Mutation in Evolutionary Game Dynamics and Learning; towards Evolving Topologies of Interaction Networks

A doctoral dissertation submitted by

Johann Bauer

in fulfilment of the requirements for the degree of Doctor of Philosophy.

CITY UNIVERSITY
LONDON

City, University of London

School of Mathematics, Computer Science and Engineering

Department of Mathematics

December 2020

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my gratitude to all the people who have contributed to making this piece of research possible, knowingly and unknowingly. This would not have been possible without the myriads of people who have steadily moved the scientific community to becoming more collaborative and more inclusive–step by tiny step, as could not be expected otherwise from science. I would like to thank my institution, City, University of London, for providing funding for me to pursue this research, and to thank all the people in the administration and behind the scenes at City making this possible. I would further like to thank my committee members, Dr Andrea Baronchelli (City), Dr Andrew Morozov (University of Leicester), and Dr Bogdan Stefanski (City), for their rigorous yet friendly examination of this thesis.

It should go without saying that I am deeply grateful to my supervisors Mark Broom and Eduardo Alonso for their continued and reliable support, the frequent discussions–scientific and otherwise–and their thorough engaging with the questions treated herein; for their contribution to my development as a researcher and for their company on research trips. I would like to thank the University of North Carolina at Greensboro and Igor Erovenko and Jan Rychtář for being such friendly hosts during two research stays and likewise to thank the University of Siena and Marco Gori and Alessandro Betti for hosting me.

Of course, none of this would have been possible without the support of my family and friends, especially without the support of my parents to whom I am very grateful. I owe special thanks to Dr Nike Dreyer for her loving support, and to Momo and to Kasimir Franz for reminding me of the important things in life and being extraordinary scientific catalysts.

Thank you.

Declaration

I, Johann Bauer, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

I further grant powers of discretion to the Director of Library Services of City, University of London, to allow the thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Signed: 

Date: 18th December, 2020

Abstract

The present thesis considers two biologically significant processes: the evolution of populations of organisms through natural selection and the change of individual behaviour through learning. More specifically, we consider evolution and learning as guided by the interactions of multiple populations or of multiple learners, respectively, where we assume that these interactions can be described in the language of game theory. While the evolution of populations is often considered in the framework of evolutionary game theory and learning is often considered in the framework of multi-agent reinforcement learning, this thesis strives to present a common perspective on these two classes of processes by analysing the relation between the multi-population replicator dynamics of evolutionary games and simple multi-agent reinforcement learning algorithms.

In particular, this thesis addresses the question of when such processes lead a system of interest, be it populations of organisms or individual learners, to states which reflect the game theoretic structure describing the interactions between populations or between individuals, respectively, and specifically when such systems converge to the Nash equilibria of the underlying game. We motivate these ideas by considering a preliminary application to learning in artificial neural networks, as a concrete multi-agent learning setting of high interest in the fields of artificial intelligence and machine learning. We address the challenges to obtain convergence to interior Nash equilibria in multi-population replicator dynamics by considering more closely the effects of weak mutation. In order to more explicitly account for mutation we specify a replicator-mutator dynamics and relate the equilibria of these dynamics to the underlying game's Nash equilibria in a precise manner, showing that this relation is independent of the choice of mutation parameters. We further prove that such mutation stabilises Nash equilibria in two-player zero-sum games. Finally, we demonstrate how our results regarding the replicator-mutator dynamics can inform the formulation of concrete multi-agent learning algorithms and provide an analytical investigation of the convergence properties of such a learning algorithm.

1 Introduction

The evolution of populations through natural selection and the change of individual behaviour through learning are biologically important but quite distinct processes. However, the latter, i.e., the ability of an individual to change its behaviour through learning, has evolved under the conditions of the former, i.e., under the conditions of natural selection. Furthermore, from a mathematical point of view, both processes tend to move towards a state or behaviour that is better "adapted," under some suitable interpretation, to a given environment. Under this premise, the current thesis strives to consider a common perspective on evolutionary and learning processes.

Commonly, learning is thought of as a process centred on an agent, i.e., a biological organism or a computational agent, while evolution concerns populations of agents and their properties as they change over time. In particular, evolutionary game theory considers the changes in populations when reproduction and other aspects are determined by interactions in a game theoretical setting and we can ask, whether and under which conditions populations might evolve towards a state which reflects the game theoretical properties of their interactions. When learning is considered in settings that include multiple agents, game theory provides a useful framework. Just as evolution can be driven by the outcomes of interactions, so can the changes in agents' behaviours and we can ask, whether a system of interacting learning agents, too, can acquire or learn a behaviour which reflects the game theoretical properties of the interactions, e.g., behave in accordance with a game's Nash equilibrium.

The evolutionary game theoretic perspective on multi-agent learning becomes particularly relevant when we require multi-agent learning under very restrictive conditions; in particular, when we consider agents to be cognitively extremely simple, departing from the often assumed perfect rationality in game theory, to only have a very limited ability to observe other agents' behaviours and to have no knowledge of their own payoffs in advance and no knowledge of other agents' payoffs at all. Such a situation might be formulated as follows:

(†) Given a set of very simple agents with a limited ability to observe each others' actions and an ability to react to those observed actions, and given some optimality criterion, under which conditions will the system converge to a state reflecting the game theoretic structure of the interactions?

Answers to questions similar to ours are actively sought in, among others, the areas of reinforcement learning, multi-agent systems and multi-agent reinforcement learning, and we will refer to the relevant results where appropriate. However our approach differs significantly from existing approaches, not least in an explicit consideration of some degree of computational simplicity of the agents and a focus on a mathematically rigorous treatment. We are not aware of any publications pursuing a line of thought sufficiently close to ours, although there are publications having relevance for individual aspects of our approach, which we refer to where appropriate. Our approach is further informed by the following considerations: First, there are (at least) two limit cases of our question. In the one case, convergence should occur if we assume that there is in fact only a single agent instead of a multitude. In the other case, we can assume that, apart from a focal agent, all agents have some form of stationary behaviour. In both cases, our question reduces to a Markov decision process and we can employ the existing results and solution approaches to consider these limit cases of our question. As we will see, reinforcement learning provides an answer to our question in these limit cases.

Second, proceeding from the single-agent case to a multi-agent case, game theory provides a formal framework and certain concepts in terms of which our question can be formulated at least partially. However, the main concepts and results of game theory are mostly static and we need to consider an extended formulation of game theory in order to include the dynamic aspects of our question. In this respect, the merits of an evolutionary game theoretic approach are that it prefers simple dynamics over more complicated ones by design and that it provides suitable concepts to investigate the dynamics and their stationary points in terms of game theory.

Furthermore, the research on multi-agent reinforcement learning investigates the possibilities of extending the methods of reinforcement learning to the multi-agent setting and we will consider in how far these extensions can be understood from an evolutionary game theoretic perspective and how they can inform our approach.

The objective of this thesis is to contribute to the mathematical foundations for the analysis of the evolution of interactions in networks, be they networks of interacting populations or of agents, describable as directed graphs with general topologies, including cyclic graphs, and the evolution of such topologies induced by the evolution of the populations or agents located at the vertices and their dependencies on other populations or agents.

A basic and important questions regarding dynamic processes is the role of fixed points and their stability. In particular, asymptotic stability allows one to analyse a system's dynamics in terms of its asymptotically stable equilibria as suitably close initial conditions lead a system to converge to such equilibria. In particular in the area of machine learning, as a subfield of mathematical optimization, questions of convergence are central.

The questions addressed in this thesis are focused on questions of stability of equilibria and the relations between deterministic continuous-time dynamics and stochastic discrete-time dynamics. The former are often, but not exclusively, considered in evolutionary game theory and tend to be easier to analyse mathematically, while the latter often result from learning algorithms. The rest of this thesis is structured as follows:

• *In chapter 2* we provide the general theoretical results from the literature required. In particular, we give an overview of results from evolutionary game theory and from the theory of Markov decision processes, focusing on the relevant parts of each, and of how our approach differs from related questions in the literature in 2.3.1.

• *In chapter 3* we provide a preliminary attempt to combine these two perspectives in a motivating application to learning in networks of agents, illustrating the general ideas in a simple artificial neural network setting. To our knowledge, the learning mechanism explored in this chapter is the first such approach to formulate learning on the level of individual neurons based on rewards derived from the network's overall behaviour. In particular, chapter 3 demonstrates that the combination of these ideas is not easily achieved in a well-informed ad-hoc adaptation of well-known basics of artificial neural networks and illustrates the limits of such an approach. It therefore motivates the pursuit of a mathematically rigorous foundation on which these ideas can be based, as attempted in the following chapters.

• *In chapter 4* we consider an extension of a standard multi-population replicator dynamics to include certain kinds of mutation in addition to replication, which was

also published as [8]. These results demonstrate that relaxing the assumption of the negligibility of mutation fundamentally affects the stability properties of equilibria in multi-population replicator dynamics and, in principle, allows for equilibria to become asymptotically stable. A central contribution of this chapter is that there are equilibria with properties which are independent of certain mutation parameters. In particular, this makes the resulting multi-population replicator-mutator dynamics a candidate for a well-founded learning dynamics in multi-agent learning.

• *In chapter 5* we consider how mutation affects the stability properties of equilibria in zero-sum settings, in preparation for publication as [7]. More specifically, we prove that mutation stabilises equilibria in 2×2 games and potentially larger two-player zero-sum games, for all mutations in the class of mutation mechanisms specified in chapter 4.

• *In chapter 6* we relate the previous results explicitly to a learning perspective and provide a stochastic algorithm which is related to the deterministic dynamics considered earlier in a precise manner, with results being prepared for publication as [9] and with parts of the numerical implementation being contributed by Sheldon West.[1] This chapter demonstrates the learning behaviour in 2×2 games and zero-sum games and how the speed of learning changes with game size, comparing results to two algorithms from the literature. These results point in the direction in which a rigorously analysable learning dynamics for larger multi-agent systems in the spirit of the ideas considered in chapter 3 might be sought.

• *In chapter 7* we discuss the implications of the presented results for understanding the role of mutation in evolutionary games and learning, their limitations based on the assumptions we made, potential future questions and the directions in which these very first steps point in order to further advance the ideas outlined in this thesis.

Further results on evolutionary games and network topologies to which the author of this thesis has contributed as a co-author are published in [31]. However, the questions considered there do not directly relate to the main question of this thesis and therefore will not be elaborated upon.

[1] The initial program code used for simulations was the result of an excellent MSc thesis by Sheldon West which the author co-supervised.

2 Evolutionary dynamics and learning

We begin with a presentation of Markov decision processes and some fundamental ideas of reinforcement learning as these allow us to consider our question in the limit case of a single agent in a stochastic but stationary environment. This will allow us to formulate a first motivating application in the framework of game theory and to approach the interaction of agents more explicitly in chapter 3.

2.1 Markov decision processes and reinforcement learning

We introduce the basic framework of Markov decision processes and the main results that will have a bearing on the understanding of our problem. The theory of Markov decision processes considers the setting of an agent acting on a stochastic system as it evolves and receiving a sequence of rewards depending on the system's states. The guiding question is whether there is a strategy for the agent that satisfies some specific optimality criterion. The classic results give the optimality conditions of a strategy given the specification of the process and guarantee the existence of optimal strategies in a subset of all possible strategies.

Corresponding well with our assumption of minimal apriori knowledge on part of the agents, one common practical assumption is that the specification of the process is not known in advance. This motivates solution approaches which adapt strategies depending on the rewards received during the evolution of the system. In this situation the classical results provide the conditions under which such approaches are guaranteed to converge to optimal strategies. This is the basic setting of reinforcement learning where behaviours yielding high rewards are strengthened, i.e., increased in frequency. In its advanced variants, reinforcement learning has been shown to be able to arrive at strategies that take into account rewards that are to be

expected in the far future, as in the situation of complex strategic board games such as backgammon [106] or the game of Go [94].

We restrict our attention to infinite horizon Markov decision processes in discrete time with discrete state space and discrete action space, being close to repeated and stochastic games and thus having the most relevance for us. The presentation mainly follows [80], to which we refer the reader for a more detailed presentation of the topic.

Definition 2.1.1 (MDP)**.** We call a tuple $(T, S, (A_s), (p_t(\cdot|s,a)), (r_t))$ with

i) T a set of *decision epochs*, with the elements of T referred to as *stages* or *times*,

ii) S a set of *states* the system can occupy,

iii) $(A_s)_{s\in S}$ a family of non-empty action sets, with $A := \bigcup_{s\in S} A_s$,

iv) $(p_t(\cdot|s,a))_{s\in S, a\in A, t\in T}$ a family of probability distributions on S, the *transition probability functions*,

v) $r_t : S \times A \times S \to \mathbb{R}$ reward functions for every time $t \in T$

a *Markov decision process* (MDP).

We further make the following assumptions:

i) $T = \mathbb{N}$, i.e., we focus on infinite-horizon discrete time MDPs;

ii) S finite;

iii) $A_s = A$ $(\forall s \in S)$ and A finite, i.e., available actions do not depend on the system's state;

iv) The transition probability functions p_t are independent of t , i.e., we can represent the family of transition probability functions (p_t) as the family of matrices $(P_a)_{a\in A}$, with $[P_a]_{s,s'} = p_t(s'|s,a)$ $(\forall s, s' \in S, a \in A)$;

v) The reward functions r_t are independent of t and only depend on the subsequent state s', i.e., $r_t(s,a,s') = r(s')$, and thus can be represented as a vector in $r \in \mathbb{R}^{|S|}$.

Under these assumptions we define a Markov decision process for our purposes as the tuple $(S, A, (P_a), r)$.

2.1.1 Decision rules and strategies

Definition 2.1.2 (History)**.** Given an MDP, $(S, A, (P_a), r)$, we call a tuple

$$h_t = (s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t)$$

a history of length t, where $s_\tau \in S$ and $a_\tau \in A$ are the state of the system and the action taken at time τ respectively. We can write recursively $h_t = (h_{t-1}, a_{t-1}, s_t)$. We denote the set of all histories of length t by H_t. Note that $H_1 = S$ and $H_{t+1} = H_t \times A \times S$, and we set $H_\infty = (S \times A)^T$, i.e., the functions $T \to (S \times A)$, as the set of all terminal histories.

Remark. Note that we could extend the set of states S to $\tilde{S} = A \times S$, and thus let an action taken at $t-1$ be absorbed by the state at $s_t \in \tilde{S}$. Apart from a separate consideration of initial states this leads to a slightly, but not significantly, differing definition of a history, which we will use when introducing repeated games.

Definition 2.1.3 (Decision rules & strategies)**.** We call a function $d_t : H_t \to A$ a *deterministic decision rule*, and D_t the set of available decision rules at t. A function $\pi : T \to \bigcup_{t \in T} D_t$ with $\pi(t) \in D_t$ then is called a *deterministic strategy* (or *policy*).

Let $\mathcal{D}(A)$ denote the set of probability distributions on A. A function $d_t : H_t \to \mathcal{D}(A)$ is called a *randomized decision rule*. Accordingly, a function π that assigns a randomized decision rule $\pi(t) \in \mathcal{D}(D_t)$ to every time t is called a *randomized strategy*. A strategy π is called *stationary* if $\pi(t) = \pi(t')$ for all $t, t' \in T$.

We can consider deterministic decision rules as randomized decision rules that assign a degenerate probability distribution, i.e., with $(d_t(h_t))(a) = 1$ for $h_t \in H_t$ and some $a \in A$. We further introduce the following special kind of decision rules and corresponding strategies:

Definition 2.1.4 (Markovian decision rule)**.** A deterministic or randomized decision rule, d_t, is called *Markovian* (or *memoryless*) if for any two histories $h_t, h_t' \in H_t$ with the same state at time t, i.e., $h_t = (h_{t-1}, a_{t-1}, s_t)$ and $h_t' = (h_{t-1}', a_{t-1}', s_t)$, we have $d_t(h_t) = d_t(h_t')$. In this case, we consider d_t a function $S \to A$ or $S \to \mathcal{D}(A)$ respectively.

We denote the set of all Markovian deterministic rules by D^{MD} and all such strategies by Π^{MD} and the set of all Markovian randomized rules by D^{MR} and all such strategies by Π^{MR}.

Markovian decision rules and strategies are of special interest because they are much simpler than strategies which properly depend on the history of the process and therefore simplify the resulting stochastic process.

Remark. We refer to the set of non-Markovian deterministic or randomized decision rules by D_t^{HD} or D_t^{HR} respectively, i.e., these rules properly depend on at least some history of the process. The sets of corresponding strategies are denoted by Π^{HD} and Π^{HR}.

An MDP $(S, A, (P_a), r)$ together with a strategy π (from Π^{HR} or Π^{MR}) generate a stochastic process

$$\{(X_t, Y_t) | t \in T, X_t \in S, Y_t \in A\}$$

on the probability space $(H_\infty, \mathfrak{B}(H_\infty), P_\pi)$, where $\mathfrak{B}(H_\infty)$ is the Borel algebra over H_∞ and, given some distribution over initial states p_1, P_π is induced by:

$$P_\pi(X_1 = s) = p_1(s)$$
$$P_\pi(Y_t = a | h_t) = \pi_t(h_t, a)$$
$$P_\pi(X_{t+1} = s | (h_{t-1}, a_{t-1}, s_t), A_t = a_t) = p(s | s_t, a_t)$$

If π is a (randomized or deterministic) Markovian strategy, the induced stochastic process is a discrete time Markov chain on a finite state space. Otherwise, the induced stochastic process is in general non-Markovian. In the following we will not require a deeper theory of stochastic processes and therefore refer the reader to the literature on MDPs for a more detailed presentation, e.g. [80].

A note on terminology. Both terms, *policy* and *strategy*, are equally valid and seem to be used synonymously in the MDP literature. However, *policy* seems to be the more dominant one in MDP theory and in reinforcement learning, while *strategy* is convenient for the embedding of MDP theory in game theory, where *strategy* is the dominant term and is defined in essentially the same way, as we will see later. We will, therefore, predominantly use the term *strategy*. However, the reader should keep in mind that *policy* is synonymous, as it might occur in special circumstances, e.g., in proper names of certain strategies.

Value of a strategy

In order to be able to compare strategies to each other, we first define the value of a strategy over a finite horizon of length $N+1$:

Definition 2.1.5. For an initial state $s \in S$, i.e., for $p_1(s) = 1$, we define the value of a strategy π over $N+1$ stages as

$$v^{\pi}_{N+1}(s) = \mathbb{E}^{\pi}_{s}\left[\sum_{t=1}^{N} r(X_t, Y_t, X_{t+1})\right].$$

We further define the value of the MDP over $N+1$ stages for an initial state s as

$$v^{*}_{N+1}(s) = \sup_{\pi \in \Pi^{HR}} v^{\pi}_{N+1}(s).$$

Thus, we will be interested in finding a strategy π^* such that

$$v^{\pi^*}_{N+1}(s) = v^{*}_{N+1}(s).$$

As we will be interested in the infinite horizon MDP, we introduce the following extensions of the notion for a strategy π:

Definition 2.1.6. For $s \in S$, we define the following three criteria:

i) Expected total reward:

$$v^{\pi}(s) = \lim_{N\to\infty} \mathbb{E}^{\pi}_{s}\left[\sum_{t=1}^{N} r(X_t, Y_t, X_{t+1})\right] = \lim_{N\to\infty} v^{\pi}_{N+1}(s)$$

The existence of the limit is not guaranteed. For optimality concepts in this case we refer to [80]. Where it exists, we define $v^*(s) = \sup\{v^{\pi}(s) \mid \pi \in \Pi^{HR}\}$.

ii) Expected total discounted reward:

$$v^{\pi}_{\gamma}(s) = \lim_{N\to\infty} \mathbb{E}^{\pi}_{s}\left[\sum_{t=1}^{N} \gamma^{t-1} r(X_t, Y_t, X_{t+1})\right]$$

for $0 \leq \gamma < 1$. The limit exists if $\{r(s,a,s') \mid s,s' \in S, a \in A\}$ is bounded, which is the case for finite S and A. We define $v^{*}_{\gamma}(s) = \sup\left\{v^{\pi}_{\gamma}(s) \;\middle|\; \pi \in \Pi^{HR}\right\}$.

iii) Average expected reward:

$$g^{\pi}(s) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}^{\pi}_{s}\left[\sum_{t=1}^{N} r(X_t, Y_t, X_{t+1})\right] = \lim_{N\to\infty} \frac{1}{N} v^{\pi}_{N+1}(s)$$

Similarly, we define $g^*(s) = \sup\left\{g^{\pi}(s) \;\middle|\; \pi \in \Pi^{HR}\right\}$.

