Stability and Chaos in Multi-Agent Reinforcement Learning

Anonymous Author(s)
Submission Id: 489

ABSTRACT

Modelling the dynamics of Q-Learning is an active and important topic for the sake of developing an *a priori* understanding of Reinforcement Learning. In this paper, we use methods from evolutionary game theory to analyse the stability of Q-Learning in p-player games, where payoffs are randomly generated. We determine the parameter range in which Q-Learning is expected to settle to a stable fixed point and the range in which the dynamics are unstable.

This study allows for parameters to be appropriately chosen to ensure the safe convergence of a learning algorithm and as a first step towards understanding the range of behaviours that can be displayed by learning using the Q-Learning algorithm. We validate our theoretical results through numerical simulation and show that, within the bounds of experimental error, the region of instability can be characterised by the learning dynamics.

KEYWORDS

Reinforcement Learning, Dynamical Systems, Game Theory.

ACM Reference Format:
Anonymous Author(s). 2021. Stability and Chaos in Multi-Agent Reinforcement Learning. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

Single agent reinforcement learning (RL) is a well-established framework for allowing agents to learn optimal strategies when trained on an iterated task [?]. For the realisation of complex tasks, such as air traffic control, market negotiations, and multi-robot coordination, however, it is required that the system be modelled as a multi-agent system (MAS). Such systems are not as well-understood as single-agent systems in RL, since a given agent is tasked with optimising a reward function which depends not only on a non-stationary environment, but also on the actions of other, possibly loosely coupled, agents [?].

It is, therefore, of paramount importance to develop a strong theoretical understanding of multi-agent reinforcement learning (MARL) to allow an *a priori* understanding of the behaviour of a given learning algorithm. Fortunately, the study of MAS is not unique to the field of AI and has been extensively investigated from the point of view of economics and game theory as well [?]. In particular, evolutionary game theory (EGT) considers the problem of a MAS which is repeatedly exposed to an iterated game. This idea shares a strong resemblance with MARL and, in fact, in [?] it was shown that techniques from EGT may be fruitfully applied to the analysis of Q-Learning [?].

An important result in modelling such multi-agent systems from the EGT perspective is that, when games are learnt and the assumptions of rationality and perfect information are lifted, games may not converge to an equilibrium. Instead, as shown in [?], the dynamics may be more complex, and even chaotic (see Sec. 2.1). The present work further describes that the emergence of such behaviours depends on the parameters of the games and the learning algorithm. For the analysis of multi-agent systems whose behaviour is inter-dependent, it is desirable that the system be such that the game converges to a stable equilibrium. Without this, it is inherently impossible to predict the outcome of learning.

Contribution. In this study, we will be considering the question: under what choice of parameters is a multi-agent system, which is trained on an iterated normal-form game using Q-Learning, likely to converge to an equilibrium as opposed to displaying complex, or unstable behaviours? In investigating this question, we will make the following assumptions:

(1) There is a finite set of agents, though its size p can be arbitrarily large.
(2) The agents have a large, but discrete, strategy space: during the theoretical study, we will make the assumption that the number of actions N goes to infinity.
(3) The agents are homogeneous. This requires that all agents have the same parameters and are trained using the same algorithm. Since this is typically the case in reinforcement learning studies, it is not an unreasonable assumption, but does present an interesting avenue for future work.
(4) The agents are trained on stateless, normal form games (i.e., the environment is static).

Through our analysis and experiments we find that likelihood of displaying unstable dynamics increases as the *step length*, α and *intensity of choice*, τ parameters in the Q-Learning algorithm increase. However, the correlation, Γ between payoffs, which measures how cooperative or competitive a game is, does not affect the stability of the system. In addition, we find that regardless of the choice of these parameters, the likelihood of unstable behaviours increases as the number of players p in the game increases.

Related Work

In this section we present an overview of the vast literature which considers the study of reinforcement learning from an evolutionary game dynamics perspective, as well as touching upon some important results and research within the field of game dynamics.

Evolutionary Game Dynamics. The theory of evolutionary game dynamics [?] considers game-like settings in which agents must repeatedly interact with one another. The outcome of this interaction depends on a payoff matrix; 'strong' strategies which maximise the reward are promoted, whilst 'weaker' strategies diminish. The *replicator dynamic* models this behaviour, and allows one to determine whether, after a number of iterations, the game is likely to

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, London, UK.

converge to some fixed equilibrium and, if so, the probabilities with which strategies are played in this equilibrium [?].

Analyses have been performed from the perspective of the replicator dynamic w.r.t. the various types of behaviours that can emerge from iterated games. Such games are considered imperfect, in that agents make decisions by attempting to anticipate their opponents' behaviour based on experience [?]. [?] presents the observation that, in the Prisoner's Dilemma, cyclic behaviour emerges in which the optimal strategy cycles across defection, cooperation and tit-for-tat despite defection being the NE. In [?], Galla builds on this observation by showing that, in fact, such behaviour is highly dependent on the presence of memory loss in the system. In the absence of memory loss, the learning converges to the NE as expected.

Taking the notion of parameter dependence further, [?] describes an analysis of two-player iterated games in which the players learn using *experience weighted attraction* (EWA), a form of reinforcement learning typically applied in experimental economics [?]. Through numerical simulations, the authors are able to show that the emergence of chaos and cycles is dependent on the choice of parameters. A rigorous theoretical analysis of the replicator dynamic corresponding to EWA, which applies the techniques in [?], results in a method for characterising the regions in parameter space in which a game is likely to converge, exhibit chaos or limit cycles. More recently, the same authors presented an extended study in [?], in which the same analysis is applied to generic p-player games, which showed that chaotic dynamics are more likely to be observed as the number p of players increases. Differently from this line, we here consider Q-Learning [?], which is one of the most popular algorithms in RL.

It is evident, therefore, that in the case of learning on iterated games, convergence to stable equilibria cannot be taken for granted and, in fact, is rare. It would therefore be fruitful to bring these analyses from EGT to better understand reinforcement learning from an AI perspective.

