Stability of Multi-Agent Learning: Convergence in Network Games with Many Players

Aamal Hussain [1 *] **Dan Leonte** [1 *] **Francesco Belardinelli** [1] **Georgios Piliouras** [2 3]

Abstract

The behaviour of multi-agent learning in many player games has been shown to display complex dynamics outside of restrictive examples such as network zero-sum games. In addition, it has been shown that convergent behaviour is less likely to occur as the number of players increase. To make progress in resolving this problem, we study Q-Learning dynamics and determine a sufficient condition for the dynamics to converge to a unique equilibrium in any network game. We find that this condition depends on the nature of pairwise interactions and on the network structure, but is explicitly independent of the total number of agents in the game. We evaluate this result on a number of representative network games and show that, under suitable network conditions, stable learning dynamics can be achieved with an arbitrary number of agents.

1. Introduction

Determining the convergence of multi-agent learning is arguably amongst the most studied problems in game theory and online learning (Anagnostides et al., 2022; Bai et al., 2021; Ewerhart & Valkanova, 2020). In this setting, agents are required to explore their state space to determine optimal actions, whilst simultaneously aiming to maximise their expected reward. To do this, each agent must react to the changing behaviour of the other agents so that, from the perspective of any given agent, the environment is non stationary. A large body of recent work has shown that this non stationarity leads to complex behaviours being displayed by learning dynamics in games with many agents (Sato et al., 2002; Andrade et al., 2021). In addition, even in games where convergent behaviour can be achieved, it can be to one of multiple equilibria (Czechowski & Piliouras, 2022; Villatoro et al., 2009; Sanders et al., 2018). In fact, recent work suggests that as the number of agents in the system increase the likelihood for chaotic dynamics increases. This presents a significant challenge for making predictions in multi-agent settings.

Nevertheless, multi-agent learning has been the major driver of a number of successes in Artificial Intelligence and Machine Learning. Such examples include strong performance in competitive games (Brown & Sandholm, 2019; Perolat et al., 2022), resource allocation (Parise et al., 2020; Amelina et al., 2015) and robotics (Hamann, 2018; Hernández et al., 2013). Due to these successes, and the increasing use of multi-agent systems, it becomes important to develop a theoretical understanding of learning algorithms in various settings.

To this end, a number of advances have been made which consider learning in multi-agent settings. Indeed, strong positive results on convergence have been found in cooperative settings, such as *potential games* (Candogan et al., 2013; Harris, 1998) and competitive settings, such as zero-sum network games (Leonardos et al., 2021; Kadan & Fu, 2021; Cai et al., 2016). However, a general framework for understanding learning behaviour must extend beyond these settings. In (Hussain et al., 2023) a strong positive result on convergence was found which states that an equilibrium can be reached in any game, given sufficient exploration by all agents. Unfortunately, the condition requires that the amount of exploration increases with the number of agents. To make matters worse, (Sanders et al., 2018) show that, as the number of agents increase in the game, learning dynamics are more likely to display chaotic behaviour.

However, both of these works did not impose any structure on the interaction between agents. Rather, (Sanders et al., 2018) assume that all agents interact with all others. In reality, agents are more likely to interact according to an underlying communication network. In economic settings this may correspond to social interactions between agents, whilst in machine learning settings, networks are often used to enforce an underlying structure to the model.

[*] Equal contribution [1] Imperial College London [2] Singapore University of Technology and Design [3] DeepMind. Correspondence to: Aamal Hussain <aamal.hussain15@imperial.ac.uk>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023. Copyright 2023 by the author(s).

Model and Contributions In light of this, we study learning in *network games*, where interactions between agents can be constrained. On this model, we study the *Q-Learning* dynamic (Sato & Crutchfield, 2003; Tuyls et al., 2006), a well studied learning dynamic captures the balance between agents who explore their state space whilst maximising their reward.

Our main result tightens the requirement of sufficient exploration found by (Hussain et al., 2023) to achieve convergence to a unique equilibrium in any network game. In particular we find that the amount of exploration depends on the nature of the interaction between agents and, more importantly, the structure of the network. We examine how our bound explicitly depends on the total number of agents in the system and find that, for certain networks, there is no explicit dependence. This enables a higher number of agents to be introduced in the system without compromising stability. In addition, our result applies to all network games, and not only network zero sum games. In fact, we show how our results relate to existing statements in the literature. Finally, we validate our findings on a number of representative classes of games and networks.

Related Work The theory of evolutionary game dynamics models multi-agent interactions in which agents improve their actions through *online learning* (Shalev-Shwartz, 2011). The premise is that popular learning algorithms such as *Hedge* (Krichene et al., 2015), online gradient descent (Kadan & Fu, 2021) and Q-Learning (Sutton & Barto, 2018; Schwartz, 2014) can be approximated in continuous time by a dynamical system (Mertikopoulos & Sandholm, 2016; Krichene, 2016; Tuyls et al., 2006). This enables tools from the study of dynamical systems to be used to analyse the behaviour of the learning algorithm. This approach has yielded a number of successes, most notably in *potential games* (Leonardos & Piliouras, 2022; Candogan et al., 2013; Monderer & Shapley, 1996), which model multi-agent cooperation, and *network zero sum games* (Cai et al., 2016; Abernethy et al., 2021), which models competition. In these settings, it is known that a number of learning dynamics converge to an equilibrium (Kadan & Fu, 2021; Ewerhart & Valkanova, 2020; Leonardos et al., 2021).