Remark. Note that the total discounted reward criterion corresponds to an expected total reward criterion over ν stages, where ν is a random variable following a geometric distribution with parameter γ, i.e.,

$$v_\nu^\pi(s) = \mathbb{E}_s^\pi \left[\mathbb{E}_\nu \left\{ \sum_{t=1}^{\nu} r\left(X_t, Y_t, X_{t+1}\right) \right\} \right]$$

where $P(\nu = n) = (1-\gamma)\gamma^{n-1}$ for $n = 1, 2, \ldots$. This is the statement of the next proposition, for the proof of which we refer to [80, proposition 5.3.1, p. 126].

Proposition 2.1.7. *Suppose S and A are finite, and ν has a geometric distribution with parameter γ. Then $v_\nu^\pi(s) = v_\gamma^\pi(s)$ for all $s \in S$.*

According to this proposition, a strategy which is optimal under the expected total discounted reward criterion is also optimal if we evaluate the value of a strategy over a finite duration where the duration has a geometric distribution. It is also equivalent to the total expected reward if we extend the state space by a special absorbing end state with zero reward and adapt the transition probabilities accordingly, i.e., by multiplying the transition probabilities by γ and setting the transition probabilities to the newly added end state to $1-\gamma$.

2.1.2 Markov strategies

The following results illustrate the importance of Markovian strategies and why we may restrict our attention to Markov strategies instead of considering all possible strategies:

Proposition 2.1.8 ([80, theorem 5.5.1])**.** *Let $\pi \in \Pi^{HR}$. Then for each $s \in S$, there exists $\pi' \in \Pi^{MR}$ such that for all $t \in T$ and all $j \in S, a \in A$:*

$$P_\pi(X_t = j, Y_t = a | X_1 = s) = P_{\pi'}(X_t = j, Y_t = a | X_1 = s)$$

In words, we can always find a process induced by a Markovian strategy, which has the same probabilities over state-action pairs for all times. From this proposition directly follows that the values of such strategies are identical:

Proposition 2.1.9 ([80, theorem 5.5.3])**.** *Let $\pi \in \Pi^{HR}$. Then for each $s \in S$, there exists $\pi' \in \Pi^{MR}$ such that (whenever the respective limits exists):*

i) $v_N^{\pi'}(s) = v_N^\pi(s)$ for all $N \in T$

ii) $v_\gamma^{\pi'}(s) = v_\gamma^{\pi}(s)$ *for all* $0 \leq \gamma < 1$

iii) $g^{\pi'}(s) = g^{\pi}(s)$

Therefore, to find an optimal strategy for an MDP, it is sufficient to consider the randomized Markov strategies, i.e., the functions $T \to D^{MR}$. This proposition does not yet imply that these strategies have to be stationary, i.e., they can assign a different randomized Markov decision rule at every point in time. However, that we can indeed an optimal strategy which is a stationary Markov strategy, is demonstrated in the following.

Let P_d denote the transition probability matrix of an MDP induced by a Markov randomized decision rule $d \in D^{MR}$ and let $\pi \in \Pi^{MR}$ be such that $\pi = (d_1, d_2, ...)$. Then we can write the probability that the process, given an initial state $s \in S$, is in some state $s' \in S$ at time $t+1$ as

$$P_\pi\left(X_{t+1} = s' \,|X_1 = s\right) = \left[P_{d_t} \cdot P_{d_{t-1}} \cdots P_{d_1}\right]_{s,s'} =: \left[P_\pi^t\right]_{s,s'}.$$

Furthermore, given a real valued function w on S (assumed finite), we have for $\pi \in \Pi^{MR}$:

$$\mathbb{E}_s^\pi\left[w\left(X_t\right)\right] = \sum_{s' \in S} \left[P_\pi^{t-1}\right]_{s,s'} w(s')$$

Therefore, for the discounted total reward criterion where the reward depends solely on the *subsequent* state, we have

$$v_\gamma^\pi(s) = \mathbb{E}_s^\pi\left[\sum_{t=1}^{\infty} \gamma^t r\left(X_{t+1}\right)\right] = \sum_{t=1}^{\infty} \gamma^t \sum_{s' \in S} \left[P_\pi^t\right]_{s,s'} r(s')$$

or in vector notation

$$v_\gamma^\pi = \sum_{t=1}^{\infty} \gamma^t P_\pi^t r.$$

If π is a stationary randomized Markov strategy, i.e., $\pi = (d, d, ...)$ we have that

$$P_\pi\left(X_{t+1} = s' \,|X_1 = s\right) = \left[\underbrace{P_d \cdot P_d \cdots P_d}_{t \text{ times}}\right]_{s,s'} = \left[P_d^t\right]_{s,s'}.$$

Depending on the context, we might write P_π instead of P_d if it is clear that π is a stationary Markov strategy.

The value of the stationary randomized Markov strategy π simplifies to

$$v_\gamma^\pi = \sum_{t=1}^{\infty} \gamma^t P_d^t r$$

and we have the recursive relationship

$$v_\gamma^\pi = \gamma P_d r + \gamma P_d \sum_{t=1}^{\infty} \gamma^t P_d^t r = \gamma P_d r + \gamma P_d v_\gamma^\pi = r_d + \gamma P_d v_\gamma^\pi$$

where $r_d = \gamma P_d r$. This is summarized in the following:

Proposition 2.1.10. *Let $0 \leq \gamma < 1$. Then for any stationary strategy $\pi = (d, d, ...)$ with $d \in D^{MR}$, the value v of π is given by the unique solution of*

$$v = r_d + \gamma P_d v \tag{2.1.1}$$

which can be written as

$$v = (I - \gamma P_d)^{-1} r_d.$$

Proof. A detailed proof is given in [80, theorem 6.1.1, p. 145] and relies on P_d being a stochastic matrix, thus $\|\gamma P_d\| < 1$. Therefore via the Neumann series, $(I - \gamma P_d)^{-1}$ exists. □

Let V be the space of bounded functions $S \to \mathbb{R}$ with supremum norm and componentwise partial order. We introduce the non-linear operator $\mathscr{L}$ on V as

$$\mathscr{L}v = \sup_{d \in D^{MD}} \{r_d + \gamma P_d v\}, \tag{2.1.2}$$

where we refer to [80, lemma 5.6.1] for the fact that $\mathscr{L}[V] \subseteq V$. Note that for finite A, as we assume, the supremum on the right-hand side is attained for all $v \in V$, and in particular for S finite, we may assume $V = \mathbb{R}^{|S|}$. One useful property of the operator $\mathscr{L}$ is summarized in the following [80, theorem 6.2.2, p. 148]:

Proposition 2.1.11. *Suppose there is $v \in V$ such that*

i) $v \geq \mathscr{L}v$, then $v \geq v_\gamma^*$;

ii) $v \leq \mathscr{L}v$, then $v \leq v_\gamma^*$;

iii) $v = \mathscr{L}v$, then v is the unique such solution and $v = v_\gamma^*$.

The main result for our purposes is that under our assumptions, there exists a stationary deterministic Markov strategy π that is optimal, i.e., for which $v_\gamma^\pi = v_\gamma^*$. The following proposition secures this existence, as the supremum in (2.1.2) is attained:

Proposition 2.1.12. *Suppose S is countable. Then $\mathscr{L}$ has a unique fixed point $v^* \in V$ and $v^* = v_\gamma^*$.*

Proof. The proof is given in [80, theorem 6.2.5] and relies on the fact that $\mathscr{L}$ is a contracting map and the unique fixed point exists due to the Banach fixed point theorem, and $v^* = v_\gamma^*$ follows from proposition 2.1.11. □

Bellman Equations. From the above it follows that a stationary Markov strategy π^* is optimal *iff*

$$v_\gamma^{\pi^*} = v_\gamma^* = \mathscr{L}v_\gamma^* = \max_{d \in D^{MD}} \left\{ r_d + \gamma P_d v_\gamma^{\pi^*} \right\}$$

or in component-wise notation

$$v_\gamma^{\pi^*}(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p\left(s' \,|\, s,a\right) v_\gamma^{\pi^*}(s') \right\}, \quad s \in S, \tag{2.1.3}$$

i.e., if it is equivalent to picking the action that solves the one-step optimization problem and following π^* thereafter, where $r(s,a) = \gamma \sum_{s' \in S} p\left(s' \,|\, s,a\right) r\left(s,a,s'\right)$. The equations (2.1.3) are also referred to as the *Bellmann equations* or *optimality equations*.

Remark. If the transition probabilities $p(s'|s,a)$ are known, we can compute v_γ^* numerically by solving $v_\gamma^* = \mathscr{L}v_\gamma^*$ iteratively, as well as v_γ^π for a Markov strategy π by solving equation (2.1.1). A range of algorithms to find an optimal strategy are known, where two prominent approaches are value iteration and policy iteration approaches which are analysed in detail in [80].

If the transition probabilities are not known, an algorithm can compute an estimate of the probabilities by sampling. Such approaches are considered model-based solutions as they rely on estimating a model of the MDP. In contrast, model-free approaches combine this sampling and the search for an optimal strategy such that an explicit estimation of a model is not necessary.

2.1.3 Solutions based on action-value learning

One class of model-free algorithms relies on an estimate of the *action-value* function, which we define as follows:

Definition 2.1.13. Given a stationary Markov strategy π, we define the *action-value function* $Q^\pi : S \times A \to \mathbb{R}$ associated with π as

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\gamma^\pi(s')$$

and Q^* corresponding to v_γ^* as

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\gamma^*(s').$$

The Q-function thus gives the value of choosing some action deterministically and following some strategy thereafter.

Remark 2.1.14. From the definition of the action-value function and the properties of the value function, we can derive a similar recursive relationship for Q:

$$\sum_{a \in A} \pi(s,a) Q^\pi(s,a) = \sum_{a \in A} \pi(s,a) \left(r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\gamma^\pi(s') \right) = v_\gamma^\pi(s)$$

and thus by substitution in the definition of Q

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(s',a') Q^\pi(s',a').$$

and similarly with (2.1.3)

$$v_\gamma^*(s) = \max_{a \in A} \{Q^*(s,a)\}$$

we have that

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \max_{a' \in A} \{Q^*(s',a')\}.$$

While we need the transition probabilities to construct an optimal strategy from a known v^*, we can construct a deterministic Markov strategy d^*, that is optimal, directly from Q^* by choosing

$$d^*(s) \in \arg\max_{a \in A} \{Q^*(s,a)\}.$$

That d^* is indeed an optimal strategy follows from the optimality equations (2.1.3). Such a strategy, i.e., one that picks the action with a maximal action-value, is called a *greedy strategy* or *greedy policy*. Closely related are the ε-greedy strategies, which with probability $1-\varepsilon$ pick an action with a maximal action-value and with some probability ε pick a random action with uniform probability, thus ensuring sufficient exploration of the action-values.

Sarsa and Q-learning algorithms

Two solution approaches based on learning an action-value function can be distinguished: *on-policy* learning, where the action-value function Q^π of the current strategy (or policy) π is estimated; and *off-policy* learning, where the optimal action-value

function Q^* is estimated. As an example of the first, we present the Sarsa algorithm (algorithm 2.1.1), first introduced in [85] and further elaborated upon in [100]. As an example of the second, we present Q-learning (algorithm 2.1.2), proposed by [116] with a convergence proof in [117]. Both algorithms can be considered special cases of temporal difference learning, introduced in [101], the convergence properties of which essentially rely on stochastic approximation theory based on [30].

Our aim is primarily to show the similarity between both approaches and that they exploit the fact that the action-value function is a fixed point of a contracting operator.

Algorithm 2.1.1 Sarsa [102, p. 146].

Let π_Q be a Markov strategy derived from an action-value function estimate Q, e.g., ε-greedy. Then the Sarsa algorithm proceeds as follows:

1. Initialize $Q(s,a)$ arbitrarily.
2. Repeat (for each episode):
   a) Initialize s;
   b) Given the state s, choose a according to π_Q;
   c) Repeat (for each time step in episode) until s is terminal:
      i. Take action a, observe reward $r(s,a,s')$ and subsequent state s';
      ii. Given s', choose a' according to π_Q;
      iii. $Q(s,a) \leftarrow Q(s,a) + \alpha\Big(r + \gamma Q(s',a') - Q(s,a)\Big)$;
      iv. $s \leftarrow s'; a \leftarrow a'$;

Under the condition that π_Q converges to a greedy strategy with respect to the estimated action-value function Q, and under further conditions on the learning rates α, [96] provides a proof that the action-value function estimated by Sarsa converges to Q^* and thus π_Q converges to an optimal strategy. It is however important that π_Q does not approach a greedy strategy too quickly in order for Q to converge to Q^*, a fact known as the exploration-exploitation trade-off in reinforcement learning.

For Q-learning, convergence of Q to Q^* is proven under similar assumptions as for Sarsa in [50] and [108]. The proofs in [96], [50], and [108] make essential use of Robbins-Monro stochastic approximation, [83], and results in [30].

We will keep in mind that Q^* can be reliably estimated without knowing the transition probabilities of the MDP, but we will abstract away from the concrete process of estimating Q^* for the time being and will instead assume that agents have enough

Algorithm 2.1.2 Q-learning [102, p. 149].

Let π_Q be a Markov strategy derived from an action-value function estimate Q, e.g., ε-greedy. Then the Q-learning algorithm proceeds as follows:

1. Initialize $Q(s,a)$ arbitrarily.
2. Repeat (for each episode):
   a) Initialize s
   b) Repeat (for each step in episode) until s is terminal:
      i. Choose a from s according to π_Q
      ii. Take action a, observe reward r, and subsequent state s'
      iii. $Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma \max_{a'}\{Q(s',a')\} - Q(s,a)\big)$
      iv. $s \leftarrow s'$

opportunity to arrive at a sufficiently good estimate. Thus, the existence of these algorithms allows us to compute Q^* explicitly for the numerical analysis without invalidating our assumptions about the computational simplicity of the agents.

As a concluding remark, reinforcement learning algorithms have been shown to be successful in a wide variety of situations. While Q-learning and Sarsa rely on a complete representation of the action-value function, in practical applications with large state spaces and action spaces, approximations of the action-value function are employed, e.g., through artificial (deep) neural networks. Such deep Q-learning has been successfully employed to play Atari games [68], and [94] shows that deep Q-learning can find a strategy for playing the game of go that surpasses that of any human player, without using expert knowledge, to name just a very few applications. However, convergence of such algorithms employing an approximation of the action-value function is not always guaranteed, [32]. As temporal difference learning in general, so can Sarsa and Q-learning be extended with so-called eligibility traces such that action-values for states that have been visited further in the past are updated as well, resulting in Sarsa(λ) and $Q(\lambda)$, and general TD(λ) algorithms, where λ is an eligibility trace parameter, [102]. The success and versatility of reinforcement learning algorithms make the extension of these algorithms to multi-agent settings an important question and a field of active research.

2.2 Game theory and repeated games

Having elaborated on the single-agent perspective and having presented the main results, we turn to a perspective that explicitly takes into account the interactions between agents. The investigation of agents interacting, such that other agents' actions have a significant influence on the rewards or payoffs of an agent, has been at the heart of game theory since its early formulations in [114] and [115].

Based on these classical formulations, game theory has been employed to study the evolution of traits, most notably in [63], and especially cooperative behaviour, by extending it to repeated games and incorporating evolutionary population dynamics, often considering the repeated Prisoner's Dilemma as in [6], [41], [42], and [79], but also with respect to the Stag Hunt in [76] and [97], and has seen a wide variety of applications in biology [17]. While these mostly investigate interactions and dynamics on infinite well-mixed populations, evolutionary game theory has also been formulated for structured finite populations as, e.g., evolutionary graph theory in [73] or [1], and further in [16].

Our objective will be to combine the perspective of population dynamics with that of each player simultaneously searching for an optimal strategy as in an MDP. The latter problem has been addressed in the context of stochastic games in [92] and more recently in [59], where it was proposed as an adequate framework to investigate multi-agent reinforcement learning. We therefore introduce the basic concepts of game theory and repeated games, following [75] and [34], and proceed to evolutionary game theory and population dynamics, following [45].

2.2.1 Finite games and Nash equilibria

We begin with the definition of the central concept of game theory:

Definition 2.2.1 (Game)**.** We call the tuple $G = (N, A, (\succsim_i)_{i\in N})$ a *game*, where

i) N is a set of players, where sometimes N will denote the cardinality of N if misunderstanding is unlikely,[1]

ii) $A = \times_{i\in N} A_i$ the set of *action profiles* where $(A_i)_{i\in N}$ is a family of non-empty sets (we call A_i the set of actions available to a player i),

[1] A rigorous approach to this would be to use von Neumann ordinals, which would however imply starting to count with 0 instead of 1.

iii) $(\succsim_i)_{i\in N}$ is a family of *preference relations* on A, i.e., a family of complete reflexive binary relations on A.

We call G a *finite game* if A and N are finite.

Remark 2.2.2. We will in general assume games to be finite unless stated otherwise. Furthermore, we will focus on games $(N, A, (\succsim_i))$, where the preference relations are induced by a family of *payoff functions* $u = (u_i)_{i\in N}$ from A to $\mathbb{R}$, such that:

$$\forall i \in N, a, b \in A : a \succsim_i b \Leftrightarrow u_i(a) \geq u_i(b)$$

In this case, we denote the game by (N, A, u).

To illustrate the definition and provide a general intuition, we introduce the following prominent examples:

Example 2.2.3 (PD)**.** Consider the following 2-player game known as the *Prisoner's Dilemma* (PD). Both players have $\{C, D\}$ as their action set and the payoff function given by the following table, where the row player is player 1:

	C	D
C	4, 4	0, 5
D	5, 0	1, 1

If both players choose C, they both receive a payoff of 4 each. However, if one player plays C, the other player is tempted to "defect", that is play D, and receive a payoff of 5, sometime called the temptation, leaving the other with 0.

As the payoffs of the game are symmetrical, its payoffs can also be expressed as a matrix of the payoffs for player 1:

$$A = \begin{pmatrix} R & S \\ T & P \end{pmatrix}$$

The payoffs for player 2 then are given by A^T. In general, a Prisoner's Dilemma is any such game with $T > R > P > S$ and $2R > T + S$.

Example 2.2.4 (SH)**.** The *Stag Hunt* (SH) game is based on a passage of Rousseau's 1754 *A Discourse on Inequality*, as related in [97, p. 1]: "If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple."

Of course, this is not a specification of a strategic game. From what we have, we infer that a stag is harder to catch than a hare and requires the hunters to focus on hunting the stag. We should also think that a share of the stag is more desirable than a hare, since otherwise we would not concern ourselves with the stag hunt to begin with. We therefore formulate the game as a symmetrical two player game with the following payoffs:

	S	H
S	$10, 10$	$0, 2$
H	$2, 0$	$2, 2$

Formulated as a payoff matrix, the game has the form

$$A = \begin{pmatrix} 10 & 0 \\ 2 & 2 \end{pmatrix}.$$

Compared to the PD payoff structure, we have $R > T = P > S$. Thus, if the two players can commit themselves to hunting the stag, there is no temptation to deviate. However, in the absence of such a commitment possibility it is risky, because that could lead to having neither stag nor hare.

