Convergence Analysis of Gradient-Based Learning in Continuous Games

Benjamin Chasnov[1], **Lillian Ratliff**[1], **Eric Mazumdar**[2], **Samuel Burden**[1]
[1]University of Washington, Seattle, WA
[2]University of California, Berkeley, Berkeley, CA

Abstract

Considering a class of gradient-based multi-agent learning algorithms in non-cooperative settings, we provide convergence guarantees to a neighborhood of a stable Nash equilibrium. In particular, we consider continuous games where agents learn in 1) deterministic settings with oracle access to their individual gradient and 2) stochastic settings with an unbiased estimator of their individual gradient. We also study the effects of non-uniform learning rates, which cause a distortion of the vector field that can alter the equilibrium to which the agents converge and the learning path. We support the analysis with numerical examples that provide insight into how games may be synthesized to achieve desirable equilibria.

1 INTRODUCTION

The characterization and computation of equilibria such as *Nash equilibria* and its refinements constitutes a significant focus in non-cooperative game theory. A natural question that arises is 'how do players find or learn such equilibria and how should the grappling process that occurs during learning be interpreted?' With this question in mind, a variety of fields have focused attention on the problem of learning in games which has lead to a plethora of learning algorithms including gradient play, fictitious play, best response, and multi-agent reinforcement learning among others (Fudenberg and Levine, 1998). While convergence has been studied for many of these algorithms, the results largely tend to be asymptotic in nature; questions of error bounds and convergence rates are often less explored, particularly in the context of non-uniform learning rates, a key feature of systems of autonomous learning agents.

From an applications point of view, another recent trend is in the adoption of game theoretic models of algorithm interaction in machine learning applications. For instance, game theoretic tools are being used to improve the robustness and generalizability of machine learning algorithms; e.g., generative adversarial networks have become a popular topic of study demanding the use of game theoretic ideas to provide performance guarantees (Daskalakis et al., 2017). In other work from the learning community, game theoretic concepts are being leveraged to analyze the interaction of learning agents—see, e.g., (Balduzzi et al., 2018; Heinrich and Silver, 2016; Mazumdar and Ratliff, 2018; Mertikopoulos and Zhou, 2019; Tuyls et al., 2018). Even more recently, the study of convergence to Nash equilibria has been called into question (Papadimitriou and Piliouras, 2018); in its place is a proposal to consider game dynamics as the *meaning of the game*. This is an interesting perspective as it is well known that in general learning dynamics do not obtain an Nash equilibrium even asymptotically—see, e.g., (Hart and Mas-Colell, 2003)—and, perhaps more interestingly, many learning dynamics exhibit very interesting limiting behaviors including periodic orbits and chaos—see, e.g., (Benaïm and Hirsch, 1999; Benaïm et al., 2012; Hofbauer, 1996; Hommes and Ochea, 2012).

Despite this activity, we still lack a complete understanding of the dynamics and limiting behaviors of coupled, competing learning algorithms. One may imagine that the myraid results on convergence of gradient descent in optimization readily extend to the game setting. Yet, they do not since gradient-based learning schemes in games *do not correspond to gradient flows*, a class of flows that are guaranteed to converge to local minimizers almost surely. In particular, the gradient-based learning dynamics for competitive, multi-agent settings have a *non-symmetric Jacobian* and, as a consequence, their dynamics may admit complex eigenvalues and non-equilibrium limiting behavior such as periodic orbits. In short, this fact makes it difficut to extend many of the optimization approaches

to convergence in single-agent optimization settings to multi-agent settings primarily due to the fact that steps in the direction of a player's individual gradient does not guarantee that the player's cost decreases. In fact, in games, as our examples highlight, a player's cost can increase when they follow the gradient of their own cost. This behavior is due to the coupling between the agents.

Some of the questions that remain unaddress, and to which we provide at least partial answers, include the derivation of error bounds and convergence rates which are important for ensuring certain performance guarantees on the collective behavior and can be used to provide guarantees on subsequent control or incentive policy synthesis. We also investigate the question of how naturally arising features of the learning process for autonomous agents, such as their learning rates, impact the learning path and limiting behavior. This further exposes interesting questions about the overall quality of the limiting behavior and the cost accumulated along the learning path—e.g., is it better to be a slow or fast learner both in terms of the cost of learning and the learned behavior?

Contributions. We use state of the art tools and techniques from dynamical systems theory and numerical methods to make new contributions to the field of multi-agent learning, the theory of continuous games, and learning in games. In particular, we study convergence of a class of gradient-based multi-agent learning algorithms in non-cooperative settings where agents have non-uniform learning rates by leveraging the framework of n-player continuous games. That is, we consider a class of learning algorithms $x_i^+ = x_i - \gamma_i g_i(x_i, x_{-i})$ in which g_i is derived from the gradient of a function that abstractly represents the cost function of player i. This class encompases a wide variety of approaches to learning in games including multi-agent policy gradient and multi-agent gradient-based online optimization. We consider two settings: (i) agents have oracle access to g_i and (ii) agents have an unbiased estimator for g_i.

To our knowledge finite time guarantees for either the stochastic or deterministic setting given non-uniform learning rates have not been provided; we provide both. Towards this end, we characterize the local structure of the game around the equilibria and exploit this local structure to obtain finite time rates by combining it with dynamical systems theory results thereby leading to convergence guarantees for settings not currently covered by the state of the art. The analysis combines insights about games and the structure of the learning dynamics near equilibria with results from numerical methods to obtain finite time bounds in the deterministic setting and very recent advancements in concentration bounds for stochastic approximation in the stochastic setting. The setting of non-uniform learning rates complicates the analysis and is well motivated, particularly for applications in which the agents are autonomous and learning their strategies through repeated interaction, as opposed to a setting in which an external entity has the goal of computing the Nash equilibria of a game.

2 PRELIMINARIES

Consider a setting in which at iteration k, each agent $i \in \mathcal{I} = \{1, \dots, n\}$ updates their choice variable $x_i \in X_i = \mathbb{R}^{d_i}$ by the process

$$x_{i,k+1} = x_{i,k} - \gamma_{i,k} g_i(x_{i,k}, x_{-i,k}) \tag{1}$$

where $\gamma_{i,k}$ is the learning rate, $x_{-i,k} = (x_{j,k})_{j \in \mathcal{I}/\{i\}} \in \prod_{j \in \mathcal{I}/\{i\}} X_j$ denotes the choices of all agents excluding the i–th agent, and $(x_{i,k}, x_{-i,k}) \in X = \prod_{i \in \mathcal{I}} X_i$. For each $i \in \mathcal{I}$, there exists a sufficiently smooth function $f_i \in C^q(X, \mathbb{R})$, $q \geq 2$ such that g_i is either $D_i f_i$ where $D_i(\cdot)$ denotes the derivative of f_i with respect to x_i or an unbiased estimator of $D_i f_i$—i.e., $g_i \equiv \widehat{D_i f_i}$ where $\mathbb{E}[\widehat{D_i f_i}] = D_i f_i$.

The collection of costs $(f_1, \dots, f_n)$ on X where $f_i : X \to \mathbb{R}$ is agent i's cost function and X_i is their action space defines a *continuous game*. In this continuous game abstraction, each player $i \in \mathcal{I}$ aims to selection an action $x_i \in X_i$ that minimizes their cost $f_i(x_i, x_{-i})$ given the actions of all other agents, $x_{-i} \in X_{-i}$. From this perspective, the learning algorithm in (1) can be interpreted as follows: players myopically update their actions by following the gradient of their cost with respect to their own choice variable.

Assumption 1. *For each $i \in \mathcal{I}$, $f_i \in C^q(X, \mathbb{R})$ for $q \geq 2$ and $\omega(x) = (D_1 f_1(x) \ \cdots \ D_n f_n(x))$ is L–Lipschitz.*

Let $D_i^2 f_i$ denote the second partial derivative of f_i with respect to x_i and $D_{ji} f_i$ denote the partial derivative of $D_i f_i$ with respect to x_j. The *game Jacobian*—i.e., the Jacobian of ω—is given by

$$J(x) = \begin{bmatrix} D_1^2 f_1(x) & \cdots & D_{1n} f_1(x) \\ \vdots & \ddots & \vdots \\ D_{n1} f_n(x) & \cdots & D_n^2 f_n(x) \end{bmatrix}.$$

The entries of the above matrix are dependent on x, however, we drop this dependence where obvious. Note that each $D_i^2 f_i$ is symmetric under Assumption 1, yet J is not. This is an important point and causes the subsequent analysis to deviate from the typical analysis of (stochastic) gradient descent.

The most common characterization of limiting behavior in games is that of a Nash equilibrium. The following definitions are useful for our analysis.

Definition 1. *A strategy $x \in X$ is a local Nash equilibrium for the game $(f_1, \ldots, f_n)$ if for each $i \in \mathcal{I}$ there exists an open set $W_i \subset X_i$ such that $x_i \in W_i$ and $f_i(x_i, x_{-i}) \leq f_i(x_i', x_{-i})$ for all $x_i' \in W_i$. If the above inequalities are strict, x is a strict local Nash equilibrium.*

Definition 2. *A point $x \in X$ is said to be a critical point for the game if $\omega(x) = 0$.*

We denote the set of critical points of a game $\mathcal{G} = (f_1, \ldots, f_n)$ as $\mathcal{C}(\mathcal{G}) = \{x \in X | \ \omega(x) = 0\}$. Analogous to single-player optimization, viewing all other players' actions as fixed, there are necessary and sufficient conditions which characterize local optimality for each player.

Proposition 1 (Ratliff et al. (2016))**.** *If x is a local Nash equilibrium of the game $(f_1, \ldots, f_n)$, then $\omega(x) = 0$ and $D_i^2 f_i(x) \geq 0$. On the other hand, if $\omega(x) = 0$ and $D_i^2 f_i(x) > 0$, then $x \in X$ is a local Nash equilibrium.*

The sufficient conditions in the above result give rise to the following definition of a differential Nash equilibrium.