Outside of these classes, the behaviour of learning is less certain (Anagnostides et al., 2022). In particular, it is known that learning dynamics can exhibit complex behaviours such as cycles (Mertikopoulos et al., 2018; Imhof et al., 2005; Pangallo et al., 2019; Shapley, 2016) and chaos (van Strien & Sparrow, 2011; Mukhopadhyay & Chakraborty, 2020; Sato et al., 2002; Pangallo et al., 2022). Indeed, (Galla & Farmer, 2013) showed that the Experience Weighted Attraction (EWA) dynamic, which is closely related to Q-Learning (Leonardos et al., 2021) achieves chaos in classes of two-player games. Advancing this result, (Sanders et al., 2018) showed that chaotic dynamics become more prevalent as the number of agents increase, regardless of the exploration rates. Similar to the work in this paper, (Hussain et al., 2023) determine a sufficient condition on the exploration rates for Q-Learning to converge in any game, yet they also find that this condition increases with the number of agents. This presents a strong barrier in placing guarantees on the behaviour in multi-agent systems with many agents, outside of restrictive settings.

Our work also employs a number of tools from the study of variational inequalities in game theory. This is a well studied framework for analysing the structure of equilibrium sets in a game (Melo, 2018; Facchinei & Pang, 2004) and for studying the convergence of algorithms equilibrium seeking algorithms (Tatarenko & Kamgarpour, 2019; Hadikhanloo et al., 2022; Mertikopoulos & Zhou, 2019; Sorin & Wan, 2016). Recent advances in this field begin to consider the properties of network games. Notably, (Parise & Ozdaglar, 2019; Melo, 2018) determine conditions under which the Nash Equilibrium of a network game is unique, and how these relate to properties of the network. Similarly, (Melo, 2021) shows the uniqueness of various formulations of the Quantal Response Equilibrium (QRE) under particular choices of payoff functions. Whilst our results use similar techniques, we do not make such assumptions on the nature of the payoffs, but rather parameterise our final condition on the nature of interactions between agents. In addition, we consider the stability of learning.

In our work, we aim to address the problem of convergence in many-agent systems by considering games which are played on a network (Cai et al., 2016). Extending the work of (Hussain et al., 2023), we are able to find a sufficient condition on exploration rates so that the Q-Learning dynamics converge to a unique equilibrium. Importantly, we show that this is independent of the total number of agents in the system. To our knowledge this is the first work which shows the convergence of Q-Learning in arbitrary network games.

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* (Cai et al., 2016). A Network Game is described by the tuple $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$, where $\mathcal{N}$ denotes a finite set of players $\mathcal{N}$ indexed by $k = 1, \ldots, N$. Each agent can choose from a finite set of actions $\mathcal{A}_k$ 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 $\Delta(\mathcal{A}_k) := \{\mathbf{x}_k \in \mathbb{R}^n : \sum_i x_{ki} = 1, x_{ki} \geq 0\}$. Each agent is also given a payoff function $u_k : \Delta(\mathcal{A}_k) \times \Delta(\mathcal{A}_{-k}) \to \mathbb{R}$ where $\mathcal{A}_{-k}$ denotes the action set of all agents other than k. 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. An equivalent way to define the network is through an *adjacency matrix* G so 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 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 $\mathbf{x} \in \Delta =: \times_k \Delta(\mathcal{A}_k)$, we can define the reward to agent k for playing action i as $r_{ki}(\mathbf{x}_{-k}) = \frac{\partial u_{ki}(\mathbf{x})}{\partial x_{ki}}$. Under this notation, $u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \langle \mathbf{x}_k, r_k(\mathbf{x})\rangle$. With this in place, we can define an equilibrium solution for the game.

Definition 2.1 (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 $i \in \mathcal{A}_k$

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

The QRE (Camerer et al., 2004) is the prototypical extension of the Nash Equilibrium to the case of agents with bounded rationality, parameterised by the *exploration rate* T_k. In particular, the limit $T_k \to 0$ corresponds exactly to the Nash Equilibrium, whereas the limit $T_k \to \infty$ corresponds to a purely irrational case, where action $i \in \mathcal{A}_k$ is played with the same probability regardless of its associated reward. The link between the QRE and the Nash Equilibrium is made stronger through the following result.

Proposition 2.2 ((Melo, 2021))**.** *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k\in\mathcal{N}})$ and let $T_1, \ldots, T_N > 0$ be exploration rates. Define the perturbed game $\mathcal{G}^H = (\mathcal{N}, \mathcal{E}, (u_k^H, \mathcal{A}_k)_{k\in\mathcal{N}})$ 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 $\mathcal{G}$ if and only if it is a Nash Equilibrium of $\mathcal{G}^H$.

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. 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 (Tuyls et al., 2006; Sato & Crutchfield, 2003) 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 (Leonardos et al., 2021).

2.3. Variational Inequalities and Game Theory

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 (Melo, 2018; Parise & Ozdaglar, 2019; Xu et al., 2019) as well as learning in games (Hadikhanloo et al., 2022; Sorin & Wan, 2016; Hussain et al., 2023). In this paper, we connect these areas of literature.

Definition 2.3 (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

$$\langle \mathbf{x} - \bar{\mathbf{x}}, F(\bar{\mathbf{x}})\rangle \geq 0, \qquad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{2}$$

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 (2).

The premise of the variational approach to game theory (Facchinei & Pang, 2004; Rosen, 1965) 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 2.4 (Pseudo-Gradient Map)**.** The pseudo-gradient map of a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$ 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 (Hussain et al., 2023; Kadan & Fu, 2021).

