Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality

Anonymous Author(s)
Affiliation
Address
email

Abstract

The interplay between exploration and exploitation in competitive multi-agent learning is still far from being well understood. Motivated by this, we study smooth Q-learning, a prototypical learning model that explicitly captures the balance between game rewards and exploration costs. We show that Q-learning always converges to the unique quantal-response equilibrium (QRE), the standard solution concept for games under bounded rationality, in weighted zero-sum polymatrix games with heterogeneous learning agents using positive exploration rates. Complementing recent results about convergence in weighted potential games [15, 34], we show that fast convergence of Q-learning in competitive settings obtains regardless of the number of agents and without any need for parameter fine-tuning. As showcased by our experiments in network zero-sum games, these theoretical results provide the necessary guarantees for an algorithmic approach to the currently open problem of equilibrium selection in competitive multi-agent settings.

1 Introduction

Zero-sum games and variants thereof are arguably amongst the most well studied settings in game theory. Indeed much attention has focused on the class of strictly competitive games [5], i.e., two player games such that when both players change their mixed strategies, then either there is no change in the expected payoffs, or one of the two expected payoffs increases and the other decreases.[1] According to Aumann [5], *"Strictly competitive games constitute one of the few areas in game theory, and indeed in the social sciences, where a fairly sharp, unique prediction is made."* The unique prediction, of course, refers to the min-max solution and the resulting values guaranteed to both agents due to the classic work of von Neumann [52].

Unfortunately, when we move away from the safe haven of two-agent strictly competitive games, a lot of these regularities disappear. For example, in multi-agent variants of zero-sum and strictly competitive games, several critical aspects of the min-max theorem collapse [13]. Critically, Nash equilibrium payoffs need not be unique. In fact, there could be continua of equilibria with the payoff range of different agents corresponding to positive measure sets. Furthermore, Nash equilibrium strategies need not be exchangeable (i.e., mixing-matching strategies from different Nash profiles does not lead to a Nash) or max-min. Thus, network competition is not only significantly harder, but poses qualitatively different questions than two-agent competition.

Nevertheless, and in spite of the intense recent interest, inspired by ML applications such as Generative Adversarial Networks (GANs) and actor-critic systems, on understanding learning dynamics in zero-sum games and even network variants thereof [18, 20, 23, 1], so far there has been no systematic study

[1] In fact, as recent work has established these strictly competitive games are formally equivalent to weighted zero-sum games, i.e., affine transformations of zero-sum games [2].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

of how agents should deal with uncertainty of the resulting payoffs in such games. In these settings, the use of purely optimization-driven regret-minimizing algorithms, is no longer equally attractive as in the two-player case. The multiplicity of equilibria and the lack of unique value give rise to a non-trivial problem of *equilibrium selection* and learning agents face the fundamental dilemma between exploration and exploitation [14, 41, 12, 58]. These considerations drive our motivating question:

Are there exploration-exploitation dynamics that provably converge in networks of strictly competitive games? How do they behave in settings with multiple, payoff diverse Nash equilibria?

Model and Results. We study a well known smooth variant of Q-learning [53, 54], with softmax or Boltzmann exploration (one of the most fundamental models of exploration-exploitation in MAS), termed Boltzmann Q-learning or *smooth Q-learning* [50, 47]. Informally (see Section 3 for the rigorous definition), each agent k updates their choice distribution $x = (x_i)$ according to the rule $\dot{x}_i/x_i = (r_i - \bar{r}) - T_k(\ln x_i - \sum_j x_j \ln x_j)$, where $r_i, \bar{r}$ denote agent k's rewards from action i and average rewards, respectively, given all other agents' actions and T_k is agent k's exploration rate.

In our main result, we show convergence of Q-learning to Quantal Response Equilibria (QRE) (the prototypical extension of NE for games with bounded rationality [36]) in arbitrary networks of strictly competitive games [13]. As long as all exploration rates are positive, we prove via a global Lyapunov argument that the Q-learning dynamic converges pointwise to a unique QRE regardless of initial conditions and regardless of the number of the Nash equilibria of the original network game (Theorem 4.1). Related to the above, we demonstrate how exploration by all agents leads to equilibrium selection. This theoretically and empirically long-standing open problem ([32, 44, 42]) becomes tractable due to the theoretical guarantees of fast convergence to QRE that we provide in Theorem 4.1 for this class of (competitive) games. In fact, Theorem 4.1 is in some sense tight, as there exist network competitive settings whose dynamics lead to limit cycles as long as not all of the agents are performing exploration (see discussion and experiments in Section 5).

Other Related Works. The variant of Q-learning that we study has recently received a lot of attention due to its connection to evolutionary game theory [46, 31, 56]. It has also been also extensively studied in the economics and reinforcement learning literature under various names, see e.g., [3, 45] and [28, 38]. Recent works demonstrate that it is possible to show convergence of the Q-learning dynamics in multi-agent *cooperative* settings and to select highly desirable equilibria via bifurcations [34].

On the other hand, competitive multi-agent systems constitute one of the current frontiers in Artificial Intelligence and Machine Learning research. Many recent works investigate the complex behavior of competitive game theoretic settings [43, 10, 9, 8, 40]. focusing on carefully designed convergent algorithms, e.g., “optimism" [19, 21, 55], extra-gradient methods [39, 4, 27], regularization [42], momentum adapted dynamics [24], or symplectic integration schemes [7]. However, despite the theoretical progress and the impressive results in the empirical front [48, 51, 33], the literature on equilibrium selection seems to have received little attention so far. Prior works have focused on equilibrium selection in cooperative AI or have theoretically studied competition in 2-agent zero-sum games in which equilibrium selection is irrelevant [16, 17]. To our knowledge, we are the first paper to explicitly study both theoretically and experimentally multi-agent competitive settings without uniquely defined values.

