decent lemma shows that for a nonconvex, \( \mu \)-SC zero-sum game, the function \( f \) acts as a potential function for the \( \tau \)-PGDA dynamics.

Lemma 19 (Descent Lemma). Consider a non-convex, \( \mu \)-SC zero-sum game defined by \( f \in C^2(\mathbb{Z}, \mathbb{R}). \) Under Assumptions 1–5, there exists an absolute constant \( c_{\max} \) such that for any fixed \( k, k_0, \varepsilon > 0 \), if \( \frac{3}{2L_1} < \gamma < 1/\ell \), then with at least \( 1 - 8e^{-1} \) probability, the sequence generated by \( \tau \)-PGDA with parameters \( \gamma, \tau, \nu \) satisfies

\[
  f(z_{k_0} + k) - f(z_{k_0}) \leq \frac{1}{8} \sum_{k=0}^{t-1} \left( \| \nabla_1 f(z_{k_0} + k) \|^2 + \|
abla_2 f(z_{k_0} + k) \|^2 \right) + c \left( \gamma (\sigma_1^2 + r^2) (\gamma \ell k + \nu) + \tau \gamma (\sigma_2^2 + r^2) (\mu \tau \gamma k + \nu) \right).
\]

Proof. Without loss of generality we can take \( t_0 = 0 \) (Algorithm 2 is Markovian).

We first argue a one step descent bound. Towards this end, take the Taylor expansion of \( f(\cdot, y_k) \) to get that

\[
  f(x_{k+1}, y_k) - f(x_k, y_k) \leq \langle \nabla_1 f(x_k, y_k), x_{k+1} - x_k \rangle + \frac{L_1}{2} \| x_{k+1} - x_k \|^2.
\]

Now we add \( \langle \nabla_2 f(x_k, y_k), y_{k+1} - y_k \rangle \) to get that

\[
  f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \leq \langle \nabla_2 f(x_{k+1}, y_k), y_{k+1} - y_k \rangle - \frac{\mu}{2} \| y_{k+1} - y_k \|^2 + \langle \nabla_2 f(x_k, y_k), y_{k+1} - y_k \rangle.
\]

By Young’s inequality with \( \epsilon > 0 \) and the update equation for the \( y \)-player, we have that

\[
  f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \leq \frac{L_2}{2\epsilon} \left( \| x_{k+1} - x_k \|^2 - \frac{\mu - \epsilon}{2} \| y_{k+1} - y_k \|^2 + \langle \nabla_2 f(x_k, y_k), y_{k+1} - y_k \rangle \right) + \frac{\mu}{2} \| y_{k+1} - y_k \|^2 + \langle \nabla_2 f(x_k, y_k), y_{k+1} - y_k \rangle + \tilde{\zeta}_{2,k}.
\]

We next put the above inequality together with (31). Towards this end, let \( \epsilon = \min \{ \frac{3}{4} L_2, \frac{1}{4} \mu \} \) and define

\[
  \ell = \begin{cases} 
    L_1 + L_2, & \text{if } \epsilon = \frac{3}{4} L_2 \\
    L_1 + 2\kappa L_2, & \text{if } \epsilon = \frac{1}{4} \mu
  \end{cases}
\]

We consider the two cases that define \( \ell \) above.
Case 1: $\epsilon = \frac{1}{3} \mu$. Since $\frac{3}{2\mu r} < \gamma < \frac{1}{3}$, plugging in $\epsilon = \frac{1}{3} \mu$, we have that
\[
\begin{align*}
f(x_{k+1}, y_{k+1}) - f(x_k, y_k) &\leq -\left(\gamma - \frac{3L_1}{4}\right) \left\|\nabla_1 f(x_k, y_k)\right\|^2 - \gamma \langle \nabla_1 f(x_k, y_k), \tilde{\zeta}_{1,k} \rangle + \frac{3}{2} \gamma^2 L_1 \left\|\tilde{\zeta}_{1,k}\right\|^2 \\
&\quad + \frac{\tau \mu}{4\mu} \left\|x_{k+1} - x_k\right\|^2 - \frac{\tau}{2} \left(\frac{\mu - \epsilon}{2} \tau \gamma - 1\right) \left\|\nabla_2 f(x_k, y_k)\right\|^2 - \frac{\mu - \epsilon}{2} \gamma \left\|\tilde{\zeta}_{2,k}\right\|^2 \\
&\quad + \tau \gamma \langle \nabla_2 f(x_k, y_k), \tilde{\zeta}_{2,k} \rangle \\
&\leq -\left(\gamma - \frac{3L_1}{4}\right) \left\|\nabla_1 f(x_k, y_k)\right\|^2 - \gamma \langle \nabla_1 f(x_k, y_k), \tilde{\zeta}_{1,k} \rangle + \frac{3}{2} \gamma^2 L_1 \left\|\tilde{\zeta}_{1,k}\right\|^2 \\
&\quad - \tau \gamma \left(\frac{\mu \gamma}{3} - 1\right) \left\|\nabla_2 f(x_k, y_k)\right\|^2 - \frac{\mu \gamma^2}{3} \left\|\tilde{\zeta}_{2,k}\right\|^2 + \tau \gamma \langle \nabla_2 f(x_k, y_k), \tilde{\zeta}_{2,k} \rangle.
\end{align*}
\]

Case 2: $\epsilon = \frac{3}{2} L_2$. Since $\frac{3}{2\mu r} < \gamma < \frac{1}{3}$ and $\frac{3}{2} L_2 \leq \frac{1}{3} \mu$, we get that
\[
\begin{align*}
f(x_{k+1}, y_{k+1}) - f(x_k, y_k) &\leq -\left(\gamma - \frac{3L_1}{4}\right) \left\|\nabla_1 f(x_k, y_k)\right\|^2 - \gamma \langle \nabla_1 f(x_k, y_k), \tilde{\zeta}_{1,k} \rangle + \frac{3}{2} \gamma^2 L_1 \left\|\tilde{\zeta}_{1,k}\right\|^2 \\
&\quad + \frac{3L_2}{4} \left\|x_{k+1} - x_k\right\|^2 - \tau \gamma \left(\frac{\mu \gamma}{3} - 1\right) \left\|\nabla_2 f(x_k, y_k)\right\|^2 - \frac{\mu \gamma^2}{3} \left\|\tilde{\zeta}_{2,k}\right\|^2 \\
&\quad + \tau \gamma \langle \nabla_2 f(x_k, y_k), \tilde{\zeta}_{2,k} \rangle \\
&\leq -\left(\gamma - \frac{3L_1}{4}\right) \left\|\nabla_1 f(x_k, y_k)\right\|^2 - \gamma \langle \nabla_1 f(x_k, y_k), \tilde{\zeta}_{1,k} \rangle + \frac{3}{2} \gamma^2 L_1 \left\|\tilde{\zeta}_{1,k}\right\|^2 \\
&\quad - \tau \gamma \left(\frac{\mu \gamma}{3} - 1\right) \left\|\nabla_2 f(x_k, y_k)\right\|^2 - \frac{\mu \gamma^2}{3} \left\|\tilde{\zeta}_{2,k}\right\|^2 + \tau \gamma \langle \nabla_2 f(x_k, y_k), \tilde{\zeta}_{2,k} \rangle.
\end{align*}
\]
In both cases, we see that we get
\[
f(x_{k+1}, y_{k+1}) - f(x_k, y_k) \leq -\frac{\gamma}{4} \left\|\nabla_1 f(x_k, y_k)\right\|^2 - \gamma \langle \nabla_1 f(x_k, y_k), \tilde{\zeta}_{1,k} \rangle + \frac{3}{4} \gamma^2 \left\|\tilde{\zeta}_{1,k}\right\|^2 \\
&\quad - \frac{\tau \gamma}{2} \left\|\nabla_2 f(x_k, y_k)\right\|^2 - \frac{\mu \gamma^2}{3} \left\|\tilde{\zeta}_{2,k}\right\|^2 + \tau \gamma \langle \nabla_2 f(x_k, y_k), \tilde{\zeta}_{2,k} \rangle
\]
where the last inequality holds since $\gamma < 1/\ell$.

Summing this inequality on both sides, we have that
\[
f(x_k, y_k) - f(x_0, y_0) \leq -\frac{\gamma}{4} \sum_{t=0}^{k-1} \left\|\nabla_1 f(x_t, y_t)\right\|^2 - \gamma \sum_{t=0}^{k-1} \langle \nabla_1 f(x_t, y_t), \tilde{\zeta}_{1,t} \rangle + \frac{3}{4} \gamma^2 \sum_{t=0}^{k-1} \left\|\tilde{\zeta}_{1,t}\right\|^2 \\
&\quad - \frac{\tau \gamma}{2} \sum_{t=0}^{k-1} \left\|\nabla_2 f(x_t, y_t)\right\|^2 - \frac{\mu \gamma^2}{3} \sum_{t=0}^{k-1} \left\|\tilde{\zeta}_{2,t}\right\|^2 + \tau \gamma \sum_{t=0}^{k-1} \langle \nabla_2 f(x_t, y_t), \tilde{\zeta}_{2,t} \rangle \\
&\leq -\frac{\gamma}{4} \sum_{t=0}^{k-1} \left\|\nabla_1 f(x_t, y_t)\right\|^2 - \gamma \sum_{t=0}^{k-1} \langle \nabla_1 f(x_t, y_t), \tilde{\zeta}_{1,t} \rangle + \frac{3}{4} \gamma^2 \sum_{t=0}^{k-1} \left\|\tilde{\zeta}_{1,t}\right\|^2 \\
&\quad - \frac{\tau \gamma}{2} \sum_{t=0}^{k-1} \left\|\nabla_2 f(x_t, y_t)\right\|^2 - \frac{\mu \gamma^2}{3} \sum_{t=0}^{k-1} \left\|\tilde{\zeta}_{2,t}\right\|^2 + \tau \gamma \sum_{t=0}^{k-1} \langle \nabla_2 f(x_t, y_t), \tilde{\zeta}_{2,t} \rangle.
\]
Now, by Lemma 18 there exists a constant $c_1$ such that with probability $1 - 2e^{-t}$ the second term on the right hand side is bounded by
\[
-\gamma \sum_{t=0}^{k-1} \langle \nabla_1 f(x_t, y_t), \tilde{\zeta}_{1,t} \rangle \leq \frac{\gamma}{8} \sum_{t=0}^{k-1} \left\|\nabla_1 f(x_t, y_t)\right\|^2 + c_1 \gamma \tilde{\sigma}_t^2
\]
and similarly, there exists a constant $c_2$ such that with probability $1 - 2e^{-\ell}$ the last term on the right hand side of the inequality is bounded by
\begin{align}
\tau \gamma \sum_{t=0}^{k-1} \langle \nabla_2 f(x_t, y_t), \hat{z}_{1,t} \rangle \leq \frac{\tau \gamma}{8} \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 + c_2 \gamma \tau \hat{\sigma}_2^2 f.
\end{align}

Applying Lemma 17 to $\| \hat{z}_{1,t} \|^2$ and $\| \hat{z}_{2,t} \|^2$ terms, with probability $1 - 2e^{-\ell}$, we have that
\begin{align}
\frac{3\gamma^2}{4} \sum_{t=0}^{k-1} \| \hat{z}_{1,t} \|^2 \leq \frac{3\gamma^2}{4} \sum_{t=0}^{k-1} \| \hat{z}_{1,t} \|^2 + 2\| \hat{z}_{1,t} \|^2 \leq c_1 \gamma^2 \tau \hat{\sigma}_2^2 (k + \ell)
\end{align}
and
\begin{align}
-\frac{\mu \tau^2 \gamma^2}{3} \sum_{t=0}^{k-1} \| \hat{z}_{2,t} \|^2 \leq \frac{\mu \tau^2 \gamma^2}{3} \sum_{t=0}^{k-1} \| \hat{z}_{2,t} \|^2 \leq \frac{2 \mu \tau^2 \gamma^2}{3} \sum_{t=0}^{k-1} (\| \hat{z}_{2,t} \|^2 + \| \hat{z}_{2,t} \|^2) \leq c_2 \frac{2 \mu \tau^2 \gamma^2}{3} \hat{\sigma}_2^2 (k + \ell)
\end{align}
where we are absorbing constants into $c_i$, $i = 1, 2$ throughout since they can always be made larger. Hence, noting that $3/(2\mu \tau) < \gamma < 1/\ell$ with probability $1 - 8e^{-\ell}$,
\begin{align}
f(x_k, y_k) - f(x_0, y_0) &\leq -\frac{\gamma}{4} \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 - \gamma \sum_{t=0}^{k-1} \langle \nabla_1 f(x_t, y_t), \hat{z}_{1,t} \rangle + \frac{3}{2} \gamma \sum_{t=0}^{k-1} \| \hat{z}_{1,t} \|^2 \\
&\quad - \frac{\tau \gamma}{4} \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 - \tau \gamma \sum_{t=0}^{k-1} \| \hat{z}_{2,t} \|^2 + \tau \gamma \sum_{t=0}^{k-1} \langle \nabla_2 f(x_t, y_t), \hat{z}_{2,t} \rangle \\
&\quad \leq -\frac{\gamma}{8} \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + \tau \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 \right) \\
&\quad + c_1 \gamma \left( \hat{\sigma}_2^2 (\gamma \ell k + \ell) + \tau \hat{\sigma}_2^2 (\mu \tau \gamma k + \ell) \right)
\end{align}
where $c = \max\{c_1, c_2\}$. This completes the proof of the stochastic descent lemma.

