Derivation Write Up

Aamal Hussain

June 16, 2020

FB: the bibfile is missing.

1 Introduction

- For use in Multi-Agent Reinforcement Learning
- Requirement to get some desired behaviours out of the system (e.g. conforming to predefined values or constraints)
- For this, it is required that we are able to predict the outcome of the learning behaviour
- Work done on determining the dynamics of MARL gives an indication of what solutions an algorithm will produce
- We can use this, alongside stability theory, to understand the nature of the games which result in stable behaviours and which result in chaotic behaviours.
- Largely follows the work presented by Sanders et al. However, whilst they consider EWA, we consider Q-Learning, which is more popular in RL communities.

1.1 Problem Statement

Given a particular learning algorithm and game, is it possible to determine whether the game is likely to reach a unique fixed point or exhibit more complex behaviour? What are the factors which affect this resulting behaviour?

1.2 Objectives and Scope

Analyse the stability of Multiagent Q-Learning on iterated games.

Assumptions:

- Focuses on stateless games
- Homogeneous agents
- Discrete action spaces, with large number of actions.
- Small, finite number of agents.

The work by Sanders et al. considers agents who have these same assumptions, but learn with a reduced version of the Experience Weighted Attraction algorithm. For the Reinforcement Learning community, it is of interest to consider the question of stability from the point of view of agents who follow the popular Q-Learning algorithm.

FB: you might discuss in which way any of the points above differs from the relevant reference.

2 Derivation

FB: it would be good to have a overall description of the goal of the derivation.

We start with the two-agent Q-Learning dynamics as presented by Tuyls et al.

$$\frac{\dot{x}(t)}{x(t)} = \alpha\tau(\sum_j a_{ij}y_j - \sum_{ij} x_i a_{ij} y_j) + \alpha \sum_j x_j ln(\frac{x_j}{x_i}) \tag{1a}$$

$$\frac{\dot{y}(t)}{y(t)} = \alpha\tau(\sum_j b_{ij}x_j - \sum_{ij} y_i b_{ij} x_j) + \alpha \sum_j y_j ln(\frac{y_j}{y_i}). \tag{1b}$$

Here, α and τ are the parameters of the agent; Sanders et al. refer to these as the memory and intensity of choice parameters respectively. Agent 1 takes action i with probability x_i while Agent 2 takes action j with probability y_j. If these actions are taken, the agents receive payoff a_{ij} and b_{ji} respectively.

2.1 Rescaling of Variables

In order to follow the conventions of spin glass theory for the analysis of disordered systems, we rescale the system so that the payoff matrix elements are of order $N^{-1/2}$. The motivation for doing this is that, along the way, we will take the limit of the number of actions N to infinity. However, doing this will result in numerical underflow of the action probabilities (i.e. the probability of each action goes to zero). To compensate for this, we adjust the system so that the sum of the probabilities add to N. The scaling goes as follows:

$$a_{ij} = \sqrt{N}\tilde{a_{ij}} \tag{2a}$$

$$b_{ji} = \sqrt{N}\tilde{b_{ji}} \tag{2b}$$

We compensate for this change with

$$x_i = \tilde{x_i}/N \tag{3a}$$

$$y_i = \tilde{y_i}/N. \tag{3b}$$

This gives the original equations as

$$\frac{\dot{\tilde{x}}_i(t)}{\tilde{x}_i(t)} = \alpha\tilde{\tau}\sum_j \tilde{a_{ij}}\tilde{y_j} - \tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij}\tilde{x_i}\tilde{a_{ij}}\tilde{y_j} + \tilde{\alpha}\sum_j \tilde{x_j} ln(\frac{\tilde{x_j}}{\tilde{x_i}}) \tag{4a}$$

$$\frac{\dot{\tilde{y}}_i(t)}{\tilde{y}_i(t)} = \alpha\tilde{\tau}\sum_j \tilde{b_{ij}}\tilde{x_j} - \tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij}\tilde{y_i}\tilde{b_{ij}}\tilde{x_j} + \tilde{\alpha}\sum_j \tilde{y_j} ln(\frac{\tilde{y_j}}{\tilde{y_i}}). \tag{4b}$$

