Beyond Strict Competition: Approximate Convergence of Multi-agent Q-Learning Dynamics

Aamal Hussain[1] , **Francesco Belardinelli**[1] , **Georgios Piliouras**[2]
[1]Imperial College London
[2]Singapore University of Technology and Design
{aamal.hussain15, francesco.belardinelli}@imperial.ac.uk, georgios@sutd.edu.sg

Abstract

The behaviour of multi-agent learning in competitive settings is often considered under the restrictive assumption of a zero-sum game. Only under this strict requirement is the behaviour of learning well understood; beyond this, learning dynamics can often display non-convergent behaviours which prevent fixed-point analysis. Nonetheless, many relevant competitive games do not satisfy the zero-sum assumption. Motivated by this, we study a smooth variant of Q-Learning, a popular reinforcement learning dynamics which balances the agents' tendency to maximise their payoffs with their propensity to explore the state space. We examine this dynamic in games which are 'close' to network zero-sum games and find that Q-Learning converges to a neighbourhood around a unique equilibrium. The size of the neighbourhood is determined by the 'distance' to the zero-sum game, as well as the exploration rates of the agents. We complement these results by providing a method whereby, given an arbitrary network game, the 'nearest' network zero-sum game can be found efficiently. As our experiments show, these guarantees are independent of whether the dynamics ultimately reach an equilibrium, or remain non-convergent.

1 Introduction

The convergence of multi-agent learning in competitive settings has long been studied under the context of zero-sum games. The ability to make strong predictions in zero-sum games follows from its enforcement of strict competition between agents. Indeed many positive results have been achieved which show the convergence, in time average, of no regret learning algorithms to a Nash Equilibrium (NE) [Nisan *et al.*, 2007; Hadikhanloo *et al.*, 2022; Bailey and Piliouras, 2019a]. Yet time average convergence does not always imply convergence of the last-iterate. Under this context, zero-sum games, and their network variants, have received much attention, showing cyclic behaviour for some algorithms [Mertikopoulos *et al.*, ; Hofbauer, 1996] and asymptotic convergence for others [Leonardos *et al.*, 2021; Ewerhart and Valkanova, 2020; Hofbauer and Sorin, 2005].

Yet in multi-agent settings the satisfaction of strict competition cannot be taken for granted. The reason for this is simple: not all competitive games are zero-sum. Another contributor is noise; in practice payoffs measured by agents may be subject to perturbations so that the underlying game no longer satisfies the zero-sum condition. It is natural, then, to ask whether the convergence structure holds as we move away from the requirement of strict competition.

Unfortunately, the general answer to this question is *no*. Learning algorithms are known to display complex, even chaotic behaviour when even slightly perturbed away from the safe haven of zero sum games [Sato *et al.*, 2002; Galla and Farmer, 2013; Galla, 2011; Cheung and Tao, 2021]. In fact, this problem becomes even more prevalent as the number of players is increased [Sanders *et al.*, 2018]. The introduction of chaos makes the exact prediction of long-term behaviours impossible in a wide class of games and we are led to a fundamental dichotomy between the need, and ability to understand multi-agent learning in competitive games.

Main Contribution. To make progress in understanding general competitive games, we consider the natural starting point of *near network zero-sum games*. The concept of 'close' games has been introduced in the context of potential games [Candogan *et al.*, 2013], which model strictly cooperative settings. Following its introduction, a number of results on the approximate convergence of learning algorithms have been determined in near-potential games [Anagnostides *et al.*, 2022; Cheng and Ji, 2022] Motivated by the success of the cooperative setting, we re-purpose the distance notion for network zero-sum games (NZSG) which form the natural extension of the zero-sum game to multi agent settings [Cai *et al.*, 2016].

In this setting, we study the *(smooth) Q-Learning dynamics* which models the popular Q-Learning algorithm with Boltzmann exploration [Sutton and Barto, 2018; Schwartz, 2014]. This learning model captures the behaviour of agents who attempt to maximise their payoffs whilst balancing a tendency to explore the space of their possible strategies.

Our first contribution is to show that, in near network zero-sum games, Q-Learning converges to a neighbourhood around the unique equilibrium of the underlying NZSG. The size of this set goes to zero as the distance from the NZSG goes to zero and/or as the exploration rate of each agent increases. Given, then, the distance from the NZSG this size of

the neighbourhood can be adjusted by manipulating the exploration rates of the agents. To assist in this process, we also provide upper bounds on the distance between network games based on the differences in payoff matrices and the network structure. Finally, in a similar light to [Candogan *et al.*, 2013; Cheng and Ji, 2022] which consider potential games, we present a quadratic optimisation formulation for determining the closest NZSG to a given network game. Taken together, these results give a picture of the approximate behaviour of Q-Learning in competitive games which do not exactly satisfy the zero-sum condition.

Related Work. Studies on learning in competitive games often occur within the context of zero-sum games [Aumann, 1989] or its network variants [Cai *et al.*, 2016]. Indeed, due to the desirable structure of these games and the increasing interest of competitive systems [Abernethy *et al.*, 2021], many positive results have been obtained concerning various learning dynamics, including Follow the Regularised Leader [Bailey and Piliouras, 2019b; Anagnostides *et al.*, 2022], Fictitious Play [Ewerhart and Valkanova, 2020], and Q-Learning [Leonardos *et al.*, 2021].

