Stability and Chaos in Multi-Agent Reinforcement Learning

Submission number 3289

Abstract

Modelling the dynamics of Q-Learning is an active and important topic for the sake of developing an *a priori* understanding of multi-agent reinforcement learning. In this paper, we apply methods from evolutionary game theory to analyse the stability of Q-Learning in p-player games with generalised payoff matrices. We determine the effect that the number of players, payoff correlations, and agent parameters have on the likelihood that the system converges to a stable fixed point, as opposed to displaying unstable behaviour. 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 using Q-learning. We validate our theoretical results through numerical simulations and show that, within the bounds of experimental error, the region of instability can be characterised by the learning dynamics.

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 [Sutton and Barto, 2018]. 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 less well-understood, since a given agent is tasked with optimising a reward function that depends not only on a, non-stationary environment, but also on the actions of other, possibly loosely coupled, agents [Schwartz, 2014].

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 [Shoham and Leyton-Brown, 2008; Leslie and Collins, 2006]. In particular, evolutionary game theory (EGT) considers the problem of a MAS that is repeatedly exposed to an iterated game. This idea shares a strong resemblance with MARL and, in fact, in [Tuyls *et al.*, 2006] it was shown that techniques from EGT may be fruitfully applied to the analysis of Q-Learning.

An important result in modelling 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 [Sanders *et al.*, 2018], the dynamics may be more complex and even chaotic [Strogatz, 2000]. In fact, chaotic behaviour has been found in a wide range of learning algorithms [van Strien and Sparrow, 2011; Chotibut *et al.*, 2020] The present work further describes to what extent the emergence of such behaviours depends on the parameters of the games and the learning algorithm.

Contribution In this study, we will be considering the question: how does the likelihood of convergent behaviour, as opposed to instability, depend on the parameters of a homogeneous multi-agent system which is trained on an iterated normal-form game using Q-learning? 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 (sometimes referred to as players) have a discrete strategy space. (3) The agents are homogeneous, i.e., they all share the same parameters.

We find that our analysis gives a strong indication for how the stability of learning depends on its parameters, and that these predictions are confirmed in experiments. Specifically, we find that unstable dynamics become more prominent as the parameters *step length* α and *intensity of choice* τ of Q-learning increase. By comparison, the correlation Γ between payoffs, which measures how cooperative or competitive a game is, has a negligible affect on stability. In addition, we find that regardless of the choice of these parameters, the likelihood of unstable behaviours increases monotonically with the number of players p.

Related Work The theory of evolutionary game dynamics [von Neumann *et al.*, 1944] considers game-like settings in which agents repeatedly interact with one another (i.e., 'iterated games'). The outcome of these interactions depends on a payoff matrix; 'strong' strategies which maximise the reward are promoted, whilst 'weaker' strategies diminish. The *replicator dynamic* models this behaviour as a differential equation, and allows one to determine whether, after a number of

iterations, the game is likely to converge to some fixed equilibrium [Shoham and Leyton-Brown, 2008].

Iterated games are considered imperfect, in that agents make decisions by attempting to anticipate and respond to their opponents' behaviour based on experience [Galla, 2011].

In such games [Galla and Farmer, 2013], it is found that cycles, as well as more complex behaviours, can emerge when two agents play using the *experience weighted attraction* (EWA) learning algorithm [Camerer and Ho, 1999]. A rigorous theoretical analysis of EWA, results in a method for characterising how complex dynamics, such as chaos or cycles, depends on the game parameters. [Sanders *et al.*, 2018] extends this analysis to generic p-player games, and show that chaotic dynamics are more likely to be observed as the number p of players increases.

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

In [Tuyls *et al.*, 2006], the authors derive a continuous time deterministic approximation of Q-Learning and show its relation to the replicator dynamic. In doing so, they present a dynamical system that has been used to accurately analyse the behaviour of Q-Learning [Bloembergen *et al.*, 2015]. One such analysis is [Leonardos and Piliouras, 2020] which analyses a variant of the Q-Learning dynamics. The authors show that changes in the exploration parameter (here called the 'intensity of choice' τ) can lead to different qualitative behaviours. Their study shares the same general conclusions as our own, though through a different methodology. In addition, we aim also to characterise how the number p of players and the agent memory α affect the behaviour of Q-learning.