We now show what is commonly known as "improve or localize" in the recent saddle avoidance literature: with high probability, either the function value decreases significantly, or the iterates stay in a small local region. In particular, the following lemma states that with high probability, if the function value does not decrease "too much" over $k$ iterations, then the iterates of Algorithm 2 remain in a "localized" region around $x_{k_0}$.

**Lemma 20.** Consider the assumptions and parameter choices of Lemma 19. Then, with probability at least $1 - 16de^{-\ell}$, there exists an absolute constant $c$ such that the sequence generated by Algorithm 2 satisfies
\begin{align}
\forall k' \leq k : \| x_{k_0 + k'} - x_{k_0} \|^2 + \| y_{k_0 + k'} - y_{k_0} \|^2 \leq c_3 \gamma \tau k (f(x_{k_0}, y_{k_0}) - f(x_{k_0 + k}, y_{k_0 + k})) \\
+ c_3 \gamma \tau k \left( \hat{\sigma}_1^2 (\gamma \ell k + \ell) + \tau \hat{\sigma}_2^2 (\mu \tau \gamma k + \ell) \right).
\end{align}

**Proof.** Since Algorithm 2 is Markovian, without loss of generality we can take $k_0 = 0$. Before proceeding, we need the following property for the remainder of the proof that derives from the Cauchy-Schwarz inequality.

**Fact 1.** Given $v_1, \ldots, v_n \in \mathbb{R}^m$, $\| \sum_{i=1}^n v_i \|^2 \leq n(\sum_{i=1}^n \| v_i \|^2)$. 

34
For any fixed $k' \leq k$ we obtain the following by applying Fact 1:

$$
\|x_{k'} - x_0\|^2 + \|y_{k'} - y_0\|^2 = \gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) + \tilde{\zeta}_{1,t} \| \right)^2 + \tau^2 \gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) + \tilde{\zeta}_{2,t} \| \right)^2
\leq 2\gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \| \right)^2 + \left( \sum_{t=0}^{k-1} \| \tilde{\zeta}_{1,t} \| \right)^2
+ 2\tau^2 \gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \| \right)^2 + \left( \sum_{t=0}^{k-1} \| \tilde{\zeta}_{2,t} \| \right)^2.
\tag{38}
$$

We now proceed by applying Lemma 15 to the terms $\| \sum_{t=0}^{k'-1} \tilde{\zeta}_{1,t} \|^2$ and $\| \sum_{t=0}^{k'-1} \tilde{\zeta}_{2,t} \|^2$ in the final quantity above. From this, we get that with probability at least $1 - 4de^{-\epsilon}$ there exists constants $c_1$ and $c_2$ such that

$$
\|x_{k'} - x_0\|^2 + \|y_{k'} - y_0\|^2 \leq 2\gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \| \right)^2 + c_1 k' \tilde{\sigma}^2_{1,t} + 2\tau^2 \gamma^2 \left( \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \| \right)^2 + c_2 k' \tilde{\sigma}^2_{2,t}.
$$

Then, applying Fact 1 to the terms $\| \sum_{t=0}^{k-1} \nabla_1 f(x_t, y_t) \|^2$ and $\| \sum_{t=0}^{k-1} \nabla_2 f(x_t, y_t) \|^2$ and using that $k' \leq k$, we get that

$$
\|x_{k'} - x_0\|^2 + \|y_{k'} - y_0\|^2 \leq 2\gamma^2 k \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \| \right)^2 + c_1 k \tilde{\sigma}^2_{1,t} + 2\tau^2 \gamma^2 k \left( \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \| \right)^2 + c_2 \tilde{\sigma}^2_{2,t}
\leq 2\gamma^2 k \left( \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \| \right)^2 + c_1 \tilde{\sigma}^2_{1,t} + 2\tau^2 \gamma^2 k \left( \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \| \right)^2 + c_2 \tilde{\sigma}^2_{2,t}
= 2\gamma^2 k \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + 2\tau^2 \gamma^2 k \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 + 2\gamma e^2 (c_1 \tilde{\sigma}^2_{1} + c_2 \tilde{\sigma}^2_{2} \tau^2).
\tag{39}
$$

Moreover, given that $\tau > 1$, we have that

$$
2\gamma^2 k \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + 2\tau^2 \gamma^2 k \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 \leq 2\gamma^2 k \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + 2\tau^2 \gamma^2 k \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2.
\tag{40}
$$

Recall from Lemma 19 that, with probability at least $1 - 8e^{-\epsilon}$, there exists a constant $c_3$ such that

$$
\sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + \tau \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2 \leq \frac{8}{\gamma} (f(x_0, y_0) - f(x_k, y_k)) + \frac{8c_3}{\gamma} (\gamma (L_1 + 2\kappa L_2) k + \epsilon) + \tau \gamma \tilde{\sigma}^2 (\mu \gamma k + \epsilon).
\tag{41}
$$

Thus, from (40) and (41), with probability at least $1 - 16de^{-\epsilon}$ we have that

$$
2\gamma^2 k \sum_{t=0}^{k-1} \| \nabla_1 f(x_t, y_t) \|^2 + 2\tau^2 \gamma^2 k \sum_{t=0}^{k-1} \| \nabla_2 f(x_t, y_t) \|^2
\leq 16 \gamma \tau k (f(x_0, y_0) - f(x_k, y_k)) + 16c_3 \gamma k (\gamma (L_1 + 2\kappa L_2) k + \epsilon) + \tau \gamma \tilde{\sigma}^2 (\mu \gamma k + \epsilon).
\tag{42}
$$

Finally, from (39), (42), and a union bound, for all $k' \leq k$ with probability at least $1 - 16dke^{-\epsilon}$ it holds that

$$
\|x_{k'} - x_0\|^2 + \|y_{k'} - y_0\|^2 \leq c \gamma k (f(x_0, y_0) - f(x_k, y_k)) + c \gamma^2 k (\tilde{\sigma}^2 (\gamma (L_1 + 2\kappa L_2) k + \epsilon) + \tau \tilde{\sigma}^2 (\mu \gamma k + \epsilon))
$$

where $c = \max\{16, 2c_1, 2c_2, 32c_3\}$.  

\[\square\]
Then, (Algorithm 2) satisfies the following:

**Lemma 21** (Escaping Saddle Points). Given Assumptions 1–6, there exists an absolute constant $c$ such that, for any fixed $k_0 > 0$, $i > c \max \log(\ell \sqrt{d}/(3 \varepsilon))$, if $r, \tau, \gamma, F, K, \mathcal{R}$ are chosen as in Table 2 and $z_{t_0}$ satisfies $\|\nabla_1 f(z_{t_0}), \nabla_2 f(z_{t_0})\| \leq \varepsilon$ and $\text{Re}(\lambda_{\min}(J_r(z_{t_0}))) \leq -\sqrt{3}\varepsilon$, then the sequence of iterates generated by $\tau$-PGDA (Algorithm 3) satisfies the following:

$$\mathbb{P}(f(z_{t_0 + \tau}) - f(z_{t_0}) \leq 0.1 \mathcal{F}) \geq 1 - 8c^{-i} \tag{43}$$

and

$$\mathbb{P}(f(z_{t_0 + \tau}) - f(z_{t_0}) \leq -\mathcal{F}) \geq \frac{1}{3} - 5d \mathcal{T}^2 \log(\sqrt{d}/(\gamma r)) e^{-i} - 32d \mathcal{T} e^{-i}. \tag{44}$$

Toward proving this Lemma 21, we need to prove several intermediate results. We begin by showing that if the function values both sequences do not decrease sufficiently, they both remain localized around the initial condition.

**Lemma 22** (Localization.). Consider the notation and assumptions of Lemma 21. Define the events

$$\mathcal{E}_1 = \{\min\{f(z_{\tau}) - f(z_0), f(z_{\tau}) - f(z_0)\} \leq -\mathcal{F}\}$$

and

$$\mathcal{E}_2 = \{\forall k \leq \mathcal{T} : \max\{\|z_k - z_0\|^2, \|z_k' - z_0\|^2\} \leq \mathcal{S}^2\}.$$

Then,

$$\mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_1) \geq \mathbb{P}(\mathcal{E}_2 | \mathcal{E}_1^c) \geq 1 - \delta_1. \tag{45}$$

where $\delta_1 = 32d \mathcal{T} e^{-i}$. 

---

<table>
<thead>
<tr>
<th>Constant</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell$</td>
<td>$c \log(d \Delta f \mathcal{R}/(3 \varepsilon \delta))$</td>
</tr>
<tr>
<td>$\mathcal{F}$</td>
<td>$\frac{\tau}{T} \frac{1}{\tau} \sqrt{\frac{\gamma^2}{\beta}}$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\tau \sqrt{\mathcal{R}}$</td>
</tr>
<tr>
<td>$\mathcal{T}$</td>
<td>$\frac{1}{\gamma \sqrt{2 \varepsilon}}$</td>
</tr>
<tr>
<td>$r$</td>
<td>$\ell \varepsilon \sqrt{\mathcal{R}}$</td>
</tr>
<tr>
<td>$\mathcal{K}$</td>
<td>$\frac{2}{\beta \tau}$</td>
</tr>
<tr>
<td>$\mathcal{S}$</td>
<td>$\frac{2}{\tau} \sqrt{\frac{\varepsilon \mathcal{R}}{\beta}}$</td>
</tr>
<tr>
<td>$\mathcal{R}$</td>
<td>$1 + \max{\frac{\max(\sigma_{1,1}^1, \sigma_{1,1}^2)}{\varepsilon \mathcal{R}}, \frac{1}{\ell \mu \sqrt{2 \varepsilon}}, \frac{\max(\sigma_{1,1}^1, \sigma_{1,1}^2)}{\varepsilon \mathcal{R}}}$</td>
</tr>
</tbody>
</table>

Table 2: Table of parameters
Proof. Let $\mathcal{E}_1^c$ denote the complement of the event $\mathcal{E}_1$. Observe that we have

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2)$$

where we use the fact that $P(\mathcal{E}_1) \geq P(\mathcal{E}_2|\mathcal{E}_1^c)P(\mathcal{E}_1^c)$ since $P(\mathcal{E}_2|\mathcal{E}_1^c) \leq 1$.

We are left to show that $P(\mathcal{E}_2|\mathcal{E}_1^c) \geq 1 - \delta_1$ where $\delta_1 = 32dTe^{-t_i}$. Let us start with the result of Lemma 20, which states that for all $k \leq T$ with probability at least $1 - 8dTe^{-t_i}$,

$$\|z_k - z_0\|^2 \leq c^2\gamma T(f(z_0) - f(z_T)) + c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i))$$

where $z_k = (x_k, y_k)$ and where we take $k = T$ and without loss of generality we take $k_0 = 0$ (since the algorithm is Markovian). We also have from above that

$$f(z_0) - f(z_T) < F = \frac{r^3}{2\beta} \sqrt{\frac{\epsilon}{\beta}}$$

so that with probability $1 - 8dke^{-t_i}$, for all $\ell \leq T$,

$$\|z_\ell - z_0\|^2 \leq c^2\gamma T(f(z_0) - f(z_T)) + c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i))$$

or

$$\|z_\ell - z_0\|^2 \leq c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i)).$$

Now, we simply need to bound the term (ii) above:

$$c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i))$$

$$\leq c^2\gamma T\frac{\ell}{\sqrt{\beta}}(\sigma^2_1 + \bar{\sigma}^2_2(\mu \gamma T + i))$$

$$\bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i))$$

$$\leq c^2\gamma T\frac{\ell}{\sqrt{\beta}}(\sigma^2_1 + \bar{\sigma}^2_2(\mu \gamma T + i))$$

where we use the fact that $r = \epsilon \sqrt{R}$. Distributing $\gamma$, we have that

$$c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i))$$

$$\leq c^2\gamma T\frac{\ell}{\sqrt{\beta}}\left(\frac{1}{\varepsilon^2\ell R^2} + \frac{\varepsilon^2}{\ell^2 R^2}\right)(\ell \frac{\ell}{\sqrt{\beta}} + i)$$

$$\frac{\varepsilon^2}{\ell^2 R^2}(\ell \frac{\ell}{\sqrt{\beta}} + i) + \tau(\sigma^2_1 + \bar{\sigma}^2_2(\mu \gamma T + i))$$

$$\leq c^2\gamma T\frac{\ell}{\sqrt{\beta}}(\sigma^2_1 + \bar{\sigma}^2_2(\mu \gamma T + i))$$

By our choice of stepsizes, $\mu > \ell / \tau$. Hence, collecting terms, we have that

$$c^2\gamma T(\gamma \bar{\sigma}^2_1(\gamma \ell T + i) + \gamma \bar{\sigma}^2_2(\mu \gamma T + i)) \leq 4c\mathcal{K}(\frac{\varepsilon}{\beta^2} + \frac{\varepsilon}{\beta^2}) \leq 8\frac{\mathcal{K}\varepsilon}{\beta^2}$$

$$\leq 8\frac{\mathcal{K}\varepsilon}{\beta^2}$$
where in the second to last inequality we have used the following facts: $\mathcal{K} = \tau^4/\ell$, $\tau > 1$, and $\mathcal{R} > \max\{\sigma_1^2, \sigma_2^2\}/\varepsilon^2$. Now, we can combine this with the term (i) in (18) to get that

$$\|z_t - z_0\|^2 \leq c\mathcal{K} \frac{c}{\ell \beta} + c\gamma T \left( c\gamma T + i \right) + \gamma T \varepsilon^2 \leq \hat{c}4\mathcal{K} \frac{c}{\beta \ell^4} = \hat{c}S^2$$

where $\hat{c} = \frac{9}{2}c$. This bound holds with probability $1 - 16dT e^{-t}$, hence taking a union bound over the same argument for $\|z'_k - z_0\|^2$, the statement in the lemma holds.