Example 2.2.5 (RPS)**.** Another frequently encountered game in the literature as well as real life is of the *Rock-Paper-Scissors* (RPS) type. The players' action sets are given as $\{R, P, S\}$ and the payoff function can again be given by the following table, where the row player is player 1:

	R	P	S
R	$0, 0$	$-1, 1$	$1, -1$
P	$1, -1$	$0, 0$	$-1, 1$
S	$-1, 1$	$1, -1$	$0, 0$

This game can be framed as a zero-sum game, where the sum of payoffs of the players is always 0, as is done here, but that is not the only sensible representation of such a game, e.g., [17, p. 31].

The following is a central concept in the theory of games which helps to understand the structure of games and characterize certain action profiles:

Definition 2.2.6 (Nash equilibrium)**.** Given a game (N, A, u), a *Nash equilibrium* is an action profile $a^* \in A$ such that:

$$\forall i \in N, a_i \in A_i : \; u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*)$$

That is, given the other players' actions, a^*_{-i}, a player i cannot increase the payoff by deviating from a^*_i. Note that (b_i, a_{-i}) denotes that action profile which is given by $a \in A$ and where a_i is replaced by b_i.

Remark 2.2.7. Note that in a Nash equilibrium $a^* \in A$, each player i's payoff is maximal given the other players' actions, i.e.,

$$u_i(a^*_i, a^*_{-i}) = \max_{a_i \in A_i} u_i(a_i, a^*_{-i}).$$

We also call $B_i : A \to \mathcal{P}(A_i), a \mapsto \{a_i \in A_i \mid \forall b_i \in A_i : u_i(a_i, a_{-i}) \geq u_i(b_i, a_{-i})\}$ the *best-response correspondence* of player i, such that $a^* \in A$ is a Nash equilibrium if and only if $\forall i \in N : a^*_i \in B_i(a^*)$.

Remark 2.2.8. Note that in PD type games the action profile (C, C) is not a Nash equilibrium, as players are tempted to deviate from it. The only Nash equilibrium of the game is (D, D). Furthermore, not every game has an action profile that is a Nash equilibrium, as can be seen from RPS type games.

Mixed strategies and mixed strategy Nash equilibria

Motivated by the observation that not all games have an action profile which is a Nash equilibrium, such as RPS, we introduce the notion of a mixed strategy in order to account for randomized behaviour and to ensure the existence of equilibria in games.

Definition 2.2.9 (Mixed strategy)**.** Given a game (N, A, u), we denote by $\mathcal{D}(A_i)$ the set of probability distributions over A_i and call the elements of $\mathcal{D}(A_i)$ player i's *mixed strategies* or just *strategies*. As usual, given $\sigma \in \mathcal{D}(A_i)$, we refer to $\{a_i \in A_i : \sigma(a_i) > 0\}$ as the support of σ, or $\text{supp}(\sigma)$.

Remark. A profile of mixed strategies $(\sigma_i)_{i \in N}$ induces a probability distribution over A. For a finite A, as will be usually assumed further on, the probability distribution induced by a strategy profile $\sigma \in (\mathcal{D}(A_i))_{i \in N}$ is a function $A \to \mathbb{R}, a \mapsto \prod_{i \in N} \sigma_i(a_i)$.

Definition 2.2.10 (Pure strategy)**.** A probability distribution $\sigma \in \mathcal{D}(A_i)$ with $\sigma(a_i) = 1$ for some $a_i \in A_i$ is called a *pure strategy*, and we sometimes refer to such an a_i itself as a pure strategy.

Definition 2.2.11 (Mixed extension)**.** Given a game $G = (N, A, u)$, we call the tuple $(N, (\mathcal{D}(A_i))_{i \in N}, (U_i)_{i \in N})$ the mixed extension of G if for each $i \in N$, $\mathcal{D}(A_i)$ is the set

of probability distributions over A_i and $U_i : \bigtimes_{j \in N} \mathscr{D}(A_j) \to \mathbb{R}$ assigns to each strategy profile $\sigma = (\sigma_j)_{j \in N}$ the expected value of $u_i(a)$ under the probability distribution induced by σ over A.

Remark. For finite A, U_i then is the function $\sigma \mapsto \sum_{a \in A} \prod_{j \in N} \sigma_j(a_j) u_i(a)$.

Definition 2.2.12 (Mixed strategy Nash equilibrium)**.** A *mixed strategy Nash equilibrium* of a game G is the Nash equilibrium of its mixed extension.

According to this definition, given $(N, (\mathscr{D}(A_i))_{i \in N}, (U_i)_{i \in N})$, a strategy profile σ^* is a Nash equilibrium of this mixed extension *iff*

$$\forall i \in N, \sigma_i \in \mathscr{D}(A_i) : U_i(\sigma_i^*, \sigma_{-i}^*) \geq U_i(\sigma_i, \sigma_{-i}^*).$$

Existence of mixed strategy Nash equilibria in finite games

We would like to delineate the conditions under which a game has at least one Nash equilibrium. It turns out that, while a game does not necessarily have a Nash equilibrium in pure strategies, the existence of a Nash equilibrium in mixed strategies is guaranteed for all finite games:

Proposition 2.2.13 (Nash, [69])**.** *Every finite game has a mixed strategy Nash equilibrium.*

Remark. Although we are concerned with finite games and their mixed extensions, the existence of mixed strategy Nash equilibria can be guaranteed for games with finite N but infinite A under certain compactness conditions on the A_i and continuity conditions on the u_i, [35].

2.2.2 Repeated games

We turn to repeated games, where we consider players playing a game repeatedly, e.g., a repeated Prisoner's Dilemma. The framework of repeated games allows us to analyse the repeated interaction of agents over time, where agents can take into account the previous behaviour of other players. This will be a first step towards analysing adaptive behaviour such as learning and evolution.

The consideration of repeated games has been a fruitful field to understand how repetition affects the behaviour of players. The repeated Prisoner's Dilemma is a very intensely studied game, as under certain conditions cooperation becomes not only a

feasible but a dominant option in contrast to the one-shot PD, where mutual defection is the only Nash equilibrium.

Repeated games can consist of either finite or infinite repetitions of a game. Note that the finite horizon usually implies that the number of repetitions is known in advance. The infinite horizon case also includes the case where the number of repetitions is finite but stochastic. A prominent case is where the number of repetitions follows a geometric distribution. The solution approach, i.e., the Nash equilibrium, differs for the two cases. In the finite horizon case, the repeated game can be solved by backwards induction from solving the last round. In the infinite horizon case, there either is no last round or it is unknown which round is the last round. Therefore, here the solution approach differs.

Thus, a repeated game consists of a game G that is repeated and sometimes called the *stage game*. And players choose their actions in each repetition simultaneously, while they know the actions chosen by the other players for the whole history of the repeated game. It is often defined as an *extensive form game* with perfect information and simultaneous moves. As this will be the only kind of extensive form games of interest to us, we refer the reader to the standard texts for a general definition of extensive form games, e.g., [75]. For reasons of clarity, we define repeated games directly as follows:

Definition 2.2.14. Let $G = (N, A, (\succsim_i))$ be a game. An *infinitely repeated game* of G is a tuple $(N, H, (\succsim_i^*))$, where

i) $H = A^{\mathbb{N}} \cup (\bigcup_{t=0}^{\infty} A^t)$ with $A^0 = \{\emptyset\}$, i.e., H is the set of all finite and infinite sequences of elements of A, i.e., action profiles of the stage game G. We will denote the set of finite sequences by $\mathring{H}$, i.e., $\mathring{H} = H \setminus A^{\mathbb{N}}$.

ii) $(\succsim_i^*)$ is a family of preference relations on $A^{\mathbb{N}}$ such that $\forall i \in N$ and $\forall (a^t) \in A^{\mathbb{N}}, a, a' \in A$ with $a \succsim_i a'$:

$$(a^1, \dots, a^{t-1}, a, a^{t+1}, \dots) \succsim_i^* (a^1, \dots, a^{t-1}, a', a^{t+1}, \dots) \quad (\forall t \geq 1)$$

Three forms for the preference relations in repeated games are of particular interest and have analogues in the theory of Markov decision processes, [80]. Although we present all three for the sake of completeness, only the discounting criterion will play a role in the further analysis. We will again only consider the case where the preference

relations in the stage game are induced by a family of functions $u = (u_i)_{i \in N}$. Thus, we assume that the stage game is given as $G = (N, A, u)$:

i) Discounting criterion. For a player $i \in N$ let $\gamma_i \in (0,1)$ and $(a^t), (b^t) \in A^{\mathbb{N}}$ be histories. Then define the preference relation $\succsim_i^*$ on $A^{\mathbb{N}}$ as

$$(a^t) \succsim_i^* (b^t) :\Leftrightarrow \sum_{t=1}^{\infty} \gamma_i^{t-1}(u_i(a^t) - u_i(b^t)) \geq 0.$$

This sum is well-defined, since G is a finite game with finite A and the sequence $(u_i(a^t) - u_i(b^t))_{t \in \mathbb{N}}$ is therefore bounded. We will further assume that instead of individual discount factors γ_i, all players have the same discount factor γ. If we have $(a^t) \succsim_i^* (b^t)$ for two histories $(a^t), (b^t) \in A^{\mathbb{N}}$, we say that (a^t) is preferred over (b^t) by the γ-discounting criterion. This criterion is a natural criterion in the case where the number of repetitions follows a geometric distribution with parameter γ and we are interested in the expected sum of payoffs, as in the MDP case. It is clear that for two histories differences in earlier repetitions have a larger influence than differences in later repetitions under the discounting criterion. This also corresponds to the setting where a finite but random number of repetitions is played, with the number of repetitions following a geometric distribution, where potential differences in later rounds play a smaller role because they are less probably to occur.

ii) Limit of means criterion. For a player $i \in N$ let $(a^t), (b^t) \in A^{\mathbb{N}}$ be histories. Then define the preference relation $\succsim_i^*$ on $A^{\mathbb{N}}$ as

$$(a^t) \succsim_i^* (b^t) :\Leftrightarrow \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \big(u_i(a^t) - u_i(b^t)\big) \geq 0.$$

Note that for any two histories $a, b \in A^{\mathbb{N}}$, where a is preferred over b by the limit of means criterion, there exists a $\gamma < 1$ such that a is preferred over b by the γ-discounting criterion. In the theory of Markov decision processes this criterion is called the *lim inf average* criterion, [80].

iii) Overtaking criterion. For a player $i \in N$ let $(a^t), (b^t) \in A^{\mathbb{N}}$ be histories. Then define the preference relation $\succsim_i^*$ on $A^{\mathbb{N}}$ as

$$(a^t) \succsim_i^* (b^t) :\Leftrightarrow \liminf_{T \to \infty} \sum_{t=1}^{T} \big(u_i(a^t) - u_i(b^t)\big) \geq 0.$$

This criterion can be used to differentiate between strategies in the case where the series over payoffs has no finite limit and it is sensitive to changes at a finite amount of times, in contrast to the limit of means criterion.

Strategies and equilibria in repeated games

In the context of repeated games, the set of strategies becomes far larger and more complex than for one-shot games. For repeated games, and extensive form games in general, a strategy should determine a player's action in every repetition depending on (potentially) the whole history of previously played action profiles. This motivates the following:

Definition 2.2.15. Given the infinitely repeated game $(N, H, (\succsim_i^*))$ of a stage game $G = (N, A, u)$ and a player $i \in N$, a *strategy for the infinitely repeated game* of G of player i is a function $\sigma_i : \mathring{H} \to A_i$. We call a family of such strategies $\sigma = (\sigma_i)_{i \in N}$ a strategy profile. We denote the set of all strategies of a player i by Π_i.

Given a strategy profile, we want to define how that profile leads to a history of the game:

Definition 2.2.16 (Outcome)**.** Given the infinitely repeated game $(N, H, (\succsim_i^*))$ of a stage game $G = (N, A, u)$ and a strategy profile $(\sigma_i)_{i \in N}$, we define the *outcome* $O(\sigma)$ of σ to be the history $(a^t) \in A^{\mathbb{N}}$ that results when each player $i \in \mathbb{N}$ follows the strategy σ_i, i.e.:

$$\forall t \in \mathbb{N} : a^{t+1} = \sigma((a^\tau)_{\tau \leq t})$$

With these definitions in place, we can extend the definition of a Nash equilibrium to repeated games:

Definition 2.2.17. Let $(N, H, (\succsim_i^*))$ be an infinitely repeated game with possible strategies $(\Pi_i)_{i \in N}$. A strategy profile $(\sigma_i^*)_{i \in N}$ is called a *Nash equilibrium* if

$$\forall i \in N, \sigma_i \in \Pi_i : O(\sigma_i^*, \sigma_{-i}^*) \succsim_i^* O(\sigma_i, \sigma_{-i}^*).$$

In the case of repeated games, we need at least one further refinement of the equilibrium concept, that of a subgame perfect equilibrium, although we will not introduce the notion of a subgame as it is not crucial to the understanding of our specific situation. Let us consider the following motivating example from [75, p. 146].

Example 2.2.18. Let a 2-player game be given by the following payoff table:

	A	D
A	$2,3$	$1,5$
D	$0,1$	$0,1$

For player 2, we have $(A,D) \succ_2 (A,A) \succ_2 (D,A) \sim (D,D)$ in the one-shot game. However, player 1 prefers (A,A) over all other action profiles. Suppose then player 1's strategy is given as

$$\sigma_1^*((a^\tau)_{\tau \leq t}) = \begin{cases} D & \text{if } \exists \tau \leq t : a_2^\tau = D \\ A & \text{otherwise} \end{cases}$$

i.e., if player 2 chooses D even once, then player 1 will play D forever as a "punishment". Let player 2's strategy be given as $\sigma_2^* : (a^\tau)_{\tau \leq t} \mapsto A$, i.e., player 2 always plays A. Then, given that player 1's strategy is σ_1^*, we have that

$$\forall \sigma_2 \in \Pi_2 : O(\sigma_2^*, \sigma_{-2}^*) \gtrsim_2^* O(\sigma_2, \sigma_{-2}^*).$$

Note that any deviation from playing A would lead to action profiles (D,A) and (D,D) and thus a payoff of 1 afterwards, for any strategy $\sigma_2 \in \Pi_2$. This would offset any gains from playing D under the criteria introduced earlier (for a sufficiently high discount factor γ). More obviously, given player 2's strategy σ_2^*, we have that

$$\forall \sigma_1 \in \Pi_1 : O(\sigma_1^*, \sigma_{-1}^*) \gtrsim_1^* O(\sigma_1, \sigma_{-1}^*) .$$

Thus, (σ_1^*, σ_2^*) is a Nash equilibrium strategy profile for the repeated game. However, note that σ_1^* implies that player 1 would receive a payoff of 0 for "punishing", i.e., for playing D. Even if player 2 switches to playing always D, then player 1 would have a strong incentive to play A, i.e., the threat of "punishment" is not credible.

This example motivates the reasoning that players should have no motivation to change their strategy later on in the game, after any previous history, i.e., they would not gain anything from changing their strategy if things turned out differently, which motivates the following definition:

Definition 2.2.19. Given an infinitely repeated game $(N, H, (\gtrsim_i^*))$, a strategy profile σ^* is a *subgame perfect equilibrium* of the game if for every player $i \in N$ and every non-terminal history $h \in \mathring{H}$

$$\forall \sigma_i \in \Pi_i : O_h(\sigma_i^*, \sigma_{-i}^*) \gtrsim_i^* O_h(\sigma_i, \sigma_{-i}^*) ,$$

where for a strategy profile σ and a non-terminal history $h = (h^t)_{t \leq t_h}$, $O_h(\sigma)$ is the terminal history $(a^t)_{t \in \mathbb{N}}$ such that

$$a^t = \begin{cases} h^t & \text{if } t \leq t_h \\ \sigma((a^\tau)_{\tau < t}) & \text{otherwise} \end{cases}.$$

Thus, a subgame perfect equilibrium strategy profile is a Nash equilibrium after *any* history, even and especially if that history would not have occurred under that strategy profile. Considering the Nash equilibrium strategy profile presented in the motivating example 2.2.18, it becomes clear that this is not a subgame perfect equilibrium, as seen after, e.g., the very short history (D, D), which would not have occurred under the given strategy profile.

2.3 Evolutionary game theory and replicator dynamics

Based on the game theoretic concepts introduced so far, we can now consider a very specific approach to analyse and establish equilibria through dynamics on a population of players. Our presentation mainly follows [45].

In evolutionary game theory the question of interest is the presence of traits in a population that is subjected to an evolutionary process where the fitness associated with a trait is determined by a game. One standard setting is a single infinite population of players who encounter each other randomly and play a game in each encounter. The expected payoffs of the game then determine the fitness of that agent and the agents reproduce according to their respective fitness such that traits which lead to a higher fitness proliferate whereas traits with lower fitness disappear from a population.

However, we are only interested in traits which have a bearing on a player's strategy, and instead of a player having certain traits, we will say that a player is of some type and we sometimes speak of strategies instead of types, employing these terms almost synonymously in the present context.

The main idea of evolutionary stability was introduced in [64], further specified in [105], and can be formulated as in [45]:

Let $W(I, Q) \in \mathbb{R}$ be the fitness of a type I in a population with composition Q and let $xJ + (1 - x)I$ denote the population with a proportion x of J-types and $1 - x$ of I-types. We call a population of I-types *evolutionarily stable* if there is $\bar{\varepsilon} > 0$ such that

for all $\varepsilon < \bar{\varepsilon}$ and all other types J ($J \neq I$) we have:

$$W(J, \varepsilon J + (1-\varepsilon)I) < W(I, \varepsilon J + (1-\varepsilon)I)$$

Thus, if we replace a small proportion of the population with some other type J, that type will disappear under an evolutionary process.

In the context of evolutionary game theory, the fitness of a type is determined by a finite game. Therefore, let $(\pi_i)_{1 \leq i \leq \nu}$ be a mixed strategy in a game, i.e., a distribution over the ν pure strategies of the game. Let further u_{ij} denote the payoff of playing the pure strategy i against the pure strategy j, giving the payoff matrix $U = (u_{ij})$. Then the expected payoff of a π-player against a ρ-player is given as:

$$\sum_{1 \leq i,j \leq \nu} \pi_i \rho_j u_{ij} \quad \text{or} \quad \pi^T U \rho.$$

Definition 2.3.1. We call π an *evolutionarily stable strategy* (ESS) if there is $\bar{\varepsilon} > 0$ such that for all strategies ρ ($\rho \neq \pi$) and all $0 < \varepsilon < \bar{\varepsilon}$ we have

$$\rho^T U(\varepsilon\rho + (1-\varepsilon)\pi) < \pi^T U(\varepsilon\rho + (1-\varepsilon)\pi).$$

Remark. This definition implies by continuity for $\varepsilon \to 0$ that a strategy π is an ESS if and only if

i) (π, π) is a Nash equilibrium, i.e. for all strategies ρ

$$\rho^T U \pi \leq \pi^T U \pi$$

ii) and for all strategies ρ ($\rho \neq \pi$)

$$\rho^T U \pi = \pi^T U \pi \Rightarrow \rho^T U \rho < \pi^T U \rho.$$

Note that π is an ESS if it is a strict Nash equilibrium, i.e., with strict inequality for all $\rho \neq \pi$ in i).