2 Game-Theoretic Preliminaries

A *polymatrix or separable network game*, $\Gamma = \Big((V,E), (S_k, w_k)_{k\in V}, (\mathbf{A}_{kl})_{[k,l]\in E}\Big)$ consists of a graph (V,E), where $V = \{1,2,\dots,n\}$ is the set of players (or agents), and E is a set of pairs, $[k,l]$, of players $k \neq l \in V$. Each player, $k \in V$, has a finite set of actions (or strategies) S_k with generic elements $s_k \in S_k$ (depending on the context, sometimes we will also write i or $j \in S_k$). Players may also use mixed strategies (or choice distributions) $\mathbf{x}_k = (x_{ki})_{i\in S_k} \in \Delta_k$, where Δ_k is the simplex in $\mathbb{R}^{|S_k|}$, i.e., $\sum_{i\in S_k} x_{ki} = 1$, and $x_{ki} \geq 0$, for any $\mathbf{x}_k \in \Delta_k$. The *interior* of Δ_k is the set of all points $\mathbf{x}_k \in \Delta_k$ with $x_{ki} \in (0,1)$ for all $i \in S_k$. All points of Δ_k that are not in the interior, are called *boundary points*. We will write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ or $\mathbf{x} = (\mathbf{x}_k, \mathbf{x}_{-k})$ for a mixed strategy

profile $\mathbf{x} \in \Delta := \prod_{k \in V} \Delta_k$, where $\mathbf{x}_{-k} \in \Delta_{-k} = \prod_{l \neq k \in V} \Delta_l$ is the vector of mixed strategies of all players $l \in V$ other than k.

Each edge $[k, l] \in E$ defines a two player game with payoff matrices $\mathbf{A}_{kl} \in \mathbb{R}^{|S_k| \times |S_l|}$ and $\mathbf{A}_{lk} \in \mathbb{R}^{|S_l| \times |S_k|}$. The elements $a_{kl}(s_k, s_l)$ of a matrix $\mathbf{A}_{kl}$ denote the payoffs of player k when players k and l use pure actions $s_k \in S_k$ and $s_l \in S_l$, respectively. Each player $k \in V$ chooses a strategy (mixed or pure) and plays that strategy in *all* games $[k, l] \in E$. Thus, the payoff of player k at the pure strategy profile $\mathbf{s} = (s_1, \ldots, s_n) \in S := \prod_{k \in V} S_k$ is $u_k(\mathbf{s}) = \sum_{[k,l] \in E} a_{kl}(s_k, s_l)$. Similarly, the expected reward of player k in the mixed strategy profile $\mathbf{x} \in \Delta$ is

$$u_k(\mathbf{x}) := \sum\nolimits_{[k,l] \in E} \mathbf{x}_k^\top \mathbf{A}_{kl} \mathbf{x}_l = \mathbf{x}_k^\top \left(\sum\nolimits_{[k,l] \in E} \mathbf{A}_{kl} \mathbf{x}_l \right). \tag{1}$$

It will be convenient to write $r_{ki}(\mathbf{x}_{-k}) := u_k\left(i, \mathbf{x}_{-k}\right) = \sum_{[k,l] \in E} \{\mathbf{A}_{kl} \mathbf{x}_l\}_i$ (where $\{v\}_i$ to denotes the i-th element of a vector v), for the reward of pure action $i \in S_k$ of player k when all other players use the strategy profile $\mathbf{x}_{-k}$ and $r_k\left(\mathbf{x}_{-k}\right) := \left(r_{ki}(\mathbf{x}_{-k})\right)_{i \in S_k}$ for the resulting reward vector of all pure actions of agent k, respectively. Using this notation, the expected reward of player $k \in V$ at the mixed strategy profile $\mathbf{x} = (\mathbf{x}_k, \mathbf{x}_{-k})$ can be compactly expressed as $u_k(\mathbf{x}) = \mathbf{x}_k^\top r_k\left(\mathbf{x}_{-k}\right)$.

Weighted Zero-Sum Polymatrix Games. Γ is called a *weighted or rescaled* zero-sum polymatrix game [13], if there exist positive constants $w_1, w_2, \ldots, w_n > 0$, so that

$$\sum\nolimits_{k \in V} w_k u_k(\mathbf{x}) = 0, \quad \text{for all } \mathbf{x} \in \Delta. \tag{2}$$

By summing over the edges in E (instead of the players in V), we can equivalently express the weighted zero-sum property as

$$\sum\nolimits_{[k,l] \in E} \left[w_k \mathbf{x}_k^\top \mathbf{A}_{kl} \mathbf{x}_l + w_l \mathbf{x}_l^\top \mathbf{A}_{lk} \mathbf{x}_k \right] = 0, \quad \text{for all } \mathbf{x} \in \Delta. \tag{3}$$

Nash Equilibrium. A strategy profile (tuple of mixed strategies), $\mathbf{p} = (p_k)_{k \in V} \in \Delta$, with $\mathbf{p}_k = (p_{ki})_{i \in S_k} \in \Delta_k$ for each $k \in V$ is a *Nash equilibrium* of Γ if

$$u_k(\mathbf{p}) \geq u_k(x_k, \mathbf{p}_{-k}), \quad \text{for all } x_k \in \Delta_k, \text{ for all } k \in V, \tag{4}$$

i.e., if there exist no profitable unilateral deviations. By linearity, it suffices to verify the condition in equation (4) only for pure actions $s_k \in S_k$ instead of all $x_k \in \Delta_k$.

3 Joint-learning Model: Q-learning Dynamics

We next discuss how we can get to the Q-learning dynamics from Q-learning agents when there are multiple learners in the system. Our goal is to identify the dynamics in competitive systems in which multiple agents are playing a rescaled zero-sum polymatrix game repeatedly over time.

