On the Stability of Learning in Network Games with Many Players

Aamal Hussain
Imperial College London
aamal.hussain15@imperial.ac.uk

Dan Leonte
Imperial College London
dan.leonte16@imperial.ac.uk

Francesco Belardinelli
Imperial College London
francesco.belardinelli@imperial.ac.uk

Georgios Piliouras
Singapore University of Technology and Design
georgios@sutd.edu.sg

ABSTRACT

Multi-agent learning algorithms have been shown to display complex, unstable behaviours in a wide array of games. In fact, previous works indicate that convergent behaviours are less likely to occur as the total number of agents increases. This seemingly prohibits convergence to stable strategies, such as Nash Equilibria, in games with many players.

To make progress towards addressing this challenge we study the Q-Learning Dynamics, a classical model for exploration and exploitation in multi-agent learning. In particular, we study the behaviour of Q-Learning on games where interactions between agents are constrained by a network. We determine a number of sufficient conditions, depending on the game and network structure, which guarantee that agent strategies converge to a unique stable strategy, called the Quantal Response Equilibrium (QRE). Crucially, these sufficient conditions are independent of the total number of agents, allowing for provable convergence in arbitrarily large games.

Next, we compare the learned QRE to the underlying NE of the game, by showing that any QRE is an ϵ-approximate Nash Equilibrium. We first provide tight bounds on ϵ and show how these bounds lead naturally to a centralised scheme for choosing exploration rates, which enables independent learners to learn stable approximate Nash Equilibrium strategies. We validate the method through experiments and demonstrate its effectiveness even in the presence of numerous agents and actions. Through these results, we show that independent learning dynamics may converge to approximate Nash Equilibria, even in the presence of many agents.

KEYWORDS

Multi-Agent Learning, Quantal Response Equilibrium, Online Learning in Games

ACM Reference Format:
Aamal Hussain, Dan Leonte, Francesco Belardinelli, and Georgios Piliouras. 2024. On the Stability of Learning in Network Games with Many Players. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 18 pages.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Game Theory (EGT) has emerged as a powerful formalism for studying learning in multi-agent settings [55, 63]. Here, agents are required to explore their state space to determine optimal actions, whilst simultaneously maximising their expected reward in the face of the changing behaviour of their opponents. By modelling these situations as idealised games it is possible to study the effect of various factors, such as payoffs and number of agents, on the dynamics of learning. An important question which is often studied from this lens is whether popular multi-agent learning algorithms converge to an equilibrium [20, 31, 41] (most often the Nash Equilibrium).

Unfortunately, it seems that the general answer to this question is *no*. Recent work has shown that, even in zero-sum games, the dynamics of no-regret learning algorithms can be cyclic [39] or chaotic [6]. In addition, even small deviations from the zero-sum setting can result in robustly non-convergent dynamics [7, 25] so that in general-sum games, non-convergent behaviour appears to be the norm [13, 14, 17, 27, 30, 44, 45, 52, 65, 66]. To make matters worse, recent findings in [51] suggest that, as the number of agents in the game increases, the likelihood for chaotic dynamics also increases when agents have low exploration rates. Similarly, the results of [26] imply that incredibly large exploration rates may be required in games with many agents in order to ensure convergence. This seemingly presents a bottleneck for strong convergence guarantees in multi-agent settings with many agents.

Despite this, many real world problems such as resource allocation [1, 46], routing [3, 8, 9] and robotics [19, 57] consider a large number of agents who continuously adapt to one another. These practical applications in conjunction with the negative results in the face of many players immediately yield the following question:

Is there any hope for independent learning agents to converge to an equilibrium in games with many players?

To make progress in answering this question, this work examines multi-agent learning in *network* games. Here, it is assumed that agents can only interact with their neighbours within an underlying communication network. Such systems are ubiquitous: machine learning architectures often impose structure between models [22, 33]; in robotic systems, agents interact through communication networks [16, 57]; in both economics and biology, agent interactions are constrained through social networks. Network games refine the setting of [26, 51], in which it was assumed that each agent is directly influenced by every other agent in the environment. This work provides strong evidence that the network structure matters, in some cases even more so than the total number of agents.

Model and Contribution. We consider agents who update via the *Q-Learning* dynamic, [54, 64], a foundational model from game theory which describes the behaviour of agents who balance exploration and exploitation. Similar to [26] we determine a number of sufficient conditions on exploration rates such that Q-Learning is guaranteed to converge to a unique equilibrium. In this work, however, we find that these conditions depend on graph theoretic properties of the interaction network. In our experiments, we examine how these conditions depend on the total number of agents and find network structures for which there is no explicit dependence. These implications are visualised on a number of representative network games and it is shown that large numbers of agents may converge to an equilibrium, so long as weakly connected network structures are used. By contrast, if the network is strongly connected, we recover the results of [26, 51] and show that stability depends on the total number of agents.

The equilibrium solution to which Q-Learning converges is the *Quantal Response Equilibrium* (QRE) [32, 35], a widely studied extension of the Nash Equilibrium for agents who explore their state space [15, 29, 31]. In this work, we quantify the 'distance' between a QRE and NE by showing that any QRE is an approximate Nash Equilibrium and providing tight bounds on this approximation. Using this, we present a procedure for choosing exploration rates so that Q-Learning agents may converge 'closer' to the Nash Equilibrium, whilst maintaining the stability of the dynamic. We validate this procedure in a number of large scale network games and show that it leads to improvements in the convergence of Q-Learning dynamics towards approximate Nash Equilibria.

Related Work. In [14] the authors showed that the Experience Weighted Attraction (EWA) dynamic, which is closely related to Q-Learning [32], achieves chaos in classes of two-player games. Advancing this result, [51] showed that chaotic dynamics become more prevalent as the number of agents increase. Similar to this work, [26] apply the framework of *monotone game* [12, 47, 61] to show that Q-Learning Dynamics converge to a unique equilibrium in any game, given sufficient exploration. However, they also find that this condition increases with the number of agents.

Besides online learning, other approaches have been developed to try to *compute* Nash Equilibria in games. For our purposes, the most relevant of these are homotopy-like methods [21, 62]. The principle of these methods is to perturb the payoff functions so that the resulting perturbed game is 'easier' to solve. Then, by iteratively annealing this perturbation, one can approximate the underlying NE. Recently [15] applies an entropy perturbation of payoffs and use gradient-descent based approach to solve for a continuum of *Quantal Response Equilibria* (QRE), which eventually leads to a NE [35]. Whilst homotopy methods present a powerful tool for computing approximate equilibria, they often lack the advantages of decentralisation provided by online learning and may not come with strong guarantees. [48] combines the entropy perturbation approach with online learning and show that, in two-player zero-sum games, this method allows independent learners to converge asymptotically to an NE. However, as with most learning strategies, its behaviour in many player, general sum games is unknown.

We address the problem of learning in many player games by examining the role of an underlying communication network. A number of works in game theory have shown that network structure affects the uniqueness and stability of NE [2, 4, 11, 37, 47]. Our main result refines that of [26] to include the network and show that Q-Learning dynamics can reach a QRE in any network game, given sufficiently high exploration rates. Crucially, these conditions are explicitly independent of the total number of agents. We also show that the QRE achieved by Q-Learning is an approximate Nash Equilibrium, and design a centralised scheme for updating exploration rates so that Q-Learning dynamics converge along the continuum of stable QRE to an approximate Nash Equilibrium.

2 PRELIMINARIES

We begin in Section 2.1 by defining the network game model, which is the setting on which we study the Q-Learning dynamics, which we describe in Section 2.2.

2.1 Game Model

In this work, we consider *network polymatrix games* [32]. A Network Game is described by the tuple $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k \in \mathcal{N}})$, where $\mathcal{N}$ denotes a finite set $\mathcal{N}$ of players, indexed by $k = 1, \ldots, N$. Each agent can choose from a finite set $\mathcal{S}_k$ of actions, indexed by $i = 1, \ldots, n$. We denote the *strategy* $\mathbf{x}_k$ of an agent k as the probabilities with which they play their actions. Then, the set of all strategies of agent k is $\Delta(\mathcal{S}_k) := \{\mathbf{x}_k \in \mathbb{R}^n : \sum_i x_{ki} = 1, x_{ki} \geq 0\}$. Each agent is also given a payoff function $u_k : \Delta(\mathcal{S}_k) \times \Delta(\mathcal{S}_{-k}) \to \mathbb{R}$. Agents are connected via an underlying network defined by $\mathcal{E}$. In particular, $\mathcal{E}$ consists of pairs $(k, l) \in \mathcal{N} \times \mathcal{N}$ of connected agents k and l. For any agent $k \in \mathcal{N}$, we denote by $\mathcal{N}_k = \{l \in \mathcal{N} : (k, l) \in \mathcal{E}\}$ the *neighbours* of k, i.e. all the agents who directly interact with agent k in the network. An equivalent way to define the network is through an *adjacency matrix* G such that

$$[G]_{k,l} = \begin{cases} 1, & \text{if agents } k, l \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} .$$

It is assumed that the network is undirected so that G is a symmetric matrix. Each edge $(k, l) \in \mathcal{E}$ corresponds to a pair of payoff matrices A^{kl}, A^{lk}. With these specifications, the payoff received by each agent k under joint strategy $\mathbf{x} = (\mathbf{x}_k, \mathbf{x}_{-k})$ is given by

$$u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \sum_{(k,l) \in \mathcal{E}} \mathbf{x}_k \cdot A^{kl} \mathbf{x}_l. \tag{1}$$

For any $\mathbf{x} \in \Delta =: \times_k \Delta(\mathcal{S}_k)$, we can define the reward to agent k for playing action i as $r_{ki}(\mathbf{x}_{-k}) = \partial u_{ki}(\mathbf{x}) / \partial x_{ki}$. Under this notation, $u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \langle \mathbf{x}_k, r_k(\mathbf{x}_{-k}) \rangle$. With this in place, we can define suitable equilibrium solutions for the game.

Definition 2.1 (Nash Equilibrium (NE)). A joint mixed strategy $\bar{\mathbf{x}} \in \Delta$ is a *Nash Equilibrium* (NE) if, for all agents k and all actions $i \in \mathcal{S}_k$

$$\bar{\mathbf{x}}_k = \arg \max_{\mathbf{y}_k \in \Delta_k} u_k(\mathbf{y}_k, \bar{\mathbf{x}}_{-k}).$$

Definition 2.2 (Quantal Response Equilibrium (QRE)). A joint mixed strategy $\bar{\mathbf{x}} \in \Delta$ is a *Quantal Response Equilibrium* (QRE) if, for all agents k and all actions $i \in \mathcal{S}_k$

$$\bar{\mathbf{x}}_{ki} = \frac{\exp(r_{ki}(\bar{\mathbf{x}}_{-k})/T_k)}{\sum_{j \in \mathcal{S}_k} \exp(r_{kj}(\bar{\mathbf{x}}_{-k})/T_k)}.$$

The QRE [5, 35] naturally extends the Nash Equilibrium through the parameter T_k, known as the *exploration rate*. In particular, the limit $T_k \longrightarrow 0$ corresponds exactly to the Nash Equilibrium, whereas the limit $T_k \longrightarrow \infty$ corresponds to the case where action $i \in \mathcal{S}_k$ is played with the same probability regardless of its associated reward. The link between the QRE and the Nash Equilibrium is made precise through the following result.

Proposition 2.3 ([38]). *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k \in \mathcal{N}})$ and let $T_1, \dots, T_N > 0$ be exploration rates. Define the perturbed game $\mathcal{G}^H = (\mathcal{N}, \mathcal{E}, (u_k^H, \mathcal{S}_k)_{k \in \mathcal{N}})$ with the payoff functions*

$$u_k^H(\mathbf{x}_k, \mathbf{x}_{-k}) = u_k(\mathbf{x}_k, \mathbf{x}_{-k}) - T_k \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle.$$

Then $\bar{\mathbf{x}} \in \Delta$ is a QRE of $\mathcal{G}$ iff it is a Nash Equilibrium of $\mathcal{G}^H$.

Game Structure. To achieve our main result, we must parameterise interactions in the network game. This allows us to consider network games which are not necessarily zero-sum. First, we define the *influence bound* for each agent k.