By contrast, non-convergent behaviour, including cycles and chaos, appears to be increasingly prevalent as the NZSG condition is lifted [Galla, 2011; Sato *et al.*, 2002; Sato and Crutchfield, 2003; Mukhopadhyay and Chakraborty, 2020] and as the number of agents increases [Sanders *et al.*, 2018]. This presents a strong barrier when attempting to engineer competitive multi-agent systems, where the network zero-sum assumption need not hold [Ewerhart and Valkanova, 2020; Roberson, 2006]. Outside of this class, results on convergence often make restrictive assumptions, such as the existence of a potential function [Leonardos and Piliouras, 2022; Monderer and Shapley, 1996; Harris, 1998] which enforces strict cooperation amongst agents, or that the game has only two players and two actions [Kianercy and Galstyan, 2012; Metrick and Polak, 1994]. Of course, these do not cover the vast majority of games encountered in practice. In fact, the strongest result regarding learning outside of NZSG is a negative one: consider [Vlatakis-Gkaragkounis *et al.*, 2020] which shows that the popular Follow the Regularised Leader dynamic cannot converge to a fully mixed Nash Equilibrium, regardless of the game structure. With all these taken together, it becomes clear that a complete picture of learning in games cannot be found by considering only convergence to a fixed point, but must include the eventuality of non-convergence.

To make progress on this, we apply the concept of 'nearness' in games. This was first introduced in the context of potential games [Candogan *et al.*, 2013; Cheng and Ji, 2022] to extend the analysis of cooperative games to those which do not satisfy the potential assumption. With this, various learning algorithms including fictitious play [Candogan *et al.*, 2013; Aydın *et al.*, 2022] and Follow the Regularised Leader [Anagnostides *et al.*, 2022], can be understood in terms of *approximate convergence*, i.e., convergence to a neighbourhood of an equilibrium. On the other hand, whilst [Cheung and Tao, 2021] shows that games which deviate from the network zero-sum setting can display chaos, little is known about how deviations from the strictly competitive setting affect the approximate convergence of learning. To our knowledge, the present work is the first to study, both theoretically and experimentally, near network zero-sum games with an aim to understand approximate convergence, even in the face of chaos.

2 Preliminaries

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

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

Network Zero-Sum Games. A *polymatrix* or *network* game also contains a graph $(\mathcal{N}, \mathcal{E})$ in which $\mathcal{N}$ still denotes the set of agents and $\mathcal{E}$ consists of pairs $(k, l) \in \mathcal{N}$ of agents, who are meant to be connected [Cai *et al.*, 2016]. Each edge has associated a pair (A^{kl}, A^{lk}) of matrices, which define the payoff to k against l and vice versa. The payoffs are then given by

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

We represent a network game as a tuple $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$. Γ is a *network zero-sum game* (NZSG) if, for all $\mathbf{x} \in \Delta$,

$$\sum_k u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = 0$$

A seminal result in the study of NZSG is that of [Cai *et al.*, 2016] which shows that any NZSG is payoff equivalent to a pairwise constant sum game, where all the constants add to zero. More formally, this is stated in the following proposition

Proposition 1 ([Cai *et al.*, 2016], [Leonardos *et al.*, 2021])**.** Let $Z = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ be a NZSG. For all $(k, l) \in \mathcal{E}$ there exist $(\hat{A}^{kl}, \hat{A}^{lk})$ and a constant $c_{kl} \in \mathbb{R}$ such that

$$[\hat{A}^{kl}]_{ij} + [\hat{A}^{lk}]_{ji} = c_{kl}, \ \forall i \in S_k, j \in S_l,$$

with

$$\sum_{(k,l)\in\mathcal{E}} c_{kl} = 0,$$

and payoffs to agent k in Z is equivalent to their payoffs in $\hat{Z} = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (\hat{A}^{kl}, \hat{A}^{lk})_{(k,l)\in\mathcal{E}})$. In particular, for all $k \in \mathcal{N}$ and all $\mathbf{x}_k \in \Delta_k$

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

Maximum Pairwise Difference. To define 'nearness' in the context of games, we require a notion of distance on the space of all games. We apply the widely used metric defined in [Candogan *et al.*, 2013], known as *Maximum Pairwise Difference*. Formally, let $\Gamma_1 = (\mathcal{N}, (S_k, A_k)_{k\in\mathcal{N}})$ and $\Gamma_2 = (\mathcal{N}, (S_k, B_k)_{k\in\mathcal{N}})$ be two games which share the same set of agents $\mathcal{N}$ and actionsets $(S_k)_{k\in\mathcal{N}}$ but differ in payoff functions. Then, the Maximum Pairwise Difference between Γ_1 and Γ_2 is

$$d(\Gamma_1, \Gamma_2) = \max |A_k(\mathbf{y}_k, \mathbf{x}_{-k}) - A_k(\mathbf{x}_k, \mathbf{x}_{-k}) - (B_k(\mathbf{y}_k, \mathbf{x}_{-k}) - B_k(\mathbf{x}_k, \mathbf{x}_{-k}))| \quad \text{(MPD)}$$

where the maximum is taken over all agents k, all $\mathbf{x}_{-k} \in \Delta_{-k}$ and all $\mathbf{x}_k, \mathbf{y}_k \in \Delta_k$. In words, (MPD) captures the similarity between two games in terms of the capacity for any agent to improve their payoff by deviating from $\mathbf{x}_k$ to $\mathbf{y}_k$ whilst their opponents maintain their strategy x_{-k}.

Q-Learning Dynamics. Q-Learning [Sutton and Barto, 2018; Schwartz, 2014] is the prototypical model for determining optimal policies in the face of uncertainty. In this model, each agent $k \in \mathcal{N}$ maintains a history of the past performance of each of their actions. This history is updated via the Q-update

$$Q_{ki}(\tau + 1) = (1 - \alpha_k) Q_{ki}(\tau) + \alpha_k r_{ki}(\mathbf{x}_{-k}(\tau))$$

where τ denotes the current time step. $Q_{ki}(\tau)$ denotes the *Q-value* maintained by agent k about the performance of action $i \in S_k$. In effect, Q_{ki} gives a discounted history of the rewards received when i is played, with $1 - \alpha_k$ as the discount factor.