We now define the dynamics of the coupling sequence difference. After we define the dynamics of the coupling sequence difference in Lemma 23, we then go on to show that localization happens with small probability, so that with high probability one of the sequences in the coupling sequence must have a direction of escape from the region around a saddle point.

**Lemma 23.** Consider coupling sequences $\{z_k\} = \{x_k, y_k\}$ and $\{z'_k\} = \{x'_k, y'_k\}$ as given in Definition 10. Moreover, let $\hat{z}_k = z_k - z'_k$, $\hat{\zeta}_k = (\zeta_{1,k} - \zeta'_{1,k}, -\zeta_{2,k} - \zeta'_{2,k})$, and $\hat{\xi}_k = (\xi_{1,k} - \xi'_{1,k}, -\xi_{2,k} - \xi'_{2,k})$. Then,

$$\hat{z}_k = -\hat{q}_z,k - \hat{q}_\zeta,k - \hat{q}_\xi,k,$$  

where

$$\hat{q}_z,k = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{J}_\tau)^{k-1-t} \Delta_t \hat{z}_t,$$  

$$\hat{q}_\zeta,k = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{J}_\tau)^{k-1-t} \hat{\zeta}_t,$$  

$$\hat{q}_\xi,k = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{J}_\tau)^{k-1-t} \hat{\xi}_t,$$

and

$$\Delta_k = \int_0^1 J_\tau(\psi x_k + (1 - \psi)x'_k, \psi y_k + (1 - \psi)y'_k) d\psi - \mathcal{J}_\tau.$$  

**Proof.** Recall that the combined update formula at each time $k$ is given by

$$x_{k+1} = x_k - \gamma (g_1(x_k, y_k; \theta_1,k) + \zeta_{1,k}) = x_k - \gamma (\nabla_1 f(x_k, y_k) + \zeta_{1,k} + \xi_{1,k}),$$

$$y_{k+1} = y_k + \gamma (g_2(x_k, y_k; \theta_2,k) + \zeta_{2,k}) = y_k + \gamma (\nabla_2 f(x_k, y_k) + \zeta_{2,k} + \xi_{2,k}),$$

where $\zeta_{1,k} = g_1(x_k, y_k; \theta_1,k) - \nabla_i f(x_k, y_k)$ and $\xi_{i,k} \sim \mathcal{N}(0, (\tau^2/d)I)$ for each player $i = 1, 2$. Taking the difference between the coupling sequences $\{z_k\} = \{x_k, y_k\}$ and $\{z'_k\} = \{x'_k, y'_k\}$, we obtain

$$\hat{x}_{k+1} = x_{k+1} - x'_{k+1} = \hat{x}_k - \gamma (\nabla_1 f(x_k, y_k) - \nabla_1 f(x'_k, y'_k) + \zeta_{1,k} - \zeta'_{1,k} + \xi_{1,k} - \xi'_{1,k}),$$

$$\hat{y}_{k+1} = y_{k+1} - y'_{k+1} = \hat{y}_k + \gamma (\nabla_2 f(x_k, y_k) - \nabla_2 f(x'_k, y'_k) + \zeta_{2,k} - \zeta'_{2,k} + \tau^{-1}(\xi_{2,k} - \xi'_{2,k})).$$

Then, by the fundamental theorem of calculus and the mean value theorem, observe that

$$\left[ \begin{array}{c} \nabla_1 f(x_k, y_k) \\ -\tau \nabla_2 f(x_k, y_k) \end{array} \right] = \left[ \begin{array}{c} \nabla_1 f(x'_k, y'_k) \\ -\tau \nabla_2 f(x'_k, y'_k) \end{array} \right] + \int_0^1 J_\tau(\psi x_k + (1 - \psi)x'_k, \psi y_k + (1 - \psi)y'_k) d\psi \left[ \begin{array}{c} x_k - x'_k \\ y_k - y'_k \end{array} \right].$$

Define

$$\hat{z}_k = z_k - z'_k = (x_k - x'_k, y_k - y'_k), \hat{\zeta}_k = (\zeta_{1,k} - \zeta'_{1,k}, -\tau(\xi_{2,k} - \xi'_{2,k})), \hat{\xi}_k = (\xi_{1,k} - \xi'_{1,k}, -(\xi_{2,k} - \xi'_{2,k})).$$

Finally, let

$$\Delta_k = \int_0^1 J_\tau(\psi x_k + (1 - \psi)x'_k, \psi y_k + (1 - \psi)y'_k) d\psi - \mathcal{J}_\tau.$$
In this notation and using (56), we get that the system of equations in (54) and (55) is equivalently given by

\[
\hat{z}_{k+1} = \hat{z}_k - \gamma((\Delta_k + \mathcal{J}_\tau)\hat{z}_k + \hat{\zeta}_k + \hat{\xi}_k) \\
= (I - \gamma\mathcal{J}_\tau)\hat{z}_k - \gamma(\Delta_k\hat{z}_k + \hat{\zeta}_k + \hat{\xi}_k).
\]

We now show that the last expression can be simplified into the form

\[
\hat{z}_{k+1} = -\gamma \sum_{t=0}^{k} (I - \gamma\mathcal{J}_\tau)^{k-t}(\Delta_t\hat{z}_t + \hat{\zeta}_t + \hat{\xi}_t).
\]

To show this, let us simplify notation by defining \(A = (I - \gamma\mathcal{J}_\tau)\) and \(w_k = \Delta_k\hat{z}_k + \hat{\zeta}_k + \hat{\xi}_k\). In this notation, we get that

\[
\hat{z}_{k+1} = A\hat{z}_k - \gamma w_k
\]

\[
= A(A\hat{z}_{k-1} - \gamma w_{k-1}) - \gamma w_k = A^2\hat{z}_{k-1} - \gamma Aw_{k-1} - \gamma w_k
\]

\[
= A^2(A\hat{z}_{k-2} - \gamma w_{k-2}) - \gamma Aw_{k-1} - \gamma w_k = A^3\hat{z}_{k-2} - \gamma A^2w_{k-2} - \gamma Aw_{k-1} - \gamma w_k
\]

\[\vdots\]

\[
= A^{k+1}\hat{z}_0 - \gamma \sum_{t=0}^{k} A^{k-t}w_t.
\]

Hence, noting that \(\hat{z}_0 = (x_0 - x'_0, y_0 - y'_0) = 0\) by definition of the coupling sequence and returning to our standard notation, we arrive at the claimed result as follows:

\[
\hat{z}_{k+1} = A^{k+1}\hat{z}_0 - \gamma \sum_{t=0}^{k} A^{k-t}w_t = -\gamma \sum_{t=0}^{k} A^{k-t}w_t = -\gamma \sum_{t=0}^{k} (I - \gamma\mathcal{J}_\tau)^{k-t}(\Delta_t\hat{z}_t + \hat{\zeta}_t + \hat{\xi}_t).
\]

Now that we have defined the dynamics of the coupling sequence difference, we move toward boundedness.

**Lemma 24.** Let \(-\eta\) be the eigenvalue with the minimum real component among eigenvalues of \(J_\tau = J_\tau(x_0, y_0)\). Moreover, let \(\text{Re}(\eta)\) denote the real component of \(\eta\) and define the quantity

\[
\alpha_k = \sqrt{\frac{|1 + \gamma \eta|^{2k} - 1}{2\gamma\text{Re}(\eta) + \gamma^2|\eta|^2}}.
\]

Then, under the notation of Lemma 23 for any fixed \(k \geq 0\):

\[
\mathbb{P}\left(\|\hat{q}_{\xi,k}\| \leq \frac{2\gamma^\gamma \alpha_k \sqrt{t}}{\sqrt{d}}\right) \geq 1 - 2e^{-t}
\]

\[
\mathbb{P}\left(\|\hat{q}_{\xi,k}\| \geq \frac{\gamma^\gamma \alpha_k}{10\sqrt{d}}\right) \geq 2/3.
\]

**Proof.** We proceed in 3 steps. In step 1, we show that \(\|\hat{q}_{\xi,k}\|\) is equivalent to the absolute value of a Gaussian distributed scalar random variable. Then, in steps 2 and 3, we apply standard concentration and anti-concentration inequalities for Gaussian random variables to arrive at the final bounds.
Step 1. Recall that

\[ \hat{q}_{\xi,k} = \gamma \sum_{t=0}^{k-1} (I - \gamma J_{\tau})^{k-1-t} \hat{\xi}_t. \]  

(57)

where

\[ \hat{\xi}_t = \begin{bmatrix} \xi_{1,t} \\ -\xi_{2,t} \end{bmatrix} - \begin{bmatrix} \xi'_{1,t} \\ -\xi'_{2,t} \end{bmatrix} = \hat{\xi}_t - \hat{\xi}_t'. \]  

(58)

Moreover, \( \xi_{i,t}, \xi'_{i,t} \sim \mathcal{N}(0, (r^2/d)I) \) for each player \( i = \{1, 2\} \). Then, observe that both \( \hat{\xi}_t \) and \( \hat{\xi}_t' \) can be written as the sum a component projected onto the \( v_1 \) direction and a component projected onto the subspace complement of \( v_1 \). Recalling that \( P_{-1} \) denotes the projection onto the subspace complement of \( v_1 \), \( \hat{\xi}_t \) and \( \hat{\xi}_t' \) are equivalently given by

\[ \hat{\xi}_t = v_1 v_1^\top \hat{\xi}_t + P_{-1} \hat{\xi}_t \]

\[ \hat{\xi}_t' = v_1 v_1^\top \hat{\xi}_t' + P_{-1} \hat{\xi}_t'. \]  

(59)

The coupling sequence from Definition \[10\] is such that \( P_{-1} \hat{\xi}_t = P_{-1} \hat{\xi}_t' \) and \( v_1^\top \hat{\xi}_t \neq -v_1^\top \hat{\xi}_t' \). This fact, together with the forms of (58) and (59), imply that \( \hat{\xi}_t \) is equivalently given by

\[ \hat{\xi}_t = v_1 v_1^\top (\hat{\xi}_t - \hat{\xi}_t'). \]

Substituting this form of \( \hat{\xi}_t \) into (57) we obtain

\[ \hat{q}_{\xi,k} = \gamma \sum_{t=0}^{k-1} (I - \gamma J_{\tau})^{k-1-t} v_1 v_1^\top (\hat{\xi}_t - \hat{\xi}_t'). \]

Since \( v_1 \) is the eigenvector of \( J_{\tau} \) with eigenvalue \(-\eta\), it is also an eigenvector of \( I - \gamma J_{\tau} \) with eigenvalue \( 1 + \gamma \eta \). Hence, \( (I - \gamma J_{\tau}) v_1 = (1 + \gamma \eta) v_1 \), so that repeatedly applying this map gives \( (I - \gamma J_{\tau})^{k-1-t} v_1 = (1 + \gamma \eta)^{k-1-t} v_1 \). Relating this back to the previous equation, we obtain

\[ \hat{q}_{\xi,k} = \gamma \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} v_1 v_1^\top (\hat{\xi}_t - \hat{\xi}_t'). \]

Recall that our goal is to derive upper and lower bounds on the norm of \( \hat{q}_{\xi,k} \), which is now given by

\[ \| \hat{q}_{\xi,k} \| = \| \gamma \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} v_1 v_1^\top (\hat{\xi}_t - \hat{\xi}_t') \|. \]

Denote by \( Z_t = v_1^\top (\hat{\xi}_t - \hat{\xi}_t') \). Then by the homogeneity property of norms and the fact that \( \| v_1 \| = 1 \),

\[ \| \hat{q}_{\xi,k} \| = \| \gamma \left( \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} Z_t \right) v_1 \| = \| \gamma \left( \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} Z_t \right) \| v_1 \| = \| \gamma \left( \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} Z_t \right) \|. \]

The random variable \( Z_t \) is distributed as \( Z_t \sim \mathcal{N}(0, 2r^2/d) \) since \( \hat{\xi}_t - \hat{\xi}_t' \sim \mathcal{N}(0, 2r^2/d \cdot I) \) and \( \| v_1 \| = 1 \). Let

\[ W = \gamma \left( \sum_{t=0}^{k-1} (1 + \gamma \eta)^{k-1-t} Z_t \right) \]

and observe that

\[ W \sim \mathcal{N} \left( 0, \frac{2 \gamma r^2}{d} \sum_{t=0}^{k-1} |(1 + \gamma \eta)^{k-1-t}|^2 \right) \]

since \( W \) is the weighted sum of each \( Z_t \) which are independent and identically distributed\[9\]. Moreover, using the fact that \( |u^s| = |u|^s \) for any \( u \in \mathbb{C} \) and \( s \in \mathbb{R} \) to simplify the variance of \( W \), we obtain

\[ W \sim \mathcal{N} \left( 0, \frac{2 \gamma r^2}{d} \sum_{t=0}^{k-1} |1 + \gamma \eta|^{2(k-1-t)} \right). \]

---

\[8\] Recall that given a fixed vector \( w \) and a random vector \( y \) with covariance matrix \( \Sigma \), the variance of \( w^\top y \) is given by \( w^\top \Sigma w \).