Lemma 2.5 ((Melo, 2021))**.** *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$ and for any $T_1, \ldots, T_N > 0$, 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 2.6. 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 2.7 (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 2.8 ((Melo, 2021))**.** *Consider a game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$ and for any $T_1, \ldots, T_N > 0$, 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 2.9 ((Hussain et al., 2023))**.** *If the game G is* monotone*, then the Q-Learning Dynamics (QLD) converge to the unique QRE with any positive exploration rates $T_1, \ldots, T_N > 0$.*

3. Convergence of Q-Learning in Network Games

In this section we determine a sufficient condition under which Q-Learning converges to a unique QRE, which is given in terms of the exploration rate and the network game structure. To do this, we determine a sufficient condition on exploration rates T_k such that the perturbed game $\mathcal{G}^H$ is strongly monotone. We find that this condition is dependent on the strength of pairwise interactions in the network, as well as its structure. We then compare our result to that of (Hussain et al., 2023) 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 (Sanders et al., 2018) which suggests that learning dynamics are increasingly unstable as the number of players increases, regardless of exploration rate.

To achieve our main result, we first parameterise pairwise interactions in a network game as follows.

Definition 3.1 (Interaction Coefficient)**.** Let $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$ be a network game whose edgeset is associated with the payoff functions $(A^{kl}, A^{lk})_{(k,l) \in \mathcal{E}}$. Then, the *interaction coefficient* δ_S of $\mathcal{G}$ is given as

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

where $\|M\|_2 = \sup_{||\mathbf{x}||_2 = 1} \|M\mathbf{x}\|_2$ denotes the operator 2-norm (Meiss, 2007).

Theorem 3.2. *Consider a network game $\mathcal{G} = (\mathcal{N}, \mathcal{E}, (u_k, \mathcal{A}_k)_{k \in \mathcal{N}})$ which has interaction coefficient δ_S and adjacency matrix G. The Q-Learning Dynamic converges to a unique QRE $\bar{\mathbf{x}} \in \Delta$ if, for all agents $k \in \mathcal{N}$,*

$$T_k > \frac{1}{2} \delta_S \left\| G \right\|_\infty, \tag{4}$$

where $\|M\|_\infty = \max_i \sum_j |[G_{ij}]|$ is the operator ∞-norm.

We defer the full proof of Theorem 3.2 to the Appendix and illustrate the main ideas here. In order to apply Lemma 2.9, we must show that under (4), the perturbed game $\mathcal{G}^H$ is monotone. To do this, we decompose $\mathcal{G}^H$ into a term which is solely parameterised by exploration rates, and another term corresponding to the payoff matrices and graph structure. We then show that the second term can be decomposed as $\frac{1}{2} \delta_S \left\| G \right\|_\infty$, which allows us to separate terms involving the payoffs and the graph structure. We use the fact that the transformation between a $\mathcal{G}$ and $\mathcal{G}^H$ is given by $T_k \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle$, which has a strongly monotone gradient (Melo, 2021) with constant T_k. Then, if the exploration rates are high enough to offset $\frac{1}{2} \delta_S \left\| G \right\|_\infty$, the resulting pseudo-gradient is monotone, and Lemma 2.9 can be applied.

The condition of Theorem 3.2 for the convergence asserts that Q-Learning dynamics is convergent in a network game given sufficient exploration. In a similar light to the result of (Hussain et al., 2023), the amount of exploration required depends on the strength of interaction. The main difference is that the condition includes a term $\left\| G \right\|_\infty$ which encodes the network structure. This term has a natural interpretation as follows. Let $\mathcal{N}_k = \{l \in \mathcal{N} : (k, l) \in \mathcal{E}\}$ be the *neighbours* of agent k, i.e. all the agents who interact with agent

k according to the network. Then $\|G\|_\infty = \max_k |\mathcal{N}_k|$, which denotes the maximum number of neighbours across all agents.

A useful point about (4) is that it does not make any assumptions regarding the nature of the interaction between games, but rather parameterises pairwise interactions by δ_S. As such, the result is not limited to restrictive settings such as *network zero sum games* (Leonardos et al., 2021). In fact, the convergence of Q-Learning dynamics in pairwise zero-sum games follows immediately from Theorem 3.2.

Corollary 3.3. *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.*

Remark 3.4. Corollary 1 is supported by the result of (Leonardos et al., 2021; Hussain et al., 2023) 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.

Discussion The main takeaway from Theorem 3.2 is that the condition on sufficient exploration depends on $\|G\|_\infty$, which is a measure of the network structure. In certain networks, such as the ring network depicted in Figure 1a, $\|G\|_\infty$ is independent of the number of agents in the system. Therefore, as the number of agents increase, the bound (4) does not increase. By contrast, in the fully connected network all agents are connected to each other and so $\|G\|_\infty$ increases with the number of agents. This illustrates the main point that the variation in the stability boundary defined by (4) depends on the structure of the network rather than solely on the total number of agents as previously found by (Hussain et al., 2023; Sanders et al., 2018). We illustrate this further in Figure 2 which plots the stability boundary defined in (Hussain et al., 2023) in various games (which we define in Section 4) as well as (4) for the ring and fully-connected network. Here, it is clear that (4) is a tighter bound than that of (Hussain et al., 2023) particularly for the ring network in all games. The advantage of using (4) is most clear in the example of the Sato game which, in (Sato et al., 2002) was shown to display chaotic behaviour in the two-agent case when exploration rates are uniformly zero. In Figure 2 it can be seen that only a small amount of exploration is required to stabilise the system.

4. Experiments

In our experiments, we visualise and exemplify the implications of Theorem 3.2 on a number of games. In particular, we simulate the Q-Learning algorithm described in Section 2.2 and show that 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 the bound (4) must hold for all agents k, this assumption does not affect the generality of the results.