Given these Q-values, each agent updates their mixed strategies according to the Boltzmann distribution, given by

$$x_{ki}(\tau) = \frac{\exp(Q_{ki}(\tau)/T_k)}{\sum_j \exp(Q_{kj}(\tau)/T_k)}$$

in which $T_k \in [0, \infty)$ is the *exploration rate* of agent k: low values of T_k allow the agent to play the action(s) with the highest Q-value with a large probability, thereby exploiting their high performance. By contrast, higher values of T_k enforce that agents play each of their strategies with or the same probability, regardless of their Q-value.

It was shown in [Tuyls *et al.*, 2006] that a continuous time approximation of the Q-learning algorithm can be written as

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

which we call the *Q-learning dynamics*. The fixed points of this dynamic coincide with the *Quantal Response Equilibria* (QRE) of the game:

Definition 1 (QRE)**.** A joint mixed strategy $\mathbf{p} \in \Delta$ is a *Quantal Response Equilibrium* of the game $\Gamma = (\mathcal{N}, (S_k, u_k)_{k\in\mathcal{N}})$ if, for all agents $k \in \mathcal{N}$, $i \in S_k$,

$$p_{ki} = \frac{\exp(r_{ki}(\mathbf{p}_{-k})/T_k)}{\sum_{j\in S_k} \exp(r_{kj}(\mathbf{p}_{-k})/T_k)} \quad (1)$$

The QRE is a well-studied equilibrium concept for games of *bounded rationality* [McKelvey and Palfrey, 1995]. This is seen in the fact that, in the limit $T_k \to 0$ for all k, (1) corresponds exactly to the Nash Equilibrium; whereas in the limit $T_k \to \infty$ for all k, the QRE is the uniform distribution, i.e., each agent plays each action with the same probability, regardless of its past performance.

Game Perturbations. In [Leonardos and Piliouras, 2022] it is shown that, for any $(T_k)_{k\in\mathcal{N}}$, the Q-learning dynamics in a game Γ is equivalent to the well-studied *replicator dynamics* (RD) in a perturbed game Γ^H. More formally, the authors show the following.

Lemma 1 ([Leonardos and Piliouras, 2022])**.** Consider a game $\Gamma = (\mathcal{N}, (S_k, u_k)_{k\in\mathcal{N}})$ and, for each agent k let $T_k > 0$. Then (QLD) can be written as

$$\frac{\dot{x}_{ki}}{x_{ki}} = r^H_{ki}(\mathbf{x}) - \langle \mathbf{x}_k, r^H_k(\mathbf{x}) \rangle,$$

where $r^H_{ki} = r_{ki}(\mathbf{x}_{-k}) - T_k(\ln x_{ki} + 1)$. In particular, (QLD) recovers the replicator dynamics in the pertubed game $\Gamma^H = (\mathcal{N}, (S_k, u^H_k)_{k\in\mathcal{N}})$ where

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

The perturbed game Γ^H has the same players and action sets as Γ but has modified utilities. The same perturbation maps the QRE of the game Γ to Nash Equilibria of Γ^H [Melo, 2021; Gemp *et al.*, 2022]

3 Near Network Zero-Sum Games

Our main results concern the competitive setting. We first show that, in near-NZSG (QLD) converges to a set around the QRE of the NZSG. This determines approximate convergence behaviour when the game is perturbed away from the NZSG assumption. We follow this with a scheme to determine, for any network game (not necessarily zero sum), the nearest NZSG. Using these results together provides a method to determine approximate convergence behaviour for arbitrary competitive network games.

3.1 Approximate Convergence

To define the convergence of the Q-Learning dynamic we need a measure of distance. To this end, we use the *Kullback-Leibler* (KL) divergence.

Definition 2. The Kullback-Leibler Divergence between a set of joint strategies $\mathbf{x}, \mathbf{y} \in \Delta$ is given by

$$D_{KL}(\mathbf{y}||\mathbf{x}) = \sum_k D_{KL}(\mathbf{y}_k||\mathbf{x}_k) = \sum_k \sum_i y_{ki} \ln \frac{y_{ki}}{x_{ki}} \quad (2)$$

The key point which we will use in our main theorem is that $D_{KL}(\mathbf{y}||\mathbf{x})$ is zero if and only if $\mathbf{x} = \mathbf{y}$ and is positive everywhere else.

Theorem 1. Let $Z = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ be a network zero-sum game which, for some $T_1, \dots, T_N > 0$, has unique QRE $\mathbf{p} \in \Delta$. Let $G = (\mathcal{N}, (S_k, u_k)_{k\in\mathcal{N}})$ be a game such that $d(Z, G) < \delta$ for some $\delta > 0$. Then, for any trajectory of mixed strategies $\mathbf{x}(t)$ generated by running (QLD) on G,

$$\lim_{t\to\infty} D_{KL}(\mathbf{p}||\mathbf{x}(t)) \leq \frac{N\delta}{T_{\min}}$$

where $T_{\min} = \min_k T_k$.

Theorem 1 provides a method whereby the behaviour of Q-Learning dynamics can be understood even if the game is slightly perturbed away from a NZSG. It is important to note that the approximate behaviour is also governed by choice of exploration rate; in particular, the region to which (QLD) converges decreases in size as $T_{\min}$ increases. This can be explained as follows: as the exploration rate increases, each agent places less importance on the rewards that each action produces when updating their mixed distribution. Therefore, perturbations away from the NZSG condition are not felt as strongly as they would be if the exploration rate were low.