\[9\] Note that given complex-valued weights \( w_1, \ldots, w_n \) and independent and identically distributed random variables \( X_1, \ldots, X_n \), then \( \text{Var} \left( \sum_{i=1}^{n} w_i X_i \right) = \text{Var} \left( X_1 \right) \sum_{i=1}^{n} |w_i|^2 \).
Computing the closed form of a geometric series, we get that
\[
\alpha_k^2 = \sum_{t=0}^{k-1} |1 + \gamma \eta|^{2(k-1-t)} = \sum_{t=0}^{k-1} |1 + \gamma \eta|^t = \frac{1 - |1 + \gamma \eta|^{2k}}{1 - |1 + \gamma \eta|^2} = \frac{|1 + \gamma \eta|^{2k} - 1}{|1 + \gamma \eta|^2 - 1} = \frac{|1 + \gamma \eta|^{2k} - 1}{2 \gamma \text{Re}(\eta) + \gamma^2 |\eta|^2}.
\]
Thus,
\[
W \sim \mathcal{N}(0, \frac{2 \gamma r^2 \alpha_k^2}{d}).
\]

**Step 2.** Recall that for random variable \(W \sim \mathcal{N}(0, \sigma^2)\), the following concentration inequality holds:
\[
P(|W| \leq \sigma \sqrt{2 \ln(2/\delta)}) \geq 1 - \delta.
\]
Now, since \(\|\hat{q}_{\xi,k}\| = |W|\) where \(W \sim \mathcal{N}(0, 2 \gamma r^2 \alpha_k^2 / d)\), when we apply the concentration inequality we get that
\[
P\left(\|\hat{q}_{\xi,k}\| \leq \frac{2 \gamma r \alpha_k \sqrt{\ln(2/\delta)}}{\sqrt{d}}\right) \geq 1 - \delta.
\]
Finally, selecting \(\delta = 2e^{-t}\) we arrived at the desired result given by
\[
P\left(\|\hat{q}_{\xi,k}\| \leq \frac{2 \gamma r \alpha_k \sqrt{t}}{\sqrt{d}}\right) \geq 1 - 2e^{-t}.
\]

**Step 3.** For a Gaussian random variable \(W \sim \mathcal{N}(0, \sigma^2)\),
\[
P(|W| \leq \delta \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\delta \sigma}^{\delta \sigma} e^{-w^2/(2\sigma^2)} dw < \frac{1}{\sqrt{2\pi}} \int_{-\delta \sigma}^{\delta \sigma} 1 dw = \frac{2\delta}{\sqrt{2\pi}} \leq \delta
\]
Hence,
\[
P(|W| \geq \sigma \delta) \geq 1 - \delta.
\]
Then, since \(\|\hat{q}_{\xi,k}\| = |W|\) where \(W \sim \mathcal{N}(0, 2 \gamma r^2 \alpha_k^2 / d)\), when we apply the anti-concentration inequality we get that
\[
P\left(\|\hat{q}_{\xi,k}\| \geq \frac{\sqrt{2} \gamma r \alpha_k \delta}{\sqrt{d}}\right) \geq 1 - \delta.
\]
So choosing \(\delta = \sqrt{2}/10\), we obtain
\[
P\left(\|\hat{q}_{\xi,k}\| \geq \frac{\gamma r \alpha_k}{10 \sqrt{d}}\right) \geq 1 - \sqrt{2}/10 \geq 2/3.
\]

We now show that there is a range of choices of learning rates such that \(\rho(I - \gamma J_\tau) = |1 + \gamma \eta|\) and \((1 + \gamma \eta)\) is the unique (simple) eigenvalue of the matrix \((I - \gamma J_\tau)\) obtaining the maximum modulus. This fact will be important for proving the result of Lemma 26.

**Lemma 25.** Let \(-\eta\) be the eigenvalue with the minimum real component among eigenvalues of \(J_\tau = J_\tau(x_0, y_0)\) and define the quantity
\[
\gamma^* = \min_{\lambda \in \text{spec}(J_\tau) \setminus \{-\eta\} : |\eta|^2 - |\lambda|^2 < 0} \frac{2 \text{Re}(-\eta) - \text{Re}(\lambda)}{|\eta|^2 - |\lambda|^2}.
\]
Then, for all \(\gamma \in (0, \gamma^*)\),
\[
\rho(I - \gamma J_\tau) = |1 + \gamma \eta|,
\]
and \((1 + \gamma \eta)\) is the eigenvalue of the matrix \((I - \gamma J_\tau)\) obtaining the maximum modulus.
Proof. Let the set of eigenvalues of the matrix $\mathcal{J}_\tau$ be denoted by
\[ \text{spec}(\mathcal{J}_\tau) = \{\lambda_1, \ldots, \lambda_d\}. \]
Similarly, the set of eigenvalues of the matrix $I - \gamma \mathcal{J}_\tau$ are given by
\[ \text{spec}(I - \gamma \mathcal{J}_\tau) = \{1 - \gamma \lambda_1, \ldots, 1 - \gamma \lambda_d\}. \]
The spectral radius of the matrix $(I - \gamma \mathcal{J}_\tau)$ in this notation is given by
\[ \rho(I - \gamma \mathcal{J}_\tau) = \max\{|1 - \gamma \lambda_1|, \ldots, |1 - \gamma \lambda_d|\}. \]
Recall that $-\eta$ denotes the eigenvalue in $\text{spec}(\mathcal{J}_\tau)$ with the minimum real component. In this notation, we need to prove that for all $\gamma \in (0, \gamma^*)$,
\[ |1 - \gamma \lambda| < |1 + \gamma \eta| \quad \forall \lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\} \]
and consequently it is immediate that
\[ \rho(I - \gamma \mathcal{J}_\tau) = |1 + \gamma \eta|. \]
To prove this, our approach will be to directly solve for the range of feasible choices of $\gamma$ for each $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$ such that
\[ |1 - \gamma \lambda|^2 = 1 - 2\gamma \text{Re}(\lambda) + \gamma^2 |\lambda|^2 < |1 + \gamma \eta|^2 = 1 - 2\gamma \text{Re}(\eta) + \gamma^2 |\eta|^2. \quad (60) \]
This will immediately imply an equivalent range of feasible choices of $\gamma$ for each $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$ such that
\[ |1 - \gamma \lambda| < |1 + \gamma \eta|. \]
Then, by taking the intersection of the feasible ranges of $\gamma$ such that the inequality above holds for each $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$, we obtain the stated result.
In particular, we prove this result as follows. To find the range of $\gamma$ such that (60) holds for some $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$, we solve for the conditions on $\gamma$ for which the following holds:
\[ (1 - 2\gamma \text{Re}(\lambda) + \gamma^2 |\lambda|^2) - (1 - 2\gamma \text{Re}(\eta) + \gamma^2 |\eta|^2) = \gamma (2\text{Re}(\lambda) - \text{Re}(\lambda)) + \gamma(|\lambda|^2 - |\eta|^2)) < 0. \]
This problem reduces to finding the conditions on $\gamma$ such that for $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$:
\[ 2(\text{Re}(\lambda) - \text{Re}(\lambda)) + \gamma(|\lambda|^2 - |\eta|^2) < 0. \quad (61) \]
We solve the problem in (61) dependent on the sign of $|\eta|^2 - |\lambda|^2$.
\begin{itemize}
  \item \textbf{Case 1.} Suppose that $|\eta|^2 - |\lambda|^2 = 0$. Then, the inequality in (61) is trivially satisfied by any $\gamma$ since $\text{Re}(\eta) < \text{Re}(\lambda)$.
  \item \textbf{Case 2.} Suppose that $|\eta|^2 - |\lambda|^2 > 0$. Then, solving for $\gamma$ in (61) results in the condition
    \[ \gamma > 2 \frac{\text{Re}(\lambda) - \text{Re}(\lambda)}{|\eta|^2 - |\lambda|^2}. \]
    This inequality is satisfied by any $\gamma \in (0, \infty)$ since $\text{Re}(\eta) < 0$ and $\text{Re}(\eta) < \text{Re}(\lambda)$ so that $\text{Re}(\eta) - \text{Re}(\lambda) < 0$ and $2(\text{Re}(\eta) - \text{Re}(\lambda))/(|\eta|^2 - |\lambda|^2) < 0$.
  \item \textbf{Case 3.} Suppose that $|\eta|^2 - |\lambda|^2 < 0$. Then, solving for $\gamma$ in (61) results in the condition
    \[ \gamma < 2 \frac{\text{Re}(\lambda) - \text{Re}(\lambda)}{|\eta|^2 - |\lambda|^2}. \]
    There exists a range of $\gamma > 0$ such that this condition is satisfied since $\text{Re}(\eta) < 0$ and $\text{Re}(\eta) < \text{Re}(\lambda)$ so that $\text{Re}(\eta) - \text{Re}(\lambda) < 0$ and $2(\text{Re}(\eta) - \text{Re}(\lambda))/(|\eta|^2 - |\lambda|^2) > 0$.
\end{itemize}
Combining the cases, we find that for any particular $\lambda \in \text{spec}(\mathcal{J}_\tau) \setminus \{-\eta\}$, the inequality
\[ |1 - \gamma \lambda|^2 < |1 + \gamma \eta|^2 \]
holds for $\gamma \in (0, \infty)$ when $|\eta|^2 - |\lambda|^2 \geq 0$ and $\gamma \in (0, 2(\text{Re}(-\eta) - \text{Re}(\lambda))/(|\eta|^2 - |\lambda|^2))$ when $|\eta|^2 - |\lambda|^2 < 0$. Moreover, observe that since $|a|^2 < |b|^2$ if and only if $|a| < |b|$ for $a, b \in \mathbb{C}$ this also implies for any particular $\lambda \in \text{spec}(\mathcal{F}_r) \setminus \{-\eta\}$,

$$[1 - \gamma \lambda] < [1 + \gamma \eta]$$

(62)

for $\gamma \in (0, \infty)$ when $|\eta|^2 - |\lambda|^2 \geq 0$ and $\gamma \in (0, 2(\text{Re}(-\eta) - \text{Re}(\lambda))/(|\eta|^2 - |\lambda|^2))$ when $|\eta|^2 - |\lambda|^2 < 0$.

By taking the intersection of the feasible ranges of $\gamma$ such that the inequality in (62) holds for each particular $\lambda \in \text{spec}(\mathcal{F}_r) \setminus \{-\eta\}$, we get that

$$[1 - \gamma \lambda] < [1 + \gamma \eta] \quad \forall \lambda \in \text{spec}(\mathcal{F}_r) \setminus \{-\eta\}$$

for all $\gamma \in (0, \gamma^*)$ where

$$\gamma^* = \min_{\lambda \in \text{spec}(\mathcal{F}_r) \setminus \{-\eta\}} \frac{2 \text{Re}(\eta) - \text{Re}(\lambda)}{|\eta|^2 - |\lambda|^2}.$$ 

Hence, for all $\gamma \in (0, \gamma^*)$,

$$\rho(I - \gamma \mathcal{F}_r) = [1 + \gamma \eta],$$

and $(1 + \gamma \eta)$ is the eigenvalue of the matrix $(I - \gamma \mathcal{F}_r)$ obtaining the maximum modulus, which is simple by assumption.

**Lemma 26.** Consider the notation and assumptions of Lemma 21. Define the events

\[
\begin{align*}
E_1 &= \left\{ \min \{ f(z_T) - f(z_0), f(z_T') - f(z_0) \} \leq -\mathcal{F} \right\}, \\
E_2 &= \{ \forall k \leq T : \max \{ \|z_k - z_0\|, \|z_k' - z_0\| \} \leq S^2 \}, \\
E_3 &= \{ \forall k \leq T, \|\hat{q}_{z,k} + \hat{q}_{\xi,k}\| \leq \frac{2\tau_{\alpha,k}}{20\sqrt{d}} \}. 
\end{align*}
\]

Then,

$$\mathbb{P}(E_3|E_1^c \cap E_2) \geq 1 - \delta_2 \quad \text{and} \quad \mathbb{P}(E_1 \cup E_3) \geq 1 - \delta_1 - \delta_2$$

(63)

where $\delta_1 = 32dT e^{-t}$ and $\delta_2 = 10dT^2 \log(S\sqrt{d}/(\gamma r)) e^{-t}$.