Definition 2.4 (Influence Bound). Let $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k \in \mathcal{N}})$ be a network game. Then, for any $k \in \mathcal{N}$, the *influence bound* is given by

$$\delta_k = \max_{i \in \mathcal{S}_k, a_{-k}, \tilde{a}_{-k} \in \mathcal{S}_{-k}} \{|r_{ki}(a_{-k}) - r_{ki}(\tilde{a}_{-k})|\}, \tag{2}$$

where the pure strategies $a_{-k}, \tilde{a}_{-k} \in \mathcal{S}_{-k}$ differ only in the action of one agent $l \in \mathcal{N}_k$.

The influence bound describes how sensitive each agent's reward is to changes in opponent strategies. As another parameterisation which is directly applicable to network games, we define the *intensity of identical interests.*

Definition 2.5 (Intensity of Identical Interests). Let $\mathcal{G}$ be a network game whose edgeset $\mathcal{E}$ is associated with the payoff matrices $(A^{kl}, A^{lk})_{(k,l) \in \mathcal{E}}$. The *intensity of identical interests* σ_I of $\mathcal{G}$ is given as

$$\sigma_I = \max_{(k,l) \in \mathcal{E}} \|A^{kl} + (A^{lk})^\top\|_2, \tag{3}$$

where $\|M\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|M\mathbf{x}\|_2$ denotes the operator 2-norm [36].

The intensity of identical interests can be thought of as a measure of how *cooperative* a network game is. The reasoning for this is as follows. Suppose A, B are the payoff matrices which maximise (3) and suppose that $B^\top = cA$ for some $c = (-1, 1)$. Then, σ_I is minimised when $c = -1$, in which case A, B is zero-sum, and is maximised at $c = 1$ so that $A = B^\top$, which defines an game of identical interests.

2.2 Learning Model

In this work, we analyse the *Q-Learning dynamic*, a prototypical model for determining optimal policies by balancing exploration and exploitation [55, 59]. In this model, each agent $k \in \mathcal{N}$ maintains a history of the past performance of each of their actions. This history is updated via the Q-update

$$Q_{ki}(\tau + 1) = (1 - \alpha_k) Q_{ki}(\tau) + \alpha_k r_{ki}(\mathbf{x}_{-k}(\tau)),$$

where τ denotes the current time step.

$Q_{ki}(\tau)$ denotes the *Q-value* maintained by agent k about the performance of action $i \in S_k$. In effect, Q_{ki} gives a discounted history of the rewards received when i is played, with $1 - \alpha_k$ as the discount factor.

Given these Q-values, each agent updates their mixed strategies according to the Boltzmann distribution, given by

$$x_{ki}(\tau) = \frac{\exp(Q_{ki}(\tau)/T_k)}{\sum_j \exp(Q_{kj}(\tau)/T_k)},$$

in which $T_k \in [0, \infty)$ is the *exploration rate* of agent k.

It was shown in [54, 64] that a continuous time approximation of the Q-Learning algorithm could be written as

$$\frac{\dot{x}_{ki}}{x_{ki}} = r_{ki}(\mathbf{x}_{-k}) - \langle \mathbf{x}_k, r_k(\mathbf{x}) \rangle + T_k \sum_{j \in S_k} x_{kj} \ln \frac{x_{kj}}{x_{ki}}, \tag{QLD}$$

which we call the *Q-Learning dynamics* (QLD). The fixed points of this dynamic coincide with the (QRE) of the game [32]. QLD can also be seen as an entropy regularised form of the well-studied *replicator dynamics* (RD) [23, 34]. Besides its importance in the study of population biology [42], RD is known to be a special case of the generalised *Follow the Regularised Leader* learning dynamic [40], which models agents who maximise their accumulated payoffs subject to a penalisation function. RD has been shown to display asymptotic convergence in potential games [23], cyclic behaviour in zero-sum games [39] and chaos in a number of other classes [17, 53]. The connection between RD and QLD is explored in [31].

3 GUARANTEED CONVERGENCE OF Q-LEARNING IN NETWORK GAMES

In this section we determine a number of sufficient conditions on the exploration rates T_k under which Q-Learning dynamics converge to a unique QRE. We find that these conditions are dependent on the structure of the rewards in the game, parameterised by the interaction coefficient or the inflence bound, and also on the structure of the network. We then compare our result to that of [26] and show that, under suitable network structures, stability can be achieved with comparatively low exploration rates, even in the presence of many players. This also refines the result of [51] which suggests that learning dynamics are increasingly unstable as the number of players increases, regardless of exploration rate. All proofs are in Appendix B.

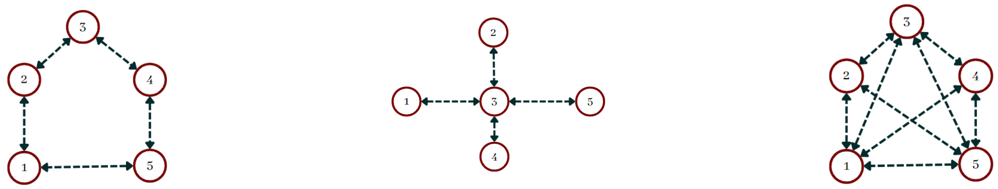
Theorem 3.1. *Consider a network game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k \in \mathcal{N}})$ which has a network adjacency matrix G. Let σ_I denote the intensity of identical interests for $\mathcal{G}$ and δ_k denote the influence bound of each agent $k \in \mathcal{N}$. Then, the Q-Learning Dynamic converges to a unique QRE $\bar{\mathbf{x}} \in \Delta$ if any of the following conditions hold for all agents $k \in \mathcal{N}$,*

$$T_k > \delta_k |\mathcal{N}_k|, \tag{C1}$$

$$T_k > \frac{1}{2} \sigma_I \|G\|_\infty, \tag{C2}$$

where $\|M\|_\infty = \max_i \sum_j |[M]_{ij}|$ is the operator ∞-norm. If, in addition, each edge defines the same bimatrix game (A, B), then asymptotic convergence of Q-Learning Dynamics holds if, for all $k \in \mathcal{N}$

$$T_k > \frac{1}{2} \sigma_I \|G\|_2. \tag{C3}$$

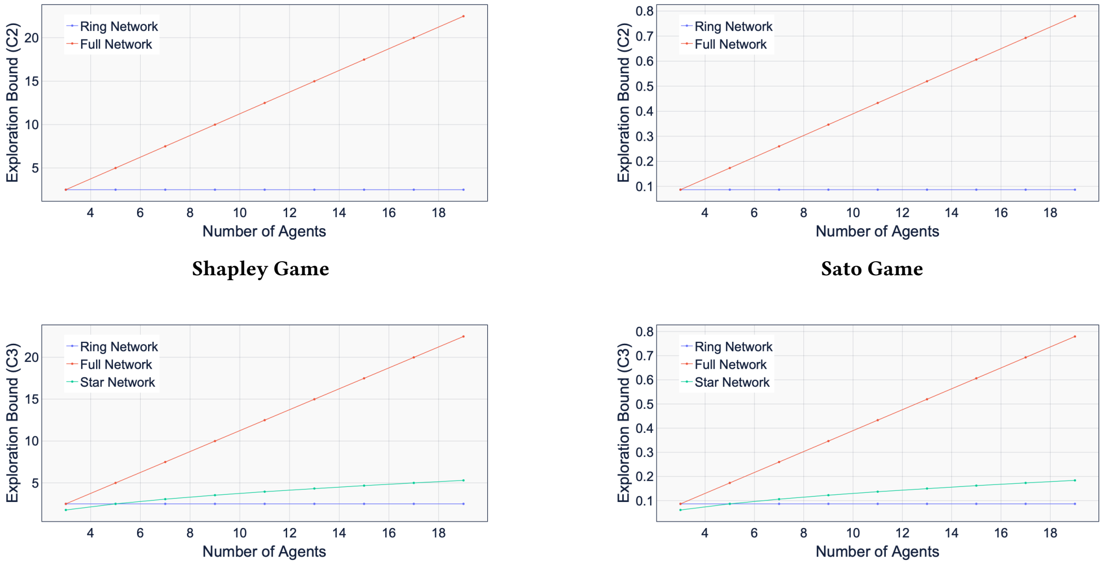


Ring Network
$\|G\|_\infty = 2, \|G\|_2 = 2$

Star Network
$\|G\|_\infty = N - 1, \|G\|_2 = \sqrt{N-1}$

Fully Connected Network
$\|G\|_\infty = N - 1, \|G\|_2 = N - 1$

Figure 1: Examples of networks with five agents and associated $\|G\|_\infty$ and $\|G\|_2$.



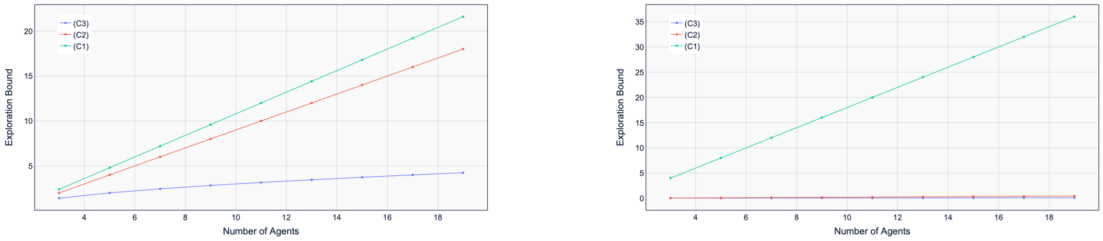
Shapley Game

Sato Game

Shapley Game

Sato Game

Figure 2: Lower Bound on sufficient exploration as defined by (Top) (C2) in a Full Network and Ring Network (Bottom) (C3) in a Full Network, Star Network and Full Network.



Shapley Game

Sato Game

Figure 3: Lower Bound on sufficient exploration as defined by (C1), (C2) and (C3) in a Star Network. For (C1), $\max_{k\in\mathcal{N}} \delta_k |\mathcal{N}_k|$ is depicted which therefore coincides with the condition defined in [26].

Remark 3.2. Condition (C1) immediately refines the result of [26] to the case of network games. In the latter work, the authors implicitly assume that the reward for each agent depends on all other agents. In our work, this corresponds exactly to the case of a fully connected network, where $\mathcal{N}_k = \mathcal{N}\backslash\{k\}$. In addition, [26] define the influence bound to be over all agents, yielding a single condition which must hold for all k. Instead (C1) allows for agents who have a lower δ_k or who are not strongly connected in the network to have lower exploration rates T_k without compromising convergence.

Remark 3.3. We can directly compare (C1) and (C2) due to the definition of the infinity norm. In particular $\|G\|_\infty = \max_k |\mathcal{N}_k|$ is the maximum number of neighbours for any agent $k \in \mathcal{N}$. Therefore, in a network where all agents are connected identically, the network dependency in (C1) is the same as that in (C2) . Next, the advantage of using the influence bound is that its definition applies in games which are not defined by matrices, and so the result generalises outside of network polymatrix games. By contrast, σ_I is often easier to compute than δ_k as it is based on matrix norms rather than pairwise differences. Furthermore, $\frac{1}{2}\sigma_I$ is less than δ_k in a number of polymatrix games (c.f. Sec. 4). In summary, (C1) presents an advantage in terms of generality , whilst (C2) is often easier to compute and can be a tighter bound in network polymatrix games where all agents are identically coupled.

Remark 3.4. Theorem 3.1 applies generally across all network polymatrix games, without making any assumptions, such as the network zero-sum condition. In fact, for networks of pairwise zero sum games, the following holds

Corollary 3.5. *If the network game $\mathcal{G}$ is a pairwise zero-sum matrix, i.e. $A^{kl} + (A^{lk})^\top = 0$ for all $(k, l) \in \mathcal{E}$, then the Q-Learning dynamics converge to a unique QRE so long as exploration rates T_k for all agents are strictly positive.*

Corollary 1 is supported by the result of [26, 32] in which it was shown that Q-Learning converges to a unique QRE in all network zero-sum games, even if they are not pairwise zero-sum , so long as all exploration rates T_k are positive.