The following then characterizes an ESS:

Proposition 2.3.2 ([45, theorem 6.4.1])**.** *A strategy π is an ESS if and only if for all $\rho \neq \pi$ in some neighbourhood of π,*

$$\pi^T U \rho > \rho^T U \rho\,.$$

A dynamic perspective. In order to clarify what we mean, when we say that a type will disappear under an evolutionary dynamics, we need to specify that dynamics for some population. Although the range of possible population dynamics is wide, as [46] shows, we focus on the replicator dynamics, which was introduced in [105] in order to provide a dynamics and relate its stationary points to evolutionarily stable strategies:

Consider a population x of individuals of types $i \in \{1, 2, \dots, n\}$, and corresponding proportions of the population, $x_i \geq 0$ with $\sum_{i=1}^{n} x_i = 1$. Let further $f_i(x)$ be the fitness of a type i in a population x, and $\bar{f}(x) = \sum_{i=1}^{n} x_i f_i(x)$ the average fitness of the population.

In the replicator dynamics, the time evolution of a population x is then given by the following differential equations:

$$\dot{x}_i = x_i (f_i(x) - \bar{f}(x)) \tag{2.3.1}$$

An alternative discrete time replicator dynamics is given by

$$x_i(t+1) = x_i(t) \frac{f_i(x(t)) + c}{\bar{f}(x(t)) + c} \tag{2.3.2}$$

where c can be interpreted as a background fitness ensuring positive numerators as suggested in [46].

Remark 2.3.3. We note that the discrete dynamics (2.3.2) approaches the continuous (2.3.1) as the background fitness tends to infinity. We can rewrite (2.3.2) as:

$$x_i(t + \Delta t) = x_i(t) \frac{w f_i(x(t)) + 1}{w \bar{f}(x(t)) + 1}$$

As $w \to 0$, we approach the weak selection case, where the fitness has essentially no influence on the selection process. Linearizing the dynamics around $w = 0$ yields

$$\frac{x_i(t + \Delta t) - x_i(t)}{w} = x_i(t) \left(f_i(x(t)) - \bar{f}(x(t)) \right) \tag{2.3.3}$$

Setting $w = \Delta t$ and taking the limit $\Delta t \to 0$ then yields the continuous dynamics (2.3.1). Thus the two dynamics coincide in the weak selection limit.

Clarifying the relation between the dynamical perspective and evolutionary stability, we next specify the relationship between the stationary points of the continuous replicator dynamics (2.3.1) and the ESS of a game. Consider again a game with ν pure strategies and a payoff matrix $U = (u_{ij})$, where u_{ij} gives the payoff of playing the pure strategy i against a pure strategy j. Assume a population $(x_i)_{1 \leq i \leq n}$ consisting of n

types, where each type i plays a (possibly mixed) strategy $(\sigma_k^i)_{1\leq k\leq \nu}$. Then the payoff of playing σ^i against σ^j is given as:

$$a_{ij} = \sum_{1\leq k,l\leq \nu} \sigma_k^i u_{kl} \sigma_l^j = (\sigma^i)^T U \sigma^j$$

We define the fitness $f_i(x)$ of a type i as the expected payoff of playing a random opponent in x:

$$f_i(x) = \sum_{1\leq j\leq n} a_{ij} x_j = [Ax]_i$$

In this case, the replicator equations (2.3.1) simplify to

$$\dot{x}_i = x_i([Ax]_i - x^T Ax)\,. \tag{2.3.4}$$

To clarify the relationship between the stationary points of (2.3.4) and evolutionarily stable strategies of the game with payoffs U, we will need the following definitions:

Definition 2.3.4. Let A be as above. Then we call a population composition x^*:

i) a *Nash equilibrium* of the game with the payoff matrix A if $x^T Ax^* \leq (x^*)^T Ax^*$ for all populations x;

ii) an *evolutionarily stable state(!)* for a payoff matrix A if $x^T Ax < (x^*)^T Ax$ for all $x \neq x^*$ in a neighbourhood of x^*;

iii) an *interior point* if $x_i^* > 0$ for all $1 \leq i \leq n$; and an orbit $(x(t))$ an interior orbit if all its points are interior points.

Definition 2.3.5 (Dynamic stability)**.** Given some dynamics, not necessarily (2.3.1) or (2.3.4), we call a stationary point x^* of that dynamics

i) *Lyapunov stable* (or *neutrally stable*) if for every neighbourhood $\mathcal{V}$ of x^*, there is a neighbourhood $\mathcal{U}$ of x^* such that for every $x \in \mathcal{U}$ its forward-orbit $(x(t))_{t\in\mathbb{R}_0^+}$ under under the dynamics is contained in $\mathcal{V}$;

ii) *asymptotically stable* if x^* is Lyapunov stable and there is a neighbourhood $\mathcal{U}$ of x^* such that for every $x \in \mathcal{U}$ its forward-orbit $(x(t))_{t\in\mathbb{R}_0^+}$ converges to x^*; and *globally asymptotically stable* if this is the case for every x in the domain of the dynamics.

Then the following result establishes the relation between these perspectives:

Proposition 2.3.6 ([45, theorems 7.2.1 and 7.2.4])**.** *Let $\tilde{x}$ be a population of n types and A a payoff matrix for* (2.3.4)*.*

i) If x^* is a Nash equilibrium of the game with payoffs A, then x^* is a stationary point of (2.3.4)*.*

ii) If x^* is the limit of an interior orbit $(x(t))_{t\in\mathbb{R}}$ for $t \to \infty$, then x^* is a Nash equilibrium.

iii) If x^* is Lyapunov stable, then x^* is a Nash equilibrium.

iv) If x^* is an evolutionarily stable state for A, then x^* is an asymptotically stable stationary point of (2.3.4)*.*

Note that this result pertains to the relationship between populations and the Nash equilibria of the game with payoff matrix A. However, it will allow us to establish the relationship between the stationary points of (2.3.4) and the ESS of the game with payoff matrix U. To this end, we introduce the following:

Definition 2.3.7. Given the types $1, \dots, n$, corresponding to the (possibly mixed) strategies $\sigma^1, \dots, \sigma^n$ in the underlying game with payoff matrix U, and a population $(x_i)_{1\leq i\leq n}$, we call the strategy $\bar{\sigma} = \sum_{1\leq i\leq n} x_i\sigma^i$ the *mean population strategy*.

We call a strategy $\bar{\sigma}$ *strongly stable*, if it is a mean population strategy and any mean population strategy in a neighbourhood of $\bar{\sigma}$ converges to $\bar{\sigma}$ under (2.3.4). Note that this is not asymptotic stability, but asymptotic stability without requiring neutral stability.

Note that the mapping from populations to mean population strategies is not necessarily injective, and further that if the types in a population are exactly the pure strategies of a game, then the possible mean population strategies cover all (pure and mixed) strategies of the game.

This leads to our main result on the relationship between evolutionarily stable strategies and the single-population replicator equation (2.3.4):

Proposition 2.3.8 ([45, theorem 7.3.2])**.** *A strategy $\bar{\sigma}$ is an ESS of the game with payoff matrix U if and only if it is strongly stable.*

We illustrate the variety of systems that can result from a replicator dynamics with an example given by [17, p. 31]:

Example. Consider a population of players in the pairwise Rock-Paper-Scissors (RPS) game with the general payoff matrix for player 1 being:

$$A = \begin{pmatrix} 0 & a_3 & -b_2 \\ -b_3 & 0 & a_1 \\ a_2 & -b_1 & 0 \end{pmatrix}$$

Let the types correspond to the deterministic strategies of RPS and the population be a 3-tuple, where the components are in the order R, S, P, representing the frequency of the corresponding type. For the parameter choices

$$A_1 = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & 2 \\ 2 & -1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 1 & -2 \\ -2 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},$$

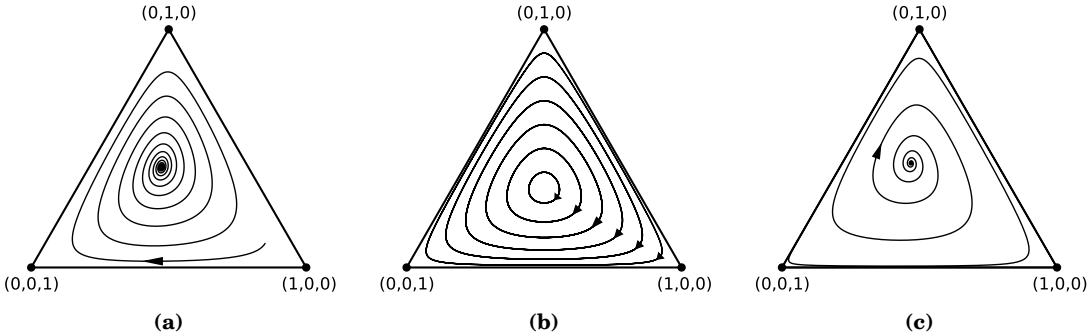
the dynamics produces three different outcomes (figure 2.1):

(a) For A_1, $(5/19, 8/19, 6/19)$ is an asymptotically stable stationary point of the dynamics and is the ω-limit of all interior orbits.

(b) For A_2, a Lyapunov stable stationary point at $(1/3, 1/3, 1/3)$ results. However, it is not asymptotically stable and all interior orbits are closed.

(c) For A_3, an unstable and globally repelling stationary point at $(4/16, 7/16, 5/16)$ results.

Figure 2.1: Three possible situations resulting from the continuous single-population replicator dynamics with different payoffs in the Rock-Paper-Scissors game. (a) An asymptotically stable and globally attracting stationary point results from payoff matrix A_1. (b) For the conventional choice for payoffs A_2, a Lyapunov stable but not asymptotically stable stationary point with closed orbits results. (c) For A_3, an unstable and globally repelling stationary point results.



Note that this formulation of evolutionary game theory considers a population of players where players from the same population interact with each other. As we will

consider several populations, where players from different populations interact and gain fitness from those interactions, we will not employ the simplified replicator equations (2.3.4) but will consider the more general and discrete form (2.3.2) instead, or rather the weak selection version (2.3.3). Therefore in our case, a type's absolute fitness f_i does not depend on the composition of its own population but rather on the composition of the other population(s). The reproductive success of a type will however still depend on the relative fitness with respect to its own population and thus on the population composition.

2.3.1 Evolutionary dynamics and multi-agent reinforcement learning

The relationship between game theory and multi-agent reinforcement learning shows parallels to that between MDPs and single-agent reinforcement learning. We want to give a short overview over those aspects of multi-agent systems that are closely linked to game theory and are concerned with extending the approach of single-agent reinforcement learning to a multi-agent setting, while a comprehensive presentation is given in, e.g., [119]. While the links between discrete-time replicator dynamics and single-agent reinforcement learning dynamics have been investigated on several occasions, as listed, e.g., in [46], the multi-agent setting raises additional questions such as the observability of the other agents' actions and possibly their rewards. Accordingly, multi-agent reinforcement learning presents a wider variety of proposed algorithms, as surveyed, e.g., in [20]. Furthermore, drawing on ideas from evolutionary game theory for the construction of multi-agent reinforcement learning algorithms has been proposed in, e.g., [110], specifically for iterated games in [109], and as a selection-mutation dynamics in [111]. A detailed analysis of the continuous time Q-learning dynamics in two-player-two-actions games is provided in [53]. Finally, [11] gives a comprehensive overview of algorithms motivated by an evolutionary dynamical approach to multi-agent learning. However, we are not aware of a comprehensive analytical treatment of the underlying dynamics for discounted repeated games, where the focus on the stationary distribution of the resulting Markov process in general is not sufficient, nor are we aware of a treatment relating multi-agent reinforcement learning to game theory in a mathematical way comparable to the relation of reinforcement learning to Markov decision processes.

3 An application to learning in networks

In order to better clarify the context in which we consider the parallels between learning and evolution and the idea of employing evolutionary game theory in understanding the learning dynamics in a network of simple interacting agents, we consider a concrete learning dynamics in a basic artificial neural network. To our knowledge, there are no previous proposals connecting the learning dynamics in artificial neural networks to evolutionary game theory in the literature. Of course, the broad area of learning has been addressed in many different ways in game theory as well as in evolutionary game theory. However, assumptions on agents' rationality and their ability to observe others are often incompatible with the perspective of artificial neural networks and the resulting directions pursued and the techniques employed cannot be easily translated.

The presented setting is a preliminary approach to incorporate learning based on game theoretic payoffs in an artificial neural network and it will become clear that such a preliminary approach, although informed by the literature on artificial neural networks, very quickly encounters problems which require a more rigorous treatment. Therefore, besides providing an illustration of the proposed ideas, the main objective here is to underscore the necessity of pursuing a mathematically rigorous treatment.

We start by considering neurons to be simple agents with a limited ability to observe each others' actions and an ability to react to those observed actions. As usual for neural networks, we assume some optimality criterion on the collective behaviour is given. We further assume that individual payoffs depend on the network's collective performance. Our main question (†) here becomes whether agents can learn individual behaviour that is collectively optimal. We formulate this question in the following terms:

– Let N be a finite set of agents, where we assume $N = \{1, 2, \dots, |N|\}$, with agents having non-empty action sets $(A_i)_{i \in N}$, such that the possible states of the system

are given by the set S with

$$S = \times_{i \in N} A_i := \left\{ a : N \to \bigcup_{i \in N} A_i \;\middle|\; \forall i \in N : \, a(i) \in A_i \right\}.$$

– Let $I \subset N$ be a set of "input" agents and $J \subset N$ a set of "output" agents with $I \cap J = \emptyset$. The optimality criterion should consider only these agents' states.

– Let $(O_i)_{i \in N}$ be a family of subsets of N, which we will call the agents' neighbourhoods. Note that $(O_i)_{i \in N}$ induces a digraph with vertices N and edges $\{(j,i) \in N \times N \mid j \in O_i\}$, which we call the *observability graph* or *interaction graph*.

– Let for each $i \in N$, Π_i be the set of the agent's strategies $\pi_i : \times_{j \in O_i} A_j \to A_i$. The agents' computational "simplicity" is expressed by limiting the sets of possible strategies. Here, we only consider Markov strategies.

The question also implies two dynamics: First, the agents form a digraph and their respective actions depend on the agents' actions they can observe. Therefore, we need to consider the time T in which the state of the system evolves, which we assume to be $\mathbb{N}$. Then we can define the sequence of states $(a_i^t)_{i \in N, t \in T}$ given an initial state $(a_i^0)_{i \in N}$ and given strategies $(\pi_i^t)_{i \in N, t \in T}$, with $(\pi_i^t)_{t \in T} \subset \Pi_i$, such that

$$\forall t \in T, i \in N : \, a_i^{t+1} = \pi_i^t((a_j^t)_{j \in O_i}).$$

Given the evolution of the system, we can consider some form of optimality criterion defined on the sequence of states, as the basis for a feedback signal, from which the agents' individual rewards are derived. Here, the question of convergence of the system's states and the agents' rewards arises. Note that if the induced digraph is acyclic and the strategies are stationary, i.e., independent of $t \in T$, and deterministic, then the system's state will converge after a finite amount of time and we can define an optimality criterion as depending on these limit states. If the graph contains cycles, then convergence in general is not guaranteed even with stationary deterministic strategies. However, it is the cyclic case in which we can speak of agents actually *interacting*, as there is no mutual dependence in behaviours without cycles. This motivates us to employ the discounted total reward as a suitable optimality criterion for the agents, as it is equivalent to evaluating the system over a random but finite

amount of time. In this case, we can define an optimality criterion and feedback signal without requiring the convergence of the system's states and we thus can consider interaction graphs with cycles.

Second, our initial question assumes that agents "arrive" at some desired behaviour, which can be accommodated by either of the following: We can either assume that there is some dynamics on the agents' strategies, i.e., functions $G_i : T \to \Pi_i$. This assumes that both dynamics, i.e., the system's evolution and the change in strategies, operate on the same time scale. Or we can assume that the agents' strategies are adapted on a separate time scale, e.g., adaptation happening after the system has been evaluated for some time and before the time is reset for a new series of evaluations. We will consider both possibilities.

In this chapter, we first consider our question in a setting that is informed by simplified artificial neural networks. It illustrates the interaction of the different aspects of our question and the possible arising complexities. We then proceed to the limit case of a single agent in a stationary environment and relax this assumption to an environment that adapts very slowly relative to the agent with the more general results allowing for proper multi-agent settings presented in later chapters.

3.1 A first numerical approach

In an initial attempt to approach the problem, we specify the following algorithm as a variant of our main question (†) and numerically investigate the system's ability to approximate the identity function on $X = (0, 10)^n$ for some $n \in \mathbb{N}$. We first present the concrete algorithm and will then go into the details as to how this algorithm relates to our initial question:

Algorithm 3.1.1

1. Initialization

   a) Choose the set of agents to be $N = \{1, \dots, |N|\}$, the number of input agents, $|I|$, and output agents, $|\!\circ\! J|$, with $|I| = |\!\circ\! J| < \frac{1}{2}|N|$, and set $I = \{1, \dots, |I|\} \subset N$, $P = \{|I| + 1, \dots, |N| - |\!\circ\! J|\} \subset N$, and $J = \{|N| - |\!\circ\! J| + 1, \dots, |N|\} \subset N$.

   b) Set $X = (0, 10)^{|I|}$.

c) Initialize a *random* upper triangular matrix $W^0 \in (0,1)^{|N|\times|N|}$ of the following block form:

$$W^0 = \begin{pmatrix} 0 & A & 0 \\ 0 & B & C \\ 0 & 0 & 0 \end{pmatrix}$$

where $A \in (0,1)^{|I|\times|P|}$, B is a $|P| \times |P|$ upper triangular matrix with zero diagonal and entries in $(0,1)$ above the diagonal, and $C \in (0,1)^{|P|\times|J|}$. W^0 is the adjacency matrix of an *acyclic* weighted digraph describing the interactions between the agents.

d) Choose the maximal number of iterations T_{max} over X and meta-parameters $\alpha, \gamma \in (0,1)$, and a smoothed characteristic function of $(0,\infty)$, σ.

2. Iterate over X, setting $t \leftarrow 1$.

a) Choose a random pair (x^t, x^t) from $X \times X$.

b) Calculate agents' actions and thus the system's state:

i. Given $x^t \in (0,10)^{|I|}$, set the system's state as:

$$a_i^t \leftarrow \begin{cases} x_i^t & \text{if } i \in I \\ 0 & \text{otherwise} \end{cases}$$

ii. Choose agents as enumerated by W^t and calculate their actions, starting with $i \leftarrow |I| + 1$:

A. Let w_i^t denote the i-th column of W^t.

B. Choose a random diagonal $|N| \times |N|$ matrix $\mathcal{H}_i^t$, where we have for the diagonal elements

$$\ln\left([\mathcal{H}_i^t]_{j,j}\right) \sim \mathcal{N}(0,\sigma^2) \quad (j \in N),$$

and calculate i's action a_i^t as

$$a_i^t \leftarrow \langle \mathcal{H}_i^t w_i^t, a^t \rangle .$$

C. If $i < |N|$, set $i \leftarrow i + 1$ and repeat.

c) Calculate agents' rewards:

i. Calculate the system's performance as

$$F^t = \left(\frac{\|(a_i^t)_{i\in J} - (a_i^t)_{i\in I}\|^2}{\|(a_i^t)_{i\in I}\|^2} + 1 \right)^{-1},$$

where we use the natural enumerations of I and J.

ii. Keep track of the "average" performance over time

$$f^t = (1-\delta) f^{t-1} + \delta F^t,$$

where $f_i^0 = 0$.

iii. Calculate the (non-input) agents' rewards as:

$$R_i^t = \frac{\sigma(a_i^t)}{\sum_{k\in N\setminus I} \sigma(a_k^t)} \max\{0, F^t - f^t\} \quad (i \in N \setminus I),$$

where σ is a smoothed characteristic function of $(0,\infty)$.

d) Adapt agents' behaviours:

i. For each active non-input agent i (i.e., $a_i > 0$) update the reward estimate:

$$r_i^t = (1-\gamma) r_i^{t-1} + \gamma R_i^t,$$

where $r_i^0 = 0$.

ii. Set

$$w_i^{t+1} = w_i^t + \alpha_i^t (\mathcal{K}_i^t w_i^t - w_i^t),$$

where

$$\alpha_i^t = \max\left\{ \alpha \frac{R_i^t - r_i^t}{|R_i^t - r_i^t| + 1}, 0 \right\}.$$

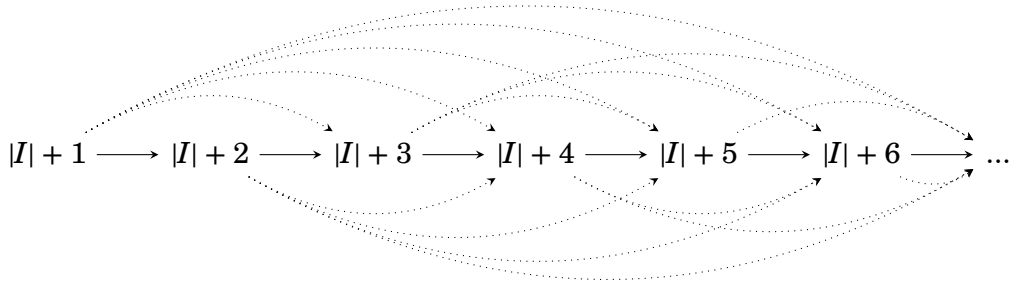
iii. W^{t+1} is then composed of the updated columns w_i^{t+1} and zero-columns for the input agents.

e) If $t < T_{max}$, set $t \leftarrow t+1$ and repeat.

