# Asymptotic Extinction in Large Coordination Games

**Desmond Chan** [1] *, **Bart de Keijzer** [1], **Tobias Galla** [2], **Stefanos Leonardos** [1], **Carmine Ventre** [1]

[1] King's College London
[2] Institute for Cross-Disciplinary Physics and Complex Systems (IFISC, CSIC-UIB)
desmond.chan@kcl.ac.uk , bart.de_keijzer@kcl.ac.uk, tobias.galla@ifisc.uib-csic.es,
stefanos.leonardos@kcl.ac.uk , carmine.ventre@kcl.ac.uk

## Abstract

We study the exploration-exploitation trade-off for large multiplayer coordination games where players strategise via Q-Learning, a common learning framework in multi-agent reinforcement learning. Q-Learning is known to have two shortcomings, namely non-convergence and potential equilibrium selection problems, when there are multiple fixed points, called Quantal Response Equilibria (QRE). Furthermore, whilst QRE have full support for finite games, it is not clear how Q-Learning behaves as the game becomes large. In this paper, we characterise the critical exploration rate that guarantees convergence to a unique fixed point, addressing the two shortcomings above. Using a generating-functional method, we show that this rate increases with the number of players and the alignment of their payoffs. For many-player coordination games with perfectly aligned payoffs, this exploration rate is roughly twice that of p-player zero-sum games. As for large games, we provide a structural result for QRE, which suggests that as the game size increases, Q-Learning converges to a QRE near the boundary of the simplex of the action space, a phenomenon we term asymptotic extinction, where a constant fraction of the actions are played with zero probability at a rate o(1/N) for an N -action game.

## Introduction

Multi-agent systems are an increasingly relevant area in AI research. They typically consist of learning agents trying to coordinate to reach specific outcomes, such scenarios are prevalent in fields ranging from economics (March 1991), robotics and distributed systems (Panait and Luke 2005). A key challenge in these settings is balancing exploration and exploitation in high-dimensional action spaces. Exploration is required for the discovery of optimal strategies; this can come at the expense of short-term rewards. Effectively exploring such complex spaces can be a critical point of failure, preventing convergence to "good" outcomes.

In multi-agent reinforcement learning (MARL), coordination scenarios consisting of interacting agents, can be represented as games. Throughout this work, we will focus on Q-Learning, one of the most widely used methods in MARL, as it provides a framework to analyse the exploitation-exploration trade-off algorithmically. The fixed points of Q-Learning are Quantal Response Equilibria (QRE) (Leonardos, Piliouras, and Spendlove 2021), which always assign positive probability to all actions of finite games and for low exploration rates approximate the Nash Equilibria (NE) for the underlying game.

Coordination games are characterised by players' payoffs being aligned in a manner to incentivise picking mutually beneficial actions. In such settings, agents following the Q-Learning algorithm over a fixed game can exhibit two different dynamical behaviours: (i) Convergence to a unique fixed point (at high exploration rates) – where agents reach the same, joint fixed point regardless of initial conditions; and, (ii) Convergence to multiple equilibria (at low exploration rates) – the final strategy profiles agents converge to is dependent on initial conditions. The effectiveness of Q-Learning is influenced by which of these two outcomes emerges during the learning process. This paper investigates the dynamical behaviour of Q-Learning over large, multi-player coordination games where the payoff matrices are randomly drawn from multivariate Gaussians. In each game, the payoffs matrices are randomly-generated and held fixed. Players are assigned random initial strategies and we study the emerging dynamics.

**Related Work and Our Contribution.** The study of random competitive games was first considered in (Galla and Farmer 2013), which was inspired by replicator models in the context of biological evolution (Opper and Diederich 1992; Galla 2006). Following this line of work, we characterise, through the use of random games, the typical behaviour in complex, coordination games a priori the learning process. Our work complements and extends the results in (Sanders, Farmer, and Galla 2018) to coordination games.

Our theoretical analysis suggest in coordination games with large action sets of size $N$, a constant, non-zero proportion of actions are played with a frequency of $o(1/N)$. We call this effect *asymptotic extinction*. This extinction rate is asymptotic and varies with the model parameters. While simulations cannot fully quantitatively confirm this effect, simulation results are broadly consistent with our theoretical analysis. Taking this effect into account, a minimum exploration rate $T_{\text{crit}}$, which guarantees convergence to a unique fixed point can be found over games with varying number of players and degree of payoff correlation.

*Corresponding author

## Preliminaries

**Multiplayer normal form games.** A $p$ player, $N$ action normal form game, $\mathcal{G}$, is defined by a tuple, $\mathcal{G} = (\mathcal{P}, \mathcal{A}, \Pi)$, where $\mathcal{P} := \{1, 2, \ldots, p\}$ is the *set of players*, and $\mathcal{A} := \{1, \ldots, N\}$ is a set of actions. Each player in $\mathcal{G}$ chooses an action, resulting in an *action profile*, i.e., an element $\mathbf{a} \in \mathcal{A}^p$. Thus, for an action profile $\mathbf{a}$, we write $a_i$ to refer to Player $i$'s chosen action in $\mathbf{a}$. Furthermore, we use $\mathbf{a}_{-i}$ to refer to vector obtained from $\mathbf{a}$ by removing the $i$th coordinate. The notation $(b, \mathbf{a}_{-i})$ then refers to the vector obtained from $\mathbf{a}$ by replacing the value at coordinate $i$ with $b$ (so that $\mathbf{a} = (a_i, \mathbf{a}_{-i})$. For each action profile, every player experiences a certain *payoff* which players want to maximise. Payoffs are specified by the *payoff function* $\Pi : \mathcal{A}^p \to \mathbb{R}^p$, where for $i \in \mathcal{P}$ and $\mathbf{a} \in \mathcal{A}^p$, the payoff for Player $i$ on action profile $\mathbf{a}$ is given by $\Pi(\mathbf{a})_i$.