Q-learning. Q-learning [53, 54] is a value-iteration method for solving the optimal strategies in Markov Decision Processes (MDPs). It can be used as a model where users learn about their optimal strategy when facing uncertainties. In Q-learning, each agent $k \in \mathcal{N}$ keeps track of the past performance of their actions $i \in S_k$ via the Q-learning update rule

$$Q_{ki}(n+1) = Q_{ki}(n) + \alpha_k \left[r_{ki}(\mathbf{x}_{-k}, n) - Q_{ki}(n) \right], \; i \in S_k \tag{5}$$

where the additional argument, $n \geq 0$, denotes (discrete) time steps and $\alpha_k \in [0, 1]$ denotes the learning rate or memory decay of agent k, cf. [47, 31]. $Q_{ki}(t)$ is called the *memory* of agent k about the performance of action $i \in S_k$ up to time step $t \geq 0$. Agent $k \in V$ updates their actions (choice distributions) according to a Boltzmann type distribution, with

$$x_{ki}(n) = \frac{\exp\left(Q_{ki}(n)/T_k\right)}{\sum_{j \in S_k} \exp\left(Q_{kj}(n)/T_k\right)}, \quad \text{for each } i \in S_k, \tag{6}$$

where $T_k \in [0, +\infty)$ denotes agent k's learning sensitivity or adaptation, i.e., how much the choice distribution is affected by the past performance. We will refer to T_k as the *exploration rate* or

temperature of player k [34, 50] (see Remark 1 for a discussion). Combining equations (5) and (6), one obtains the recursive equation of player k's mixed strategy (or choice distribution)

$$x_{ki}(n+1) = \frac{x_{ki}(n)\exp\left((Q_{ki}(n+1) - Q_{ki}(n))/T_k\right)}{\sum_{j\in S_k} x_{kj}(n)\exp\left((Q_{kj}(n+1) - Q_{kj}(n))/T_k\right)},$$

for each $i \in S_k$. In practice, agents perform a large number of actions (updates of Q-values) for each choice-distribution (update of Boltzmann selection probabilities). This motivates to consider a continuous time version of the learning process for each agent $k \in V$ which results in the following update rules for both the memories Q_{ki} and the selection probabilities x_{ki} of each action $i \in S_k$

$$\dot{Q}_{ki} = \alpha_k \left[r_{ki}(\mathbf{x}_{-k}) - Q_{ki}\right], \text{ and } \dot{x}_{ki} = x_{ki}\left(\dot{Q}_{ki} - \sum\nolimits_{j\in S_k} \dot{Q}_{kj}\right)/T_k,$$

where we suppressed the dependence on (the continuous) time steps $t \geq 0$. Combining the last two equations under the assumptions that the relationship between pairs of actions is constant over time and that the choice distributions of the various agents are independently distributed, yields the *Q-learning dynamics*

$$\dot{x}_{ki} = x_{ki}\left[r_{ki}(\mathbf{x}_{-k}) - \mathbf{x}_k^\top r_k(\mathbf{x}_{-k}) - T_k\left(\ln(x_{ki}) - \mathbf{x}_k^\top \ln(\mathbf{x}_k)\right)\right], \tag{7}$$

for all $i \in S_k$ and all players $k \in V$. The next property provides intuition for the meaning of Q-values in a practical context (see also Remark 1 for a discussion). As is the case for all technical materials of this section, the proof of Proposition 3.1 is deferred to Appendix A.

Proposition 3.1 (Q-value Updates and Exponential Discounting)**.** *Consider the Q-learning updates*

$$Q_i(n+1) = Q_i(n) + \alpha[r_i(n) - Q_i(n)] \tag{8}$$

where $n \geq 0$ are discrete time steps and assume that $Q_i(0) = 0$ for all $i \in A$. Then, in continuous time, the updates are given by

$$Q_i(t) = \alpha\int_0^t e^{-\alpha s} r_i(t-s)ds = \alpha e^{-\alpha t}\int_0^t e^{\alpha s} r_i(s)\,ds, \quad \textit{for any } t > 0.$$

Remark 1. Equation (6) implies that higher values of T_k indicate a higher exploration rate, i.e., randomization among the agent's available choices in S_k, whereas values of T_k close to 0 indicate a higher exploitation rate, i.e., preference of the agent towards the best performing action. Specifically, for $T_k \to 0$, (6) corresponds to the best-response update rule, and agent k selects the action $i \in S_k$ with the highest Q-value with probability one. On the other hand, as $T_k \to \infty$, equation (6) corresponds to the case in which player k selects their actions uniformly at random.

From a behavioral perspective, this leads to an equivalent interpretation of the T_k's as the degrees of (bounded) rationality of the agents [25, 26, 22]. Similarly, Proposition 3.1 implies that the Q-learning updates can be intuitively interpreted as exponential discounting in the payoffs, with payoffs further back in the past being less important. In other words, the Q-value of an agent for an action i shows how beneficial the (discounted) strategy i is for that agent. On the other hand, from a purely algorithmic perspective, the T_k's and the corresponding entropy terms in the players' choice distributions updates can be viewed as regularization terms that prevent the learning dynamics from reaching the boundary or from getting trapped in local optima, see [11, 15, 37] and [29, 30].

Quantal Response Equilibria. Using the convention $x\ln x := 0$ whenever $x = 0$ (recall that $\lim_{x\to 0^+} x\ln x = 0$), it is immediate to infer from equation (7) that $x_{ki}(t) = x_{ki}(0)$ for all $t > 0$, for all $i \in S_k$ such that $x_{ki}(0) \in \{0, 1\}$. In other words, if the dynamics start on the boundary, then they will remain there, i.e., the boundary of Δ_k is invariant for the Q-learning dynamics. This implies, in particular, that all corner points of the simplex Δ_k, i.e, all points $x_k = (x_{ki})_{i\in S_k}$ with one coordinate equal to 1 (and consequently, all other coordinates equal to 0) are (trivially) fixed points of the Q-learning dynamics. Thus, the interesting part concerns the fixed-point analysis of the dynamics for *interior* starting points.

The *interior* fixed points of the Q-learning dynamics are called *Quantal Response Equilibria*. From equation (7), it is apparent that whenever $T_k > 0$ for all $k \in V$, then such points can only lie in the interior of Δ_k due to the entropy term ($\ln$ of the choice probabilities). Thus, at such fixed points the rate of change of x_{ki} must be equal to zero. Specifically, a strategy profile $\mathbf{p} = (p_k)_{k\in V}$ with $\mathbf{p}_k = (p_{ki})_{i\in S_k}$ for each $k \in V$ is a *Quantal Response Equilibrium (QRE)* of Γ if

$$r_{ki}(\mathbf{p}_{-k}) = \mathbf{p}_k^\top r_k(\mathbf{p}_{-k}) + T_k\left(\ln(p_{ki}) - \mathbf{p}_k^\top \ln(\mathbf{p}_k)\right), \quad \text{for all } k \in V. \tag{9}$$

The following properties are immediate from the characterization of the QRE in equation (9).