Remark 3.6. Whilst (C3) requires a stronger assumption, namely that each edge corresponds to the same bimatrix game, this setting is well motivated in the literature [17, 60]. In addition, it holds that $\|G\|_2 \leq \|G\|_\infty$ for all symmetric matrices G. Therefore, (C3) provides a stronger bound than (C2). Figure 2 depicts (C2) and (C3) on various network games, whilst a direct comparison is visualised in Figure 3.

3.1 QRE as approximate Nash Equilibria

In the following section we compare the QRE as an equilibrium solution to the Nash Equilibrium (NE) condition. In particular we show that the QRE of any game, which no longer needs to be a network game, is close to an NE in the following sense

Definition 3.7 (ϵ-approximate Nash Equilibrium)**.** A strategy $\bar{\mathbf{x}} \in \Delta$ is an *ϵ-approximate Nash Equilibrium* for the game $\mathcal{G}$ if, for all agents k, and all strategies $\mathbf{y}_k \in \Delta_k$

$$u_k(\mathbf{y}_k, \bar{\mathbf{x}}_{-k}) - u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k}) \leq \epsilon.$$

Proposition 3.8. *Consider a game $\mathcal{G}$ and let $T_1, \ldots, T_N > 0$ denote positive exploration rates. Then any QRE $\bar{\mathbf{x}} \in \Delta$ is an ϵ-approximate Nash Equilibrium where*

$$\epsilon = \max_{k \in \mathcal{N}} T_k A_k(\bar{\mathbf{x}}_k), \tag{4}$$

$$A_k(\mathbf{x}_k) = \max_{i \in S_k} \ln x_{ki} - \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle. \tag{5}$$

Remark 3.9. Comparing (4) with (QLD), it can be seen that ϵ denotes the maximum amount of entropy regularisation applied to the payoffs at the QRE $\bar{\mathbf{x}}$. Of course, this depends on the value of $\bar{\mathbf{x}}$ itself. As an example, if the QRE is the uniform distribution, i.e. $\bar{\mathbf{x}}_k = (1/n_k, \ldots, 1/n_k)$ for all agents k, then $A_k(\bar{\mathbf{x}}_k) = 0$. In this case, $\bar{\mathbf{x}}$ is exactly an NE of the game.

Remark 3.10. It is also important to note that value of ϵ given by any QRE $\bar{\mathbf{x}}$ holds exactly. This gives the tightest possible approximation of Nash for any given QRE $\bar{\mathbf{x}}$. Whilst it is largely known that QRE can be considered as approximations of Nash [15, 35, 62], to our knowledge Proposition 3.8 is the first which exactly quantifies the 'distance' between the two equilibrium concepts.

We plot $A_k(\mathbf{x})$ for the case $n_k = 3$ and $n_k = 2$ in the Appendix (Figure 9). To determine its upper bounds, note that $A_k(\bar{\mathbf{x}}_k) \leq \max_{\mathbf{x}_k \in \Delta_k} A_k(\mathbf{x}_k) =: \bar{A}_k$. The form for $\bar{A}_k$ is in general unavailable in closed form and so we give exact values in the Appendix, focusing here on sharp bounds.

Lemma 3.11 (Full version in Lemma C.1)**.**

$$\bar{A}_k := \max_{\mathbf{x}_k \in \Delta_k} \left(\max_{i \in S_k} \ln x_{ki} - \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle \right) = O(\ln n_k).$$

3.2 Updating Exploration Rates

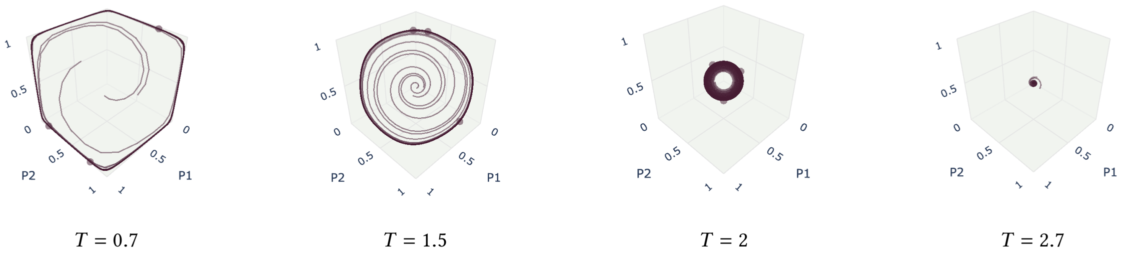
In this section, we use Theorem 3.1 and Proposition 3.8 to devise a scheme to update exploration rates so that which Q-Learning dynamics are driven 'close' to a NE. The full algorithm is provided in the Appendix, with the main ideas discussed here. Starting with a choice of T_k which satisfies any of the conditions in Theorem 3.1, it is clear that agents will achieve an ϵ-NE where ϵ is given by (4). First, we notice that the value of ϵ depends only on the agent who maximises $T_k A_k(\bar{\mathbf{x}}_k)$. Therefore, it is natural to decrease the exploration rate for only this agent. We repeat this process until another agent maximises $T_k A_k(\bar{\mathbf{x}}_k)$, in which case this becomes the agent whose exploration rate is decreased, or the learning dynamics no longer achieve asymptotic convergence, at which point the learning process stops, and the last found QRE is chosen as the final joint strategy of all agents. To evaluate whether the system achieves asymptotic convergence for any choice of T_k, a window of the final H-iterations of learning is recorded and, for each $k \in \mathcal{N}$, $i \in \mathcal{S}_k$ the relative difference between the maximum and minimum value of x_{ki} across the window is determined. If this value is less some tolerance, the system is said to have converged. More formally the dynamics are said to have converged if

$$\left(\frac{\max_{t \in H} x_{ki}(t) - \min_{t \in H} x_{ki}(t)}{\max_{t \in H} x_{ki}(t)} \right) < l. \tag{6}$$

By following this process, agents iteratively reach QRE which are closer approximations of an NE. We evaluate this process in our experiments and show that, even in large scale games, the ϵ-approximation of the NE improves leading to optimal, and stable, learned joint strategies.

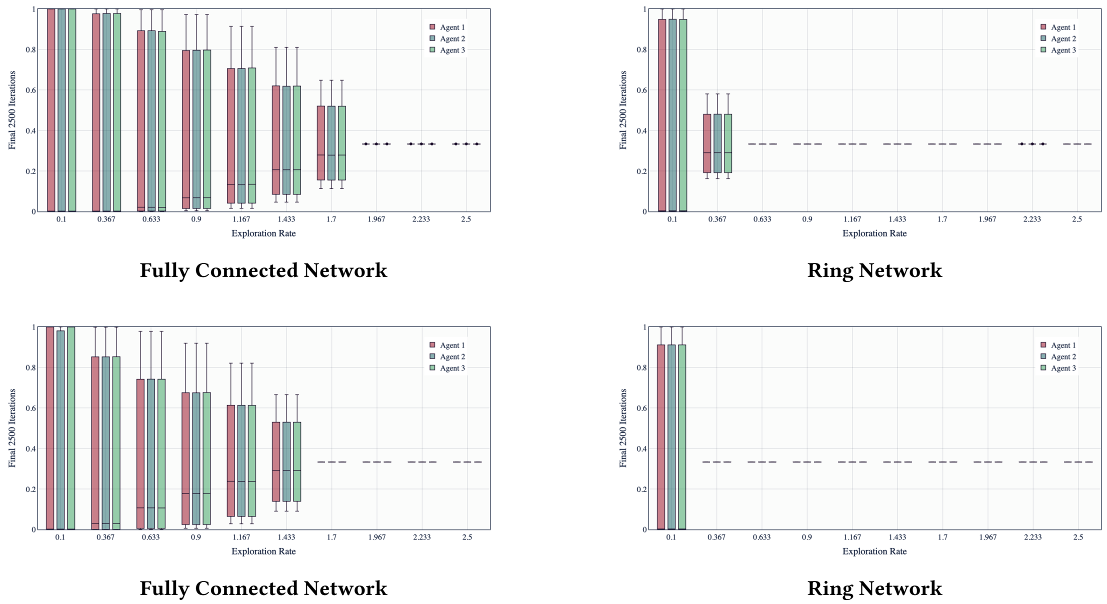
4 EXPERIMENTS

We first visualise and exemplify the implications of our main result, Theorem 3.1, on a number of games. In particular, we simulate the Q-Learning algorithm described in Section 2.2 and show that



$T = 0.7$ $T = 1.5$ $T = 2$ $T = 2.7$

Figure 4: Trajectories of Q-Learning in a three agent Network Chakraborty Game with $\alpha = 7, \beta = 8.5$. Axes denote the probabilities with which each player chooses their first action.



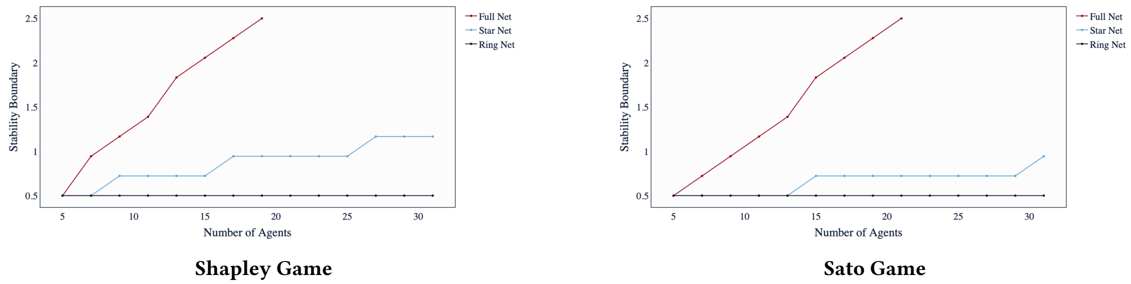
Fully Connected Network

Ring Network

Fully Connected Network

Ring Network

Figure 5: Q-Learning in the (Top) Network Shapley Game (Bottom) Network Sato Game with 15 agents. The boxplot depicts the probabilities with which three of the agents play their first action in the final 2500 iterations of learning. This is depicted for varying choices of exploration rate T.



Shapley Game

Sato Game

Figure 6: Empirically determined stability boundary of Q-Learning measured against the number of agents. Q-Learning is iterated with 10 initial conditions and the game is considered to have converged if, for all agents and initial conditions (6) holds with $l = 1 \times 10^{-5}$. The Fully Connected Network, Star Network and Ring Networks are considered.

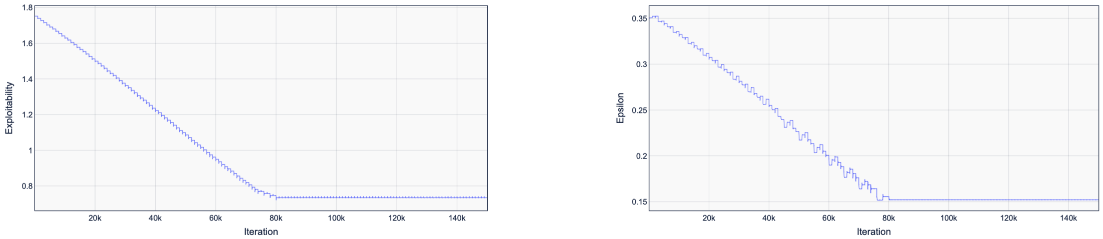


Figure 7: Measures of 'closeness' to Nash Equilibrium as the exploration update scheme is applied to the Network Chakraborty Game with five agents and $\alpha = 2.5, \beta = 1.5$. (Left) Distance to NE measured by exploitability (7) of the joint strategy $\mathrm{x}(t)$. (Right) ϵ as defined by (4). Both metrics decreases as exploration rates are updated until condition (6) fails at approx. 8×10^4 iterations, after which learning is halted.