Dynamics of RL. In [?], the authors are able to derive a relation between Q-Learning and the replicator dynamic. In doing so, the authors present an evolutionary model which accurately predicts the dynamics of Q-Learning. This sprung forward a vast array of literature which performs similar derivations for various RL algorithms [?]. In particular, many of these studies focus on normal form games, in which the payoff matrix does not change (i.e., the environment is static). Recent work, such as [?], extends this framework towards multi-state games. Similarly, [?] extends the work in [?] towards games with continuous action spaces. More recently, [?] considers Q-learning in the mean-field limit, in which the strategies of populations of agents are considered, rather than individuals.

The relation between reinforcement learning and the replicator dynamic allows for the assumption of convergence in RL to be lifted and instead for the emergence of more complex behaviours to be examined. In particular, this study aims to perform an analysis similarly to [?] and establish the regions in parameter space in which Q-Learning converges to stable fixed points. In doing so, we extend the dynamics considered by Tuyls et al. to a general p-player setting, rather than two players only, thereby allowing an analysis of how the number of players in a game affects its stability.

2 PRELIMINARIES

In this section, we establish some of the preliminaries which are required in order to follow the subsequent sections. We first consider *dynamical behaviours*, which describes the evolution of agent (note we use the terms *agent* and *player* interchangeably) strategies as they learn how to play an iterated game. This study aims to classify, based on the agent parameters and the payoff matrices, which of these behaviours will be observed. The second is the *dynamics of Q-Learning*, which are a set of equations that model the aforementioned strategy evolution. It is on these dynamics that we perform our analysis.

2.1 Dynamical Behaviours

As discussed in the Sec. 1, when learning on iterated games, player strategies may exhibit much more complex behaviours than convergence to a Nash Equilibrium (NE), including convergence to a unique equilibrium (though not always to an NE), convergence to one of multiple equilibria, limit cycles, and chaos. These behaviours are illustrated in Fig.1. To be able to predict the behaviour of a learning algorithm, it should ideally converge to a stable equilibrium although it is still possible to study systems with multiple equilibria or limit cycles [?]. It would be difficult, however, to control systems whose dynamics are governed by chaos (though research into controlling chaos is ongoing and rife with opportunity [?]). It would, therefore, be a useful endeavour to determine the conditions under which the different sorts of behaviours arise.

2.2 Dynamics of Q-Learning

The behaviour of a system may be studied given a model of its dynamics. It is through this process that a wide array of physical systems, from harmonic pendulums to geophysical fluids, can be understood. A growing body of research aims to understand multi-agent reinforcement learning through the lens of its dynamics. In this light, Tuyls et al. [?] present a derivation of a continuous-time dynamical system describing how agents following a Q-Learning approach adjust the probabilities of choosing actions as they iteratively play a game.

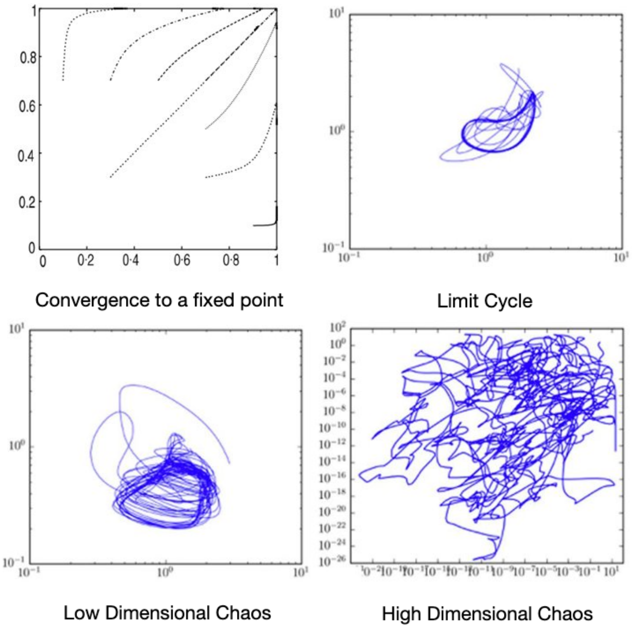
The games we consider are *normal form games*. These consist of: a finite set of players, with individual strategy spaces $\mathcal{A}$ (though we assume that all players share the same strategy space) and payoff functions for each player. The agents choose an action from their strategy space and receive a reward from their payoff matrix dependent on the actions of all players. The payoffs remain unchanged across iterations. Examples of these normal form games include the popular Prisoner's Dilemma or Matching Pennies games [?].

The Q-Learning approach considered requires an agent to choose an action i at step $k+1$ with probability

$$x_i(k) = \frac{e^{\tau Q_i(k)}}{\sum_j e^{\tau Q_j(k)}} \tag{1}$$

where $\tau \in [0, \infty)$ is the *intensity of choice* as described at the end of this section, and Q_i denotes the *Q-value* of an action i, which is to be updated at each step according to

$$Q_i(k+1) = (1-\alpha)Q_i(k) + \alpha(r + \gamma \max_j Q_j(k)) \tag{2}$$



Convergence to a fixed point

Limit Cycle

Low Dimensional Chaos

High Dimensional Chaos

Figure 1: Different types of dynamical behaviour displayed by learning agents. Here the x-axis (resp. y-axis) is the probability with which agent 1 (resp. agent 2) chooses a given action. (Top Left) Figure drawn from [?]. Convergence to a unique fixed point at the point (1, 1). (Top Right) Limit Cycle, the trajectories converge to cyclic behaviour (Bottom) Chaotic behaviour, here small deviations in the initial conditions can grow exponentially. (Top Right) and (Bottom) drawn from [?].

where $\alpha \in [0,1]$ is the *step length* parameter described below, r is the immediate reward received, and $\gamma \in [0,1]$ is the discount factor.