**Proof.** We begin this proof by recalling some of our notation and given an overview of the analysis method.

**Notation and Setup.** Consider the coupling sequences $\{z_k\} = \{x_k, y_k\}$ and $\{z_k'\} = \{x_k', y_k\}$ as given in Definition 10. Following the notation of Lemma 23, let $\hat{z}_k = z_k - z_k'$, $\xi_k = (\zeta_{1,k} - \zeta_{1,k}', -(\zeta_{2,k} - \zeta_{2,k}'))$, and $\hat{\xi}_k = (\xi_{1,k} - \xi_{1,k}', -(\xi_{2,k} - \xi_{2,k}'))$. Recall from Lemma 23 that

$$\hat{z}_k = -\hat{q}_{z,k} - \hat{q}_{\xi,k} - \hat{q}_{\xi,k},$$

where

$$\hat{q}_{z,k} = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{F}_r)^{k-1-t} \Delta_t \hat{z}_t, \quad \hat{q}_{\xi,k} = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{F}_r)^{k-1-t} \hat{\xi}_t, \quad \hat{q}_{\xi,k} = \gamma \sum_{t=0}^{k-1} (I - \gamma \mathcal{F}_r)^{k-1-t} \hat{\xi}_t, \quad$$

(64)

and

$$\Delta_k = \int_0^1 J_r(\psi x_k + (1 - \psi)x_k', \psi y_k + (1 - \psi)y_k')d\psi - \mathcal{F}_r.$$ 

**Proof Approach.** Our proof approach will be to show that

$$\mathbb{P}(E_3|E_1^c \cap E_2) \geq 1 - 10dT^2 \log(S\sqrt{d}/(\gamma r)) e^{-t} := 1 - \delta_2.$$ 

(65)

Then, combined with the fact from Lemma 22 that

$$\mathbb{P}(E_1 \cup E_2) \geq \mathbb{P}(E_2|E_1^c) \geq 1 - 32dT e^{-t} := 1 - \delta_1,$$

(66)
we can near immediately conclude
\[ \mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_1) \geq 1 - \delta_1 - \delta_2. \]  
(67)

In particular, given the inequalities from (65) and (66), the inequality in (67) can be obtained via the following steps that are explained below:

\[
\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3) = \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
= \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2') + \mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 + \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
\geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2') + \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
= \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2') + \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
\geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2') + (1 - \delta_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
= \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2') + \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) - \delta_2\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
= \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) - \delta_2\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \\
\geq 1 - \delta_1 - \delta_2. \\
(75)

Observe that (68) follows from the law of total probability and then (68) is obtained by simplifying the expression using the facts that \( \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) = 1 \) and \( \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) = 1 \). The inequality in (70) uses that \( \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 0 \) and (71) holds since \( \mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) \). The inequality in (72) is a result of applying the assumed inequality \( \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \delta_2 \) from (65). Finally, (73) and (74) are direct simplifications and the last inequality in (75) holds by applying the bound \( \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) \geq 1 - \delta_1 \) from (66) and using that \( \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \leq 1 \).

Thus, we can now focus on showing that
\[ \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \delta_2 \]  
(76)

and then the previous analysis will then allow us to immediately conclude that \( \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_3) \geq 1 - \delta_1 - \delta_2 \).

**Inductive Argument.** We prove the statement in (76) using induction. Let us define the relevant events in terms of the final time index \( t \) as follows:

\[ \mathcal{E}_1(t) = \left\{ \min \left\{ f(z_t) - f(z_0), f(z'_t) - f(z_0) \right\} \leq -\mathcal{F}_{\text{st}} \right\}, \]

\[ \mathcal{E}_2(t) = \left\{ \forall k \leq t : \max \left\{ \|z_k - z_0\|_2^2, \|z'_k - z_0\|_2^2 \right\} \leq \mathcal{S}_2^2 \right\}, \]

\[ \mathcal{E}_3(t) = \left\{ \forall k \leq t, \|q_{z,k} + \hat{q}_{z,k}\| \leq \frac{\gamma r \alpha_k}{20 \sqrt{d}} \right\}. \]

Observe that \( \mathcal{E}_1(T) = \mathcal{E}_1, \mathcal{E}_2(T) = \mathcal{E}_2, \) and \( \mathcal{E}_3(T) = \mathcal{E}_3 \). Similarly, let \( \delta_2(t) = 10dT \log(S \sqrt{d}/(\gamma r))e^{-t} \) and observe that, by a slight abuse of notation, \( \delta_2(T) = \delta_2 \). The claim we prove by induction is that \( \mathbb{P}(\mathcal{E}_3(T)|\mathcal{E}_1(t) \cap \mathcal{E}_2(t)) \geq 1 - \delta_2(T) \).

**Base Case.** For the base case, we consider \( t = 0 \) and show that given the events \( \{\mathcal{E}_1(0) \cap \mathcal{E}_2(0)\} \), then \( \mathcal{E}_3(0) \) holds with probability one. Indeed, this trivially holds since \( \hat{q}_{z,0} = 0 \) and \( \hat{q}_{z,0} = 0 \) so that

\[ \|\hat{q}_{z,0} + \hat{q}_{z,0}\| = 0 \leq \frac{\gamma r \alpha_0}{20 \sqrt{d}}. \]

Thus, \( \mathbb{P}(\mathcal{E}_3(0)|\mathcal{E}_1(0) \cap \mathcal{E}_2(0)) \geq 1 - \delta_2(0) \).

**Inductive Hypothesis.** For our inductive hypothesis, we consider any iterate \( 0 < t < T \). Assume that

\[ \mathbb{P}(\mathcal{E}_3(t)|\mathcal{E}_1(t) \cap \mathcal{E}_2(t)) \geq 1 - \delta_2(t). \]
Toward proving the inductive step that follows, we derive implications of the inductive hypothesis. To begin, define the following event in terms of the final time index $t$:

$$
\mathcal{E}_4(t) = \{ \forall k \leq t, \|\hat{q}_{\xi,k}\| \leq \frac{2\gamma r\alpha k \sqrt{t}}{\sqrt{d}} \}.
$$

Recall that by Lemma 24 for any fixed $k \geq 0$,

$$
\mathbb{P}\left( \|\hat{q}_{\xi,k}\| \leq \frac{2\gamma r\alpha k \sqrt{t}}{\sqrt{d}} \right) \geq 1 - 2e^{-t}.
$$

Thus, by a union bound it immediately follows that

$$
\mathbb{P}\left( \forall k \leq t, \|\hat{q}_{\xi,k}\| \leq \frac{2\gamma r\alpha k \sqrt{t}}{\sqrt{d}} \right) \geq 1 - 2te^{-t}. \tag{77}
$$

Furthermore, observe that on the event that both $\mathcal{E}_3(t)$ and the event in (77) hold, then by the triangle inequality, for all $0 \leq k \leq t$:

$$
\|\hat{z}_k\| = \|\hat{q}_{z,k} - \hat{q}_{\xi,k} - \hat{q}_{\xi,k}\| \leq \|\hat{q}_{z,k} + \hat{q}_{\xi,k}\| + \|\hat{q}_{\xi,k}\| \leq \frac{\gamma r\alpha_k}{20\sqrt{d}} + \frac{2\gamma r\alpha_k \sqrt{t}}{\sqrt{d}} \leq \frac{c\gamma r\alpha_k \sqrt{t}}{\sqrt{d}}.
$$

Thus, defining the event

$$
\mathcal{E}_4(t) = \{ \forall k \leq t, \|\hat{z}_k\| \leq \frac{c\gamma r\alpha_k \sqrt{t}}{\sqrt{d}} \},
$$

by a union bound we also have that under the inductive hypothesis:

$$
\mathbb{P}(\mathcal{E}_3(t) \cap \mathcal{E}_4(t) | \mathcal{E}_1(t) \cap \mathcal{E}_2(t)) \geq 1 - \delta_2(t) - 2te^{-t}.
$$

**Inductive Step.** Using the inductive hypothesis and its implications, we now prove that

$$
\mathbb{P}(\mathcal{E}_3(t + 1) | \mathcal{E}_1(t + 1) \cap \mathcal{E}_2(t + 1)) \geq 1 - \delta_2(t + 1).
$$

Our high level approach will be to show that on the event that $\mathcal{E}_3(t) \cap \mathcal{E}_4(t)$ holds and conditioned on $\mathcal{E}_1(t + 1) \cap \mathcal{E}_2(t + 1)$, then with high probability

$$
\|\hat{q}_{z,t+1}\| \leq \frac{\gamma r\alpha_{t+1}}{40\sqrt{d}} \quad \text{and} \quad \|\hat{q}_{\xi,t+1}\| \leq \frac{\gamma r\alpha_{t+1}}{40\sqrt{d}}.
$$

The final result will then follow from the triangle inequality and a union bound.

**Norm Construction.** Before diving into the bounds, we give provide some details and notation for a technical component of the remainder of the proof, specifically regarding the norm constructions from Appendix B.4.2. In particular, by Lemma 25 $(1 + \gamma\eta)$ is the unique eigenvalue of the matrix $(I - \gamma J_r)$ obtaining the maximum modulus. By Assumption 6, the eigenvalue $-\eta$ is simple so that the eigenvalue $(1 + \gamma\eta)$ is also simple. Therefore, the conditions of Lemma 9 are satisfied to ensure that there exists a matrix norm $\| \cdot \|_*$ such that $\rho(I - \gamma J_r) = \| I - \gamma J_r \|_*$ and then Lemma 25 gives that $\rho(I - \gamma J_r) = |1 + \gamma\eta|$. Specifically, from Lemma 9 there exists a non-singular matrix $M \in \mathbb{C}^{n \times n}$ such that $\| I - \gamma J_r \|_* = \| M(I - \gamma J_r)M^{-1} \|$ where $\| \cdot \|$ denotes the induced operator 2-norm. Moreover, the matrix norm $\| \cdot \|_*$ is induced by the vector norm $\| \cdot \|_*$ such that for $x \in \mathbb{C}^n$, $\| x \|_* = \| Mx \|$, where $\| \cdot \|$ denotes the vector 2-norm. In the remainder of the proof, the matrix $M$ denotes the matrix from this norm construction and $\sigma_{\min}(M)$ and $\sigma_{\max}(M)$ denote its minimum and maximum singular values, respectively. We now proceed to bound $\|\hat{q}_{\xi,t+1}\|$ and $\|\hat{q}_{\xi,t+1}\|$ before combining the bounds to reach the conclusion.
Bounding $\|q_{z,t+1}\|$. Assume that the event $E_3(t) \cap E_4(t)$ holds and recall that we are already conditioning on the event $E_2(t + 1) \cap E_2(t + 1)$. To bound this quantity we need the norm construction from Lemma 26. In particular, given the construction from Lemma 26, we can upper bound $\|q_{z,t+1}\|$ by $\sigma_{\min}(M)^{-1} \|q_{z,t+1}\|$ where the matrix $M$ comes from the norm construction. In particular, using Lemma 12, we have

$$\|q_{z,t+1}\| = \frac{\sigma_{\min}(M)}{\sigma_{\min}(M)} \|q_{z,t+1}\| \leq \frac{1}{\sigma_{\min}(M)} \|Mq_{t,k+1}\| = \frac{1}{\sigma_{\min}(M)} \|q_{t,k+1}\|$$  (78)

Moreover, by the definition of $q_{z,k+1}$ from (64), the triangle inequality, and the fact from Lemma 10 that the vector norm $\|\cdot\|$ is subordinate, we get the following bound:

$$\|q_{z,t+1}\| = \|\sum_{t=0}^{\infty} (I - \gamma J^t \Delta_k \hat{z}_k)\| \leq \gamma \sum_{k=0}^{\infty} \|(I - \gamma J^t \Delta_k \hat{z}_k)\| \leq \gamma \sum_{k=0}^{\infty} \|(I - \gamma J^t \Delta_k \hat{z}_k)\|_\ast.$$  (79)

We now provide bounds on both $\|(I - \gamma J^t \Delta_k \hat{z}_k)\|_\ast$ and $\|\Delta_k \hat{z}_k\|_\ast$ that hold for any $0 \leq k \leq t$. In particular, using the fact from Lemma 11 that the matrix norm $\|\cdot\|_\ast$ is submultiplicative, along with Lemma 9, we have that for any $0 \leq k \leq t$,

$$\|(I - \gamma J^t \Delta_k \hat{z}_k)\|_\ast \leq \|(I - \gamma J^t \Delta_k \hat{z}_k)\|_\ast = \rho(I - \gamma J^t \Delta_k \hat{z}_k) = \|1 + \gamma\|^{-k}.$$  (80)

Toward bounding $\|\Delta_k \hat{z}_k\|_\ast$ for all $0 \leq k \leq t$, recall that we have assumed that $E_4(t)$ holds so for all $0 \leq k \leq t$ we have

$$\|\hat{z}_k\| \leq \frac{c\gamma R\alpha_k}{\sqrt{d}}.$$  (81)

Now, using the fact we are conditioning on the event $E_2(t + 1)$ so that $\max\{\|z_k - z_0\|^2, \|z'_k - z_0\|^2\} \leq S^2$ for all $0 \leq k \leq t$ along with the Hessian Lipschitz assumption, we have that for all $0 \leq k \leq t$,

$$\|\Delta_k\| \leq \|\int_0^1 J^\tau_\rho(x_k + (1 - \rho)x'_k, \psi y_k + (1 - \rho)\psi' y'_k) d\rho - J\| \leq \beta \max\{\|z_k - z_0\|, \|z'_k - z_0\|\} \leq \beta S.$$  (82)

Furthermore, by definition of the vector norm $\|\cdot\|_\ast$, Lemma 12, the fact that the vector norm $\|\cdot\|$ is subordinate, and (81) and (82), we have that for all $0 \leq k \leq t$,

$$\|\Delta_k \hat{z}_k\|_\ast = \|M \Delta_k \hat{z}_k\|_\ast \leq \sigma_{\max}(M) \|\Delta_k \hat{z}_k\|_\ast \leq \sigma_{\max}(M) \|\Delta_k\| \|\hat{z}_k\| \leq \frac{c\gamma \sigma_{\max}(M) \beta S \sqrt{\gamma}}{\sqrt{d}}$$  (83)

Thus, combining (78), (79), (80), and (83), we have

$$\|q_{z,t+1}\| \leq \frac{c\gamma^2 \sigma_{\max}(M) \beta S \sqrt{\gamma}}{\sigma_{\min}(M) \sqrt{d}} \sum_{k=0}^{t} |1 + \gamma\|^{-k} \alpha_k.$$  (84)