Convergence of Q-Learning. We first illustrate the convergence of Q-Learning using two representative examples: the *Network Chakraborty Game* and the *Mismatching Pennies Game*. The former was first analysed in (Pandit et al., 2018) to characterise chaos in learning dynamics. Formally, the payoff to each agent k is defined as

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

The latter was first analysed in (Kleinberg et al., 2011) in which it was shown that learning dynamics reach a cycle around the boundary of the simplex. Here, the payoffs to each agent are given by

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

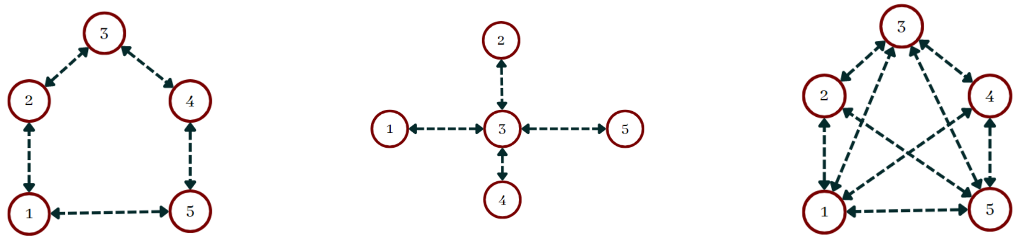
We visualise the trajectories generated by running Q-Learning in Figure 3 in both games for a three agent network and choosing $\alpha = 7, \beta = 8.5, M = 2$. It can be seen that, for low exploration rates, the dynamics reach a limit cycle around the boundary of the simplex. However, as exploration increases, the dynamics are eventually driven towards a fixed point for all initial conditions. The higher requirement on exploration in the Chakraborty Game as compared to the Mismatching Game can be seen as stemming from the higher $\delta_S \approx 8.67$ in the former compared to $\delta_S = 2$ in the latter.

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 (Shapley, 2016), where it was shown that the *Fictitious Play* learning dynamic do not converge to an equilibrium. (Hussain et al., 2023) 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



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 $\|G\|_2$.

exploration. In Figure 4 we examine both a fully connected network and a ring network with 15 agents. Figure 4 depicts the final 2500 iterations of learning for three agents and 35 initial conditions. It can be seen that, as exploration rates increase Q-Learning is driven towards an equilibrium for all initial conditions. 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 $\|G\|_\infty = 2$.

Network Sato Game We also analyse the behaviour of Q-Learning in a variant of the game introduced in (Sato et al., 2002), 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 (Sato et al., 2002) 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 $\delta_S = 0.05$ as compared to the Shapley game $\delta_S = 2$.

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 record the final 2500 iterations and check whether the relative difference between the maximum and minimum strategy components x_{ki} is less than some tolerance l for all agents k, actions i and initial conditions. More formally we aim to determine if

$$\lim_{t \to \infty} \left(\frac{\max_t x_{ki}(t) - \min_t x_{ki}(t)}{\max_t x_{ki}(t)} \right) < l \qquad (5)$$

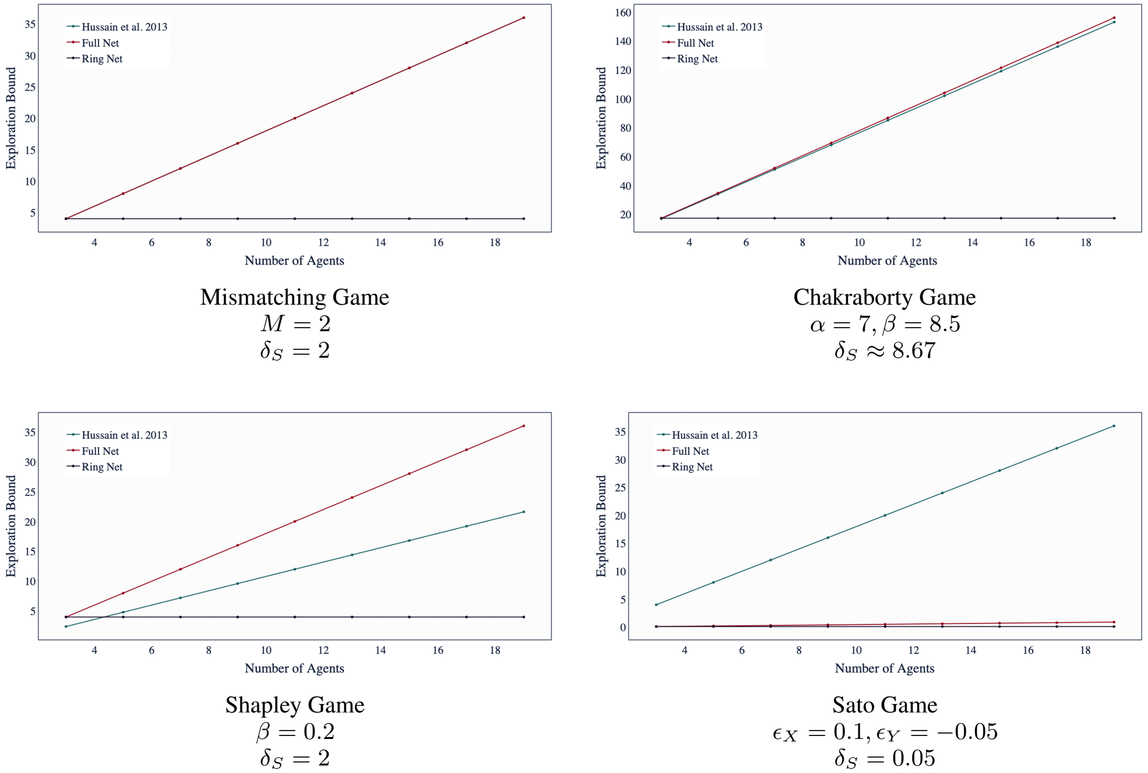
holds for all $k \in \mathcal{N}$ and all $i \in \mathcal{A}_k$. In Figure 5 we plot the smallest exploration rate T for which (5) holds for varying choices of N, using $l = 1 \times 10^{-5}$. It can be seen that the prediction of (4) holds, in that the number of agents plays no impact for the ring network 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). We therefore conjecture that a tighter lower bound on exploration can be obtained using the 2-norm, which we consider an important avenue for future work.