Players can choose their actions probabilistically. This gives rise to the notion of a *strategy* $\mathbf{x}$, which is a probability distribution over $\mathcal{A}$. Thus, $\mathbf{x}$ is a point on the $(N-1)$-simplex $\Delta_N = \{\mathbf{x} \in \mathbb{R}^n_{\geq 0} \ : \ x_1 + \cdots + x_N = 1\}$. *Interior points* of the simplex correspond to strategies where all actions are played with positive probability ($x_i > 0, \forall i$). Points not in the interior are known as *boundary points*. Similar to the notion of an action profile, a *strategy profile* is a choice of strategy by each of the players, and is hence given by an element $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^p) \in \Delta_N^p$. A strategy profile $\mathbf{x}$, induces a probability distribution over action profiles, and we define the payoff $R(\mathbf{x})^i$ of Player $i$ for $\mathbf{x}$ from $\Pi$, as the expected value of the payoff of the random action profile:

$$R(\mathbf{x})^i = \sum_{\mathbf{a} \in \mathcal{A}^p} \Pi(\mathbf{a})_i \prod_{i \in \mathcal{P}} x^i_{a_i}. \tag{1}$$

Similar to our notation for action profiles, for a strategy profile $\mathbf{x}$ we use $\mathbf{x}^{-i}$ to refer to the vector obtained from $\mathbf{x}$ by removing the strategy of Player $i$ from it. The notation $(y, \mathbf{x}^{-i})$ then refers to the vector obtained from $\mathbf{x}$ by replacing $\mathbf{x}^i$ with $y \in \Delta_N$. Furthermore, we sometimes abuse notation and write $(a, \mathbf{x}^{-i})$, for an action $a \in \mathcal{A}$, to denote $(\mathbf{e}_a, \mathbf{x}^{-i})$, where $\mathbf{e}_a$ denotes the vector with a 1 at coordinate $a$ and 0s at all other coordinates.

**Constructing payoff matrices.** To generate a game, we draw the payoff matrix, $\Pi$, from a multivariate Gaussian with mean 0 which treats all players symmetrically [1].The covariance matrix of the distribution is determined by parameter $\Gamma \in (-1, p-1)$, which captures the *pairwise correlations* between the players' payoffs. Fixing all but two players $i, j \in \mathcal{P}$, we have the following pairwise-correlation structure for each action profile $\mathbf{a}, \mathbf{b} \in \mathcal{A}^p$.

$$\mathbb{E}\left[\Pi(\mathbf{a})_i \cdot \Pi(\mathbf{b})_j\right] = \begin{cases} 1, & \text{if } \mathbf{a} = \mathbf{b}, i = j, \\ \Gamma/(p-1), & \text{if } \mathbf{a} = \mathbf{b}, i \neq j \\ 0, & \text{if } \mathbf{a} \neq \mathbf{b} \end{cases}$$

$\Gamma$ acts as a measure of the level of cooperativeness-competitiveness of a game. For every additional unit of reward that Player $i$ receives by changing their strategy, the sum of all other players' payoffs will change by $\Gamma$ in expectation. $\Gamma = -1$ represents a $p$-player zero-sum game, while $\Gamma = p - 1$ represents an identical payoff game. In general, $\Gamma < 0$, corresponds to *competitive games* where players can only benefit at the expense of others. Conversely, $\Gamma > 0$, corresponds to *coordination games* in which the player's payoffs are positively aligned to a degree given by $\Gamma$.

**Q-Learning.** Given a game and an initial set of strategies, we wish to analyse how players learn and how their strategies evolve over time. Players following a learning algorithm turn games into dynamical systems with strategies evolving in a state space. Our focus is on the Q-Learning model (Watkins and Dayan 1992). [2] Here each player $i \in \mathcal{P}$ keeps track of a Q-value corresponding to each action $a \in \mathcal{A}$, which estimates the *quality* of the given action. At each time step $t$, the Q-value corresponding to action $a$ are updated as follows:

$$Q^i_a(t+1) = \underbrace{(1-\alpha)Q^i_a(t)}_{\text{discounted previous Q-value}} + \underbrace{R(a, \mathbf{x}^{-i}(t))^i}_{\text{current reward}} \tag{2}$$

where $\alpha \in (0, 1)$ denotes the discount parameter. [3] The discount rate is then given by $(1 - \alpha)$ which indicates experience (the previous Q-value) is prioritised against the current reward. With the Q-values, player select mixed strategies according to the softmax distribution parameterised by $\beta > 0$

$$x^i_a(t) = \frac{\exp\left[\beta Q^i_a(t)\right]}{\sum_{b \in \mathcal{A}} \exp\left[\beta Q^i_b(t)\right]} \tag{3}$$

We refer to parameter $T := \alpha/\beta$ as the *exploration rate*. Taking $\alpha, \beta \to 0$, but keeping $T$ constant is equivalent to taking smaller step sizes in each update until we reach the continuous limit. See the Appendix in an arXiv paper under the same name (Chan et al. 2024) for details on how this limit is obtained from the discrete equations (2) (3). Thus, we obtain the continuous Q-Learning equations (Sato and Crutchfield 2003; Tuyls, Hoen, and Vanschoenwinkel 2006):

$$\frac{\dot{x}^i_a(t)}{x^i_a(t)} = R(a, \mathbf{x}^{-i}(t))^i - T \ln x^i_a(t) - \rho^i(t) \tag{4}$$

where $\rho^i := R(\mathbf{x}^i(t), \mathbf{x}^{-i}(t))^i - T\langle \mathbf{x}^i(t), \ln \mathbf{x}^i(t) \rangle$ is a normalisation parameter, which ensures strategies stay within the simplex. Here, $\langle \cdot, \cdot \rangle$ denotes the inner product. The fixed points of Q-Learning dynamics (both discrete and continuous variant) are *Quantal Response Equilibria*.

