The Impact of Exploration on Convergence and Performance of Multi-Agent Q-Learning Dynamics

Aamal Hussain [1] **Francesco Belardinelli** [1] **Dario Paccagnan** [1]

Abstract

Understanding the impact of exploration on the behaviour of multi-agent learning has, so far, benefited from the restriction to potential, or network zero-sum games in which convergence to an equilibrium can be shown. Outside of these classes, learning dynamics rarely converge and little is known about the effect of exploration in the face of non-convergence. To progress this front, we study the smooth Q-Learning dynamics. We show that, in any network game, exploration by agents results in the convergence of Q-Learning to a neighbourhood of an equilibrium. This holds independently of whether the dynamics reach the equilibrium or display complex behaviours. We show that increasing the exploration rate decreases the size of this neighbourhood and also decreases the ability of all agents to improve their payoffs. Furthermore, in a broad class of games, the payoff performance of Q-Learning dynamics, measured by Social Welfare, decreases when the exploration rate increases. Our experiments show this to be a general phenomenon, namely that exploration leads to improved convergence of Q-Learning, at the cost of payoff performance.

1. Introduction

Learning in games requires agents to explore their state space. Whilst it is known that the rate of exploration impacts the behaviour of the learning dynamic (Tuyls et al., 2006), understanding precisely this impact often relies on placing restrictions on the structure of the interaction. A primary example of this are network zero-sum games, which model perfect competition between agents. Here it is known that, without exploration, agents cannot reach an equilibrium (Mertikopoulos et al., 2018; Ganesh Nagarajan et al., 2020), whilst arbitrarily small exploration leads to convergence (Leonardos et al., 2021). Similarly studied classes of games include potential games (Leonardos & Piliouras, 2022) and games with two players and two actions (Kianercy & Galstyan, 2012).

Beyond these restrictive classes however, the picture is much less clear. Whilst recent work has shown that high exploration rates always lead to convergence (Hussain et al., 2023), it is also well known that learning dynamics often display complex behaviours, including limit cycles (Mertikopoulos et al., 2018; Galla, 2011) and chaos (Bielawski et al., 2021; Galla & Farmer, 2013; Sanders et al., 2018; Mukhopadhyay & Chakraborty, 2020; Sato et al., 2002). The latter case represents a particularly strong barrier towards understanding the asymptotic behaviour of learning. In short, the prospect of convergence to an equilibrium cannot be taken for granted. Unfortunately, little is known about the impact of exploration in the face of non-convergent dynamics.

Yet understanding exploration remains an important endeavour as it allows agents to avoid suboptimal, or potentially unsafe areas of their state space (Leonardos & Piliouras, 2022; Bura et al., 2022; Bai et al., 2021). In addition, it is empirically known that the choice of exploration rate impacts the expected total reward (Cai et al., 2020). This has lead to an increased interest in studying the influence of exploration on learning in single-agent settings (Schäfer et al., 2021; Piliouras, 2020). It is then paramount to understand the effect of exploration also in *multi-agent environments* outside the restrictions of potential, or network zero-sum settings. This leads us to our central question:

How does exploration affect reinforcement learning dynamics in arbitrary games, even if convergence to an equilibrium cannot be guaranteed?

Main Contributions To answer this question, we consider the *smooth Q-Learning* (SQL) dynamic, a learning dynamic which quantifies the tendency for agents to explore their state space whilst also seeking to maximise their payoffs.

We then lift the assumption of convergence to an equilib-

[1]Department of Computing, Imperial College London, London, United Kingdom. Correspondence to: Aamal Hussain <aamal.hussain15@imperial.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

rium. In doing so, we show that, in network games with unique Nash Equilibria, SQL remains contained in a neighbourhood of an equilibrium for any positive exploration rate. The effect of increased exploration is to decrease the size of this neighbourhood. As such, exploration rates can be tuned to control the 'degree' of convergence of SQL. By considering convergence to a set, we are able to include several dynamical behaviours, including convergence to a fixed point, but also to a cycle or chaotic behaviour. In addition, this greatly widens the class of games in which the effect of exploration can be understood.

Next, we analyse the effect of exploration on system performance. Namely, we show that, by increasing exploration, the ability for agents to improve their payoffs whilst following SQL decreases. We exemplify this by considering a class of games in which Q-Learning does not converge for low exploration rates, but non-convergent SQL dynamics *outperform* the equilibrium, in terms of the sum of payoffs of all agents. Therefore, whilst increasing exploration drives the system towards convergence, the payoff performance of SQL suffers. Our experiments show this to be a general property of SQL dynamics. As a result, when system performance is desirable, it may be beneficial to restrict exploration, even if the dynamics do not converge to an equilibrium.

Related Work Since its conception, the smooth Q-Learning dynamic (Tuyls et al., 2006; Sato & Crutchfield, 2003) has received significant attention as it provides a fundamental model of exploration in multi-agent learning and is closely related to the well-studied *replicator dynamics* (Maynard Smith, 1974; Hofbauer & Sigmund, 1998; Leonardos & Piliouras, 2022), as well as the popular Q-Learning algorithm (Sutton & Barto, 2018; Schwartz, 2014). Interestingly, a number of studies have shown that, for various choices of exploration rates, the Q-Learning dynamic can display a number of complex behaviours, such as limit cycles (Mertikopoulos et al., 2018; Kleinberg et al., 2011; Hofbauer, 1996) or even chaos (Galla & Farmer, 2013; Sato et al., 2002; Ganesh Nagarajan et al., 2020). To make matters worse, as the number of players increases the likelihood of these complex behaviours increases (Sanders et al., 2018). In fact, when there is no exploration on the part of every agents, the Q-Learning dynamics cannot converge to an interior equilibrium (Vlatakis-Gkaragkounis et al., 2020). This behaviour is not limited to the Q-Learning dynamics; a wide array of online learning algorithms are known to exhibit chaos, including *Fictitious Play* (van Strien & Sparrow, 2011; Ewerhart & Valkanova, 2020) and *Follow the Regularised Leader* (Bielawski et al., 2021; Andrade et al., 2021; Anagnostides et al., 2022; Cheung & Tao, 2021).

The presence of these non-convergent behaviours result in an inherent challenge in understanding the effect of exploration on multi-agent learning (Klos et al., 2010). As a result studies considering this topic from a theoretical standpoint are often limited to understanding equilibrium behaviours. For instance, (Leonardos & Piliouras, 2022) shows that different exploration rates can result in different stability properties, which can lead to potentially unbounded gains and losses in system performance. Similar phase transitions are also found in (Kaisers & Tuyls, 2011; Kianercy & Galstyan, 2012; Piliouras, 2020). By contrast, (Hussain et al., 2023) shows that sufficiently high exploration rates yields convergence of the Q-Learning dynamic to a unique equilibrium, regardless of the number or stability of Nash Equilibria in the game. To our knowledge ours is the first study to analyse, theoretically and empirically, the effect of exploration in arbitrary multi-agent settings in terms of dynamics and system performance, *without* assuming convergence to an equilibrium.

2. Preliminaries

In this paper we study a network game $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$, where $\mathcal{N}$ denotes a finite set of agents. Each agent $k \in \mathcal{N}$ has a finite set S_k of actions which are indexed by $i = 1, \dots, n_k$, and can play a mixed strategy $\mathbf{x}_k$, i.e. a discrete probability distribution over its set of actions. The set of all such mixed strategies is the unit simplex in $\mathbb{R}^{n_k}$, i.e. $\Delta_k = \{\mathbf{x}_k \in \mathbb{R}^{n_k} \mid \sum_{i\in S_k} x_{ki} = 1, \text{ and } x_{ki} \geq 0 \text{ for all } i \in S_k\}$. We denote with $\Delta = \times_{k\in\mathcal{N}}\Delta_k$ the joint simplex over all agents, with $\mathbf{x} = (\mathbf{x}_k)_{k\in\mathcal{N}}$ the joint mixed strategy of all agents and, for any k, with $\mathbf{x}_{-k} = (\mathbf{x}_l)_{l\in\mathcal{N}\setminus\{k\}} \in \Delta_{-k}$ the joint strategy of all agents other than k.

Agents in a network game interact according to an edgeset $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. Each edge corresponds to a bimatrix game (A^{kl}, A^{lk}). The expected payoff for each agent $k \in \mathcal{N}$ who plays mixed strategy $\mathbf{x}_k$ against joint mixed strategy $\mathbf{x}_{-k}$ is given by

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

For any $\mathbf{x} \in \Delta$, the reward to agent k when they play action $i \in S_k$ as $r_{ki}(\mathbf{x}) := \partial u_{ki}(\mathbf{x})/\partial x_{ki}$. With this, we can write $r_k(\mathbf{x}) = (r_{ki}(\mathbf{x}))_{k\in\mathcal{N}}$ as the concatenation of all rewards to agent k. In this notation, $u_k(\mathbf{x}) = \mathbf{x}_k^\top r_k(\mathbf{x})$. Using this notation, we can define the equilibrium of 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 $i \in S_k$,

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

In the definition of the QRE, T_k denotes the *exploration rate* of the agent. Note that as $T_k \to \infty$, the QRE is unique and

is given by $\bar{\mathbf{x}} = (\frac{1}{n_k}\mathbf{1})_{k\in\mathcal{N}}$. This equilibrium corresponds to the case in which each agent plays each action with the same probability, regardless of the payoff received. In the other limit, $T_k \to 0$, the QRE corresponds to the Nash Equilbrium, which we now define.

Definition 2.2 (Nash Equilibrium (NE))**.** A joint mixed strategy $\bar{\mathbf{x}} \in \Delta$ is an *Nash Equilibrium* if, for all agents k and all $\mathbf{x}_k \in \Delta_k$,

$$\langle \mathbf{x}_k, r_k(\bar{\mathbf{x}}_{-k})\rangle \leq \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k})\rangle \tag{1}$$

Informally, at a Nash Equilibrium, no agent can increase their utility by means of unilateral deviations, i.e. agents are considered perfectly rational. With this in mind, the QRE can be thought of as an equilibrium notion for agents with bounded rationality.

In this work, we will be making the following assumption which will allow for a comparative analysis between non-convergent behaviours and a unique equilibrium.

Assumption 2.3. The network game $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ has a unique, interior Nash Equilibrium

We can extend this assumption towards the QRE of the game using the following result.

Proposition 2.4. *If a game Γ has a unique Nash Equilibrium then, for any $T_1, \dots, T_N > 0$, there exists a unique interior QRE.*

Influence Bound Our convergence results on Q-Learning will depend on a suitable notion of *size* of the game. To define this formally, we introduce the *influence bound* (Melo, 2021) of the game which is defined as

Definition 2.5 (Influence Bound)**.** A game Γ has the *influence bound* δ given by

$$\delta = \max_{k\in\mathcal{N}, i\in S_k, s_{-k}, \tilde{s}_{-k}\in S_{-k}} \{|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|\}$$

where the pure strategies $s_{-k}, \tilde{s}_{-k} \in S_{-k}$ differ only in the strategies of one agent $l \neq k$.

Since $|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|$ measures the change in reward to agent k for playing action i due to a change the other players' actions, the influence bound δ defines the maximum influence (in terms of reward) that any agent could receive from their opponents.

Remark 2.6. In the case of a network game, the influence bound of the game is simply

$$\max_{k\in\mathcal{N}, i\in S_k, s_{-k}, \tilde{s}_{-k}\in S_{-k}} |\left(A^k\right)_{i,s_{-k}} - \left(A^k\right)_{i,\tilde{s}_{-k}}|.$$

In other words, it is the maximum difference between any row elements across the payoff matrices for all agents.

When considering performance, we use two closely related measures. The first is the total *exploitability*. Informally, exploitability measures an agent's ability to improve their current payoff by deviating to another strategy. More formally, we define exploitability of a strategy $\mathbf{x}$ with respect to a set $S = \times_k S_k$ as

$$\Phi_S(\mathbf{x}) = \sum_k \max_{\mathbf{y}_k \in S_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - u_k(\mathbf{x}_k, \mathbf{x}_{-k}). \tag{2}$$

Our second metric for performance is *Social Welfare*, which measures the total payoff received by all agents. Formally, the Social Welfare of a mixed strategy $\mathbf{x} \in \Delta$ is given by

$$SW(\mathbf{x}) = \sum_k u_k(\mathbf{x}_k, \mathbf{x}_{-k}) \tag{3}$$

Social Welfare is a stronger measure of performance than Exploitability as the latter considers an agent's ability to improve their payoff, whereas the former measures the realised payoff that each agent receives.