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



Mismatching Game
$M = 2$
$\delta_S = 2$

Chakraborty Game
$\alpha = 7, \beta = 8.5$
$\delta_S \approx 8.67$

Shapley Game
$\beta = 0.2$
$\delta_S = 2$

Sato Game
$\epsilon_X = 0.1, \epsilon_Y = -0.05$
$\delta_S = 0.05$

Figure 2. Lower Bound on sufficient exploration as defined by (Hussain et al., 2023) and by (4) in a fully connected network and ring network. The star network is not plotted as it has the same ∞-norm as the fully connected network. These are repeated for various games who are defined in Section 4.

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.

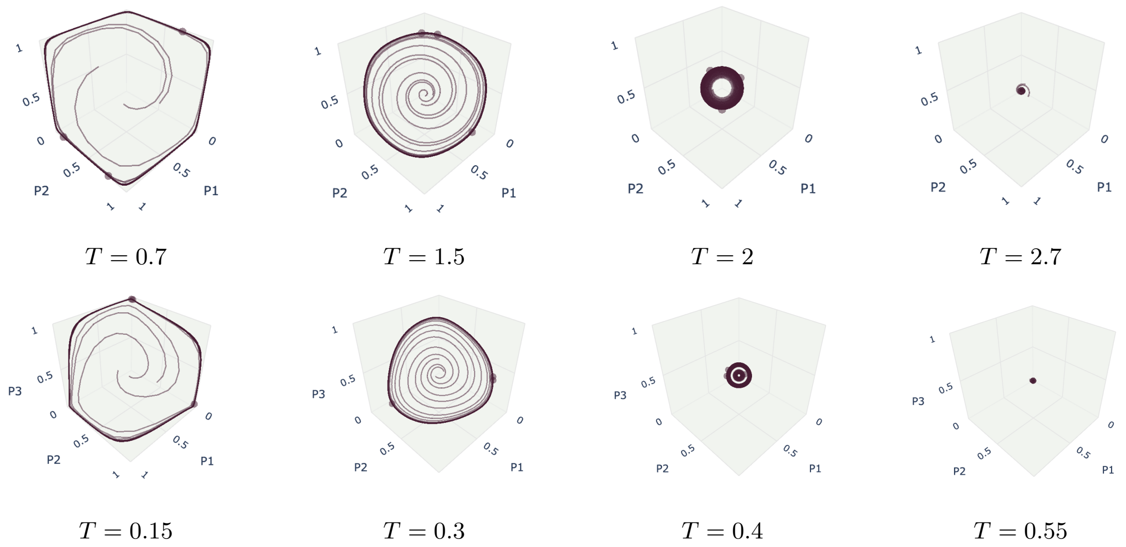
Acknowledgements

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/project is 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), NRF 2018 Fellowship NRF-NRFF2018-07, NRF2019-NRF-ANR095 ALIAS grant, grant PIESGP-AI-2020-01, AME Programmatic Fund (Grant No.A20H6b0151) from the Agency for Science, Technology and Research (A*STAR) and Provost's Chair Professorship grant RGEPPV2101.

References

Abernethy, J., Lai, K. A., and Wibisono, A. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In Feldman, V., Ligett, K., and Sabato, S. (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 3–47. PMLR, 16–19 Mar 2021. URL `https://proceedings.mlr.press/v132/abernethy21a.html`.

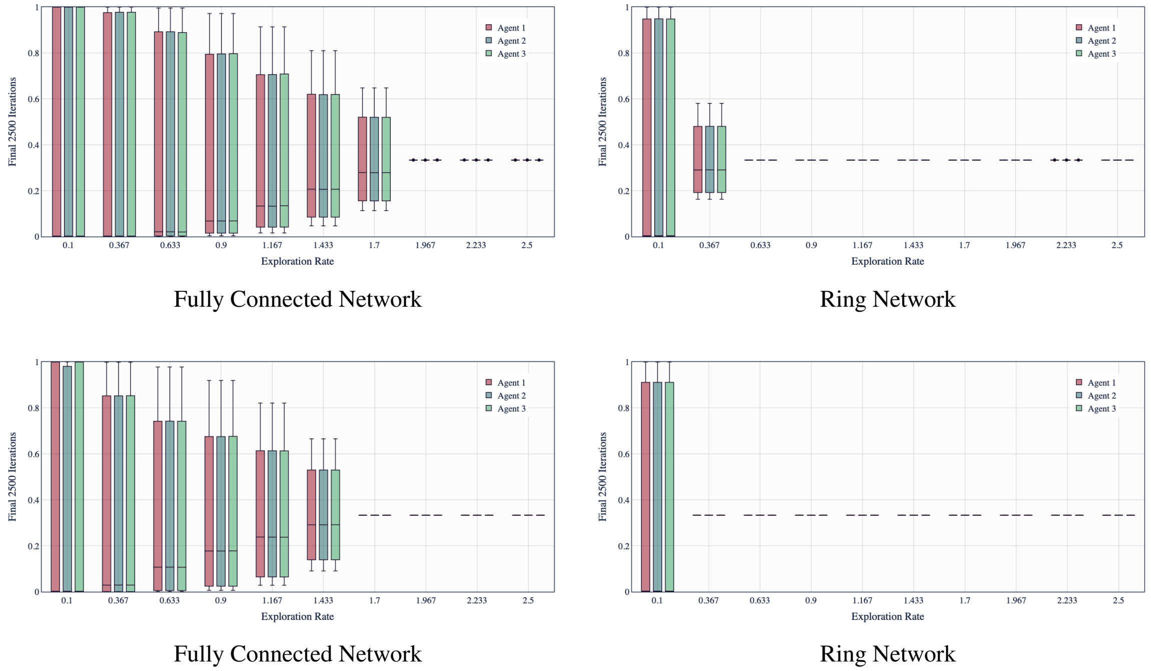
Amelina, N., Fradkov, A., Jiang, Y., and Vergados, D. J. Approximate Consensus in Stochastic Networks with Application to Load Balancing. *IEEE Transactions on Information Theory*, 61(4):1739–1752, 4 2015. ISSN 00189448. doi: 10.1109/TIT.2015.2406323.