**Definition 1 (Quantal Response Equilibrium (QRE))** *A strategy profile $\bar{\mathbf{x}} \in \Delta$ is a* Quantal Response Equilibrium *(QRE) if, for all players $i \in \mathcal{P}$ and all actions $a \in \mathcal{A}$*

$$\bar{x}^i_a = \frac{\exp(R(a, \bar{\mathbf{x}}^{-i})^i/T)}{\sum_{j \in N} \exp(R(j, \bar{\mathbf{x}}^{-i})^i/T)}.$$

[1] The choice of Gaussian distribution can be motivated by a maximum entropy and universality argument (Tao and Vu 2011) See the Appendix for more details.

[2] The use of Q-Learning is widespread in multi-agent learning and game theory literature where it appears under various names and variants including Experience Weighted Attraction (EWA) (Camerer and Hua Ho 1999), Boltzmann Q-Learning (Kianercy and Galstyan 2012; Bloembergen et al. 2015) etc. See (Pangallo et al. 2017) for a general overview.

[3] Note that we could include a factor $\alpha$ in front of the reward term, this should extend the possible range in which one can trade off exploration with exploitation.

*where $T \in [0, \infty)$ denotes the* exploration rate*, which is assumed to be equal for all players.*

**QRE Interpretation** QREs are a natural equilibrium solution concept, which takes into account the risk-reward management of the players, and are related to NE. At $T = 0$, only the actions which yield the highest payoff are played. Here the QRE corresponds to the NE. For $T > 0$, players mix actions, with players converging to the uniform distribution as $T \to \infty$. Thus, $T$ acts as a *risk aversion* parameter. Crucially, for any finite game, any initial strategy in the interior of the simplex, continuous Q-Learning (4) will converge to a unique interior fixed point given a sufficiently high $T$ (Hussain, Belardinelli, and Piliouras 2023).

**Rescaling of $T$** As we vary the number of actions, $N$, available to each player, intuitively, we expect the 'typical action' $x_a^i$ to scale at a rate of $1/N$. Hence, the expected payoff across different actions scales at a rate of $\sqrt{1/N^{(p-1)}}$. We dedicate a segment in the Appendix to discuss this rescaling. To facilitate a fair comparison across games of different sizes, we have to take these effects into account and rescale $T$ as follows: $T = \tilde{T}/\sqrt{N^{(p-1)}}$, where $\tilde{T}$ represents the previous unscaled exploration rate. Thus, we will henceforth be working with the scaled exploration rate $T$.

**Overview of Numerical Results** For all values of $\Gamma$, Q-Learning converges to a unique fixed point at sufficiently high exploration rates $T$. Below some critical exploration rate $T_{\text{crit}}$, which increases with $\Gamma$ and $p$, we observe dynamics of varying nature, given in Table 1.

| Condition | Dynamical Behaviour at $T < T_{\text{crit}}$ |
|---|---|
| $\Gamma > 0$ | Convergence to multiple fixed points |
| $\Gamma \approx 0$ | Occasional limit cycles |
| $\Gamma < 0$ | Chaotic behaviour |

Table 1: Dynamical behaviour at $T < T_{\text{crit}}$ for varying $\Gamma$. A brief overview (and supporting figures from simulation) of the possible dynamical behaviours can be found in the Appendix. A similar overview for this model can be found in (Sanders, Farmer, and Galla 2018).

## Analytic Background

We provide an overview of the generating functional method, which was first introduced in (Galla and Farmer 2013) to study the dynamical behaviour of Q-Learning in the $N \to \infty$ limit. Instead of focusing on the outcome of a single initialisation of Q-Learning, we study the evolution of ensembles of possible initialisations. Thus, we will work with the distributions of possible Q-Learning trajectories and how they evolve over time. Borrowing methods from Dynamical mean-field theory (DMFT) [4], we consider the distribution of trajectories in the $N \to \infty$ limit; here the statistics of the Q-Learning trajectories satisfy a stochastic relation, which we refer to as the *effective dynamics*. As $N$ increases, the statistics of the Q-learning trajectories obey the effective dynamics with increasing accuracy. Solving for the fixed points of the effective dynamics and its corresponding stability will allow us to identify the critical exploration rate, $T_{\text{crit}}$, required for convergence to a unique fixed point and the rate of extinction in the unique fixed point regime.

**Effective Dynamics.** Deriving the effective dynamics relation can be broken up into the following two steps: (i) Defining a probability measure over possible trajectories under Q-Learning; (ii) Averaging over all possible payoff matrices, by considering the large action space limit ($N \to \infty$).

The calculations in each of these steps are lengthy and relies on path integral methods from disordered systems theory and a rescaling of variables, we have relegated the details of derivation from (Sanders, Farmer, and Galla 2018) into the Appendix alongside references. The result of this analysis, which holds for any given value of $\Gamma$ and $T$, is the following effective dynamics:

$$\frac{\dot{x}(t)}{x(t)} = \Gamma \int_{t_0}^{t} G\left(t,t'\right) C\left(t,t'\right)^{p-2} x\left(t'\right) \mathrm{d}t' - T \ln x(t) - \rho(t) + \eta(t). \tag{5}$$

The term $\rho(t)$ is a function corresponding to the normalisation term $\rho^i(t)$ of (4), and $\eta(t)$ is a coloured (i.e., time-correlated) Gaussian random variable satisfying: $\langle \eta(t)\eta(t') \rangle_* = C\left(t,t'\right)^{p-1}$ for all $t' < t$, where $\langle \dots \rangle_*$ denotes the expected value over realisations.

The $\eta$ term can be thought of as the randomness at the fixed point phase originating from the initialisation of the payoff matrices. Lastly, $C$ and $G$ are given by:

$$C\left(t,t'\right) = \langle x(t)x(t') \rangle_*, \; G\left(t,t'\right) = \left\langle \frac{\delta x(t)}{\delta \eta\left(t'\right)} \right\rangle_*.$$

Here, $C$ describes time correlations between strategies and $G$ acts a 'response' function that links how strategies are correlated over time and how this varies with $\eta$.