Definition 3 (Ratliff et al. (2013); Ratliff et al. (2016))**.** *A strategy $x \in X$ is a differential Nash equilibrium if $\omega(x) = 0$ and $D_i^2 f_i(x) > 0$ for each $i \in \mathcal{I}$.*

Differential Nash need not be isolated. However, if $J(x)$ is non-degenerate—meaning that $\det J(x) \neq 0$—for a differential Nash x, then x is an *isolated strict local Nash equilibrium*. Non-degenerate differential Nash are *generic* amongst local Nash equilibria and they are *structurally stable* (Ratliff et al., 2014) which ensures they persist under small perturbations. This also implies an asymptotic convergence result: if the spectrum of J is strictly in the right-half plane (i.e. $\mathrm{spec}(J(x)) \subset \mathbb{C}_+^\circ$), then a differential Nash equilibrium x is (exponentially) attracting under the flow of $-\omega$ (Ratliff et al., 2016, Prop. 2). We say such equilibria are *stable*.

3 DETERMINISTIC SETTING

Let us first consider the setting in which each agent i has oracle access to g_i and their learning rates are non-uniform, but constant—i.e., $\gamma_{i,k} \equiv \gamma_i$. The learning dynamics are given by

$$x_{k+1} = x_k - \Gamma\omega(x_k) \tag{2}$$

where $\Gamma = \mathrm{blockdiag}(\gamma_1 I_{d_1}, \ldots, \gamma_n I_{d_n})$ with I_{d_i} denoting the $d_i \times d_i$ identity matrix.

3.1 ASYMPTOTIC ANALYSIS

For a stable differential Nash x^*, let $\mathcal{R}(x^*)$ denote the region of attraction for x^*. Denote by $\rho(A)$ the spectral radius of the matrix A.

Proposition 2. *Consider an n–player game $\mathcal{G} = (f_1, \ldots, f_n)$ satisfying Assumption 1. Let $x^* \in X$ be a stable differential Nash equilibrium. Suppose agents use the gradient-based learning rule $x_{k+1} = x_k - \Gamma\omega(x_k)$ with learning rates γ_i such that $\rho(I - \Gamma J(x)) < 1$ for all $x \in \mathcal{R}(x^*)$. Then, for $x_0 \in \mathcal{R}(x^*)$, $x_k \to x^*$ exponentially.*

The proof is a direct application of Ostrowski's theorem (Ostrowski, 1966).

Remark 1. *If all the agents have the same learning rate—i.e., for each $i \in \mathcal{I}$, $\gamma_i = \gamma$—then the condition that $\rho(I - \Gamma J(x)) < 1$ on $\mathcal{R}(x^*)$ can be written as $0 < \gamma < \tilde{\gamma}$ where $\tilde{\gamma}$ is the smallest positive h such that $\max_j |1 - h\lambda_j(J(x^*))| = 1$. If the game is a potential game—i.e., there exists a function ϕ such that $D_i f_i = D_i \phi$ for each i which occurs if and only if $D_{ij} f_i = D_{ji} f_j$—then convergence analysis coincides with gradient descent so that any $\gamma < 1/L$ where L is the Lipschitz constant of ω results in local asymptotic convergence.*

Mazumdar and Ratliff (2018) show that (2) will almost surely avoid strict saddle points of the dynamics, some of which are Nash equilibria in non-zero sum games. Moreover, except on a set of measure zero, (2) will converge to a stable attractor of $\dot{x} = -\omega(x)$ which includes stable local non-Nash critical points. Since ω is not a gradient flow, the set of attractors may also include limit cycles.

3.2 FINITE SAMPLE ANALYSIS

Throughout this subsection we need the following notation. For a symmetric matrix $A \in \mathbb{R}^{d \times d}$, let $\lambda_d(A) \leq \cdots \leq \lambda_1(A)$ be its eigenvalues. Let $S(x) = \frac{1}{2}(J(x) + J(x)^T)$ be the symmetric part of $J(x)$. Define $\alpha = \min_{x \in B_r(x^*)} \lambda_d\big(S(x)^T S(x)\big)$ and $\beta = \max_{x \in B_r(x^*)} \lambda_1(J(x)^T J(x))$ where $B_r(x^*)$ is a r–radius ball around x^*. Let $B_{r_0}(x^*)$ with $0 < r_0 < \infty$ be the largest ball contained in the region of attraction of x^*—i.e. $B_{r_0}(x^*) \subset \mathcal{R}(x^*)$.

Letting $g(x) = x - \Gamma\omega(x)$, since $\omega \in C^q$ for some $q \geq 1$, $g \in C^q$, the expansion $g(x) = g(x^*) + (I - \Gamma J(x))(x - x^*) + R(x - x^*)$ holds, where R satisfies $\lim_{x \to x^*} \|R(x - x^*)\| / \|x - x^*\| = 0$ so that given $c > 0$, there exists an $r > 0$ such that $\|R(x - x^*)\| \leq c\|x - x^*\|$, $\forall\, x \in B_r(x^*)$.

Proposition 3. *Suppose that $\|I - \Gamma J(x)\| < 1$ for all $x \in B_{r_0}(x^*) \subset \mathcal{R}(x^*)$ so that there exists r', r'' such that $\|I - \Gamma J(x)\| \leq r' < r'' < 1$ for all $x \in B_{r_0}(x^*)$. For $1 - r'' > 0$, let $0 < r < \infty$ be the largest r such that $\|R(x - x^*)\| \leq (1 - r'')\|x - x^*\|$ for all $x \in B_r(x^*)$. Furthermore, let $x_0 \in B_{r^*}(x^*)$ where $r^* = \min\{r, r_0\}$*

be arbitrary. Then, given $\varepsilon > 0$, gradient-based learning with learning rates Γ obtains an ε–differential Nash equilibrium in finite time—i.e., $x_t \in B_\varepsilon(x^*)$ for all $t \geq T = \lceil \frac{1}{\delta} \log(r^*/\varepsilon) \rceil$ where $\delta = r'' - r'$.

With some modification, the proof follows the proof of Theorem 1 in (Argyros, 1999); we provide it in Appendix A.1 for completeness.

Remark 2. *We note that the proposition can be more generally stated with the assumption that $\rho(I - \Gamma J(x)) < 1$, in which case there exists some δ defined in terms of bounds on powers of $I - \Gamma J$. We provide the proof of this in Appendix A.1. We also note that these results hold even if Γ is not a diagonal matrix as we have assumed as long as $\rho(I - \Gamma J(x)) < 1$.*

A perhaps more interpretable finite bound stated in terms of the game structure can also be obtained. Consider the case in which players adopt learning rates $\gamma_i = \sqrt{\alpha}/(\beta k_i)$ with $k_i \geq 1$. Given a stable differential Nash equilibrium x^*, let $B_r(x^*)$ be the largest ball of radius r contained in the region of attraction on which $\tilde{S} \equiv \frac{1}{2}(\tilde{J}^T + \tilde{J})$ is positive definite where $\tilde{\omega} = (D_i f_i / k_i)_{i \in \mathcal{I}}$ so that $\tilde{J} \equiv D\tilde{\omega}$, and define $\tilde{\alpha} = \min_{x \in B_r(x^*)} \lambda_d\big(\tilde{S}(x)^T \tilde{S}(x)\big)$ and $\tilde{\beta} = \max_{x \in B_r(x^*)} \lambda_1(\tilde{J}(x)^T \tilde{J}(x))$.

Theorem 1. *Suppose that Assumption 1 holds and that $x^* \in X$ is a stable differential Nash equilibrium. Let $x_0 \in B_r(x^*)$, $\alpha < k_{\min}\beta$, $\sqrt{\alpha}/k_{\min} \leq \sqrt{\tilde{\alpha}}$, and for each i, $\gamma_i = \sqrt{\alpha}/(\beta k_i)$ with $k_i \geq 1$. Then, given $\varepsilon > 0$, the gradient-based learning dynamics with learning rates γ_i obtain an ε–differential Nash such that $x_t \in B_\varepsilon(x^*)$ for all $t \geq \lceil 2\frac{\beta k_{\min}}{\alpha} \log(\frac{r}{\varepsilon}) \rceil$.*

Proof First, note that $\|x_{k+1} - x^*\| = \|g(x_k) - g(x^*)\|$ where $g(x) = x - \Gamma\omega(x)$. Now, $x_0 \in B_r(x^*)$ so that by the mean value theorem, $\|g(x_0) - g(x^*)\| = \| \int_0^1 Dg(\tau x_0 + (1-\tau)x^*)(x_0 - x^*) d\tau \| \leq \sup_{x \in B_r(x^*)} \|Dg(x)\| \|x_0 - x^*\|$. Hence, it suffices to show that for the choice of Γ, the eigenvalues of $I - \Gamma J(x)$ live in the unit circle, and then use an inductive argument. Let $\Lambda = \operatorname{diag}(1/k_1, \ldots, 1/k_n)$ so that we need to show that $I - \gamma\Lambda D\omega$ has eigenvalues in the unit circle. Since $\omega(x^*) = 0$, we have that $\|x_{k+1} - x^*\|_2 = \|x_k - x^* - \gamma\Lambda(\omega(x_k) - \omega(x^*))\|_2 \leq \sup_{x \in B_r(x^*)} \|I - \gamma\Lambda J(x)\|_2 \|x_k - x^*\|_2$ If $\sup_{x \in B_r(x^*)} \|I - \gamma\Lambda J(x)\|_2$ is less than one, where the norm is the operator 2–norm, then the dynamics are contracting. Indeed, observe that the singular values of $\Lambda J^T J \Lambda$ are the same as those of $J^T \Lambda^2 J$ since the latter is positive-definite symmetric. By noting that $\|A\|_2 = \sigma_{\max}(A)$ and employing Cauchy-Schwartz, we get that $\|\Lambda\|_2^2 \|J^T J\|_2 \geq \|\Lambda J^T J \Lambda\|_2$. Thus,

$$
\begin{aligned}
&(I - \gamma\Lambda J)^T (I - \gamma\Lambda J) \leq (1 - 2\gamma\lambda_d(\tilde{S}) + \tfrac{\gamma^2 \lambda_1(J^T J)}{k_{\min}^2}) I \\
&\leq (1 - 2\gamma\sqrt{\alpha}/k_{\min} + \alpha/(\beta k_{\min})) I \\
&= (1 - \alpha/(\beta k_{\min})) I.
\end{aligned}
$$

Using the above to bound $\sup_{x \in B_r(x^*)} \|I - \gamma\Lambda J(x)\|_2$, we have $\|x_{k+1} - x^*\|_2 \leq (1 - \frac{\alpha}{\beta k_{\min}})^{1/2} \|x_k - x^*\|_2$. Since $\alpha < k_{\min}\beta$, $(1 - \alpha/(\beta k_{\min})) < e^{-\alpha/(\beta k_{\min})}$ so that $\|x_{k+1} - x^*\|_2 \leq e^{-T\alpha/(2k_{\min}\beta)} \|x_0 - x^*\|_2$. This, in turn, implies that for alls $t \geq \lceil 2\frac{\beta k_{\min}}{\alpha} \log(r/\varepsilon) \rceil$, $x_t \in B_\varepsilon(x^*)$.