By the definition of $\alpha_k$ and using the fact that $|1 + \gamma\| > 1$, we have

$$\sum_{k=0}^{t} |1 + \gamma\|^{-k} \alpha_k = \frac{1}{\sqrt{2\gamma} \Re(\eta) + \gamma^2 |\eta|^2} \sum_{k=0}^{t} |1 + \gamma\|^{-k} \sqrt{|1 + \gamma\|^{2k} - 1}
= \frac{1}{\sqrt{2\gamma} \Re(\eta) + \gamma^2 |\eta|^2} \sum_{k=0}^{t} \sqrt{|1 + \gamma\|^{2k} - 1}
\leq \frac{1}{\sqrt{2\gamma} \Re(\eta) + \gamma^2 |\eta|^2} \sum_{k=0}^{t} \sqrt{|1 + \gamma\|^{2k} - 1}
= (t + 1) \alpha_t \leq T \alpha_{t+1}.$$  (85)
Hence, returning to bounding \(|\hat{q}_{z,t+1}\|\), using the previous inequality and step explained below, we get that
\[
|\hat{q}_{z,t+1}| \leq \frac{c_{\alpha}t + 1 \gamma^2 \sigma_{\max}(M) \beta S \sqrt{d}}{\sigma_{\min}(M) \sqrt{d}} \leq \frac{c_{\alpha}t + 1 \gamma \sigma_{\max}(M) \sqrt{T}}{\sigma_{\min}(M) \sqrt{T d}} \leq \frac{\alpha_{t+1} \gamma^r}{40 \sqrt{d}}.
\]
The final inequalities are a result of our choice of constants and using that with a sufficiently large constant in \(t\) we have
\[
\gamma \beta S T = \gamma \beta \left(\frac{2}{\beta^2} \sqrt{\frac{K}{\beta}}\right) \left(\frac{t}{\gamma \beta^2}\right) = \frac{2 \sqrt{K}}{t} \quad \text{and} \quad \frac{c_{\sigma_{\max}(M) \sqrt{T}}}{\sigma_{\min}(M) \sqrt{T d}} \leq \frac{1}{40}.
\]
Thus we have shown
\[
|\hat{q}_{z,t+1}| \leq \frac{\alpha_{t+1} \gamma^r}{40 \sqrt{d}}
\]
deterministically on the event \(E_3(t) \cap E_4(t)\) holds and conditioning on the event \(E_1(t + 1) \cap E_2(t + 1)\).

**Bounding \(\|\hat{q}_{t,t+1}\|\).** To begin, we only assume the event \(E_1(t + 1) \cap E_2(t + 1)\) that is being conditioned on. By definition, the quantity we need to bound is
\[
\|\hat{q}_{t,t}|| = \|\sum_{k=0}^{t} (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k ||.
\]
We bound this using Lemma 16. To invoke Lemma 16, we show that for each \(0 \leq k \leq t\), \(\gamma (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k\) is norm-subGaussian. Observe that \(\hat{\zeta}_k\) is a \(\tilde{\ell}_{\alpha} ||\hat{\zeta}||\)-norm-subGaussian random vector where \(\tilde{\ell}_1 + \tilde{\ell}_2 \leq \ell\) by Assumptions 4 and 5. By Definition 9 this means that
\[
\mathbb{P}(||\hat{\zeta}_t - \mathbb{E}[\hat{\zeta}_t]]|| \geq \nu) \leq 2 \exp \left(-\frac{-\nu^2}{2(\tilde{\ell}_\alpha ||\hat{\zeta}||)^2}\right) \quad \forall \nu \in \mathbb{R}. \quad (84)
\]
Toward showing that \(\gamma (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k\) is norm-subGaussian for all \(0 \leq k \leq t\), observe that using Lemma 12 the definition of the norm construction from Lemma 9 and (80) we have that
\[
\gamma (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k = \gamma \sigma_{\min}(M) \gamma_{\max}(M)^{-1} ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]||
\]
\[
\leq \gamma \sigma_{\min}(M) ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]||
\]
\[
\leq \gamma \sigma_{\min}(M)^{-1} ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]||
\]
\[
\leq \gamma \sigma_{\min}(M)^{-1} ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]||
\]
\[
\leq \gamma \sigma_{\max}(M) \sigma_{\min}(M)^{-1} ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]||
\]
Thus, combining (84) and (85), we get that for \(0 \leq k \leq t\),
\[
\mathbb{P}(\gamma (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]|| \geq \nu) \leq \mathbb{P}(\gamma \sigma_{\max}(M) \sigma_{\min}(M)^{-1} ||(I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k - \mathbb{E}[\hat{\zeta}_k]|| \geq \nu)
\]
\[
\leq 2 \exp \left(-\frac{-\nu^2}{2(\gamma \sigma_{\max}(M) \sigma_{\min}(M)^{-1} ||(1 + \gamma \eta)^{l-k} \hat{\zeta}_k||)^2}\right).
\]
Thus, for all \(0 \leq k \leq t\), \(\gamma (I - \gamma \mathcal{J}_t)^{l-k} \hat{\zeta}_k\) is norm-subGaussian with \(\sigma_k = \gamma \sigma_{\max}(M) \sigma_{\min}(M)^{-1} ||(1 + \gamma \eta)^{l-k} \hat{\zeta}_k||\).
Now observe that for all $0 \leq k \leq t$,
\[
\|\hat{z}_k\| = \|z_k - z'_k\| = \|(z_k - z_0) - (z_0 - z'_k)\| \leq \|z_k - z_0\| + \|z_0 - z'_k\| \leq 2 \max\{\|z_k - z_0\|, \|z_0 - z'_k\|\}.
\]
Thus, using the fact that we are conditioning on the event $E_2(t + 1)$ so that $\max\{\|z_k - z_0\|^2, \|z'_k - z_0\|^2\} \leq S^2$ for all $0 \leq k \leq t$ we have,
\[
\|\hat{z}_k\|^2 \leq 4 \max\{\|z_k - z_0\|^2, \|z_0 - z'_k\|^2\} \leq 4S^2.
\]
Thus, we have that the sum of the norm-subGaussian parameters are given and bounded as follows:
\[
\sum_{k=0}^{t} \sigma_k^2 = \gamma^2 \tilde{\ell}^2 \sigma_{\max}(M)^2 \sigma_{\min}(M)^{-2} \sum_{k=0}^{t} |1 + \gamma \eta|^2(t-k) \|\hat{z}_k\|^2 \leq 4 \gamma^2 \tilde{\ell}^2 \sigma_{\max}(M)^2 \sigma_{\min}(M)^{-2} \sum_{k=0}^{t} |1 + \gamma \eta|^2(t-k) S^2
\]
\[
= 4 \gamma^2 \tilde{\ell}^2 \sigma_{\max}(M)^2 \sigma_{\min}(M)^{-2} t \alpha_{k+1}^2 S^2.
\]
(86)
We now define
\[
B = 4 \gamma^2 \tilde{\ell}^2 \sigma_{\max}(M)^2 \sigma_{\min}(M)^{-2} t \alpha_{k+1}^2 S^2
\]
and
\[
b = \frac{4 \gamma^2 \tilde{\ell}^2 r^2 \sigma_{\max}(M)^2 \sigma_{\min}(M)^{-2} \alpha_{t+1}^2}{d}.
\]
(87)
Finally, by definition and applying Lemma 16 with $B$ and $b$ defined in (88) and (89), with probability at least $1 - 4d \log(S\sqrt{d}/(\gamma r)) e^{-t}$,
\[
\|\hat{\zeta}_{t+1}\| = \|\gamma \sum_{k=0}^{t} (I - \gamma J_r)^{t-k} \hat{\zeta}_k\|
\]
\[
\leq c \sqrt{\max\{\sum_{k=0}^{t} \sigma_k^2, b\} t}
\]
\[
= c \gamma \tilde{\ell} \sigma_{\max}(M) \sigma_{\min}(M)^{-1} \sqrt{t} \sqrt{\max\{\sum_{k=0}^{t} |1 + \gamma \eta|^2(t-k) \|\hat{z}_k\|^2, \frac{\gamma^2 r^2 \alpha_{t+1}^2}{d}\}}.
\]
We denote this event (the concentration inequality holding) as $E_5$ and note that it holds with probability at least $1 - 4d \log(S\sqrt{d}/(\gamma r)) e^{-t}$ conditioned on $E_2(t + 1) \cap E_2(t + 1)$. Continuing on, assume that the event $E_3(t) \cap E_4(t)$ holds and recall that we are already conditioning on the event $E_2(t + 1) \cap E_2(t + 1)$. Then, we have
\[
\|\hat{\zeta}_{t+1}\| \leq c \gamma \tilde{\ell} \sigma_{\max}(M) \sigma_{\min}(M)^{-1} \sqrt{t} \sqrt{\max\{\frac{\gamma^2 r^2 \alpha_{t+1}^2}{d}, \frac{\gamma^2 r^2 \alpha_{t+1}^2}{d}\}}
\]
\[
= \frac{c \gamma^2 \tilde{\ell} \sigma_{\max}(M) \sigma_{\min}(M)^{-1} \alpha_{t+1} \sqrt{t}}{\sqrt{d}}
\]
\[
\leq \frac{\alpha_{t+1} \gamma r}{40 \sqrt{d}}.
\]
(90)
(91)
Observe that (90) holds since we have assumed that $E_4(t)$ holds so for all $0 \leq k \leq t$ we have
\[
\|\hat{z}_k\| \leq \frac{c \gamma \alpha_k \sqrt{t}}{\sqrt{d}}.
\]
(93)
Furthermore, we obtain (91) using the following bound that follows from the definition of $\alpha_t$ and the fact
that \(|1 + \gamma \eta| > 1\):

\[
\sum_{k=0}^{t} |1 + \gamma \eta|^{2(t-k)} \alpha_k^2 = \frac{1}{2} \sum_{k=0}^{t} |1 + \gamma \eta|^{2(t-k)} (|1 + \gamma \eta|^{2k} - 1) = \frac{1}{2} \sum_{k=0}^{t} (|1 + \gamma \eta|^{2t} - |1 + \gamma \eta|^{2t-2k}) \leq \frac{1}{2} \sum_{k=0}^{t} (|1 + \gamma \eta|^{2t} - 1) = (t+1) \alpha_t^2 \leq T \alpha_{t+1}^2.
\]

To see how \((\ref{eq:alpha-t})\) was obtained, observe that by our choice of constants we have

\[
\gamma \ell \sqrt{T} = \gamma (\tilde{\ell} t) \left( \frac{1}{\sqrt{\eta/(\beta \varepsilon)^{1/4}}} \right) = \sqrt{\gamma} (\tilde{\ell} t) \left( \frac{1}{(\beta \varepsilon)^{1/4}} \right) = \left( \frac{1}{\ell \eta^{9/4} \sqrt{\mu \beta \varepsilon}} \right) \left( \frac{1}{\sqrt{T}} \right).
\]

Now, using that

\[
R := 1 + \max \left\{ \frac{\max\{\sigma_1^2, \sigma_2^2\}}{\varepsilon^2}, \frac{\ell^2}{\ell \mu \sqrt{\beta \varepsilon}} \right\} \geq \frac{\ell^2}{\ell \mu \sqrt{\beta \varepsilon}}
\]

we get the bound

\[
\gamma \ell \sqrt{T} \leq \left( \frac{\tilde{\ell}}{\ell \eta^{9/4} \sqrt{\mu \beta \varepsilon}} \right) \left( \frac{1}{\sqrt{R}} \right) \leq \left( \frac{\sqrt{\mu \beta \varepsilon}^{1/4}}{\tilde{\ell}} \right) \left( \frac{\sqrt{\mu \beta \varepsilon}^{1/4}}{\tilde{\ell}} \right) = \left( \frac{1}{\ell \eta^{9/4}} \right).
\]

Thus, using that with a sufficiently large constant in \(\ell\) we have

\[
\frac{c \gamma^2 \ell \sigma_{\max}(M) \sigma_{\min}(M)^{-1} \alpha_{t+1} \sqrt{T}}{\sqrt{d}} \leq \alpha_{t+1} \gamma r \leq \frac{c \gamma r \sigma_{\max}(M) \sigma_{\min}(M)^{-1}}{d \ell^{7/2}} = \frac{\alpha_{t+1} \gamma r}{40 \sqrt{d}}.
\]

Thus we have shown that the event \(E_5\) holds with probability at least \(1 - 4d \log(S \sqrt{d}/(\gamma r)) e^{-t}\) conditioning on the event \(E_1(t + 1) \cap E_2(t + 1)\) and

\[
\|\hat{q}_{z,t+1}\| \leq \frac{\alpha_{t+1} \gamma r}{40 \sqrt{d}}
\]

deterministically on the event \(E_3(t) \cap E_4(t)\) also holds.