Equation (5) describes the evolution of the marginal probability of playing a given action $x(t)$ as a stochastic process. It is scaled by a factor of $N$ such that each action is played with mean 1, $\langle x(t) \rangle_* = 1$.[5] We note that (5) does not depend on the player nor the index of the actions. This a due to the a-priori symmetry among players in the initial conditions. By solving for the fixed-point of (5), we are able to find the marginal probability distributions of playing each action at the unique fixed point regime of Q-Learning.

**Fixed Points of the Effective Dynamics.** For high exploration rates $T$, Q-Learning converges to a unique fixed point regardless of initial conditions. For large $N$, this should align with the presence of a *stable* fixed point for (5). Thus, to check the dynamical behaviour of Q-Learning, given $\Gamma$ and $T$, we would have to (i) identify the fixed points corresponding to (5) and (ii) analyse the respective stability.

Thus, to find a fixed point of the effective dynamics in a large game, we consider the following. First, a *fixed point of* (5) is defined as a solution where $\dot{x}(t) = 0$ (for all $t$), so that $x(t) = x$ is constant across $t$ for each realisation of

[4] A step-by-step guide of the method on replicator models, we refer the reader to (Galla 2024).

[5] This is discussed in more detail in the Appendix, along with the scaling of the exploration rate $T$.

the random variable $\eta$. In this stationary regime, we note: $C$ becomes a constant; $C(t,t') = \left\langle (x)^2 \right\rangle_* = q$, and $\eta$ turns out to be a static realisation of a Gaussian $\eta \sim \mathcal{N}\left(0,\, q^{p-1}\right)$, while $G$ is now a function of the time difference; $G(t,t') = G(t-t')$. We let $z \sim \mathcal{N}(0,1)$ such that $\eta = q^{(p-1)/2} z$. Thus, we have:

$$0 = x(z)\left[\Gamma q^{(p-2)} x(z)\chi - T\ln x(z) - \rho + q^{(p-1)/2} z\right] \tag{6}$$

where we note that $x$ in (6) is a function of the Gaussian realisation $z$, $\chi = \int_0^\infty G(\tau)d\tau$ is assumed to be finite, and $\rho$ corresponds to $\rho(t)$ in (5), which must be constant in $t$ as well by the requirement that $\dot{x}(t) = 0$. By the definitions of $q, \chi, C, G$ and $\eta$, the following self-consistency relations must hold:

$$\left\langle \frac{\delta x(z)}{\delta z} \right\rangle_* = q^{(p-1)/2}\chi,\ \left\langle x(z)^2 \right\rangle_* = q, \left\langle x(z) \right\rangle_* = 1 \tag{7}$$

where $\langle x(z)\rangle_*$ can be replaced by the integral $\int_{-\infty}^{\infty} x(z)\exp(-z^2/2)/\sqrt{2\pi}dz$. This gives us a 4-equation, 4-unknown problem (with unknowns $q, \chi, \rho, x$ and Equations (7), and (6)), and solving this problem yields us a fixed point corresponding to (5).

## Beyond Competitive Games

Up to this point, we have summarised previous work by (Galla and Farmer 2013; Sanders, Farmer, and Galla 2018). While the effective dynamics (5) and fixed point equations (6) hold for any $\Gamma$, only the competitive case ($\Gamma < 0$) has been studied analytically. Now, we will explore the implications of allowing $\Gamma$ to be positive.

The structure is as follows: We show that theory suggests asymptotic extinction occurs when $\Gamma \geq 0$, and this is consistent with numerical simulations. We solve $x(z)$ for the fixed point equations (6) using an *ansatz* that accounts for this effect, determining the frequency of asymptotic extinction. We will then determine $T_{\text{crit}}$, which we will compare against numerical simulations. Additional details, derivations, and supporting figures are provided in the Appendix.

### Asymptotic Extinction

To satisfy (6), we need either of the following to hold for all values of $z$: $x(z) = 0$ or the bracketed term in (6) $= 0$, whilst satisfying the self-consistency relations (7). We can classify our fixed points into two distinct cases: i) Interior , where: $\left[\Gamma q^{(p-2)} x(z)\chi - T\ln x(z) - \rho + q^{(p-1)/2} z\right] = 0, \forall z \in \mathbb{R}$ ii) Boundary, where $x(z) = 0$: for some $z$.

Points on the boundary correspond to strategies with extinct actions. When $\Gamma > 0$, (6) has no corresponding internal fixed points, only boundary fixed points, even at high exploration rates. How could (6) suggest that Q-Learning would converge to a boundary point?

It is known that strategies on the boundary of the simplex are unstable under Q-Learning in any finitely sized game. However, boundary points, as defined here, are not necessarily unstable since we are working in the large action size limit $N \to \infty$. (Recall $x(z)$ is rescaled by a factor of $N$ in the effective dynamics (5)). Simulating games with fixed parameter values ($\Gamma \geq 0$ and varying $N$, we find that a proportion of actions are played with near 0 probability. When $T < T_{\text{crit}}$, this proportion can be significant. This occurs to a much smaller extent in the unique fixed point regime $T > T_{\text{crit}}$, typically affecting less than $1\%$ of actions. (See Figure 1, where a probability mass of actions are played with near 0 probability as $N$ increases.) This provides the basis for *asymptotic extinction*, where points can be internal for any finite game, but asymptotically approaches the boundary in the large action size $N \to \infty$ limit.