2 Background on Q-learning

We now provide the specifics of the types of games and algorithm that we analyse in the subsequent sections. We consider *normal form games* consisting of: a finite set of players, with individual action 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 depending 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 [Tuyls *et al.*, 2006].

The Q-learning approach requires an agent to choose an action $i \in \mathcal{A}$ at step k 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

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

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. Note we use the terms 'payoff' and 'reward' interchangeably.

By analysing this algorithm, Tuyls et al. derive a continuous-time, deterministic model of Q-learning, and verify empirically its accuracy. This model (which can be found in [Tuyls *et al.*, 2006] and is reproduced in the supplementary material Sec S2.1 for completeness) makes it clear that the long term strategy selection of the agents is dependent on parameters α and τ, as well as the payoffs themselves. In this study, therefore, we aim to establish the nature of this dependence. Specifically, the parameters we consider are:

1. The *step length* $\alpha \in [0, 1]$. 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. The *intensity of choice* $\tau \in [0, \infty)$, which is sometimes written as β in the literature [Leonardos and Piliouras, 2020]. $\tau = 0$ 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.

3. The *payoff correlation* $\Gamma \in [-1, p-1]$. Since there are an infinite number of realisations of the payoff matrices, we parameterise the payoffs by their correlations. 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.In the subsequent sections, we will take an average over this Gaussian to determine the expected behaviour a MAS will take where the payoffs have a particular choice of Γ. $\Gamma = -1$ indicates a zero-sum game, in which the sum of payoffs for a given action across all agents is zero, resulting in a purely competitive game. $\Gamma = p - 1$ indicates 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 (4) and follows the same procedure as outlined in [Sanders *et al.*, 2018].

3 Stability Analysis of Q-Learning

In this section, we determine how the choice of parameters α and τ, alongside the choice of payoff matrix and the number p of players affects the stability of the continuous-time model of Q-learning. As the values in the payoff matrices 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 [Coolen, 2005; Galla and Farmer, 2013] to average over these realisations (Sec 3.2). This yields the *effective dynamics*, the dynamics averaged over all realisations of payoff matrices. Then, in Sec 3.3, we perform a linear stability analysis on these dynamics around equilibrium points to determine the conditions under which a fixed point is stable. As this process is complex, 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 Generalised Dynamics

Our first point of call is to write the dynamics of Tuyls et al. for a general p-player game. This yields the dynamics (3),

where 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 $\mathcal{A}\setminus i_\mu$.

$$\frac{\dot{x^\mu_{i_\mu}}(t)}{x^\mu_{i_\mu}(t)} = \alpha\tau\Big(\sum_{i_{-\mu}} P^\mu_{i_\mu,i_{-\mu}} \prod_{\kappa\neq\mu} x^\kappa_{i_\kappa}(t) - \sum_{i_\mu i_{-\mu}} x^\mu_{i_\mu}(t) P^\mu_{i_\mu,i_{-\mu}} \prod_{\kappa\neq\mu} x^\kappa_{i_\kappa}(t)\Big) + \alpha \sum_{j_\mu\in -i_\mu} x^\mu_{j_\mu}(t) ln \frac{x^\mu_{j_\mu}(t)}{x^\mu_{i_\mu}(t)} \quad (3)$$

For the sake of brevity, we sometimes drop the explicit dependence on t in the notation of $x^\mu_{i_\mu}(t)$, and allow it to be inferred.

To average over all payoff elements , we assert that they are generated by a multi-variate Gaussian distribution with mean zero and covariance given as

$$\mathbb{E}\left[P^\mu_{i_\mu,i_{-\mu}} P^\nu_{i_\nu,i_{-\nu}}\right] = \begin{cases} \frac{1}{N^{p-1}} & \text{if } \nu=\mu \\ \frac{\Gamma}{(p-1)N^{p-1}} & \text{otherwise.} \end{cases} \quad (4)$$

The motivation for choosing a Gaussian distribution is to allow for the use of Gaussian identities when determining the average [Zinn-Justin, 2002]. Furthermore, the domain of the multi-variate Gaussian is over all possible realisations of the payoff elements (i.e., over $\mathbb{R}^{p\cdot N^p}$). This means that we can accurately consider any set of payoff matrices to be drawn from a Gaussian.