■

Multiple learning rates lead to a scaling rows which can have a significant effect on the eigenstructure of the matrix, thereby making it difficult to reason about the relationship between ΓJ and J. None-the-less, there are numerous approaches to solving nonlinear systems of equations that employ *preconditioning* (i.e., coordinate scaling). The purpose of using a preconditioning matrix is to rescale the problem and achieve more stable or faster convergence. In games, however, the interpretation is slightly different since each of the coordinates of the dynamics corresponds to minimizing a different cost function along the respective coordinate axis. The resultant effect is a distortion of the vector field in such a way that it has the effect of leading the joint action to a point which has a lower value in general for the slower player relative to the flow of the dynamics given a uniform learning rate and the same initialization. In this sense, it seems that it is most beneficial for an agent to have the slower learning rate, which is suggestive of desirable qualities for synthesized algorithms. In the case of autonomous learning agents, perhaps this reveals an interesting direction of future research in terms of synthesizing games or learning rules via incentivization (Ratliff and Fiez, 2018) or reward shaping (Ng et al., 1999) for either coordinating agents or improving the learning process.

4 STOCHASTIC SETTING

Consider the setting where agents no longer have oracle access to their individual gradients, but rather have an unbiased estimator $g_i \equiv \widehat{D_i f_i}$ and a timevarying learning rate $\gamma_{i,k}$. For the sake of brevity, we show the convergence result in detail for the two agent case—that is, where $\mathcal{I} = \{1, 2\}$. We note that the extension to n agents is straightforward.

The gradient-based learning rules are given by

$$x_{i,k+1} = x_{i,k} - \gamma_{i,k}(\omega(x_k) + w_{i,k+1}) \tag{3}$$

so that within $\gamma_{2,k} = o(\gamma_{1,k})$, in the limit $\tau \to 0$, the above system can be thought of as approximating the singularly perturbed system defined as follows:

$$\dot{x}_1(t) = -D_1 f_1(x_1(t), x_2(t)) \tag{4}$$
$$\dot{x}_2(t) = -\tau D_2 f_2(x_1(t), x_2(t)) \tag{5}$$

Indeed, since $\lim_{k\to\infty} \gamma_{2,k}/\gamma_{1,k} \to 0$—i.e., $\gamma_{2,k} \to 0$ at a faster rate than $\gamma_{1,k}$—updates to x_1 appear to be equilibriated for the current quasi-static x_2.

We require some modified assumptions in this section on the learning process structure.

Assumption 2. *For the gradient-based learning rule* (3)*, suppose that the following hold:*

A2a. *Given the filtration $\mathcal{F}_k = \sigma(x_s, w_{1,s}, w_{2,s}, s \leq k)$, $\{w_{i,k+1}\}_{i\in\mathcal{I}}$ are conditionally independent, $\mathbb{E}[w_{i,k+1}|\ \mathcal{F}_k] = 0$ almost surely (a.s.), and $\mathbb{E}[\|w_{i,k+1}\||\ \mathcal{F}_k] \leq c_i(1 + \|x_k\|)$ a.s. for some constants $c_i \geq 0$, $i \in \mathcal{I}$.*

A2b. *The stepsize sequences $\{\gamma_{i,k}\}_t$, $i \in \mathcal{I}$ are positive scalars satisfying: (i) $\sum_i \sum_k \gamma_{i,k}^2 < \infty$; (ii) $\sum_k \gamma_{i,k} = \infty$, $i \in \mathcal{I}$; (iii) $\gamma_{2,k} = o(\gamma_{1,k})$.*

A2c. *Each $f_i \in C^q(\mathbb{R}^d, \mathbb{R})$ for some $q \geq 3$ and each f_i and ω are L_i– and L_ω–Lipschitz, respectively.*

Assumption 3. *For fixed x_2, $\dot{x}_1(t) = -D_1 f_1(x_1(t), x_2)$ has a globally asymptotically stable equilibrium $\lambda(x_2)$.*

4.1 ASYMPTOTIC GUARANTEES

The following lemma follows from classical analysis (see, e.g., Borkar (2008, Chap. 6) or Bhatnagar and Prasad (2013, Chap. 3)). Define the event $\mathcal{E} = \{\sup_k \sum_i \|x_{i,k}\|_2 < \infty\}$.

Lemma 1. *Under Assumptions 2 and 3, conditioned on the event $\mathcal{E}$, $(x_{1,k}, x_{2,k}) \to \{(\lambda(x_2), x_2)|\ x_2 \in \mathbb{R}^{d_2}\}$ a.s.*

Let $t_k = \sum_{l=0}^{k-1} \gamma_{2,k}$ be the continuous time accumulated after k samples of x_2. Define $x_2(t, s, x_s)$ for $t \geq s$ to be the trajectory of $\dot{x}_2 = -D_2 f_2(\lambda(x_2), x_2)$.

Theorem 2. *Under Assumptions 2 and 3 hold, for any $K > 0$, $\lim_{k\to\infty} \sup_{0\leq h\leq K} \|x_{2,k+h} - x_2(t_{k+h}, t_k, x_k)\|_2 = 0$ conditioned on $\mathcal{E}$.*

Proof The proof invokes Lemma 1 above and (Benaïm, 1999, Prop. 4.1 and 4.2). Indeed, by Lemma 1, $(\lambda(x_{2,k}) - x_{2,k}) \to 0$ a.s. Hence, we can study the sample path generated by $x_{2,k+1} = x_{2,k} - \gamma_{2,k}(D_2 f_2(\lambda(x_{2,k}), x_{2,k}) + w_{2,k+1})$. Since $D_2 f_2 \in C^{q-1}$ for some $q \geq 3$, it is locally Lipschitz and, on the event $\mathcal{E}$, it is bounded. It thus induces a continuous globally integrable vector field, and therefore satisfies the assumptions of Prop. 4.1 of (Benaïm, 1999). Moreover, under Assumption 2, the assumptions of Prop. 4.2 of (Benaïm, 1999) are satisfied. Invoking said propositions gives the desired result. ■

This result essentially says that the slow player's sample path asymptotically tracks the flow of $\dot{x}_2 = -D_2 f_2(\lambda(x_2), x_2)$. If we additionally assume that the slow component also has a global attractor, then the above theorem gives rise to a stronger convergence result.

Assumption 4. *Given $\lambda(\cdot)$ as in Assumption 3, $\dot{x}_2(t) = -\tau D_2 f_2(\lambda(x_2(t)), x_2(t))$ has a globally asymptotically stable equilibrium x_2^*.*

Corollary 1. *Under the assumptions of Theorem 2 and Assumption 4, conditioned on $\mathcal{E}$, gradient-based learning converges a.s. to a stable attractor (x_1^*, x_2^*) where $x_1^* = \lambda(x_2^*)$, the set of which contains the stable differential Nash equilibria.*

More generally, the process $(x_{1,k}, x_{2,k})$ will converge almost surely to the internally chain transitive set of the limiting dynamics (5) and this set contains the stable Nash equilibria. If the only internally chain transitive sets for (5) are isolated equilibria (this occurs, e.g., if the game is a potential game), then x_k converges almost surely to a stationary point of the dynamics, a subset of which are stable local Nash equilibria. It is also worth commenting on what types of games will satisfy these assumptions. To satisfy Assumption 3, it is sufficient that the fastest player has a convex cost in their choice variable.

Proposition 4. *Suppose Assumption 2 and 4 hold and that $f_1(\cdot, x_2)$ is convex. Conditioned on $\mathcal{E}$, the sample points of gradient-based learning satisfy $(x_{1,k}, x_{2,k}) \to \{(\lambda(x_2), x_2)|\ x_2 \in \mathbb{R}^{d_2}\}$ a.s. Moreover, $(x_{1,k}, x_{2,k}) \to (x_1^*, x_2^*)$ a.s., where $x_1^* = \lambda(x_2^*)$.*

Note that (x_1^*, x_2^*) could still be a spurious stable non-Nash point still since the above implies $D(D_2 f_2(\lambda(\cdot), \cdot))|_{x_2^*} > 0$ which does not imply that necessarily $D_2^2 f_2(\lambda(x_2^*), x_2^*) > 0$.