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

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

Figure 3. Trajectories of Q-Learning in a three agent (Top) Network Chakraborty Game with $\alpha = 7, \beta = 8.5$ (Bottom) Mismatching Game with $M = 2$. Axes denote the probabilities with which each player chooses their first action.



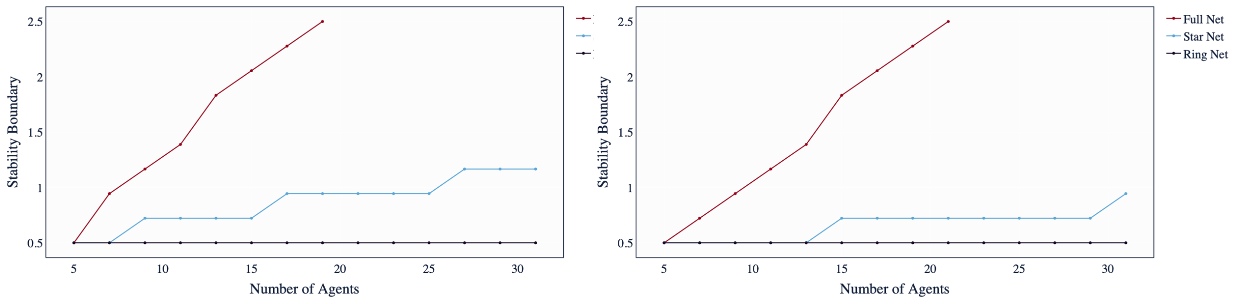
Fully Connected Network Ring Network

Fully Connected Network Ring Network

Figure 4. 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

Anagnostides, I., Panageas, I., Farina, G., and Sandholm, T. On Last-Iterate Convergence Beyond Zero-Sum Games. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 536–581. PMLR, 8 2022. URL https://proceedings.mlr.press/v162/anagnostides22a.html.

Andrade, G. P., Frongillo, R., Belkin, M., and



Shapley Game

Sato Game

Figure 5. 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 (5) holds with $l = 1 \times 10^{-5}$. The Fully Connected Network, Star Network and Ring Networks are considered.

Kpotufe, S. Learning in Matrix Games can be Arbitrarily Complex, 7 2021. ISSN 2640-3498. URL https://proceedings.mlr.press/v134/andrade21a.html.

Bai, C., Wang, L., Han, L., Hao, J., Garg, A., Liu, P., and Wang, Z. Principled exploration via optimistic bootstrapping and backward induction. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 577–587. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/bai21d.html.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. doi: 10.1126/science.aay2400. URL https://www.science.org/doi/abs/10.1126/science.aay2400.

Cai, Y., Candogan, O., Daskalakis, C., and Papadimitriou, C. Zero-sum polymatrix games: a generalization of minmax. *Mathematics of Operations Research*, 41(2):648–656, 5 2016. ISSN 0364765X. URL https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=0364765X&v=2.1&it=r&id=GALE%7CA451940436&sid=googleScholar&linkaccess=fulltexthttps://go.gale.com/ps/i.do?p=AONE&sw=w&issn=0364765X&v=2.1&it=r&id=GALE%7CA451940436&sid=googleScholar&linkaccess=abs.

Camerer, C. F., Ho, T. H., and Chong, J. K. Behavioural game theory: Thinking, learning and teaching. *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality*, pp. 120–180, 1 2004. doi: 10.1057/9780230523371{_}8/COVER. URL https://link.springer.com/chapter/10.1057/9780230523371_8.

Candogan, O., Ozdaglar, A., and Parrilo, P. A. Dynamics in near-potential games. *Games and Economic Behavior*, 82:66–90, 11 2013. ISSN 0899-8256. doi: 10.1016/J.GEB.2013.07.001.

Czechowski, A. and Piliouras, G. 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*, 2022. URL www.ifaamas.org.

Ewerhart, C. and Valkanova, K. Fictitious play in networks. *Games and Economic Behavior*, 123:182–206, 9 2020. ISSN 10902473. doi: 10.1016/j.geb.2020.06.006.

Facchinei, F. and Pang, J. S. Finite-Dimensional Variational Inequalities and Complementarity Problems. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, 2004. doi: 10.1007/B97543.

Galla, T. and Farmer, J. D. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences of the United States of America*, 110(4):1232–1236, 2013. ISSN 00278424. doi: 10.1073/pnas.1109672110.

Hadikhanloo, S., Laraki, R., Mertikopoulos, P., and Sorin, S. Learning in nonatomic games part I Finite action spaces and population games. *Journal of Dynamics and Games. 2022*, 0(0):0, 2022. ISSN 2164-6066. doi: 10.3934/JDG.2022018.

Hamann, H. *Swarm Robotics: A Formal Approach*. Springer International Publishing, 2018. doi: 10.1007/978-3-319-74528-2.