We now provide the idea required to prove Theorem 1, leaving the details to the Supplementary Material. To do so, we adapt the proof technique of [Leonardos *et al.*, 2021], in which it was shown that Q-Learning converges to a unique QRE in any NZSG. We extend this to consider games which do not fall into this rather restrictive class through the following Lemma.

Lemma 2. Let Z and G be games in the setting of Theorem 1. Then, if agents playing in game G update their mixed strategies according to (QLD), $D_{KL}(\mathbf{p}||\mathbf{x}(t))$ is strictly decreasing whenever $\mathbf{x}(t)$ satisfies

$$\frac{N\delta}{T_{\min}} < D_{KL}(\mathbf{p}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\mathbf{p}) \tag{3}$$

Proof Sketch of Theorem 1. From Lemma 2 it follows that when (3) is satisfied, the distance (as measured by KL-Divergence) is decreasing. Therefore, the trajectory converges to the region where $D_{KL}(\mathbf{p}||\mathbf{x}(t)) + D_{KL}(\mathbf{x}(t)||\mathbf{p}) \leq \frac{N\delta}{T_{\min}}$. Let us denote this region as S and let $D_S := \sup_{\mathbf{x}\in S} D_{KL}(\mathbf{p}||\mathbf{x})$. Then, if $\mathbf{x}(t)$ leaves S, by Lemma 2, $D_{KL}(\mathbf{p}||\mathbf{x}(t))$ cannot increase past D_S. It follows, then, that $\limsup_{t\to\infty} D_{KL}(\mathbf{p}||\mathbf{x}(t)) \leq D_S$. Finally, we note that $D_S = \sup_{\mathbf{x}\in S} D_{KL}(\mathbf{p}||\mathbf{x}) \leq \sup_{\mathbf{x}\in S} D_{KL}(\mathbf{p}||\mathbf{x}) + D_{KL}(\mathbf{x}||\mathbf{p}) \leq \frac{N\delta}{T_{\min}}$. □

Remark. It is important to note that Theorem 1 makes no statement on whether Q-Learning in a near NZSG will itself converge to a QRE. In fact such counter-examples are demonstrated in Figure 5, and complex behaviour is known to be prevalent in multi-agent learning (e.g. [Sanders *et al.*, 2018; Cheung and Tao, 2021]). Nonetheless, Theorem 1 provides a complete picture on the *approximate* last iterate behaviour of Q-Learning. It does this by determining a region to which Q-Learning dynamics must remain trapped, even if it does not ultimately reach a QRE within this region. This region is defined with respect to the QRE of an NZSG, which is unique and can be found by running Q-Learning.

3.2 Finding the Closest NZSG

In order use Theorem 1 to determine the approximate behaviour of an arbitrary competitive (but not zero sum) game, it is first required that we find the nearest network zero-sum game. In this section we show that this process can be solved efficiently. In particular, given any network game $\Gamma = (\mathcal{N}, \mathcal{E}, (S_k)_{k\in\mathcal{N}}, (A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}})$ which is not necessarily zero sum, the problem of finding the 'nearest' NZSG can be written as a quadratic minimisation problem with linear constraints. In doing so, the approximate behaviour of Q-Learning in the original game can be determined.

This formulation manipulates Proposition 1: that any NZSG is payoff equivalent to a pairwise constant-sum game, where the constants add to zero. As such, given the network game Γ, we can write the problem of finding the 'nearest' NZSG as finding the nearest pairwise constant-sum game. This is formulated as

$$\begin{cases} \min\limits_{(\hat{A}^{kl}, \hat{A}^{lk}, c_{kl})} & \sum_{(k,l)\in\mathcal{E}} ||\hat{A}^{kl} - A^{kl}||_2^2 + ||\hat{A}^{lk} - A^{lk}||_2^2 \\ \text{s.t.} & [A^{kl}]_{ij} + [A^{lk}]_{ji} = c_{kl}, \\ & \sum_{(k,l)\in\mathcal{E}} c_{kl} = 0 \end{cases} \tag{P1}$$

where A^{kl}, A^{lk} are the payoff matrices which define Γ. As the objective function in (P1) is quadratic, and the constraints are linear, (P1) is a quadratic optimisation problem which can be solved efficiently.

To connect the minimisation of the $2-$norm to (MPD), we have the following results.

Proposition 2. Suppose $\Gamma_1 = (\mathcal{N}, (S_k, A_k)_{k\in\mathcal{N}})$, $\Gamma_2 = (\mathcal{N}, (S_k, B_k)_{k\in\mathcal{N}})$ are games which have rewards $a_{ki}(\mathbf{x}_{-k}) = \partial A_{ki}(\mathbf{x})/\partial x_{ki}$ and $b_{ki}(\mathbf{x}_{-k}) = \partial B_{ki}(\mathbf{x})/\partial x_{ki}$ respectively. Suppose also that, for all $k \in \mathcal{N}$, $i \in S_k$ and $\mathbf{x}_{-k} \in \Delta_{-k}$,

$$|a_{ki}(\mathbf{x}_{-k}) - b_{ki}(\mathbf{x}_{-k})| \leq \frac{\delta}{2n_k}$$

where $\delta > 0$. Then $d(\Gamma_1, \Gamma_2) \leq \delta$

From Proposition 2 we immediately obtain the following corollary for the particular case of network games.