Learning Model We study a smooth variant of Q-Learning with Boltzmann exploration, called smooth Q-Learning (SQL) (Tuyls et al., 2006). This requires that each agent k updates their mixed strategy x_k according to the dynamic

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

in which $T_k \in [0, \infty)$ denotes the exploration rate of agent k. At the limit of zero exploration rates we recover the replicator dynamic in which agents maximise their payoff at every time step (Sato & Crutchfield, 2003). At the other limit, $T_k \to \infty$, we recover an entropy maximising dynamic in which the unique fixed point $\bar{\mathbf{x}} = (\frac{1}{n_k}\mathbf{1})_{k\in\mathcal{N}}$ is globally asymptotically stable. This allows us to capture, with the parameter T_k, how the exploration rate affects the dynamics.

3. Convergence and Performance

In this section, we first show how the imposition of exploration by all agents forces the learning dynamics to converge to a neighbourhood of the QRE $\mathbf{x}$. Using this, we define a lower bound on each x_{ki}, independently of whether the dynamics converge to an equilibrium or display more complex behaviour.

Convergence To define convergence, we require a measure of distance. To this end, we employ the *Kullback-Leibler* (KL) Divergence.

Definition 3.1 (Kullback-Leibler Divergence)**.** The KL Divergence between a set of joint mixed strategies $\mathbf{x}, \mathbf{y} \in \Delta$

is given by

$$D_{KL}(\mathbf{y}||\mathbf{x}) = \sum_k D_{KL}(\mathbf{y}_k||\mathbf{x}_k) = \sum_{ki} y_{ki} \ln \frac{y_{ki}}{x_{ki}} \quad (4)$$

Notice that the KL-Divergence does not formally define a metric as it is not symmetric (i.e. in general $D_{KL}(\mathbf{y}||\mathbf{x}) \neq D_{KL}(\mathbf{x}||\mathbf{y})$). Rather, the KL-Divergence can be thought of as measuring the overlap between probability distributions $\mathbf{y}$ and $\mathbf{x}$. The key point which we will use in our main theorem is that $D_{KL}(\mathbf{y}||\mathbf{x})$ is zero if and only if $\mathbf{x} = \mathbf{y}$ and is positive everywhere else.

Theorem 3.2. *Let δ be influence bound of the game Γ and let $\bar{\mathbf{x}} \in int\Delta$ denote the QRE of the game for some $T_1, \dots, T_N$. Then, the Q-Learning dynamics remain asymptotically within the set*

$$S_T = \{\mathbf{x} \in \Delta \,|\, D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) \leq \frac{\delta}{T_{\min}} \sum_k n_k\}$$

Here, we provide a sketch of the proof of Theorem B.2 whilst deferring the full proof to the supplementary material. We first show that, close to the boundary of the simplex, Q-Learning strictly minimises the KL-Divergence between the current mixed strategy $\mathbf{x}(t)$ and the unique QRE $\bar{\mathbf{x}}$. This is formalised in the following Lemma

Lemma 3.3. *Consider a game Γ with influence bound δ. Let $\mathbf{x}(t)$ denote the joint strategy generated by (SQL) at some time t for some initial condition $\mathbf{x}(0)$. Also let $\bar{\mathbf{x}}$ denote the QRE for the game for some T_k. Then $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ is a decreasing function along trajectories of the Q-Learning dynamic (SQL) for all $t \in [0, \infty)$ such that*

$$D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\bar{\mathbf{x}}) > \frac{\delta}{T_{\min}} \sum_k n_k \quad (5)$$

We apply this Lemma to prove Theorem B.2 in the following manner. Since $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ is also bounded below, it immediately follows that $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ must decrease until $\mathbf{x}(t)$ reaches a region where $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\bar{\mathbf{x}}) \leq \frac{\delta}{T_{\min}} \sum_k n_k$. Let us denote this region as S. Once $\mathbf{x}(t)$ enters S, which occurs in finite time, the KL-Divergence is no longer guaranteed to be a decreasing function. Therefore, the dynamics can leave S. Regardless of whether $\mathbf{x}(t)$ remains in S or ultimately leaves it, the decreasing property shown by Lemma 3.3 enforces that $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ cannot increase further than $\sup_{\mathbf{x} \in S} D_{KL}(\bar{\mathbf{x}}||\mathbf{x}) =: D_S$. Using also that $D_S \leq \frac{\delta}{T_{\min}} \sum_k n_k$, it follows that

$$\lim_{t \to \infty} \mathbf{x}(t) \in S_T$$

Remark 3.4. At first glance, it appears that the size of the set defined by Theorem B.2 increases with the number of players and number of actions due to the presence of the term $\sum_k n_k$. On closer inspection, however, we see that the KL-Divergence itself is given by a summation over agents and actions. Therefore, both sides of the inequality which define S_T grow at the same rate.

Theorem B.2 determines the convergence structure of Q-Learning in arbitrary games. In particular, it finds a set in which Q-Learning remains asymptotically trapped, independently of whether the dynamics are ultimately chaotic, cyclic or converge to an equilibrium. As T_k increases, the size of this set decreases, tightening the convergence bound. On the other hand, the size of the set increases with the size of the game, as measured by the influence bound. Indeed, in the limit $T_k \to \infty$ for all k, the set defined by Theorem B.2 is a singleton, so that Q-Learning must converge to the QRE.

Another implication of Theorem B.2 is that, for any $T_1, \dots, T_N > 0$ the Q-Learning dynamics must remain bounded away from the boundary of the simplex $\partial\Delta$ for all $t > 0$. In addition, this bound increases with T_k, resulting in the dynamics being forced further in the interior of Δ.

Corollary 3.5. *In the setting of Theorem B.2, when each agent k has exploration rate $T_k \geq 0$ and follows the Q-Learning dynamic, there exists an $\epsilon_T \in [0, 1/(\min_k n_k)]$ which grows with $T_{\min} = \min_k T_k$ such that for any $k \in \mathcal{N}$, $i \in S_k$*

$$\liminf_{t \to \infty} x_{ki}(t) \geq \epsilon_T$$

Performance Next, we examine an important implication of Theorem B.2, namely that, as the learning dynamics remain asymptotically bounded within the interior of the simplex, the *exploitability* of the system decreases. We go further by placing an upper bound on this reduction of exploitability in terms of the lower bound defined by Corollary B.5.

As all results depend on $T_{\min}$, we ease notation by assuming the exploration rates T_k for all agents are equal. Then we drop the $\min$ notation and just write T. Next, we define ϵ_T as the lower bound in Corollary B.5 for some choice of T,

Now, for any T define $\Omega_k = \times_{i \in S_k}[\epsilon_T, 1 - \epsilon_T]$ and $\Omega = \times_k \Omega_k$. Then it is clear that $S_T \subset \Omega$ and, as $T \to 0$, $\Omega \to \Delta$.

With this, we can apply the definition of exploitation (9) with respect to Ω.

$$\Phi_\Omega(\mathbf{x}) = \sum_k \max_{\mathbf{y}_k \in \Omega_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - u_k(\mathbf{x}_k, \mathbf{x}_{-k}) \quad (6)$$

The motivation for defining this metric is as follows: the definition of exploitability which is most widely applied in the context of online learning corresponds to Φ_Δ (Perrin

et al., 2020; Gemp et al., 2022), which measures the best payoff any agent could receive by deviating to any other strategy in the simplex, assuming that all other agents keep their strategies fixed. However, if agents are following the Q-Learning dynamic, with some positive exploration rates, the whole simplex becomes unavailable. Rather agents can only improve their strategy from within the set S_T. We can compare Φ_Ω against the case of zero exploration using $\Delta\Phi(\mathbf{x}) = \Phi_\Omega(\mathbf{x}) - \Phi_\Delta(\mathbf{x})$. As such, $\Delta\Phi$ measures the change in exploitability as exploration is introduced.

Theorem 3.6. *Let Γ be a game with a unique, interior Nash Equilibrum. Then, for any $T \geq 0$, $\Delta\Phi(\mathbf{x}) \leq 0$ with equality holding iff $\mathbf{x}$ is the NE of the game, denoted $\bar{\mathbf{x}}$. In addition, for all $\mathbf{x} \neq \bar{\mathbf{x}}$, $\Delta\Phi(\mathbf{x}) \leq -\alpha\epsilon_T < 0$, for some $\alpha > 0$ i.e. exploitability decreases as T increases.*

Discussion. Theorem B.7 shows that the ability of an agent to improve their payoffs strictly decreases as exploration increases. As a caveat, Theorem B.2 shows that higher exploration rates leads to a greater certainty in convergence of Q-Learning. In summary, we show that stronger guarantees on the convergence due to exploration may come at the price of decreased system performance.

Whilst Theorems B.2 and B.7 paint a broad stroke on convergence and performance, a limitation is that Theorem B.2 does not have anything to say about the behaviour of Q-Learning within the defined set. Indeed it may be the case, as it is for network zero-sum games, that Q-Learning converges to a QRE for all $T_k > 0$. Similarly, whilst Lemma B.7 shows that agents cannot improve their payoffs as exploration increases, it does not show that their payoffs decrease as the agents move from non-convergent to convergent behaviours.

4. Exploration Reduces Payoff Performance

In the previous sections, we showed that exploration leads to a decreased ability of each agent to improve their payoff. In this section we tackle the limitations discussed in the previous section. To do this we focus a specific class of games for which we show that non-convergent behaviours strictly outperform convergence. In this case, performance is measured through social welfare so that the agents' realised payoffs are considered, rather than their *ability* to improve their payoffs.

Shapley Network Game In the first example we examine a network of agents, where each edge is equipped with a Shapley game. In particular, the payoff to each agent k is given by

$$u_k = u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \mathbf{x}_k \mathbf{A} \mathbf{x}_{k-1} + \mathbf{x}_k \mathbf{B}^\top \mathbf{x}_{k+1}$$

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

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

In the two agent case, (Shapley, 2016) showed that the popular *Fictitious Play* dynamics (Brown P, 1949; Hofbauer & Sigmund, 2003) do not converge to an NE, but rather reach a limit cycle. In (Ostrovski & van Strien, 2014), the authors show that the non-convergent cycle outperforms the Nash Equilibrium. In (Hussain et al., 2023), the multi-agent extension was experimentally examined and it was suggested that the performance of Q-Learning decreases as the system moves from a limit cycle to an equilibrium. We make this statement rigorous by showing that the dynamics, whilst initially non-convergent, can be made convergent through a sufficiently high exploration rate. However, we find that the result is accompanied by a strict decrease in the social welfare along trajectories.

Lemma 4.1. *For any $\beta \in (0, 1)$ and $T_k \geq 0$, the Network Shapley game has a QRE at the uniform distribution $\bar{\mathbf{x}} = (\frac{1}{3})_{k \in \mathcal{N}, i \in S_k}$ which is globally repelling under (SQL) at $T_k = 0$ and locally attracting if, for all $k \in \mathcal{N}$, $T_k > 1 + \frac{\beta}{3}$. The QRE is globally asymptotically stable if $T_k > (N - 1)(1 + \beta)$.*

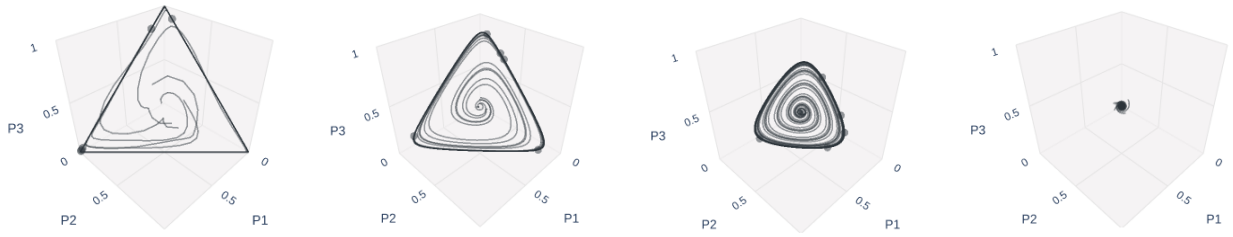
We report the proof of Lemma C.1 in the supplementary material. The main takeaway is that (SQL) does not reach the QRE in the case of zero exploration, whilst for sufficiently high exploration rates the QRE is globally attracting. As a result, exploration drives the Q-Learning dynamics from initially non-convergent behaviour to globally convergent. Whilst, at first this seems like a positive result, in the following theorem, we show that the non-convergent behaviours *strictly* outperform the equilibrium in terms of Social Welfare.