Theorem 3.2 (Interior Fixed Points of the Q-learning Dynamics)**.** *Let Γ be an arbitrary game, with positive exploration rates T_k and consider the associated Q-learning dynamics*

$$\dot{x_{ki}} = x_{ki}\left[r_{ki}(\mathbf{x}_{-k}) - \mathbf{x}_k^\top r_k\left(\mathbf{x}_{-k}\right) - T_k\left(\ln\left(x_{ki}\right) - \mathbf{x}_k^\top \ln\left(\mathbf{x}_k\right)\right)\right], \quad i \in S_k, k \in V.$$

The interior fixed points, $\mathbf{p} = (p_k)_{k\in V}$ of the Q-learning dynamics are the solutions of the system

$$p_{ki} = \frac{\exp\left(r_{ki}(\mathbf{p}_{-k})/T_k\right)}{\sum_{j\in S_k}\exp\left(r_{kj}(\mathbf{p}_{-k})/T_k\right)}, \quad \textit{for all } i \in S_k. \tag{10}$$

Such fixed points always exists and coincide with the Quantal Response Equilibria (QRE) of Γ. Given any such fixed point $\mathbf{p}$, we have, for all $\mathbf{x}_k \in \Delta_k$ and for all $k \in V$, that

$$\left(\mathbf{x}_k - \mathbf{p}_k\right)^\top\left[r_k(\mathbf{p}_{-k}) - T_k \ln\left(\mathbf{p}_k\right)\right] = 0. \tag{11}$$

The proof of Theorem 3.2 does not make use of the weighted zero-sum property, cf. equation (2), and thus, it holds for arbitrary network games (not necessarily weighted network zero-sum).

Exploration Rates. Equation (11) in Theorem 3.2 is critical for the main technical step in our proof of convergence of the Q-learning dynamics in weighted zero-sum polymatrix games. Thus, it is important to note that this equation only holds when *all* exploration rates, $T_k, k \in V$ are positive (and not merely non-negative). The intuition is that when all $T_k > 0$, every QRE becomes interior, i.e., has full support. This implies that all deviations (to pure or mixed strategies) give the *same* expected reward to the deviating agent. If some T_k are equal to 0, then (11) holds with (non-strict) inequality which is not generally enough for the proof of Theorem 4.1.

When $T_k = 0$ for all $k \in V$, the Q-learning dynamics reduce to the standard replicator dynamics and the QRE coincide with the Nash equilibria of Γ, cf. equation (4). The case in which only some T_k are equal to 0 exhibits more interesting behavior and will be discussed in detail later. In brief, if this is the case, then the dynamics may also converge to a boundary point (for the non-exploring players, i.e., for the players $k \in V$ with $T_k = 0$), even if the dynamics start in the interior.

4 Convergence of Q-learning in Weighted Zero-sum Polymatrix Games

Next, we consider the joint learning model (Q-learning) of the previous section in rescaled zero-sum polymatrix games, Γ, as defined above. Our main result in this section is that Q-learning converges to the QRE of Γ. The key step of the proof is to show that the *distance* between an interior QRE, $\mathbf{p} \in \Delta$ and the sequence of play, $\mathbf{x}(t), t \geq 0 \in \Delta$, that is generated by the Q-learning model is monotonically decreasing. To measure this distance in a meaningful way, we will use the notion of KL-divergence which is formally defined next.

Definition 1 (Kullback-Leibler (KL) Divergence)**.** The Kullback-Leibler or simply *KL-Divergence* (also called *relative entropy*), D_{KL}, between two strategy profiles $\mathbf{p} = (\mathbf{p}_k)_{k\in V}, \mathbf{x}(t) = (\mathbf{x}_k(t))_{k\in V} \in \Delta$ with $\mathbf{p}_k = (p_{ki})_{i\in S_k}$ and $\mathbf{x}_k(t) = (x_{ki}(t))_{i\in S_k} \in \Delta_k$ for all $k \in V$, is defined as

$$D_{\text{KL}}\left(\mathbf{p} \parallel \mathbf{x}(t)\right) := \sum\nolimits_{k\in V} D_{\text{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k(t)) = \sum\nolimits_{k\in V} \mathbf{p}_k^\top \ln\left(\frac{\mathbf{p}_k}{\mathbf{x}_k(t)}\right), \tag{12}$$

where $\mathbf{p}_k^\top \ln\left(\frac{\mathbf{p}_k}{\mathbf{x}_k(t)}\right) = \sum_{i\in S_k} p_{ki}\ln\left(p_{ki}/x_{ki}(t)\right)$. If $\mathbf{w} = (w_k)_{k\in V}$ is a k-dimensional vector of positive scalars, then the *weighted or rescaled KL-divergence*, $D_{\text{KL}(w)}$, is defined as

$$D_{\text{KL}(w)}\left(\mathbf{p} \parallel \mathbf{x}(t)\right) := \sum\nolimits_{k\in V} w_k D_{\text{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k(t)). \tag{13}$$

The KL-divergence between $\mathbf{p}$ and $\mathbf{x}(t)$ can be thought of as a measurement of how far the distribution $\mathbf{p}$ is from the distribution $\mathbf{x}(t)$. However, the KL-divergence is not symmetric, i.e., in general it holds that $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{x}(t)) \neq D_{\text{KL}}(\mathbf{x}(t) \parallel \mathbf{p})$.

Main Result. Using the notion of KL-divergence we can now formulate our main result.

Theorem 4.1 (Convergence of Q-learning dynamics)**.** *Let Γ be a rescaled zero-sum polymatrix game, with positive exploration rates T_k. There exists a unique QRE $\mathbf{p}$ such that if $\mathbf{x}(t)$ is any trajectory of the associated Q-learning dynamics $\dot{\mathbf{x}} = f(\mathbf{x})$, with f_i is given via* (7)*, where $\mathbf{x}(0)$ is an interior point, then $\mathbf{x}(t)$ converges to $\mathbf{p}$ exponentially fast. In particular, we have that*

$$\frac{d}{dt} D_{KL(w)}(\mathbf{p} \parallel \mathbf{x}(t)) = -\sum\nolimits_{k \in V} w_k T_k \left[D_{KL}(\mathbf{p}_k \parallel \mathbf{x}_k) + D_{KL}(\mathbf{x}_k \parallel \mathbf{p}_k) \right]. \tag{14}$$

Sketch of Proof. Let $\mathbf{p} = (\mathbf{p}_k)$ be a QRE equilibrium for the Q-learning dynamics (existence of $\mathbf{p}$ is guaranteed via Theorem 3.2), and let $\mathbf{x}(t)$ be a trajectory with $\mathbf{x}(0)$ an interior point. We will first establish (14), from which the other statements follow. The first step is to derive an explicit formula for the time derivative of the KL divergence between $\mathbf{p}_k$ and $\mathbf{x}_k(t)$ for a given player k.