We want to comment on certain parts of the algorithm and clarify how certain aspects relate to our initial general question.

Agents and observability. We interpret the matrix W^0 in the algorithm as the adjacency matrix of a weighted directed graph, where we only consider edges with non-zero weights. It is clear then that the neighbourhoods for the input agents ($i \in I$) are given as $O_i = \emptyset$, i.e., these agents cannot observe any part of the system's state. The output agents in J cannot observe any of the input agents or output agents, but observe all agents in P, i.e., $\forall j \in J : O_j = P$. It is further clear from the definition of W^0 that for all agents $i \in P$ we have $I \subset O_i$ and that there is a linear sorting of the agents in P such that every agent has outgoing edges to every subsequent agent in P, as illustrated in figure 3.1. Indeed, the introduced enumeration of agents is a topological ordering of the whole interaction graph, i.e., for all edges (i,j) we have $i < j$. For the more general case, we may relax the requirement that the digraph be acyclic.

Figure 3.1: Schematic illustration of the induced subgraph of agents in P.



Time. Let us consider the evolution of the system's state. Note that any agent's action only depends on the agents that precede it in the ordering given by W^0. Therefore, an agent's action does not change if its preceding agents' actions do not change and if we used deterministic strategies, e.g., by reusing the same disturbances $\mathcal{K}_i^t$. Thus, in such a deterministic case, agents reach their final action as soon as their preceding agents reach their respective final actions, since the interaction digraph is acyclic and we have a topological sorting of the graph. By computing the agents' actions in the order given by W^0 or equivalently by W^t, cf. algorithm 3.1.1, step 2. b) ii., we arrive at the system's final state after computing each agent's action once.

Note that if agents were evaluated in some arbitrary sequence, all agents would reach their final state after at most $|N|^2$ evaluations. Therefore, if at each point in time all agents were evaluated simultaneously, the system would reach its final state not later than time $|N|$. Therefore in the stochastic case, any variation after time $|N|$

would be due solely to the stochastic disturbances. By computing the agents' actions in the order given by W^0, we effectively ignore the role of the time it would take to reach the system's final state and consider the agents' rewards only for the final state.

The only remaining time scale is the time in which the input agents' states change due to iterations over our input set X, and in which the agents' strategies are adapted. Effectively, we have two nested time scales: On the inner time scale, strategies are constant for input agents and deterministic for all other agents and we compute the system's final state in that time scale, effectively ignoring or "compressing" this time scale. On the outer time scale, input agents have stationary stochastic strategies and all other agents' strategies are adapted stochastically, and this is the time scale that we keep track of in the algorithm.

Agents' strategies. The specific choice of agents' strategies is motivated by the activation function of rectified linear units in deep artificial neural networks, e.g., in [57] and [5]. There, instead of the linear function used here, an affine linear function is used with values below zero replaced by zero, e.g. for some $b_i \in \mathbb{R}$ a unit's activation function would be

$$(a_j)_{j\in O_i} \mapsto \max\left\{0, \left\langle (a_j), w_i \right\rangle + b_i\right\}.$$

For simplicity in this first approach, we ignore the threshold parameter b_i and, as our actions are all non-negative, we do not need the maximum.

Signals and rewards. We investigate the system's ability to approximate the identity function on $(0,10)^{|I|}$. Therefore, F measures the relative distance between input and output state. Furthermore, as the system is sequentially presented with random values from $(0,10)^{|I|}$ a (biased) approximation f of the average value of F is calculated, which is then used to calculate agents' rewards.

Agents' rewards are non-negative and each active agent receives the same reward if its action is larger than some ε, which is the smoothing cutoff of σ, i.e., $\sigma(x) = 1$ for $x > \varepsilon$. Analogous to F, agents have an approximation of their average rewards and agents' policies are only changed if their received reward is higher than this average reward.

Dynamics on strategies. The strategy updates are based on the realised actual reward, R_i^t, and past rewards, r_i^t, as well as on the parameters, w_i^t, and the disturbance

matrix, $\mathcal{H}_i^t$. The strategy dynamics is motivated by the reasoning that the action produced by w_i^t would have been $\langle w_i^t, a^t \rangle$ and was expected to result in a reward of r_i^t, as $\mathbb{E}[\mathcal{H}_i^t] = Id$. However, the random disturbance matrix $\mathcal{H}_i^t$ causes i's action to be a_i^t and results in the possibly different reward, R_i^t. Therefore, if the actual reward is higher than the expected reward, the strategy is adapted such that, given the same observed state, it produces an action that is closer to the actual action a_i^t. Similar approaches employing randomly disturbed strategies are presented in [84].

3.1.1 Numerical results and discussion

We consider the algorithm's behaviour in the following variants:

1. For $|I| = |J| = 1$ and $|P| = 2$, $|P| = 20$, and $|P| = 40$, respectively, we consider the system's ability to approximate the identity on $(0, 10)$.

2. For $|I| = |J| = 2$ and $|P| = 1$, $|P| = 2$, $|P| = 10$, and $|P| = 20$, respectively, we consider the system's ability to approximate the identity on $(0, 10)^2$.

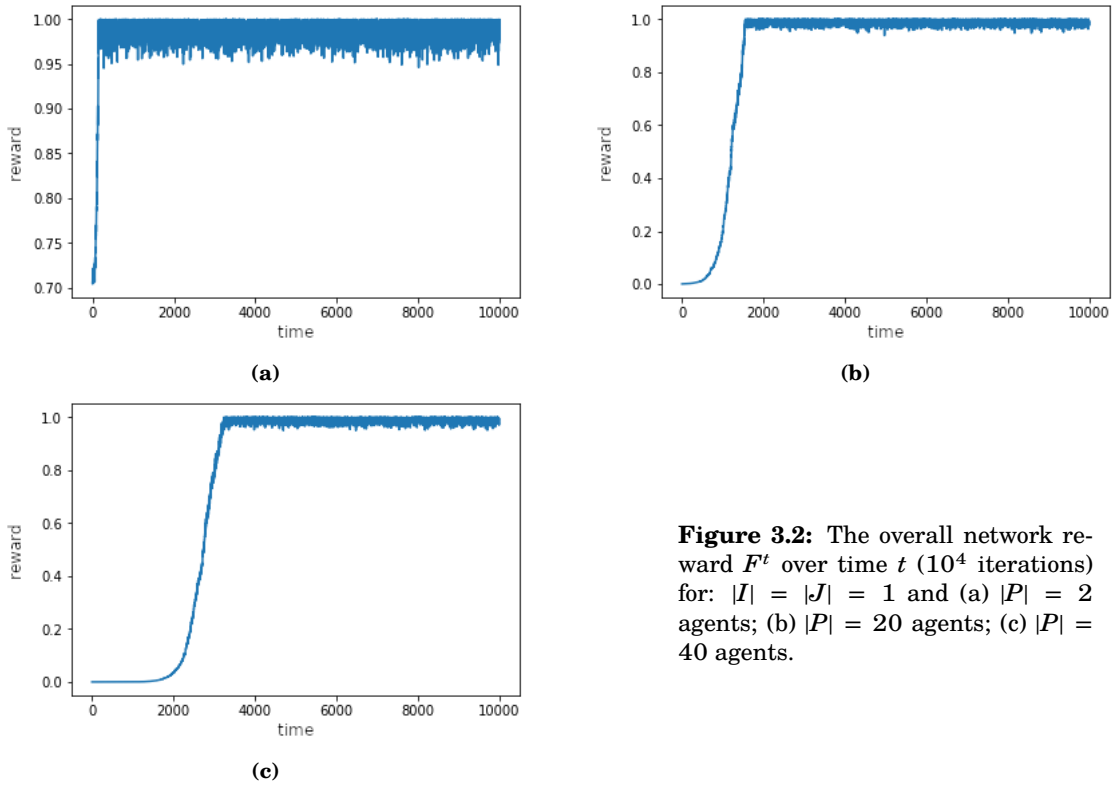
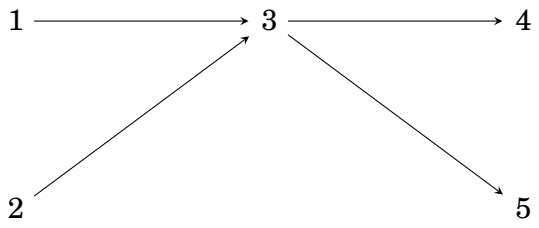


Figure 3.2: The overall network reward F^t over time t (10^4 iterations) for: $|I| = |J| = 1$ and (a) $|P| = 2$ agents; (b) $|P| = 20$ agents; (c) $|P| = 40$ agents.

For the first set of variants, i.e., $|I| = |J| = 1$, figures 3.2a-3.2c, show the values of F^t over the course of 10^4 iterations. In all three cases the system approaches the maximum of F, and the remaining deviation is due only to the stochastic disturbance.

The number of iterations required to reach the limit behaviour increases with the number of agents, as would be expected given the larger search space.

Figure 3.3: Interaction graph in the case of $|I| = |J| = 2$ and $|P| = 1$.



For the second set of variants, i.e., $|I| = |J| = 2$, figures 3.4a-3.4d, show the values of F^t over the course of 10^4 iterations (10^5 in figure 3.4c). The variant with $|P| = 1$ is included as a negative test case, as it is impossible to sufficiently approximate the identity on $(0, 10)^2$ in this case, as is clear from the interaction graph (figure 3.3). As expected, the approximation in this case does not reach a comparable level to the one dimensional case (figure 3.4a).

In the case of $|P| = 2$, approximation is expected to be straight-forward, and the interaction graph is expected to approach one of two possible graphs (figure 3.5). However, as figure 3.4b shows, the approximation does not improve compared to $|P| = 1$, and indeed the interaction graph does not approach any of the two expected variants. For the sake of completeness, figures 3.4c and 3.4d show the system's behaviour for larger numbers of agents. Both cases show a qualitatively similar behaviour to the smaller cases, although considerably more iterations are required to reach that behaviour.

The investigation of the system's behaviour therefore shows that the presented naive approach encounters serious problems in an intuitively simple case, i.e., the approximation of the identity on a two dimensional bounded set.[1] This result thus motivates us to approach the basic question in a more systematic way and to consider two perspectives on the problem:

1) A minimum requirement on a sensible dynamics is that, assuming that all except one agents' policies are held constant, it results in a maximization of the system's performance by optimizing over that one agent's parameter space. Thus, the dynamics should solve the single agent optimization problem if all other

[1] An extension of X to $[0, 10]^2$ produces no significant difference in this respect.

Figure 3.4: The overall network reward F^t over time t (10^4 iterations) for: $|I| = |J| = 2$ and (a) $|P| = 1$ agent; (b) $|P| = 2$ agents; (c) $|P| = 10$ agents (over 10^5 iterations); and (d) $|P| = 20$ agents.

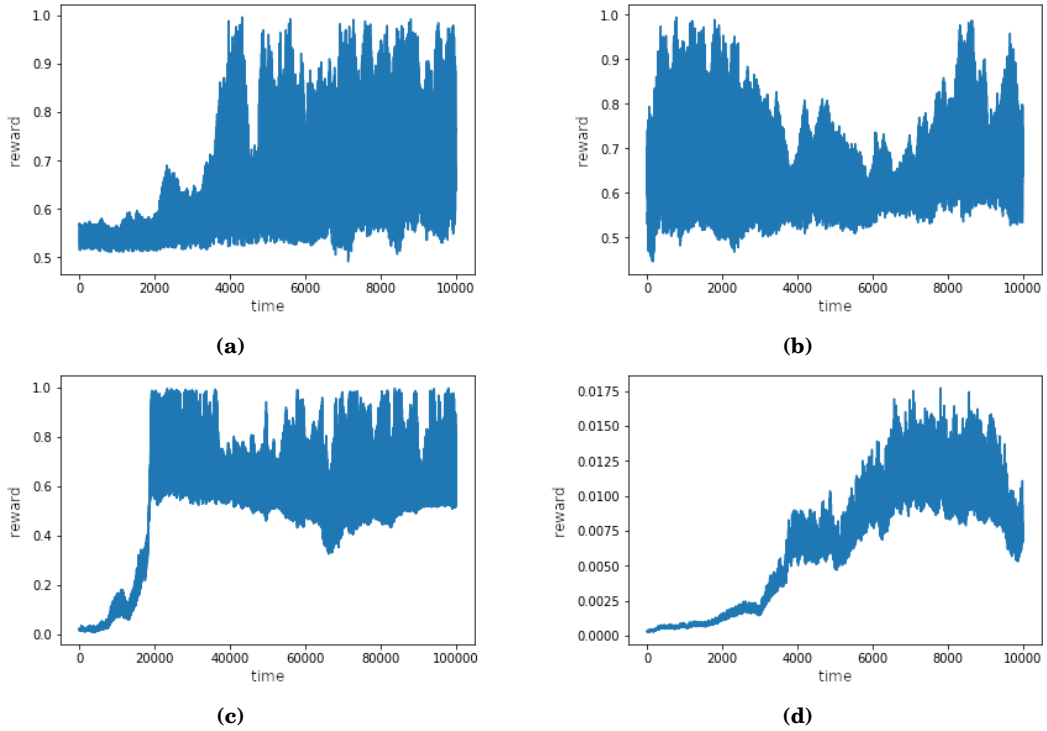
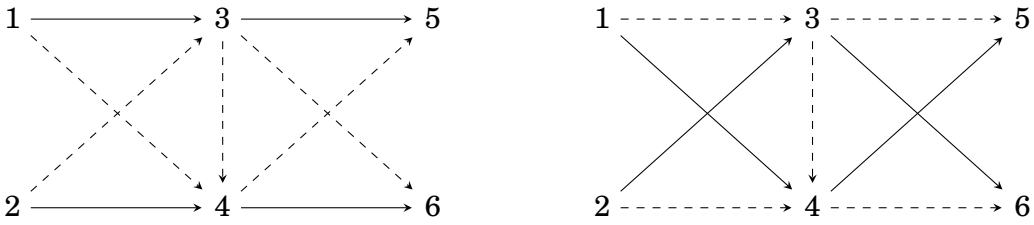


Figure 3.5: Expected interaction graphs for $|I| = |J| = 2$, $|P| = 2$, where dashed edges have weights close to zero.



agents' policies are held constant. This property of a learning algorithm is called *rationality* in [14] in the context of multi-agent learning.

2) The two agent setting, with $N = \{1, 2\}$ and $I = J = N$, with a complete interaction graph should be analysable in the language of standard and evolutionary game theory.

3.2 A single agent in a stationary environment

As a result of the observations stated above, we consider a special case of our initial problem. We assume that the dynamics on all agents except one, i, are constant, i.e., all other agents have stationary strategies. We further assume that we have some

function $F : \times_{j \in N} A_j \to \mathbb{R}$ which measures the desirability of a system's state and that we are interested in maximizing

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t F(a^t)\right] \tag{3.2.1}$$

for some $\gamma \in [0,1)$ where $(a^t)_{t \in \mathbb{N}}$ is the sequence of states generated by the agents' strategies. We further assume that for all sequences of states $(a^t), (b^t) \in (\times_{j \in N} A_j)^{\mathbb{N}}$ the following relation between the reward function R_i and the function F holds:

$$\sum_{t=1}^{\infty} \gamma^t R_i(a^t) > \sum_{t=1}^{\infty} \gamma^t R_i(b^t) \Leftrightarrow \sum_{t=1}^{\infty} \gamma^t F(a^t) > \sum_{t=1}^{\infty} \gamma^t F(b^t)$$

Under this assumption, a dynamics G on the focal agent's strategy maximizes (3.2.1) if it solves the optimization problem of the focal agent i, i.e., if it solves:

$$\max \ \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^t R_i(a^t)\right] \quad \text{s. t. } \pi \in \Pi_i \tag{3.2.2}$$

Under our assumption that for all $j \in N$ we have $|A_j| < \infty$, i.e., that we have finite action sets and therefore a finite state space for the system, and that Π_i contains all stationary Markov strategies, the focal agent's maximization problem is a Markov decision process and a strategy $\hat{\pi}_i$ solves (3.2.2) if and only if it satisfies the optimality conditions for the Markov decision process (2.1.3). Accordingly, a dynamics G maximizes (3.2.1) if and only if its limit is a strategy that satisfies these optimality conditions, such that, e.g., Q-learning would be one such dynamics.

Thus, this situation provides two crucial conditions under which we can hope for a sensible answer to our initial question. First, R should preserve the ordering relation on $\times_{j \in N} A_j$ induced by F and the chosen optimality condition, the discounted total reward in our case. Second, G should converge to strategies that solve the individual agents' Markov decision processes if all other agents are assumed stationary, i.e., G should be rational.

Note that by choosing the discounted total reward as optimality criterion, we do not need the sequence of the system's states to be convergent and we can estimate the expected discounted total reward by estimating total rewards over finite times with times following a geometric distribution, as noted earlier.

An outlook on adaptation on different time scales

In order to approach a dynamics where all agents' strategies are adapted, we relax our assumption about the stationarity of all other agents and instead assume that

our focal agent's strategy adapts fast whereas the other agents' strategies adapt very slowly. In the extreme, this means that the focal agent adapts instantly relative to the adaptation speed of the other agents, while the other agents' strategies have a slow drift. We take a tentative look at the numerical behaviour of such a system with two agents playing the repeated Stag Hunt with the following payoffs:

	a_S	a_H
a_S	$1,1$	$-2,0$
a_H	$0,-2$	$0,0$

with the action set $A = \{a_S, a_H\}$. For simplicity, we consider the case of a complete interaction graph, where all agents can observe all actions. Before we introduce the algorithm, we point out the following properties of the problem: In a repeated two player game with a given payoff vector and a given stationary strategy ρ for player 2, we can specify a Markov decision process for player 1, where the state space S consists of $A \times A$ and the probability to transition from some state $s \in S$ to a state $(a_1, a_2) \in S$ if player 1 chooses action $a \in A$ is induced by ρ as follows:

$$p((a_1,a_2)|s,a) = \begin{cases} \rho(s,a_2) & \text{if } a_1 = a, \\ 0 & \text{otherwise.} \end{cases}$$

We define $\pi^* : \Pi^{MR} \to \Pi^{MR}$ such that

$$[\pi^*(\rho)](s,a) = \begin{cases} \frac{1}{\left|\arg\max_{a' \in A}\{Q^*_\rho(s,a')\}\right|} & \text{if } a \in \arg\max_{a' \in A}\{Q^*_\rho(s,a')\}, \\ 0 & \text{otherwise} \end{cases}$$

where Q^*_ρ is the optimal action-value function for the MDP induced by player 2's strategy ρ. Thus, $\pi^*(\rho)$ is an optimal strategy for the induced MDP. We can then introduce the following algorithm for the repeated two player Stag Hunt game:

In this situation, player 1's strategy adapts faster than player 2's. In the extreme case of $\alpha_\pi \gg \alpha_\rho$, player 1's strategy would immediately reach an optimal strategy π^*. Apart from border cases, in the SH game this means that $\pi^* \in \Pi^{MD}$, i.e., in each state s, player 1 will deterministically chose one of the two available actions a_S or a_H. We could exclude these border cases by defining π^* such that it always returns a stationary deterministic Markov strategy. As we have $|S| = 4$ and $|A| = 2$, this results in $\left|\Pi^{MD}\right| = 2^4$. Table 3.1 shows the optimal strategies for player 1, given deterministic strategies for player 2. Note that player 1 always has a *deterministic*

Algorithm 3.2.1

Let α_π and α_ρ denote the adaptation speed of players 1 and 2 respectively and π and ρ the strategies of players 1 and 2 respectively.