In Network Shapley Games, let the *time-average social welfare* (TSW) along Q-Learning trajectories be defined as

$$TSW = \lim_{t \to \infty} \frac{1}{t} \int_0^t SW(\mathbf{x}(s))\, ds$$

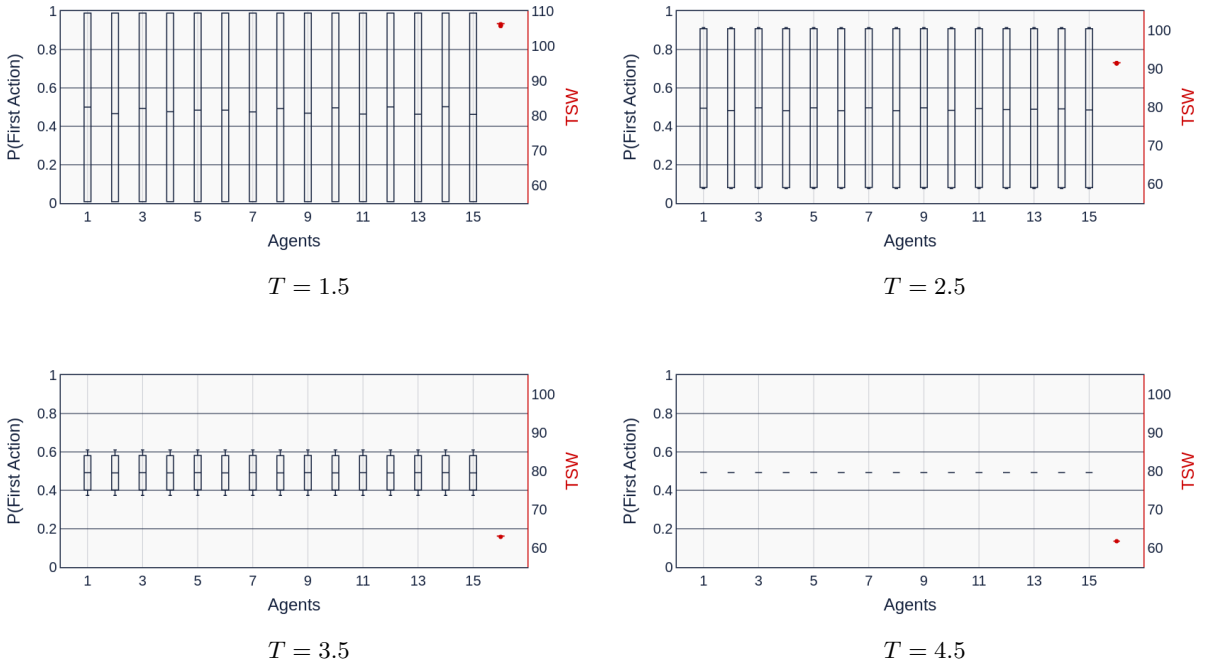
where $\mathbf{x}(t)$ is a trajectory of mixed strategies generated according to the Q-Learning dynamic for some initial condition $\mathbf{x}_0 \in \Delta$.

Theorem 4.2. *Non-convergent trajectories of Q-Learning strictly outperform the social welfare of the unique QRE $\bar{\mathbf{x}} \in \Delta$. In particular, $TSW \geq SW(\bar{\mathbf{x}})$ with equality holding if and only if the trajectory converges to the QRE.*



$T = 0.1, TSW = 3.22$ $T = 0.2, TSW = 2.71$ $T = 0.25, TSW = 2.34$ $T = 0.3, TSW = 2.0$

Figure 1. Trajectories of Q-Learning in the Network Shapley Game, for $\beta = 0.2$ with five agents alongside experimentally obtained TSW. Axes the probability with which three agents play their first action. For low values of T, the system reaches a limit cycle whose size of this limit cycle decreases as exploration forces the dynamics into the interior of the simplex. Eventually, the system equilibriates at the uniform distribution. All non convergent dynamics strictly outperform the QRE.



$T = 1.5$ $T = 2.5$

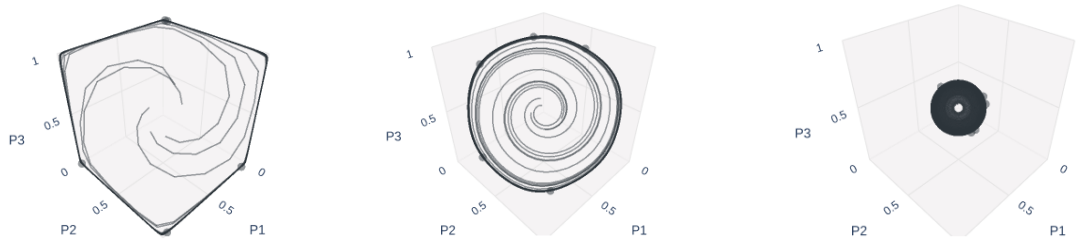
$T = 3.5$ $T = 4.5$

Figure 2. Asymptotic Behaviour and Performance of Q-Learning in the Network Chakraborty Game, for $U = 7.0, V = 8.5$ with 15 agents. Boxplots show the spread of probabilities with which agents play their first action in the final 25% of 1×10^5 iterations of learning. The red plot shows the TSW asymptotically achieved. For low values of T, the asymptotic dynamics are spread across the entire simplex, whilst achieving a high TSW. As T increases, the dynamics eventually reach a fixed point, but consistently decrease TSW as a result.

5. Experiments on Exploration

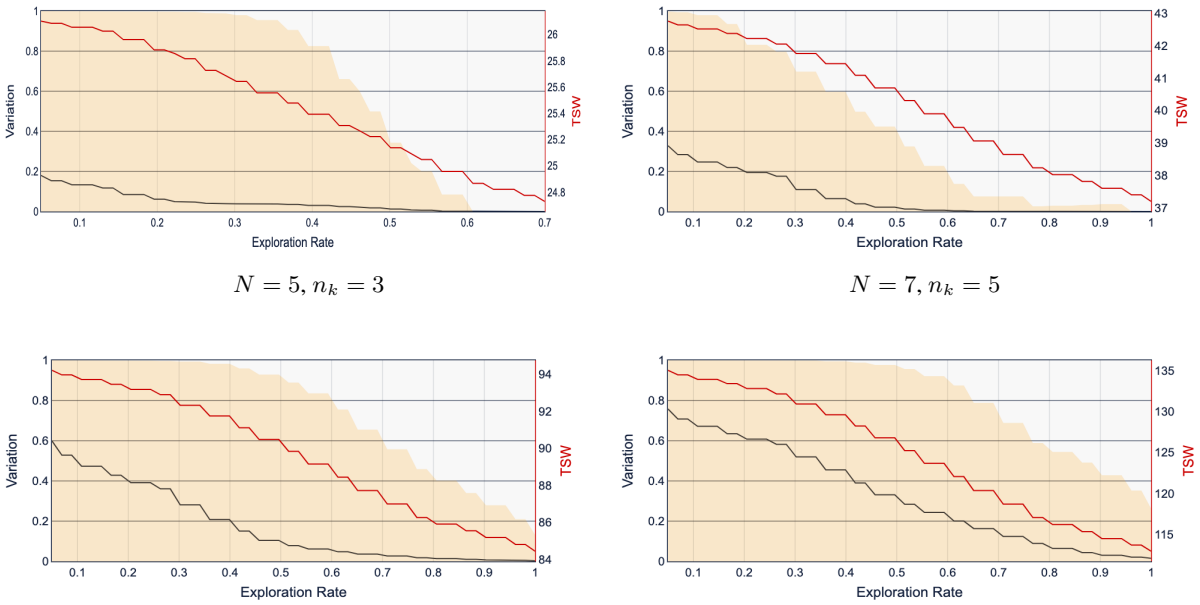
Our experiments further analyse the phenomenon that we observe in our results - namely that increased exploration results in Q-Learning dynamics asymptotically reaching a set in the interior of the simplex, whose size decreases with T. We also analyse the effect on Social Welfare, to test whether the phenomenon shown for the Network Shapley Game, namely that exploration leads to a decrease in payoff performance, holds more generally.

Network Shapley Game In Figure 5 we visualise the effect of exploration on the Network Shapley game, examined in Section C. We generate a network game with five players and run Q-Learning on the game. To be able to visualise the trajectory, we select three agents and plot their first action on the space $[0, 1]^3$. In Figure 5, we keep β fixed at 0.2, which yields a fixed $\delta = 1 + \beta = 1.2$. This process is repeated for increasing choices of T. TSW is calculated as the final time averaged social welfare after running Q-Learning for 1×10^5 iterations.



$T = 0.5$ $T = 1.5$ $T = 2.0$

Figure 3. Trajectories of Q-Learning dynamics in the Network Chakraborty Game for $U = 7.0, V = 8.5$ and three agents. Trajectories show the probability with which the first action is selected. For low T, the dynamics cycle on the boundary of the simplex, leading to a large variation in each strategy component in the asymptotic limit cycle. As T increases, the size of this cycle decreases, resulting in a smaller variation of strategy components.



$N = 5, n_k = 3$ $N = 7, n_k = 5$

$N = 15, n_k = 5$ $N = 20, n_k = 8$

Figure 4. Convergence and Performance of Q-Learning in Randomly Generated Games with payoff elements in $[0, 5]$. Convergence is measured by maximum variation in mixed strategies, given by (21). The black line depicts the average taken over all 50 games and initial conditions, whilst the confidence interval depicts the maximum and minimum variation. The red line shows the TSW achieved by Q-Learning, averaged over all games.

It can be seen that, for small values of T, Q-Learning does not converge, but rather reaches a limit cycle. As predicted by Theorem B.2, the size of this limit cycle decreases as T increases, until eventually Q-Learning converges asymptotically to the QRE at the uniform distribution. Furthermore, as predicted by Theorem C.3, all non-convergent behaviours strictly outperform the Social Welfare of the uniform distribution which, for our choice of β, is 2.

Network Chakraborty Game. Next, we consider a class of two-action network games which we call the *Network Chakraborty* game. In this game, each agent responds only to the 'previous' agent in a circular chain. More formally, the payoff to each agent k is

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

This game was analysed in (Pandit et al., 2018) under the context of an infinitely large population of agents with identical payoffs. It was shown that, for certain combinations of

U, V, a discrete time analog of replicator dynamics shows chaotic behaviour. Here, we extend this game to the multi agent case and analyse it under Q-Learning.

The results of this investigation are presented in Figure 6. We examine a 15 agent game with $U = 7.0, V = 8.5$, which appears in (Pandit et al., 2018) as a case which shows chaos by discrete replicator. We run Q-Learning, with 100 initial conditions, on this system for 1×10^5 iterations and isolate the final 25,000 iterations. By examining this window, we include the possibility of complex asymptotic behaviours. In fact, we visualise the trajectories in Figure 7 and find that, in a three agent network, the dynamics reach a limit cycle. The boxplot depicts, for each agent the probability of choosing their first action taken across the entire 25,000 final iterations, and all initial conditions. Beside this, we plot the TSW achieved by Q-Learning, taken across all initial conditions.

Once again it is clear that, for small exploration rates, the dynamics do not converge to an equilibrium. Rather, the asymptotic behaviour is spread across the entire state space. As exploration increases, the spread of probability distributions is bounded within the interior of the simplex, until eventually the dynamics equilibriate when $T = 4.5$. This process is again accompanied by a decrease in TSW achieved by Q-Learning, which reaches its minimum when the dynamics converge to an equilibrium.

Arbitrary Games. Finally, we look to extend our investigation beyond specific classes of games. To do this, we analyse the effect of exploration in randomly generated games, which do not follow any specific payoff or network structure. In addition, these games are not required to satisfy Assumption 2.3. To ensure like comparison, we generate payoffs that are positive and upper bounded. This ensures that the difference in payoffs between two randomly drawn games are not so significant as to affect plotting the results. Neither assumption impacts the generality of the results as the dynamics of Q-Learning are invariant to additions and multiplications by positive constants to all elements of the payoffs. A proof of this statement appears in the supplementary material.

To generate Figure 8, we run Q-Learning in 50 randomly generated games, for 5 initial conditions and record the final 25,000 iterations. Then we determine the largest variation in mixed strategies across all agents and all actions. More formally, this process estimates

$$\max_{ki} \lim_{t\to\infty} \left(\max_t x_{ki}(t) - \min_t x_{ki}(t) \right) \tag{7}$$

Figure 8 shows that the variation decreases as exploration rates are increased. Taken together, our results present strong evidence that the relation between convergence and performance does not just occur in the special cases already examined, but rather holds in the vast majority of network games.

6. Conclusion

Understanding the effect of exploration in multi-agent learning faces a significant challenge due to the fact that, outside of a restrictive class of games, online learning often does not display asymptotic convergence to an equilibrium. We made a first contribution at solving this by showing that, in all network games with unique Nash Equilibrium (NE), smooth Q-Learning converges to a neighbourhood of the equilibrium. This occurs independently of the behaviour of the learning dynamics within this neighbourhood. The size of this neighbourhood can be decreased by increasing the exploration rates of the agents. As such, controlling the degree to which Q-Learning converges amounts to parameter tuning.

The downside of this process is reduced asymptotic performance of learning. We show that, in all games, increased exploration leads naturally to a reduced ability for agents to improve their payoff through learning. As our results place upper bounds on this phenomena, they give a manner in which exploration rates can be tuned to balance convergence and payoff performance. To take this further, we show that non convergent Q-Learning dynamics strictly outperform convergence in a multi-agent extension of the Shapley game. As our experiments confirm, this turns out to be a general phenomena across a large number of games.