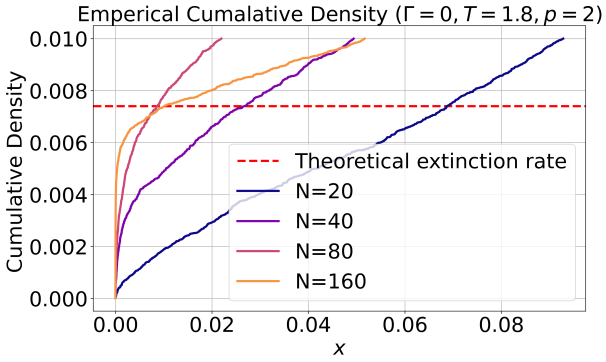


Figure 1: Empirical cumulative density plot representing the marginal likelihood of playing an action at unique fixed point for randomly generated games following the Q-Learning dynamic where $\Gamma = 0, T = 1.8, p = 2$. The plot is zoomed in at the bottom $1\%$ of least played actions and $x$ is rescaled such that $x = 1$ would represent the average likelihood $(1/N)$. As $N$ increases, a probability mass appears to form near 0, representing actions going asymptotically extinct. The red line represents the theoretical estimate of the extinction rate $(0.74\%)$ in the $N \to \infty$ limit. In this limit, the cumulative density plot would begin on the red line.

| Condition | Stable Interior fixed point? |
|---|---|
| $\Gamma > 0$ | no |
| $\Gamma = 0$ | depends on $T$ |
| $\Gamma < 0$ | yes |

Table 2: Existence condition for a stable interior fixed point for finite exploration rates.

The fixed point solution of $x(z)$ allows us to characterise this behaviour in the unique fixed point regime $T > T_{\text{crit}}$. Depending on the sign of $\Gamma$, the fixed point solution given by $x(z)$ varies significantly (See Table 2). We will provide a case-by-case ansatz of $x(z)$ [6], which takes the effect of asymptotic extinction into account for $\Gamma \geq 0$:

**Competitive Games ($\Gamma < 0$).** In this regime, there exists a unique interior fixed point. At this fixed point we have:

$$x(z) = Ke^{bz + ax(z)} \tag{8}$$

where: $a = \Gamma q^{(p-2)}\chi T^{-1} < 0$, $b = q^{(p-1)/2}T^{-1} > 0$ and

[6] A guide to solving these relations numerically is provided in the Appendix.

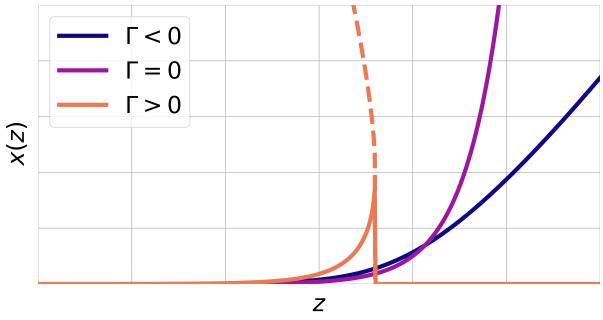


Figure 2: Sketch of $x(z)$ for different values of $\Gamma$. The solution for $\Gamma > 0$ is double-valued below a critical $z$, as seen by the dotted lines. The bottom (solid) branch is of interest here.

$K$ is a normalisation constant, ensuring $\langle x(z) \rangle_* = 1$. These are determined by solving the self-consistency relation (7) and is in agreement with (Sanders, Farmer, and Galla 2018).

**Cooperative Games ($\Gamma > 0$).** For cooperative games $\Gamma > 0$, Equation (6) does not yield an interior fixed point, but rather a boundary point. Fixed points here take the following form:

$$x(z) = \begin{cases} Ke^{bz+ax(z)} & , z < z_{\text{crit}} \\ 0 & , z \geq z_{\text{crit}} \end{cases} \tag{9}$$

where: $a = \Gamma q^{(p-2)} \chi T^{-1} > 0$ and $z_{\text{crit}} = -1/b\,(1 + \ln(aK))$. Values $a, b, K$ are the same as in (8), except $a$ is now positive. We note (9) is double-valued for $x(z)$ below $z_{\text{crit}}$. This is represented by the two branches in Figure 2: the bottom branch (the solid line) and the top branch (the dotted-line). We take the bottom branch as our value for $x(z)$.

**Uncorrelated Games ($\Gamma = 0$).** Uncorrelated games represent a special case, where the existence of an internal fixed point is dependent on $T$. When, $T \geq \sqrt{3e(p-1)/2}$, we have an internal fixed point [7] of the form:

$$x(z) = Ke^{bz} \tag{10}$$

While, when $T < \sqrt{3e(p-1)/2}$, the fixed point is on the boundary, taking the form:

$$x(z) = \begin{cases} Ke^{bz} & , z < z_{\text{crit}} \\ 0 & , z \geq z_{\text{crit}} \end{cases} \tag{11}$$

With the disappearance of $a$, $z_{\text{crit}}$ is to be determined directly from self-consistency (7) as the third unknown.

**How likely is extinction?** For lower exploration rates $T < T_{\text{crit}}$, the solutions of $x(z)$ are unstable and we are unable to characterise likelihood of extinction. Our solution $x(z)$ to the fixed point relations given by (9) and (10) can only predict the distribution of actions when $T > T_{\text{crit}}$ (i.e. when there is a unique fixed point), with $T_{\text{crit}}$ and the corresponding regime will be identified in the next section. For now, we will discuss the extinction likelihood obtained by solving $x(z)$, given by $P(z < z_{\text{crit}})$, where $z \sim \mathcal{N}(0, 1)$.

Figure 3 displays the theoretical extinction rate for games with $p \in \{2, 3, 5\}$ and varying $T$ and $\Gamma$. As $T$ is increased beyond $T_{\text{crit}}$, our fixed point relations suggests the likelihood of a randomly chosen strategy going extinct asymptotically decreases drastically. In the large game limit, $N \to \infty$, for any finite exploration rate $T$, theory suggests that a non-zero proportion of strategies is expected to go asymptotically extinct in coordination games.