Lemma 4.2. *For any QRE equilibrium $\mathbf{p}$ and any player $k \in V$ we have that*

$$\frac{d}{dt} D_{KL}(\mathbf{p}_k \parallel \mathbf{x}_k(t)) = (\mathbf{x}_k - \mathbf{p}_k)^\top \left[r_k(\mathbf{x}_{-k}) - r_k(\mathbf{p}_{-k}) \right] - T_k \left[D_{KL}(\mathbf{p}_k \parallel \mathbf{x}_k) + D_{KL}(\mathbf{x}_k \parallel \mathbf{p}_k) \right].$$

It follows from Lemma 4.2 that

$$\begin{aligned}
\frac{d}{dt} D_{\mathrm{KL}}(\mathbf{p} \parallel \mathbf{x}(t)) = & \sum_{k \in V} \frac{d}{dt} w_k D_{\mathrm{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k) \\
= & \sum_{k \in V} w_k \left[(\mathbf{x}_k - \mathbf{p}_k)^\top \left[r_k(\mathbf{x}_{-k}) - r_k(\mathbf{p}_{-k}) \right] - T_k \left(D_{\mathrm{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k) + D_{\mathrm{KL}}(\mathbf{x}_k \parallel \mathbf{p}_k) \right) \right] \\
= & \sum_{k \in V} w_k (\mathbf{x}_k - \mathbf{p}_k)^\top \left[r_k(\mathbf{x}_{-k}) - r_k(\mathbf{p}_{-k}) \right] - \sum_{k \in V} w_k T_k \left[D_{\mathrm{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k) + D_{\mathrm{KL}}(\mathbf{x}_k \parallel \mathbf{p}_k) \right].
\end{aligned} \tag{15}$$

Notice that (14) follows from (15) if the first term vanishes. This can be shown by utilizing the following result regarding rescaled zero-sum polymatrix games.

Lemma 4.3. *Given any QRE equilibrium $\mathbf{p} = (\mathbf{p}_k)$ and $\mathbf{x} = (\mathbf{x}_k)$ we have that*

$$\sum\nolimits_{k \in V} w_k \left[\mathbf{x}_k^\top r_k(\mathbf{p}_{-k}) + \mathbf{p}_k^\top r_k(\mathbf{x}_{-k}) \right] = 0. \tag{16}$$

Returning to (15), we have that the first term in the right hand side of the last equation vanishes since

$$\begin{aligned}
\sum_{k \in V} w_k (\mathbf{x}_k - \mathbf{p}_k)^\top \left[r_k(\mathbf{x}_{-k}) - r_k(\mathbf{p}_{-k}) \right] = & \sum_{k \in V} w_k \mathbf{x}_k^\top r_k(\mathbf{x}_{-k}) + \sum_{k \in V} w_k \mathbf{p}_k^\top r_k(\mathbf{p}_{-k}) \\
& + \sum_{k \in V} w_k \left[\mathbf{x}_k^\top r_k(\mathbf{p}_{-k}) + \mathbf{p}_k^\top r_k(\mathbf{x}_{-k}) \right] = 0,
\end{aligned}$$

where the last equality follows from both the zero-sum property (2) and Lemma 4.3. Thus,

$$\frac{d}{dt} D_{\mathrm{KL}}(\mathbf{p} \parallel \mathbf{x}(t)) = -\sum\nolimits_{k \in V} w_k T_k \left[D_{\mathrm{KL}}(\mathbf{p}_k \parallel \mathbf{x}_k) + D_{\mathrm{KL}}(\mathbf{x}_k \parallel \mathbf{p}_k) \right].$$

This equality implies that $D_{\mathrm{KL}(w)}$ is a Lyapunov function for the Q-learning dynamics. Moreover, it implies that $\mathbf{p}$ is unique, and that $\mathbf{x}(t)$ is converging to $\mathbf{p}$ at an exponential rate. □

Equilibrium Selection. Theorem 4.1 suggests a tractable, algorithmic approach to the problem of equilibrium selection in weighted zero-sum polymatrix games. In particular, given an interior initial condition and positive exploration rates T_k, Theorem 4.1 guarantees the system will converge to a unique QRE. Moreover, it will do so exponentially fast. Thus, one can use the following exploration policy set up a fast-slow system for equilibrium selection: for each player, the exploration rates T_k are brought to zero slowly (the slow system), while the Q-learning dynamics (the fast system) converge at an exponentially rate to the unique QRE; this process will approximate a Nash equilibrium of Γ.

5 Experiments: Equilibrium Selection in Competitive Games

We study the performance of Q-learning in (weighted) network zero-sum and (as a subcase thereof) in 2-agent zero-sum games. We also investigate numerically the case in which only *some* of the agents have positive exploration rates.