The results in this paper brings improves the understanding of exploration in learning dynamics outside the realm of potential and network zero sum games. An interesting avenue for future work would be to continue developing this direction by lifting the assumption of network interactions and by considering the effect of exploration in games with multiple Nash Equilibria.

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).

References

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 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.

Bielawski, J., Chotibut, T., Falniowski, F., Kosiorowski, G., Misiurewicz, M., and Piliouras, G. Follow-the-regularized-leader routes to chaos in routing games. 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. 925–935. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/bielawski21a.html.

Brown P, G. W. SOME NOTES ON COMPUTATION OF GAMES SOLUTIONS. Technical report, 4 1949. URL https://apps.dtic.mil/sti/citations/AD0603823.

Bura, A., Hasanzadezonuzy, A., Kalathil, D., Shakkottai, S., and Chamberland, J.-F. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=U4BUMoVTrB2.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1283–1294. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

Cheung, Y. K. and Tao, Y. Chaos of learning beyond zero-sum and coordination via game decompositions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a3wKPZpGtCF.

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.

Galla, T. Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(8), 8 2011. ISSN 17425468. doi: 10.1088/1742-5468/2011/08/P08007.

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.

Ganesh Nagarajan, S., Balduzzi, D., and Piliouras, G. From Chaos to Order: Symmetry and Conservation Laws in Game Dynamics. Technical report, 11 2020. URL http://proceedings.mlr.press/v119/nagarajan20a.html.

Gemp, I., Savani, R., Lanctot, M., Bachrach, Y., Anthony, T., Everett, R., Tacchetti, A., Eccles, T., and Kramár, J. 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, pp. 507–515, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Hofbauer, J. Evolutionary dynamics for bimatrix games: A Hamiltonian system? *Journal of Mathematical Biology*, 34(5-6):675–688, 1996. ISSN 14321416. doi: 10.1007/BF02409754.

Hofbauer, J. and Sigmund, K. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 5 1998. ISBN 9780521623650. doi: 10.1017/CBO9781139173179. URL https://www.cambridge.org/core/books/evolutionary-games-and-population-dynamics/A8D94EBE6A16837E7CB3CED24E1948F8.

Hofbauer, J. and Sigmund, K. Evolutionary Game Dynamics. *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY*, 40(4):479–519, 2003.

Hussain, A. A., Belardinelli, F., and Piliouras, G. Asymptotic convergence and performance of multi-agent q-learning dynamics, 2023. URL https://arxiv.org/abs/2301.09619.

Kaisers, M. and Tuyls, K. FAQ-learning in matrix games: Demonstrating convergence near Nash equilibria, and bifurcation of attractors in the Battle of Sexes. Technical report, 2011. URL www.aaai.org.

Kianercy, A. and Galstyan, A. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(4):041145, 4 2012. ISSN

15393755. doi: 10.1103/PhysRevE.85.041145. URL https://journals.aps.org/pre/abstract/10.1103/PhysRevE.85.041145.

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

Klos, T., van Ahee, G. J., and Tuyls, K. Evolutionary dynamics of regret minimization. In Balcázar, J. L., Bonchi, F., Gionis, A., and Sebag, M. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 82–96, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15883-4.

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.

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

McKelvey, R. D. and Palfrey, T. R. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38, 7 1995. ISSN 0899-8256. doi: 10.1006/GAME.1995.1023.

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., 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.

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.

Ostrovski, G. and van Strien, S. Payoff performance of fictitious play. *Journal of Dynamics and Games*, 1(4):621–638, 8 2014. ISSN 21646074. doi: 10.3934/jdg.2014.1.621. URL http://arxiv.org/abs/1308.4049.

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.

Perrin, S., Perolat, J., Lauriere, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13199–13213. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/995ca733e3657ff9f5f3c823d73371e1-Paper.pdf.

Piliouras, G. Catastrophe by design in population games: Destabilizing wasteful locked-in technologies. *Web and Internet Economics LNCS 12495*, pp. 473, 2020.

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.

Schäfer, L., Christianos, F., Hanna, J., and Albrecht, S. V. Decoupling exploration and exploitation in reinforcement learning. In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021. URL https://openreview.net/forum?id=NwuIIOcznYt.

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

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.

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

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.

Vlatakis-Gkaragkounis, E.-V., Flokas, L., Lianeas, T., Mertikopoulos, P., and Piliouras, G. No-Regret Learning and Mixed Nash Equilibria: They Do Not Mix. *Advances in Neural Information Processing Systems*, 33:1380–1391, 2020.

A. Introduction

Learning in games requires agents to explore their state space. Whilst it is known that the rate of exploration impacts the behaviour of the learning dynamic (Tuyls et al., 2006), understanding precisely this impact often relies on placing restrictions on the structure of the interaction. A primary example of this are network zero-sum games, which model perfect competition between agents. Here it is known that, without exploration, agents cannot reach an equilibrium (Mertikopoulos et al., 2018; Ganesh Nagarajan et al., 2020), whilst arbitrarily small exploration leads to convergence (Leonardos et al., 2021). Similarly studied classes of games include potential games (Leonardos & Piliouras, 2022) and games with two players and two actions (Kianercy & Galstyan, 2012).

Beyond these restrictive classes however, the picture is much less clear. Whilst recent work has shown that high exploration rates always lead to convergence (Hussain et al., 2023), it is also well known that learning dynamics often display complex behaviours, including limit cycles (Mertikopoulos et al., 2018; Galla, 2011) and chaos (Bielawski et al., 2021; Galla & Farmer, 2013; Sanders et al., 2018; Mukhopadhyay & Chakraborty, 2020; Sato et al., 2002). The latter case represents a particularly strong barrier towards understanding the asymptotic behaviour of learning. In short, the prospect of convergence to an equilibrium cannot be taken for granted. Unfortunately, little is known about the impact of exploration in the face of non-convergent dynamics.

Yet understanding exploration remains an important endeavour as it allows agents to avoid suboptimal, or potentially unsafe areas of their state space (Leonardos & Piliouras, 2022; Bura et al., 2022; Bai et al., 2021). In addition, it is empirically known that the choice of exploration rate impacts the expected total reward (Cai et al., 2020). This has lead to an increased interest in studying the influence of exploration on learning in single-agent settings (Schäfer et al., 2021; Piliouras, 2020). It is then paramount to understand the effect of exploration also in *multi-agent environments* outside the restrictions of potential, or network zero-sum settings. This leads us to our central question:

How does exploration affect reinforcement learning dynamics in arbitrary games, even if convergence to an equilibrium cannot be guaranteed?

Main Contributions To answer this question, we consider the *smooth Q-Learning* (SQL) dynamic, a learning dynamic which quantifies the tendency for agents to explore their state space whilst also seeking to maximise their payoffs.

We then lift the assumption of convergence to an equilibrium. In doing so, we show that, in network games with unique Nash Equilibria, SQL remains contained in a neighbourhood of an equilibrium for any positive exploration rate. The effect of increased exploration is to decrease the size of this neighbourhood. As such, exploration rates can be tuned to control the 'degree' of convergence of SQL. By considering convergence to a set, we are able to include several dynamical behaviours, including convergence to a fixed point, but also to a cycle or chaotic behaviour. In addition, this greatly widens the class of games in which the effect of exploration can be understood.

Next, we analyse the effect of exploration on system performance. Namely, we show that, by increasing exploration, the ability for agents to improve their payoffs whilst following SQL decreases. We exemplify this by considering a class of games in which Q-Learning does not converge for low exploration rates, but non-convergent SQL dynamics *outperform* the equilibrium, in terms of the sum of payoffs of all agents. Therefore, whilst increasing exploration drives the system towards convergence, the payoff performance of SQL suffers. Our experiments show this to be a general property of SQL dynamics. As a result, when system performance is desirable, it may be beneficial to restrict exploration, even if the dynamics do not converge to an equilibrium.

Related Work Since its conception, the smooth Q-Learning dynamic (Tuyls et al., 2006; Sato & Crutchfield, 2003) has received significant attention as it provides a fundamental model of exploration in multi-agent learning and is closely related to the well-studied *replicator dynamics* (Maynard Smith, 1974; Hofbauer & Sigmund, 1998; Leonardos & Piliouras, 2022), as well as the popular Q-Learning algorithm (Sutton & Barto, 2018; Schwartz, 2014). Interestingly, a number of studies have shown that, for various choices of exploration rates, the Q-Learning dynamic can display a number of complex behaviours, as limit cycles (Mertikopoulos et al., 2018; Kleinberg et al., 2011; Hofbauer, 1996) or even chaos (Galla & Farmer, 2013; Sato et al., 2002; Ganesh Nagarajan et al., 2020). To make matters worse, as the number of players increases the likelihood of these complex behaviours increases (Sanders et al., 2018). In fact, when there is no exploration on the part of every agents, the Q-Learning dynamics cannot converge to an interior equilibrium (Vlatakis-Gkaragkounis et al., 2020). This behaviour is not limited to the Q-Learning dynamics; a wide array of online learning algorithms are known to exhibit chaos, including *Fictitious Play* (van Strien & Sparrow, 2011; Ewerhart & Valkanova, 2020) and *Follow the Regularised Leader*

(Bielawski et al., 2021; Andrade et al., 2021; Anagnostides et al., 2022; Cheung & Tao, 2021).

The presence of these non-convergent behaviours result in an inherent challenge in understanding the effect of exploration on multi-agent learning (Klos et al., 2010). As a result studies considering this topic from a theoretical standpoint are often limited to understanding equilibrium behaviours. For instance, (Leonardos & Piliouras, 2022) shows that different exploration rates can result in different stability properties, which can lead to potentially unbounded gains and losses in system performance. Similar phase transitions are also found in (Kaisers & Tuyls, 2011; Kianercy & Galstyan, 2012; Piliouras, 2020). By contrast, (Hussain et al., 2023) shows that sufficiently high exploration rates yields convergence of the Q-Learning dynamic to a unique equilibrium, regardless of the number or stability of Nash Equilibria in the game. To our knowledge ours is the first study to analyse, theoretically and empirically, the effect of exploration in arbitrary multi-agent settings in terms of dynamics and system performance, *without* assuming convergence to an equilibrium.

In this paper we study a network game $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$, where $\mathcal{N}$ denotes a finite set of agents. Each agent $k \in \mathcal{N}$ has a finite set S_k of actions which are indexed by $i = 1, \dots, n_k$, and can play a mixed strategy $\mathbf{x}_k$, i.e. a discrete probability distribution over its set of actions. The set of all such mixed strategies is the unit simplex in $\mathbb{R}^{n_k}$, i.e. $\Delta_k = \{\mathbf{x}_k \in \mathbb{R}^{n_k} \mid \sum_{i\in S_k} x_{ki} = 1, \text{ and } x_{ki} \geq 0 \text{ for all } i \in S_k\}$. We denote with $\Delta = \times_{k\in\mathcal{N}}\Delta_k$ the joint simplex over all agents, with $\mathbf{x} = (\mathbf{x}_k)_{k\in\mathcal{N}}$ the joint mixed strategy of all agents and, for any k, with $\mathbf{x}_{-k} = (\mathbf{x}_l)_{l\in\mathcal{N}\setminus\{k\}} \in \Delta_{-k}$ the joint strategy of all agents other than k.

Agents in a network game interact according to an edgeset $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. Each edge corresponds to a bimatrix game (A^{kl}, A^{lk}). The expected payoff for each agent $k \in \mathcal{N}$ who plays mixed strategy $\mathbf{x}_k$ against joint mixed strategy $\mathbf{x}_{-k}$ is given by

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

Also associated to each agent k is a payoff function $u_k : \Delta_k \times \Delta_{-k} \to \mathbb{R}$. Then, for any $\mathbf{x} \in \Delta$, the reward to agent k when they play action $i \in S_k$ as $r_{ki}(\mathbf{x}) := \partial u_{ki}(\mathbf{x})/\partial x_{ki}$. With this, we can write $r_k(\mathbf{x}) = (r_{ki}(\mathbf{x}))_{k\in\mathcal{N}}$ as the concatenation of all rewards to agent k. In this notation, $u_k(\mathbf{x}) = \mathbf{x}_k^\top r_k(\mathbf{x})$. Using this notation, we can define the equilibrium of the game.

Definition A.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 $i \in S_k$,

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

In the definition of the QRE, T_k denotes the *exploration rate* of the agent. Note that as $T_k \to \infty$, the QRE is unique and is given by $\bar{\mathbf{x}} = (\frac{1}{n_k}\mathbf{1})_{k\in\mathcal{N}}$. This equilibrium corresponds to the case in which each agent plays each action with the same probability, regardless of the payoff received. In the other limit, $T_k \to 0$, the QRE corresponds to the Nash Equilbrium, which we now define.