The need for the factor of $\frac{1}{\sqrt{N}}$ is to follow the conventions of the saddle point method of integration. This will become clear after taking the expectation of the generating functional. Note that we may drop the notation on time dependence for x and y. However, these will always be functions of time. Henceforth, we shall not write the tildes on x, y, a_{ij}, b_{ij}. We shall also abbreviate the final term as

$$\rho_{x,i}(t) = \sum_j x_j ln(\frac{x_j}{x_i})$$

2.2 Generating Functional

The generating functional allows us to take a path integral over all possible realisations of learning [**?**]. This is given as:

$$Z = \int D[\vec{x}, \vec{y}] \prod_i \delta(\textit{equation of motion}_i) exp(i \int dt[x_i(t)\psi_i(t) + y_i(t)\phi_i(t)]), \tag{5}$$

where the equations of motion are the Lagrange equations of motion given in (4) andSince the fields $\psi_i\vec{(t)}$ and $\phi_i\vec{(t)}$ will be set to zero at the end of the calculation. δ denotes the Dirac delta function. We write this in its Fourier representation, which yields

$$Z(\vec{\psi},\vec{\phi}) = \int D[\vec{x},\vec{\hat{x}},\vec{y},\vec{\hat{y}}] exp(i\sum_i \int dt[\hat{x}_i(\frac{\dot{x_i}(t)}{x_i(t)} - \alpha\tilde{\tau}\sum_j a_{ij}y_j + \tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} x_i a_{ij} y_j - \tilde{\alpha}\rho_{x,i}(t) + h_{x,i}(t))])$$
$$\times exp(i\sum_i \int dt[\hat{y}_i(\frac{\dot{y_i}(t)}{y_i(t)} - \alpha\tilde{\tau}\sum_j b_{ij}x_j + \tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} y_i b_{ij} x_j - \tilde{\alpha}\rho_{y,i}(t) + h_{y,i}(t))]) \quad (6)$$
$$\times exp(i\sum_i \int dt[x_i(t)\psi_i(t) + y_i(t)\phi_i(t))]),$$

Here, the terms a_{ij}, b_{ij} are the payoffs in the game and the term h denotes a field which will be set to zero at the end of the calculation. We recall that these are randomly generated using a multivariate gaussian and then held fixed for the rest of the game. We call this 'quenched disorder'. Isolating these terms allows us to rearrange the above as

$$Z(\vec{\psi},\vec{\phi}) = \int D[\vec{x},\vec{\hat{x}},\vec{y},\vec{\hat{y}}] exp(i\sum_i \int dt[\hat{x}_i(\frac{\dot{x_i}(t)}{x_i(t)} - \tilde{\alpha}\rho_{x,i}(t) + h_{x,i}(t))])$$
$$\times exp(i\sum_i \int dt[\hat{y}_i(\frac{\dot{y_i}(t)}{y_i(t)} - \tilde{\alpha}\rho_{y,i}(t) + h_{y,i}(t))]) \quad (7)$$
$$\times exp(i\sum_i \int dt[x_i(t)\psi_i(t) + y_i(t)\phi_i(t))])$$
$$\times exp(i\sum_i \int dt[-\hat{x}_i\alpha\tilde{\tau}\sum_j a_{ij}y_j + \hat{x}_i\tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} x_i a_{ij} y_j - \hat{y}_i\alpha\tilde{\tau}\sum_j b_{ij}x_j + \hat{y}_i\tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} y_i b_{ij} x_j]).$$

The only difference between (6) and (7) is that we moved the term FB: which one exactly? containing the quenched disorder into a separate exponential. Since our aim is to take an average over all possible realisations of this disorder, we will only need to focus on the last exponential which we rewrite as

$$Q = exp(i\sum_i \int dt[-\hat{x}_i\alpha\tilde{\tau}\sum_j a_{ij}y_j + \hat{x}_i\tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} x_i a_{ij} y_j - \hat{y}_i\alpha\tilde{\tau}\sum_j b_{ij}x_j + \hat{y}_i\tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ij} y_i b_{ij} x_j]). \quad (8)$$

We will then separate the terms so that like sums are paired together

$$Q = exp(-i\alpha\tilde{\tau}\sum_{ij}\int dt[\hat{x}_i a_{ij} y_j + \hat{y}_j b_{ji} x_i]) \times exp(i\tilde{\alpha}\tau\frac{1}{\sqrt{N}}\sum_{ijk}\int dt[\hat{x}_i x_j a_{jk} y_k + \hat{y}_i y_k b_{kj} x_j]) \quad (9)$$

It should be noted that we have changed some of the subscripts on the terms. Since these terms are all multiplied together and we sum over the subscripts, the letters we choose are of no importance and we can exchange them freely FB: perhaps some more justification might be helpful. We will define both exponentials in Q as Q_1 and Q_2 respectively.