1. Choose $\alpha_\pi, \alpha_\rho \in (0,1)$ with $\alpha_\pi > \alpha_\rho$, and $\pi_0, \rho_0 \in \Pi^{MR}$.
2. For $t \in \mathbb{N}_0$:
   a) Calculate the optimal strategy for player 1: $\pi_t^* = \pi^*(\rho_t)$
   b) $\pi_{t+1} = (1-\alpha_\pi)\pi_t + \alpha_\pi \pi_t^*$
   c) Calculate the optimal strategy for player 2: $\rho_{t+1}^* = \rho^*(\pi_{t+1})$
   d) $\rho_{t+1} = (1-\alpha_\rho)\rho_t + \alpha_\rho \rho_{t+1}^*$

strategy that is optimal and reaches that strategy instantly under $\alpha_\pi \gg \alpha_\rho$. The drift of player 2's strategy is therefore a response to player 1's deterministic optimal strategy, in this case. Due to the symmetry of the game, the table therefore shows the optimal strategy $\pi^*(\rho)$ of *any* player, given that the *other* player's strategy is ρ. We also note that player 1's optimal strategy $\pi^*(\rho)$ is the same if player 2's strategy is close enough to the corresponding deterministic strategy. for illustration, we consider the following:

Table 3.1: Optimal deterministic strategies for player 1, $\pi^*(\rho)$, given the strategy of player 2, ρ. The states are enumerated as given at the top. The strategies are represented by a tuple of probabilities of playing a_S in the respective state.

Player 2's strategy ρ	Player 1's optimal strategy $\pi^*(\rho)$
(a_S,a_S), (a_S,a_H), (a_H,a_S), (a_H,a_H)	(a_S,a_S), (a_S,a_H), (a_H,a_S), (a_H,a_H)
$(1,1,1,1)$	$(1,1,1,1)$
$(1,1,1,0)$	$(1,1,1,1)$
$(1,1,0,1)$	$(1,1,0,1)$
$(1,1,0,0)$	$(1,1,1,1)$
$(1,0,1,1)$	$(1,0,1,1)$
$(1,0,1,0)$	$(1,1,0,1)$
$(1,0,0,1)$	$(1,0,0,1)$
$(1,0,0,0)$	$(1,0,0,0)$
$(0,1,1,1)$	$(0,1,1,1)$
$(0,1,1,0)$	$(0,1,1,0)$
$(0,1,0,1)$	$(0,1,0,1)$
$(0,1,0,0)$	$(0,1,0,0)$
$(0,0,1,1)$	$(0,0,1,1)$
$(0,0,1,0)$	$(0,0,1,0)$
$(0,0,0,1)$	$(0,0,0,1)$
$(0,0,0,0)$	$(0,0,0,0)$

Example. Let us represent a strategy by a tuple in $\mathbb{R}^4$, where each entry represents the probability of playing a_S in the respective state and the states are enumerated

in the order $(a_S, a_S), (a_S, a_H), (a_H, a_S), (a_H, a_H)$, as in table 3.1. Suppose then that player 2's initial strategy ρ_0 is close to $(1, 0, 0, 1)$. Then player 1 will adopt the strategy $\pi^*(\rho_0) = (1, 0, 0, 1)$, independent of its initial strategy π_0. In response, player 2's strategy will slowly drift towards its optimal strategy, $(1, 0, 0, 1)$, as well. If player 1's optimal strategy does not change as long as ρ does not leave a neighbourhood of $(1, 0, 0, 1)$, both players' strategies converge to $(1, 0, 0, 1)$.

Remark. In the case where algorithm 3.2.1 is applied to the repeated Prisoner's Dilemma game, the role of adaptation speeds α_π and α_ρ becomes more prominent, as their relationship affects whether players tend to play the overall beneficial but non-Nash-equilibrium (C, C) or the Nash equilibrium (D, D).

Although numerical results seem to confirm the reasoning presented above, two aspects need to be clarified in order to further investigate the concrete relationships and the concrete dynamics. First, although the adaptation dynamics proposed in algorithm 3.2.1 is comparable to the Policy-Hill-Climbing algorithm investigated in [14], it is not directly comparable to Q-learning or evolutionary dynamics, and does not explicitly account for the stochasticity of the underlying problem. Second, in the intuition presented above, player 2's strategy would be fixed, while in the presented algorithm it is changing slowly. Therefore, an explicit formulation for this dynamics has to be found which allows to analytically investigate and check the asymptotic behaviour suggested by intuition. To address these issues, we consider the implications of employing an evolutionary dynamics.

3.3 A single agent as a population with a replicator dynamics

In order to specify a sound dynamics on the agents' strategies, we investigate the single agent perspective further. As noted, for a single agent in a stationary environment the problem is equivalent to solving an MDP. As proven, there is always a deterministic Markov strategy that solves the MDP. We note that given a distribution $(\pi^\sigma)_{\sigma \in \Pi^{MD}} \in \mathcal{D}(\Pi^{MD})$ over the stationary deterministic Markov strategies, the following defines a stationary randomized Markov strategy $\pi \in \Pi^{MR}$:

$$\pi(s, a) = \sum_{\sigma \in \Pi^{MD}} \pi^\sigma \chi_\sigma(s, a)$$

where χ_σ is the characteristic function on the graph of σ, i.e., $\chi_\sigma(s,a) = 1 \Leftrightarrow \sigma(s) = a$.

We consider $(\pi^\sigma)_{\sigma\in\Pi^{MD}}$ to represent a population of types $\sigma \in \Pi^{MD}$ and the randomized Markov strategy π the resulting mean population strategy corresponding to definition 2.3.7. This perspective allows us to define a dynamics on a stationary randomized Markov strategy $\pi \in \Pi^{MR}$ that corresponds to an evolutionary dynamics on a population of stationary deterministic Markov strategies $(\pi^\sigma) \in \mathscr{D}(\Pi^{MD})$.

Optimality of a population. It is clear from the optimality conditions for MDPs (2.1.3) that a stationary Markov strategy π is optimal *iff* $\forall s \in S, a \in A$:

$$\pi(s,a) > 0 \Rightarrow a \in \arg\max_{a'\in A} \{Q^*(s,a')\}$$

This yields that a population $(\pi^\sigma) \in \mathscr{D}(\Pi^{MD})$ corresponds to an optimal strategy π if and only if

$$\pi^\sigma > 0 \Rightarrow \forall s \in S:\ \sigma(s) \in \arg\max_{a'\in A} \{Q^*(s,a')\}\,.$$

In other words, $\pi^\sigma > 0$ only if σ is itself an optimal deterministic strategy. If follows from proposition 2.1.12 that under the assumption of a stationary environment, i.e., the stationarity of all other players, there exists at least one stationary deterministic Markov strategy $\sigma \in \Pi^{MD}$, that is optimal, and thus that there exists an optimal population composition.

Type fitness. We define the fitness $f^\sigma(s)$ of a type $\sigma \in \Pi^{MD}$ for a state $s \in S$ as

$$f^\sigma(s) = Q^*(s,\sigma(s))\,,$$

i.e., as the optimal action-value for that action $\sigma(s)$ which the deterministic strategy σ chooses in state s. The average population fitness,

$$\bar{f}(s) = \sum_{\sigma\in\Pi^{MD}} \pi^\sigma Q^*(s,\sigma(s)) = \sum_{a\in A} \pi(s,a) Q^*(s,a)\,,$$

then is the value of following the mean population strategy π in the current state s and an optimal strategy thereafter. Note that, as there is always a stationary deterministic strategy that is optimal, for at least one of the types $\sigma \in \Pi^{MD}$ we already have $f^\sigma(s) = v^*(s) \geq \bar{f}(s)$ $(\forall s \in S)$.

We consider how the population changes under the discrete replicator dynamics (2.3.2), i.e.,

$$\pi^\sigma(t+1) = \pi^\sigma(t)\frac{1+wf^\sigma(s_t)}{1+w\bar{f}(s_t)}\,,$$

under weak selection, i.e., for sufficiently small w:

Lemma 3.3.1. *Let π_0 be an interior population, and $(\pi(t))_{t\in\mathbb{N}}$ a realization of the discrete replicator dynamics* (2.3.2) *with sufficiently small w. Then every accumulation point $\bar{\pi} \in \Pi^{MR}$ of $(\pi(t))_{t\in\mathbb{N}}$ in the sense that there is a subsequence $(t_k)_{k\in\mathbb{N}}$ of times such that $(s_{t_k})_{k\in\mathbb{N}}$ visits all states in S infinitely often and for all $a \in A$,*

$$\left(\sum_{\rho\in\Pi^{MD},\, \rho(s_{t_k})=a} \pi^{\rho}(t_k) \right) \to \bar{\pi}(s_{t_k}, a) \text{ as } k \to \infty\,,$$

is a solution to the corresponding MDP almost certainly.

Proof. Let π_0 be an interior point, i.e., $\forall\sigma \in \Pi^{MD} : \pi_0^{\sigma} > 0$, and let there be a non-optimal deterministic strategy for the problem, i.e., there is $\sigma \in \Pi^{MD}$ such that for some state $s \in S$ we have $f^{\sigma}(s) < v^*(s)$. (Otherwise, π_0 is already an optimal population and a rest point of (2.3.2).) Note further that (2.3.2) produces a sequence $(\pi(t))_{t\in\mathbb{N}}$ of interior populations if π_0 is in the interior, and therefore almost certainly a sequence of states $(s_t)_{t\in\mathbb{N}}$ such that every state is recurrent, i.e., is an accumulation point of the sequence. For notational simplicity, we set $f_t^{\sigma} := f^{\sigma}(s_t)$ and $\bar{f}_t := \bar{f}(s_t)$. Then:

$$\bar{f}_t = \sum_{\sigma\in\Pi^{MD}} \pi^{\sigma}(t) f_t^{\sigma} < \sum_{\sigma\in\Pi^{MD}} \pi^{\sigma}(t) v^*(s_t) = v^*(s_t)$$

Let us denote the set of optimal *deterministic* strategies by Π^* ($\neq \emptyset$ as noted above) and let $\sigma^* \in \Pi^*$. Then,

$$f_t^{\sigma^*} = v^*(s_t) > \bar{f}_t$$

and thus for $w > 0$ but small enough to ensure positivity, we have

$$\frac{1 + w f_t^{\sigma^*}}{1 + w\bar{f}_t} > 1\,.$$

By the replicator dynamics, we then have:

$$\pi^{\sigma^*}(t+1) = \pi^{\sigma^*}(t) \frac{1 + w f_t^{\sigma^*}}{1 + w\bar{f}_t} > \pi^{\sigma^*}(t)$$

Thus, as an increasing sequence bounded by 1 from above, $(\pi^{\sigma^*}(t))_{t\in\mathbb{N}}$ is convergent, and hence for $t \to \infty$ we have

$$\frac{1 + w f_t^{\sigma^*}}{1 + w\bar{f}_t} = \frac{\pi^{\sigma^*}(t+1)}{\pi^{\sigma^*}(t)} \to 1 \text{ and so } |f_t^{\sigma^*} - \bar{f}_t| \to 0\,.$$

Let $\rho \notin \Pi^*$ and let s be a state for which $f^{\rho}(s) < v^*(s)$. As every state is revisited infinitely often almost surely, there is a subsequence $(s_{t_k})_{k\in\mathbb{N}} \subset (s_t)_{t\in\mathbb{N}}$ which contains exactly the occurrences of s, i.e., $s_{t_k} = s$ $(\forall k \in \mathbb{N})$.

Then we have

$$\begin{aligned} f_{t_k}^{\sigma^*} - \bar{f}_{t_k} &= f_{t_k}^{\sigma^*} - \sum_{\sigma \in \Pi^{MD}} \pi^{\sigma}(t_k) f_{t_k}^{\sigma} + v^*(s_{t_k}) - v^*(s_{t_k}) \\ &= \underbrace{f_{t_k}^{\sigma^*} - v^*(s_{t_k})}_{=0} - \sum_{\sigma \in \Pi^*} \pi^{\sigma}(t_k) \underbrace{(f_{t_k}^{\sigma} - v^*(s_{t_k}))}_{=0} - \sum_{\sigma \notin \Pi^*} \pi^{\sigma}(t_k) (f_{t_k}^{\sigma} - v^*(s_{t_k})) \\ &= \sum_{\sigma \notin \Pi^*} \pi^{\sigma}(t_k) (v^*(s_{t_k}) - f^{\sigma}(s_{t_k})) > \pi^{\rho}(t_k) \underbrace{(v^*(s) - f^{\rho}(s))}_{>0} \geq 0 \end{aligned}$$

and so $(\pi^{\rho}(t_k))_{k \in \mathbb{N}} \to 0$. Hence for $a \notin \arg\max_{a' \in A} \{Q^*(s, a')\}$, we have that

$$\left(\sum_{\rho \in \Pi^{MD}, \rho(s_{t_k}) = a} \pi^{\rho}(t_k) \right) \to 0 \text{ as } k \to \infty ,$$

and further

$$\sum_{a \notin \arg\max_{a' \in A} \{Q^*(s_t, a')\}} \left(\sum_{\rho \in \Pi^{MD}, \rho(s_t) = a} \pi^{\rho}(t) \right) \to 0 \text{ as } t \to \infty .$$

Therefore, all accumulation points in the sense of the lemma must be solutions to the MDP if $(s_t)_{t \in \mathbb{N}}$ visits all states in S infinitely often, which is almost certainly the case. □

Simulation of the replicator dynamics

We have simulated the discrete replicator dynamics (2.3.2) for a discounted repeated Prisoner's Dilemma with discount factor $\gamma = 0.8$. The state space is given as

$$S = \{(C,C), (C,D), (D,C), (D,D)\}$$

and is enumerated in this order. The state represents the action profile chosen by the players in the preceding round. The payoffs for player 1 are given as $r = (4, 0, 5, 1)^T$. We fix player 2's Markov strategy as presented in table 3.2. It amounts to player 2 playing a stochastic version of a tit-for-tat strategy where the player plays the action chosen by the opponent in the preceding round as discussed in [6].

Table 3.2: Randomized Markov strategy of player 2. The table shows the probability that player 2 will play C or D in a given state.

	C	D
(C,C)	0.7	0.3
(C,D)	0.7	0.3
(D,C)	0.01	0.99
(D,D)	0.01	0.99

The optimal action-value function Q^* for the resulting MDP is given in table 3.3. It shows that player 1 maximizes the expected total reward or equivalently the discounted total reward given player 2's strategy by playing C in each state.

Table 3.3: Q^*-values for player 1 for actions C and D in the corresponding state (given player 2's strategy).

	C	D
(C,C)	11.2	10.2336
(C,D)	11.2	10.2336
(D,C)	8.992	8.0256
(D,D)	8.992	8.0256

We set the uniform distribution over states, $(1/4, 1/4, 1/4, 1/4)$, as the initial state probabilities. In order to incorporate the population perspective, we consider that the set of stationary deterministic Markov strategies Π^{MD} in this case consists of 2^4 strategies. Therefore, we set our population to consist of 16 types corresponding to those strategies, as given in table 3.4. We set the initial population to consist of equal proportions of all types.

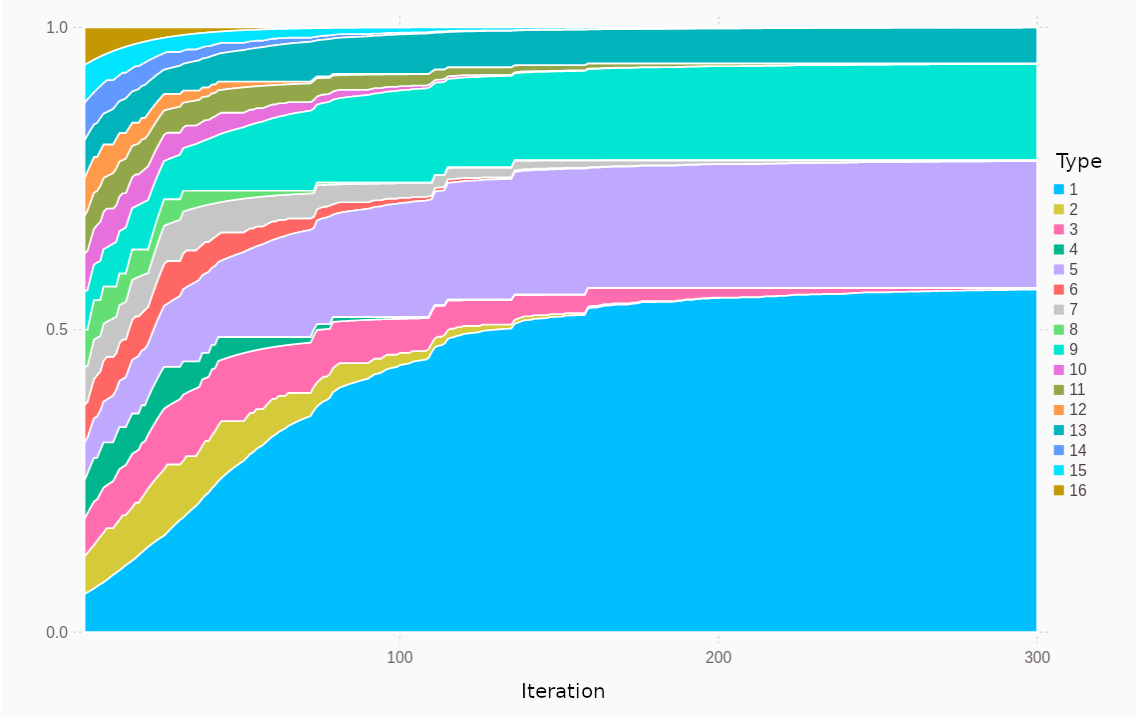
Table 3.4: Types corresponding to deterministic Markov strategies for player 1.

Type No.	(C,C)	(C,D)	(D,C)	(D,D)
1	C	C	C	C
2	D	C	C	C
3	C	D	C	C
4	D	D	C	C
5	C	C	D	C
6	D	C	D	C
7	C	D	D	C
8	D	D	D	C
9	C	C	C	D
10	D	C	C	D
11	C	D	C	D
12	D	D	C	D
13	C	C	D	D
14	D	C	D	D
15	C	D	D	D
16	D	D	D	D

The simulation results are presented in figure 3.6. We see that after 300 generations only four types remain in significant proportions, while all other types' proportions approach 0. Table 3.5 gives the population compositions after 300 and 1000 generations respectively and illustrates that the proportions of the four surviving types remain almost constant. We further see that the four remaining types correspond to

strategies choosing C in the states (C,C) and (C,D). Although the only optimal type is type 1, states (D,C) and (D,D) are not visited often enough to see a diminishing of the proportions of the suboptimal types 5, 9, and 13, after 1000 generations. However from lemma 3.3.1, we know that only strategy 1 is ever played in the limit.