Through their analysis, Tuyls et al. were able to arrive at the following dynamical model of multi-agent Q-Learning

$$\frac{\dot{x}_i(t)}{x_i(t)} = \alpha\tau\Big(\sum_j a_{ij}y_j - \sum_{ij} x_i a_{ij} y_j\Big) + \alpha \sum_j x_j ln\Big(\frac{x_j}{x_i}\Big) \tag{3a}$$

$$\frac{\dot{y}_i(t)}{y_i(t)} = \alpha\tau\Big(\sum_j b_{ij}x_j - \sum_{ij} y_i b_{ij} x_j\Big) + \alpha \sum_j y_j ln\Big(\frac{y_j}{y_i}\Big). \tag{3b}$$

Here, agent 1 (resp. agent 2) takes action i with probability x_i (resp. y_i). When these actions are taken, the agent receives payoff a_{ij} (resp. b_{ji}). With these equations, it is possible to predict the expected behaviour of Q-Learning agents, which they go on to empirically verify.

It is clear from (3) that the long-term strategy selection of these agents is determined by the parameters α, τ and the payoffs a_{ij}, b_{ij}. In light of this, this study aims to address the question presented in the introduction: how do these parameters influence the types of behaviours seen during learning on an iterated game? Specifically, the parameters we will consider are:

(1) $\alpha \in [0,1]$: the *step length*. Low values of α denote smaller updates. Heuristically, we can consider this to be the memory of the agent: lower α denotes longer memory.

(2) $\tau \in [0,\infty)$: the *intensity of choice*, as termed by Sanders et al. [?]. This is sometimes written as β in the literature [?]. $\tau = 0$ results in all actions being selected with equal probability, regardless of their Q-value, whilst $\tau \to \infty$ results in the action with the highest Q-value chosen at every step. This may be seen as the *exploration-exploitation parameter*. For greater details into these parameters, the interested reader should consult [?].

(3) $\Gamma \in [-1, p-1]$: the *payoff correlation*. Since there are an infinite number of realisations of the payoffs a_{ij}, b_{ij}, we instead analyse the behaviour for the average payoff case. To do this, we assume that the payoff matrices are drawn from a multi-variate Gaussian with mean zero and covariance matrix parameterised by Γ. We then average over this Gaussian. $\Gamma = -1$ indicates a zero-sum game, in which the sum of payoffs for a given action across all agents is 0, resulting in a purely competitive game. $\Gamma = p-1$ gives a purely cooperative game in which all agents share the same payoffs. The manner in which a game is generated from the choice of Γ is described in (6) and follows the same procedure as outlined in [?].

3 STABILITY ANALYSIS OF Q-LEARNING

In this section, we aim to determine how the choice of parameters α and τ, alongside the choice of payoff matrix, affect the stability of the dynamics of Q-learning. As the values in the payoff matrix can take any real number, however, there is an infinite number of possible realisations of games. Of course, it would not be possible to analyse every possible game, so we instead follow the procedure outlined in [?] and [?] to average over these realisations and, instead, analyse the *effective dynamics*, the dynamics averaged over all realisations of payoff matrices. Then, we perform a linear stability analysis around equilibrium points to determine the conditions under which a fixed point is stable. As this process is rather involved, we focus on the main steps in this paper, and refer to the Supplementary Material for the technical details. Sections from the Supplementary Material are denoted with the prefix S.

3.1 Rescaling of Variables

Our first point of call is to extend the dynamics of Tuyls et al. in (3a) and (3b) to a general p-player game. This yields the dynamics (4). Here, player μ chooses action i_μ from its strategy space $\mathcal{A}$ at time t with probability $x^{\mu}_{i_\mu}(t)$ and receives a reward $P^{\mu}_{i_\mu, i_{-\mu}}$ from its payoff matrix P^μ depending on its own action and the actions of all other agents $i_{-\mu}$, where $i_{-\mu}$ denotes the set $\{i_\kappa : \kappa \in \{1, 2, \ldots, p\} \setminus \{\mu\}\}$ and $-i_\mu$ denotes the set $\mathcal{A} \setminus i_\mu$.

$$\frac{\dot{x}^{\mu}_{i_\mu}(t)}{x^{\mu}_{i_\mu}(t)} = \alpha\tau\left(\sum_{i_{-\mu}} P^{\mu}_{i_\mu, i_{-\mu}} \prod_{\kappa \neq \mu} x^{\kappa}_{i_\kappa}(t) - \sum_{i_\mu i_{-\mu}} x^{\mu}_{i_\mu}(t) P^{\mu}_{i_\mu, i_{-\mu}} \prod_{\kappa \neq \mu} x^{\kappa}_{i_\kappa}(t)\right)$$
$$+\alpha \sum_{j_\mu \in -i_\mu} x^{\mu}_{j_\mu}(t) ln \frac{x^{\mu}_{j_\mu}(t)}{x^{\mu}_{i_\mu}(t)} \tag{4}$$

We now scale the system so that the probabilities sum to N. The reason for this scaling is that, as the number of agents p and number of actions N increases, the expected payoffs that agent

$x^{\mu}_{i_\mu}(t)$ receives all converge to the same value across its action space. Therefore, the system must be scaled to ensure appreciable differences across actions, so that the agent can choose its actions appropriately (see the discussion in [?] or Sec. S2.2 for further details). Accordingly we choose scaled variables (denoted with a tilde or hat) as

$$\begin{aligned}
P^{\mu}_{i_\mu,i_{-\mu}} &= \tilde{P}^{\mu}_{i_\mu,i_{-\mu}}\sqrt{N^{p-1}} \\
x^{\mu}_{i_\mu}(t) &= \frac{\tilde{x}^{\mu}_{i_\mu}(t)}{N} \\
\tilde{\alpha} &= \frac{\alpha}{N},\ \tilde{\tau} = N^{-(p-1)/2}\tau,\ \hat{\tau} = N^{-p/2}\tau.
\end{aligned}$$

The substitution in (4) yields the scaled equation

$$\begin{aligned}
\frac{\dot{\tilde{x}}^{\mu}_{i_\mu}(t)}{\tilde{x}^{\mu}_{i_\mu}(t)} &= \alpha\tilde{\tau}\left(\sum_{i_{-\mu}}\tilde{P}^{\mu}_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu}\tilde{x}^{\kappa}_{i_\kappa}(t)\right) \\
&\quad -\alpha\hat{\tau}\left(\frac{1}{\sqrt{N}}\sum_{i_\mu i_{-\mu}}\tilde{x}^{\mu}_{i_\mu}(t)\tilde{P}^{\mu}_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu}\tilde{x}^{\kappa}_{i_\kappa}(t)\right) \\
&\quad +\tilde{\alpha}\sum_{j_\mu\in -i_\mu}\tilde{x}^{\mu}_{j_\mu}(t)ln\frac{\tilde{x}^{\mu}_{j_\mu}(t)}{\tilde{x}^{\mu}_{i_\mu}(t)} \qquad (5)
\end{aligned}$$

Note that the scaled probabilities now satisfy the constraint $\sum_{i_\mu}\tilde{x}^{\mu}_{i_\mu}(t) = N$. For the remainder of this derivation we concern ourselves only with the scaled system so we drop the tilde notation on P and x.