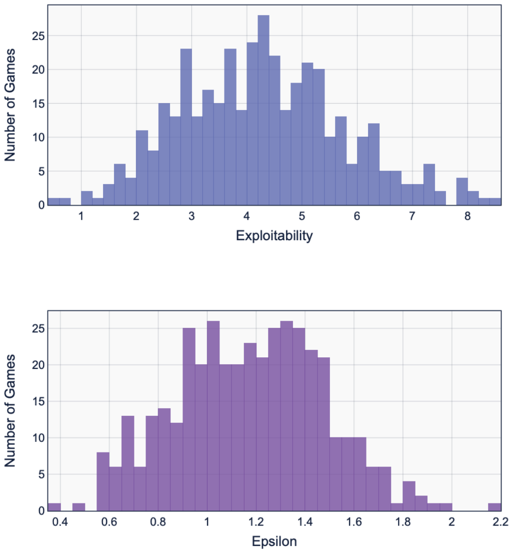


Figure 8: Histograms depicting the decrease of (Top) Exploitability and (Bottom) ϵ over $150,000$ iterations of learning across 500 randomly generated network games with payoffs bounded in $[0, 5]$.

Q-Learning asymptotically approaches a unique QRE so long as the exploration rates are sufficiently large. We show, in particular, that the amount of exploration required depends on the structure of the network rather than the total number of agents.

Remark 4.1. In our experiments, we take all agents k to have the same exploration rate T and so drop the k notation. As all bounds in Theorem 3.1 must hold for all agents k, this assumption does not affect the generality of the results.

4.1 Convergence of Q-Learning

We first illustrate the convergence of Q-Learning using the *Network Chakraborty Game*, which was analysed in [43] to characterise chaos in learning dynamics. Formally, the payoff to each agent k is defined as

$$u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \mathbf{x}_k^\top \mathbf{A} \mathbf{x}_l, \; l = k - 1 \mod N,$$
$$A = \begin{pmatrix} 1 & \alpha \\ \beta & 0 \end{pmatrix}, \; \alpha, \beta \in \mathbb{R}.$$

We visualise the trajectories generated by running Q-Learning in Figure 4 for a three agent network and choosing $\alpha = 7, \beta = 8.5$. It can be seen that, for low exploration rates, the dynamics reach a limit cycle around the boundary of the simplex. However, as exploration increases, the dynamics are eventually driven towards a fixed point for all initial conditions.

Network Shapley Game. In the following example, each edge of the network game has associated the same pair of matrices A, B where

$$A = \begin{pmatrix} 1 & 0 & \beta \\ \beta & 1 & 0 \\ 0 & \beta & 1 \end{pmatrix}, \; B = \begin{pmatrix} -\beta & 1 & 0 \\ 0 & -\beta & 1 \\ 1 & 0 & -\beta \end{pmatrix},$$

where $\beta \in (0, 1)$.

This has been analysed in the two-agent case in [56], where it was shown that the *Fictitious Play* learning dynamic do not converge to an equilibrium. [26] analysed the network variant of this game for the case of a ring network and numerically showed that convergence can be achieved by Q-Learning through sufficient exploration.

In Figure 5 we examine both a fully connected network and a ring network with 15 agents. In this case, the dynamics evolve in $\mathbb{R}^{45}$ which prohibits a visualisation of the complete dynamics. To resolve this, we instead take three representative agents and depict the *spread* of their strategies in the final 2500 iterations of learning. A bar which stretches from 0 to 1 indicates that the dynamics are spread across the simplex which may occur in a limit cycle or chaotic orbit that approaches the boundary of the simplex (c.f. Figure 4). These are seen to occur for low exploration rates. By contrast, when exploration rates are increased beyond a certain threshold, a flat line is seen which indicates that the dynamics are stationary, i.e. a fixed point has been reached. Importantly, the boundary at which equilibrium behaviour occurs is higher in the fully connected network, where $\|G\|_\infty = 14$ than in the ring network, where $\|G\|_\infty = 2$. This indicates that larger numbers of agents may be introduced in the environment without impacting stability, so long as a weakly connected network is chosen.

Network Sato Game. We also analyse the behaviour of Q-Learning in a variant of the game introduced in [53], where it was shown that chaotic behaviour is exhibited by learning dynamics in the two-agent case. We extend this towards a network game by associating each edge with the payoff matrices A, B given by

$$A = \begin{pmatrix} \epsilon_X & -1 & 1 \\ 1 & \epsilon_X & -1 \\ -1 & 1 & \epsilon_X \end{pmatrix}, \; B = \begin{pmatrix} \epsilon_Y & -1 & 1 \\ 1 & \epsilon_Y & -1 \\ -1 & 1 & \epsilon_Y \end{pmatrix},$$

where $\epsilon_X, \epsilon_Y \in \mathbb{R}$. Notice that for $\epsilon_X = \epsilon_Y = 0$, this corresponds to the classic Rock-Paper-Scissors game which is zero-sum so that, by Corollary 1, Q-Learning will converge to an equilibrium with any positive exploration rates. We choose $\epsilon_X = 0.01, \epsilon_Y = -0.05$ in order to stay consistent with [53] which showed chaotic dynamics for this choice. The boxplot once again shows that sufficient exploration leads to convergence of all initial conditions. However, the amount of exploration required is significantly smaller than that of the Network Shapley Game. This can be seen as being due to the significantly lower interaction coefficient of the Sato game $\sigma_I = 0.05$ as compared to the Shapley game $\sigma_I = 2$.

4.2 Stability Boundary

In these experiments we empirically determine the dependence of the stability boundary w.r.t. the number of agents. For accurate comparison with Figure 2, we consider the Network Sato and Shapley Games in a fully-connected network, star network and ring network. We iterate Q-Learning for various values of T and determine whether the dynamics have converged. To evaluate convergence, we apply (6) with $|H| = 2500$ iterations and $l = 1 \times 10^{-5}$. In Figure 6, we plot the smallest exploration rate T for which (6) holds for varying choices of N. It can be seen that the prediction of Theorem 3.1 holds, in that the number of agents plays no impact for the ring network whereas the increase in the fully-connected network is linear in N. In addition, it is clear that the stability boundary increases slower in the Sato game than in the Shapley game, owing to the smaller interaction coefficient.

An additional point to note is that the stability boundary for the star network increases slower than the fully-connected network in all games. We anticipate that this is due to the fact that the 2-norm $\|G\|_2$ in the star network is smaller than that of the fully-connected network (c.f. Figure 1).

4.3 Effectiveness of Exploration Update Scheme

In these experiments, we evaluate the exploration update scheme outlined in Section 3.2. using $|H| = 500$ and $l = 1 \times 10^{-5}$. In Figure 7 we consider the Network Chakraborty Game with $\alpha = 2.5, \beta = 1.5$ We measure the 'distance' between the strategy $\mathbf{x}(t)$ and the NE using two metrics: first by ϵ as given in (4) and second through *exploitability* $\phi(\mathbf{x})$ given as

$$\phi(\mathbf{x}) = \sum_k \max_{\mathbf{y}_k \in \Delta_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - u_k(\mathbf{x}_k, \mathbf{x}_{-k}). \qquad (7)$$

The exploitability is used, sometimes under different names, as a measure of distance to the NE [15, 49] and, from (4) it can be seen that $\phi(\bar{\mathbf{x}}) = \sum_k T_k A_k(\bar{\mathbf{x}}_k)$ for any QRE $\bar{\mathbf{x}}$. The reason for examining ϕ is that its definition holds for any strategy $\mathbf{x} \in \Delta$, whilst (4) only holds at a QRE $\bar{\mathbf{x}} \in \Delta$. It can be seen in all cases that both metrics decrease as agents learn, until condition (6) is no longer satisfied. To examine the generality of this performance, we evaluate the exploration update scheme in 500 randomly generated network games with 15 agents, two actions and a ring structure. Exploitability and ϵ are evaluated at the first iteration and final iteration and the difference is recorded. Figure 8 plots the decrease of both metrics as a histogram across all 500 games. These experiments (as well as additional presented in Appendix D) suggest that, if exploration rates are updated according the scheme in Section 3.2, independent learning agents may learn stable equilibrium strategies which closely approximate Nash Equilibria.

5 CONCLUSION

In this paper we show that the Q-Learning dynamics is guaranteed to converge in arbitrary network games, independent of any restrictive assumptions such as network zero-sum or potential. This allows us to make a branching statement which applies across all network games.

In particular, our analysis shows that convergence of the Q-Learning dynamics can be achieved through sufficient exploration, where the bound depends on the pairwise interaction between agents and the structure of the network. Overall, compared to the literature, we are able to tighten the bound on sufficient exploration and show that, under certain network interactions, the bound does not increase with the total number of agents. This allows for stability to be guaranteed in network games with many players.

A fruitful direction for future research would be to capture the effect of the payoffs through a tighter bound than the interaction coefficient and to explore further how properties of the network affect the bound. In addition, whilst there is still much to learn in the behaviour of Q-Learning in stateless games, the introduction of the state variable in the Q-update is a valuable next step.

ACKNOWLEDGMENTS

Aamal Hussain and Francesco Belardinelli are partly funded by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (grant number EP/S023356/1). Dan Leonte acknowledges support from the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1). This research was supported in part by the National Research Foundation, Singapore and DSO National Laboratories under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016), grant PIESGP-AI-2020-01, AME Programmatic Fund (Grant No.A20H6b0151) from A*STAR.

REFERENCES

[1] Natalia Amelina, Alexander Fradkov, Yuming Jiang, and Dimitrios J Vergados. 2015. Approximate Consensus in Stochastic Networks with Application to Load Balancing. *IEEE Transactions on Information Theory* 61, 4 (9 2015), 1739–1752. https://doi.org/10.1109/TIT.2015.2406323

[2] Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. 2006. Who's Who in Networks. Wanted: The Key Player. *Econometrica* 74, 5 (2006), 1403–1417. http://www.jstor.org/stable/3805930

[3] Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Michał Misiurewicz, and Georgios Piliouras. 2021. Follow-the-Regularized-Leader Routes to Chaos in Routing Games. (2 2021). http://arxiv.org/abs/2102.07974

[4] Yann Bramoullé, Rachel Kranton, and Martin D'Amours. 2014. Strategic Interaction and Networks. *The American Economic Review* 104, 3 (2014), 898–930. http://www.jstor.org/stable/42920723

[5] Colin F. Camerer, Teck Hua Ho, and Juin Kuan Chong. 2004. Behavioural game theory: Thinking, learning and teaching. *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality* (1 2004), 120–180. https://doi.org/10.1057/9780230523371{_}8/COVER

[6] Yun Kuen Cheung and Georgios Piliouras. 2019. Vortices Instead of Equilibria in MinMax Optimization: Chaos and Butterfly Effects of Online Learning in Zero-Sum Games. In *Proceedings of the Thirty-Second Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 99)*, Alina Beygelzimer and Daniel Hsu (Eds.). PMLR, 807–834. https://proceedings.mlr.press/v99/cheung19a.html

[7] Yun Kuen Cheung and Yixin Tao. 2020. Chaos of Learning Beyond Zero-sum and Coordination via Game Decompositions. (8 2020). http://arxiv.org/abs/2008.00540

[8] Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. 2019. The route to chaos in routing games: Population increase drives period-doubling instability, chaos & inefficiency with Price of Anarchy equal to one. (2019). http://arxiv.org/abs/1906.02486

[9] Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. 2019. The route to chaos in routing games: When is Price of Anarchy too optimistic? (2019). http://arxiv.org/abs/1906.02486

[10] Roberto Cominetti, Emerson Melo, and Sylvain Sorin. 2010. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior* 70, 1 (2010), 71–83. https://doi.org/10.1016/j.geb.2008.11.012

[11] Aleksander Czechowski and Georgios Piliouras. 2022. Poincaré-Bendixson Limit Sets in Multi-Agent Learning; Poincaré-Bendixson Limit Sets in Multi-Agent Learning. In *International Conference on Autonomous Agents and Multiagent Systems*. www.ifaamas.org

[12] Francisco Facchinei and Jong Shi Pang. 2004. Finite-Dimensional Variational Inequalities and Complementarity Problems. *Finite-Dimensional Variational Inequalities and Complementarity Problems* (2004). https://doi.org/10.1007/B97543

[13] Tobias Galla. 2011. Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment* 2011, 8 (8 2011). https://doi.org/10.1088/1742-5468/2011/08/P08007

[14] Tobias Galla and J. Doyne Farmer. 2013. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences of the United States of America* 110, 4 (2013), 1232–1236. https://doi.org/10.1073/pnas.1109672110

[15] Ian Gemp, Rahul Savani, Marc Lanctot, Yoram Bachrach, Thomas Anthony, Richard Everett, Andrea Tacchetti, Tom Eccles, and János Kramár. 2022. Sample-Based Approximation of Nash in Large Many-Player Games via Gradient Descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 507–515.