We are now ready to take the expectation of Q. To do this, we will exploit the fact that the payoff elements are Gaussian distributed and use the identity [**?**]

$$\int dz[e^{-A_2(z)+\vec{b}\cdot\vec{z}}] = (2\pi)^{k/2}(det(A))^{-1/2}e^{\omega(b)}, \quad (10)$$

where

$$A_2(z) = 1/2\sum_{ij} z_i A_{ij} z_j$$
$$\omega_2(z) = 1/2\sum_{ij} b_i (A)^{-1}_{ij} b_j$$

2.2.1 Expectation of Q_1

We can rewrite Q_1 as

$$Q_1 = \prod_{ij} exp(\vec{b} \cdot \vec{z}),$$

where

$$\begin{aligned} b &\coloneqq [-i\alpha\tilde{\tau} \int dt[\hat{x}_i y_j], -i\alpha\tilde{\tau} \int dt[\hat{y}_j x_i]]^T \\ z &\coloneqq [a_{ij}, b_{ji}]^T \\ A &\coloneqq \Sigma^{-1} \\ \Sigma_{ij} &\coloneqq Cov[z_i, z_j], \end{aligned}$$

where Σ is the covariance of $\vec{z}$. We recall that the scaled system has payoff elements chosen so that

$$\begin{aligned} \mathbb{E}[a_{ij}] = \mathbb{E}[b_{ji}] &= 0 \\ \mathbb{E}[a_{ij}^2] = \mathbb{E}[b_{ji}^2] &= 1/N \\ \mathbb{E}[a_{ij} b_{ji}] &= \Gamma/N. \end{aligned}$$

Applying identity (10) gives:

$$\begin{aligned} \mathbb{E}[Q_1] = \prod_{ij} exp(-\alpha^2\tilde{\tau}^2 \frac{1}{2N} \int dtdt'[&\hat{x}_i(t)\hat{x}_i(t')y_j(t)y_j(t') + \hat{y}_j(t)\hat{y}_j(t')x_i(t)x_i(t') \\ &+\Gamma\hat{x}_i(t)x_i(t')y_j(t)\hat{y}_j(t') + \Gamma\hat{y}_j(t)y_j(t')x_i(t)\hat{x}_i(t')]). \end{aligned} \tag{11}$$

2.2.2 Expectation of Q_2

We take a similar approach with the following definitions

$$\begin{aligned} b &\coloneqq [i\tilde{\alpha}\tau \int dt[\hat{x}_i x_j y_k], i\tilde{\alpha}\tau \int dt[\hat{y}_i x_j y_k]]^T \\ z &\coloneqq [a_{jk}, b_{kj}]^T \\ A &\coloneqq \Sigma^{-1} \\ \Sigma_{ij} &\coloneqq Cov[z_i, z_j], \end{aligned}$$

Following the same procedure as for Q_1 yields

$$\begin{aligned} \mathbb{E}[Q_2] = \prod_{ij} exp(-\tilde{\alpha}^2\tau^2 \frac{1}{2N^2} \int dtdt'[&\hat{x}_i(t)\hat{x}_i(t')x_j(t)x_j(t')y_k(t)y_k(t') \\ &+\hat{y}_i(t)\hat{y}_i(t')x_j(t)x_j(t')y_k(t)y_k(t') \\ &+\Gamma\hat{x}_i(t)\hat{y}_i(t')x_j(t)x_j(t')y_k(t)y_k(t') \\ &+\Gamma\hat{y}_i(t)\hat{x}_i(t')x_j(t)x_j(t')y_k(t)y_k(t')]). \end{aligned} \tag{12}$$

We now define the correlation functions

$$C_x(t,t') = N^{-1}\sum_i x_i(t)x_i(t') \qquad C_y(t,t') = N^{-1}\sum_i y_i(t)y_i(t')$$
$$L_x(t,t') = N^{-1}\sum_i \hat{x}_i(t)\hat{x}_i(t') \qquad L_y(t,t') = N^{-1}\sum_i \hat{y}_i(t)\hat{y}_i(t')$$
$$K_x(t,t') = N^{-1}\sum_i x_i(t)\hat{x}_i(t') \qquad K_y(t,t') = N^{-1}\sum_i y_i(t)\hat{y}_i(t')$$
$$A_{xy}(t,t') = N^{-1}\sum_i \hat{x}_i(t)\hat{y}_i(t').$$

We then rewrite $\mathbb{E}[Q]$ as

$$\begin{aligned}\mathbb{E}[Q] = exp(-\alpha^2\tilde{\tau}^2\frac{N}{2}\int dtdt'[L_x(t,t')C_y(t,t') + L_y(t,t')C_x(t,t') + 2\Gamma K_x(t,t')K_y(t',t)] \\ -\tilde{\alpha}^2\tau^2\frac{N}{2}\int dtdt'[L_x(t,t')C_x(t,t')C_y(t,t') + L_y(t,t')C_x(t,t')C_y(t,t') \\ +\Gamma A_{xy}(t,t')C_x(t,t')C_y(t,t') + \Gamma A_{xy}(t',t)C_x(t,t')C_y(t,t')])\end{aligned} \tag{13}$$