Remark 3 (Relaxing Assumptions: Local Asymptotical Stability)**.** *Under relaxed assumptions on* global asymptotic stability *(i.e., if Assumptions 3 and 4 are relaxed to local asymptotic stability), we can obtain high-probability results on convergence to locally asymptotically stable attractors. However, this requires conditioning on an unverifiable event—i.e. the high-probability bound in this case is conditioned on the event $\{\{x_{1,k}\}$ belongs to a compact set B, which depends on the sample point, of $\cap_{x_2} \mathcal{R}(\lambda(x_2))\}$ where $\mathcal{R}(\lambda(x_2))$ is the region of attraction of $\lambda(x_2)$. None-the-less, it is possible to leverage results from stochastic approximation (Karmakar and*

Bhatnagar, 2018), (Borkar, 2008, Chap. 2) to prove local versions of the results for non-uniform learning rates. Further investigation is required to provide concentration bounds for not only games but stochastic approximation in general.

4.2 CONCENTRATION BOUNDS

In the stochastic setting, the learning dynamics are stochastic approximation updates, and non-uniform learning rates lead to a multi-timescale setting. The concentration bounds we derive leverage very recent results—e.g., (Borkar and Pattathil, 2018)—from stochastic approximation and we note that our objective here is to show that they apply to games and provide commentary on the interpretation of the results in this context.

For a stable differential Nash equilibrium $x^* = (\lambda(x_2^*), x_2^*)$, using the bounds in Lemma 1 and Lemma 2 in Appendix A.2, we can provide a high-probability guarantee that $(x_{1,k}, x_{2,k})$ gets locked in to a ball around $(\lambda(x_2^*), x_2^*)$.

Let $\bar{x}_i(\cdot)$ denote the linear interpolates between sample points $x_{i,k}$ and, as in the preceding sub-section, let $x_i(\cdot, t_{i,k}, x_k)$ denote the continuous time flow of $\dot{x}_i$ with initial data $(t_{i,k}, x_k)$ where $t_{i,k} = \sum_{l=0}^{k-1} \gamma_{i,k}$. Define also $\tau_k = \gamma_{2,k}/\gamma_{1,k}$. Alekseev's formula is a nonlinear variation of constants formula that provides solutions to perturbations of differential equations using a local linear approximation. We can apply this to the *asymptotic pseudo-trajectories* $\bar{x}_i(\cdot)$ in each timescale. For these local approximations, linear systems theory let's us find growth rate bounds for the perturbations, which can, in turn, be used to bound the normed difference between the continuous time flow and the asymptotic pseudo-trajectories. More detail is provided in Appendix A.2.

Towards this end, fix $\varepsilon \in [0, 1)$ and let N be such that $\gamma_{1,n} \leq \varepsilon/(8K)$, $\tau_n \leq \varepsilon/(8K)$ for all $n \geq N$. Define $t_{1,k} = \tilde{t}_k$ and $t_{2,k} = \hat{t}_k$. Let $n_0 \geq N$ and with K as in Lemma 1 (Appendix A.2), let T be such that $e^{-\kappa_1(\tilde{t}_n - \tilde{t}_{n_0})} H_{n_0} \leq \varepsilon/(8K)$ for all $n \geq n_0 + T$ where $\kappa_1 > 0$ is a constant derived from Alekseev's formula applied to $\bar{x}_1(\cdot)$. Moreover, with $\bar{K}$ as in Lemma 2 (Appendix A.2), let $e^{-\kappa_2(\hat{t}_n - \hat{t}_{n_0})}(\|\bar{x}_2(\hat{t}_{n_0}) - x_2(\hat{t}_{n_0})\| \leq \varepsilon/(8\bar{K})$, $\forall n \geq n_0 + T$ where $\kappa_2 > 0$ is a constant derived from Alekseev's formula applied to $\bar{x}_2(\cdot)$.

Theorem 3. *Suppose that Assumptions 2–4 hold and let $\gamma_{2,k} = o(\gamma_{1,k})$. Given a stable differential Nash equilibrium $x^* = (\lambda(x_2^*), x_2^*)$, the player 2's sample path generated by* (3) *for $i = 1$ will asymptotically track $z_k = \lambda(x_{2,k})$, and given $\varepsilon \in [0, 1)$, x_k will get 'locked in' to a ε–neighborhood with high probability conditioned on reaching $B_{r_0}(x^*)$ by iteration n_0. That is, letting $\bar{n} = n_0 + T + 1$, for some $C_1, C_2 > 0$,*

$$\begin{aligned}
&\mathrm{P}(\|x_{1,n} - z_n\| \leq \varepsilon, \forall n \geq \bar{n} | x_{1,n_0}, z_{n_0} \in B_{r_0}) \\
&\quad \geq 1 - \textstyle\sum_{n=n_0}^{\infty} C_1 \exp\big(-C_2\sqrt{\varepsilon}/\sqrt{\gamma_{1,n}}\big) \\
&\qquad - \textstyle\sum_{n=n_0}^{\infty} C_2 \exp\big(-C_2\sqrt{\varepsilon}/\sqrt{\tau_n}\big) \\
&\qquad - \textstyle\sum_{n=n_0}^{\infty} C_1 \exp\big(-C_2\varepsilon^2/\beta_n\big). \qquad (6)
\end{aligned}$$

with $\beta_n = \max_{n_0 \leq k \leq n-1} e^{-\kappa_1(\sum_{i=k+1}^{n-1} \gamma_{1,i})} \gamma_{1,k}$. Moreover, for some constants $\tilde{C}_1, \tilde{C}_2 > 0$,

$$\begin{aligned}
&\mathrm{P}(\|x_{2,n} - x_2(\hat{t}_n)\| \leq \varepsilon, \forall n \geq \bar{n} | x_{n_0}, z_{n_0} \in B_{r_0}) \\
&\quad \geq 1 + \textstyle\sum_{n=n_0}^{\infty} \tilde{C}_1 \exp\big(-\tilde{C}_2\sqrt{\varepsilon}/\sqrt{\gamma_{1,n}}\big) \\
&\qquad - \textstyle\sum_{n=n_0}^{\infty} \tilde{C}_1 \exp\big(-\tilde{C}_2\sqrt{\varepsilon}/\sqrt{\tau_n}\big) \\
&\qquad - \textstyle\sum_{n=n_0}^{\infty} \tilde{C}_1 \exp\big(-\tilde{C}_2\varepsilon^2/\beta_n\big) \\
&\qquad - \textstyle\sum_{n=n_0}^{\infty} \tilde{C}_1 \exp\big(-\tilde{C}_2\varepsilon^2/\eta_n\big) \qquad (7)
\end{aligned}$$

with $\eta_n = \max_{n_0 \leq k \leq n-1} \big(e^{-\kappa_2(\sum_{i=k+1}^{n-1} \gamma_{2,i})} \gamma_{2,k}\big)$.

Corollary 2. *Fix $\varepsilon \in [0, 1)$ and suppose that $\gamma_{1,n} \leq \varepsilon/(8K)$ for all $n \geq 0$. With K as in Lemma 1 (Appendix A.2), let T be such that $e^{-\kappa_1(\tilde{t}_n - \tilde{t}_0)} H_0 \leq \varepsilon/(8K)$ for all $n \geq T$. Furthermore, with $\bar{K}$ as in Lemma 2 (Appendix A.2), let $e^{-\kappa_2(\hat{t}_n - \hat{t}_0)}(\|\bar{x}_2(\hat{t}_0) - x_2(\hat{t}_0)\| \leq \varepsilon/(8\bar{K})$, $\forall n \geq T$. Under the assumptions of Theorem 3, x_k will will get 'locked in' to a ε–neighborhood with high probability conditioned on $x_0 \in B_{r_0}(x^*)$ where the high-probability bounds in* (6) *holds with $n_0 = 0$.*

The key technique in proving the above theorem (which is done in detail in Borkar and Pattathil (2018) which is, in turn, leveraging results from Thoppe and Borkar (2018)), is first to compute the errors between the sample points from the stochastic learning rules and the continuous time flow generated by initializing the continuous time limiting dynamics at each sample point and flowing it forward for time $t_{n+1} - t_n$, doing this for each $x_{1,k}$ and $x_{2,k}$ separately and in their own timescale, and then take a union bound over all the continuous time intervals defined for $n \geq n_0$.

In Appendix A.3, we specialize to the case of uniform learning rates for which we can tighter bounds leveraging the results of (Thoppe and Borkar, 2018).

5 NUMERICAL EXAMPLES

We consider several examples that illustrate the effect that non-uniform learning rates have on the stability of the learning dynamics, its vector field and resulting equilibria of continuous games. These examples highlight the importance of studying the convergence properties of game dynamics in non-cooperative continuous games where agents may learn at different rates. Additional examples are provided in Appendix B.

5.1 EFFECTS OF PRECONDITIONG

Although the fixed points of the game dynamics are invariant under change of learning rates, i.e. the solutions to $\omega = 0$ and $\Gamma\omega = 0$ are the same for diagonal Γ, the stability properties near such fixed points may change. The following example illustrates a counter-intuitive but non-degenerate situation in which an agent that decides to learn slower causes a stable differential Nash equilibrium to become unstable.

Consider a three-player continuous game where the Jacobian at a fixed point has positive-definite block diagonals and strictly positive eigenvalues. This implies the fixed point is a Nash equilibrium and that the dynamics $\dot{x} = -\omega(x)$ are stable in a neighborhood around the fixed point, say x^*. Now suppose an agent decides to slow down its learning by five times, from γ to $\gamma/5$. We show via this simple example that this change can cause the learning dynamics to become unstable.

Suppose the Jacobian of the continuous time learning dynamics at the fixed point is

$$J(x^*) = \begin{bmatrix} 2 & 9 & 0 \\ 0 & 2 & 6 \\ 9 & 0 & 12 \end{bmatrix},$$

whose spectrum lives on the right half complex plane with eigenvalues $\{14.9,\ 0.5 \pm 6.0i\}$. However, precondtioning the Jacobian with $\Gamma = \text{diag}(1,\ 1,\ 1/5)$, the eigenvalues of ΓJ are $\{6.7,\ -0.2 \pm 4.0i\}$, which indicate a saddle point.