3.2 The Effective Dynamics

Our aim is first find to the corollary of (3) for the case of averaged payoff elements, we call this the 'effective dynamics'. In order to find these dynamics we take a similar approach to that outlined in [Mezard, 1987]. The idea is as follows. We first take a path integral over the entire trajectory as the *generating functional* Z. This gives a single value which can be assigned to the trajectory. We can then find an expression for the expected value of Z by averaging over all payoff elements, giving $\mathbb{E}[Z]$. Finally, we notice that, in the same way that Z was generated from an equation of motion, so $\mathbb{E}[Z]$ may be considered as generated from an 'averaged' equation of motion, which we call the *effective dynamics*. These dynamics, then, form the required average case of (3).

The generating functional of the dynamics (3) is given as

$$\begin{aligned} Z(\vec{\psi}) = \int D[\vec{x},\vec{\hat{x}}] exp(i\sum_{i,\mu}\int dt[\hat{x}^\mu_{i_\mu}(\frac{\dot{x}^\mu_{i_\mu}}{x^\mu_{i_\mu}} - \tilde{\alpha}\rho^\mu_i(t) - h^\mu_i(t))]) \\ \times exp(-i\alpha\tilde{\tau}\sum_\mu\sum_{i_\mu,i_{-\mu}}\int dt[\hat{x}^\mu_{i_\mu} P^\mu_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu} x^\kappa_{i_\kappa})])) \\ \times exp(-i\alpha\tilde{\tau}\sum_\mu\sum_{j_\mu,i_\mu,i_{-\mu}}\int dt[\hat{x}^\mu_{j_\mu} x^\mu_{i_\mu} P^\mu_{i_\mu,i_{-\mu}}\prod_{\kappa\neq\mu} x^\kappa_{i_\kappa}])) \\ \times exp(i\sum_{i,\mu}\int dt[x^\mu_{i_\mu}\psi^\mu_i(t)]), \end{aligned} \quad (5)$$

where $\vec{\psi_i}(t)$ and $\vec{\phi_i}(t)$ generate the necessary correlation functions. These act simply as a mathematical tool and will be set to zero at the end of the calculation [Coolen, 2005].

The last two exponentials contain the payoff elements of the game. These are randomly generated using a multi-variate gaussian and then held fixed for the rest of the game. We employ the mean and covariance expressions given in (4) to average over all possible realisations of these elements. This gives us a new form for an averaged generating functional. The technical details of how this averaging is performed is given in the supplementary material (Sec S2.3-S2.5).

As mentioned, in the same way that (3) gives rise to the functional $Z(\vec{\psi})$, so the averaged functional can be thought of as having arisen from averaged dynamics, the aforementioned 'effective dynamics':

$$\begin{aligned} \frac{1}{x}\frac{d}{dt}x(t) &= \alpha^2\tilde{\tau}^2\Gamma\int dt'\left[G(t,t')C^{p-2}(t,t')x(t')\right] \\ &\quad +\sqrt{2}\alpha\tilde{\tau}\eta_1(t) + \sqrt{2}\alpha\hat{\tau}\eta_0(t) + \tilde{\alpha}\rho(t), \end{aligned} \quad (6)$$

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, 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 that the substantial impact of the assumption that all actions of all agents are i.i.d. This is a strong assumption that removes the interdependency between agents in the analysis and is required to ensure that (6) 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 parameters p, α, τ, and Γ.

3.3 Linear Stability Analysis

In this section we determine the condition under which the effective dynamics will converge to a stable fixed point. The process is as follows. We first find the expression for a fixed point x_∞ of (6) by letting $\dot{x}(t) = 0$ and find a linear approximation of (6) close to this point. This gives an expression for the dynamics as a function of time. By taking the Fourier transform of this expression, we can examine the behaviour of the system at low frequencies (i.e. at large times) [Sewell, 2005].

$$\begin{aligned} 0 &= x_\infty[\alpha^2\tilde{\tau}^2\Gamma x_\infty q^{p-2}\chi + \sqrt{2}\alpha\tilde{\tau}q^{(p-1)/2}z \\ &\quad +\sqrt{2}\alpha\hat{\tau}q^{p/2}z^{p/p-1} + \tilde{\alpha}\rho] \end{aligned} \quad (7)$$

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 (7) we can calculate x_∞. We disregard the choice $x_\infty = 0$,

since no action will be chosen with exactly zero probability [Coolen, 2005]. 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 $\Gamma \in [-1, 0]$.