As mentioned, we now generalise these dynamics to account for all the possible realisations of the payoff matrix. To do this, we assert that the payoff elements (after scaling) are generated by a multi-variate Gaussian distribution with mean zero and covariance given as

$$\mathbb{E}\left[P^{\mu}_{i_\mu,i_{-\mu}}P^{\nu}_{i_\nu,i_{-\nu}}\right] = \begin{cases}\frac{1}{N^{p-1}} & \text{if } \nu = \mu \\ \frac{\Gamma}{(p-1)N^{p-1}} & \text{otherwise.}\end{cases} \qquad (6)$$

The motivation for choosing that the payoffs is generated by a Gaussian is that it allows for the use of Gaussian identities when determining the average.

3.2 The Effective Dynamics

We then use a generating functional approach as outlined in [?] to average the dynamics over this multi-variate Gaussian. The generating functional is a method derived from statistical mechanics which allows for expectations to be taken over equations of motion and, in particular, allows for the analysis of systems with 'quenched disorder': these are variables which are random but held fixed in time as the system evolves (in our case these variables are the payoff elements). The generating functional for the equation of motion (5) is given as [?]

$$\begin{aligned}
Z(\vec{\psi}) = \int D[\vec{x},\vec{\hat{x}}]exp(i\sum_{i,\mu}\int dt[\hat{x}^{\mu}_{i_\mu}(t)(\frac{\dot{x}^{\mu}_{i_\mu}(t)}{x^{\mu}_{i_\mu}(t)} - \tilde{\alpha}\rho^{\mu}_{i}(t) - h^{\mu}_{i}(t))])\times \\
exp(-i\alpha\tilde{\tau}\sum_{\mu}\sum_{i_\mu,i_{-\mu}}\int dt[\hat{x}^{\mu}_{i_\mu}(t)P^{\mu}_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu}x^{\kappa}_{i_\kappa}(t))]))\times \\
exp(-i\alpha\tilde{\tau}\sum_{\mu}\sum_{j_\mu,i_\mu,i_{-\mu}}\int dt[\hat{x}^{\mu}_{j_\mu}(t)x^{\mu}_{i_\mu}(t)P^{\mu}_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu}x^{\kappa}_{i_\kappa}(t)]))\times \\
exp(i\sum_{i,\mu}\int dt[x^{\psi}_{i_\psi}(t)^{\mu}_{i}(t)]),
\end{aligned} \qquad (7)$$

where the 'generating fields' [?] $\psi_i\vec{(t)}$ and $\phi_i\vec{(t)}$ will be set to zero at the end of the calculation.

The last two exponentials contain the payoffs of the game. These are randomly generated using a multi-variate gaussian and then held fixed for the rest of the game. As such these comprise the aforementioned 'quenched disorder'. We will average over this quenched disorder (and therefore average over all possible realisations of payoff matrices) using the mean and covariance expressions given in (6). We define Q to be the product of the second and third exponentials in (7)

$$\begin{aligned}
\mathbb{E}[Q] = exp(-\frac{\alpha^2\tilde{\tau}^2}{2}N\sum_{\mu}\Big[\int dtdt'L^{\mu}(t,t')\prod_{\kappa\neq\mu}C^{\kappa}(t,t')+ \\
\Gamma\sum_{\nu\neq\mu}K^{\mu}(t,t')K^{\nu}(t,t')\prod_{\kappa\notin\{\mu,\nu\}}C^{\kappa}(t,t')\Big])\times \\
exp(-\frac{\alpha^2\hat{\tau}^2}{2}N\sum_{\mu}\Big[\int dtdt'L^{\mu}(t,t')C^{\mu}(t,t')\prod_{\kappa\neq\mu}C^{\kappa}(t,t')+ \\
\Gamma\sum_{\nu\neq\mu}A^{\mu\nu}(t,t')C^{\nu}(t,t')C^{\mu}(t,t')\prod_{\kappa\notin\{\mu,\nu\}}C^{\kappa}(t,t')\Big])
\end{aligned} \qquad (8)$$

in which

$$\begin{aligned}
C^{\mu}(t,t') &:= N^{-1}\sum_{i}x^{\mu}_{i_\mu}(t)x^{\mu}_{i_\mu}(t') \\
L^{\mu}(t,t') &:= N^{-1}\sum_{i}\hat{x}^{\mu}_{i_\mu}(t)\hat{x}^{\mu}_{i_\mu}(t') \\
K^{\mu}(t,t') &:= N^{-1}\sum_{i}x^{\mu}_{i_\mu}(t)\hat{x}^{\mu}_{i_\mu}(t') \\
A^{\mu,\nu}(t,t') &:= N^{-1}\sum_{i}\hat{x}^{\mu}_{i_\mu}(t)\hat{x}^{\nu}_{i_\nu}(t').
\end{aligned}$$