[16] Sergio Grammatico, Francesca Parise, Marcello Colombino, and John Lygeros. 2016. Decentralized Convergence to Nash Equilibria in Constrained Deterministic Mean Field Control. *IEEE Trans. Automat. Control* 61, 11 (11 2016), 3315–3329. https://doi.org/10.1109/TAC.2015.2513368

[17] Christopher Griffin, Justin Semonsen, and Andrew Belmonte. 2022. Generalized Hamiltonian Dynamics and Chaos in Evolutionary Games on Networks. *Physica A: Statistical Mechanics and its Applications* 597 (7 2022).

[18] Saeed Hadikhanloo, Rida Laraki, Panayotis Mertikopoulos, and Sylvain Sorin. 2022. Learning in nonatomic games part I Finite action spaces and population games. *Journal of Dynamics and Games. 2022* 0, 0 (2022), 0. https://doi.org/10.3934/JDG.2022018

[19] Heiko Hamann. 2018. *Swarm Robotics: A Formal Approach*. Springer International Publishing. https://doi.org/10.1007/978-3-319-74528-2

[20] Christopher Harris. 1998. On the Rate of Convergence of Continuous-Time Fictitious Play. *Games and Economic Behavior* 22, 2 (2 1998), 238–259. https://doi.org/10.1006/game.1997.0582

[21] P Jean-Jacques Herings and Ronald Peeters. 2010. Homotopy methods to compute equilibria in game theory. *Economic Theory* 42, 1 (2010), 119–156. https://doi.org/10.1007/s00199-009-0441-5

[22] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. 2018. MGAN: Training Generative Adversarial Nets with Multiple Generators. In *International Conference on Learning Representations*.

[23] Josef Hofbauer and Karl Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173179

[24] Abdolhossein Hoorfar and Mehdi Hassani. 2008. Inequalities on the Lambert W function and hyperpower function. *J. Inequal. Pure and Appl. Math* 9, 2 (2008), 5–9.

[25] Aamal Hussain, Francesco Belardinelli, and Georgios Piliouras. 2023. Beyond Strict Competition: Approximate Convergence of Multi Agent Q-Learning Dynamics. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*. https://www.ijcai.org/proceedings/2023/0016.pdf

[26] Aamal Abbas Hussain, Francesco Belardinelli, and Georgios Piliouras. 2023. Asymptotic Convergence and Performance of Multi-Agent Q-Learning Dynamics. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1578–1586.

[27] Lorens A. Imhof, Drew Fudenberg, and Martin A. Nowak. 2005. Evolutionary cycles of cooperation and defection. In *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 102. 10797–10800. https://doi.org/10.1073/pnas.0502589102

[28] Amit Kadan and Hu Fu. 2021. Exponential Convergence of Gradient Methods in Concave Network Zero-Sum Games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12458 LNAI (2021), 19–34. https://doi.org/10.1007/978-3-030-67661-2{_}2/FIGURES/3

[29] Ardeshir Kianercy and Aram Galstyan. 2012. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 85, 4 (4 2012), 041145. https://doi.org/10.1103/PhysRevE.85.041145

[30] Robert Kleinberg, Katrina Ligett, Georgios Piliouras, and Eva Tardos. 2011. Beyond the Nash Equilibrium Barrier. *Innovations in Computer Science* (2011).

[31] Stefanos Leonardos and Georgios Piliouras. 2022. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence* 304 (2022), 103653. https://doi.org/10.1016/j.artint.2021.103653

[32] Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. 2021. Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality. *Advances in Neural Information Processing Systems* 34 (12 2021), 26318–26331.

[33] Chongxuan LI, Taufik Xu, Jun Zhu, and Bo Zhang. 2017. Triple Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/86e78499eeb33fb9cac16b7555b50767-Paper.pdf

[34] J. Maynard Smith. 1974. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology* 47, 1 (9 1974), 209–221. https://doi.org/10.1016/0022-5193(74)90110-6

[35] Richard D. McKelvey and Thomas R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10, 1 (7 1995), 6–38. https://doi.org/10.1006/GAME.1995.1023

[36] James D. Meiss. 2007. *Differential Dynamical Systems*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718232

[37] Emerson Melo. 2018. A Variational Approach to Network Games. *SSRN Electronic Journal* (11 2018). https://doi.org/10.2139/SSRN.3143468

[38] Emerson Melo. 2021. On the Uniqueness of Quantal Response Equilibria and Its Application to Network Games. *SSRN Electronic Journal* (6 2021). https://doi.org/10.2139/SSRN.3631575

[39] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. 2018. Cycles in adversarial regularized learning. *Proceedings* (2018), 2703–2717. https://doi.org/10.1137/1.9781611975031.172

[40] Panayotis Mertikopoulos and William H. Sandholm. 2016. Learning in Games via Reinforcement and Regularization. *https://doi.org/10.1287/moor.2016.0778* 41, 4 (8 2016), 1297–1324. https://doi.org/10.1287/MOOR.2016.0778

[41] Andrew I Metrick and Ben Polak. 1994. Fictitious play in 2 • 2 games: a geometric proof of convergence*. *Econ. Theory* 4 (1994), 923–933.

[42] Archan Mukhopadhyay and Sagar Chakraborty. 2020. Deciphering chaos in evolutionary games. *Chaos* 30, 12 (12 2020), 121104. https://doi.org/10.1063/5.0029480

[43] Varun Pandit, Archan Mukhopadhyay, and Sagar Chakraborty. 2018. Weight of fitness deviation governs strict physical chaos in replicator dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28, 3 (3 2018), 033104. https://doi.org/10.1063/1.5011955

[44] Marco Pangallo, Torsten Heinrich, and J. Doyne Farmer. 2019. Best reply structure and equilibrium convergence in generic games. *Science Advances* 5, 2 (2 2019). https://doi.org/10.1126/SCIADV.AAT1328/SUPPL{_}FILE/AAT1328{_}SM.PDF

[45] Marco Pangallo, James B.T. Sanders, Tobias Galla, and J. Doyne Farmer. 2022. Towards a taxonomy of learning dynamics in 2 × 2 games. *Games and Economic Behavior* 132 (3 2022), 1–21. https://doi.org/10.1016/J.GEB.2021.11.015

[46] Francesca Parise, Sergio Grammatico, Basilio Gentile, and John Lygeros. 2020. Distributed convergence to Nash equilibria in network and average aggregative games. *Automatica* 117 (2020), 108959. https://doi.org/10.1016/j.automatica.2020.108959

[47] Francesca Parise and Asuman Ozdaglar. 2019. A variational inequality framework for network games: Existence, uniqueness, convergence and sensitivity analysis. *Games and Economic Behavior* 114 (3 2019), 47–82. https://doi.org/10.1016/j.geb.2018.11.012

[48] 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. 2020. *From poincaré recurrence to convergence in imperfect information games: finding equilibrium via regularization*. Technical Report.

[49] Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. In *Advances in Neural Information Processing Systems*, H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (Eds.), Vol. 33. Curran Associates, Inc., 13199–13213. https://proceedings.neurips.cc/paper/2020/file/

995ca733e3657ff9f5f3c823d73371e1-Paper.pdf

[50] J Rosen. 1965. Existence and Uniqueness of Equilibrium Points for Concave N-Person Games. *Econometrica* 33, 3 (1965).

[51] James B T Sanders, J Doyne Farmer, and Tobias Galla. 2018. The prevalence of chaotic dynamics in games with many players. *Scientific Reports* 8, 1 (2018), 4902. https://doi.org/10.1038/s41598-018-22013-5

[52] Yuzuru Sato, Eizo Akiyama, and James P Crutchfield. 2004. Stability and diversity in collective adaptation. *Physica D: Nonlinear Phenomena* 210 (2004), 21–57.

[53] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. 2002. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences of the United States of America* 99, 7 (4 2002), 4748–4751. https://doi.org/10.1073/pnas.032086299

[54] Yuzuru Sato and James P. Crutchfield. 2003. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E* 67, 1 (1 2003), 015206. https://doi.org/10.1103/PhysRevE.67.015206

[55] Howard M. Schwartz. 2014. *Multi-Agent Machine Learning: A Reinforcement Approach*. Wiley. 1–242 pages. https://doi.org/10.1002/9781118884614

[56] L. S. Shapley. 2016. Some Topics in Two-Person Games. In *Advances in Game Theory. (AM-52)*. Princeton University Press, 1–28. https://doi.org/10.1515/9781400882014-002

[57] Mohammad Shokri and Hamed Kebriaei. 2020. Leader-Follower Network Aggregative Game with Stochastic Agents' Communication and Activeness. *IEEE Trans. Automat. Control* 65, 12 (12 2020), 5496–5502. https://doi.org/10.1109/TAC.2020.2973807

[58] Sylvain Sorin and Cheng Wan. 2016. Finite composite games: Equilibria and dynamics. *Journal of Dynamics and Games* 3, 1 (2016), 101–120.

[59] R Sutton and A Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press. http://incompleteideas.net/book/the-book-2nd.html

[60] György Szabó and Gábor Fáth. 2007. Evolutionary games on graphs. *Physics Reports* 446, 4 (2007), 97–216. https://doi.org/10.1016/j.physrep.2007.04.004

[61] Tatiana Tatarenko and Maryam Kamgarpour. 2019. Learning Nash Equilibria in Monotone Games. *Proceedings of the IEEE Conference on Decision and Control* 2019-December (12 2019), 3104–3109. https://doi.org/10.1109/CDC40024.2019.9029659

[62] Theodore L Turocy. 2005. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior* 51, 2 (2005), 243–263. https://doi.org/10.1016/j.geb.2004.04.003

[63] Karl Tuyls. 2023. Multiagent Learning: From Fundamentals to Foundation Models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1.

[64] Karl Tuyls, Pieter Jan T Hoen, and Bram Vanschoenwinkel. 2006. An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games. *Autonomous Agents and Multi-Agent Systems* 12, 1 (2006), 115–153. https://doi.org/10.1007/s10458-005-3783-9

[65] Sebastian van Strien and Colin Sparrow. 2011. Fictitious play in 3×3 games: Chaos and dithering behaviour. *Games and Economic Behavior* 73, 1 (2011), 262–286. https://doi.org/10.1016/j.geb.2010.12.004

[66] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. 2020. No-Regret Learning and Mixed Nash Equilibria: They Do Not Mix. *Advances in Neural Information Processing Systems* 33 (2020), 1380–1391.

A PRELIMINARIES

In this section we outline the various tools and properties that we will use in our proofs.

A.1 Variational Inequalities and Monotone Games

Our aim in this work is to analyse the Q-Learning dynamics in network games without invoking any particular structure on the payoffs (e.g. zero-sum). To do this, we employ the *Variational Inequality* approach, which has been successfully applied towards the analysis of network games [37, 47] as well as learning in games [18, 26, 58].

Definition A.1 (Variational Inequality). Consider a set $\mathcal{X} \subset \mathbb{R}^d$ and a map $F : \mathcal{X} \to \mathbb{R}^d$. The Variational Inequality (VI) problem $VI(\mathcal{X}, F)$ is given as

$$\langle \mathbf{x} - \bar{\mathbf{x}}, F(\bar{\mathbf{x}}) \rangle \geq 0, \qquad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{8}$$

We say that $\bar{\mathbf{x}} \in \mathcal{X}$ belongs to the set of solutions to a variational inequality problem $VI(\mathcal{X}, F)$ if it satisfies (8).

The premise of the variational approach to game theory [12, 50] is that the problem of finding equilibria of games can be reformulated as determining the set of solutions to a VI problem. This is done by choosing associating the set $\mathcal{X}$ with Δ and the map F with the *pseudo-gradient* of the game.