We can introduce these correlation functions into the expectation which gives

$$\mathbb{E}[Q] = \int D[C_x,\hat{C}_x,L_x,\hat{L_x},K_x,\hat{K_x},C_y,\hat{C}_y,L_y,\hat{L_y},K_y,\hat{K_y},A_{xy},\hat{A}_{xy}]exp(N(\Psi,\Phi,\Lambda)), \tag{14}$$

where

$$\begin{aligned}\Psi = i\int dtdt'[\hat{C}_x(t,t')C_x(t,t') + \hat{L}_x(t,t')L_x(t,t') + \hat{K}_x(t,t')K_x(t,t') \\ +\hat{C}_y(t,t')C_y(t,t') + \hat{L}_y(t,t')L_y(t,t') + \hat{K}_y(t,t')K_y(t,t') + \hat{A}_{xy}(t,t')A_{xy}(t,t')]\end{aligned} \tag{15}$$

$$\begin{aligned}\Phi = -\alpha^2\tilde{\tau}^2\frac{N}{2}\int dtdt'[L_x(t,t')C_y(t,t') + L_y(t,t')C_x(t,t') + 2\Gamma K_x(t,t')K_y(t',t)] \\ -\tilde{\alpha}^2\tau^2\frac{N}{2}\int dtdt'[L_x(t,t')C_x(t,t')C_y(t,t') + L_y(t,t')C_x(t,t')C_y(t,t') \\ +\Gamma A_{xy}(t,t')C_x(t,t')C_y(t,t') + \Gamma A_{xy}(t',t)C_x(t,t')C_y(t,t')]\end{aligned} \tag{16}$$

$$\begin{aligned}\Lambda = i\sum_i\int dtdt'[\hat{C}_x(t,t')x_i(t)x_i(t') + \hat{L}_x(t,t')\hat{x}_i(t)\hat{x}_i(t') + \hat{K}_x(t,t')x_i(t)\hat{x}_i(t') \\ +\hat{C}_y(t,t')y_i(t)y_i(t') + \hat{L}_y(t,t')\hat{y}_i(t)\hat{y}_i(t') + \hat{K}_y(t,t')y_i(t)\hat{y}_i(t') \\ +\hat{A}_{xy}(t,t')\hat{x}_i(t)\hat{y}_i(t')].\end{aligned} \tag{17}$$

We insert this expectation back into the original generating functional which gives

$$\mathbb{E}[Z(\vec{\psi},\vec{\phi})] = \int D[C_x,\hat{C}_x,L_x,\hat{L_x},K_x,\hat{K_x},C_y,\hat{C}_y,L_y,\hat{L_y},K_y,\hat{K_y},A_{xy},\hat{A}_{xy}]exp(N(\Psi,\Phi,\Omega + (\mathcal{O}(N^{-1})))), \tag{18}$$

where Ω includes all terms describing the time evolution of the system and is given by

$$\Omega = N^{-1}\sum_i log\int D[x_i,\hat{x}_i,y_i,\hat{y}_i]exp(i\int dt[\hat{x}_i(\frac{\dot{x_i}(t)}{x_i(t)}-\tilde{\alpha}\rho_{x,i}(t)+h_{x,i}(t))])$$
$$\times exp(i\int dtdt'[\hat{C}_x(t,t')x_i(t)x_i(t')+\hat{L}_x(t,t')\hat{x}_i(t)\hat{x}_i(t')+\hat{K}_x(t,t')x_i(t)\hat{x}_i(t')])$$
$$\times exp(i\int dt[\hat{y}_i(\frac{\dot{y_i}(t)}{y_i(t)}-\tilde{\alpha}\rho_{y,i}(t)+h_{y,i}(t))]) \quad (19)$$
$$\times exp(i\int dtdt'[\hat{C}_y(t,t')y_i(t)y_i(t')+\hat{L}_y(t,t')\hat{y}_i(t)\hat{y}_i(t')+\hat{K}_y(t,t')y_i(t)\hat{y}_i(t')])$$
$$\times exp(i\int dtdt'[\hat{A}_{xy}(t,t')\hat{x}_i(t)\hat{y}_i(t')]])\times exp(i\sum_i\int dt[x_i(t)\psi_i(t)+y_i(t)\phi_i(t))]).$$

We will evaluate the path integral using the saddle point method for integration [**?**]. In this method, we consider that the integration is dominated by the maximum of the function

$$f=\Psi+\Phi+\Omega,$$

and we take the limit as N extends to infinity. We therefore determine the relations which maximise this function. We find