By taking this average (Sec. S2.3), we obtain the *effective dynamics*

$$\begin{aligned}
\frac{1}{x}\frac{d}{dt}x(t) &= \alpha^2\tilde{\tau}^2\Gamma\int dt'\left[G(t,t')C^{p-2}(t,t')x(t')\right] \\
&\quad +\sqrt{2}\alpha\tilde{\tau}\eta_1(t) + \sqrt{2}\alpha\hat{\tau}\eta_0(t) + \tilde{\alpha}\rho(t), \qquad (9)
\end{aligned}$$

in which we have assumed that all players' actions are independent and drawn from the same initial distribution (i.i.d) and therefore dropped the distinction between players and strategy components. The terms G, C, η_1, η_0 are correlation functions, generated when averaging the Gaussian. These are given as

$$\begin{aligned} C(t,t') &= \mathbb{E}[x(t)x(t')] \\ \mathbb{E}[\eta_1(t)] &= 1,\ \mathbb{E}[\eta_1(t)\eta_1(t')] = C^{p-1}(t,t') \\ \mathbb{E}[\eta_0(t)] &= 1,\ \mathbb{E}[\eta_0(t)\eta_0(t')] = C^{p}(t,t') \\ G(t,t') &= \mathbb{E}\left[\frac{\delta x(t)}{\delta \eta_1(t')}\right]. \end{aligned}$$

It is important to note the effect of the former assumption (that all actions of all agents are i.i.d) is substantial. With this assumption, we identify the coupled term $A^{\mu,\nu}(t,t')$ in (9) with the uncoupled term $L^{\mu}(t,t')$ (since there is no distinction between the actions of agents μ and ν). This is a strong assumption which removes the interdependency between agents in the analysis and is required to ensure that (9) yields real-valued action probabilities. However, as shown by the experimental evaluation, it does not produce a strong discrepancy in describing the qualitative effect on stability caused by the parameters α, Γ, and τ.

3.3 Linear Stability Analysis

We first find a fixed point x_∞ of (9) by letting $\dot{x}(t) = 0$. This yields the expression

$$0 = x_\infty[\alpha^2\tilde{\tau}^2\Gamma x_\infty q^{p-2}\chi + \sqrt{2}\alpha\tilde{\tau}q^{(p-1)/2}z + \sqrt{2}\alpha\hat{\tau}q^{p/2}z^{p/p-1} + \tilde{\alpha}\rho] \tag{10}$$

where $\chi = \int dt' G(t-t')$, $\eta_0(t) = q^{(p-1)/2}z$, z is drawn from a Gaussian of zero mean and unit variance. By using (10) we can calculate x_∞. We disregard the choice $x_\infty = 0$, since we assume that no action will be chosen with exactly zero probability, an intuitive assumption which is also considered in [?] and [?]. The expression inside the squared bracket admits a positive solution only in the region $\Gamma \in [-1, 0]$, whilst for positive $\Gamma \leq p-1$, the nature of solutions may not be guaranteed. Therefore, we restrict our analysis to the region $[-1, 0]$.

We now analyse the stability of the system (9) in a neighbourhood around this fixed point. This follows a similar procedure as in [?], in which the fixed point dynamics are proposed to be perturbed by a disturbance $\xi(t)$ which is drawn from a Gaussian of zero mean and unit variance. The disturbance causes the values of $x(t)$ and $\eta_0(t), \eta_1(t)$ to deviate from their fixed point position $x_\infty, \eta_{0,\infty}, \eta_{1,\infty}$ by an amount $\hat{x}(t), \hat{\eta}_0(t), \hat{\eta}_1(t)$. If we consider only the terms which are linear in these perturbations, we arrive at

$$\begin{aligned} \frac{d}{dt}\hat{x}(t) &= (x_\infty + \hat{x}(t))[\alpha^2\tilde{\tau}^2\Gamma x_\infty \int dt'[G(t,t')C^{p-2}(t,t')] \\ &\quad +\sqrt{2}\alpha\tilde{\tau}\eta_{1,\infty} + \sqrt{2}\alpha\hat{\tau}\eta_{0,\infty} + \tilde{\alpha}\rho] \\ &\quad +x_\infty[\alpha^2\tilde{\tau}^2\Gamma \int dt'[G(t,t')C^{p-2}(t,t')\hat{x}(t')] \\ &\quad +\sqrt{2}\alpha\tilde{\tau}\hat{\eta}_1(t) + \sqrt{2}\alpha\hat{\tau}\hat{\eta}_0(t) + \xi(t)]. \end{aligned} \tag{11}$$

We then examine the long-term behaviour of the perturbation $\hat{x}(t)$. by taking the Fourier transform of (11) and analysing its behaviour at $\omega = 0$. After some manipulation (Sec. S4), this gives

$$\mathbb{E}[|\mathcal{X}(\omega=0)|^2] = \frac{1}{\left[(-\alpha^2\tilde{\tau}^2\Gamma q^{p-2}\chi)^2 - 2(\alpha\tilde{\tau} + \alpha\hat{\tau})^2(p-1)q^{p-2}\right]} \tag{12}$$

As the left hand side of this equation must be greater than zero, we arrive at the condition that, at a fixed point, the long term perturbations must satisfy

$$0 \leq \left[(\alpha^2\tilde{\tau}^2\Gamma q^{p-2}\chi)^2 - 2(\alpha\tilde{\tau} + \alpha\hat{\tau})^2(p-1)q^{p-2}\right] \tag{13}$$

It should be noted that, to arrive at this result, we have had to take the assumption that p is large enough so that $p/p - 1 \approx 1$ so that the fractional power on z in (10) is reduced to 1. As z can take any value, including negative values, this again ensures that the expression for x_∞ remains real valued.

3.4 Discussion

The fixed point condition (13) yields a number of testable implications, of which we subsequently test the validity. These implications are as follows.

(1) A trivial result is that the game is everywhere convergent (i.e. regardless of the choice of Γ, N, p) for the cases of $\alpha = 0$ and/or $\tau = 0$. We see that choosing these values would result in the right hand side of (13) evaluating to zero, which falls within the stable region.

(2) Convergence is rare. This is seen by the fact that the analytic result overestimates the region of instability. In fact, for all allowed choices of Γ, α, τ, the right hand side of (13) is negative, which violates the stability criterion. This implies that games, in general, will not converge.

(3) The likelihood of convergence decreases with increasing α and τ, regardless of the choice of N and p. We can see this since the right hand side of (13) tends further away from zero (i.e. further from the region of stability).