Definition A.2 (Pseudo-Gradient Map). The pseudo-gradient map of a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k\in\mathcal{N}})$ is given by $F(\mathbf{x}) = (F_k(\mathbf{x}))_{k\in\mathcal{N}} = (-D_{\mathbf{x}_k} u_k(\mathbf{x}_k, \mathbf{x}_{-k}))_{k\in\mathcal{N}}$.

The advantage of this formulation is that we can apply results from the study of Variational Inequalities to determine properties of the game. These results rely solely on the form of the pseudo-gradient map and so can generalise results which assume a potential or zero-sum structure of the game [26, 28].

Lemma A.3 ([38]). *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k\in\mathcal{N}})$ and for any $T_1, \dots, T_N > 0$, let F be the pseudo-gradient map of $\mathcal{G}^H$. Then $\bar{\mathbf{x}} \in \Delta$ is a QRE of $\mathcal{G}$ if and only if $\bar{\mathbf{x}}$ is a solution to $VI(\Delta, F)$.*

With this correspondence in place, we can analyse properties of the pseudo-gradient map and its relation to properties of the game and the learning dynamic. One important property is *monotonicity.*

Definition A.4. A map $F : \mathcal{X} \to \mathbb{R}^d$ is

(1) *Monotone* if, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0.$$

(2) *Strongly Monotone* with constant $\alpha > 0$ if, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \alpha ||\mathbf{x} - \mathbf{y}||_2^2.$$

Definition A.5 (Monotone Game). A game $\mathcal{G}$ is *monotone* if its pseudo-gradient map is monotone.

A large part of our analysis will be in determining conditions under which the pseudo-gradient map is monotone. Upon doing so, we are able to employ the following results.

Lemma A.6 ([38]). *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k\in\mathcal{N}})$ and for any $T_1, \dots, T_N > 0$, let F be the pseudo-gradient map of $\mathcal{G}^H$. $\mathcal{G}$ has a unique QRE $\bar{\mathbf{x}} \in \Delta$ if F is strongly monotone with any $\alpha > 0$.*

Lemma A.7 ([26]). *If the game G is* monotone*, then the Q-Learning Dynamics (QLD) converge to the unique QRE with any positive exploration rates $T_1, \dots, T_N > 0$.*

Finally, recall that an operator $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}^n$ is *strongly convex* with constant α if, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + Df(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2} ||\mathbf{x} - \mathbf{y}||_2^2.$$

It is known that, if $f(\mathbf{x})$ is strongly convex, then its Hessian $D_{\mathbf{x}}^2 f(\mathbf{x})$ is strongly positive definite with constant α. Thus, all eigenvalues of $D_{\mathbf{x}}^2 f(\mathbf{x})$ are larger than α. To apply this in our setting, we use the following result.

Proposition A.8 ([38]). *The function $f(\mathbf{x}_k) = T_k \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle$ is strongly convex with constant T_k.*

A.2 Matrix Norms

In addition, the following definitions and properties hold for any matrix A.

(1) $||A||_2 = \sqrt{\lambda_{\max}(A^\top A)}$ where $\lambda_{\max}$ is the largest eigenvalue of A,
(2) $||A||_\infty = \max_i \sum_j |[A]_{ij}|$,
(3) $\rho(A) = \max_i |\lambda_i(A)|$ where $\lambda_i(A)$ denotes an eigenvalue of A.

Proposition A.9 (Weyl's Inequality). *Let $J = D + N$ where D and N are symmetric matrices. Then it holds that*

$$\lambda_{\min}(J) \geq \lambda_{\min}(D) + \lambda_{\min}(N).$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a matrix.

Proposition A.10. *Let G, A be matrices and $\otimes$ denote the Kronecker product. Then*

$$\|G \otimes A\|_2 = \|G\|_2 \|A\|_2.$$

Proposition A.11. *Let A be a symmetric matrix. Then*

$$|\lambda_{\min}(A)| \leq \rho(A) = \|A\|_2.$$

The following result is used in our proof to be able to parameterise the effect of pairwise interactions by σ_I.

Lemma A.12. *Let $G \in \mathcal{M}_N(\mathbb{R})$ be matrix for which each entry $g_{ij} := [G]_{ij}$ is either 0 or 1. Let $N \in \mathcal{M}_{Nn}(\mathbb{R})$ be a block matrix such that*

$$[N]_{ij} = \begin{cases} A^{ij} & \textit{if } g_{ij} = 1 \\ \mathbf{0} & \textit{otherwise} \end{cases},$$

where $A^{ij} \in \mathcal{M}_n(\mathbb{R})$ are matrices of the same dimension. let $A \in M_n(\mathbb{R})$ be a matrix which satisfies $\|A\|_2 \geq \|B^{ij}\|_2$ for all $(i, j) \in \mathcal{E}$. Finally let $\tilde{N} \in M_{Nn}(\mathbb{R})$ be a block matrix given by Then

$$\|N\|_2 \leq \sqrt{\|G\|_1 \, \|G\|_\infty} \max_{1 \leq i,j \leq n} \left\|A_{ij}\right\|_2.$$

Proof. Let $v = (v^1, \ldots, v^n) \in \mathbb{R}^{Nn}$ where $v^i \in \mathbb{R}^N$ for $1 \leq i \leq n$. Then

$$\|Nv\|_2^2 = \left\| \begin{pmatrix} g_{11}A^{11} & \ldots & g_{1n}A^{1n} \\ \vdots & & \vdots \\ g_{n1}A^{n1} & \ldots & g_{nn}A^{nn} \end{pmatrix} \begin{bmatrix} v^1 \\ \vdots \\ v^n \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sum_{1j} g_{1j} A^{1j} v^j \\ \vdots \\ \sum_{ni} g_{nj} A^{nj} v^j \end{bmatrix} \right\|_2^2 \leq \sum_{i=1}^n \left\| \sum_{j=1}^n g_{ij} A^{ij} v^j \right\|_2^2. \tag{9}$$

For each fixed $i \in \{1, \ldots, n\}$, we have the upper bound

$$\left\| \sum_{j=1}^n g_{ij} A^{ij} v^j \right\|_2 \leq \sum_{j=1}^n g_{ij} \left\|A^{ij} v^j\right\|_2 \leq \sum_{j=1}^n g_{ij} \left\|A^{ij}\right\|_2 \left\|v^j\right\|_2 \leq \max_{1 \leq i,j \leq n} \left\|A^{ij}\right\|_2 \sum_{j=1}^n g_{ij} \left\|v^j\right\|_2. \tag{10}$$

By plugging (10) in (9) and expanding the squared bracket, we obtain that

$$\begin{aligned} \|Nv\|_2^2 \leq \sum_{i=1}^n \left(\max_{1 \leq i,j \leq n} \left\|A^{ij}\right\|_2 \sum_{j=1}^n g_{ij} \left\|v^j\right\|_2 \right)^2 &= \max_{1 \leq i,j \leq n} \left\|A^{ij}\right\|_2^2 \sum_{i=1}^n \sum_{k,l=1}^n g_{ik} g_{il} \left\|v^k\right\|_2 \left\|v^l\right\|_2 \\ &\leq \max_{1 \leq i,j \leq n} \left\|A^{ij}\right\|_2^2 \sum_{i=1}^n \sum_{k,l=1}^n g_{ik} g_{il} \left(\frac{1}{2} \left\|v^k\right\|_2^2 + \frac{1}{2} \left\|v^l\right\|_2^2 \right), \end{aligned}$$

where the last inequality follows by completing the square. Notice that the two sums above are identical, hence

$$\|Nv\|_2^2 \leq \max_{1 \leq i,j \leq n} \left\|A^{ij}\right\|_2^2 \sum_{i=1}^n \sum_{k,l=1}^n g_{ik} g_{il} \left\|v^k\right\|_2^2.$$

It remains the upper bound the RHS in the above inequality. Indeed, we have that

$$\begin{aligned} \sum_{i=1}^n \sum_{k,l=1}^n g_{ik} g_{il} \left\|v^k\right\|_2^2 = \sum_{i=1}^n \sum_{k=1}^n g_{ik} \left\|v^k\right\|_2^2 \left(\sum_{l=1}^n g_{il} \right) &\leq \|G_\infty\| \sum_{i=1}^n \sum_{k=1}^n g_{ik} \left\|v^k\right\|_2^2 \\ &\leq \|G_\infty\| \sum_{k=1}^n \left(\sum_{i=1}^n g_{ik} \right) \left\|v^k\right\|_2^2 \leq \|G_\infty\| \, \|G_1\| \sum_{k=1}^n \left\|v^k\right\|_2^2 = \|G_\infty\| \, \|G_1\|. \end{aligned}$$

Thus

$$\sup_{v:\ \|v\|=1} \|Nv\|_2^2 \leq \|G_\infty\| \, \|G_1\| \max_{i,j} \left\|A^{ij}\right\|_2^2,$$

and the conclusion follows.

□

B PROOF OF THEOREM 3.1

In this section we provide the full proof of Theorem 3.1. First, we prove the following result, which will be used to parameterise interactions by the influence bound δ_k.

Lemma B.1. *In a network game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k\in\mathcal{N}})$, the following holds for any agent $k \in \mathcal{N}$ action $i \in \mathcal{S}_k$ and strategies $\mathbf{x}_{-k}, \mathbf{y}_{-k} \in \Delta_{-k}$*

$$|r_{ki}(\mathbf{x}_{-k}) - r_{ki}(\mathbf{y}_{-k})| \leq \delta_k \sum_{l\in\mathcal{N}_k} \|\mathbf{x}_l - \mathbf{y}_l\|_1.$$

Proof. Fix an agent k and define the dummy game $\tilde{\mathcal{G}}_k = (\mathcal{N}_k \cup \{k\}, (\tilde{u}_k, \mathcal{S}_k)_{k\in\mathcal{N}_k\cup\{k\}})$ so that $\tilde{\mathcal{G}}_k$ is composed of only agent k and its neighbours. In addition, the rewards are chosen so that $\tilde{r}_{ki}(\mathbf{x}_{-k}) = r_{ki}(\mathbf{x}_{-k})$ for all $\mathbf{x}_{-k}$ and $\tilde{r}_{li}(\mathbf{x}_{-l}) = \tilde{r}_{lj}(\mathbf{x}_{-l})$ for all $l \in \mathcal{N}_k$ and all $i, j \in \mathcal{S}_k$. In $\tilde{\mathcal{G}}_k$, the maximum influence bound $\delta := \max_{l\in\mathcal{N}_k\cup\{k\}} \delta_l$ is exactly δ_k. Then, from [10] Proposition 5, the following holds in $\tilde{\mathcal{G}}_k$

$$|r_{ki}(\mathbf{x}_{-k}) - r_{ki}(\mathbf{y}_{-k})| \leq \sum_{l\neq k} \delta \|\mathbf{x}_l - \mathbf{y}_l\|_1.$$

This translates in the original network game $\mathcal{G}$ to the statement of the Lemma. □

With these results in place, we can prove Theorem 3.1 in the main paper.