$$\frac{\partial f}{\partial C_x(t,t')} \implies i\hat{C}_x(t,t')=\frac{\alpha^2\tilde{\tau}^2}{2}L_y(t,t')+\frac{\tilde{\alpha}^2\tau^2}{2}(L_x(t,t')C_y(t,t')+L_y(t,t')C_y(t,t')+2\Gamma A_{xy}(t,t')C_y(t,t'))$$
$$\frac{\partial f}{\partial L_x(t,t')} \implies i\hat{L}_x(t,t')=\frac{\alpha^2\tilde{\tau}^2}{2}C_y(t,t')+\frac{\tilde{\alpha}^2\tau^2}{2}(C_x(t,t')C_y(t,t'))$$
$$\frac{\partial f}{\partial K_x(t,t')} \implies i\hat{K}_x(t,t')=\alpha^2\tilde{\tau}^2\Gamma K_y(t',t)$$
$$\frac{\partial f}{\partial C_y(t,t')} \implies i\hat{C}_y(t,t')=\frac{\alpha^2\tilde{\tau}^2}{2}L_x(t,t')+\frac{\tilde{\alpha}^2\tau^2}{2}(L_x(t,t')C_x(t,t')+L_y(t,t')C_x(t,t')+2\Gamma A_{xy}(t,t')C_x(t,t'))$$
$$\frac{\partial f}{\partial L_y(t,t')} \implies i\hat{L}_y(t,t')=\frac{\alpha^2\tilde{\tau}^2}{2}C_x(t,t')+\frac{\tilde{\alpha}^2\tau^2}{2}(C_x(t,t')C_y(t,t'))$$
$$\frac{\partial f}{\partial K_y(t,t')} \implies i\hat{K}_y(t,t')=\alpha^2\tilde{\tau}^2\Gamma K_x(t',t)$$
$$\frac{\partial f}{\partial A_{xy}(t,t')} \implies i\hat{A}_{xy}(t,t')=\tilde{\alpha}^2\tau^2\Gamma C_x(t,t')C_y(t,t').$$

Similarly,

$$
\begin{aligned}
\frac{\partial f}{\partial \hat{C}_x(t,t')} &\implies C_x(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle x_i(t)x_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial\psi_i(t)\partial\psi_i(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{L}_x(t,t')} &\implies L_x(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle \hat{x}_i(t)\hat{x}_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial h_{x,i}(t)\partial h_{x,i}(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{K}_x(t,t')} &\implies K_x(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle x_i(t)\hat{x}_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial\vec{\psi}(t)\partial h_{x,i}(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{C}_y(t,t')} &\implies C_y(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle y_i(t)y_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial\phi_i(t)\partial\phi_i(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{L}_y(t,t')} &\implies L_y(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle \hat{y}_i(t)\hat{y}_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial h_{y,i}(t)\partial h_{y,i}(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{K}_y(t,t')} &\implies K_y(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle y_i(t)\hat{y}_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial\vec{\phi}(t)\partial h_{x,i}(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}\\
\frac{\partial f}{\partial \hat{A}_{xy}(t,t')} &\implies A_{xy}(t,t') = \lim_{N\to\infty} N^{-1}\sum_i \langle \hat{x}_i(t)\hat{y}_i(t')\rangle_\Omega = -\lim_{N\to\infty}\sum_i \frac{\partial^2 \mathbb{E}[Z(\psi,\phi)]}{\partial\partial h_{x,i}(t)\partial h_{y,i}(t')}\Big|_{\vec{\phi}=\vec{\psi}=\vec{h}=0}.
\end{aligned}
$$

We implement these relations into the expression of Ω, and make the assumption that all actions i are independent and identically distributed (i.i.d.), which gives

$$
\begin{aligned}
\Omega = log\int D[x,\hat{x},y,\hat{y}]exp(i\int dt[\hat{x}(t)(\frac{\dot{x}(t)}{x(t)}-\tilde{\alpha}\rho_x(t))])\\
\times exp(-\int dtdt'[\frac{1}{2}\alpha^2\tilde{\tau}^2C_y(t,t')\hat{x}(t)\hat{x}(t')+\frac{1}{2}\tilde{\alpha}^2\tau^2C_x(t,t')C_y(t,t')\hat{x}(t)\hat{x}(t')+i\alpha^2\tilde{\tau}^2\Gamma G_y(t',t)x(t)\hat{x}(t')])\\
\times exp(i\int dt[\hat{y}(t)(\frac{\dot{y}(t)}{y(t)}-\tilde{\alpha}\rho_y(t))])\\
\times exp(-\int dtdt'[\frac{1}{2}\alpha^2\tilde{\tau}^2C_x(t,t')\hat{y}(t)\hat{y}(t')+\frac{1}{2}\tilde{\alpha}^2\tau^2C_x(t,t')C_y(t,t')\hat{y}(t)\hat{y}(t')+i\alpha^2\tilde{\tau}^2\Gamma G_x(t,t')y(t)\hat{y}(t')])\\
\times exp(-\int dtdt'[\tilde{\alpha}^2\tau^2\Gamma C_x(t,t')C_y(t,t')\hat{x}(t)\hat{y}(t')]).
\end{aligned}
\tag{20}
$$