5.2 TORUS GAME

The second example is a two-player game with agents' joint strategy space on a torus. This example serves as a useful benchmark example because it has multiple equilibria and they are completely characterizable. We visualize the warping of the region of attraction of these equilibria under different learning rates, and the affinity of the "faster" player to its own zero line.

Each player's strategy space is the unit circle $\mathbb{S}^1$ and have cost $f_i : \mathbb{S}^1 \times \mathbb{S}^1$ given by

$$\begin{bmatrix} f_1(\theta_1, \theta_2) \\ f_2(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} -\alpha_1 \cos(\theta_1 - \phi_1) + \cos(\theta_1 - \theta_2) \\ -\alpha_2 \cos(\theta_2 - \phi_2) + \cos(\theta_2 - \theta_1) \end{bmatrix}$$

where α_i and ϕ_i are constants and θ_i is player i's choice variable. An interpretation of these costs is that each player wishes to be near ϕ_i but far from each other. This game has many applications including those which abstract nicely to coupled oscillators. The vector of individual gradients is given by

$$\omega(\theta_1, \theta_2) = \begin{bmatrix} \alpha_1 \sin(\theta_1 - \phi_1) - \sin(\theta_1 - \theta_2) \\ \alpha_2 \sin(\theta_2 - \phi_2) - \sin(\theta_2 - \theta_1) \end{bmatrix}, \quad (8)$$

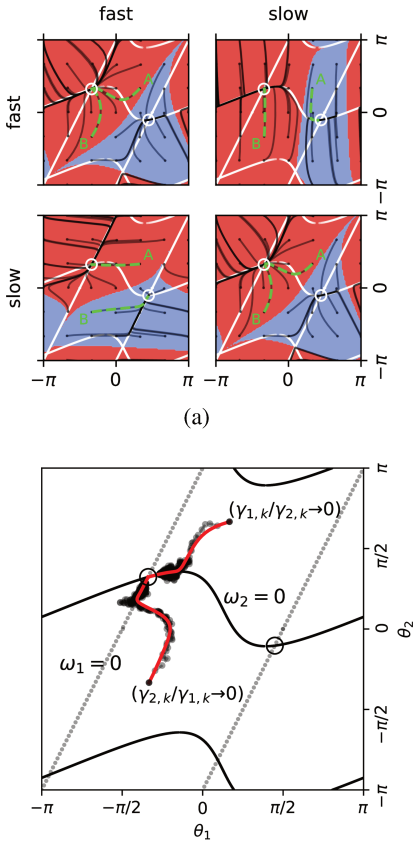

Figure 1: The effects of non-uniform learning rates on the path of convergence to the equilibria. The zero lines for each player ($D_1 f_1 = 0$ or $D_2 f_2 = 0$) are plotted as the diagonal and curved lines, and the two stable Nash equilibria as circles (where $D_1^2 f_1 > 0$ and $D_2^2 f_2 > 0$). (a) In the deterministic setting, the region of attractions for each equilibrium can be computed numerically. Four scenarios are shown, with a combination of fast and slow agents. The region of attractions for each Nash equilibrium are warped under different learning rates. (b) In the stochastic setting, the samples (in black) approximate the differential equation $\dot{\theta} = -\Lambda\omega(\theta)$ (in red), where $\Lambda = \text{diag}(1, \tau)$ for $\gamma_{2,k} = o(\gamma_{1,k})$ and $\Lambda = \text{diag}(\tau, 1)$ for $\gamma_{1,k} = o(\gamma_{2,k})$. Two initializations and learning rate configurations are plotted.

and its game Jacobian has terms $\alpha_i \cos(\theta_i - \phi_i) - \cos(\theta_i - \theta_{-i})$, $i = 1, 2$ on the diagonal and $\cos(\theta_i - \theta_{-i})$, $i = 1, 2$ on the off diagonal.

The Nash equilibria of this game occur where $\omega(\theta_1, \theta_2) = 0$ and where the diagonals of the game Jacobian are positive. Using constants $\phi = (0,\ \pi/8)$ and $\alpha = (1.0,\ 1.5)$, there are two Nash equilibria situated at $(-1.063,\ 1.014)$ and $(1.408,\ -0.325)$, respectively. These equilibria happen to also be stable differential Nash equilibria, thus we expect the gradient dynamics to converge to them. Which equilibrium they converge to, however, depends on the

initialization and learning rates of agents.

In the deterministic setting, the fastest (largest) learning rate which guarantee convergence to either Nash is determined to be $\gamma_{\texttt{fast}} = 0.171$ by Theorem 1. We set the slow agent to learn $10\times$ slower, i.e. $\gamma_{\texttt{slow}} = 0.017$. Figure 1a shows the convergence of agents' strategies to the Nash equilibria using non-uniform learning rates. Each of the four squares depicts the full strategy space on the torus from $-\pi$ to π for both agents' actions, with agent 1 on the x-axis and agent 2 on the y-axis. The labels "fast" and "slow" indicate the learning rate of the corresponding agent. For example, in the bottom left square, player 1 is the fast player and player 2 is the slow player. Hence, the non-uniform update equation for that square is $\theta_{k+1} = \theta_k - \text{diag}(\gamma_{\texttt{fast}}, \gamma_{\texttt{slow}})\omega(\theta_k)$. Analogous dynamics can be constructed for the other three squares using their indicated learning rates.

The white lines indicate the points in which $\omega_i \equiv 0$, and the intersection of the white lines indicate where $\omega \equiv 0$. Two of such intersections, marked as circles, are the Nash equilibria; the unmarked intersections are either saddle points or other unstable equilibria. The black lines show paths of the update equations under the non-uniform update equation, with initial points selected from a equally spaced 7×7 grid. Two of such paths are highlighted in green (labeled A and B), beginning at $(\pi/3,\ \pi/3)$ and $(-\pi/3, -\pi/3)$. In the case where agents both learn at the same rate, $(\gamma_{\texttt{fast}}, \gamma_{\texttt{fast}})$ and $(\gamma_{\texttt{slow}}, \gamma_{\texttt{slow}})$, paths A and B both converge to the Nash equilibrium at $(-1.063,\ 1.014)$. However, when agents learn at different rates, the equilibrium that the paths converge to are no longer the same. This phenomena can also be captured by displaying the region of attraction for both Nash equilibria. The red region corresponds to points when initialized there, will converge to one specific equilibrium; the blue region corresponds to the region of attraction of the other equilibria.

In the stochastic setting, the learning rates for each agent must satisfy Assumption 2. We choose scaled learning rates $\gamma_{\texttt{slow},k} = \frac{1}{1+k\log(k+1)}$ and $\gamma_{\texttt{fast},k} = \frac{1}{1+k}$ such that $\gamma_{\texttt{slow},k}/\gamma_{\texttt{fast},k} \to 0$ as $k \to \infty$. Figure 1b shows the result of such game, initialized at two different points, each with different flipped learning rate configurations. We observe that the sample points approximate the differential equation $\dot{\theta} = -\Lambda\omega(\theta)$ shown in red, where $\Lambda = \text{diag}(1, \tau)$ for $\gamma_{2,k} = o(\gamma_{1,k})$ and $\Lambda = \text{diag}(\tau, 1)$ for $\gamma_{1,k} = o(\gamma_{2,k})$.

In both deterministic and stochastic settings, we observe the affinity of the faster agent to its own zero line. For example, the bottom left square (in Figure 1a) and bottom left path (in Figure 1b) both have agent 1 as the faster agent, and the learning paths both tend to arrive to the line $\omega_1 \equiv 0$ before finally converging to the Nash equilibrium.

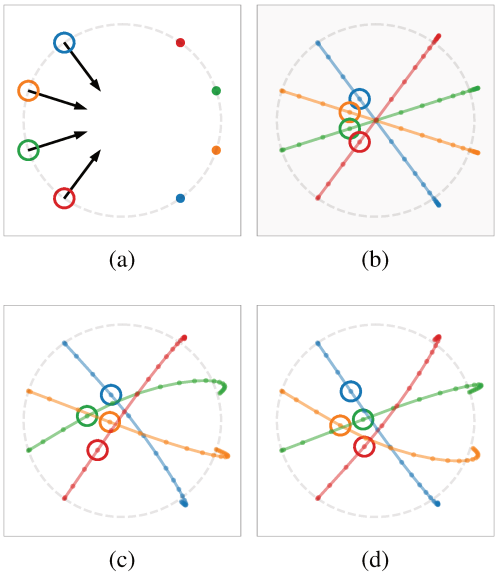

Figure 2: Minimum-fuel particle avoidance control example. (a) Each particle seeks to reach the opposite side of the circle using minimum fuel while avoiding each other. The circles represent the approximate boundaries around each particle at time $t = 5$. (b) The joint strategy $x = (u_1, \cdots, u_4)$ is initialized to the minimum fuel solution ignoring interaction between particles. (c) Equilibrium solution achieved by setting the blue agent to have a slower learning rate. (d) Another equilibrium, where the red agent has the slower learning rate.

An interpretation of this is that the faster agent tries to be situated at the bottom of the "valley" of its own cost function. The faster agent tends to be at its *own* minimum while it waits for the slower agent to change its strategy. As a Stackelberg interpretation, where there are followers and leaders, the slower agent would be the leader and faster agent the follower. In a sense, the slower agent has an advantage.

5.3 CONTROL AND COLLISION AVOIDANCE

The last example is a non-cooperative game between four collision-avoiding agents, in which they seek to arrive a destination with minimum fuel while avoiding each other. In this example, different scaling between agents' learning rate can dictate which equilibrium solution it converges to. This can be useful in designing non-cooperative open-loop controllers where agents may choose to learn slower in order to deviate less from its initial plan, perhaps in an attempt incurs less risk.