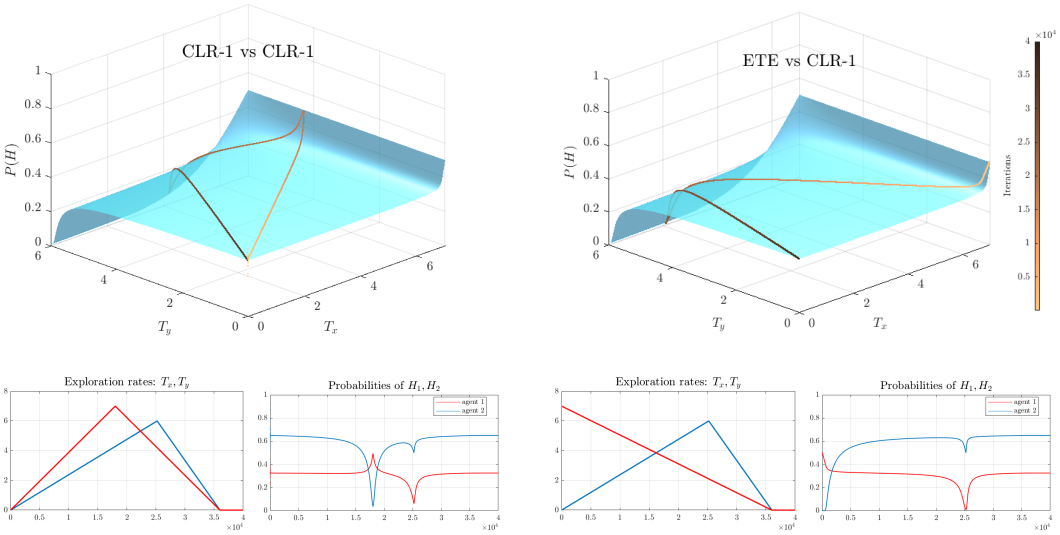

Figure 1: QRE surface and exploration paths (upper panels) for two different exploration policy profiles (lower panels) in the Asymmetric MPs game. For any combination of exploration policies (CLR-1 and ETE), the sequence of play converges to the unique QRE and as the exploration rates decrease zero, the sequence of play converges to the unique Nash equilibrium of the game.

Two-agent Weighted Zero-Sum Games. The visualizations that we obtain in this low-dimensional case allow us to build intuition that carries over to the higher dimensional cases (with more agents) that we treat later.

Experimental setup: We consider the Asymmetric Matching Pennies (AMPs) (a variation of the well-known Matching Pennies game [57]) which is a 2-agent, weighted zero-sum game. Each agent has two actions $\{H, T\}$ and the payoff matrices are $\mathbf{A} = \begin{pmatrix} 2 & -2 \\ 0 & 2 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 4 & 0 \\ -4 & -4 \end{pmatrix}$. The AMPs game is a weighted zero-sum game since $\mathbf{A} + 0.5 \cdot \mathbf{B}^\top = 0$ (each agent is the row agent in their matrix), with a unique interior Nash equilibrium, $(\mathbf{p}, \mathbf{q}) = ((1/3, 2/3), (2/3, 1/3))$.

Results: In Figure 1, we visualize the QRE surfaces (light blue manifolds) for different exploration rates T_x, T_y of the two agents ($x - y$ plane). The vertical axis shows the probability with which agent 1 chooses H at the unique QRE of the game. We plot the exploration path along two representative exploration-exploitation policies: *Explore-Then-Exploit* (ETE) [6], which starts with (relatively) high exploration that gradually reduces to zero and *Cyclical Learning Rate with one cycle* (CLR-1) [49], which starts with low exploration, increases to high exploration around the half-life of the cycle and then decays to zero. For each pair of exploration rates, the learning dynamics converge to the unique QRE that corresponds to these exploration rates. As the exploration rates decay to zero, the dynamics converge to (i.e., select) the unique QRE (i.e, the Nash equilibrium in this case) of the original game.

By contrast, the first panel of Figure 2 shows that the dynamics cycle around the unique Nash equilibrium when both agents do not use exploration. The rest of the panels of Figure 2 show that in

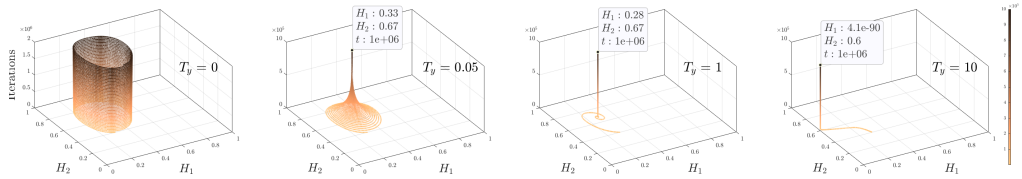

Figure 2: Q-learning dynamics in the AMPs game for $T_x = 0$ (no exploration by x-agent) and four different exploration rates, $T_y \geq 0$ by y-agent.

this case, exploration by only one agent suffices to lead the joint-learning dynamics to convergence. In this case, for higher values of exploration by the exploring agent, the QRE component of the non-exploring agent may lie at the boundary (see panel 4 in Figure 2 and Appendix C).

Network Zero-Sum Games. As the next experiment shows, equilibrium selections works also in larger networks, provided that all agents maintain a positive exploration rate. In contrast to the previous example, in this case, exploration by only some agents is not sufficient for convergence.

Experimental setup: We consider a zero-sum polymatrix game (ZSPG) with $n+2$ agents ($n \in \mathbb{N}$ is arbitrary) which is depicted in Figure 3. Each agent $p_1, p_2, \ldots, p_n$ has two actions $\{H, T\}$ and

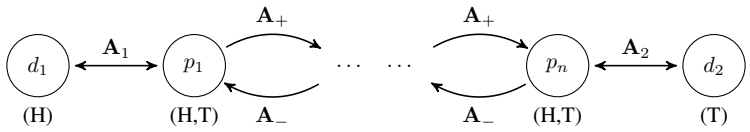

Figure 3: The match-mismatch network zero-sum game.

receives $+1$ if they match the action of the next agent (otherwise they receive -1) and $+1$ if they mismatch the strategy of the previous agent (otherwise they receive -1). There are two dummy agents, d_1, d_2, who have the same payoffs but only one action: H for d_1 and T for d_2. In particular, the payoff matrices for the games between non-dummy agents are given by

$$\mathbf{A}_+ = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{A}_- = -\mathbf{A}_+ \tag{ZSPG}$$

whereas the payoff matrices of dummy agents d_1 and d_2 are given by $\mathbf{A}_1 = \mathbf{A}_2 = (1, -1)$. The column agents in the payoff matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are p_1 and p_n, and their payoffs in the encounters against d_1 and d_2 are given by $-\mathbf{A}_1^\top$ and $-\mathbf{A}_2^\top$, respectively. This game has multiple (infinite many) Nash equilibria of the following form: the odd-numbered agents play pure strategy T (which, for agent 1 ensures $+1$ against d_1) whereas the even agents are indifferent between H, T (since they are certain to have the opposing result in the two games against the previous and the next agent, both being odd and playing T).