Figure 3.6: Evolution of population composition (abscissa) over time for first 300 generations, where the vertical distance along the ordinate between type boundaries gives the proportion in the population.



3.4 Towards multi-population dynamics

We have so far essentially only considered cases where only a single population or a single player's strategy was changing. In evolutionary game theory, the multi-population case has been considered in detail for certain classes of dynamics, the multi-population replicator dynamics being one of the most prominent, e.g., in [118]. Furthermore, simultaneous learning has been considered in game theory and a prominent negative result is given in [39], which also applies to the multi-population replicator dynamics, among others.

In particular, [118, proposition 5.13] excludes the possibility that an interior[2] equilibrium can be asymptotically stable under the standard multi-population replicator

[2]To be precise, this extends to relatively interior equilibria, i.e., points such that any population consist of more than one type.

Table 3.5: Population compositions after 300 and 1000 generations, respectively, for each type. Rows set in bold face mark the types present after 1000 generations.

Type No.	Strategy	proportion after 300 generations	proportion after 1000 generations
1	**(C, C, C, C)**	**0.566226**	**0.569213**
2	(D,C,C,C)	2.49848e-5	3.21206e-16
3	(C,D,C,C)	0.00296228	5.61113e-8
4	(D,D,C,C)	1.30711e-7	3.16635e-23
5	**(C, C, D, C)**	**0.209617**	**0.210723**
6	(D,C,D,C)	9.24938e-6	1.18911e-16
7	(C,D,D,C)	0.00109664	2.07725e-8
8	(D,D,D,C)	4.83893e-8	1.17219e-23
9	**(C, C, C, D)**	**0.159764**	**0.160607**
10	(D,C,C,D)	7.04958e-6	9.06301e-17
11	(C,D,C,D)	0.000835825	1.58321e-8
12	(D,D,C,D)	3.68808e-8	8.93404e-24
13	**(C, C, D, D)**	**0.0591447**	**0.0594567**
14	(D,C,D,D)	2.60976e-6	3.35514e-17
15	(C,D,D,D)	0.000309423	5.86106e-9
16	(D,D,D,D)	1.36533e-8	3.30739e-24

dynamics, which we will consider in the following chapter. Furthermore, [118, proposition 5.14] gives a negative result regarding constant player-dependent rescalings of the dynamics, proving that such rescalings do not affect the stability of interior equilibria. This is extended by [82] to non-constant rescalings that are non-decreasing in the type fitness. While these results concern the case of multi-population replicator dynamics, [39] relates to so-called uncoupled dynamics, i.e., where players' strategies change only depending on their own strategies, their own payoffs, or other players' strategies, but not depending on other players' payoffs. For every such dynamics there is a game and a neighbourhood of that game such that the dynamics does not converge to the Nash equilibrium for all games in the neighbourhood. This is a basic impossibility result which does not depend on further properties besides uncoupledness, e.g., the dynamics being payoff-increasing.

It is these fundamental challenges regarding stability that we hope to address in the further course of this thesis, in order to gain a better understanding of the conditions for a common perspective on evolutionary and learning dynamics in games and to provide useful insights for multi-agent learning.

4 Replicator dynamics and mutation limits

4.1 Introduction

Evolutionary game theory has contributed significantly to our understanding of a wide range of biological, e.g., [17, 64], and social phenomena, as shown by the vast research into the evolution of cooperation and eusociality, e.g., [6], or the problem of collective action, e.g., [76]. The evolutionary game theoretic approach, formulated in [64], initially assumed a single population with intrapopulation interaction and competition for reproduction, resulting in the concept of the evolutionarily stable strategy (ESS), a refinement of the Nash equilibrium concept, where a strategy is said to be evolutionarily stable if it outperforms any other newcomer strategy in a population consisting almost entirely of players playing the former. While the intuition underlying the notion of an ESS is dynamic, its main definition is usually given in static terms. In an effort to capture the dynamic intuition of the ESS concept, the continuous time replicator dynamics (RD)[1], provided by [105], relates the ESS to certain stationary points, [45], albeit lacking a complete characterization. In its usual formulation, it captures the single population setting with pairwise intrapopulation interactions. However, just as the concept of an ESS has been extended to the multi-population, or multi-species, setting, e.g., [26], so has RD been formulated and analysed in the multi-population setting with intrapopulation competition (for reproduction) but *inter*population interactions (determining reproductive advantage), e.g., [118]. Forms of multi-population RD have been employed in the analysis of coevolutionary systems, such as mutualism [10], antagonistic coevolution of host-parasite systems [70, 99], of institutional ecosystems [40], of the evolution of a population's sex ratio [4], or the coevolution of social behaviour and recognition [98]. It has further been linked to Cross' learning, a simple type of reinforcement learning [13].

[1]The term 'replicator dynamics' and its abbreviation 'RD' are unspecific and can refer to a range of different dynamics. We later define a specific system of equations of the same name and we use '(RD)' (note also the different typeface) when referring to these equations.

In the context of potentially very large systems, e.g., complex ecosystems or multi-agent systems, multi-population RD is of special interest because a population's composition evolves exclusively depending on the payoffs from interactions, but independent of any information about the other populations' payoffs, their compositions, or indeed their very existence. The latter specifics affect a population's composition only through their effect on its payoffs. Borrowing the term from [39], we call this property of RD its *uncoupledness*.

In spite of RD leading to payoff-improving or even equilibrium states in certain cases, there are intuitively simple games, for which neither an ESS exists nor RD reaches any Nash equilibrium, exhibiting periodic limit or general non-convergent behaviour instead: In the usual rock-paper-scissors (RPS) game, RD has exclusively periodic orbits in the single population case and the only Nash equilibrium, an interior point, is not approached from any initial state, e.g., [17], and a range of (un)-stable situations can result [47]. Further, the two population setting results in periodic orbits, as well, and therefore does not reach the interior Nash equilibrium either. An analogue result holds for the matching pennies game, e.g., [118]. Indeed, it has been shown in [39] that no uncoupled dynamics, in particular RD, can be converging to a Nash equilibrium for all possible games. For our understanding of actual biological populations, this periodicity is not necessarily problematic. On the contrary, periodic population dynamics similar to the single-population RPS case have been observed in nature, e.g., in the common side-blotched lizard (*Uta stansburiana*) [95]. For our understanding of the conditions of behavioural convergence in multi-agent systems and their ability to solve large-scale problems such periodic behaviour is less desirable.

Although RD is intended to capture the idea of evolutionary selection, and thus is inspired by evolution, it treats mutation, an arguably central process of evolution and one of the main generators of the diversity on which selection operates, as an extremely rare event, to the degree that it is actually absent from the formulation of the dynamics, especially in the case of multiple populations, e.g., [118]. Approaches which include mutation mainly focus on the single population case [2, 12, 15, 18, 44, 49, 54, 77], consider a payoff-adjusted RD, or a discrete time process [19], or a single discrete population [48, 113], while we are not aware of an analysis of continuous-time multi-population RD with mutation, apart from [81] where certain approximations to multi-population RD are considered, with a different focus however and not linked to mutation.

We demonstrate that introducing a class of mutations, which allows simple non-uniform mutations, in multi-population RD can fundamentally change the properties of the dynamics, i.e., preclude any periodicity in certain cases and, furthermore, guarantee convergence to states close to Nash equilibria, which would not be reachable under standard RD. Note that the non-existence result in [39] does not directly apply to such mutation dynamics, as it only considers Nash-convergence.

Our main interest, therefore, lies with the derivation of an uncoupled dynamics, which, on the one hand, explicitly considers mutation and, on the other hand, is as close as possible to standard RD, and with the analysis of how this mutation mechanism affects the position and stability of equilibria compared to the standard (multi-population) RD. The resulting mutation mechanism with spontaneous mutations from one type to another is of course not appropriate for all biological mutation processes In a biological population, such spontaneous mutation between a finite number of types occurs, e.g., for single nucleotide polymorphisms, where alleles differ by only one nucleotide, with the number of possible single nucleotide polymorphisms at that position restricted to four. Furthermore, such point mutations are known to occur with a non-negligible probability [23, 28] and can be significant factors in diseases, [28, 71], e.g., sickle cell anaemia, [25, 61], which also interacts with malaria parasites, [60], cystic fibrosis, [36], or β thalassemia, [21, 93], and further in human cancer cells, [29, 67]. There is further evidence that in *Drosophila* most such non-synonymous point mutations are deleterious, while the rest are slightly deleterious, near-neutral, or weakly beneficial, [91], suggesting that a weak selection assumption as we employ can be reasonable for persisting polymorphisms. Considered as a learning dynamics, modifications of multi-population RD have been shown to be linked to so-called Q-learning, a more sophisticated reinforcement learning algorithm, [111]. In particular, the resulting modification can be interpreted as a mutation-like term.

The inclusion of mutation should not only further our understanding of coevolutionary multi-population systems, such as ecosystems. Its ability in certain cases to stabilise equilibria for any non-zero mutation rate, and thereby make them approachable under an uncoupled dynamics, should also be useful in the study of game theoretical solution concepts, such as ε-Nash equilibria, [33], and the formulation of conditions for the convergence of learning in multi-agent systems.

We proceed by introducing the standard multi-population RD, i.e., without mutation, and recounting some stability properties of its equilibria and their relation to

game theoretic concepts, such as Nash equilibria and evolutionary stability.

We then introduce mutation, specifically a simple class of mutations one might call memoryless, which includes non-uniform mutations, and give a heuristic derivation of the specific form of mutation we consider, defining a replicator-mutator dynamics (RMD), the equilibria of which we call *mutation equilibria*. For fixed mutation parameters, we prove the existence of equilibria of RMD, their ε-Nash property, and their uniqueness and asymptotic stability under very high mutation.

We proceed by defining the concept of limits of mutation equilibria for vanishing mutation, which we call *mutation limits*. Mutation limits and their properties are independent of any choice of specific mutation parameters. We prove the existence of mutation limits for all systems with continuously differentiable fitness functions and give a sufficient condition for a Nash equilibrium to be a mutation limit.

In order to address the question of reachability of mutation limits, we define the notion of an *attracting* mutation limit based on the asymptotic stability of the mutation equilibria by which it is approximated. Such attracting mutation limits are reachable in the sense that for any choice of mutation parameters there is an asymptotically stable mutation equilibrium arbitrarily close to the mutation limit.

We further provide a sufficient condition for a Nash equilibrium to be an attracting mutation limit. In particular, all evolutionarily stable states are attracting mutation limits, but not all attracting mutation limits are evolutionarily stable, showing the notion to be a strictly weaker property than evolutionary stability. We conclude by giving a necessary condition for attracting mutation limits, ruling out hyperbolic interior equilibria.

4.2 Multi-population replicator dynamics

In the following we consider the situation where we have a finite set of populations $I = \{1, 2, \dots, N\}$ and each population i consists of a finite number of types which we enumerate and denote by $S_i = \{1, 2, \dots, n_i\}$. Note that types are population-specific and numbers do not identify types across populations. The composition of a population i is then given as a vector x_i such that $x_{ih} \geq 0$ gives the frequency of a type $h \in S_i$ in population i. Thus, the set of possible compositions of population i is given as:

$$\Delta_i = \left\{ x_i \in \mathbb{R}^{n_i}_{\geq 0} \middle| \sum_{h \leq n_i} x_{ih} = 1 \right\}$$

For convenience, we denote the Cartesian product of the Δ_i $(i = 1 \dots N)$ by Δ, i.e., $\Delta = \times_{i \leq N} \Delta_i$, and denote by Δ^o the interior of Δ, i.e., $\forall i \leq N, h \leq n_i : x_{ih} > 0$. Furthermore, we set $S = \{(i,h) | i \in I \text{ and } h \in S_i\}$, such that $\Delta \subset \mathbb{R}^S$, where $\mathbb{R}^S$ denotes the set of tuples of reals indexed by S. The state of the multi-population model then is a description of the frequencies of the different types in the populations, i.e., it is given by some $x \in \Delta$.

We assume that for each population $i \in I$ and each type in that population $h \in S_i$ we have a function $f_{ih} \in \mathscr{C}^1(U, \mathbb{R})$, for $U \supset \Delta$ open, describing the reproductive rate or fitness $f_{ih}(x)$ of that type in a given state $x \in \Delta$ and we define population i's average fitness as $\bar{f}_i(x) = \sum_{h \leq n_i} x_{ih} f_{ih}(x)$. It should be noted that fitness is frequency-dependent in replicator dynamics models and not affected by population sizes. We further assume that there is no intraspecific interaction affecting fitness in a type-specific manner, i.e., the fitness values of types in population i are independent of the composition of population i or $\frac{\partial}{\partial x_{ik}} f_{ih}(x) = 0$ $(i \in I, h, k \in S_i)$ in keeping with the classic normal-form game settings.[2] The standard multi-population replicator dynamics, based on [104] and developed later, e.g., [118], is given by the following system of differential equations:

$$\dot{x}_{ih} = \varphi_{ih}(x) := x_{ih} \left(f_{ih}(x) - \bar{f}_i(x) \right) \quad (i \in I, h \in S_i) \tag{RD}$$

We denote by $\Phi : \mathbb{R} \times \Delta \to \Delta$ the flow of (RD), i.e., for $x \in \Delta$, $\Phi(\cdot, x) : \mathbb{R} \to \Delta, t \mapsto \Phi(t,x)$ is a solution of (RD) with $\Phi(0,x) = x$. Due to our continuity assumption on f, the existence and uniqueness of Φ is clear, e.g., [107, theorem 6.1].

4.2.1 Stationary points of the replicator dynamics

We give a short recount of some well-known properties of (RD) with regards to game theory, beginning with the main concept of game theory:

Definition 4.2.1 (Nash equilibrium)**.** We call a state $x^* \in \Delta$ a *Nash equilibrium* if

$$\forall i \in I, z_i \in \Delta_i \backslash \{x_i^*\} : \bar{f}_i(x^*) \geq \bar{f}_i(x^*_{-i}, z_i),$$

where (x^*_{-i}, z_i) denotes the state such that

$$[x^*_{-i}, z_i]_{jk} = \begin{cases} z_{ik} & \text{if } j = i, \\ x^*_{jk} & \text{otherwise} \end{cases}.$$

[2] Note that this assumption is not essential for all results.

We call $x^* \in \Delta$ a *strict Nash equilibrium* if all inequalities in the Nash equilibrium condition are strict.

Remark. It is clear that $x^* \in \Delta$ is a Nash equilibrium if and only if

$$\forall i \in I, h \leq n_i : g_{ih}(x^*) := f_{ih}(x^*) - \bar{f}_i(x^*) \leq 0.$$

Note that $g_{ih}(x)$ is exactly the coefficient of x_{ih} in (RD). Therefore, we can denote the set of Nash equilibria by $\mathscr{E} = \{x \in \Delta \,|\, g(x) \leq 0\}$, where the inequality is componentwise. A strict Nash equilibrium $x^* \in \Delta$ in particular is a state where each population consists of exactly one type, i.e., for each population $i \in I$ there is exactly one type h_i such that $x^*_{ih_i} = 1$.

The following results on Nash equilibria and stationary points of (RD) are straightforward and well-known, e.g., [118, p. 173]:

Proposition 4.2.2. *If $x \in \Delta$ is a Nash equilibrium, then x is a stationary point of* (RD)*, i.e., $\varphi(x) = 0$.*

Proposition 4.2.3. *If $x \in \Delta^0$ is a stationary point of* (RD)*, then x is a Nash equilibrium.*

Stability properties of equilibria

Our special interest lies with the attainability of Nash equilibria. Therefore, we restate a few stability properties of Nash equilibria and stationary points of (RD) respectively.

Definition 4.2.4. We call a stationary point $x \in \Delta$ *stable*, if for every neighbourhood U of x there is a neighbourhood $V \subset U$ such that $\Phi(\mathbb{R}_{\geq 0}, V) \subset U$. We further call a stationary point $x \in \Delta$ *asymptotically stable* if x is stable and there is a neighbourhood V of x such that for all $y \in V$ we have $\Phi(t, y) \to x$ for $t \to \infty$.

For stable stationary points we have the following:

Proposition 4.2.5. *If $x \in \Delta$ is a stable stationary point of* (RD)*, then x is a Nash equilibrium.*

A proof of this statement can be found in [118, theorem 5.2]. Note that this further characterization is interesting if $x \in \partial\Delta$, as stationary points on the boundary of Δ are

not necessarily Nash equilibria. Furthermore, it implies that stationary points that are not Nash equilibria must be unstable and thus are harder to attain under (RD). However, note that Nash equilibria do not have to be stable. We have the following stronger characterization of asymptotically stable stationary points (with a proof in, e.g., [118, proposition 5.13]):

Proposition 4.2.6. *A stationary point $x \in \Delta$ is asymptotically stable under* (RD) *if and only if x is a strict Nash equilibrium.*

For completeness, we would like to mention the relationship between stationary points of (RD) and evolutionarily stable states, where we define evolutionary stability as in [118, p. 166], equivalently to [26], as follows:

Definition 4.2.7 (Evolutionary Stability)**.** We call a state $x^* \in \Delta$ *evolutionarily stable* if for all $y \in \Delta$ ($y \neq x^*$) there is some $\bar{\varepsilon}_y > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_y)$ and $w = \varepsilon y + (1-\varepsilon)x^*$ we have some $i \in I$ with $\bar{f}_i(x_i, w_{-i}) > \bar{f}_i(y_i, w_{-i})$.

It is well known that in the multi-population case the concept of evolutionary stability is equivalent to that of a strict Nash equilibrium, e.g., [118, proposition 5.1]:

Proposition 4.2.8. *$x \in \Delta$ is evolutionarily stable if and only if x is a strict Nash equilibrium.*

Therefore, we have that strict Nash equilibria are exactly the evolutionarily stable states and exactly the asymptotically stable stationary points of (RD). The dynamics (RD) will therefore not have any asymptotically stable points if the underlying game does not have any strict Nash equilibria. Furthermore, no mixed Nash equilibrium can be asymptotically stable, such that there is no guarantee that any Nash equilibrium will be approached under (RD) if the game has only mixed Nash equilibria.

4.3 Introducing mutation

We consider the effect of mutation for two reasons. First, the idea of evolution is intricately linked with mutation and mutation does not seem to be an extraordinary event but is to be expected. Second, a central idea in the proof that the dynamics (RD) has no interior asymptotically stable states relies on the fact that (RD) is divergence free (after suitable modification) and therefore volume preserving, [45]. However, some games, such as the matching pennies game and the standard rock-paper-scissors

game, have only interior equilibria, while describing biologically relevant interspecies interactions such as host-parasite systems. The kind of mutation we consider results quite clearly in a dynamics with negative divergence. Of course, this does not guarantee asymptotically stable interior equilibria, but it opens up the possibility of such equilibria.

We will first give a motivational heuristic derivation of our specific replicator-mutator dynamics from a more general form. Afterwards, we will consider the properties of our specific dynamics and of its equilibria.

4.3.1 Replicator-mutator dynamics

General mutation

In the standard replicator dynamics (RD), we assume that the offspring of individuals of some type inherit that same type. In contrast, we consider mutation as a process by which the offspring of a certain individual changes into another type (of the same population) with some probability. More precisely, we assume that the offspring of an h-type in population i mutates to a k-type in the same population with some probability $\mu_{ikh} > 0$, with $\sum_{k \leq n_i} \mu_{ikh} = 1$ for all populations i, and therefore:

$$\mu_{ihh} = 1 - \sum_{k \neq h} \mu_{ikh}$$

In order to represent overall mutation more clearly, we introduce *relative mutation probabilities* c_{ikh} and an overall mutation rate μ_i such that $\mu_{ikh} = \mu_i c_{ikh}$ $(h \neq k)$ and thus:

$$\mu_{ihh} = 1 - \mu_i \sum_{k \neq h} c_{ikh}$$

Here, μ_i controls the overall strength of mutation, such that for $\mu_i = 0$ there is no mutation at all, without affecting relative probabilities. We derive our specific dynamics from the general multi-population replicator-mutator dynamics as given in, e.g., [77],

$$\dot{x}_{ih} = \sum_{k \leq n_i} \mu_{ihk} x_{ik} f_{ik}(x) - x_{ih} \bar{f}_i(x) \tag{4.3.1}$$

yielding after substitution:

$$\dot{x}_{ih} = x_{ih}(f_{ih}(x) - \bar{f}_i(x)) + \mu_i \sum_{k \leq n_i} (c_{ihk} x_{ik} f_{ik}(x) - c_{ikh} x_{ih} f_{ih}(x)) \tag{4.3.2}$$

This formulation emphasizes the similarity to the standard replicator dynamics (RD) and how μ_i determines the extent to which (4.3.1) deviates from (RD).