Definition A.2 (Nash Equilibrium (NE))**.** A joint mixed strategy $\bar{\mathbf{x}} \in \Delta$ is an *Nash Equilibrium* if, for all agents k and all $\mathbf{x}_k \in \Delta_k$,

$$\langle \mathbf{x}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle \leq \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k}) \rangle \tag{8}$$

Informally, at a Nash Equilibrium, no agent can increase their utility by means of unilateral deviations, i.e. agents are considered perfectly rational. With this in mind, the QRE can be thought of as an equilibrium notion for agents with bounded rationality.

In this work, we will be making the following assumption which will allow for a comparative analysis between non-convergent behaviours and a unique equilibrium.

Assumption A.3. The network game $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ has a unique, interior Nash Equilibrium

We can extend this assumption towards the QRE of the game using the following result.

Proposition A.4. *If a game Γ has a unique Nash Equilibrium then, for any $T_1, \dots, T_N > 0$, there exists a unique interior QRE.*

Proof. Existence of an interior QRE follows from (Leonardos et al., 2021) Theorem 3.2. The fact that the QRE is unique follows from (McKelvey & Palfrey, 1995) which states that

1. The limit of the set $\{\mathbf{x} \in \Delta \,:\, \mathbf{x} \text{ is a QRE for } T_1, \ldots, T_N\}$ as $T_k \to 0$ for all k is the set of Nash Equilibria of the game.

2. There is a unique path of QRE between the uniform distribution in the limit $T_k \to \infty$ for all k, to a unique Nash Equilibrium.

If, therefore, the game admits a unique NE, then it must admit unique QRE as, if for some $T_1, \ldots, T_N > 0$, there are multiple QRE, then each of these QRE must generate a path which leads to the unique NE. This contradicts the uniqueness of the path. □

Influence Bound Our convergence results on Q-Learning will depend on a suitable notion of *size* of the game. To define this formally, we introduce the *influence bound* (Melo, 2021) of the game which is defined as

Definition A.5 (Influence Bound)**.** A game Γ has the *influence bound* δ given by

$$\delta = \max_{k \in \mathcal{N}, i \in S_k, s_{-k}, \tilde{s}_{-k} \in S_{-k}} \{|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|\}$$

where the pure strategies $s_{-k}, \tilde{s}_{-k} \in S_{-k}$ differ only in the strategies of one agent $l \neq k$.

Since $|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|$ measures the change in reward to agent k for playing action i due to a change the other players' actions, the influence bound δ defines the maximum influence (in terms of reward) that any agent could receive from their opponents.

Remark A.6. In the case of a network game, the influence bound of the game is simply

$$\max_{k \in \mathcal{N}, i \in S_k, s_{-k}, \tilde{s}_{-k} \in S_{-k}} | \left(A^k\right)_{i, s_{-k}} - \left(A^k\right)_{i, \tilde{s}_{-k}} |.$$

In other words, it is the maximum difference between any row elements across the payoff matrices for all agents.

When considering performance, we use two closely related measures. The first is the total *exploitability*. Informally, exploitability measures an agent's ability to improve their current payoff by deviating to another strategy. More formally, we define exploitability of a strategy $\mathbf{x}$ with respect to a set $S = \times_k S_k$ as

$$\Phi_S(\mathbf{x}) = \sum_k \max_{\mathbf{y}_k \in S_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - u_k(\mathbf{x}_k, \mathbf{x}_{-k}). \tag{9}$$

Our second metric for performance is *Social Welfare*, which measures the total payoff received by all agents. Formally, the Social Welfare of a mixed strategy $\mathbf{x} \in \Delta$ is given by

$$SW(\mathbf{x}) = \sum_k u_k(\mathbf{x}_k, \mathbf{x}_{-k}) \tag{10}$$

Social Welfare is a stronger measure of performance than Exploitability as the latter considers an agent's ability to improve their payoff, whereas the former measures the realised payoff that each agent receives.

Learning Model We study a smooth variant of Q-Learning with Boltzmann exploration, called smooth Q-Learning (SQL) (Tuyls et al., 2006). This requires that each agent k updates their mixed strategy x_k according to the dynamic

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

in which $T_k \in [0, \infty)$ denotes the exploration rate of agent k. At the limit of zero exploration rates we recover the replicator dynamic in which agents maximise their payoff at every time step (Sato & Crutchfield, 2003). At the other limit, $T_k \to \infty$, we recover an entropy maximising dynamic in which the unique fixed point $\bar{\mathbf{x}} = (\frac{1}{n_k}\mathbf{1})_{k \in \mathcal{N}}$ is globally asymptotically stable. This allows us to capture, with the parameter T_k, how the exploration rate affects the dynamics.

B. Convergence and Performance

In this section, we first show how the imposition of exploration by all agents forces the learning dynamics to converge to a neighbourhood of the QRE $\mathbf{x}$. Using this, we define a lower bound on each x_{ki}, independently of whether the dynamics converge to an equilibrium or display more complex behaviour.

Convergence To define convergence, we require a measure of distance. To this end, we employ the *Kullback-Leibler* (KL) Divergence.

Definition B.1 (Kullback-Leibler Divergence)**.** The KL Divergence between a set of joint mixed strategies $\mathbf{x}, \mathbf{y} \in \Delta$ is given by

$$D_{KL}(\mathbf{y}||\mathbf{x}) = \sum_k D_{KL}(\mathbf{y}_k||\mathbf{x}_k) = \sum_{ki} y_{ki} \ln \frac{y_{ki}}{x_{ki}} \tag{11}$$

Notice that the KL-Divergence does not formally define a metric as it is not symmetric (i.e. in general $D_{KL}(\mathbf{y}||\mathbf{x}) \neq D_{KL}(\mathbf{x}||\mathbf{y})$). Rather, the KL-Divergence can be thought of as measuring the overlap between probability distributions $\mathbf{y}$ and $\mathbf{x}$. The key point which we will use in our main theorem is that $D_{KL}(\mathbf{y}||\mathbf{x})$ is zero if and only if $\mathbf{x} = \mathbf{y}$ and is positive everywhere else.

Theorem B.2. *Let δ be influence bound of the game Γ and let $\bar{\mathbf{x}} \in int\Delta$ denote the QRE of the game for some $T_1, \ldots, T_N$. Then, the Q-Learning dynamics remain asymptotically within the set*

$$S_T = \{\mathbf{x} \in \Delta \,|\, D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) \leq \frac{\delta}{T_{\min}} \sum_k n_k\}$$

To prove this, we begin with the following

Lemma B.3. *Consider a game Γ with influence bound δ. Let $\mathbf{x}(t)$ denote the joint strategy generated by (SQL) at some time t for some initial condition $\mathbf{x}(0)$. Also let $\bar{\mathbf{x}}$ denote the QRE for the game for some T_k. Then $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ is a decreasing function along trajectories of the Q-Learning dynamic (SQL) for all $t \in [0, \infty)$ such that*

$$D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\bar{\mathbf{x}}) > \frac{\delta}{T_{\min}} \sum_k n_k \tag{12}$$

Proof. First we have that

$$\begin{aligned}
&(\mathbf{x}_k - \bar{\mathbf{x}}_k)^\top [r_k(\mathbf{x}_{-k}) - r_k(\bar{\mathbf{x}}_{-k})] \\
\leq &| (\mathbf{x}_k - \bar{\mathbf{x}}_k)^\top [r_k(\mathbf{x}_{-k}) - r_k(\bar{\mathbf{x}}_{-k})] | \\
\leq &\sum_i |x_{ki} - \bar{x}_{ki}||r_{ki}(\mathbf{x}_{-k}) - r_{ki}(\bar{\mathbf{x}}_{-k})| \\
\leq &n_k \delta,
\end{aligned} \tag{13}$$

where the final inequality holds from the fact that $|x_{ki} - \bar{x}_{ki}| \in [0, 1]$ and $|r_{ki}(\mathbf{x}_{-k}) - r_{ki}(\bar{\mathbf{x}}_{-k})| \leq \delta$ due to the separability of payoffs in network games and the definition of the influence bound δ.

Next, it holds (due to (Leonardos et al., 2021) Lemma 4.1) that, for any game, the KL divergence along trajectories of QL satisfies

$$\frac{d}{dt} D_{KL}(\bar{\mathbf{x}}_k||\mathbf{x}_k(t)) = (\mathbf{x}_k(t) - \bar{\mathbf{x}}_k)^T [r_k(\mathbf{x}_{-k}) - r_k(\bar{\mathbf{x}}_{-k})] - T_k(D_{KL}(\bar{\mathbf{x}}_k||\mathbf{x}_k(t)) + D_{KL}(\mathbf{x}_k(t)||\bar{\mathbf{x}}_k)) \tag{14}$$

Combining (13) and (14) we see that $\frac{d}{dt} D_{KL}(\mathbf{p}_k||\mathbf{x}_k(t)) < 0$ when

$$n_k \delta - T_k[D_{KL}(\bar{\mathbf{x}}_k||\mathbf{x}_k(t)) + D_{KL}(\mathbf{x}_k(t)||\bar{\mathbf{x}}_k)] < 0.$$

Since $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) = \sum_k D_{KL}(\bar{\mathbf{x}}_k||\mathbf{x}_k(t))$, we have that $\frac{d}{dt}D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ is strictly negative when

$$D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\bar{\mathbf{x}}) > \frac{\delta}{T_{\min}} \sum_k n_k \tag{15}$$

□

Proof of Theorem B.2. Let S denote the set

$$S = \left\{ \mathbf{x} \in \Delta \,|\, D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\bar{\mathbf{x}}) \leq \frac{\delta}{T_{\min}} \sum_k n_k \right\}$$

For any $\mathbf{x}(t) \notin S$, we know that $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ is a decreasing function outside of the set S, and is also bounded below by zero. It follows, then, that $\mathbf{x}(t)$ reaches S in finite time, at which point $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t)) \leq \sup_{\mathbf{x}\in S} D_{KL}(\bar{\mathbf{x}}||\mathbf{x}) =: D_S$. Furthermore, the decreasing property of $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ outside of S implies that, if $\mathbf{x}(t)$ leaves S, $D_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ cannot increase further than D_S. It follows that $\mathbf{x}(t)$ must remain in the set $\{\mathbf{x} \in \Delta \,:\, D_{KL}(\bar{\mathbf{x}}||\mathbf{x}) \leq D_S\}$. Finally, we note that $D_S = \sup_{\mathbf{x}\in S} D_{KL}(\bar{\mathbf{x}}||\mathbf{x}) \leq \sup_{\mathbf{x}\in S} D_{KL}(\bar{\mathbf{x}}||\mathbf{x}) + D_{KL}(\mathbf{x}||\bar{\mathbf{x}}) \leq \frac{\delta}{T_{\min}} \sum_k n_k$. □

Remark B.4. At first glance, it appears that the size of the set defined by Theorem B.2 increases with the number of players and number of actions due to the presence of the term $\sum_k n_k$. On closer inspection, however, we see that the KL-Divergence itself is given by a summation over agents and actions. Therefore, both sides of the inequality which define S_T grow at the same rate.

Theorem B.2 determines the convergence structure of Q-Learning in arbitrary games. In particular, it finds a set in which Q-Learning remains asymptotically trapped, independently of whether the dynamics are ultimately chaotic, cyclic or converge to an equilibrium. As T_k increases, the size of this set decreases, tightening the convergence bound. On the other hand, the size of the set increases with the size of the game, as measured by the influence bound. Indeed, in the limit $T_k \to \infty$ for all k, the set defined by Theorem B.2 is a singleton, so that Q-Learning must converge to the QRE.

Another implication of Theorem B.2 is that, for any $T_1, \ldots, T_N > 0$ the Q-Learning dynamics must remain bounded away from the boundary of the simplex $\partial\Delta$ for all $t > 0$. In addition, this bound increases with T_k, resulting in the dynamics being forced further in the interior of Δ.

Corollary B.5. *In the setting of Theorem B.2, when each agent k has exploration rate $T_k \geq 0$ and follows the Q-Learning dynamic, there exists an $\epsilon_T \in [0, 1/(\min_k n_k)]$ which grows with $T_{\min} = \min_k T_k$ such that for any $k \in \mathcal{N}$, $i \in S_k$*

$$\liminf_{t\to\infty} x_{ki}(t) \geq \epsilon_T$$

Proof. From Theorem B.2, it holds that the KL-Divergence is bounded by $\delta n_k / T_{\min}$ in the limit $t \to \infty$. This immediately yields the existence of an $\epsilon_T \geq$ such that the result holds. That $\epsilon_T \geq 0$ follows from the fact that $D_{KL}(\mathbf{p}||\mathbf{x}) \to \infty$ as $\mathbf{x} \to \partial\Delta$. In words the KL Divergence approaches infinity towards the boundary of the simplex. The increase of ϵ_T with $T_{\min}$ holds due to the convexity of the KL-Divergence, and that $\bar{x}_{ki} \to 1/n_k$ as $T_{\min} \to \infty$. □

Performance Next, we examine an important implication of Theorem B.2, namely that, as the learning dynamics remain asymptotically bounded within the interior of the simplex, the *exploitability* of the system decreases. We go further by placing an upper bound on this reduction of exploitability in terms of the lower bound defined by Corollary B.5.