Corollary 1. Suppose that, in the setting of Proposition 2, Γ_1 and Γ_2 are network games whose rewards are defined through the payoff matrices $(A^{kl}, A^{lk})_{(k,l)\in\mathcal{E}}$, $(B^{kl}, B^{lk})_{(k,l)\in\mathcal{E}}$ respectively. Suppose also that, for all $(k,l) \in \mathcal{E}$, $i \in S_k$ and $j \in S_l$,

$$\left|(A^{kl})_{ij} - (B^{kl})_{ij}\right| \leq \frac{\delta}{2n_k \sum_{(k,l)\in\mathcal{E}} n_l}$$

where $\delta > 0$. Then $d(\Gamma_1, \Gamma_2) \leq \delta$

Corollary 2. Suppose that, in the setting of Proposition 1, Γ_1, Γ_2 are such that for all $(k,l) \in \mathcal{E}$

$$||A^{kl} - B^{kl}||_2 \leq \frac{\delta}{2n_k \sum_{(k,l)\in\mathcal{E}} n_l}$$

where the matrix norm for a matrix $A \in M_{m\times n(\mathbb{R})}$ is given by $||A||_2 = \sup_{\mathbf{x}\in\mathbb{R}^n : ||x||_2\neq 0}(||A\mathbf{x}||_2/||\mathbf{x}||_2)$. Then $d(\Gamma_1, \Gamma_2) \leq \delta$.

Using the process outlined in this section, it is possible to determine approximate convergence in competitive, but not zero sum, network games. Its advantage lies in the fact that the QRE of NZSGs are unique for any $T_k > 0$ and it is known that Q-Learning, for any initial condition must converge to this QRE [Leonardos *et al.*, 2021]. Therefore, the aforementioned process provides a method to determine approximate convergence of Q-Learning in Γ for any initial condition.

4 Experiments on Near NZSG

In our experiments we examine the implications of Theorem 1. In particular we confirm that Q-Learning in near NZSG asymptotically remain close to the QRE of the NZSG. We also examine the implication of this finding for the introduction of noise in the payoffs.

Visualising Theorem 1. In Figure 1 we visualise the region to which Q-Learning converges as predicted by Theorem 1. In particular, we generate a two-action, zero-sum network game for a given number of agents. We then plot the KL-divergence from the QRE $\mathbf{p}$ for a given exploration rate, using the dimensionality reduction technique of [Li *et al.*, 2018], which was adapted for the KL-Divergence by [Leonardos *et al.*, 2021]. The procedure is outlined in the Supplementary Material. We then plot, on the $x - y$ plane, the contour $D_{KL}(\mathbf{p}||\mathbf{x}) = N\delta/T_{\min}$ for some choice of $\delta, T_{\min}$. It is clear that this forms a neighbourhood around the QRE of the NZSG; the implication of Theorem 1 is that, in games which are at most δ away from the NZSG, Q-Learning will asymptotically remain trapped in this neighbourhood.

Three Player Chain. We examine a 'chain' network with three agents where each agent has two actions. We generate a zero-sum game and run Q-Learning on this game to find its QRE [Leonardos *et al.*, 2021]. For the sake of simplicity we assume that all agents have the same exploration rate T_k so that we replace the notation T_k or $T_{\min}$ with simply T. Then, we perturb the payoff matrices to generate five near zero-sum games. We can use Corollary 1 to determine an upper bound on the distance between these games from the NZSG in terms of (MPD).

The results from this experiment are shown in Figure 2. The figures plot the probabiliy by which each player plays their first action. In all cases, the near-NZSG converge to equilibria (depicted with black markers) who are close to the QRE of the NZSG (red marker). The distance of the QRE of the pertubed games from the original increases as δ is increased from 0.75 to 2.

When examining the effect of noise, we take the same network game setup and periodically (every 50 iterations) add noise to the payoff matrices to perturb the game away from the zero sum. By ensuring that the perturbations satisfy Corollary 1 for some δ, we can determine an upper bound on (MPD). The results are shown in Figure 3. The power of Theorem 1 is apparent in this setting since, in this case, Q-Learning will not converge to an equilibrium. Despite this, since the perturbations are upper bounded by δ, Theorem 1 enforces that the trajectories remain within the neighbourhood of the QRE of the original game. This ensures the robustness of Q-Learning under the presence of noise. Note that, whilst in the experiments we use additive noise, Theorem 1 makes no such assumption. The only requirement is that the perturbations are bounded. Of course, the larger this bound is, the larger the neighbourhood, as evidenced by the increase in spread of the Q-Learning trajectories as δ is increased.

Ten Player Network. Finally, we extend our analysis to a 10-agent network where agents have two actions. In this case, the Q-Learning dynamics evolve in $\mathbb{R}^{20}$ and so it is not possible to visualise the trajectories. Rather, we generate a NZSG and 100 near-NZSG which are generated in the same manner as the three-player chain network. Figure 4 shows a summary of the last iterates of Q-Learning in 100 randomly generated near zero-sum games after 1×10^6 iterations. The behaviour agrees with the results in the lower dimensional case. In particular, it is clear that the last iterations of Q-Learning for all nearby games is within a bounded region around the QRE of the NZSG.

Conflict Network. We now examine competitive games that do not satisfy the network zero-sum assumption. To do this we consider *conflict networks* as considered in [Ewerhart and Valkanova, 2020], which cover a wide array of competitive games including the widely studied Colonel Blotto game [Roberson, 2006]. The details of conflict networks can be found in the Supplementary Material.