We analyse the stability of (6) in a neighbourhood around this supposed fixed point, by following a similar procedure as in [Opper and Diederich, 1992], in which the fixed point dynamics are 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 (since the deviations are considered to be small), we obtain

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

We then examine the long-term behaviour of the perturbation $\hat{x}(t)$ by taking the Fourier transform of this linearisation, and analysing its behaviour at $\omega = 0$. After some manipulation (Sec. S4), we arrive at the condition that, at a fixed point, the long term perturbations must satisfy

$$0 \leq \left[(\alpha^2\tilde{\tau}^2\Gamma q^{p-2}\chi)^2 - 2(\alpha\tilde{\tau} + \alpha\hat{\tau})^2(p-1)q^{p-2}\right] \quad (8)$$

It should be noted that, to derive (8), we assumed that the number p of players is large enough so that $p/p - 1 \approx 1$ and therefore the fractional power on z in (7) is reduced to 1. As z can take any value, including negative values, this again ensures that the expression for x_∞ remains real-valued. We discuss the effects of taking this limit, as well as $N \to \infty$ in Sec. 4.

3.4 Discussion

The fixed point condition (8) yields a number of verifiable implications, of which we test empirically the validity in Sec. 4. These implications are as follows.

1. 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 (8) evaluating to zero, which falls within the stable region.

2. Convergence is rare in the limit $N \to \infty$. 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 (8) is negative, which violates the stability criterion. As such, in the limit of infinite actions and a large number of players, games will not converge. Of course, this is not typically required in most normal form games. We are interested, therefore, in what the condition (8) predicts about the dependence of stability on the parameters rather than the size of the stable region itself.

3. The likelihood of convergence decreases for increasing α and τ, regardless of the choice of N and p. We can see this since the right hand side of (8) tends further away from zero (i.e., further from the region of stability).

4. The choice of Γ has a negligible effect on 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 Fig. 1.

5. The likelihood of convergence decreases as p increases. We see this by noticing the $(p-1)$ in the second term of (8). As p increases, this term drives the system further from the stability boundary.

We illustrate these implications in Fig. 1 which plots the value obtained by the right hand side of (8). In order to determine these values, we first solve (7) for q and χ. We determine the value of x_∞ that solves the expression inside the square brackets of (7) using a Newton-Raphson root-finding approach. It should be noted that this method of finding x_∞ yields only an approximate value, and therefore the q and χ which are found are estimates.

Finally, we consider each of the implications (2)-(5) in turn, choosing to forego illustrating (1) as it is immediate from (8). 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 1), we see that, for all choices of parameters, (8) evaluates to a value less then zero, which signals instability. As such, the analysis suggests that learning is everywhere unstable. 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. Indeed, this result holds also for the limiting case of $N \to \infty$. It would, therefore, be important to empirically assess whether this holds in games of finite actions and a lower number of players, a point which we discuss in Section 4.

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 (8). We go on to verify that the effect of p, α, Γ and τ are accurately captured by (8).

We interpret (8), 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 Fig. 1 should be considered to show the likelihood of convergence for a choice of (α, Γ). Lighter regions denote greater convergence and vice versa.

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 independently from 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 (8). 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

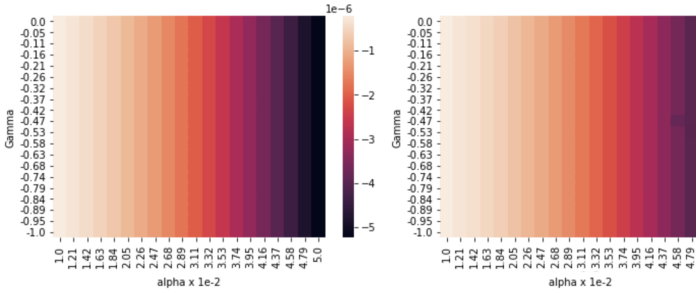

Figure 1: Evaluation of (8) over a range of α and Γ values. Light regions indicate where (8) is close to zero (i.e. close to the stabili boundary) whilst darker regions are further away. We vary α in th range $[0.01, 0.05]$ and Γ in the range $[-1, 0]$. $\tau = 0.05$. (Lef $p = 2, N = 2$, (Right) $p = 3, N = 8$.

and τ^4 terms appearing in (8).