Since this contains all of the information of the learning evolution, we consider Ω as an effective generating functional (and in fact we see that it has a similar structure to (2.2), without the existence of the fields ψ,ϕ). In particular we recognise this as the generating functional of the 'effective dynamics' given as

$$
\begin{aligned}
\dot{x}(t) &= x(t)(\Gamma\alpha^2\tilde{\tau}^2\int dt'[G_y(t,t')x(t')]+\tilde{\alpha}\rho_x(t)+\alpha\tilde{\tau}\eta_x(t)+\tilde{\alpha}\tau\eta_x(t)\eta_y(t)+\sqrt{\Gamma}\tilde{\alpha}\tau\mu_x)\\
\dot{y}(t) &= y(t)(\Gamma\alpha^2\tilde{\tau}^2\int dt'[G_x(t,t')y(t')]+\tilde{\alpha}\rho_y(t)+\alpha\tilde{\tau}\eta_y(t)+\tilde{\alpha}\tau\eta_x(t)\eta_y(t)+\sqrt{\Gamma}\tilde{\alpha}\tau\mu_y),
\end{aligned}
\tag{21}
$$

with the self-consistency relations

$$
G_x(t,t') = \langle\frac{\delta x(t)}{\delta\eta_x(t')}\rangle \qquad G_y(t,t') = \langle\frac{\delta y(t)}{\delta\eta_y(t')}\rangle \tag{22}
$$

$$
\langle\eta_x(t)\eta_x(t')\rangle = C_y(t,t') \qquad \langle\eta_y(t)\eta_y(t')\rangle = C_x(t,t') \tag{23}
$$

$$
\langle\mu_x(t)\mu_y(t')\rangle = C_x(t,t')C_y(t,t') \qquad \langle\mu_x(t)\mu_x(t')\rangle = \langle\mu_y(t)\mu_y(t')\rangle = 0 \tag{24}
$$

3 Stability Analysis

We are now in a position to take the effective dynamics, which describes the evolution of the learning dynamics after averaging over all possible realisations of the payoff elements, and determine the stability of the system at

fixed points. We will follow the procedure laid out by Opper et al [**?**]. First, we rewrite $x(t)$, $y(t)$ as perturbations about their fixed points. We will then analyse the stability of these fixed points.

Let

$$x(t) = x^* + \hat{x}(t) \tag{25}$$
$$y(t) = y^* + \hat{y}(t) \tag{26}$$
$$\eta_x(t) = \eta_x^* + \hat{\nu}_x(t) \tag{27}$$
$$\eta_y(t) = \eta_y^* + \hat{\nu}_y(t) \tag{28}$$
$$\mu_x(t) = \mu_x^* + \hat{\delta}_x(t) \tag{29}$$
$$\mu_y(t) = \mu_y^* + \hat{\delta}_y(t), \tag{30}$$

where terms denoted by $\cdot^*$ are the values that are taken at the fixed point and terms denoted by $\hat{\cdot}$ refer to deviations from the fixed point. This is not to be confused with the conjugate variable notation that we used in the previous section. We assume that these perturbations arise from additive noise, $\xi(t)$, $\zeta(t)$, drawn from the unit normal distribution which are applied to the dynamics. Rewriting the dynamics with all of these considerations gives

$$\begin{aligned}
\frac{d}{dt}(x^* + \hat{x}(t)) = (x^* + \hat{x}(t))(\Gamma\alpha^2\tilde{\tau}^2 \int dt'[G_y(t,t')(x^* + \hat{x}(t'))] + \tilde{\alpha}\rho_x(t) + \alpha\tilde{\tau}(\eta_x^* + \hat{\nu}_x(t)) \\
+\tilde{\alpha}\tau(\eta_x^* + \hat{\nu}_x(t))(\eta_y^* + \hat{\nu}_y(t)) + \sqrt{\Gamma}\tilde{\alpha}\tau(\mu_x^* + \hat{\delta}_x(t)) + \xi(t)) \\
\frac{d}{dt}(y^* + \hat{y}(t)) = (y^* + \hat{y}(t))(\Gamma\alpha^2\tilde{\tau}^2 \int dt'[G_x(t,t')(y^* + \hat{y}(t'))] + \tilde{\alpha}\rho_y(t) + \alpha\tilde{\tau}(\eta_y^* + \hat{\nu}_y(t)) \\
+\tilde{\alpha}\tau(\eta_x^* + \hat{\nu}_x(t))(\eta_y^* + \hat{\nu}_y(t)) + \sqrt{\Gamma}\tilde{\alpha}\tau(\mu_y^* + \hat{\delta}_y(t)) + \zeta(t)).
\end{aligned} \tag{31}$$