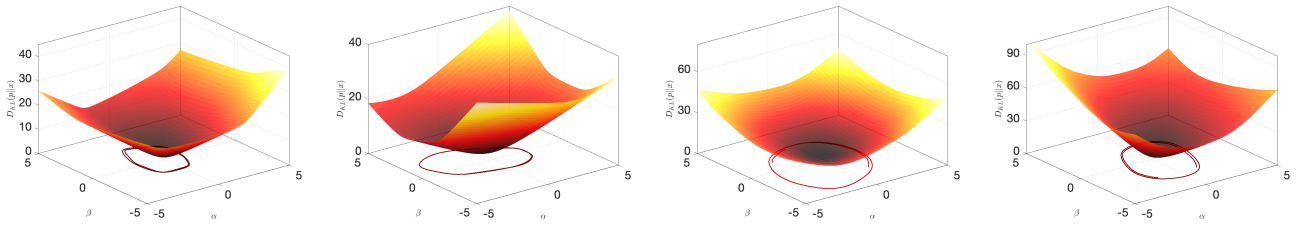
In our experiments, we generate a conflict network game with three agents, which we call Γ_C. The network is fully connected and, for each agent k,

$$A^{k,k+1} = \begin{pmatrix} 2.4 & 6.6 \\ 4.5 & 3.1 \end{pmatrix}, \quad A^{k,k-1} = \begin{pmatrix} 2.8 & 1.0 \\ 4.2 & 7.2 \end{pmatrix},$$

and the sums $k+1, k-1$ are taken $\mod N$. As this game does not satisfy the network zero-sum assumption, it is not necessary that (QLD) converges to a QRE, as shown in our experiments in Figure 5. By applying the procedure in Section 3.2, we find the nearest network zero-sum game which we call Γ_Z. Next, by using Corollary 2, it is possible to show that $d(\Gamma_C, \Gamma_Z) \leq 7.2$. With this and the fact that Q-Learning converges in Γ_Z [Leonardos *et al.*, 2021], it is possible to use Theorem 1 to determine the approximate convergence of (QLD) in Γ_C. For low values of T, the region to which Q-Learning converges is large, and takes up the entire simplex. However, this region becomes smaller as T is increased, so that Q-learning converges to a smaller neighbourhood close to the QRE of Γ_Z.

5 Conclusions

In this paper we begin developing an understanding of the smooth Q-Learning dynamics beyond strictly competitive many player games. We show that in games which are sufficiently close to satisfying the network zero-sum assumption, Q-Learning converges to within a region of a unique Quantal Response Equilibrium (QRE). The size of this region can be adjusted by controlling either the distance from the strictly competitive setting, or the exploration rates of the agents. Whilst the latter amounts to parameter tuning, we consider the former by determining a method to find, for a given network game, the nearest network zero-sum game (NZSG). In

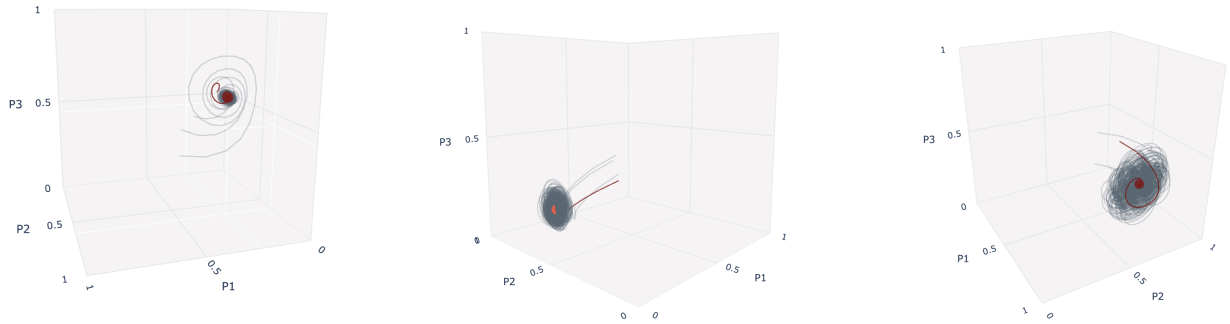


N = 5 N = 7 N = 13 N = 17

Figure 1: Visualisation of the KL-divergence between the unique QRE and a mixed strategy in an NZSG, alongside a depiction of the region to which Q-Learning converges in nearby games. The minimum of the KL-divergence occurs at zero and the region is a neighbourhood around the minimiser (i.e., the QRE of the NZSG). In all cases, we choose $\delta = 1$ and $T = 0.75$, whilst we vary the number of players N.

$\delta \leq 0.75$ $\delta \leq 2$

Figure 2: Trajectories of Q-Learning in near NZSG. In each plot, the red line depicts Q-Learning in an NZSG and the black depicts Q-Learning in a nearby game which is not zero sum. Q-Learning converges to an equilibrium in the near-NZSG (black marker), where the equilibrium is 'close' to the QRE of the NZSG (red marker). In all cases $T = 0.75$.