**Combining Bounds.** Recall that we showed under the inductive hypothesis

\[
P(E_3(t) \cap E_4(t) | E_1(t) \cap E_2(t)) \geq 1 - \delta_2(t) - 2e^{-t}.
\]

We also have that

\[
P(E_3(t) \cap E_4(t) | E_1(E_1(t + 1) \cap E_2(t + 1)) \geq 1 - \delta_2(t) - 2e^{-t} - 4d \log(S \sqrt{d}/(\gamma r)) e^{-t} \geq 1 - \delta_2(t + 1).
\]

Moreover, on this event holding, we showed that it must be the case that

\[
\|\hat{q}_{z,t+1}\| \leq \frac{\alpha_{t+1} \gamma r}{40 \sqrt{d}} \text{ and } \|\hat{q}_{z,t+1}\| \leq \frac{\alpha_{t+1} \gamma r}{40 \sqrt{d}}.
\]

Thus, with probability at least \(1 - \delta_2(t + 1)\), we get

\[
\|\hat{q}_{z,t+1} + \hat{q}_{z,t+1}\| \leq \|\hat{q}_{z,t+1}\| + \|\hat{q}_{z,t+1}\| \leq \frac{\alpha_{t+1} \gamma r}{20 \sqrt{d}}.
\]

This finishes the inductive proof. Moreover, by the steps at the beginning of the proof, proves the claim of the result.
Lemma 21 (Escaping Saddle Points). Given Assumptions [19], there exists an absolute constant $c$ such that, for any fixed $k_0 > 0$, $\nu > \max \log(\ell \sqrt{d}/(\beta \epsilon))$, if $\eta, \tau, F, K, T$ are chosen as in Table 2 and $z_{t_0}$ satisfies $\|\nabla f(z_{t_0}), \nabla^2 f(z_{t_0})\| \leq \epsilon$ and $\text{Re}(\lambda_{\min}(J_{r}(z_{t_0}))) \leq -\sqrt{\beta \epsilon}$, then the sequence of iterates generated by $\tau$-PGDA (Algorithm 3) satisfies the following:

$$\mathbb{P}(f(z_{t_0} + T) - f(z_{t_0}) \leq 0.1 F) \geq 1 - 8e^{-\nu}$$

(43)

and

$$\mathbb{P}(f(z_{t_0} + T) - f(z_{t_0}) \leq -F) \geq \frac{1}{3} - 5d T^2 \log(8\sqrt{d}/(\gamma T)) e^{-\nu} - 32d T e^{-\nu}.$$  

(44)

Proof of Lemma 21. We prove that each of the claimed bounds, (43) and (44), hold separately.

Claim 1. The bound in (43) holds.

By the choice of stepsize and Lemma 19, we have that with probability $1 - 8e^{-\nu}$,

$$f(z_T) - f(z_0) \leq +c_\gamma (\tilde{\sigma}_1^2 (\gamma \ell k + \nu) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma k + \nu))$$

Indeed, from Lemma 19 with probability at least $1 - 8e^{-\nu}$,

$$f(z_k) - f(z_0) \leq -\frac{\gamma}{8} \sum_{i=0}^{k-1} \|\nabla_1 f(x_i, y_i)\|^2 - \frac{T \gamma}{8} \sum_{i=0}^{k-1} \|\nabla_2 f(x_i, y_i)\|^2 + c_\gamma (\tilde{\sigma}_1^2 (\gamma \ell k + \nu) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma k + \nu))$$

Hence, with probability at least $1 - 8e^{-\nu}$,

$$f(z_T) - f(z_0) \leq c_\gamma (\tilde{\sigma}_1^2 (\gamma \ell T + \nu) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma T + \nu))$$

Using the fact that $\mu > \ell / \tau$ and $\ell \geq \sqrt{\beta \epsilon}$, we have that

$$f(z_T) - f(z_0) \leq c\left(\frac{\tau \tilde{\sigma}_1^2}{\epsilon^3 \ell R \sqrt{\beta \epsilon}} + \frac{\tau \epsilon^2}{\epsilon^3 \ell \sqrt{\beta \epsilon}} + 2 \frac{\tau \tilde{\sigma}_2^2}{\epsilon^3 \ell R \sqrt{\beta \epsilon}} + 2 \frac{\tau^2 \epsilon^2}{\epsilon^3 \ell R \sqrt{\beta \epsilon}} + 2 \frac{\tau^2 \epsilon^2}{\epsilon^3 \ell R \sqrt{\beta \epsilon}}\right)$$

$$\leq c\left(\frac{\tau \tilde{\sigma}_1^2}{\epsilon^3 \ell} \sqrt{\frac{\epsilon^3}{\beta}} + \frac{\tau \epsilon^2}{\epsilon^3 \ell} \sqrt{\frac{\epsilon^3}{\beta}}\right)$$

$$\leq c\left(4K \frac{1}{\ell^8} \sqrt{\frac{\epsilon^3}{\beta}} + 4K \frac{1}{\ell^5} \sqrt{\frac{\epsilon^3}{\beta}}\right)$$

$$\leq 8cK \frac{1}{\ell^6} \sqrt{\frac{\epsilon^3}{\beta}} = 8c \frac{F}{\ell}$$

where in the second inequality we use the choice of $R$, and in the third inequality the choice of $K$. This proves the bound in (43).

Claim 2. The bound in (44) holds.

For the bound in (44), we take $t_0 = 0$ without loss of generality. Then, we need to show that

$$\mathbb{P}(f(z_T) - f(z_0) \leq -F) \geq 2/3 - \delta_1 - \delta_2$$

(94)
Consider coupling sequences \( \{z_t\} \) and \( \{z'_t\} \) as defined in Definition 10. Define the following events:

\[
\begin{align*}
\mathcal{E}_1 &= \left\{ \min \{ f(z_T) - f(z_0), f(z'_T) - f(z_0) \} \leq -\mathcal{F} \right\}, \\
\mathcal{E}_2 &= \left\{ \forall k \leq T: \max \{ \|z_k - z_0\|^2, \|z'_k - z_0\|^2 \} \leq S^2 \right\}, \\
\mathcal{E}_3 &= \left\{ \forall k \leq T, \|\hat{q}_{z,k} + \hat{q}_{z,k}\| \leq \frac{\gamma \alpha r_k}{20\sqrt{d}} \right\}, \\
\mathcal{E}_0 &= \left\{ \|\hat{q}_{\gamma,T}\| \geq \frac{\gamma \alpha r}{10\sqrt{d}} \right\}.
\end{align*}
\]

We know from lemmas proved above we have the following bounds:

\[
1 - \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2^c) = \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1^c) \geq 1 - \delta_1
\]

\[
\mathbb{P}(\mathcal{E}_0) \geq 2/3
\]

\[
\mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_2^c) \geq \mathbb{P}(\mathcal{E}_3|\mathcal{E}_2) \geq 1 - \delta_2
\]

\[
\mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_1) \geq 1 - \delta_1 - \delta_2
\]

where \( \delta_1 = 32dT e^{-t} \) and \( \delta_2 = 10dT^2 \log((S\sqrt{d}/(\gamma r))e^{-t}) \). Moreover,

\[
\mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_0|\mathcal{E}_1^c) \leq \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_0|\mathcal{E}_1^c) \leq \delta_1 + \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1^c)
\]

so that

\[
\mathbb{P}(\mathcal{E}_3|\mathcal{E}_1^c) \geq \mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_3|\mathcal{E}_1^c) - \delta_1
\]

\[
= 1 - \mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}_3|\mathcal{E}_1^c) - \delta_1
\]

\[
= 1 - \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1^c \cap \mathcal{E}_2)\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1^c) - \delta_1
\]

\[
\geq 1 - \delta_2 - \delta_1.
\]

Hence,

\[
\mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_0|\mathcal{E}_1^c) \leq \mathbb{P}(\mathcal{E}_3|\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_0|\mathcal{E}_1^c) \leq \delta_2 + \delta_1 + 1/3.
\]

Since

\[
\mathbb{P}(\mathcal{E}_3 \cup \mathcal{E}_0|\mathcal{E}_1^c) = 1 - \mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_0|\mathcal{E}_1^c),
\]

we have that

\[
\mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_0|\mathcal{E}_1^c) \geq 1 - 1/3 - \delta_2 - \delta_1 = 2/3 - \delta_2 - \delta_1.
\]

Hence, given \( \mathcal{E}_1^c \), then with probability at least \( 2/3 - \delta_2 - \delta_1 \) we have that both of the following hold:

\[
\|\hat{q}_{\gamma,T}\| \geq \frac{\alpha \gamma r}{10\sqrt{d}} \quad \text{and} \quad \|\hat{q}_{z,T} + \hat{q}_{\gamma,T}\| \leq \frac{\alpha \gamma r}{20\sqrt{d}}.
\]

Observe that by the definition of \( \hat{z}_T \) and the triangle inequality,

\[
\|\hat{z}_T\| = \|z_T - z'_T\| = \|z_T - z_0 + z_0 - z'_T\| \leq \|z_T - z_0\| + \|z'_T - z_0\| \leq 2 \max\{\|z_T - z_0\|, \|z'_T - z_0\|\}.
\]

Hence, by this relationship and an application of the reverse triangle inequality\(^{10}\) we have

\[
\max\{\|z_T - z_0\|, \|z'_T - z_0\|\} \geq \frac{1}{4} \|\hat{z}(T)\|
\]

\[
= \frac{1}{2} \|\hat{q}_{z,k} - \hat{q}_{z,k} - \hat{q}_{z,k}\|
\]

\[
\geq \frac{1}{2} \|\hat{q}_{\gamma,T}\| - \|\hat{q}_{z,T} + \hat{q}_{\gamma,T}\|
\]

\[
= \frac{1}{2} \left( \|\hat{q}_{\gamma,T}\| - \|\hat{q}_{z,T} + \hat{q}_{\gamma,T}\| \right)
\]

\[
\geq \frac{1}{2} \left( \frac{\alpha \gamma r}{10\sqrt{d}} - \frac{\alpha \gamma r}{20\sqrt{d}} \right)
\]

\[
= \frac{\alpha \gamma r}{40\sqrt{d}}
\]

\(^{10}\)That is, \( \|x - y\| \geq ||x|| - ||y|| \).

51
Observe that (95) follows from the fact that the lower bound on $\|\hat{q}_{t, r}\|$ is greater than the upper bound on $\|\hat{q}, \tau + \hat{q}, \tau\|$ so $\|\hat{q}, \tau\| - \|\hat{q}, \tau + \hat{q}, \tau\|$ is positive and (96) is obtained by applying the given upper and lower bounds on the quantities.

We now claim that
$$\max\{\|z_T - z_0\|, \|z'_T - z_0\|\} \geq \frac{\alpha_T \gamma r}{40\sqrt{d}} \geq S.$$  

Recall that $T = \frac{j}{\gamma \sqrt{d}}$ and $\text{Re}(\eta) \geq \sqrt{\beta \varepsilon}$ so that
$$\alpha_T = \frac{\sqrt{1 + \gamma \eta^2 T} - 1}{\sqrt{2} \gamma \text{Re}(\eta) + \gamma^2 |\eta|^2} \geq \frac{\sqrt{1 + 2\gamma \text{Re}(\eta) + \gamma^2 |\eta|^2}}{2 \sqrt{2} \gamma \text{Re}(\eta) + \gamma^2 |\eta|^2} \geq 2^{1/2}.\frac{T}{2}.$$  

Thus, with our choices of constants and when $\ell \geq c \log(\tau^2 \ell^2 / d \mu / (\beta \varepsilon))$
$$\frac{\alpha_T \gamma r}{40\sqrt{d}} \geq \frac{2^{1/3} 2/ \gamma r}{80\sqrt{d}} = \frac{2^{1/3} 2/ \gamma r}{80\sqrt{d} \ell \mu d} \geq \frac{2\varepsilon}{80\varepsilon^2 / \mu d} \geq \frac{2\varepsilon}{\mu d} \geq \frac{\tau^4 \varepsilon}{\beta \ell} = S.$$  

This shows that
$$1 - \mathbb{P}(E_2 \cap E_1^c) \geq \mathbb{P}(E_2^c \cup E_1) \geq \mathbb{P}(E_2^c | E_1) \geq \mathbb{P}(E_2^c | E_3 \cap E_0 | E_0^c) = \mathbb{P}(E_3 \cap E_0 | E_0^c) \geq 2/3 - \delta_1 - \delta_2$$  

so that
$$\mathbb{P}(E_1^c) = \mathbb{P}((E_1^c \cap E_2^c) \cup (E_1^c \cap E_2)) = \mathbb{P}(E_2^c | E_1^c) + \mathbb{P}(E_2^c \cap E_2) \leq 1/3 + 2\delta_1 + \delta_2$$  

and so we can conclude that
$$\mathbb{P}(E_1) \geq 2/3 - 2\delta_1 - \delta_2.$$  

That is,
$$\mathbb{P}(\min\{f(z_T) - f(z_0), f'(z_T) - f(z_0)\} \leq -F) \geq \frac{2}{3} - 10dT^2 \log(S \sqrt{d} / (\gamma r)) e^{-t} - 64dT e^{-t}.$$  

We also know that the marginal distributions of $z_T$ and $z'_T$ are the same. Hence, they escape a saddle point with the same probability. Indeed,
$$\mathbb{P}(f(z_T) - f(z_0) \leq -F) \geq \frac{\mathbb{P}(E_1)}{2} \geq \frac{1}{3} - \delta_1 - \delta_2 = \frac{1}{3} - 5dT^2 \log(S \sqrt{d} / (\gamma r)) e^{-t} - 32dT e^{-t}.$$  

\begin{thebibliography}{99}
\end{thebibliography}

E.3 Convergence Analysis

To prove the main convergence theorem we use the descent lemma for $\tau$-PGD which we restate here for convenience.