Results: We provide visualizations from two instances of the (ZSPG) game. In the first, with $n = 3$ non-dummy agents (cf. panels 1-3 in Figure 4), we visualize the projections of the QRE surfaces on the H coordinate at QRE for agents 1 to 3 for fixed $T_3 = 3$ and all possible combinations of T_1, T_2 (x-y planes). The QRE manifolds approach continuously the boundary (case with $T_i = 0$ for $i = 1, 2$) which leads to a unique equilibrium selection (also when $T_1 = T_2 = 0$). The fourth panel shows a snapshot of the Lyapunov function (KL-divergence between a choice distribution $\mathbf{x}$ and the unique QRE $\mathbf{q}$) for a fixed exploration profile $T_k > 0$ for all $k = 1, \ldots, n$ in an instance with $n = 7$ non-dummy agents (the plot is similarly for all n). For this plot, we use the dimension reduction technique (to visualize high-dimensional surfaces along two randomly chosen directions) of [35] (see also Appendix C). As expected, the KL-divergence is convex and decreasing for all $\mathbf{x}$ with a unique minimizer.

Figure 5 shows summary statistics for the $n = 7$ instance of the (ZSPG) game for three different exploration profiles. The panels in the leftmost column show the equilibria (top) and utilities (bottom)

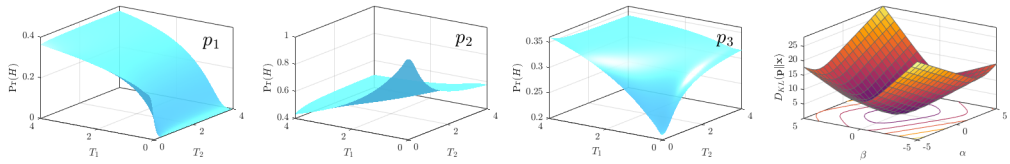

Figure 4: Projections of the QRE surfaces on the H coordinate of players p_1, p_2, p_3 (panels 1-3) and a snapshot of the Lyapunov function, KL-divergence, (panel 4) in two instances of the (ZSPG) game with 3 and 7 non-dummy agents, respectively, as discussed in Section 5.

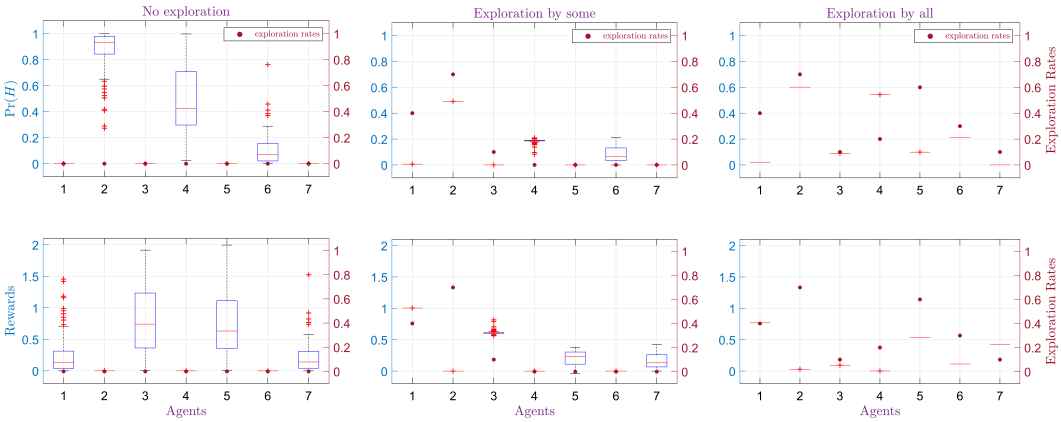

Figure 5: Different rates of exploration in an instance of the match-mismatch ZSPG with 7 agents. The boxplots show summery statistics from 100 runs. No exploration leads to multiple equilibria (first column), exploration by all agents leads to a unique QRE (last column), whereas equilibrium by some agents is not enough to ensure convergence to a unique outcome which shows the tightness of Theorem 4.1 (middle column).

of the 7 agents in 100 runs when $T_k = 0$ for all $k = 1, 2, \ldots, n$ (no exploration). In this case, the dynamics converge in all runs to the pure action for the odd agents and to some arbitrary (and different every time) mixed action for the even agents. This behavior of the learning dynamics is in line with previous results in adversarial learning when equilibria lie on one face (relative boundary) of the high-dimensional simplex (cf. [37]).

The panels in the middle column correspond to a case with exploration by agents $1, 2, 3$ and no exploration by agents $4, \ldots, 7$. Exploration by the first group of agents ensures convergence of their individual dynamics to the corresponding component of the QRE (in line with the theoretical prediction of Theorem 4.1) but even-numbered agents that lie further away from that group fail to converge to a unique outcome (boxplot for agents 4 and 6). The outcome in this case, shows that the statement of Theorem 4.1 is tight, and exploration by only a subset of agents may not be enough to ensure a unique outcome in general settings. Finally, the panels in the rightmost column show convergence to a unique QRE when all agents have positive exploration rates.

6 Conclusions

In this paper, we studied a commonly used smooth variant of the Q-learning dynamics in multi-agent competitive systems. Given that each agent's strategy includes at least some amount of exploration, we showed that the dynamics always converge to a unique Quantal Response Equilibrium (QRE). In particular, the convergence of Q-learning in this (competitive) setting is a remarkably robust phenomenon, occurring regardless of the number of agents or degree of competition, and without any need for parameter fine-tuning.

Our theoretical results have important implications from an algorithmic perspective. In contrast to two-agent zero-sum games in which agents have the same reward in all equilibria, multi-agent systems may exhibit a multiplicity of equilibria in which each agent may have different rewards. This inherent complexity within multi-agent systems leads to the non-trivial, and currently open problem, of equilibrium selection, which has previously been studied only within *cooperative* settings. The bounds we prove on the speed of convergence of the Q-learning dynamics to the QRE provide the necessary theoretical guarantees for a simple algorithm to select (approximate) Nash equilibria, in turn implying that the problem of equilibrium selection is tractable within competitive multi-agent settings as well. Our results complement previous advances on the convergence of Q-learning and equilibrium selection in cooperative settings and provide a promising starting point for exploring general multi-agent settings which combine elements from both cooperative and competitive systems.