Implication 5 can, as discussed, be seen directly from (8 itself. The prediction is that (8) will decrease (i.e. move awa from the stability boundary) monotonically with p. This due to the $(p-1)$ term which appears in the equation.

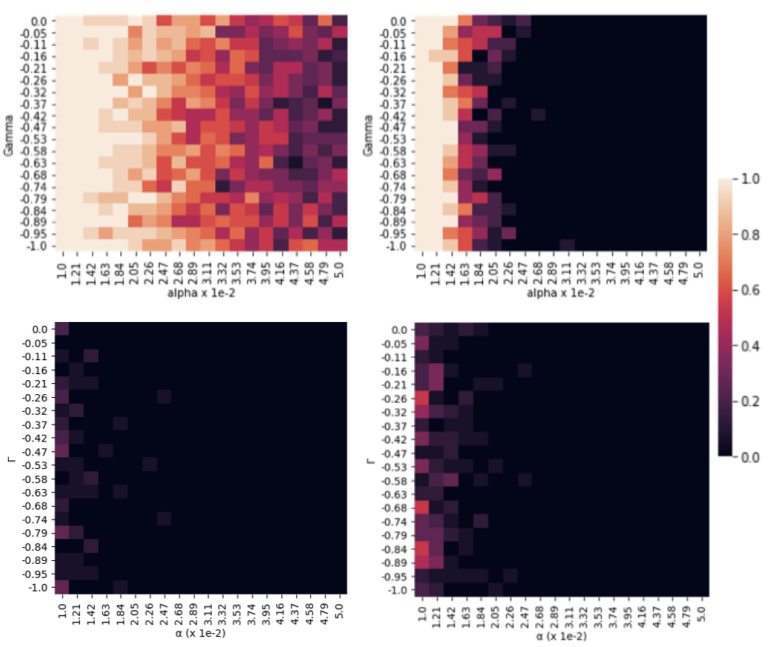

Figure 2: Results of numerical experiments in which α is varied. The heatmaps show the fraction of games which converge to an equilibrium. Lighter values show higher convergence. Each experiment is run with $\tau = 0.05$. (Top Left) $p = 2, N = 2$, (Top Right) $p = 2, N = 5$, (Bottom Left) $p = 3, N = 5$, (Bottom Right) $p = 3, N = 8$.

4 Experimental Evaluation

As mentioned, the analytic result (8) holds in the limit of a large number of players and an infinite number of actions. It is therefore prudent to assess whether Implications 2–5 hold for games which are not in this limit. In this section, we describe and discuss the numerical experiments to verify the implications in Sec. 3.4. We first describe how these experiments were conducted, and then we discuss the correlation with the theory.

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 Fig. 2, with parameters chosen to match those in the analytic assessment (Fig. 1).

To generate the numerical simulations in Fig. 2 we used the following procedure.

1. Fix the parameters τ and γ. The former is held at 0.05 and the latter at 0.1.
2. Initialise values of Γ and α. These will be swept over in the experiment.
3. Generate 15 payoff matrices by sampling from a multivariate 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 5×10^4 iterations.
6. Keep track of the action probabilities over a window of 5000 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 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 Fig. 2 that the stability of the system is highly dependent on the value of α and τ, but not on Γ.

Finally, in Fig. 3 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

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

This gives a measure of the degree of variability about the fixed point and, therefore, gives some notion of the 'degree of instability'.

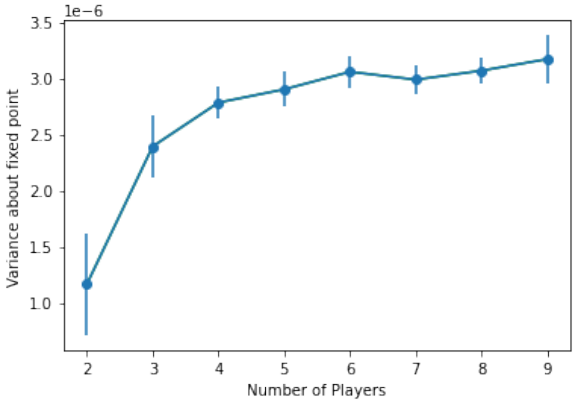

Figure 3: Variation $V(t)$ about a fixed point as p ranges from 2 - 9. $N = 2$, $\alpha = 0.02$, $\tau = 0.05$, $\Gamma \in [-1, 0]$. The experiment is conducted for 10 choices of Γ and the mean (line) is plotted alongside the standard deviation (vertical bars).