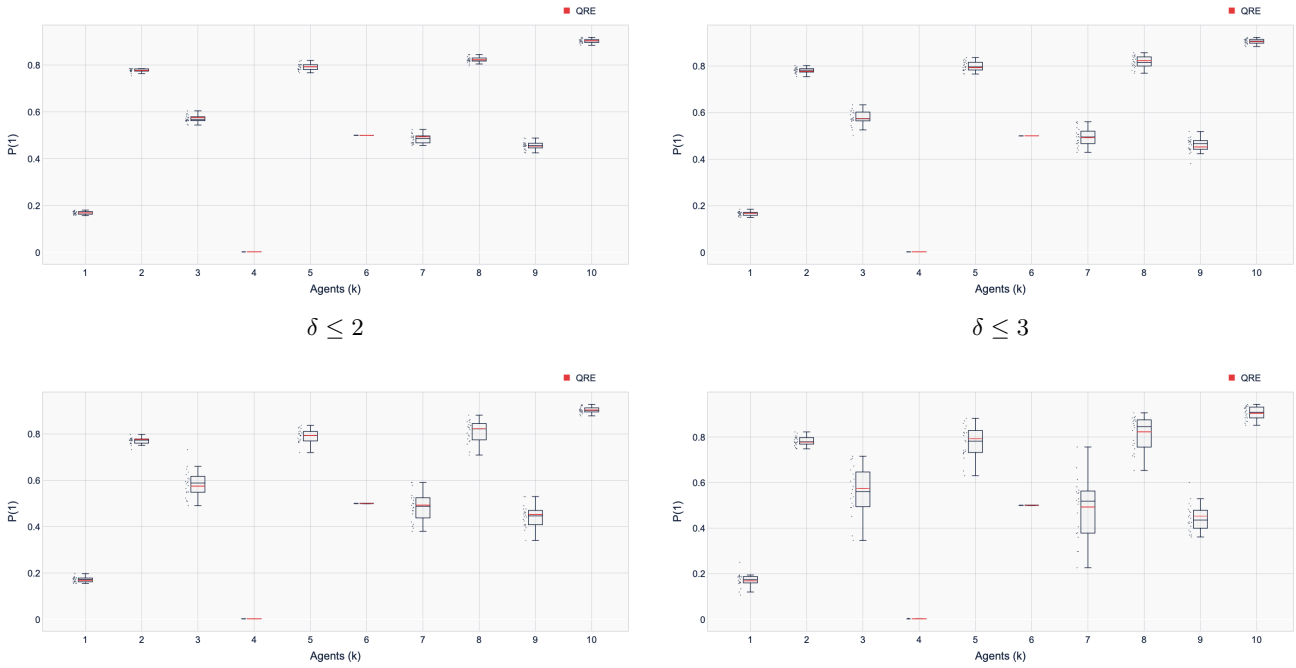


$\delta \leq 0.75$ $\delta \leq 2$ $\delta \leq 3$

Figure 3: Trajectories of Q-Learning on an NZSG in the presence of additive noise. The noise is such that the perturbed game is always close to the NZSG. In this case, Q-Learning does not reach an fixed point, but will still remain asymptotically within a region surrounding the QRE of the NZSG (red marker). In all cases $T = 0.75$.

such a manner, the approximate behaviour of Q-Learning can be understood in arbitrary competitive games. In our experiments we demonstrate the utility of our results in practice, in particular showing that, even in the presence of noise, the asymptotic behaviour of Q-Learning can be understood in terms of distance from the QRE of an underlying NZSG.

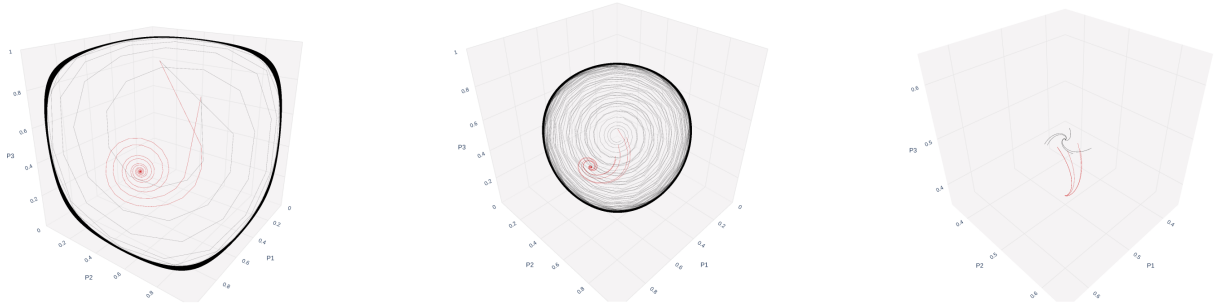
Our results also present an avenue for extending beyond strictly cooperative settings. In particular, the approximate behaviour of Q-Learning in near-potential games can be examined, thus beginning to bridge the gap between strictly competitive and strictly cooperative games. Another interesting direction would be to extend towards *weighted* NZSGs, which comprise a larger set of games than the *exact* NZSG setting considered in this work. Finally, our method for find-



$\delta \leq 2$ $\delta \leq 3$

$\delta \leq 5$ $\delta \leq 9$

Figure 4: Summary Statistics of Last-Iterates of Q-Learning in 100 near-NZSGs. The y-axis depicts the probability which each agent assigns to their first action. The red line depicts the QRE of the NZSG whilst the box depicts the spread of last iterates in the near-NZSG after 1×10^6 iterations. In all cases $T = 0.75$.



$T = 0.35$ $T = 0.5$ $T = 2.5$

Figure 5: Dynamics of Q-Learning in a Conflict Network. Axes indicate the probability with which each agent plays their first action. Black trajectories denote the dynamics in the conflict network. Red trajectories denote the dynamics in the nearest network zero sum game. Q-Learning dynamics in the NZSG converge to a QRE. In the conflict network, they converge to a neighbourhood of the QRE, whose size decreases with increasing T

ing the nearest NZSG requires the original game itself to be a bidirectional network game. Lifting this assumption would allow for the approximate behaviour of a wider class of multi-agent settings (e.g. leader-follower) games to be understood.

Ethical Statement

There are no ethical issues.