Considering only terms which are linear in the deviations yields

$$\begin{aligned}
\frac{d}{dt}\hat{x}(t) = (x^* + \hat{x}(t)) \left[\Gamma\alpha^2\tilde{\tau}^2 x^* \int G_y(t-t')dt' + \tilde{\alpha}\rho_x^* + \alpha\tilde{\tau}\eta_x^* + \tilde{\alpha}\tau\eta_x^*\eta_y^* + \Gamma\tilde{\alpha}\tau\mu_x\right] \\
+x^* \left[\Gamma\alpha^2\tilde{\tau}^2 \int G_y(t-t')\hat{x}(t')dt' + \alpha\tilde{\tau}\hat{\nu}_x(t) + \tilde{\alpha}\tau\eta_x^*\hat{\nu}_y(t) + \tilde{\alpha}\tau\eta_y^*\hat{\nu}_x(t) + \sqrt{\Gamma}\tilde{\alpha}\tau\hat{\delta}_x(t) + \xi(t)\right] \\
\frac{d}{dt}\hat{y}(t) = (y^* + \hat{y}(t)) \left[\Gamma\alpha^2\tilde{\tau}^2 x^* \int G_x(t-t')dt' + \tilde{\alpha}\rho_y^* + \alpha\tilde{\tau}\eta_y^* + \tilde{\alpha}\tau\eta_x^*\eta_y^* + \Gamma\tilde{\alpha}\tau\mu_y\right] \\
+y^* \left[\Gamma\alpha^2\tilde{\tau}^2 \int G_x(t-t')\hat{y}(t')dt' + \alpha\tilde{\tau}\hat{\nu}_y(t) + \tilde{\alpha}\tau\eta_x^*\hat{\nu}_y(t) + \tilde{\alpha}\tau\eta_y^*\hat{\nu}_x(t) + \sqrt{\Gamma}\tilde{\alpha}\tau\hat{\delta}_y(t) + \zeta(t)\right].
\end{aligned} \tag{32}$$

If we consider the long term behaviour of the system near a stable fixed point, then $\hat{x}(t)$ goes to zero as $t \to \infty$. This yields the requirement that a fixed point x^* must satisfy

$$0 = x^* \left[\Gamma\alpha^2\tilde{\tau}^2 x^* \int G_y(t-t')dt' + \tilde{\alpha}\rho_x^* + \alpha\tilde{\tau}\eta_x^* + \tilde{\alpha}\tau\eta_x^*\eta_y^* + \sqrt{\Gamma}\tilde{\alpha}\tau\mu_x\right]. \tag{33}$$

The implication here is that the resulting value of x^* can take two solutions. However, one of these is $x^* = 0$, which is found to rarely occur, while the second solution contains a term $\sqrt{\Gamma}$, which would yield complex values. If this is true, then it would imply that the learning algorithm will rarely converge for values of $\Gamma < 0$. Numerical experiments will shed more light on the likelihood of convergence. If this ansatz turns out to be incorrect (i.e. we see significant convergence for $\Gamma < 0$), it is unlikely that we will be able to solve for the stability line for this case.

We continue to follow the method of Opper et al. by taking the Fourier transform of (32). For both agents, the first term in square brackets is constant with respect to time and so will not affect the long term behaviour of the system. We can, therefore, ignore this and focus on the second term. For the sake of brevity, we will only write the equations for Agent 1 (i.e. for x), though the equations for y can be equivalently obtained. Taking the Fourier transform yields

$$\left[\frac{i\omega}{x^*} - \Gamma\alpha^2\tilde{\tau}^2\mathcal{G}_y(\omega)\right]\mathcal{X}(w) = \Gamma\alpha\tilde{\tau}\mathcal{V}_x(\omega) + \tilde{\alpha}\tau\eta_x^*\mathcal{V}_y(\omega) + \tilde{\alpha}\tau\eta_y^*\mathcal{V}_x(\omega) + \sqrt{\Gamma}\tilde{\alpha}\tau\Delta(\omega) + \Xi(\omega). \tag{34}$$