Proof of Theorem 3.1. In order to apply Lemma A.7 we show that, under any of the conditions, the perturbed game $\mathcal{G}^H$ is strongly monotone. To this end, we take the derivative of the pseudo-gradient of $\mathcal{G}^H$ which we call the *pseudo-Hessian* given by

$$[J(\mathbf{x})]_{k,l} = D_{\mathbf{x}_l} F_k(\mathbf{x}).$$

It follows that, if $\frac{J(\mathbf{x})+J^\top(\mathbf{x})}{2}$ is strongly positive definite for all $\mathbf{x} \in \Delta$ with any $\alpha > 0$, i.e. $\mathbf{x}^\top J(\mathbf{x})\mathbf{x} \geq \alpha$ for all $\mathbf{x} \in \Delta$, then $F(\mathbf{x})$ is strongly monotone with the same constant α. We can rewrite the pseudo-Hessian as

$$J(\mathbf{x}) = D(\mathbf{x}) + N(\mathbf{x}),$$

where $D(\mathbf{x})$ is a block diagonal matrix with $-D^2_{\mathbf{x}_k\mathbf{x}_k} u^H_k(\mathbf{x}_k, \mathbf{x}_{-k})$ along the diagonal. $N(\mathbf{x})$ is an off-diagonal block matrix with

$$[N(\mathbf{x})]_{k,l} = \begin{cases} -D_{\mathbf{x}_k,\mathbf{x}_l} u^H_k(\mathbf{x}_k, \mathbf{x}_{-k}) & \text{if } (k,l) \in \mathcal{E} \\ \mathbf{0} & \text{otherwise} \end{cases}.$$

In words, $N(x)$ shares the same structure of the adjacency matrix G of the game, except that it has $-D_{\mathbf{x}_k,\mathbf{x}_l} u^H_k(\mathbf{x}_k, \mathbf{x}_{-k})$ wherever G takes the value 1 and the block matrix $\mathbf{0}$ wherever G has 0. Next we evaluate these partial differentials. Recall that

$$-u^H_k(\mathbf{x}_k, \mathbf{x}_{-k}) = T_k\langle \mathbf{x}_k, \ln \mathbf{x}_k\rangle - \sum_{(k,l)\in\mathcal{E}} \mathbf{x}_k \cdot A^{kl}\mathbf{x}_l.$$

As a result, for all $(k,l) \in \mathcal{E}$, $[N(\mathbf{x})]_{k,l} = -A^{kl}$, so that $N(\mathbf{x})$ represents the network interaction. By contrast, $D(\mathbf{x})$ depends on T_k and is independent of the payoffs u_k. As such, it measures the strength of the game perturbation. Now, let $\bar{J}(\mathbf{x})$ be defined as

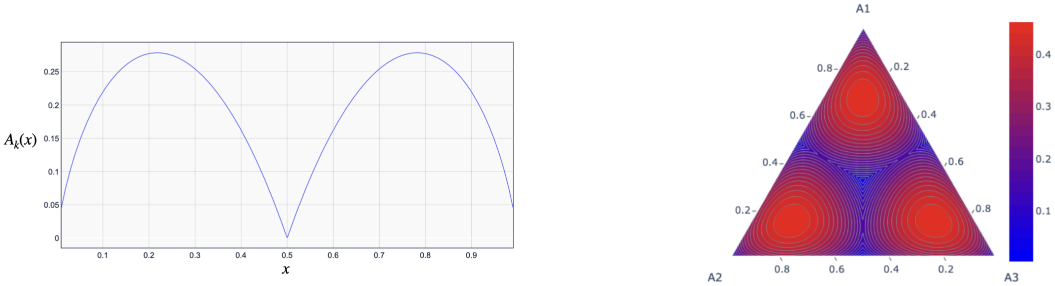
$$\bar{J}(\mathbf{x}) = \frac{J(\mathbf{x}) + J^\top(\mathbf{x})}{2} = D(\mathbf{x}) + \frac{N(\mathbf{x}) + N^\top(\mathbf{x})}{2}.$$

Then, from Proposition A.8 it follows that $D(\mathbf{x})$ is strongly positive definite with constant $T = \min_k T_k$. In particular, this means that $\lambda_{\min} D(\mathbf{x}) \geq T$. Finally, applying Weyl's inequality

$$\begin{aligned}
\lambda_{\min}(\bar{J}) &\geq T + \lambda_{\min}\left(\frac{N + N^\top}{2}\right) \\
&\geq T - \rho\left(\frac{N + N^\top}{2}\right) \\
&= T - \left\|\frac{N + N^\top}{2}\right\|_2 \\
&\geq T - \frac{1}{2}\|A + B\|_2 \sqrt{\|G\|_\infty \|G\|_1} \\
&= T - \frac{1}{2}\left\|A + B^\top\right\|_2 \|G\|_\infty \\
&= T - \frac{1}{2}\sigma_I \|G\|_\infty,
\end{aligned}$$

where we employ Propositions A.11, Lemma A.12 and the fact that G is symmetric so that $\|G\|_\infty = \|G\|_1$. The matrices A, B are chosen so that

$$\left\|A + B^\top\right\|_2 = \max_{(k,l)\in\mathcal{E}} \left\|A^{kl} + (A^{lk})^\top\right\|_2 = \sigma_I.$$



$n_k = 2$ $\qquad$ $n_k = 3$

Figure 9: $A_k(\mathbf{x}_k)$ plotted on a unit simplex Δ_k

Then, under (C2), $\lambda_{\min}(\bar{J}(\mathbf{x})) \geq T - \frac{1}{2}\sigma_I \|G\|_\infty > 0$ and, therefore $F(\mathbf{x})$ is strongly monotone with constant $T - \frac{1}{2}\sigma_I \|G\|_\infty$. Using Lemma A.7, it follows that Q-Learning Dynamics converge to a unique QRE.

To achieve (C3) we apply Proposition A.10 which yields that

$$\begin{aligned} & T - \left\| \frac{N + N^\top}{2} \right\|_2 \\ =& T - \left\| \frac{(A + B^\top) \otimes G}{2} \right\|_2 \\ =& T - \frac{1}{2} \left\| A + B^\top \right\|_2 \|G\|_2 \\ =& T - \frac{1}{2} \sigma_I \|G\|_2 . \end{aligned}$$

Finally, we prove (C1). In this case, it holds that, for any k and any $\mathbf{x}, \mathbf{y} \in \Delta$

$$\begin{aligned} (\mathbf{x}_k - \mathbf{y}_k)^\top (F_k(\mathbf{x}) - F_k(\mathbf{y})) &= (\mathbf{x}_k - \mathbf{y}_k)^\top (T_k \ln \mathbf{x}_k - T_k \ln \mathbf{y}_k) - (\mathbf{x}_k - \mathbf{y}_k)^\top (T_k r_k(\mathbf{x}_{-k}) - T_k r_k(\mathbf{y}_{-k})) \\ &\geq T_k \|\mathbf{x}_k - \mathbf{y}_k\|_1^2 - \left| (\mathbf{x}_k - \mathbf{y}_k)^\top \left(r_k(\mathbf{x}_{-k}) - r_k(\mathbf{y}_{-k}) \right) \right| \\ &\geq T_k \|\mathbf{x}_k - \mathbf{y}_k\|_1^2 - \|\mathbf{x}_k - \mathbf{y}_k\|_1 \delta_k \sum_{l \in \mathcal{N}_k} \|\mathbf{x}_l - \mathbf{y}_l\|_1 \\ &\geq T_k \|\mathbf{x}_k - \mathbf{y}_k\|_1^2 - \|\mathbf{x}_k - \mathbf{y}_k\|_1 \delta_k \sum_{l \neq k} [G]_{kl} \|\mathbf{x}_l - \mathbf{y}_l\|_1 \\ &= \xi_k (M\xi)_k, \end{aligned}$$

where $\xi = (\mathbf{x}_k - \mathbf{y}_k)_{k \in \mathcal{N}}$ and $M = (diag(T_k)_{k \in \mathcal{N}} - diag(\delta_k)_{k \in \mathcal{N}} \cdot G)$. Notice that, to achieve the third inequality, we applied Lemma B.1 Then under (C1), M is strictly diagonally dominant and so is strictly positive definite. Then

$$\sum_k (\mathbf{x}_k - \mathbf{y}_k)^\top (F_k(\mathbf{x}) - F_k(\mathbf{y})) \geq \xi^\top M \xi > 0,$$

so that $F(\mathbf{x})$ is strictly monotone. Then, Lemma A.7 can be applied to yield convergence of Q-Learning Dynamics. □

C PROOFS FROM SECTION 3.1

First we show that any QRE $\bar{\mathbf{x}}$ is an approximate Nash Equilibrium.

Proof of Proposition 3.8. We first notice that, for some $\epsilon > 0$, the definition of an ϵ-Nash Equilibrium in Definition 3.7 holds if

$$\max_{k \in \mathcal{N}} \max_{i \in \mathcal{S}_k} r_{ki}(\bar{\mathbf{x}}_{-k}) - \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle = \epsilon.$$

Next, recall that a QRE $\bar{\mathbf{x}} \in \Delta$ corresponds to an interior fixed point of the Q-Learning Dynamics [32]. From this it holds that, for any k and any $i \in \mathcal{S}_k$

$$0 = r_{ki}(\bar{\mathbf{x}}_{-k}) - \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle + T_k \sum_{j \in \mathcal{S}_k} \bar{x}_{kj} \ln \frac{x_{kj}}{x_{ki}}$$

$$\begin{aligned} r_{ki}(\bar{\mathbf{x}}_{-k}) - \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle &= -T_k \sum_{j \in \mathcal{S}_k} \bar{x}_{kj} \ln \frac{x_{kj}}{x_{ki}} \\ &= T_k \ln x_{ki} - \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle. \end{aligned}$$

As this holds for any $i \in \mathcal{S}_k$, the following holds

$$\begin{aligned} \max_{k \in \mathcal{N}} \max_{i \in \mathcal{S}_k} r_{ki}(\bar{\mathbf{x}}_{-k}) - \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle &= \max_{k \in \mathcal{N}} T_k \max_{i \in \mathcal{S}_k} \ln x_{ki} - \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle \\ &= \max_{k \in \mathcal{N}} T_k A_k(\bar{\mathbf{x}}_k), \end{aligned}$$

so that $\bar{\mathbf{x}}$ is an ϵ-Nash Equilibrium with $\epsilon = \max_{k \in \mathcal{N}} T_k A_k(\bar{\mathbf{x}}_k)$ □

Lemma C.1 (Full version of Lemma 3.11). *Let $\mathbf{u}_k = (1/n_k, \dots, 1/n_k) \in \Delta_k$ and $\mathbf{e}_{ki} \in \Delta_k$ be the canonical basis vector with i^{th} entry equal to* 1 *and* 0 *elsewhere. Then*

$$\bar{A}_k := \max_{\mathbf{x}_k \in \Delta_k} \left(\max_{i \in \mathcal{S}_k} \ln x_{ki} - \sum_{j \in \mathcal{S}_k} x_{kj} \ln x_{kj} \right) = \frac{\ln(n_k - 1) - \ln\left(W\left(\frac{n_k - 1}{e}\right)\right)}{1 + 1/W\left(\frac{n_k-1}{e}\right)}, \tag{11}$$

with equality if $\mathbf{x}^* = c\,\mathbf{e}_{ki} + (1-c)\mathbf{u}_k$ for any $i \in \mathcal{S}_k$, where $c = 1/\left(W\left(\frac{n_k-1}{e}\right) + 1\right)$ and $W(\cdot)$ is the Lambert W function.

Proof. The max over $i \in \{1, \dots, n_k\}$ can be eliminated, as

$$\max_{\mathbf{x}_k \in \Delta_k} \left(\max_{i \in \mathcal{S}_k} \ln x_{ki} - \sum_{j \in \mathcal{S}_k} x_{kj} \ln x_{kj} \right) = \max_{\mathbf{x}_k \in \Delta_k} \left(\ln x_{ki} - \sum_{j \in \mathcal{S}_k} x_{kj} \ln x_{kj} \right). \tag{12}$$

Notice that any value of the LHS realized by $\mathbf{x}_k$ can also be realized by the RHS for $\mathbf{y}_k$, where $\mathbf{y}_k$ is obtained from $\mathbf{x}_k$ by swapping the largest and i^{th} entries of $\mathbf{x}_k$. Explicitly, let $i_0 = \arg\max_{j \in \mathcal{S}_k} x_{kj}$ and

$$y_{kj} = \begin{cases} x_{ki_0} & \text{if } j = i \\ x_{ki} & \text{if } j = i_0 \\ x_{kj} & \text{otherwise} \end{cases}$$

Having eliminated the inner maximization, we can formulate the RHS of (12) as a constrained optimization problem, which we solve in two steps.

Firstly, we show that the $\mathbf{x}^*$ maximizing the RHS of (12) lies on one of the line segments connecting the vertices $\mathbf{e}_{ki}$ of the simplex Δ_k with the uniform distribution $\mathbf{u}_k$ (see Figure 9). Next, we determine the position of $\mathbf{x}^*$ on this segment, and substitute the value of $\mathbf{x}^*$ to obtain the RHS of (11). We then derive tight bounds on this quantity.