References

[1] Jacob Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 3–47. PMLR, 16–19 Mar 2021.

[2] Ilan Adler, Constantinos Daskalakis, and Christos H Papadimitriou. A note on strictly competitive games. In *WINE*, pages 471–474. Springer, 2009.

[3] C. Alós-Ferrer and N. Netzer. The logit-response dynamics. *Games and Economic Behavior*, 68(2):413–427, 2010.

[4] Kimon Antonakopoulos, E Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. *arXiv preprint arXiv:2010.12100*, 2020.

[5] Robert John Aumann. Game theory. London (UK) Macmillan, 1987.

[6] Yu Bai and Chi Jin. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, Madison, WI, USA, 2020. Omnipress.

[7] James P. Bailey, Gauthier Gidel, and Georgios Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. *ArXiv e-prints*, 2019.

[8] James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *ACM Conference on Economics and Computation*, 2018.

[9] James P. Bailey and Georgios Piliouras. Multi-agent learning in network zero-sum games is a hamiltonian system. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

[10] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *ICML*, 2018.

[11] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[12] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[13] Y. Cai, O. Candogan, C. Daskalakis, and C. Papadimitriou. Zero-Sum Polymatrix Games: A Generalization of Minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

[14] Caroline Claus and Craig Boutilier. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. AAAI '98/IAAI '98, page 746–752, 1998.

[15] P. Coucheney, B. Gaujal, and P. Mertikopoulos. Penalty-Regulated Dynamics and Robust Learning Procedures in Games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

[16] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI. *arXiv e-prints*, December 2020.

[17] C. Daskalakis, D.J. Foster, and N. Golowich. Independent Policy Gradient Methods for Competitive Reinforcement Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540. Curran Associates, Inc., 2020.

[18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

[19] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *ICLR*, 2018.

[20] Constantinos Daskalakis and Ioannis Panageas. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. In Avrim Blum, editor, *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 27:1–27:18, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[21] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

[22] T. Galla and J. D. Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

[23] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

[24] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Rémi Lepriol, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS*, 2019.

[25] T.-H. Ho and C. F. Camerer. Experience-weighted attraction learning in normal form games. *Econometrica*, 67:827–874, 1999.

[26] Teck-Hua Ho and Colin Camerer. Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *Journal of mathematical psychology*, 42:305–326, 1998.

[27] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*, 2019.

[28] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4(1):237–285, 1996.

[29] Michael Kaisers and Karl Tuyls. Frequency adjusted multi-agent q-learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, page 309–316, 2010.

[30] Michael Kaisers and Karl Tuyls. FAQ-Learning in Matrix Games: Demonstrating Convergence near Nash Equilibria, and Bifurcation of Attractors in the Battle of Sexes. In *Proceedings of the 13th AAAI Conference on Interactive Decision Theory and Game Theory*, AAAIWS'11-13, page 36–42. AAAI Press, 2011.

[31] A. Kianercy and A. Galstyan. Dynamics of Boltzmann Q learning in two-player two-action games. *Phys. Rev. E*, 85:041145, Apr 2012.

[32] Y. Kim. Equilibrium Selection in n-Person Coordination Games. *Games and Economic Behavior*, 15(2):203–227, 1996.

[33] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games, 2020.

[34] S. Leonardos and G. Piliouras. Exploration-Exploitation in Multi-Agent Learning: Catastrophe Theory Meets Game Theory. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, AAAI-21, 2021.

[35] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6389–6399. Curran Associates, Inc., 2018.

[36] R. D. McKelvey and T. R. Palfrey. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38, 1995.

[37] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in Adversarial Regularized Learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 2703–2717, USA, 2018. Society for Industrial and Applied Mathematics.

[38] P. Mertikopoulos and W. H. Sandholm. Learning in Games via Reinforcement and Regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

[39] P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *ArXiv e-prints*, July 2018.

[40] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019.

[41] Liviu Panait and Sean Luke. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, Nov 2005.

[42] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization, 2020.

[43] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[44] J. Romero. The effect of hysteresis on equilibrium selection in coordination games. *Journal of Economic Behavior & Organization*, 111:88–105, 2015.

[45] J. B.T. Sanders, J. D. Farmer, and T. Galla. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 8(1):4902, 2018.

[46] Y. Sato, E. Akiyama, and J. P. Crutchfield. Stability and diversity in collective adaptation. *Physica D:Nonlinear Phenomena*, 210(1):21–57, 2005.

[47] Y. Sato and J. P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E*, 67:015206, Jan 2003.

[48] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016.

[49] Leslie N. Smith and Nicholay Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv e-prints*, page arXiv:1708.07120, 2017.

[50] K. Tuyls, K. Verbeeck, and T. Lenaerts. A Selection-mutation Model for Q-learning in Multi-agent Systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, pages 693–700, 2003.

[51] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu

Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Nov 2019.

[52] John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–300, 1928.

[53] C. J.C.H. Watkins and P. Dayan. Technical Note: Q-Learning. *Machine Learning*, 8(3):279–292, May 1992.

[54] C.J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.

[55] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021.

[56] D. H. Wolpert, M. Harré, E. Olbrich, N. Bertschinger, and J. Jost. Hysteresis effects of changing the parameters of noncooperative games. *Phys. Rev. E*, 85:036102, Mar 2012.

[57] M. Wunder, M. Littman, and M. Babes. Classes of Multiagent Q-Learning Dynamics with ϵ-Greedy Exploration. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1167–1174, Madison, WI, USA, 2010. Omnipress.

[58] K. Zhang, Z. Yang, and T. Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv e-prints*, page arXiv:1911.10635, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   - This is a theoretical work which does not present any foreseeable societal consequence.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
   - Yes, this work includes theoretical results. In the main part, we provide proof sketches of all our results. In the supplementary material, we also provide the complete technical details.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
   - Our simulations concern theoretical abstractions of network games. The total amount of computation is not a concern and all our experiments can be reproduced in any conventional machine.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]