Harris, C. On the Rate of Convergence of Continuous-Time Fictitious Play. *Games and Economic Behavior*, 22(2): 238–259, 2 1998. ISSN 08998256. doi: 10.1006/game.1997.0582.

Hernández, E., Barrientos, A., Del Cerro, J., and Rossi, C. A multi-robot system for patrolling task via Stochastic Fictitious Play. In *ICAART 2013 - Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, volume 1, pp. 407–410, 2013. ISBN 9789898565389. doi: 10.5220/0004259504070410.

Hussain, A. A., Belardinelli, F., and Piliouras, G. 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, pp. 1578–1586, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Imhof, L. A., Fudenberg, D., and Nowak, M. A. Evolutionary cycles of cooperation and defection. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, pp. 10797–10800. 8 2005. doi: 10.1073/pnas.0502589102.

Kadan, A. and Fu, H. 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:19–34, 2021. ISSN 16113349. doi: 10.1007/978-3-030-67661-2{_}2/FIGURES/3. URL https://link.springer.com/chapter/10.1007/978-3-030-67661-2_2.

Kleinberg, R., Ligett, K., Piliouras, G., and Tardos, E. Beyond the Nash Equilibrium Barrier. *Innovations in Computer Science*, 2011.

Krichene, W. *Continuous and Discrete Dynamics for Online Learning and Convex Optimization*. PhD thesis, University of California, Berkeley, 2016. URL https://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-156.html.

Krichene, W., Drighès, B., and Bayen, A. M. Online Learning of Nash Equilibria in Congestion Games. *http://dx.doi.org/10.1137/140980685*, 53(2): 1056–1081, 4 2015. ISSN 03630129. doi: 10.1137/140980685. URL https://epubs.siam.org/doi/10.1137/140980685.

Leonardos, S. and Piliouras, G. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence*, 304:103653, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103653. URL https://www.sciencedirect.com/science/article/pii/S0004370221002046.

Leonardos, S., Piliouras, G., and Spendlove, K. Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality. *Advances in Neural Information Processing Systems*, 34:26318–26331, 12 2021.

Meiss, J. D. *Differential Dynamical Systems*. Society for Industrial and Applied Mathematics, 1 2007. doi: 10.1137/1.9780898718232.

Melo, E. A Variational Approach to Network Games. *SSRN Electronic Journal*, 11 2018. doi: 10.2139/SSRN.3143468. URL https://papers.ssrn.com/abstract=3143468.

Melo, E. On the Uniqueness of Quantal Response Equilibria and Its Application to Network Games. *SSRN Electronic Journal*, 6 2021. doi: 10.2139/SSRN.3631575. URL https://papers.ssrn.com/abstract=3631575.

Mertikopoulos, P. and Sandholm, W. H. Learning in Games via Reinforcement and Regularization. *https://doi.org/10.1287/moor.2016.0778*, 41(4):1297–1324, 8 2016. ISSN 15265471. doi: 10.1287/MOOR.2016.0778. URL https://pubsonline.informs.org/doi/abs/10.1287/moor.2016.0778.

Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019. doi: 10.1007/s10107-018-1254-8. URL https://doi.org/10.1007/s10107-018-1254-8.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. *Proceedings*, pp. 2703–2717, 2018. doi: 10.1137/1.9781611975031.172. URL https://epubs.siam.org/doi/10.1137/1.9781611975031.172.

Monderer, D. and Shapley, L. S. Potential games. *Games and Economic Behavior*, 14(1):124–143, 5 1996. ISSN 08998256. doi: 10.1006/game.1996.0044.

Mukhopadhyay, A. and Chakraborty, S. Deciphering chaos in evolutionary games. *Chaos*, 30(12): 121104, 12 2020. ISSN 10897682. doi: 10.1063/5.0029480. URL http://aip.scitation.org/doi/10.1063/5.0029480.

Pandit, V., Mukhopadhyay, A., and Chakraborty, S. Weight of fitness deviation governs strict physical chaos in replicator dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3):033104, 2018. doi: 10.1063/1.5011955. URL https://doi.org/10.1063/1.5011955.

Pangallo, M., Heinrich, T., and Farmer, J. D. Best reply structure and equilibrium convergence in generic games. *Science Advances*, 5(2), 2 2019. ISSN 23752548. doi: 10.1126/SCIADV.AAT1328/SUPPL{_}FILE/AAT1328{_}SM.PDF. URL https://www.science.org/doi/10.1126/sciadv.aat1328.

Pangallo, M., Sanders, J. B., Galla, T., and Farmer, J. D. Towards a taxonomy of learning dynamics in 2×2 games. *Games and Economic Behavior*, 132:1–21, 3 2022. ISSN 0899-8256. doi: 10.1016/J.GEB.2021.11.015.

Parise, F. and Ozdaglar, A. A variational inequality framework for network games: Existence, uniqueness, convergence and sensitivity analysis. *Games and Economic Behavior*, 114:47–82, 3 2019. ISSN 10902473. doi: 10.1016/j.geb.2018.11.012.

Parise, F., Grammatico, S., Gentile, B., and Lygeros, J. Distributed convergence to Nash equilibria in network and average aggregative games. *Automatica*, 117:108959, 2020. doi: 10.1016/j.automatica.2020.108959. URL https://doi.org/10.1016/j.automatica.2020.108959.

Perolat, J., Vylder, B. D., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., McAleer, S., Elie, R., Cen, S. H., Wang, Z., Gruslys, A., Malysheva, A., Khan, M., Ozair, S., Timbers, F., Pohlen, T., Eccles, T., Rowland, M., Lanctot, M., Lespiau, J.-B., Piot, B., Omidshafiei, S., Lockhart, E., Sifre, L., Beauguerlange, N., Munos, R., Silver, D., Singh, S., Hassabis, D., and Tuyls, K. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679. URL https://www.science.org/doi/abs/10.1126/science.add4679.