(4) The choice of Γ does not affect the stability of the game in the negatively-correlated regime. Therefore, the only dependent factors are α, τ, N and p. This is more easily interpreted from the heatmaps in Figure 2.

(5) The likelihood of convergence decreases as p increases. We see this by noticing the $(p-1)$ in the second term of (13). As p increases, this term drives the system further from the stability boundary.

We illustrate these implications in Figure 2 which plots the value obtained by the right hand side of (13). In order to determine these values, we first had to solve (10) for q and χ. We determine the x_∞ which solves the expression inside the square brackets of (10) using a Newton-Raphson root-finding approach and use the following expressions for q and χ to determine their value.

$$\begin{aligned} 1 &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_\infty(z)exp(-\frac{z^2}{2})dz \\ q &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x_\infty(z)^2 exp(-\frac{z^2}{2})dz \\ \chi &= \frac{1}{q^{n/2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \frac{\partial x_\infty(z)}{\partial z} exp(-\frac{z^2}{2})dz \end{aligned}$$

It should be noted that this method of finding x_∞ yields only an approximate value and so the q and χ which are found are estimates.

We consider each of the implications (2)-(5) in turn, choosing to forego illustrating (1) as it can be seen immediately from (13). τ is held fixed at 0.05 and Γ is varied in the range $[-1, 0]$ and α in the range $[0.01, 0.05]$.

Implication 2: By looking at the values on the heatmap (Fig 2), we see that, for all choices of the parameters, (13) evaluates to a value less then zero, which signals instability. As such, the analysis over-estimates the region of instability. We believe that this is due to the assumptions made during the derivation of the analytic result, the strongest of which was to drop the discrepancy between players and actions, treating each action of each player to be independent and identically distributed. However, even within the error that this assumption generates, it is still clear (and is verified experimentally) that (13) predicts convergence in Q-Learning to be rare. Indeed, as the value of N increases, the stability boundary determined by experiment (Figure 4) tends towards $\alpha = 0, \tau = 0$.

Similarly, as the value of N increases, the values in the heatmap (seen on the right of the map) reduce. This is due to the assumption made in the derivation of the analytic result that $N \to \infty$. As the value of N increases, the discrepancy between the true behaviour of the system and analytic result decreases. This does, however, give rise to a limitation of the result - namely that the effect of the number of actions, N, cannot be accurately inferred from (13). However, it should still be noted that, even with the increase of N, the values in parameter space remain negative and predict instability. We go on to verify that the effect of p, α, Γ and τ are accurately captured by (13).

We interpret (13), for given parameters, by determining how close the right hand side is to the stability boundary (i.e. how close the value is to zero). A choice of parameters which evaluates close to the stability boundary has a higher chance (by the analytic error) of being convergent than one which lies further from the boundary. Therefore, the heatmaps in Figure 2 should be considered to show the probability of convergence for a choice of (α, Γ). Lighter regions denote a higher probability of convergence whereas darker regions are those whose values are further from zero and therefore show a lower probability of convergence.

Implication 3: can be visualised by noticing that the likelihood of convergence decreases (the heatmap becomes darker) as α increases, and similarly for τ. This occurs independent of the choice of N or p.

Implication 4: can similarly be inferred by the trend of convergence, which occurs only in varying α. The choice of Γ makes almost no difference in the evaluation of (13). We anticipate that this is due to the restriction of Γ to the range $[-1, 0]$. In this range, the scaling effect of Γ^2 is dominated by the α^4 and τ^4 terms which appear in (13).

Implication 5: is considered in Figure 3. In this, we keep all parameters except for p fixed, and evaluate the right hand side of (13) different values of p, taking the absolute value for convenience. We see that the trend is to tend away from the stability boundary indicating that, as the number of players in the system increases, so too does the region of instability.

4 EXPERIMENTAL EVALUATION

In this section, we describe and discuss the numerical experiments we use to verify the implications described in the discussion of Sec. 3. We first describe how these experiments were conducted and then, having presented the results in Figures 4 and 6, we discuss the correlation with the theory.

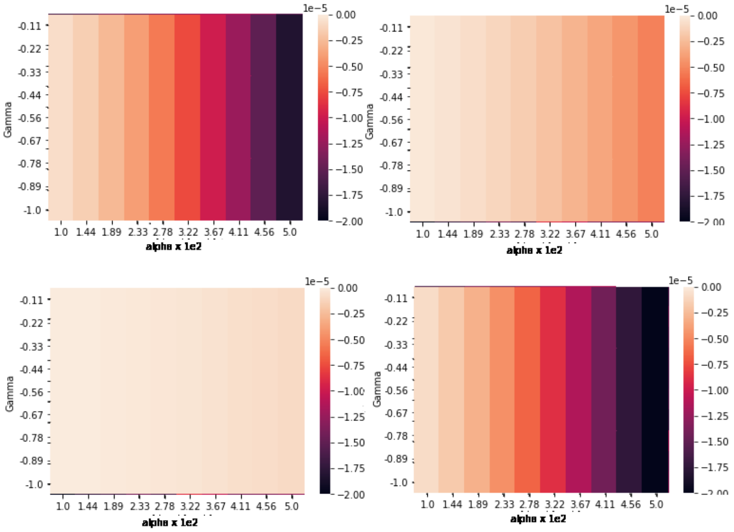


Figure 2: Evaluation of (13) over a range of α and Γ values. Ligher regions indicate where (13) is close to zero (i.e. close to the stability boundary) whilst darker regions are further away. We vary α in the range $[0.01, 0.05]$ and Γ in the range $[-1, 0]$. $\tau = 0.05$. (Top Left) $p = 2, N = 2$, (Top Right) $p = 2, N = 5$, (Bottom Left) $p = 2, N = 20$, (Bottom Right) $p = 3, N = 2$.