We present an example with $n = 4$ collision-avoiding particles traversing across the unit circle. Each particle follows discrete-time linear dynamics

$$z_i(t+1) = Az_i(t) + Bu_i(t)$$

where

$$A = \begin{bmatrix} I & hI \\ 0 & I \end{bmatrix} \in \mathbb{R}^{4\times 4},\ B = \begin{bmatrix} h^2 I \\ hI \end{bmatrix} \in \mathbb{R}^{4\times 2},$$

I is the identity matrix and $h = 0.1$. These dynamics represent a typical discretized version of the continuous dynamics $\ddot{r} = u$ in which u represents a $\mathbb{R}^2$ force vector used to accelerate the particle, and the state $z = [r, \dot{r}]$ represents its position and velocity. Let $u = (u_1, \cdots, u_n)$ and u_i be the concatenated vector of control vectors for all time, i.e $u_i = (u_i(1), \cdots, u_i(N))$ Each particle has cost

$$J_i(u_i, u_{-i}) = \sum_{t=1}^{N} \|u_i(t)\|_R^2 + \sum_{t=1}^{N+1} \|z_i(t) - \bar{z_i}\|_Q^2 + \sum_{j\neq i} \sum_{t=1}^{N+1} \rho e^{-\sigma \|z_i(t) - z_j(t)\|_S^2}$$

where the norm $\|\cdot\|_P$ is defined for positive semi-definite P by $\|z\|_P^2 = z^T P z$. The first two terms of the cost correspond to the minimum fuel objective and quadratic cost from desired final state $\bar{z}_i$, a typical setup for optimal control problems. We use $R = \mathrm{diag}(0.1, 0.1)$ and $Q = \mathrm{diag}(1, 1, 0, 0)$. The final term of the cost function is the sum of all pairwise interaction terms between the particles, modeled after the shape of a Gaussian which encodes smooth boundaries around the particles. We use constants $\rho = 10$ and $\sigma = 100$.

Figure 2(a) is a visualization of the problem setup. Each particles' initial position $z_i(0)$ is located on the left side of a unit circle, separated by $\pi/5$, and their desired final positions $\bar{z}_i$ are located directly opposite. The particles begin with zero velocity and must solve for a minimum control solution that also avoids collision with other particles.

We first initialize the problem with the optimal solution for each agent ignoring the pairwise interaction terms, shown in Figure 2(b) . This can be computed using classical discrete-time LQR methods or by gradient descent. Then each agent descends their own gradient of the full cost,

$$u_{i,k+1} = u_{i,k} - \gamma_i D_i J_i(u_{i,k}, u_{-i,k}),$$

with different learning rates γ_i to converge to the differential Nash equilibrium. These equilibria are shown in Figure 2(c) and 2(d).

6 DISCUSSION

We analyze the convergence of gradient-based learning for non-cooperative agents with continuous costs and non-uniform learning rates. In the deterministic setting where agents have oracle gradient access, we provide non-asymptotic rates of convergence. In the stochastic setting where agents have unbiased gradient estimates, we leverage dynamical systems theory and stochastic approximation analysis techniques to provide concentration bounds.

By preconditioning the gradient dynamics by Γ, a diagonal matrix composed o the agents' learning rates, we can begin to understand how changing a learning rate relative to others can alter the properties of the fixed points of the dynamics. Different learning rates amongst agents also affect the region of attraction of the game, hence starting from the same initial condition one may converge to a different equilibria. Agents may use this to their benefit, as shown in the last example. Such insights into the learning behavior of agents will be useful for providing guarantees on the design of control or incentive policies to coordinate agents.

While we present the work in the context of gradient-based learning in games, there is nothing that precludes the results from applying to update rules in other frameworks. Our results will apply to many other settings where agents myopically update their decision using a process of the form $x_{k+1} = x_k - \Gamma g(x_k)$. In this paper, we consider the special case where $g \equiv [D_1 f_1 \cdots D_n f_n]$. In the stochastic setting, variants of multi-agent Q-learning conform to this setting since Q-learning can be written as a stochastic approximation update.

As pointed out in (Mazumdar and Ratliff, 2018), not all critical points of the dyanamics $\dot{x} = -\omega(x)$ that are attracting are necessarily Nash equilibria; one can see this simply by constructing a Jacobian with positive eigenvalues with at least one $D_i^2 f_i$ with a non-positive eigenvalue. Understanding this phenomena will help us develop computational techniques to avoid them. Recent work has explored this in the context of zero-sum games (Mazumdar et al., 2019), requiring coordination amongst the learning agents. However, when our objective is to study the learning behavior of autonomous agents seeking an equilibrium, an alternative perspective is needed.

References

I. K. Argyros. A generalization of ostrowski's theorem on fixed points. *Applied Mathematics Letters*, 12:77–79, 1999.

David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The Mechanics of n-Player Differentiable Games. *CoRR*, abs/1802.05642, 2018.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de Probabilites XXXIII*, pages 1–68, 1999.

Michel Benaïm and Morris W. Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in

perturbed games. *Games and Economic Behavior*, 29 (1-2):36–72, 1999.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Perturbations of set-valued dynamical systems, with applications to game theory. *Dynamic Games and Applications*, 2(2):195–205, 2012.

S. Bhatnagar and H. L. Prasad. *Stochastic Recursive Algorithms for Optimization*. Springer, 2013.

Vivek S. Borkar and Sarath Pattathil. Concentration bounds for two time scale stochastic approximation. *arxiv:1806.10798*, 2018.

V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2008.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Traning GANs with Optimism. *arxiv:1711.00141*, 2017.

Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, December 2003.

J. Heinrich and D. Silver. Deep reinforcement learning from self-play in imperfect-information games. *arxiv:1603.01121*, 2016.

Josef Hofbauer. Evolutionary dynamics for bimatrix games: A hamiltonian system? *Journal of Mathematical Biology*, 34(5):675, May 1996.

Cars H. Hommes and Marius I. Ochea. Multiple equilibria and limit cycles in evolutionary games with logit dynamics. *Games and Economic Behavior*, 74(1):434 –441, 2012.

Prasenjit Karmakar and Shalabh Bhatnagar. Two timescale stochastic approximation with controlled markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research*, 2018.

E. Mazumdar and L. J. Ratliff. On the convergence of competitive, multi-agent gradient-based learning algorithms. *arxiv:1804.05464*, 2018.

E. Mazumdar, M. Jordan, and S. S. Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arxiv:1901.00838*, 2019.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173 (1–2):456–507, 2019.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. 16th Intern. Conf. Machine Learning*, pages 278–287, 1999.

A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, 1966.

Christos H. Papadimitriou and G. Piliouras. Game dynamics as the meaning of a game. *Sigecom*, 2018.

L. J. Ratliff and T. Fiez. Adaptive incentive design. *arxiv:1806.05749*, 2018.

L. J. Ratliff, S. A. Burden, and S. S. Sastry. Characterization and computation of local Nash equilibria in continuous games. In *Proc. 51st Ann. Allerton Conf. Communication, Control, and Computing*, pages 917–924, 2013.

L. J. Ratliff, S. A. Burden, and S. S. Sastry. On the Characterization of Local Nash Equilibria in Continuous Games. *IEEE Transactions on Automatic Control*, 61 (8):2301–2307, Aug 2016.

Lillian J. Ratliff, Samuel A. Burden, and S. Shankar Sastry. Generictiy and Structural Stability of Non–Degenerate Differential Nash Equilibria. In *Proc. 2014 Amer. Controls Conf.*, 2014.

G. Thoppe and V. S. Borkar. A concentration bound for stochastic approximation via alekseev's formula. *arXiv:1506.08657v3*, 2018.

Karl Tuyls, Julien Pérolat, Marc Lanctot, Georg Ostrovski, Rahul Savani, Joel Z Leibo, Toby Ord, Thore Graepel, and Shane Legg. Symmetric decomposition of asymmetric games. *Scientific Reports*, 8(1):1015, 2018.

A PROOFS

A.1 Deterministic Setting

The following proof follows nearly the same proof as the main result in Argyros (1999) with a few minor modifications in the conclusion; we provide it here for posterity.

Proof [Proof Proposition 3] Since $\|I-\Gamma D\omega(x)\| < 1$ for each $x \in B_{r_0}(x^*)$, as stated in the proposition statement, there exists $0 < r' < r'' < 1$ such that $\|I-\Gamma D\omega(x)\| \leq r' < r'' < 1$ for all $x \in B_r(x^*)$. Since

$$\lim_{x\to x^*} \|R(x-x^*)\|/\|x-x^*\| = 0,$$

for $0 < 1 - r'' < 1$, there exists $\tilde{r} > 0$ such that

$$\|R(x-x^*)\| \leq (1-r'')\|x-x^*\|, \ \ \forall\, x \in B_{\tilde{r}}(x^*).$$

As in the proposition statement, let r be the largest, finite such $\tilde{r}$. Note that for arbitrary $c > 0$, there exists $\tilde{r} > 0$ such that the bound on $\|R(x-x^*)\|$ holds; hence, we choose $c = 1 - r''$ and find the largest such $\tilde{r}$ for which the bound holds. Combining the above bounds with the definition of g, we have that

$$\|g(x) - g(x^*)\| \leq (1-\delta)\|x-x^*\|, \ \ \forall\, x \in B_{r^*}(x^*)$$

where $\delta = r'' - r'$ and $r^* = \min\{r_0, r\}$. Hence, applying the result iteratively, we have that

$$\|x_t - x^*\| \leq (1-\delta)^t\|x_0 - x^*\|, \ \ \forall\, x_0 \in B_{r^*}(x^*).$$

Note that $0 < 1-\delta < 1$. Using the approximation $1-\delta < \exp(-\delta)$, we have that

$$\|x_T - x^*\| \leq \exp(-T\delta)\|x_0 - x^*\|$$

so that $x_t \in B_\varepsilon(x^*)$ for all $t \geq T = \lceil \delta^{-1}\log(r^*/\varepsilon)\rceil$.