As all results depend on $T_{\min}$, we ease notation by assuming the exploration rates T_k for all agents are equal. Then we drop the $\min$ notation and just write T. Next, we define ϵ_T as the lower bound in Corollary B.5 for some choice of T,

Now, for any T define $\Omega_k = \times_{i\in S_k}[\epsilon_T, 1-\epsilon_T]$ and $\Omega = \times_k \Omega_k$. Then it is clear that $S_T \subset \Omega$ and, as $T \to 0$, $\Omega \to \Delta$.

With this, we can apply the definition of exploitation (9) with respect to Ω.

$$\Phi_\Omega(\mathbf{x}) = \sum_k \max_{\mathbf{y}_k \in \Omega_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - u_k(\mathbf{x}_k, \mathbf{x}_{-k}) \tag{16}$$

The motivation for defining this metric is as follows: the definition of exploitability which is most widely applied in the context of online learning corresponds to Φ_Δ (Perrin et al., 2020; Gemp et al., 2022), which measures the best payoff any agent could receive by deviating to any other strategy in the simplex, assuming that all other agents keep their strategies fixed. However, if agents are following the Q-Learning dynamic, with some positive exploration rates, the whole simplex becomes unavailable. Rather agents can only improve their strategy from within the set S_T. We can compare Φ_Ω against the case of zero exploration using $\Delta\Phi(\mathbf{x}) = \Phi_\Omega(\mathbf{x}) - \Phi_\Delta(\mathbf{x})$. As such, $\Delta\Phi$ measures the change in exploitability as exploration is introduced. To show that exploitability decreases, we begin with the following Proposition.

Proposition B.6. *Let $\epsilon_T \in [0, 1/(\min_k n_k)]$ be such that $x_{ki} \geq \epsilon_T$ for all $k \in \mathcal{N}$, $i \in S_k$. Then $\mathbf{y}_k \in \arg\max_{y_k \in \Omega_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k})$ where $\Omega_k = \times_{i \in S_k}[\epsilon_T, 1 - \epsilon_T]$ if*

$$y_{ki} = \begin{cases} 1 - (n_k - 1)\epsilon_T, & \text{if } i \in \arg\max_{i \in S_k} r_{ki}(\mathbf{x}_{-k}) \\ \epsilon_T, & \text{otherwise.} \end{cases}$$

Proof. The proof begins with the fact that, as $\epsilon_T \in [0, 1/n_k]$ for all $k \in \mathcal{N}$, it follows that $1 - (n_k - 1)\epsilon_T > \epsilon_T$, for all k. Next, we write the maximisation problem as

$$\begin{cases} & \max_{\mathbf{y}_k \in S_k} \mathbf{y}_k^\top r_k(\mathbf{x}_{-k}) \\ \text{s.t.} & y_{ki} \geq \epsilon_T \text{ for all } i \in S_k \\ & \sum_i y_{ki} = 1 \end{cases}$$

Which has Lagrangian

$$L(\mathbf{y}_k, \mu_k, \rho_k) = \sum_{i \in S_k} y_{ki} r_{ki}(\mathbf{x}_{-k}) - \mu_{ki}(\epsilon_T - y_{ki}) + \rho_k(1 - \sum_i y_{ki})$$

where $\mu_{ki} \geq 0$ is such that $\mu_{ki}(\epsilon_T - y_{ki}) = 0$ for all k, i.

Let $j = \arg\max_{j \in S_k} r_{kj}(\mathbf{x}_{-k})$. From the KKT conditions, and the form of the objective function it holds that a maximiser satisfies $y_{ki} = \epsilon_T$ for all $i \neq j$ and $y_{kj} = 1 - (n_k - 1)\epsilon_T$. □

Theorem B.7. *Let Γ be a game with a unique, interior Nash Equilibrum. Then, for any $T \geq 0$, $\Delta\Phi(\mathbf{x}) \leq 0$ with equality holding iff $\mathbf{x}$ is the NE of the game, denoted $\bar{\mathbf{x}}$. In addition, for all $\mathbf{x} \neq \bar{\mathbf{x}}$, $\Delta\Phi(\mathbf{x}) \leq -\alpha\epsilon_T < 0$, for some $\alpha > 0$ i.e. exploitability decreases as T increases.*

Proof. Let $\bar{\mathbf{y}}_k$ be the maximiser of the function $u_k(\mathbf{y}_k, \mathbf{x}_{-k})$ on Ω_k, i.e. it is the *exploration-induced* best response of agent k to the strategy $\mathbf{x}_{-k}$. Furthermore, as we can write any $\mathbf{y}_k$ as a convex combination of strategies i, we can write $\langle \bar{\mathbf{y}}_k, r_{ki}(\mathbf{x}_{-k}) \rangle = \sum_{i \in S_k} \lambda_{ki} r_{ki}(\mathbf{x}_{-k})$. Finally, let $j \in S_k$ satisfy $j = \arg\max_{i \in S_k} r_{ki}(\mathbf{x}_{-k})$.

$$\begin{aligned} \max_{\mathbf{y}_k \in \Omega_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) - \max_{\mathbf{y}_k \in \Delta_k} u_k(\mathbf{y}_k, \mathbf{x}_{-k}) &= \left(\sum_{i \in S_k} \lambda_{ki} r_{ki}(\mathbf{x}_{-k}) \right) - r_{kj}(\mathbf{x}_{-k}) \\ &= \left(\sum_{i \neq j} \lambda_{ki} r_{ki}(\mathbf{x}_{-k}) \right) + \left(1 - \sum_{i \neq j} \lambda_{ki} \right) r_{kj}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k}) \\ &= \sum_{i \neq j} \lambda_{ki} \left(r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k}) \right) \end{aligned} \tag{17}$$

At the Nash Equilibrium, which we have assumed to be unique and interior, we have that $r_{ki}(\bar{\mathbf{x}}_{-k}) - r_{kj}(\bar{\mathbf{x}}_{-k}) = 0$ for all i. At all other points $\mathbf{x}$, we have that $r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k}) < 0$ for all i, except for on a finite number of hyperplanes on which $r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k}) = 0$ for some i and $r_{km}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k}) < 0$ for all other $m \in S_k$. In addition, due to the construction of Ω_k, $\lambda_{ki} \geq \epsilon_T > 0$ for all k, i. By taking the sum over all k, i, we have that

$$\Delta\Phi(\mathbf{x}) = \Phi_\Omega(\mathbf{x}) - \Phi_\Delta(\mathbf{x}) < 0$$

To show the second part of the Theorem, we use Proposition B.6 which states that, when playing a best response $\bar{\mathbf{y}}_k$ within Ω_k, each agent places as little weight as possible on suboptimal actions, and all remaining weight on (one of) the best performing action j. In the case that the $\arg\max$ is not single valued, $\bar{\mathbf{y}}_k$ will produce the same reward as any other exploration induced best response. As such, we can write

$$\sum_{i\neq j} \lambda_{ki} \left(r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k})\right) = \epsilon_T \sum_{i\neq j} \left(r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k})\right) =: -\alpha_k \epsilon_T < 0.$$

where

$$\alpha_k = \sum_{i\neq j} \left(r_{ki}(\mathbf{x}_{-k}) - r_{kj}(\mathbf{x}_{-k})\right) \geq 0$$

Taking the sum over all agents k, and defining $\alpha = \sum_k \alpha_k$ we get the final result.

□

Discussion on Results Theorem B.7 shows that the ability of an agent to improve their payoffs strictly decreases as exploration increases. By contrast, Theorem B.2 shows that higher exploration rates leads to a greater certainty in convergence of Q-Learning. In summary, we show that stronger guarantees on the convergence due to exploration come at the price of decreased system performance.

Whilst Theorems B.2 and B.7 paint a broad stroke on convergence and performance, a limitation is that Theorem B.2 does not have anything to say about the behaviour of Q-Learning within the defined set. Indeed it may be the case, as it is for network zero-sum games, that Q-Learning converges to a QRE for all $T_k > 0$. Similarly, whilst Lemma B.7 shows that agents cannot improve their payoffs as exploration increases, it does not show that their payoffs decrease as the agents move from non-convergent to convergent behaviours.

C. Exploration Reduces Payoff Performance

In the previous sections, we showed that exploration leads to a decreased ability of each agent to improve their payoff. This is a strong statement on the tradeoff between convergence and performance.

In this section we tackle the limitations discussed in the previous section. To do this we focus a specific class of games for which we show that non-convergent behaviours strictly outperform convergence. In this case, performance is measured through social welfare so that the agents' realised payoffs are considered, rather than their *ability* to improve their payoffs.

Shapley Network Game In the first example we examine a network of agents, where each edge is equipped with a Shapley game. In particular, the payoff to each agent k is given by

$$u_k = u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \mathbf{x}_k \mathbf{A} \mathbf{x}_{k-1} + \mathbf{x}_k \mathbf{B}^\top \mathbf{x}_{k+1}$$
$$A = \begin{pmatrix} 1 & 0 & \beta \\ \beta & 1 & 0 \\ 0 & \beta & 1 \end{pmatrix}, B = \begin{pmatrix} -\beta & 1 & 0 \\ 0 & -\beta & 1 \\ 1 & 0 & -\beta \end{pmatrix},$$

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

In the two agent case, (Shapley, 2016) showed that the popular *Fictitious Play* dynamics (Brown P, 1949; Hofbauer & Sigmund, 2003) do not converge to an NE, but rather reach a limit cycle. In (Ostrovski & van Strien, 2014), the authors show that the non-convergent cycle outperforms the Nash Equilibrium. In (Hussain et al., 2023), the multi-agent extension was experimentally examined and it was suggested that the performance of Q-Learning decreases as the system moves from a limit cycle to an equilibrium. We make this statement rigorous by showing that the dynamics, whilst initially non-convergent, can be made convergent through a sufficiently high exploration rate. However, we find that the result is accompanied by a strict decrease in the social welfare along trajectories.

Lemma C.1. *For any $\beta \in (0,1)$ and any $T_k \geq 0$, the Network Shapley game has a QRE at the uniform distribution $\bar{\mathbf{x}} = (\frac{1}{3})_{k\in\mathcal{N}, i\in S_k}$ which is globally repelling under (SQL) at $T_k = 0$ and locally attracting if, for all $k \in \mathcal{N}$, $T_k > 1 + \frac{\beta}{3}$. The QRE is globally asymptotically stable if $T_k > (N-1)(1+\beta)$.*

Proof. For ease of presentation, we break the proof statement into two separate components. First we show that $\bar{\mathbf{x}} = (\frac{1}{3})_{k\in\mathcal{N}, i\in S_k}$ is indeed a QRE, then address its stability.

QRE The rewards to each agent k in the Network Shapley Game are given by

$$r_k(\mathbf{x}_{-k}) = A\mathbf{x}_{k-1} + B^\top \mathbf{x}_{k+1}$$

At $\bar{\mathbf{x}}$, for each agent k

$$r_k(\bar{\mathbf{x}}_{-k}) = \begin{pmatrix} (1/3)(1+\beta) + (1/3)(1-\beta) \\ (1/3)(1+\beta) + (1/3)(1-\beta) \\ (1/3)(1+\beta) + (1/3)(1-\beta) \end{pmatrix}$$

So that, for all k, $\bar{\mathbf{x}}_k \in \arg\max_{\mathbf{x}_k \in \Delta_k} u_k(\mathbf{x}_k, \bar{\mathbf{x}}_{-k})$. Therefore, $\bar{\mathbf{x}}$ is a Nash Equilibrium of the Network Shapley Game. To show that it is also a QRE, we show that it is a fixed point of the Q-Learning dynamics (SQL). Since, by (Leonardos et al., 2021), fixed points of (SQL) coincide with the QRE of the game, the result follows.

As $\bar{\mathbf{x}}$ is an interior Nash Equilibrium, for all k and all $i \in S_k$, $r_{ki}(\bar{\mathbf{x}}_{-k}) = \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}_{-k})\rangle$. Furthermore, for all $i, j \in S_k$, $\ln x_{kj}/x_{ki} = \ln 1 = 0$. Therefore, all terms on the right hand side of (SQL) is zero.