Around the stability boundary, extinctions occur to around $1\%$ to $0.01\%$ of actions. Checking selected parameter combinations of $T$ and $\Gamma$ on the stability boundary for games with more players, we find this roughly holds true for higher values of $p$. Away from the boundary, the probability of a randomly selected action going extinct becomes very rare. Verifying the likelihood of asymptotic extinction experimentally in this parameter range with experiments is difficult, as extinctions become extraordinarily rare events.

## Stability Analysis

Having found the fixed points distributions, we have to check their corresponding stability to determine if Q-Learning converges to it. We show the following result:

**Proposition 1** *Q-Learning converges to a unique fixed point when the parameters and corresponding fixed point fulfils the following relation:*

$$\phi \left\langle \left| \frac{T}{x(z)} - \Gamma q^{p-2} \chi \right|^{-2} \right\rangle_* < \left( (p-1) q^{p-2} \right)^{-1} \tag{12}$$

*where $\phi$ is the proportion of non-extinct strategies, given by $P(z < z_{crit})$ where $z \sim \mathcal{N}(0, 1)$.*

When $\Gamma < 0$, $\phi = 1$ (as the fixed points are internal); we have the same relation as (Sanders, Farmer, and Galla 2018). What we have done here is added a $\phi$-term, which takes into effect when $\Gamma \geq 0$. [8]

Thus, (12) extends the analysis to the coordination setting $\Gamma \geq 0$. In the unique fixed point regime, only a small fraction ($< 1\%$) of strategies go extinct, thus the relevant $\phi$s are almost always close to 1. More detailed guidance on obtaining (12) is attached in the Appendix; it is obtained by a somewhat standard procedure [9] used to determine the linearised stability in dynamical systems, as follows: i) linearising the dynamics at the fixed point ii) taking a frequency transform and identifying a stability criterion which guarantees the stability of the whole system under all possible perturbation modes[10].

[7] See the Appendix for the derivation, alongside figures demonstrating the consistency with numerical simulations.

[8] The inclusion of $\phi$ when species go extinct is standard for DMFT analysis of replicator models, see (Opper and Diederich 1992; Galla 2018). The possibility of including a $\phi$-term for 2-player setup has been discussed in (Galla and Farmer 2013), but not explored in further detail due to the apparent contradiction between theory suggesting actions going extinct, but no actions going to 0 in finite, numerical simulations.

[9] See (Drazin and Reid 2004) for a textbook introduction and application of this method to fluid systems.

[10] There is a subtle deviation in the derivation of the stability rela-

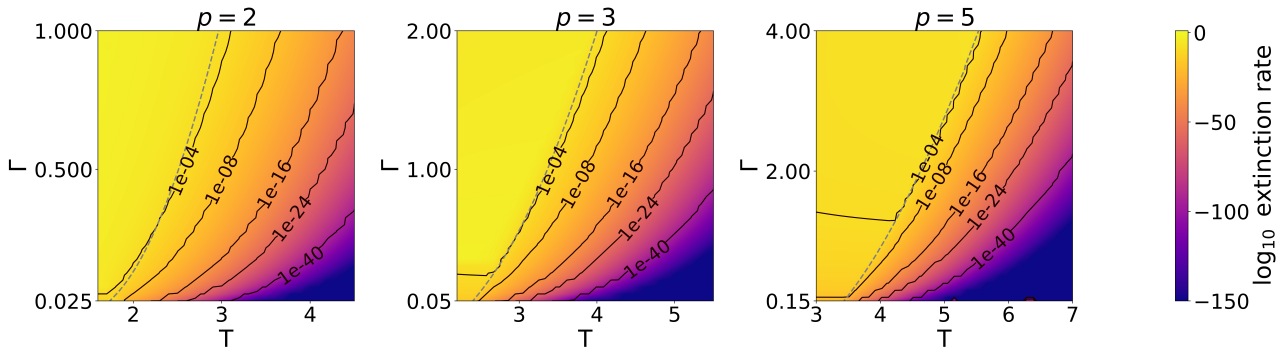


Figure 3: Theoretical asymptotic extinction rate for varying numbers of players $p$ obtained from estimates from the fixed point relations (9). These estimations are only for the unique fixed point regime (right of the yellow dotted line representing the stability boundary, which is solved in the next segment) . There are some numerical instability in the estimations (namely when $\Gamma < 0.1$, thus the axes not starting at 0), but the figure roughly shows the scale of the expected extinction rate for varying $T$ away from the boundary.

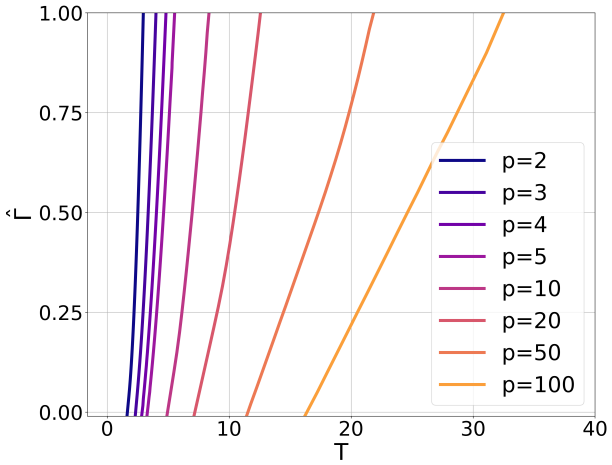


Figure 4: Stability boundary obtained by solving (12) for varying values of $p$, as a function of $T$ for $\hat{\Gamma} > 0$, where $\hat{\Gamma} = \Gamma/(p-1)$. To the right of the boundary, all Q-Learning trajectories converge to a unique fixed point in the large action size limit, $N \to \infty$. When $\hat{\Gamma} < 0$, we recover the results from (Sanders, Farmer, and Galla 2018). Our work extends the stability boundary to cover $\hat{\Gamma} > 0$.