■

As noted in the remark, a similar result holds under the relaxed assumption that $\rho(I-\Gamma D\omega(x)) < 1$ for all $x \in B_{r_0}(x^*)$. To see this, we first note that $\rho(I-\Gamma D\omega(x)) < 1$ implies there exists $c > 0$ such that $\rho(I-\Gamma D\omega(x)) \leq c < 1$. Hence, given any $\epsilon > 0$, there is a norm on $\mathbb{R}^d$ and a $c > 0$ such that $\|I-\Gamma D\omega\| \leq c + \epsilon < 1$ on $B_{r_0}(x^*)$ (Ortega and Rheinboldt, 1970, 2.2.8). Then, we can apply the same argument as above using $r' = c + \varepsilon$.

A.2 Stochastic Setting

A key tool used in the finite-time two-timescale analysis is the nonlinear variation of constants formula of Alekseev Alekseev (1961), Borkar and Pattathil (2018).

Theorem 1. *Consider a differential equation*

$$\dot{u}(t) = f(t, u(t)),\ t \geq 0,$$

and its perturbation

$$\dot{p}(t) = f(t, p(t)) + g(t, p(t)),\ t \geq 0$$

where $f, g : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$, $f \in C^1$, and $g \in C$. Let $u(t, t_0, p_0)$ and $p(t, t_0, p_0)$ denote the solutions of the above nonlinear systems for $t \geq t_0$ satisfying $u(t_0, t_0, p_0) = p(t_0, t_0, p_0) = p_0$, respectively. Then,

$$p(t, t_0, p_0) = u(t, t_0, p_0) + \int_{t_0}^{t} \Phi(t, s, p(s, t_0, p_0)) \cdot g(s, p(s, t_0, p_0))\, ds,\ t \geq t_0$$

where $\Phi(t, s, u_0)$, for $u_0 \in \mathbb{R}^d$, is the fundamental matrix of the linear system

$$\dot{v}(t) = \frac{\partial f}{\partial u}(t, u(t, s, u_0))v(t),\ t \geq s \tag{1}$$

with $\Phi(s, s, u_0) = I_d$, the d–dimensional identity matrix.

Consider a locally asymptotically stable differential Nash equilibrium $x^* = (\lambda(x_2^*), x_2^*) \in X$ and let $B_{r_0}(x^*)$ be an $r_0 > 0$ radius ball around x^* contained in the region of attraction. Stability implies that the Jacobian $J_{\mathcal{S}}(\lambda(x_2^*), x_2^*)$ is positive definite and by the converse Lyapunov theorem (Sastry, 1999, Chap. 5) there exists local Lyapunov functions for the dynamics $\dot{x}_2(t) = -\tau D_2 f_2(\lambda(x_2(t)), x_2(t))$ and for the dynamics $\dot{x}_1(t) = -D_1 f_1(x_1(t), x_2)$, for each fixed x_2. In particular, there exists a local Lyapunov function $V \in C^1(\mathbb{R}^{d_1})$ with $\lim_{\|x_2\|\uparrow\infty} V(x_2) = \infty$, and $\langle \nabla V(x_2), D_2 f_2(\lambda(x_2), x_2)\rangle < 0$ for $x_2 \neq x_2^*$. For $r > 0$, let $V^r = \{x \in \text{dom}(V) :\ V(x) \leq r\}$. Then, there is also $r > r_0 > 0$ and $\epsilon_0 > 0$ such that for $\epsilon < \epsilon_0$, $\{x_2 \in \mathbb{R}^{d_2} |\ \|x_2 - x_2^*\| \leq \epsilon\} \subseteq V^{r_0} \subset \mathcal{N}_{\epsilon_0}(V^{r_0}) \subseteq V^r \subset \text{dom}(V)$ where $\mathcal{N}_{\epsilon_0}(V^{r_0}) = \{x \in \mathbb{R}^{d_2} |\ \exists x' \in V^{r_0}$ s.t.$\|x' - x\| \leq \epsilon_0\}$. An analogously defined $\tilde{V}$ exists for the dynamics $\dot{x}_1$ for each fixed x_2.

For now, fix n_0 sufficiently large; we specify this a bit further down. Define the event $\mathcal{E}_n = \{\bar{x}_1(t) \in V^r\ \forall t \in [\tilde{t}_{n_0}, \tilde{t}_n]\}$ where $\bar{x}_1(t) = x_{1,k} + \frac{t-\tilde{t}_k}{\gamma_{1,k}}(x_{1,k+1} - x_{1,k})$ are linear interpolates defined for $t \in (\tilde{t}_k, \tilde{t}_{k+1})$ with $\tilde{t}_{k+1} = \tilde{t}_k + \gamma_{1,k}$ and $\tilde{t}_0 = 0$. The basic idea of the proof is to leverage Alekseev's formula (Theorem 1) to bound the difference between the linearly interpolated trajectories (i.e., *asymptotic psuedo-trajectories*) and the flow of the corresponding limiting differential equation on each continuous time interval between each of the successive iterates k and $k+1$ by a number that decays asymptotically. Then, for large enough n, a union bound is used

over all the remaining time intervals to construct a concentration bound. This is done first for fast player (i.e. player 1), to show that $x_{1,k}$ tracks $\lambda(x_{2,k})$, and then for the slow player (i.e. player 2).

Following Borkar and Pattathil (2018), we can express the linear interpolates for any $n \geq n_0$ as $\bar{x}_1(\tilde{t}_{n+1}) = \bar{x}_1(\tilde{t}_{n_0}) - \sum_{k=n_0}^{n} \gamma_{1,k}(D_1 f_1(x_k) + w_{1,k+1})$ where

$$\gamma_{1,k} D_1 f_1(x_k) = \int_{\tilde{t}_k}^{\tilde{t}_{k+1}} D_1 f_1(\bar{x}_1(\tilde{t}_k), x_{2,k})$$

and similarly for the $w_{1,k+1}$ term. Adding and subtracting $\int_{\tilde{t}_{n_0}}^{\tilde{t}_{n+1}} D_1 f_1(\bar{x}_1(s), x_2(s),)$, Alekseev's formula can be applied to get

$$\bar{x}_1(t) = x_1(t) + \Phi_1(t, s, \bar{x}_1(\tilde{t}_{n_0}), x_2(\tilde{t}_{n_0}))(\bar{x}_1(\tilde{t}_{n_0}) - x_1(\tilde{t}_{n_0})) + \int_{\tilde{t}_{n_0}}^{t} \Phi_2(t, s, \bar{x}_1(s), x_2(s))\zeta_1(s)\, ds$$

where $x_2(t) \equiv x_2$ is constant (since $\dot{x}_2 = 0$), $x_1(t) = \lambda(x_2)$, $\zeta_1(s) = -D_1 f_1(\bar{x}_1(\tilde{t}_k), x_2(\tilde{t}_k)) + D_1 f_1(\bar{x}_1(s), x_2(s)) + w_{1,k+1}$, and where for $t \geq s$, $\Phi_1(\cdot)$ satisfies linear system

$$\dot{\Phi}_1(t, s, x_0) = J_1(x_1(t), x_2(t))\Phi_1(t, s, x_0),$$

with initial data $\Phi_1(t, s, x_0) = I$ and $x_0 = (x_{1,0}, x_{2,0})$ and where J_1 the Jacobian of $-D_1 f_1(\cdot, x_2)$.

Given that $x^* = (\lambda(x_2^*), x_2^*)$ is a stable differential Nash equilibrium, $J_1(x^*)$ is positive definite. Hence, as in (Thoppe and Borkar, 2018, Lem. 5.3), we can find M, $\kappa_1 > 0$ such that for $t \geq s$, $x_{1,0} \in V^r$, $\|\Phi_1(t, s, x_{1,0}, x_{2,0})\| \leq Me^{-\kappa_1(t-s)}$; this result follows from standard results on stability of linear systems (see, e.g., Callier and Desoer (1991, §7.2, Thm. 33)) along with a bound on $\int_s^t \|D_1^2 f_1(x_1, x_2(\tau, s, \tilde{x}_0)) - D_1^2 f_1(x^*)\| d\tau$ for $\tilde{x}_0 \in V^r$ (see, e.g., (Thoppe and Borkar, 2018, Lem 5.2)).

Consider $z_k = \lambda(x_{2,k})$—i.e., where $D_1 f_1(x_{1,k}, x_{2,k}) = 0$. Then, using a Taylor expansion of the implicitly defined λ, we get

$$z_{k+1} = z_k + D\lambda(x_{2,k})(x_{2,k+1} - x_{2,k}) + \delta_{k+1} \quad (2)$$

where $\|\delta_{k+1}\| \leq L_r \|x_{2,k+1} - x_{2,k}\|^2$ is the error from the remainder terms. Plugging in $x_{2,k+1}$,

$$z_{k+1} = z_k + \gamma_{1,k}(-D_1 f_1(z_k, x_{2,k}) + \tau_k \lambda(x_{2,k}) \cdot (w_{2,k+1} - D_2 f_2(x_{1,k}, x_{2,k})) + \gamma_{1,k}^{-1}\delta_{k+1})$$

The terms after $-D_1 f_1$ are $o(1)$, and hence asymptotically negligible, so that this z sequence tracks dynamics as $x_{1,k}$. We show that with high probability, they asymptotically contract to one another.

Now, let us bound the normed difference between $x_{1,k}$ and z_k.

Define constant $H_{n_0} = (\|\bar{x}_1(\tilde{t}_{n_0} - x_1(\tilde{t}_{n_0})\| + \|\bar{z}(\tilde{t}_{n_0}) - x_1(\tilde{t}_{n_0})\|)$ and

$$S_{1,n} = \sum_{k=n_0}^{n-1} \Big(\int_{\tilde{t}_k}^{\tilde{t}_{k+1}} \Phi_1(\tilde{t}_n, s, \bar{x}_1(\tilde{t}_k), x_2(\tilde{t}_k))ds \Big) \cdot w_{2,k+1}.$$

Let $\tau_k = \gamma_{2,k}/\gamma_{1,k}$.