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

Sanders, J. B. T., Farmer, J. D., and Galla, T. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 8(1):4902, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-22013-5. URL https://doi.org/10.1038/s41598-018-22013-5.

Sato, Y. and Crutchfield, J. P. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, 1 2003. ISSN 1063651X. doi: 10.1103/PhysRevE.67.015206. URL https://journals.aps.org/pre/abstract/10.1103/PhysRevE.67.015206.

Sato, Y., Akiyama, E., and Farmer, J. D. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (7):4748–4751, 4 2002. ISSN 00278424. doi: 10.1073/pnas.032086299.

Schwartz, H. M. *Multi-Agent Machine Learning: A Reinforcement Approach*. Wiley, 2014. ISBN 9781118884614. doi: 10.1002/9781118884614.

Shalev-Shwartz, S. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*, 4(2), 2011. doi: 10.1561/2200000018. URL http://dx.doi.org/10.1561/2200000018.

Shapley, L. S. Some Topics in Two-Person Games. In *Advances in Game Theory. (AM-52)*, pp. 1–28. Princeton University Press, 5 2016. doi: 10.1515/9781400882014-002.

Sorin, S. and Wan, C. Finite composite games: Equilibria and dynamics. *Journal of Dynamics and Games*, 3(1): 101–120, 2016.

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

Tatarenko, T. and Kamgarpour, M. Learning Nash Equilibria in Monotone Games. *Proceedings of the IEEE Conference on Decision and Control*, 2019-December: 3104–3109, 12 2019. ISSN 25762370. doi: 10.1109/CDC40024.2019.9029659.

Tuyls, K., T Hoen, P. J., and Vanschoenwinkel, B. An evolutionary dynamical analysis of multi-agent learning in iterated games, 1 2006. ISSN 13872532.

van Strien, S. and Sparrow, C. Fictitious play in 3×3 games: Chaos and dithering behaviour. *Games and Economic Behavior*, 73(1):262–286, 2011. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2010.12.004. URL http://www.sciencedirect.com/science/article/pii/S089982561000196X.

Villatoro, D., Sen, S., and Sabater-Mir, J. Topology and Memory Effect on Convention Emergence; Topology and Memory Effect on Convention Emergence. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2, 2009. doi: 10.1109/WI-IAT.2009.155.

Xu, J., Zenou, Y., and Zhou, J. Networks in conflict: A variational inequality approach. *Available at SSRN 3364087*, 2019.

A. Proof of Theorem 3.2

Preliminaries We begin in this section by defining the various tools that we will use in our proof. 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 $D^2_{\mathbf{x}} f(\mathbf{x})$ is strongly positive definite with constant α. Thus, all eigenvalues of $D^2_{\mathbf{x}} f(\mathbf{x})$ are larger than α. To apply this in our setting, we use the following result.

Proposition A.1 ((Melo, 2021))**.** *The function $f(\mathbf{x}_k) = T_k \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle$ is strongly convex with constant T_k.*

Then, $D^2_{\mathbf{x}_k} f(\mathbf{x}_k)$ has eigenvalues larger than T_k.

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.2 (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.3. *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 δ_S.

Lemma A.4. *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*

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

where $A^{ij} \in \mathcal{M}_n(\mathbb{R})$ are matrices of the same dimension. Then

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

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

$$\|Nv\|_2^2 = \left\| \begin{pmatrix} g_{11}A^{11} & \ldots & g_{1n}A^{1n} \\ \vdots & & \vdots \\ g_{n1}A^{n1} & \ldots & g_{nn}A^{nn} \end{pmatrix} \begin{bmatrix} v^1 \\ \vdots \\ v^n \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sum_{1j} g_{1j} A^{1j} v^j \\ \vdots \\ \sum_{ni} g_{nj} A^{nj} v^j \end{bmatrix} \right\|_2^2 \leq \sum_{i=1}^n \left\| \sum_{j=1}^n g_{ij} A^{ij} v^j \right\|_2^2 . \tag{6}$$

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

$$\left\| \sum_{j=1}^n g_{ij} A^{ij} v^j \right\|_2 \leq \sum_{j=1}^n g_{ij} \left\| A^{ij} v^j \right\|_2 \leq \sum_{j=1}^n g_{ij} \left\| A^{ij} \right\|_2 \left\| v^j \right\|_2 \leq \max_{1 \leq i,j \leq n} \left\| A^{ij} \right\|_2 \sum_{j=1}^n g_{ij} \left\| v^j \right\|_2 . \tag{7}$$

By plugging (7) in (6) 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.

□

With these results in place, we can prove Theorem 3.2 in the main paper.

Proof of Theorem 3.2. In order to apply Lemma 2.9 we show that, under the condition (4), 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 $-D^2_{\mathbf{x}_k \mathbf{x}_k} u_k^H(\mathbf{x}_k, \mathbf{x}_{-k})$ along the diagonal. $N(\mathbf{x})$ is an off-diagonal block matrix with

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

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

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

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

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

Then we apply the following results.

Then, from Proposition A.1 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

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

where we employ Propositions A.3, Lemma A.4 and the fact that G is symmetric so that $\|G\|_\infty = \|G\|_1$. The matrices A, B are chosen so that

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

Then, under (4), $\lambda_{\min}(\bar{J}(\mathbf{x})) \geq T - \frac{1}{2}\delta_S\|G\|_\infty > 0$ and, therefore $F(\mathbf{x})$ is strongly monotone with constant $T - \frac{1}{2}\delta_S \|G\|_\infty$. Using Lemma 2.9, it follows that Q-Learning Dynamics converge to a unique QRE. □