Acknowledgments

Aamal Hussain and Francesco Belardinelli are partly funded by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (grant number EP/S023356/1). This research/project is supported in part by the National Research Foundation, Singapore and DSO National Laboratories under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016), NRF 2018 Fellowship NRF-NRFF2018-07, NRF2019-NRF-ANR095 ALIAS grant, grant PIESGP-AI-2020-01, AME Programmatic Fund (Grant No.A20H6b0151) from the Agency for Science, Technology and Research (A*STAR) and Provost's Chair Professorship grant RGEPPV2101.

References

[Abernethy *et al.*, 2021] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-Iterate Convergence Rates for Min-Max Optimization: Convergence of Hamiltonian Gradient Descent and Consensus Optimization. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 3–47. PMLR, 1 2021.

[Anagnostides *et al.*, 2022] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On Last-Iterate Convergence Beyond Zero-Sum Games. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 536–581. PMLR, 8 2022.

[Aumann, 1989] Robert J Aumann. Game Theory. In *Game Theory*, pages 1–53. Palgrave Macmillan UK, London, 1989.

[Aydın *et al.*, 2022] Sarper Aydın, Sina Arefizadeh, and Ceyhun Eksin. Decentralized Fictitious Play in Near-Potential Games With Time-Varying Communication Networks. *IEEE Control Systems Letters*, 6:1226–1231, 2022.

[Bailey and Piliouras, 2019a] James Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Bailey and Piliouras, 2019b] James P Bailey and Georgios Piliouras. Multi-Agent Learning in Network Zero-Sum Games is a Hamiltonian System. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 233–241, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[Cai *et al.*, 2016] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: a generalization of minmax. *Mathematics of Operations Research*, 41(2):648–656, 5 2016.

[Candogan *et al.*, 2013] Ozan Candogan, Asuman Ozdaglar, and Pablo A. Parrilo. Dynamics in near-potential games. *Games and Economic Behavior*, 82:66–90, 11 2013.

[Cheng and Ji, 2022] Daizhan Cheng and Zhengping Ji. Weighted and near weighted potential games with application to game theoretic control. *Automatica*, 141:110303, 7 2022.

[Cheung and Tao, 2021] Yun Kuen Cheung and Yixin Tao. Chaos of learning beyond zero-sum and coordination via game decompositions. In *International Conference on Learning Representations*, 2021.

[Ewerhart and Valkanova, 2020] Christian Ewerhart and Kremena Valkanova. Fictitious play in networks. *Games and Economic Behavior*, 123:182–206, 9 2020.

[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*, 110(4):1232–1236, 2013.

[Galla, 2011] Tobias Galla. Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(8), 8 2011.

[Gemp *et al.*, 2022] Ian Gemp, Rahul Savani, Marc Lanctot, Yoram Bachrach, Thomas Anthony, Richard Everett, Andrea Tacchetti, Tom Eccles, and János Kramár. Sample-based Approximation of Nash in Large Many-Player Games via Gradient Descent. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '22, pages 507–515, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.

[Hadikhanloo *et al.*, 2022] Saeed Hadikhanloo, Rida Laraki, Panayotis Mertikopoulos, and Sylvain Sorin. Learning in nonatomic games part I Finite action spaces and population games. *Journal of Dynamics and Games. 2022*, 0(0):0, 2022.

[Harris, 1998] Christopher Harris. On the Rate of Convergence of Continuous-Time Fictitious Play. *Games and Economic Behavior*, 22(2):238–259, 2 1998.

[Hofbauer and Sorin, 2005] Josef Hofbauer and Sylvain Sorin. Best response dynamics for continuous zero–sum games. *Discrete and Continuous Dynamical Systems - B. 2006, Volume 6, Pages 215-224*, 6(1):215, 10 2005.

[Hofbauer, 1996] Josef Hofbauer. Evolutionary dynamics for bimatrix games: A Hamiltonian system? *Journal of Mathematical Biology*, 34(5-6):675–688, 1996.

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

[Leonardos and Piliouras, 2022] Stefanos Leonardos and Georgios Piliouras. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence*, 304:103653, 2022.

[Leonardos *et al.*, 2021] Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality. *Advances in Neural Information Processing Systems*, 34:26318–26331, 12 2021.

[Li *et al.*, 2018] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems*, 2018.

[McKelvey and Palfrey, 1995] Richard D. McKelvey and Thomas R. Palfrey. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38, 7 1995.

[Melo, 2021] Emerson Melo. On the Uniqueness of Quantal Response Equilibria and Its Application to Network Games. *SSRN Electronic Journal*, 6 2021.

[Mertikopoulos *et al.*,] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. *Cycles in Adversarial Regularized Learning*.

[Metrick and Polak, 1994] Andrew I Metrick and Ben Polak. Fictitious play in 2 • 2 games: a geometric proof of convergence*. *Econ. Theory*, 4:923–933, 1994.

[Monderer and Shapley, 1996] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 5 1996.

[Mukhopadhyay and Chakraborty, 2020] Archan Mukhopadhyay and Sagar Chakraborty. Deciphering chaos in evolutionary games. *Chaos*, 30(12):121104, 12 2020.

[Nisan *et al.*, 2007] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.

[Roberson, 2006] Brian Roberson. The Colonel Blotto game. *Economic Theory*, 29(1):1–24, 2006.

[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*, 8(1):4902, 2018.

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

[Sato *et al.*, 2002] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4748–4751, 4 2002.

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

[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. *Autonomous Agents and Multi-Agent Systems*, 12:115–153, 2006.

[Vlatakis-Gkaragkounis *et al.*, 2020] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-Regret Learning and Mixed Nash Equilibria: They Do Not Mix. *Advances in Neural Information Processing Systems*, 33:1380–1391, 2020.