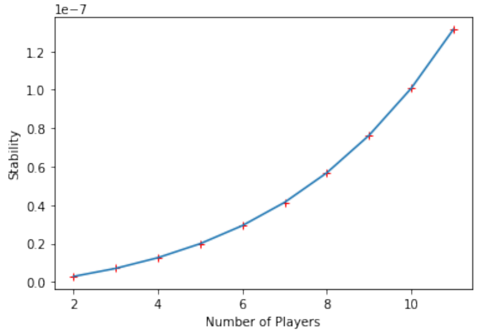


Figure 3: Evaluation of (13) for varying numbers of players p. Here, α is fixed at 0.02, $\tau = 0.05$, $N = 2$ and $\Gamma = -1$. p is varied from 2 to 12. We see that the value obtained by (13) tends away from zero monotonically.

4.1 Construction of Numerical Experiments

To verify experimentally the theoretical results, and to examine the underlying structure of stability and chaos in Multi-Agent Q-Learning, we perform a series of numerical experiments by varying the parameters Γ and α whilst keeping τ fixed. The aim is to determine the regions in which games learnt using Q-Learning converge to an equilibrium. The results of these experiments are shown in

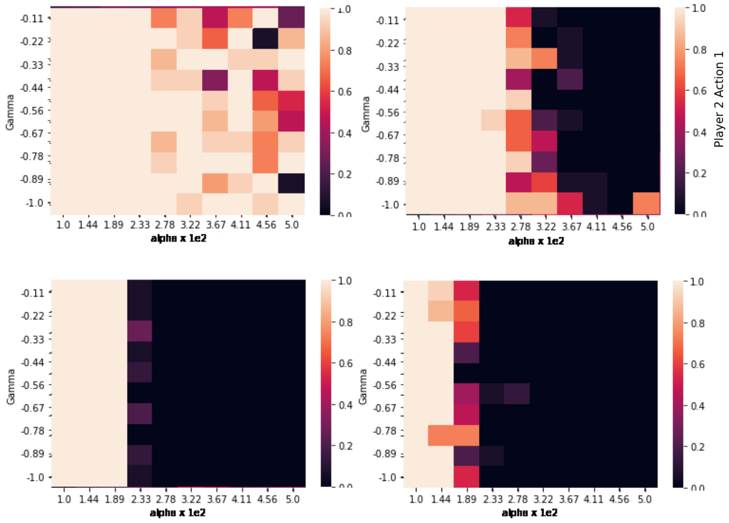


Figure 4: Results of numerical experiments in which α is varied. The heatmaps show the fraction of games which converge to an equilibrium. Lighter values show higher convergence. Each experiment is run with $\tau = 0.05$. (Top Left) $p = 2, N = 2$, (Top Right) $p = 2, N = 5$, (Bottom Left) $p = 2, N = 20$, (Bottom Right) $p = 3, N = 2$.

Figure 4, with parameters chosen to match those in the analytic assessment (Figure 2).

To generate the numerical simulations in Figure 4 we used the following procedure.

(1) Fix the parameters τ, γ. The former is held at 0.05 and the latter at 0.1.
(2) Initialise values of Γ, α (or τ appropriately). These will be swept over in the experiment.
(3) Generate 15 payoff matrices by sampling from a multi-variate Gaussian (variables are the payoff elements) with mean zero and covariance parameterised by Γ.
(4) For each of these payoff matrices, initialise a set of agents with random initial conditions (i.e., random action probabilities).
(5) Allow both sets of agents to learn over a maximum of 1.5×10^4 iterations.
(6) Keep track of the action probabilities over a window of 500 iterations. At the end of each window, determine the percentage difference between the maximum and minimum values of each strategy component
(7) If the difference is less than 1% consider the game converged. Otherwise continue to the next window.
(8) If the game reaches 1.5×10^4 iterations without satisfying the relative distance criterion, consider it to be non-convergent. Determine the fraction of these 15 games which have converged.

We see in Figure 4 that the stability of the system is highly dependent on the value of α and τ but not on Γ.

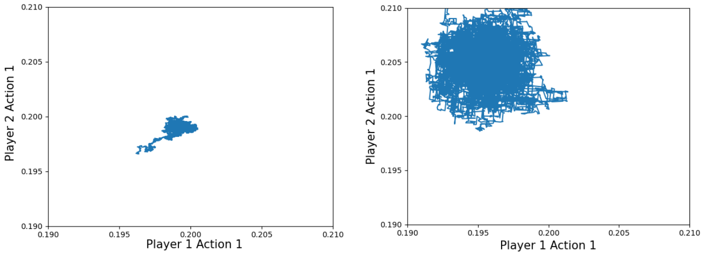


Figure 5: The trajectories of action probabilities for 2 players with 5 actions trained on a game of $\Gamma = -0.5$ and $\tau = 0.05$. (Left) $\alpha = 0.01$, (Right) $\alpha = 0.05$.

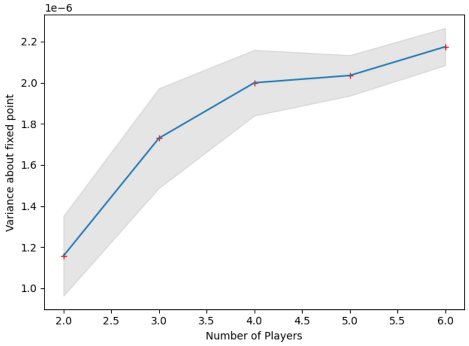


Figure 6: Variation $V(t)$ about a fixed point as p ranges from 2 - 6. $N = 2$, $\alpha = 0.02$, $\tau = 0.05$, $\Gamma = [-1, 0]$. The experiment is conducted for 10 choices of Γ and the mean (line) is plotted alongside the standard deviation (shaded region).

In Figure 5, we explicitly plot the trajectories taken by the action probabilities as two agents are trained on a game of $\Gamma = -0.5$. In the left figure, we choose $\alpha = 0.01$ whilst in the right we choose $\alpha = 0.05$. We note that the typical behaviour of the system is that the system first moves towards a fixed point and then varies in a pseudo-random manner within a vicinity of the fixed point. Critically though, the amount of variation in the case $\alpha = 0.05$ is greater than that of $\alpha = 0.01$.