Lemma 1. *For any $n \geq n_0$, there exists $K > 0$ such that*

$$\|x_{1,n} - z_n\| \leq K\big(\|S_{1,n}\| + e^{-\kappa_1(\tilde{t}_n - \tilde{t}_{n_0})}H_{n_0} + \sup_{n_0 \leq k \leq n-1} \gamma_{1,k} + \sup_{n_0 \leq k \leq n-1} \gamma_{1,k}\|w_{1,k+1}\|^2 + \sup_{n_0 \leq k \leq n-1} \tau_k + \sup_{n_0 \leq k \leq n-1} \tau_k \|w_{2,k+1}\|^2\big)$$

conditioned on $\mathcal{E}_n$.

In order to construct a high-probability bound for $x_{2,k}$, we need a similar bound as in Lemma 1 can be constructed for $x_{2,k}$.

Define the event $\hat{\mathcal{E}}_n = \{\bar{x}_2(t) \in V^r \ \forall t \in [\hat{t}_{n_0}, \hat{t}_n]\}$ where $\bar{x}_2(t) = x_{2,k} + \frac{t - \hat{t}_k}{\gamma_{2,k}}(x_{2,k+1} - x_{2,k})$ is the linear interpolated points between the samples $\{x_{2,k}\}$, $\hat{t}_{k+1} = \hat{t}_k + \gamma_{1,k}$, and $\hat{t}_0 = 0$. Then as above, Alekseev's formula can again be applied to get

$$\bar{x}_2(t) = x_2(t, \hat{t}_{n_0}, x_2(\hat{t}_{n_0})) + \Phi_2(t, \hat{t}_{n_0}, \bar{x}_2(\hat{t}_{n_0})) \cdot (\bar{x}_2(\hat{t}_{n_0}) - x_2(\hat{t}_{n_0})) + \int_{\hat{t}_{n_0}}^{t} \Phi_2(t, s, \bar{x}_2(s))\zeta_1(s)\, ds$$

where $x_2(t) \equiv x_2^*$,

$$\zeta_1(s) = D_2 f_2(\lambda(x_{2,k}), x_{2,k}) - D_2 f_2(\lambda(\bar{x}_2(s)), \bar{x}_2(s)) + D_2 f_2(x_k) - D_2 f_2(\lambda(x_{2,k}), x_{2,k}) + w_{2,k+1},$$

and Φ_2 is the solution to a linear system with dynamics $J_2(\lambda(x_2^*), x_2^*)$, the Jacobian of $-D_2 f_2(\lambda(\cdot), \cdot)$, and with initial data $\Phi_2(s, s, x_{2,0}) = I$. This linear system, as above, has bound $\|\Phi_2(t, s, x_{2,0})\| \leq M_2 e^{\kappa_2(t-1)}$ for some $M_2, \kappa_2 > 0$. Define

$$S_{2,n} = \sum_{k=n_0}^{n-1} \int_{\hat{t}_k}^{\hat{t}_{k+1}} \Phi_2(\hat{t}_n, s, \bar{x}_2(\hat{t}_k))ds \cdot w_{2,k+1}.$$

Lemma 2. *For any $n \geq n_0$, there exists $\bar{K} > 0$ such that*

$$\|\bar{x}_2(\hat{t}_n) - x_2(\hat{t}_n)\| \leq \bar{K}\big(\|S_{2,n}\| + \sup_{n_0 \leq k \leq n-1} \|S_{1,k}\| + \sup_{n_0 \leq k \leq n-1} \gamma_{1,k} + \sup_{n_0 \leq k \leq n-1} \gamma_{1,k}\|w_{1,k+1}\|^2 + \sup_{n_0 \leq k \leq n-1} \tau_k + \sup_{n_0 \leq k \leq n-1} \tau_k \|w_{2,k+1}\|^2 + e^{\kappa_2(\hat{t}_n - \hat{t}_{n_0})}\|\bar{x}_2(\hat{t}_{n_0}) - x_2(\hat{t}_{n_0})\| + \sup_{n_0 \leq k \leq n-1} \tau_k H_{n_0}\big)$$

conditioned on $\tilde{\mathcal{E}}_n$.

Using this lemma, we can get the desired guarantees on $x_{1,k}$.

A.3 Uniform Learning Rates

Before concluding, we specialize to the case in which agents have the same learning rate sequence $\gamma_{i,k} = \gamma_k$ for each $i \in \mathcal{I}$.

Theorem 2. *Suppose that x^* is a stable differential Nash equilibrium of the game $(f_1, \ldots, f_n)$. and that Assumption 2 holds (excluding A2b.iii). For each n, let $n_0 \geq 0$ and $\zeta_n = \max_{n_0 \leq k \leq n-1} \left(\exp(-\lambda \sum_{\ell=k+1}^{n-1} \gamma_\ell)\gamma_k\right.$. Given any $\varepsilon > 0$ such that $B_\varepsilon(x^*) \subset B_r(x^*) \subset \mathcal{V}$, there exists constants $C_1, C_2 > 0$ and functions $h_1(\varepsilon) = O(\log(1/\varepsilon))$ and $h_2(\varepsilon) = O(1/\varepsilon)$ so that whenever $T \geq h_1(\varepsilon)$ and $n_0 \geq N$, where N is such that $1/\gamma_n \geq h_2(\varepsilon)$ for all $n \geq N$, the samples generated by the gradient-based learning rule satisfy*

$$\begin{aligned}&\Pr\left(\bar{x}(t) \in B_\varepsilon(x^*)\ \forall t \geq t_{n_0} + T + 1 |\ \bar{x}(t_{n_0}) \in B_r(x^*)\right)\\ &\geq 1 - \textstyle\sum_{n=n_0}^{\infty} \left(C_1 \exp(-C_2 \varepsilon^{1/2}/\gamma_n^{1/2})\right.\\ &\quad \left. + C_1 \exp(-C_2 \min\{\varepsilon, \varepsilon^2\}/\zeta_n)\right)\end{aligned}$$

where the constants depend only on parameters λ, r, τ_L and the dimension $d = \sum_i d_i$. Then stochastic gradient-based learning in games obtains an ε–stable differential Nash x^* in finite time with high probability.

The above theorem implies that $x_k \in B_\varepsilon(x^*)$ for all $k \geq n_0 + \lceil \log(4\tilde{K}/\varepsilon)\lambda^{-1} \rceil + 1$ with high probability for some constant $\tilde{K}$ that depends only on λ, r, τ_L, and d.

Proof Since x^* is a stable differential Nash equilibrium, $J(x^*)$ is positive definite and $D_i^2 f_i(x^*)$ is positive definite for each $i \in \mathcal{I}$. Thus x^* is a locally asymptotically stable hyperbolic equilibrium point of $\dot{x} = -\omega(x)$. Hence, the assumptions of Theorem 1.1 Thoppe and Borkar (2018) are satisfied so that we can invoke the result which gives us the high probability bound for stochastic gradient-based learning in games. ■

B ADDITIONAL EXAMPLES

In this section, we provide additional numerical examples.

B.1 Matching pennies

This example is a classic bimatrix game—matching pennies—where agents have zero-sum costs associated with the matrices (A, B) below. We parameterize agents with a "soft" $\arg\max$ policy where they play smoothed best-response. This game has been well studied in the game theory literature and we use this example illustrate the warping of agent's vector field under non-uniform learning rates.

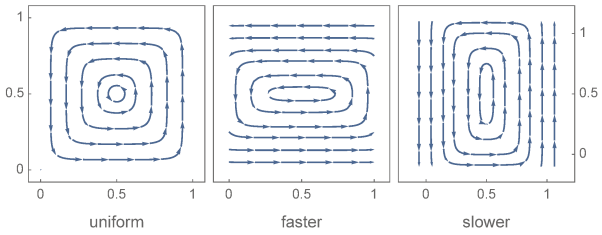

Figure 1: Gradient dynamics of the matching pennies game where agents learning have different learning rates. The vector field of the gradient dynamics are stretched along the faster agent's coordinate.

Consider the zero-sum bimatrix game with

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

where each agent's costs are $f_1(x, y) = \pi(y)^T A \pi(x)$ and $f_2(x, y) = \pi(x)^T B \pi(y)$, and soft $\max$ policy as

$$\pi(z) = \left[\frac{e^{10z}}{e^{10z} + e^{10(1-z)}}, \quad \frac{e^{10(1-z)}}{e^{10z} + e^{10(1-z)}}\right].$$

The mixed Nash equilibrium for this game is $(x^*,\ y^*) = (0.5, 0.5)$, but the Jacobian of the gradient dynamics at this fixed point is

$$J(x^*,\ y^*) = \begin{bmatrix} 0 & 100 \\ -100 & 0 \end{bmatrix}$$

and has purely imaginary eigenvalues $\pm 100i$, therefore admits a limit cycle. Regardless, we can visualize the effects of non-uniform learning rates to the gradient dynamics in Figure 1. We notice that the gradient flow stretches along the axes of the faster agent (the agent with a larger learning rate). However, the fixed point of these dynamics does not change.

References

V. M. Alekseev. An estimate for the perturbations of the solutions of ordinary differential equations. *Vestnik Moskov. Univ. Ser. I. Mat. Meh.*, 2:28–36, 1961.

I. K. Argyros. A generalization of ostrowski's theorem on fixed points. *Applied Mathematics Letters*, 12:77–79, 1999.

Vivek S. Borkar and Sarath Pattathil. Concentration bounds for two time scale stochastic approximation. *arxiv:1806.10798*, 2018.

F. Callier and C. Desoer. *Linear Systems Theory*. Springer, 1991.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solutions to Nonlinear Equations in Several Variables*. Academic Press, 1970.

Shankar Sastry. *Nonlinear Systems*. Springer New York, 1999. doi: 10.1007/978-1-4757-3108-8.

G. Thoppe and V. S. Borkar. A concentration bound for stochastic approximation via alekseev's formula. *arXiv:1506.08657v3*, 2018.