Stability To find the local stability, we begin by following (Hofbauer, 1996; Sato & Crutchfield, 2003) which considered the replicator dynamics and q-learning respectively in two-player games. We make the following transformation of variables.

$$\alpha_{ki} = \ln \frac{x_{k,i+1}}{x_{k1}},\ i = 1, \dots, n_k - 1$$

Note that this α has no relation to that used in Theorem B.7. Applying this transformation to (SQL) on the Network Shapley Game yields the dynamics

$$\dot{\alpha}_{ki} = \frac{\sum_{j=1}^{n_k-1} \tilde{a}_{ij} e^{\alpha_{k-1,j}} + \tilde{a}_{1j}}{1 + \sum_{j=1}^{n_k-1} e^{\alpha_{k-1,j}}} + \frac{\sum_{j=1}^{n_k-1} \tilde{b}_{ij} e^{\alpha_{k+1,j}} + \tilde{b}_{1j}}{1 + \sum_{j=1}^{n_k-1} e^{\alpha_{k+1,j}}} - T_k \alpha_{ki}$$

in which $\tilde{a}_{ij} = (A)_{i+1,j} - (A)_{1,j}$ and $\tilde{b}_{ij} = (B^\top)_{i+1,j} - (B^\top)_{1,j}$.

As the transformation of variables is a homeomorphism, stability results which hold for the dynamics of α hold also for $\mathbf{x}$. In the transformed system, the QRE $\bar{\mathbf{x}}$ maps to $\bar{\alpha}_{ki} = 0$ for all k, i. The Jacobian at the QRE takes the form

$$J = \begin{pmatrix} T_1\mathbf{I}_{2\times 2} & \mathbf{f}_+ & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & \mathbf{f}_- \\ \mathbf{f}_- & T_2\mathbf{I}_{2\times 2} & \mathbf{f}_+ & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{f}_+ & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & T_N\mathbf{I}_{2\times 2} & \mathbf{f}_+ \end{pmatrix}$$

$$\mathbf{f}_- = \frac{\partial \dot{\alpha}_k}{\partial \alpha_{k-1}} = \frac{1}{9}\begin{pmatrix} \beta - 2 & 4 - 2\beta \\ -\beta - 1 & 2\beta + 2 \end{pmatrix}$$

$$\mathbf{f}_+ = \frac{\partial \dot{\alpha}_k}{\partial \alpha_{k+1}} = \frac{1}{9}\begin{pmatrix} 2\beta + 1 & -4\beta - 2 \\ \beta - 1 & 2 - 2\beta \end{pmatrix}$$

Eigenvalues analysis of J at $T_k = 0$ yields a positive eigenvalue of $(\beta+1)/3$. Therefore, the system is unstable, i.e. locally repelling. This result can also be achieved by recognising that, at $T_k = 0$, (SQL) corresponds to the replicator dynamics which is a sub-class of the *Follow the Regularised Leader* dynamics (Mertikopoulos & Sandholm, 2016). In (Vlatakis-Gkaragkounis et al., 2020), it is shown that interior fixed points for this dynamic are always unstable. The result follows immediately. To show local stability, we apply Gershgorin's disc theorem

Theorem C.2 (Gershgorin). *For an $n \times n$ matrix A, let*

$$R_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|. \tag{18}$$

Then every eigenvalue λ of A lies in one of the circles

$$\{z : |z - a_{ii}| \leq R_i\} \tag{19}$$

Using this, and the fact that T_k lies on the diagonal of J, it follows that the system has all negative eigenvalues if $T_k > \max\{(1+\beta)/3, 2/3\}$, where the terms in the max function arise by summing the off diagonal terms in each row of J. The result then follows. Global asymptotic stability can be shown due to (Hussain et al., 2023) and the fact that the influence bound of the Shapley game is $1 + \beta$. □

Theorem C.3. *In Network Shapley Games, let the* time-average social welfare *(TSW) along Q-Learning trajectories be defined as*

$$TSW = \lim_{t\to\infty} \frac{1}{t} \int_0^t SW(\mathbf{x}(s))\, ds \tag{20}$$

where $\mathbf{x}(t)$ is a trajectory of mixed strategies generated according to the Q-Learning dynamic for some initial condition $\mathbf{x}_0 \in \Delta$. Then, non-convergent trajectories of Q-Learning strictly outperform the social welfare of the unique QRE $\bar{\mathbf{x}} \in \Delta$. In particular, $TSW \geq SW(\bar{\mathbf{x}})$ with equality holding if and only if $\lim_{t\to\infty} \mathbf{x}(t) = \bar{\mathbf{x}}$.

Proof. As Q-Learning is an *no-regret* learning dynamic (Leonardos & Piliouras, 2022), it holds that $R = 0$ where

$$R = \lim_{t\to\infty} \sum_k \frac{1}{t} R_k(t) = \lim_{t\to\infty} \sum_k \frac{1}{t} \int_0^t \max_{\mathbf{x}'_k \in \Delta_k} [u_k(\mathbf{x}'_k, \mathbf{x}_{-k}(s)) - u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds$$

where R_k is the *regret* of each agent k. Next, we note that the reasoning of (Ostrovski & van Strien, 2014) extends to the Network Shapley Game. In particular, for all agents k and all $\mathbf{x} \in \Delta$

$$\begin{aligned}
\max_{\mathbf{x}'_k \in \Delta_k} u_k(\mathbf{x}'_k, \mathbf{x}_{-k}) &= \max_{i \in S_k} r_{ki}(\mathbf{x}_{-k}) \\
&= \max\{(A\mathbf{x}_{k-1})_1 + (\mathbf{x}_{k+1}^\top B)_1, (A\mathbf{x}_{k-1})_2 + (\mathbf{x}_{k+1}^\top B)_2, (A\mathbf{x}_{k-1})_3 + (\mathbf{x}_{k+1}^\top B)_3\} \\
&\geq \frac{1}{3}(1+\beta) + \frac{1}{3}(1-\beta) = u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k})
\end{aligned}$$

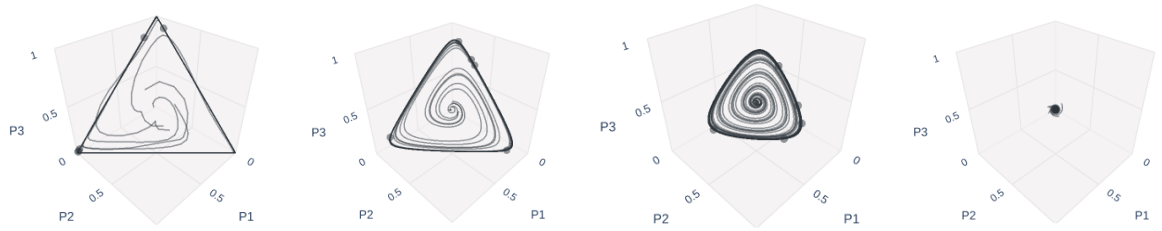
with equality holding if and only if $\mathbf{x} = \bar{\mathbf{x}}$. With this it holds that

$$\begin{aligned}
\frac{1}{t} R_k(t) &= \frac{1}{t} \int_0^t \max_{\mathbf{x}'_k \in \Delta_k} [u_k(\mathbf{x}'_k, \mathbf{x}_{-k}(s)) - u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds \\
&\geq \frac{1}{t} \int_0^t [u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k}) - u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds \\
&= u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k}) - \frac{1}{t} \int_0^t u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds \\
\implies \frac{1}{t} \int_0^t u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds &\geq u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k}) - \frac{1}{t} R_k(t) \\
\implies \lim_{t\to\infty} \frac{1}{t} \int_0^t u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))]\, ds &\geq u_k(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_{-k}) - \lim_{t\to\infty} \frac{1}{t} R_k(t) \\
\implies TSW &\geq SW(\bar{\mathbf{x}})
\end{aligned}$$

where equality holds if and only if $\lim_{t\to\infty} \mathbf{x}(t) = \bar{\mathbf{x}}$, i.e. if the trajectory converges to the QRE. □

D. Experiments on Convergence and Performance

Our experiments further analyse the phenomenon that we observe in our results - namely that increased exploration results in Q-Learning dynamics asymptotically reaching a set in the interior of the simplex, whose size decreases with T. We also analyse the effect on Social Welfare, to test whether the phenomenon shown for the Network Shapley Game, namely that exploration leads to a decrease in payoff performance, holds more generally.



$T = 0.1, TSW = 3.22$ $T = 0.2, TSW = 2.71$ $T = 0.25, TSW = 2.34$ $T = 0.3, TSW = 2.0$

Figure 5. Trajectories of Q-Learning in the Network Shapley Game, for $\beta = 0.2$ considered in Section C with five agents alongside experimentally obtained TSW. The trajectories plot the probability with which three agents play their first action. For low values of T, the system reaches a limit cycle. The size of this limit cycle decreases as exploration enforces the dynamics further into the interior of the simplex. Eventually, at $T = 0.3$, the system equilibriates at the uniform distribution. All non convergent dynamics strictly outperform the QRE.

Network Shapley Game In Figure 5 we visualise the effect of exploration on the Network Shapley game, examined in Section C. We generate a network game with five players and run Q-Learning on the game. To be able to visualise the trajectory, we select three agents and plot their first action on the space $[0, 1]^3$. In Figure 5, we keep β fixed at 0.2, which yields a fixed $\delta = 1 + \beta = 1.2$. This process is repeated for increasing choices of T. TSW is calculated as the final time averaged social welfare after running Q-Learning for 1×10^5 iterations.

It can be seen that, for small values of T, Q-Learning does not converge, but rather reaches a limit cycle. As predicted by Theorem B.2, the size of this limit cycle decreases as T increases, until eventually Q-Learning converges asymptotically to the QRE at the uniform distribution. Furthermore, as predicted by Theorem C.3, all non-convergent behaviours strictly outperform the Social Welfare of the uniform distribution which, for our choice of β, is 2.

Network Chakraborty Game Next, we consider a class of two-action network games which we call the *Network Chakraborty* game. In this game, each agent responds only to the 'previous' agent in a circular chain. More formally, the payoff to each agent k is

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

This game was analysed in (Pandit et al., 2018) under the context of an infinitely large population of agents with identical payoffs. It was shown that, for certain combinations of U, V, a discrete time analog of replicator dynamics shows chaotic behaviour. Here, we extend this game to the multi agent case and analyse it under Q-Learning.

The results of this investigation are presented in Figure 6. We examine a 15 agent game with $U = 7.0, V = 8.5$, which appears in (Pandit et al., 2018) as a case which shows chaos by discrete replicator. We run Q-Learning, with 100 initial conditions, on this system for 1×10^5 iterations and isolate the final 25,000 iterations. By examining this window, we include the possibility of complex asymptotic behaviours. In fact, we visualise the trajectories in Figure 7 and find that, in a three agent network, the dynamics reach a limit cycle. The boxplot depicts, for each agent the probability of choosing their first action taken across the entire 25,000 final iterations, and all initial conditions. Beside this, we plot the TSW achieved by Q-Learning, taken across all initial conditions.

Once again it is clear that, for small exploration rates, the dynamics do not converge to an equilibrium. Rather, the asymptotic behaviour is spread across the entire state space. As exploration increases, the spread of probability distributions

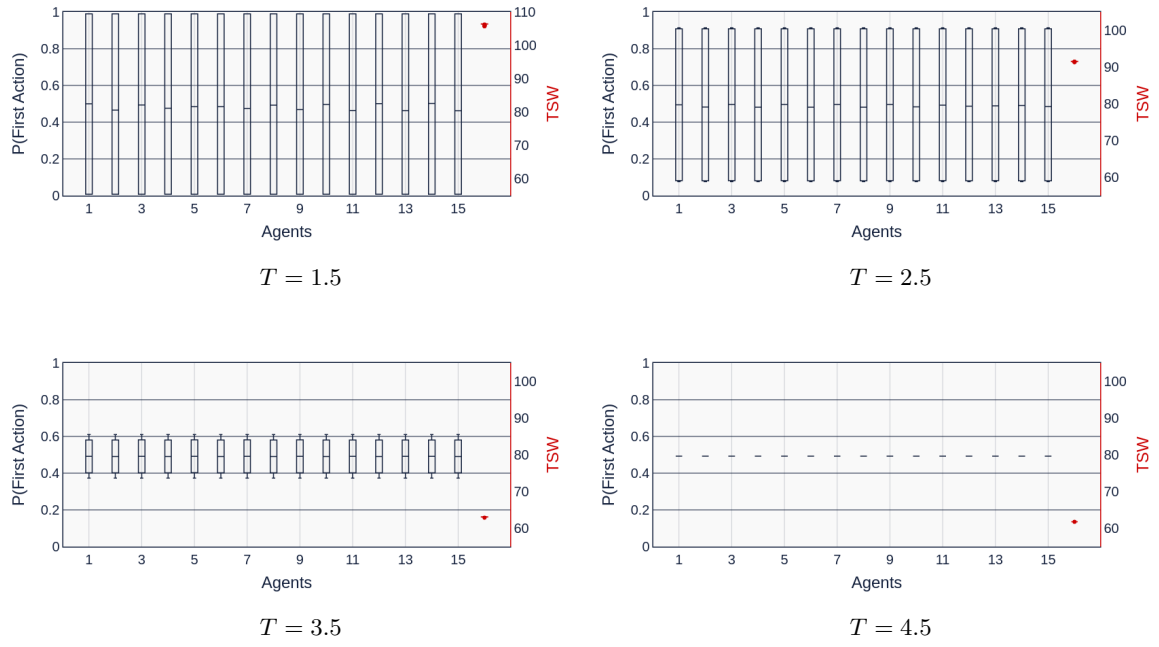


Figure 6. Asymptotic Behaviour and Performance of Q-Learning in the Network Chakraborty Game, for $U = 7.0, V = 8.5$ with 15 agents. Boxplots show the spread of probabilities with which agents play their first action in the final 25% of 1×10^5 iterations of learning. The red plot shows the TSW asymptotically achieved. For low values of T, the asymptotic dynamics are spread across the entire simplex, whilst achieving a high TSW. As T increases, the dynamics eventually reach a fixed point, but consistently decrease TSW as a result.