\textbf{Lemma 21 (Descent Lemma).} Consider a non-convex, $\mu$-SC zero-sum game defined by $f \in C^2(\mathbb{Z}, \mathbb{R})$. Under Assumptions 2 and 3 there exists an absolute constant $c_{\text{max}}$ such that for any fixed $t, t_0, \iota > 0$, if $\frac{3}{\mu \ell} < \gamma < 1/\ell$, then with at least $1 - 8e^{-t}$ probability, the sequence generated by $\tau$-PGD with parameters $\gamma, \tau, r$ satisfies

$$f(z_{t_0 + \ell t}) - f(z_{t_0}) \leq -\frac{\gamma}{8} \sum_{k=0}^{t-1} (\|\nabla_1 f(z_{t_0 + k})\|^2 + \|\nabla_2 f(z_{t_0 + k})\|^2)$$  

$$+ c (\gamma (\sigma_1^2 + \tau^2)(\gamma \ell t + \iota) + \tau \gamma (\sigma_1^2 + \tau^2)(\mu \tau \gamma t + \iota))$$  

Note that this lemma is the same as Lemma 3 from the main text, just restated with the precise constants.
Theorem 5. Consider a non-convex, $\mu$-strongly concave zero-sum game defined by $f \in C^2(Z, \mathbb{R})$ and suppose that Assumptions 1-6 hold. For any $\varepsilon, \delta > 0$, there exists $\gamma$ and $\tau$ such that, with probability $1-\delta$ for some $\delta > 0$, starting from any $z_0 = (x_0, y_0)$, at least half the iterates of $\tau$-PGDA will be $\varepsilon$-differential Stackelberg equilibria after $\tilde{O}(\varepsilon^{-4})$ iterations and $\tilde{O}(\varepsilon^{-2})$ in the stochastic and deterministic settings, respectively.

Proof. Consider an arbitrary $\varepsilon, \delta > 0$. Before proceeding to the main proof, observe that when $\|g(z_t)\| \leq \varepsilon$, we have that $\|\nabla f(z_t)\|$ is also small. Indeed, if $\|g(z_t)\| \leq \varepsilon$ then $\|\nabla_1 f(z_t)\| \leq \varepsilon$ and $\|\nabla_2 f(z_t)\| \leq \varepsilon$. Hence,

$$
\|\nabla f(z_t)\| = \|\nabla_1 f(z_t) - \nabla_2 f(z_t)(\nabla_2^2 f(z_t))^{-1}\nabla_2 f(z_t)\|
\leq \|\nabla_1 f(z_t)\| + \|\nabla_2 f(z_t)\|\|(\nabla_2^2 f(z_t))^{-1}\|\|\nabla_2 f(z_t)\|
\leq \varepsilon \left(1 + \frac{2L}{\mu}\right)
$$

where recall that $L$ is the Lipschitz constant for $g(z)$.

Let $\tilde{\varepsilon} = \varepsilon/(1 + 2\mu/L)$ and apply Lemma 19 and Lemma 21 with $\tilde{\varepsilon}$. Then, Lemma 19 gives us a high probability bound on the iterates of $\tau$-PGDA decreasing the function value and Lemma 21 gives high probability guarantees on escaping saddle by similarly arguing that the function value will decrease rapidly when we have either large gradients (i.e, $\|g(x, y)\| \geq \tilde{\varepsilon}$) or strictly negative eigenvalues of the game Jacobian (negative curvature in the game space).

Let the total number of iterations be given by

$$
T = 100 \max \left\{ \frac{\Delta_f T}{\|f\|}, \frac{\Delta_f}{\gamma \tilde{\varepsilon}} \right\} = O \left( \frac{\ell \Delta_f}{\varepsilon^2 \gamma \tilde{\varepsilon}} \right).
$$

where $\Delta_f = f(z_0) - f^*$ and $f^*$ is the global minimum of the function $f$.

The following two claims hold simultaneously with probability $1-\delta$:

Claim 1. At most $T/4$ iterates have $\|g(z_t)\| \geq \tilde{\varepsilon}$;

Claim 2. At most $T/4$ iterates are close to saddle points—that is, $\|g(z_t)\| \leq \tilde{\varepsilon}$ and $\lambda_{\min}(J_\tau(z_t)) \leq -\sqrt{\varepsilon \tilde{\varepsilon}}$.

If the two claims hold, then at least $T/2$ iterates are $\tilde{\varepsilon}$-local minmax points.

Proof of Claim 1. Suppose that within $T$ steps, we have more than $T/4$ iterates for which $\|g(z_t)\| \geq \tilde{\varepsilon}$. By Lemma 19 we have with probability $1-8e^{-\gamma}$:

$$
f(z_T) - f(z_0) \leq \frac{\gamma}{8} \sum_{k=0}^{t-1} (\|\nabla_1 f(z_{t_0+k})\|^2 + \tau \|\nabla_2 f(z_{t_0+k})\|^2)
+ c (\gamma(\sigma_1^2 + r^2)(\gamma \ell t + \iota) + \tau \gamma(\sigma_2^2 + r^2) (\mu \tau \gamma t + \iota))
\leq -\gamma \left( \frac{T(1 + \tau)\varepsilon^2}{32} - c (\sigma_1^2(\gamma \ell T + \iota) + \tau \sigma_2^2 (\mu \tau \gamma T + \iota)) \right).
$$

Observe that by our choice of constants (again recalling that in the definition of the constants we use $\tilde{\varepsilon}$ in place of $\varepsilon$), we have that

$$
\frac{T}{\|f\|} = \frac{\ell \varepsilon^6}{\tau^3 \gamma^3 \sqrt{\beta} \tilde{\varepsilon}} \sqrt{\frac{\beta}{\tilde{\varepsilon}^3}} = \frac{\ell \varepsilon^6}{\tau^3 \gamma^3 \sqrt{\beta} \tilde{\varepsilon}^4 \sqrt{\frac{\beta}{\tilde{\varepsilon}^3}}} = \frac{\ell \varepsilon^6}{\tau^3 \gamma^3 \varepsilon^2 \tilde{\varepsilon}^2}.
$$
and \( T = \geq 100\Delta_f T / F \). Hence,

\[
c_{\gamma} (\tilde{\sigma}^2 (\gamma T + i) + \tau \tilde{\sigma}^2 (\mu \tau \gamma T + i)) \leq c_{\gamma} \left( \tilde{\sigma}^2 (\gamma T \frac{\Delta_f}{F} + i) + \tau \tilde{\sigma}^2 (\mu \tau \gamma T \frac{\Delta_f}{F} + i) \right)
\]

\[
\leq c_{\gamma} \left( \tilde{\sigma}^2 (\gamma T \frac{\ell}{\sqrt{3} \sqrt{\delta} F} + i) + \tau \tilde{\sigma}^2 (\mu \tau \gamma T \frac{\ell}{\sqrt{3} \sqrt{\delta} F} + i) \right)
\]

\[
\leq 200c\Delta_f \left( \frac{\tau \tilde{\sigma}^2}{\ell^8 \sqrt{3} \delta F R} + \frac{\tau \tilde{\sigma}^2}{\ell^8 \ell \sqrt{3} \delta F R} + \frac{\tau^2 \sigma^2}{i^6 \ell \sqrt{3} \delta F R} + \frac{\tilde{\sigma}^2}{i^6 \ell \sqrt{3} \delta F R} \right)
\]

(102)

\[
\leq 200c\Delta_f \left( \frac{1}{\ell^3} + \frac{1}{\ell^3} + \frac{\tau}{\ell^3} \right)
\]

\[
\leq \gamma T (1 + \tau) \tilde{\sigma}^2
\]

\[
\leq \frac{\gamma T (1 + \tau) \tilde{\sigma}^2}{64}
\]

where \( \sigma^2 = \max \{ \sigma_1^2, \sigma_2^2 \} \) and in (101) and (102) we used the fact that \( \mu \tau > \ell \). The last inequality holds since \( T \gamma \tilde{\sigma}^2 / \ell^6 \geq 100 \Delta_f \) and by choosing \( \ell \) large enough. Returning to (100), we have that

\[
f(z_T) - f(z_0) = - \gamma \frac{T (1 + \tau) \tilde{\sigma}^2}{64}
\]

which implies that

\[
f(z_T) \leq f(z_0) - \gamma \frac{T (1 + \tau) \tilde{\sigma}^2}{64} < f^*
\]

which is not possible. Hence, claim 1 holds.

**Proof of Claim 2.** Define the following stopping times that enable us to use Lemma [21]

\[
\alpha_1 = \inf \{ s \mid \| g(z_s) \| \leq \tilde{\delta} \text{ and } \text{Re}(\lambda_{\min}(J_{\tau}(z_s))) \leq -\sqrt{3} \delta \}
\]

\[
\alpha_i = \inf \{ s > \alpha_{i-1} + T \mid \| g(z_s) \| \leq \tilde{\delta} \text{ and } \text{Re}(\lambda_{\min}(J_{\tau}(z_s))) \leq -\sqrt{3} \delta \}, \forall i > 1
\]

where \( \text{Re}(\lambda_{\min}(\cdot)) \) is the real part of the eigenvalue with the minimum real part. Each \( \alpha_i \) is a stopping time random variable indicating the \( i \)-th time in the sequence (generated by \( \tau \text{-PGDA} \)) at which Lemma [21] can be invoked. Define the random variable \( M = \max \{ i \mid \alpha_i + T \leq T \} \). Then, we decompose \( f(z_T) - f(z_0) \) as follows:

\[
f(z_T) - f(z_0) = \sum_{i=1}^{M} (f(z_{\alpha_i + T}) - f(z_{\alpha_i}))
\]

\[
+ (f(z_T) - f(z_{\alpha_M})) + (f(z_{\alpha_1}) - f(z_0)) + \sum_{i=1}^{M-1} (f(z_{\alpha_i + 1}) - f(z_{\alpha_i + T}))
\]

Now, by Lemma [21] and Azuma-Hoeffding supermartingale concentration inequality, we have for each fixed \( m \leq T \) that

\[
P \left( \sum_{i=1}^{m} (f(z_{\alpha_i + T}) - f(z_{\alpha_i})) \leq -(0.9m - c\sqrt{m})F \right) \geq 1 - 5dT^2 T \log(S\sqrt{d}/(\gamma r))e^{-c}.
\]

Since \( M \leq T/T \leq T \), taking a union bound, we know that

\[
T_1 = -(0.9M - c\sqrt{M})F
\]

54
with probability $1 - 5dT^2 T^2 \log(S \sqrt{d/(\gamma r)}) e^{-t}$. Now, for $T_2$, we take a union bound and apply Lemma 19 over all $t_1, t_2 \in [0, T]$. Indeed, with probability $1 - 8T^2 e^{-t}$,

$$T_2 \leq c\gamma (\tilde{\sigma}_1^2 (\gamma \ell T + 2M\ell) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma T + 2M\ell))$$

Arguing in a similar fashion as for Claim 1, if within $T$ steps we have more than $T/4$ saddle points, then, with probability $1 - 10dT^2 T^2 \log(S \sqrt{d/(\gamma r)}) e^{-t}$,

$$f(zt) - f(z_0) \leq -(0.9M - c\sqrt{M_1}f + c\gamma (\tilde{\sigma}_1^2 (\gamma \ell T + 2M\ell) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma T + 2M\ell)) \cdot (103)$$

To further upper bound the above expression, let us focus on the last three terms:

$$c\sqrt{M_1}f + c\gamma (\tilde{\sigma}_1^2 (\gamma \ell T + 2M\ell) + \tau \tilde{\sigma}_2^2 (\mu \tau \gamma T + 2M\ell))$$

$$\leq c\sqrt{M_1}f + c\gamma \left( \frac{\tilde{\sigma}_1^2 (\gamma \ell T + 2M\ell)}{F} + 2M\ell \right)$$

$$\leq c\sqrt{100\Delta_f^2 \ell^2 / \beta^2} + \frac{1}{2} \left( \sigma^2 + \ell^2 \right) + \frac{1}{2} \left( \ell^2 \right) + \frac{1}{2} \left( \ell^2 \right)$$

$$\leq 100\Delta_f^2 \ell^2 / \beta^2 + c \left( \frac{\tilde{\sigma}_1^2 (\gamma \ell T + 2M\ell)}{F} + 2M\ell \right)$$

$$\leq c4Mf \frac{\tau \sqrt{\tilde{\sigma}}}{\ell^2 \sqrt{\beta}} + c \frac{24Mf}{\ell}$$

where in (104) we use the fact that $\tau \mu > \ell \geq \sqrt{\beta \tilde{\epsilon}}$. Now, we choose $\ell$ to be a large constant so that

$$f(zt) - f(z_0) \leq -0.4Mf \leq -0.1T\frac{F}{T}$$

since $M \geq T/(4T)$. This implies that

$$f(zt) \leq f(z_0) - 0.1T\frac{F}{T} < f^*$$

which is not possible. Hence, choosing absolute constant $c$ large enough in $\ell = c \log(d\Delta_f \beta \hat{R} / (\beta \tilde{\epsilon} \hat{d}))$, both claims hold with probability $1 - \delta$.

**Obtaining final convergence guarantee.** When $\|g(z_t)\| \leq \tilde{\epsilon} = \epsilon / (1 + 2\mu / L)$, from the observations at the beginning of the proof, we have that

$$\|\nabla_2 f(z_t)\| \leq \epsilon \quad \text{and} \quad \|\nabla f(z_t)\| \leq \epsilon$$

so that $z_t$ is a $\epsilon$-differential Stackelberg equilibrium (strict local minmax). Hence, the claimed iteration complexity holds. □