Due to computational constraints, which we discuss in the following section, we are only able to increase p to a maximum value of 9 . We maintain $\alpha = 0.02$ and $\tau = 0.05$. Finally, in Fig. 3, we are averaging over 10 choices of Γ in the range $[-1, 0]$.

4.2 Discussion

We observe that the numerical experiments confirm the predictions suggested by the analytic results in Fig. 3. We notice that convergence occurs almost only for low values of α. This observation stands as we increase the number N of actions. As noted in Sec. 3.4, the analytic result over-estimates the region of instability due to the assumptions $N \to \infty$ and $\frac{1}{p-1} \ll 2$ 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.

A key point which we wish to discuss here is a comparison with the result found in [Sanders *et al.*, 2018], which considers 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. This highlights the importance of performing analyses such as the present work; it allows for a method to analytically compare the differences 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 (8) provides a strong assessment of the effects that parameters α, τ, Γ, and p have on the likelihood of convergence of Q-learning. Furthermore, though (8) overestimates the region of instability by assuming $N \to \infty$, it accurately conveys that convergence of Q-learning is increasingly 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.

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 [Tuyls *et al.*, 2006]. 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 for negatively correlated payoff matrices, the strength of correlation Γ does not influence stability; whereas, as α and τ increase, the likelihood of convergence decreases. This behaviour differs from the stability of EWA, suggesting that different learning methods produce different qualitative behaviours. Similarly to EWA, however our analysis also shows that the likelihood of convergence decreases, regardless of parameter choice as the number of players p in the system increases.

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

[Bloembergen *et al.*, 2015] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53, 2015.

[Camerer and Ho, 1999] Colin Camerer and Teck Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, (4), 1999.

[Chotibut *et al.*, 2020] Thiparat Chotibut, Fryderyk Falniowski, Michael Misiurewicz, and Georgios Piliouras. Family of chaotic maps from game theory. *Dynamical Systems*, 2020.

[Coolen, 2005] A.C.C. Coolen. *The Mathematical Theory of Minority Games: Statistical Mechanics of Interacting Agents (Oxford Finance Series)*. Oxford University Press, Inc., 2005.

[Galla and Farmer, 2013] Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences of the United States of America*, (4), 2013.

[Galla, 2011] Tobias Galla. Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2011.

[Leonardos and Piliouras, 2020] Stefanos Leonardos and Georgios Piliouras. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. 12 2020.

[Leslie and Collins, 2006] David S Leslie and E J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 2006.

[Mezard, 1987] N. Mezard. *Spin Glass Theory and Beyond*. World Scientific, 1987.

[Opper and Diederich, 1992] Manfred Opper and Sigurd Diederich. Phase transition and 1/f noise in a game dynamical model. *Physical Review Letters*, 69(10):1616–1619, 1992.

[Sanders *et al.*, 2018] James B. T. Sanders, J. Doyne Farmer, and Tobias Galla. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 2018.

[Schwartz, 2014] Howard M. Schwartz. *Multi-Agent Machine Learning: A Reinforcement Approach*. 2014.

[Sewell, 2005] Granville Sewell. *Appendix B - The Fourier Stability Method*. John Wiley and Sons Inc., 2005.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[Strogatz, 2000] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, 2000.

[Sutton and Barto, 2018] R Sutton and A Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[Tuyls *et al.*, 2006] Karl Tuyls, Pieter Jan T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. 2006.

[van Strien and Sparrow, 2011] Sebastian van Strien and Colin Sparrow. Fictitious play in 3×3 games: Chaos and dithering behaviour. *Games and Economic Behavior*, 73, 2011.

[von Neumann *et al.*, 1944] John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944.

[Zinn-Justin, 2002] J Zinn-Justin. *Quantum field theory and critical phenomena*. Clarendon Press, 2002.