For the first step, consider the Lagrangian $L\colon \Delta_k \times (0, -\infty) \to \mathbb{R} \cup \{\infty\}$ given by

$$L(\mathbf{x}_k) = \ln x_{ki} - \sum_{j \in \mathcal{S}_k} x_{kj} \ln x_{kj} + \lambda \left(\sum_{j \in \mathcal{S}_k} x_{kj} - 1 \right).$$

Setting the partial derivatives to 0, we obtain that

$$\begin{aligned} \frac{\partial L}{\partial x_{ki}} &= -1 + \lambda + 1/x_{kj} - \ln x_i = 0, \\ \frac{\partial L}{\partial x_{kj}} &= -1 + \lambda - \ln x_{kj} = 0 \text{ for all } j \in \mathcal{S}_k \text{ with } j \neq i, \\ \frac{\partial L}{\partial \lambda} &= \sum_{j \in \mathcal{S}_k} x_{kj} - 1 = 0, \end{aligned}$$

with the solution given by $\mathbf{x}^*$ with $x_{kj} = 1/W(e^{1-\lambda})$ and $x_{ki} = e^{-1+\lambda}$ for $j \neq i$, where W is the Lambert W function. Note that $L(\mathbf{x}_k) = -\infty$ for $\mathbf{x}_k$ on the boundary of Δ_k and further that the mapping $\mathbf{x}_k \to \ln x_{ki} - \sum_{j \in \mathcal{S}_k} x_{kj} \ln x_{kj}$ is concave, as its Hessian is diagonal with negative entries, hence the stationary point of L gives a maximum. Finally, determining λ is equivalent to solving

$$\sum_{j\in\mathcal{S}_k} x_{kj} = \frac{1}{W(e^{1-\lambda})} + (n_k-1)e^{-1+\lambda} = 1,$$

which is intractable to the best of our knowledge, even with modern software such as Mathematica. Nevertheless, we have proved that the maximizer $\mathbf{x}^*$ of (12) lies on the line segment connecting a vertex and the centre of the simplex, which reduces our initial constrained optimization problem to one dimension. Without loss of generality, pick the first vertex and let $\mathbf{x}^* = (x, \frac{1-x}{n_k-1}, \dots, \frac{1-x}{n_k-1})$ for some $x \in [0,1]$. We have that

$$\ln x_i - \sum_{j\in\mathcal{S}_k} x_{kj}\ln x_{kj} = \ln x + x\ln x + (1-x)\ln\frac{1-x}{n_k-1},$$

which involves no special functions. By setting the derivative to 0, we find that this expression is maximized for $x = 1/\left(1+W\left(\frac{n_k-1}{e}\right)\right)$ and that the maximum value is given by

$$\bar{A}_k = \frac{\ln(n_k-1) - \ln\left(W\left(\frac{n_k-1}{e}\right)\right)}{1+1/W\left(\frac{n_k-1}{e}\right)}.$$

Finally, we give bounds for $\bar{A}_k$. [24] proves the sharp bound

$$\ln x - \ln\ln x < W(x) < \ln x - \ln\ln x + \ln(1+1/e),$$

which translates to

$$\bar{A}_k < \frac{\ln n_k - \ln(\ln n_k - \ln\ln n_k)}{1+\frac{1}{\ln n_k - \ln\ln n_k}} < \ln n_k - \ln(\ln n_k - \ln\ln n_k) < \ln n_k.$$

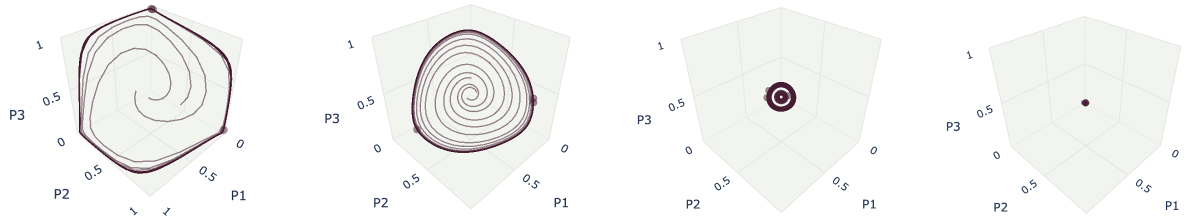
□

Algorithm 1 Iterative improvement of QRE

Input: Network game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{S}_k)_{k\in\mathcal{N}})$; Exploration Rate annealing step ΔT; Maximum number of anneals M; Q-Learning horizon H; Convergence Window Length h; Tolerance `tol`.

Output: Learned QRE $\bar{\mathbf{x}} \in \Delta$

```
$T_k \leftarrow \delta_k|\mathcal{N}_k|$ for all $k \in \mathcal{N}$                                  ▷ or (C2), (C3)
for $\tau = 1 : H$ do
    for k = 1, …, N do
        $Q_{ki} \leftarrow (1-\alpha_k)Q_{ki} + \alpha_k r_{ki}(\mathbf{x}_{-k})$
        $\mathbf{x}_k(\tau) \leftarrow \mathtt{softmax}(Q_k/T_k)$
    end for
end for
$\bar{\mathbf{x}} \leftarrow \mathbf{x}(H)$
for t = 1:M do                                                  ▷ or until break statement is reached
    for $k = 1, \dots, N$ do
        $\epsilon_k \leftarrow T_k A_k(\bar{\mathbf{x}}_k)$                                         ▷ from (5)
    end for
    $l = \arg\max_{k\in\mathcal{N}} \epsilon_k$                                       ▷ ties broken arbitrarily
    $T_l \leftarrow T_l - \Delta T$
    for $\tau = 1 : H$ do
        for k = 1, …, N do
            $Q_{ki} \leftarrow (1-\alpha_k)Q_{ki} + \alpha_k r_{ki}(\mathbf{x}_{-k})$
            $\mathbf{x}_k(\tau) \leftarrow \mathtt{softmax}(Q_k/T_k)$
        end for
    end for
    $V \leftarrow \max_{k,i}\left\{\frac{\max_{\tau\in H} x_{ki}(\tau) - \min_{\tau\in H} x_{ki}(\tau)}{\min_{\tau\in H} x_{ki}(\tau)}\right\}$
    if $V <$ tol then
        $\bar{\mathbf{x}} \leftarrow \mathbf{x}(H)$
    else
        break
    end if
end for
```



$T = 0.15$ $T = 0.3$ $T = 0.4$ $T = 0.55$

Figure 10: Trajectories of Q-Learning in a three agent Network Mismatching Game with $M = 2$. Axes denote the probabilities with which each player chooses their first action.

D ADDITIONAL EXPERIMENTS

In this section, we present additional experiments on the behaviour of Q-Learning in Network Games, as well as on the exploration update scheme. In Figure 10, we examine a Network Mismatching Game, which was analysed in [30] as an example of limit cycle behaviour in replicator dynamics. Here, the payoff to each agent k is given as

$$u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \mathbf{x}_k^\top \mathbf{A} \mathbf{x}_l, \; l = k - 1 \mod N,$$
$$A = \begin{pmatrix} 0 & 1 \\ M & 0 \end{pmatrix}, \; M \geq 1$$

From Figure 10 it is clear that, as exploration rates increase, the dynamics are driven towards a QRE from all initial conditions.

Next, we present additional experiments on the exploration updating scheme in Section 3.2. In particular, we apply the scheme to a Network Mismatching Game with 5 agents. We plot the exploitability (7) and ϵ (4) over $150,000$ iterations of learning. In both cases it is again clear that the distance to Nash decreases as the exploration updating scheme is applied. In the case that $M = 2$, the scheme is applied until (6) fails at approx. $60,000$ iterations, whilst in the case $M = 4$, agents learn for $80,000$ iterations before the dynamics are considered unstable. In Figure 12 we plot the trajectories of Q-Learning using the first action played by three representative agents. The dynamics move between QRE as the exploration rates are adjusted, however stability of the dynamic is maintained.

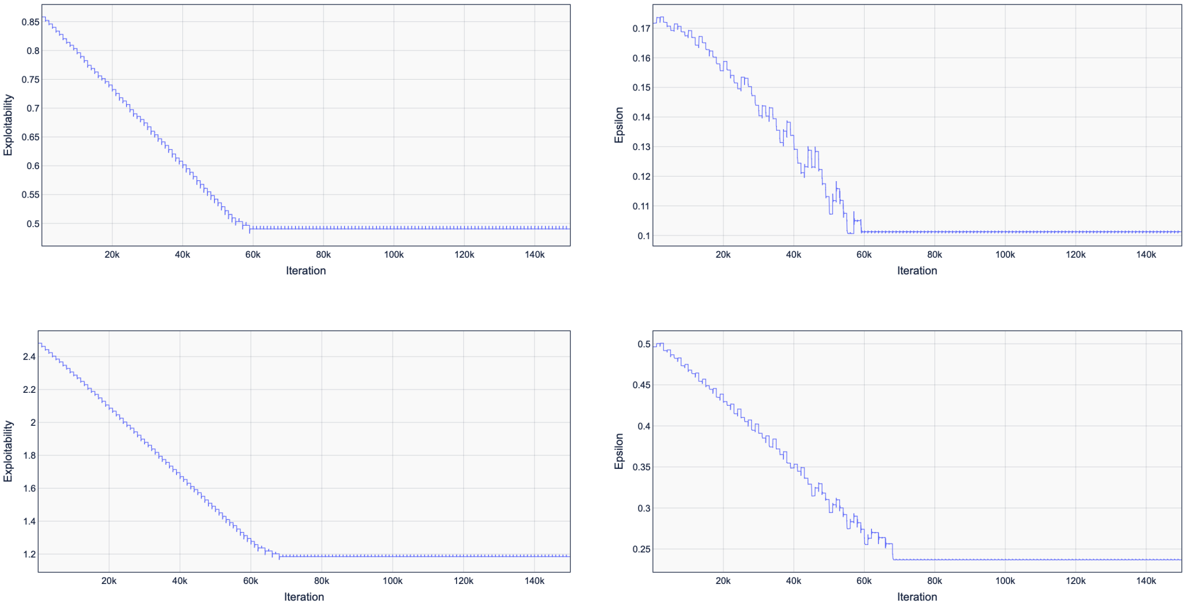


Figure 11: Measures of 'closeness' to Nash Equilibrium as the exploration update scheme is applied to the Network Mismatching Game with five agents and (Top) $M = 2$ (Bottom) $M = 4$. (Left) Distance to NE measured by exploitability (7) of the joint strategy $\mathbf{x}(t)$. (Right) ϵ as defined by (4). Both metrics decreases as exploration rates are updated until condition (6) fails, after which learning is halted.

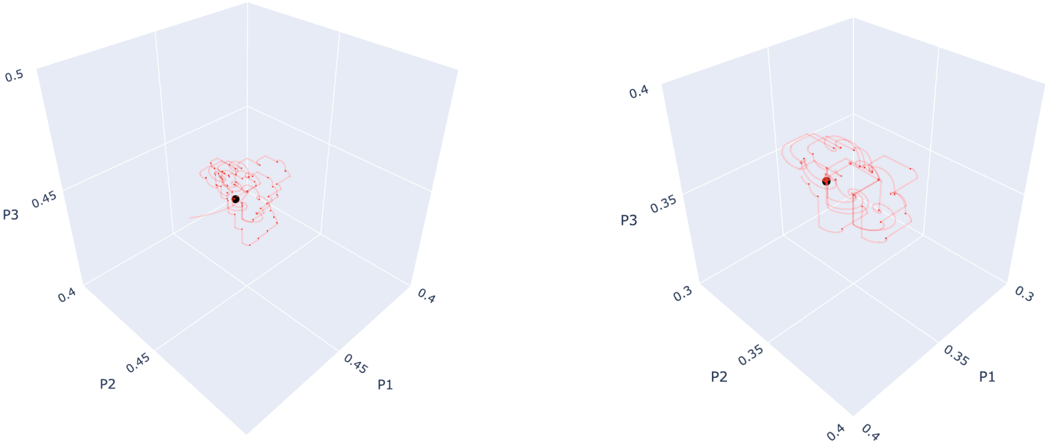


Figure 12: Trajectories of Q-Learning generated as the centralised scheme is applied to (Left) Mismatching Game with $M = 2$ (Right) Mismatching Game with $M = 4$.