Finally, in Figure 6 we plot the degree variation about the fixed point displayed as the number of players p increases. The variation is determined by first allowing the game to iterate for 5000 steps, so that it can be assumed to have reached the vicinity of the fixed point. Then, the action probabilities are recorded for a further 5000 iterations. At the end of this second period, the variance of the actions is calculated as

$$V(t) = \frac{1}{N} \sum_i \frac{1}{5000} \sum_t x_i(t)^2 - \left[\frac{1}{5000} \sum_t x_i(t) \right]^2.$$

This gives a measure of the degree of variability about the fixed point and, therefore, gives some notion of the 'degree of instability'. For example, the right hand plot of Figure 5 has a higher $V(t)$ than the left and so is considered to show a higher degree of instability.

Due to computational restraints, which we discuss further in the discussion, we are only able to increase p to a maximum value of 6 and must restrict also to $N = 2$. We maintain $\alpha = 0.02$ and $\tau = 0.05$ to match with the parameters used for Figure 3. Finally, in Figure 6, we are averaging over 10 choices of Γ in the range $[-1, 0]$.

4.2 Discussion

We note that the numerical experiments confirm the predictions suggested by the analytic results in Figure 3. We notice that convergence occurs almost only for low values of α. This observation remains as we increase the value of N. As noted, the analytic result over-estimates the region of instability due to the assumptions of $N \to \infty$ and $\frac{1}{p-1} << 2$ made during the derivation. Remaining discrepancies are due to the aforementioned approximation in the calculation of q, χ and the fact that the analytic result considers a continuous time approximation of the expected behaviour of Q-Learning. We expect that testing over a greater number of payoff realisations and initial conditions will yield a more representative assessment of the average behaviour of Q-Learning. However, running these experiments is a computationally expensive procedure: for p players and N actions we require operations on p matrices with N^p elements. As such, due to a reduced availability of computational facilities, a large scale averaging was not possible.

However, due to the prevalence of non-convergent behaviour, shown by the dark regions in Figure 4 other than at low values of α, we conclude that implications (1), (2) are verified. Indeed, implication (3) and (4) are also verified as the proportion of convergent games decreases monotonically with α, but there does not appear to be any discernible dependence on Γ. Finally, we can conclude that implication 5 is verified by comparing Figures 3 and 6. It is clear that the instability of the system increases monotonically with p. A point to note is that Figure 6 shows an almost logarithmic dependence of the instability on p, whilst Figure 3 suggests that the dependence should be exponential. Due to the limited range of p that was used to generate Figure 6, it cannot yet be concluded that the dependence on p is logarithmic. However, the key point is that, as p increases, the unstable behaviour increases, regardless of parameter choice, which is corroborated by both theory and experiments. This phenomenon was referred to in [**?**] as 'the prevalence of chaotic dynamics in games with many players'.

A point which we wish to discuss here is a comparison with the result found in [**?**], which considered the stability of *experience weighted attraction* (EWA). Namely, Sanders et al. observe that convergence is seen for higher values of α, whereas lower values give rise to chaos, the opposite of what is found here. The reason for this can be seen in the update equation (2) for Q-Learning, whereby smaller values of α result in the agent placing a lower weight on the reward received at each step. As such, lower values of α result in the agent taking more conservative steps and yields a higher probability of convergence. In contrast, the update for EWA does not discount the reward received and, instead, only discounts the previous knowledge of the Q-value based on higher choices of α. In essence, the difference is the fact that Q-Learning discounts new information by a factor of α, whilst EWA does not. Yet, the difference in stability caused by this change is significant. This highlights the importance of performing analyses such as the present work; it allows for a method to analytically compare the difference between learning algorithms and, for practitioners, to ensure that the appropriate algorithm is chosen for the parameters of their specific task.

To summarise, we have shown that the analytic result (13) provides a strong assessment for the effect that α, τ, Γ and p have on the likelihood of convergence of Q-Learning. We showed that, at least for negatively correlated payoffs, the strength of correlation plays no role in determining the stability of the system. Furthermore, though (13) overestimates the region of instability, it accurately conveys that convergence of Q-Learning is unlikely for games with many players and actions. In fact, our experimental results also verify that stability may only be guaranteed for the limiting case of two-player, two-action games. Once either of these parameters increase, the region of instability increases rapidly. Finally, we showed that the behaviour of the stability line differs from that of EWA and suggest that this shows that a stability analysis is an important mode of analysis for reinforcement learning algorithms to ensure that parameters are being chosen appropriately to guarantee the safe convergence of learning.

5 CONCLUSION

In this study, we made a first contribution towards the characterisation of the behaviours of agents learning how to play p-player, N-action games through Q-Learning. To this end, we analysed the replicator model of Q-Learning derived in [**?**]. Specifically, we searched for the regions in parameter space where the dynamics are expected to converge to a stable equilibrium and those where learning is unstable. This yielded a number of important results. We showed that convergence to a unique fixed point is found for low values of α, the *step length* of the algorithm, and for negatively correlated payoff matrices, though the strength of correlation Γ does not influence stability. As α increases, the likelihood of convergence decreases. Our analysis also shows that the likelihood of convergence decreases, regardless of parameter choice as the number of players p in the system increases.

Future Work. Our first aim is to verify the analytic results at a finer resolution and by averaging over a greater number of payoff realisations. This will give a greater insight into the boundary between convergent and non-convergent behaviour and allow for practitioners to choose their parameters in such a way that convergence to a (possibly unique) fixed point can be expected. In addition we noted that one of the conclusions of our study is that convergence to a unique fixed point is rare in Q-Learning. In fact, complex behaviours (such as the existence of multiple fixed points, limit cycles, and chaos) are more likely. We, therefore, aim to characterise these complex behaviours in parameter space both theoretically and numerically.

As research into the dynamics of RL algorithms progresses, it would be prudent to apply this analysis to various other algorithm. Algorithms whose dynamics are established, and are therefore open to a stability analysis, include piecewise Q-Learning and Cross Learning. This would provide a strong method by which to compare and provide safety guarantees to different algorithms for a particular use case.

REFERENCES