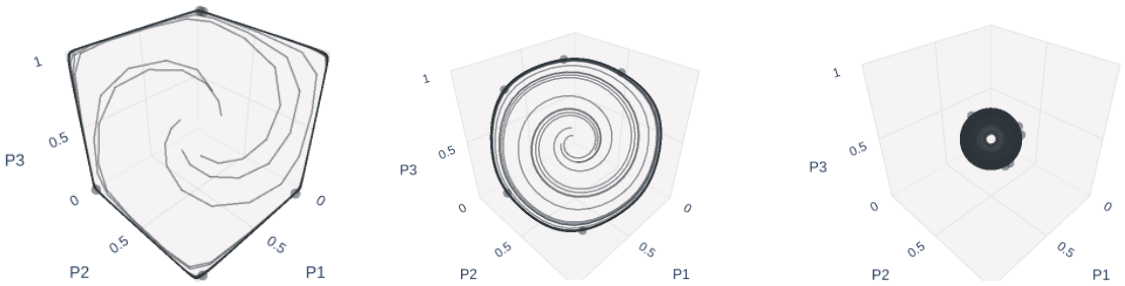
is bounded within the interior of the simplex, until eventually the dynamics equilibriate when $T = 4.5$. This process is again accompanied by a decrease in TSW achieved by Q-Learning, which reaches its minimum when the dynamics converge to an equilibrium.

Arbitrary Games Finally, we look to extend our investigation beyond specific classes of games. To do this, we analyse the effect of exploration in randomly generated games, which do not follow any specific payoff or network structure. To ensure like comparison, we generate payoffs that are positive and upper bounded. This ensures that the difference in payoffs between two randomly drawn games are not so significant as to affect plotting the results. Neither assumption impacts the generality of the results as the dynamics of Q-Learning are invariant to additions and multiplications by positive constants to all elements of the payoffs. As such, positivity can be ensured by simply adding the minimum payoff to all matrices so that the new minimum is at least zero. Similarly, boundedness can be enforced by dividing all payoff elements by a positive constant. As far as the Q-Learning dynamics are concerned, multiplication by a positive constant is equivalent to a rescaling of the exploration rate T. Recall, however, that the important factor is the effect of *increasing* exploration rates, rather than the absolute value. Finally, since all operations are being applied in the same manner to all payoff matrices, the Social Welfare is adjusted uniformly across the simplex, and the location of equilibria once again only affected up to a rescaling in T. We make these statements rigorous through the following propositions, the first of which mirrors a well known result in the replicator dynamics (Hofbauer & Sigmund, 1998).

Proposition D.1. *Consider two network games* $\Gamma_1 = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ *and* $\Gamma_2 = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (\tilde{A}^{kl}, \tilde{A}^{lk})_{(k,l)\in\mathcal{E}})$ *in which, for all* $(k,l) \in \mathcal{E}$

$$(\tilde{A}^{kl})_{ij} = c(A^{kl})_{ij} + d$$

where $c, d > 0$ *for all* i*. Then, for any initial condition* $\mathbf{x}_0 \in \Delta$*, a trajectory of (SQL) in* Γ_1 *for some* $T > 0$ *is equivalent to a trajectory of (SQL) in* Γ_2 *for* $cT > 0$*.*



$T = 0.5$ $T = 1.5$ $T = 2.0$

Figure 7. Trajectories of Q-Learning dynamics in the Network Chakraborty Game for $U = 7.0, V = 8.5$ and three agents. Trajectories show the probability with which the first action is selected. For low T, the dynamics cycle on the boundary of the simplex, leading to a large variation in each strategy component in the asymptotic limit cycle. As T increases, the size of this cycle decreases, resulting in a smaller variation of strategy components.

Proof. For any $k \in \mathcal{N}$ and any $\mathbf{x} \in \Delta$

$$\begin{aligned}
\sum_{(k,l)\in\mathcal{E}} \left(\tilde{A}^{kl}\mathbf{x}_l\right)_i &= \sum_{(k,l)\in\mathcal{E}} \sum_j \left(\tilde{A}^{kl}\right)_{ij} x_{lj} \\
&= \sum_{(k,l)\in\mathcal{E}} \sum_j (c\left(A^{kl}\right)_{ij} + d) x_{lj} \\
&= c \sum_{(k,l)\in\mathcal{E}} \sum_j \left(A^{kl}\right)_{ij} x_{lj} + d \sum_j x_{lj} \\
&= c \sum_{(k,l)\in\mathcal{E}} \sum_j \left(A^{kl}\right)_{ij} x_{lj}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{(k,l)\in\mathcal{E}} \mathbf{x}_k^\top \tilde{A}^{kl}\mathbf{x}_l &= \sum_{(k,l)\in\mathcal{E}} \sum_{ij} \left(\tilde{A}^{kl}\right)_{ij} x_{ki}x_{lj} \\
&= \sum_{(k,l)\in\mathcal{E}} \sum_{ij} (c\left(A^{kl}\right)_{ij} + d) x_{ki}x_{lj} \\
&= c \sum_{(k,l)\in\mathcal{E}} \sum_{ij} \left(A^{kl}\right)_{ij} x_{ki}x_{lj} + d \sum_i x_{ki} \sum_j x_{lj} \\
&= c \sum_{(k,l)\in\mathcal{E}} \sum_{ij} \left(A^{kl}\right)_{ij} x_{ki}x_{lj}.
\end{aligned}$$

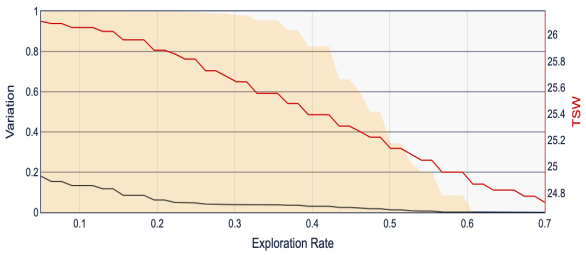
Let r_k denote the rewards in Γ_1. Then (SQL) in Γ_2 takes the form

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

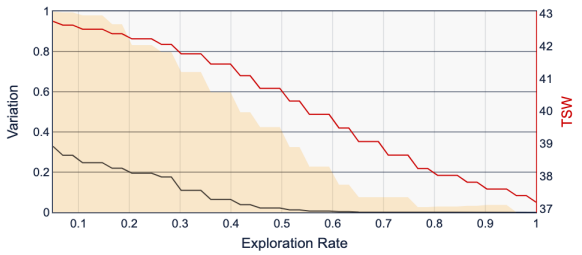
Dividing by c yields the desired result, up to a rescaling in time. □

Proposition D.2. *Let Γ_1, Γ_2 be network games in the setting of Proposition D.1. Then*

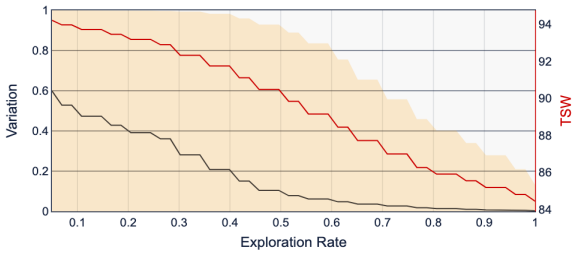
1. *for any T, $\bar{\mathbf{x}} \in \Delta$ is a QRE of Γ_1 if and only if it is a QRE of Γ_2 for cT.*
2. *for any $\mathbf{x}, \mathbf{y} \in \Delta$, $c(SW_1(\mathbf{x}) - SW_1(\mathbf{y})) = SW_2(\mathbf{x}) - SW_2(\mathbf{y})$ where SW_1 (resp. SW_2) is the social welfare determined in Γ_1 (resp. Γ_2).*



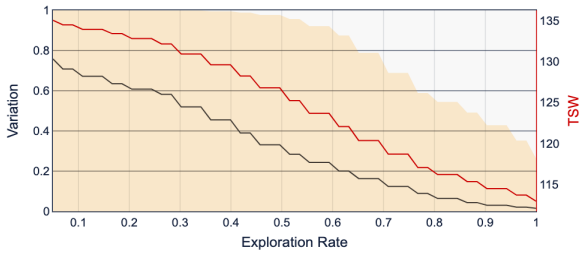
$N = 5, n_k = 3$



$N = 7, n_k = 5$



$N = 15, n_k = 5$



$N = 20, n_k = 8$

Figure 8. Convergence and Performance of Q-Learning in Randomly Generated Games with payoff elements in $[0, 5]$. Convergence is measured by the difference between mixed strategy components, defined by (21) and Performance by TSW. The black line depicts the average taken over all 50 games and initial conditions, whilst the confidence interval depicts the maximum and minimum differences. The red line shows the TSW achieved by Q-Learning, averaged over all games.

Proof. 1. It holds from (Melo, 2021) that $\bar{\mathbf{x}}$ is a QRE of Γ_1 if and only if it is an NE of the perturbed game $\Gamma_1^H = (\mathcal{N}, (S_k, u_k^H)_{k\in\mathcal{N}})$ where

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

Applying the definition of a Nash Equilibrium, it is required that, for all $y \in \Delta$,

$$\sum_k \langle \mathbf{y}_k - \bar{\mathbf{x}}_k, r_k^H(\bar{\mathbf{x}})\rangle \leq 0$$

It follows then, that , for any $c > 0$

$$\sum_k \langle \mathbf{y}_k - \bar{\mathbf{x}}_k, cr_k^H(\bar{\mathbf{x}})\rangle \leq 0$$

From Proposition D.1, $cr_k^H(\bar{\mathbf{x}})$ are the rewards in Γ_2, so that the result follows.

2. Using Proposition D.1 it holds that

$$\begin{aligned} SW_2(\mathbf{x}) &= \sum_k \sum_{(k,l)\in\mathcal{E}} \mathbf{x}_k^\top \tilde{A}^{kl} \mathbf{x}_l \\ &= c\sum_k \sum_{(k,l)\in\mathcal{E}} \mathbf{x}_k^\top A^{kl} \mathbf{x}_l \\ &= cSW_1(\mathbf{x}) \end{aligned}$$

from which the result follows immediately.

□

As our interest is in the *change* in Social Welfare due to an *change* in exploration, and not on absolute values, our results are, thereby unaffected by the assumptions of boundedness and positivity.

To generate Figure 8, we run Q-Learning in 50 randomly generated games, for 25 initial conditions and record the final 25,000 iterations. Then we determine the largest variation in mixed strategies across all agents and all actions. More formally, this process estimates

$$\max_{ki} \lim_{t\to\infty} \left(\max_{t} x_{ki}(t) - \min_{t} x_{ki}(t) \right) \tag{21}$$

Figure 8 shows that the variation decreases as exploration rates are increased. Taken together, our results present strong evidence that the tradeoff between convergence and performance does not just occur in the special cases already examined, but rather holds in the vast majority of network games.

E. Conclusion

Understanding the effect of exploration in multi-agent learning faces a significant challenge due to the fact that, outside of a restrictive class of games, online learning often does not display asymptotic convergence to an equilibrium. We resolve this by showing that, in all network games with unique Nash Equilibrium (NE), smooth Q-Learning converges to a neighbourhood of the equilibrium. This occurs independently of the complexity the learning dynamics within this neighbourhood. The size of this neighbourhood can be decreased by increasing the exploration rates of the agents. As such, controlling the degree to which Q-Learning converges amounts to parameter tuning.

The downside of this process is reduced asymptotic performance of learning. We show that, in all games, increased exploration leads naturally to a reduced ability for agents to improve their payoff through learning. As our results place upper bounds on this phenomena, they give a manner in which exploration rates can be tuned to balance convergence and payoff performance. To take this further, we show that non convergent Q-Learning dynamics strictly outperform convergence in a multi-agent extension of the Shapley game. As our experiments confirm, this turns out to be a general phenomena across a large number of games.

The results in this paper brings the understanding of exploration in learning dynamics outside the realm of potential and network zero sum games. An interesting avenue for future work would be to continue developing this direction by lifting the assumption of network interactions and by considering the effect of exploration in games with multiple Nash Equilibria.