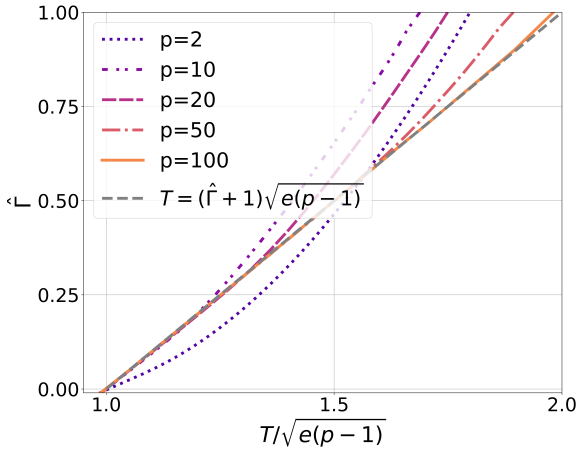


Figure 5: Rescaled stability curves for selected values of $p$. For clarity, the curves are set increasingly solid as $p$ increases. The exploration rate, $T$, is rescaled by a factor of $\sqrt{e(p-1)}$ and the grey-dashed line represents the straight line given by $T_{\text{crit}} = (\hat{\Gamma}+1)\sqrt{e(p-1)}$, which appears to be the limiting behaviour at $p \to \infty$. The increasing agreement with the grey line for large curves with larger values of $p$ suggests this linear relationship is valid, in the large $p$ limit.

**Discussion.** We are interested in the coordination setting $\Gamma > 0$. To generate comparison for games of varying number of players, $p$, we rescale the correlation term as $\hat{\Gamma} = \Gamma/(p-1)$. We will be looking at multi-player games for $\hat{\Gamma} \in (0, 1)$.

Figure 4 displays the stability curves for varying values of $p$ obtained by solving (12). The curve represents the boundary, which separate the multiple fixed point regime from the unique fixed point regime for varying $\hat{\Gamma}$ and $T$. The key result is as $p$ and $\hat{\Gamma}$ increases, so does the critical exploration rate. What does the boundary look like as $p$ gets larger? Upon a rescaling the exploration rate by $1/\sqrt{e(p-1)}$ in the stability plots, Figure 5 suggests, as $p$ increases, a direct linear relationship between the critical exploration rate $T_{\text{crit}}$ and how correlated the game is, $\hat{\Gamma}$, emerges given by [11]:

$$T_{crit} = (\hat{\Gamma} + 1)\sqrt{e(p-1)}. \tag{13}$$

tion from (Sanders, Farmer, and Galla 2018) and other similar work (Galla and Farmer 2013). We dedicate a short segment discussing this in the Appendix.

[11] This linear relationship between critical parameters on the stability boundary is similar to what is observed in DMFT analysis of random replicator predator-prey models in (Galla 2018, 2024), subject to a transformation of $\Gamma$ and $T$.

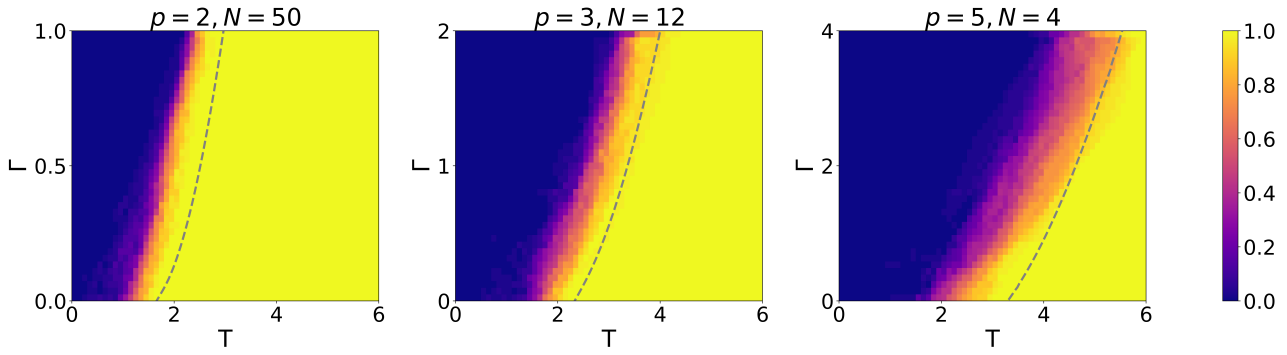


Figure 6: Heat maps showing the proportion of 40 independent payoff matrices for which all trajectories of Q-Learning converges to a unique fixed point, for varying parameter values. Yellow corresponds to all initial conditions converging to a unique fixed point, while indigo indicates there are multiple equilibria. The grey dashed line is computed from the generating functional method. It represents the stability boundary, which separates the two regimes in the large action size limit $N \to \infty$.

In (Sanders, Farmer, and Galla 2018), it is shown that in the large-$p$ limit of uncorrelated games ($\hat{\Gamma} = 0$) and $p$-player zero sum games ($\hat{\Gamma} = 1/(p-1)$) the critical exploration rate is given by $T_{\text{crit}} = \sqrt{e(p-1)}$. This, in combination with (13), suggests the following statement:

**Observation 1** *The critical exploration rate which guarantees the convergence of Q-Learning to a unique fixed point in pure coordination (identical-payoffs) games is* twice *that of p-player, zero-sum games in the large p-limit.*

**Comparison between theory and numerical results.** We compare how our theoretical results fare against numerical experiments of finite-sized, coordination games. We used the default SciPy Runga-Kutta 4(5) solver (Virtanen et al. 2020) with max stepsize set to $0.5$ to be approximate continuous Q-Learning (4) as closely as possible. A point is classified as fixed when the derivative of (4) drops below $|10^{-8}|$.

To determine if a given game converges to a unique fixed point, 100 random initial strategies are drawn and simulated for up to 5000 time units, or until it reaches a fixed point. If all 100 final points, are within a relative distance of $0.01$ of each other, we assume there is a unique fixed point.