Then,

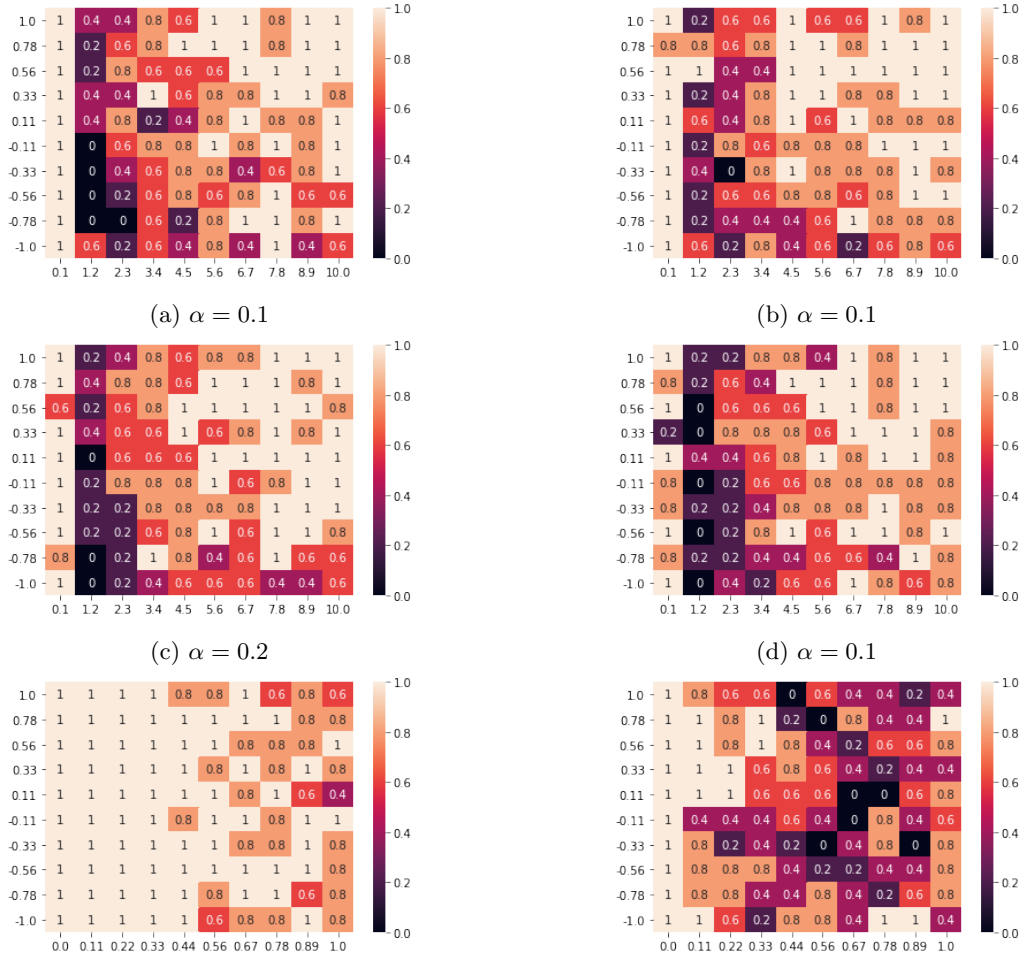
$$\left\langle [|\mathcal{X}(\omega)|^2 \right\rangle = \left\langle |\Gamma\alpha\tilde{\tau}\mathcal{V}_x(\omega) + \tilde{\alpha}\tau\eta_x^*\mathcal{V}_y(\omega) + \tilde{\alpha}\tau\eta_y^*\mathcal{V}_x(\omega) + \sqrt{\Gamma}\tilde{\alpha}\tau\Delta(\omega) + \Xi(\omega)|^2 \right\rangle \frac{1}{\langle |\mathcal{A}(\omega, x^*)|^2 \rangle}, \tag{35}$$

where

$$\mathcal{A}(\omega, x^*) = \frac{i\omega}{x^*} - \Gamma\alpha^2\tilde{\tau}^2\mathcal{G}_y(\omega) \tag{36}$$

The behaviour of this line for $\omega = 0$ gives the long term behaviour of the system. By considering that, for a stable fixed point to exist, $\left\langle [|\mathcal{X}(\omega = 0)|^2 \right\rangle$, we arrive at an expression for a stability line (i.e. the phase transition between the existence of stable fixed points and unstable behaviour).

4 Numerical Simulations



(a) $\alpha = 0.1$ (b) $\alpha = 0.1$

(c) $\alpha = 0.2$ (d) $\alpha = 0.1$

(e) $\tau = 0.0$ (f) $\tau = 0.15$

Figure 1: $\alpha = 0.1$, $\gamma = 0.1$, $\tau \in [0.1, 10]$, $\Gamma \in [-1, 1]$. Each simulation is run for 1e5 iterations and tested for convergence. The game is said to have converged based on a tolerance of 1% difference between action probabilities. For each combination of τ, Γ the game is played 5 times, each with random payoff matrices and initial conditions. The average number of converged games (giving an indication of probability of convergence) is shown in each cell of the heatmap.

To generate the numerical simulations in Figure 1 we used the following procedure.

1. Fix the parameters α and γ. The latter is held fixed at 0.1 since it does not affect the long term behaviour of the system (it does not appear in (1)).

2. We initialise values of Γ and τ. These will be the variables which we sweep over.

3. Generate payoff matrices for both agents by sampling from a multi-variate Gaussian (variables are the payoff elements) with covariance parameterised by Γ.

4. Iniitialise the agents with random initial conditions (i.e. random action probabilities).

5. Allow the agents to learn over a maximum of 1×10^5 iterations.

6. Every 100 iterations, check to see if the action probabilities have changed significantly. If not (i.e. the change over the last 100 iterations is less than 1%) the learning is considered to have converged.

7. This process is repeated 5 times with random payoff matrices generated based on the value of Γ. The probability of convergence is then recorded as $\frac{\text{number of times converged}}{5}$.

8. The values of τ and Γ are then modified and the process is repeated. The heatmap shows the probability of convergence for all values of τ and Γ which are tested.