For a $p$-player, $N$-action game, we have $p \times N^p$ payoff elements. Selecting $p = 2, 3, 5$ and respectively $N = 50, 12, 4$, yields games with approximately 5000 payoff elements each. For each of the three $(p, N)$-pairs, we perform a parameter search or 'mesh-grid' evaluation for $\Gamma \geq 0$ and $T \in (0, 6)$. For each $\Gamma$ and $T$, 40 independent games are generated, and we record the proportion of games for which Q-Learning converges to a unique fixed point. Figure 6 displays a heat map, displaying the likelihood Q-Learning convergences to a unique fixed point, given the chosen parameters. This is plotted in contrast to the theoretical stability boundary in grey. We refer to (Sanders, Farmer, and Galla 2018) for a similar comparison between the theoretical and numerical results, for $\Gamma < 0$.

We can identify a correspondence between the theoretical curve and the simulation results, which validate the generating functional approach. The simulation plots for $p = 3$ $N = 12$, $p = 5$ $N = 4$ has greater variation between sample rounds than $p = 2$ $N = 50$. We suspect increasing the number of actions $N$ should reduce the variation in the dynamics between large games drawn from the same parameters. Similar to previous work on competitive games (Sanders, Farmer, and Galla 2018), the theoretical critical exploration rate, $T_{\text{crit}}$, appears to be an overestimate, especially near $\Gamma = 0$. We assume (as in previous work) this is a finite-size effect, which disappears as $N$ increases, as the theoretical prediction is in the limit $N \to \infty$.

## Conclusion

Throughout this paper, we have studied the dynamical behaviour of Q-Learning over large, multi-player coordination games, generated from a multivariate Gaussian. This work builds on the model and analysis introduced in (Sanders, Farmer, and Galla 2018) used to study competitive games, to cover the coordination setting.

Q-Learning in large coordination games exhibits a phenomenon that we call asymptotic extinction, where a non-zero fraction of strategies are played with zero probability in the large action size limit $N \to \infty$. Asymptotic extinction is most noticeable at lower exploration rates $T$, but also occurs at high values of $T$. Taking this effect into account, a critical exploration rate $T_{\text{crit}}$ can be identified above which a unique equilibrium exists, and where all trajectories of Q-Learning from all initial points end up in the same equilibrium.

The problem of choosing the 'optimum' exploration rate remains confounding question. Picking the rate $T_{\text{crit}}$, ensures convergence to a unique distribution, avoiding ending up in worst-case scenarios of converging to bad equilibria. $T_{\text{crit}}$ could be taken as a reasonable choice of exploration rate because it is the smallest one where such a unique fixed point is guaranteed, but we emphasize that there are further intriguing questions around the topic of determining the ideal $T$, and there are potentially reasonable alternative choices for $T$. One can consider the problem of finding the exploration rate that maximises any arbitrary objective function (such as maximising total utility, or maximising the minimum utility among the players ). This gives rise to a number of interesting questions to consider for future research.

## Acknowledgments

The authors would like to thank Aamal Hussain and Edward Plumb for the useful discussions throughout the project. This work was partially supported by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1). Partial financial support has been received from the Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (FEDER, UE) under project APASOS (PID2021-122256NB-C21/PID2021-122256NB-C22), and the Maria de Maeztu project CEX2021-001164-M, funded by MCIN/AEI/10.13039/501100011033. Bart de Keijzer was partially supported by EPSRC grant EP/X021696/1.

## References

Bloembergen, D.; Tuyls, K.; Hennes, D.; and Kaisers, M. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53: 659–697.

Camerer, C.; and Hua Ho, T. 1999. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4): 827–874.

Chan, D.; De Keijzer, B.; Galla, T.; Leonardos, S.; and Ventre, C. 2024. Asymptotic Extinction in Large Coordination Games. *arXiv e-prints*, arXiv–2412.

Drazin, P.; and Reid, W. 2004. *Hydrodynamic Stability*. Cambridge Mathematical Library. Cambridge University Press. ISBN 9780521525411.

Galla, T. 2006. Random replicators with asymmetric couplings. *Journal of Physics A: Mathematical and General*, 39(15): 3853–3869.

Galla, T. 2018. Dynamically evolved community size and stability of random Lotka-Volterra ecosystems (a). *Europhysics Letters*, 123(4): 48004.

Galla, T. 2024. Generating-functional analysis of random Lotka-Volterra systems: A step-by-step guide. arXiv:2405.14289.

Galla, T.; and Farmer, J. D. 2013. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4): 1232–1236.

Hussain, A. A.; Belardinelli, F.; and Piliouras, G. 2023. Asymptotic convergence and performance of multi-agent q-learning dynamics. *arXiv preprint arXiv:2301.09619*.

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

Leonardos, S.; Piliouras, G.; and Spendlove, K. 2021. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34: 26318–26331.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization science*, 2(1): 71–87.

Opper, M.; and Diederich, S. 1992. Phase transition and 1/f noise in a game dynamical model. *Physical review letters*, 69(10): 1616.

Panait, L.; and Luke, S. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11: 387–434.

Pangallo, M.; Sanders, J.; Galla, T.; and Farmer, D. 2017. Towards a taxonomy of learning dynamics in 2 x 2 games. *arXiv preprint arXiv:1701.09043*.

Sanders, J. B.; Farmer, J. D.; and Galla, T. 2018. The prevalence of chaotic dynamics in games with many players. *Scientific reports*, 8(1): 4902.

Sato, Y.; and Crutchfield, J. P. 2003. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1): 015206.

Tao, T.; and Vu, V. 2011. Random matrices: universality of local eigenvalue statistics. *Acta Math*, 206: 127–204.

Tuyls, K.; Hoen, P. J. T.; and Vanschoenwinkel, B. 2006. An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games. *Autonomous Agents and Multi-Agent Systems*, 12(1): 115–153.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8: 279–292.