Weak selection-weak mutation limit

Recall that (RD) is invariant under the addition of a background fitness for all types of a population, a property which (4.3.1) does not have. We therefore derive a version which is invariant under the addition of a constant background fitness. For convenience, let ${s_i}^{-1}$ denote some background fitness, where s_i can be seen as representing the selection pressure on that particular trait. Formulating (4.3.1) with a modified fitness function $\tilde{f}_{ih} : x \mapsto f_{ih}(x) + {s_i}^{-1}$ and suitable substitution yields a dynamics with explicit background fitness:

$$\dot{x}_{ih} = \varphi_{ih}(x) + \frac{\mu_i}{s_i} \sum_{k \leq n_i} \left(s_i \left(c_{ihk} x_{ik} f_{ik}(x) - c_{ikh} x_{ih} f_{ih}(x) \right) + c_{ihk} x_{ik} - c_{ikh} x_{ih} \right)$$

Analogous to [45], we consider a weak selection-weak mutation limit, where the background fitness tends to infinity, i.e., the selection pressure goes to zero $s_i \to 0$, and mutation occurs on the same order as selection, i.e., $\mu_i \to 0$, such that overall:

$$\frac{\mu_i}{s_i} \to M_i > 0$$

This yields the following weak selection-weak mutation limit of (4.3.1), which is invariant under addition of background fitness,

$$\dot{x}_{ih} = x_{ih} \left(f_{ih}(x) - \bar{f}_i(x) \right) + M_i \sum_{k \leq n_i} \left(c_{ihk} x_{ik} - c_{ikh} x_{ih} \right) \tag{4.3.3}$$

where we refer to M_i as the *mutation rate* in population i. Note that (4.3.3) can also be derived from a discrete selection-mutation equation, [45]. Additionally, we assume that mutation is memoryless, i.e., $c_{ihk} = c_{ihl}$ $(k, l \neq h)$, akin to Kingman's house-of-cards model [54]. Note that this explicitly allows non-uniform mutation to occur. Then we can write c_{ih} instead of c_{ihk} and assuming that the mutation rate is the same for every population, replacing M_i with M, this yields the following: [3]

Replicator-Mutator Dynamics

For some fixed $c \in \Delta^o$ and $M \geq 0$, the replicator-mutator dynamics (RMD) is given by:

$$\dot{x}_{ih} = \varphi_{ih}^M(x) := x_{ih} \left(f_{ih}(x) - \bar{f}_i(x) \right) + M \left(c_{ih} - x_{ih} \right) \tag{RMD}$$

[3] Note that we can choose M_i such that $\sum_{h \leq n_i} c_{ih} = 1$ holds. Although we consider M as independent of the population, population-dependent mutation parameters M_i are mostly compatible with the present arguments, but would render proofs overly technical.

It is clear that we obtain (RD) for $M = 0$. We denote by $\Phi^M : \mathbb{R} \times \Delta \to \Delta$ the flow of (RMD), i.e., for $x \in \Delta$, $\Phi^M(\cdot, x) : \mathbb{R} \to \Delta, t \mapsto \Phi^M(t,x)$ is a solution of (RMD) with $\Phi^M(0,x) = x$.

Remark. Note that Φ^M also depends on our choice of c. Throughout this section, we will consider some arbitrary but *fixed* $c \in \Delta^o$ and the defined concepts will depend on that choice. However, we will proceed to properties of (RMD) which are invariant under the choice of c later on.

Definition 4.3.1. We call $x \in \Delta$ with $\varphi^M(x) = (\varphi^M_{ih}(x))_{(i,h)\in S} = 0$ a *mutation equilibrium* for M. For shortness, we call x^M a mutation equilibrium if it is a mutation equilibrium for M.

Definition 4.3.2. We call a sequence $(x_n)_{n\in\mathbb{N}} \subset \Delta$ a sequence of mutation equilibria if there is a sequence $(M_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$ with

i) $M_n \to 0$ for $n \to \infty$

ii) and x_n is a mutation equilibrium for M_n, i.e., $\varphi^{M_n}(x_n) = 0$, for all $n \in \mathbb{N}$.

For ease of notation, we write such a sequence as $(x^M)_{M>0}$.

Under suitable assumptions, such sequences represent the change of a coevolutionary system under decreasing mutation rates, and we will be especially interested in the limits of such sequences of mutation equilibria and in their properties.

4.3.2 Existence of stationary points with mutation

Lemma 4.3.3. *For all $M > 0$ and $c \in \Delta^o$ there is $x \in \Delta^o$, such that x is a stationary point of the replicator-mutator dynamics* (RMD)*, i.e., $\varphi^M(x) = 0$.*

Proof. Note that the vector field φ^M points towards the interior of Δ for all $x \in \partial\Delta$. We thus have that for all $x \in \partial\Delta$ and all $t > 0$, $\Phi^M(t,x) \in \Delta^o$, and thus Δ is forward-invariant under the flow Φ^M, in particular, $\Phi^M(\mathbb{R}_{>0}, \Delta) \subset \Delta^o$. Furthermore, it is clear that Δ is nonempty, convex and compact. Using Brouwer's fixed point theorem, we can now use that if a nonempty, convex compact set is forward-invariant under a flow, then it contains a fixed point, e.g., [107, lemma 6.8]. With $\Phi^M(\mathbb{R}_{>0}, \Delta) \subset \Delta^o$, we have that the fixed point has to be in Δ^o. □

The following definition, e.g., as given by [33], will be useful in our later investigation:

Definition 4.3.4 (ε-Equilibrium)**.** For some $\varepsilon > 0$, we call a state $x^\varepsilon \in \Delta$ an *ε-equilibrium* if

$$\forall i \in I, h \leq n_i : f_{ih}(x^\varepsilon) - \bar{f}_i(x^\varepsilon) \leq \varepsilon\,.$$

In relation to ε-equilibria we state the following property:

Lemma 4.3.5. *Let x^M be a mutation equilibrium, then x^M is an ε-equilibrium of the underlying game for $\varepsilon = M$, and in particular $\forall i \in I, h \leq n_i : f_{ih}(x^M) - \bar{f}_i(x^M) < M$.*

Proof. For $(i,h) \in S$, we have that

$$\begin{aligned} 0 = \varphi^M_{ih}(x^M) &= x^M_{ih}(f_{ih}(x^M) - \bar{f}_i(x^M)) + M(c_{ih} - x^M_{ih}) \\ &> x^M_{ih}(f_{ih}(x^M) - \bar{f}_i(x^M)) - Mx^M_{ih} \end{aligned}$$

and thus, with $x^M \in \Delta^o$, we have $f_{ih}(x^M) - \bar{f}_i(x^M) < M$. □

Together with the continuity of f, we have the following:

Corollary 4.3.6. *Let $(x^M)_{M>0}$ be a sequence of mutation equilibria and x^* an accumulation point for $M \to 0$. Then x^* is a Nash equilibrium.*

4.3.3 Mutation equilibria for high mutation rates

We consider some specific properties under high mutation rates which illustrate the effect of mutation on the number and stability of equilibria through its effect on the Jacobian of the replicator dynamics. Note that all equilibria of (RMD), irrespective of the specific choice of $M > 0$, lie in the interior of Δ and that φ^M points inward on $\partial\Delta$. We can therefore consider (RMD) as a dynamics on Δ^o. We can further, for all populations i, replace x_{in_i} with $(1 - \sum_{k<n_i} x_{ik})$, and thus proceed to the resulting reduced system $\tilde{\varphi}^M$ (with an analogous procedure to obtain $\tilde{\varphi}$ from φ), which is then defined on the Cartesian product of the $(n_i - 1)$-simplices. For ease of notation, we will still use Δ to denote this reduced space. Thus, questions regarding the stability of a mutation equilibrium $x^M \in \Delta^o$ can be treated by considering the eigenvalues of the Jacobian $D\tilde{\varphi}^M$. In particular, due to the Hartman-Grobman theorem, e.g., [78, 107], we have the following useful characterization:

Remark 4.3.7. Let x^M be a *hyperbolic* equilibrium of (RMD), and of the reduced system $\tilde{\varphi}^M$ equivalently, i.e., all eigenvalues of $D\tilde{\varphi}^M(x^M)$ have non-zero real part. Then x^M is asymptotically stable if and only if all eigenvalues of $D\tilde{\varphi}^M(x^M)$ have negative real part, e.g., [107, theorem 6.10]. In particular, all eigenvalues of $D\tilde{\varphi}^M(x^M)$ have negative real part, if and only if all eigenvalues of $D\tilde{\varphi}(x^M)$ have real part smaller than M, due to $D\tilde{\varphi}^M = D\tilde{\varphi} - M \cdot I$, where I is the identity matrix.

With this observation, we obtain the following:

Lemma 4.3.8. *There is $\underline{M} \geq 0$ such that for all $M > \underline{M}$ the stationary points of the replicator-mutator dynamics* (RMD) *are asymptotically stable. In particular, $D\tilde{\varphi}^M$ is invertible everywhere on Δ.*

Proof. Note that all eigenvalues of $D\tilde{\varphi}$ are bounded on Δ, in particular the real parts of the eigenvalues are bounded, as well. Then let $\underline{M}$ be an upper bound on all real parts of the eigenvalues of $D\tilde{\varphi}$ on Δ^o, i.e.:

$$\underline{M} = \sup\{\Re(\lambda) \mid \lambda \in \sigma(D\tilde{\varphi}(x)), x \in \Delta\}$$

Let $x^M \in \Delta^o$ be a mutation equilibrium for some $M > \underline{M}$. As noted, the Jacobian of $\tilde{\varphi}^M$ satisfies $D\tilde{\varphi}^M(x) = D\tilde{\varphi}(x) - M \cdot I$ for all $x \in \Delta$. In particular, for all eigenvalues $\lambda^M \in \sigma(D\tilde{\varphi}^M(x^M))$ we have that $\lambda^M + M \in \sigma(D\tilde{\varphi}(x^M))$ and hence $\Re(\lambda^M) + M \leq \underline{M}$, and thus $\Re(\lambda^M) < 0$. Therefore, all eigenvalues of $D\tilde{\varphi}^M(x^M)$ have strictly negative real parts and with remark 4.3.7, x^M is asymptotically stable. □

Remark. Note that that the $\underline{M}$ in the previous lemma 4.3.8 is independent of the choice of $c \in \Delta^o$, thus giving a lower bound on the mutation rate above which all equilibria are asymptotically stable independent of $c \in \Delta^o$.

Uniqueness of mutation equilibria for high mutation rates

For very high mutation ($M > \underline{M}$) we further obtain that mutation equilibria are unique and that there is a continuously differentiable function mapping mutation rates to mutation equilibria. We first consider the following lemma (proven in as corollary 4.6.4 in section 4.6):

Lemma 4.3.9. *Let $c \in \Delta^o$ and $\underline{M}$ from lemma 4.3.8. Let x^M be a mutation equilibrium for some $M > \underline{M}$. Then there is a unique function $\mathscr{M} : (\underline{M}, \infty) \to \Delta$ such that $\mathscr{M}(M) =$*

x^M *and for all* $m \in (\underline{M}, \infty)$, $\mathscr{M}(m)$ *is a mutation equilibrium for* m. *In particular,* $\mathscr{M}$ *is continuously differentiable and* $\mathscr{M}(m) \overset{m\to\infty}{\longrightarrow} c$.

Note that this does not guarantee any uniqueness of equilibria, yet, only the uniqueness of functions passing through a given equilibrium. The uniqueness of mutation equilibria for high mutation rates is then obtained in the next step from the fact that we have uniqueness at least for some mutation rate (proven in as proposition 4.6.5 in section 4.6):

Proposition 4.3.10. *Let* $c \in \Delta^o$ *and* $\underline{M}$ *from lemma 4.3.8. For all* $M > \underline{M}$, *the replicator-mutator dynamics* (RMD) *has a unique mutation equilibrium. The unique map* $\mathscr{M} : M \mapsto x^M$ *is continuously differentiable on* $(\underline{M}, \infty)$.

Remark 4.3.11. Note that the main achievement of proposition 4.3.10 is to extend the uniqueness of equilibria beyond any Lipschitz constant of $\tilde{\varphi}$ to $(\underline{M}, \infty)$, i.e., to the interval where $D\tilde{\varphi}^M$ is guaranteed to be invertible. Furthermore, if $D\tilde{\varphi}^M(x^M)$ is invertible for all $M \in (a, \infty)$ and corresponding mutation equilibria x^M then the uniqueness extends to (a, ∞). In fact, if $a = 0$ then there is a unique sequence of mutation equilibria $(x^M)_{M>0}$ for $c \in \Delta^o$ since it is induced by the function $\mathscr{M}$.

For a fixed $c \in \Delta^o$ and a sufficiently high mutation rate, the unique mutation equilibrium will be arbitrarily close to c. Therefore, if we were interested in finding the mutation equilibrium for a sufficiently high mutation rate, we could choose an initial point close to c and the dynamics (RMD) would converge to the asymptotically stable mutation equilibrium. The uniqueness on $(\underline{M}, \infty)$ further enables us to lower the mutation rate almost to $\underline{M}$ without losing uniqueness and asymptotic stability.

4.4 Mutation limits

In our previous considerations, we assumed fixed relative mutation probabilities $c \in \Delta^o$. In particular, certain effects could depend on the specific choice of c, e.g., if we picked c to coincide with a Nash equilibrium $x^* \in \mathscr{E}$ of the underlying game. However, we are interested in properties that are independent of the specific choice of c. To this end, we introduce the following definition:

Definition 4.4.1 (Mutation Limit)**.** We call a connected compact set $X \subset \mathscr{E}$ a *mutation limit*, if for all $c \in \Delta^o$ there is a sequence of mutation equilibria $(x^M)_{M>0} \subset \Delta$ that

converges to an element of X for $M \to 0$ and X contains no proper subset with these properties. We call $x \in \Delta$ a *mutation limit point* if the singleton set $\{x\}$ is a mutation limit.

4.4.1 General existence of mutation limits

A question that arises from the definition is that of the existence of mutation limit points. While we have shown that for any fixed $c \in \Delta^{\circ}$ and any mutation rate $M > 0$ there is a corresponding mutation equilibrium and therefore the Bolzano-Weierstrass theorem guarantees the existence of a limit for vanishing mutation, this limit need not be independent of the choice of c, and indeed it could be possible that there is no mutation limit at all, neither a singleton set nor otherwise. The question, therefore, is whether every game has at least one mutation limit point. To this question, we can give a negative answer, as the following example shows:

Example 4.4.2. Consider a two-player game with the following payoff structure:

	C_1	C_2
R_1	$1, 0$	$0, 1$
R_2	$0, 1$	$1, 0$
R_3	$0, 1$	$1, 0$

It is clear that any Nash equilibrium of the game has the form $\left(\left(\frac{1}{2}, \frac{t}{2}, \frac{1-t}{2}\right), \left(\frac{1}{2}, \frac{1}{2}\right)\right)$ with $t \in [0, 1]$, where we give the strategy of the row player first. Excluding a few special choices of $c \in \Delta^{\circ}$, for any generic c given as $((c_{R,1}, c_{R,2}, c_{R,3}), (c_{C,1}, c_{C,2}))$, every sequence of mutation equilibria will converge to a Nash equilibrium of the above form with $t = c_{R,2}\left(c_{R,2} + c_{R,3}\right)^{-1}$. It is therefore evident that this game has no mutation limit point, i.e., there is no Nash equilibrium that is approached by mutation equilibria for all choices $c \in \Delta^{\circ}$. However, for any Nash equilibrium x of the above form with $t \in (0, 1)$ there is a $c \in \Delta^{\circ}$ such that x is approached by a sequence of mutation equilibria. Therefore, the set of Nash equilibria is indeed a mutation limit.

In the above example, the set of all Nash equilibria turns out to be a mutation limit. However in general, the set of Nash equilibria need not be connected. In this context, the following result answers the question about the general existence of mutation limits (proven in subsection 4.6.2):

Proposition 4.4.3. *For every $f \in \mathscr{C}^1(U \supset \Delta, \mathbb{R}^S)$ there is a mutation limit $X \subset \mathscr{E}$.*

Note that this result does not require that there is no intraspecies interaction, i.e., it does not require $\frac{\partial}{\partial x_{ik}} f_{ih}(x) = 0$ $(\forall i \in I, h, k \in S_i, x \in \Delta)$. In fact, the proof can be quite easily generalized to other, not necessarily replicator dynamics. From proposition 4.4.3, we obtain the following existence result for dynamics with only a finite number of Nash equilibria:

Corollary 4.4.4. *Let $f \in \mathscr{C}^1(U \supset \Delta, \mathbb{R}^S)$ such that the set of Nash equilibria, $\mathscr{E}$, is finite. Then all mutation limits are mutation limit points and there is at least one mutation limit point.*

Note that the finiteness condition is particularly important for fitness functions that are not derived from finite normal-form games.

A sufficient condition for mutation limits

We can further guarantee that regular Nash equilibria, introduced in [38], cf. also [112], are mutation limit points, where we employ the following equivalent definition, [81]:

Definition 4.4.5. We call a Nash equilibrium $x \in \Delta$ a *regular equilibrium* if the reduced Jacobian of (RD) at x, $D\tilde{\varphi}(x)$, is invertible.

In particular, all strict Nash equilibria are regular, [112, corollary 2.5.3].

Lemma 4.4.6. *Let x^* be a regular equilibrium. Then x^* is a mutation limit, i.e., for all $c \in \Delta^o$, there is a sequence of mutation equilibria, $(x^M)_{M>0}$, such that $x^M \to x^*$ for $M \to 0$.*

Proof. Note that $D\tilde{\varphi}(x^*)$ is invertible and therefore, by the implicit function theorem, for every $c \in \Delta^o$, there is a continuously differentiable $\mu : (-\varepsilon, \varepsilon) \to \mathbb{R}^N$ for some $\varepsilon > 0$, such that for $M \in (-\varepsilon, \varepsilon)$ we have that $\tilde{\varphi}^M(\mu(M)) = 0$. Of course, negative values of M are not interpretable as mutation rates and we consider them here only for technical reasons of differentiability at 0.

If $x^* \in \Delta^o$, then it is clear that we can choose ε such that $\mu([0, \varepsilon]) \subset \Delta$, and therefore a sequence of mutation equilibria $(x^M)_{M>0} \subset \Delta$ with $x^M \to x^*$ for $M \to 0$.

Suppose that $x^* \in \partial\Delta$ and for some $(i,h) \in S$ we have $x^*_{ih} = 0$. Note that μ is continuously differentiable and therefore for $M \in (-\varepsilon, \varepsilon)$,

$$\begin{aligned} 0 = \frac{d}{dM}\varphi^M_{ih}(\mu(M)) &= \frac{d}{dM}\Big(\mu_{ih}(M) g_{ih}(\mu(M))\Big) + \frac{d}{dM}\Big(M(c_{ih} - \mu_{ih}(M))\Big) \\ &= (g_{ih}(\mu(M)) - M)\,\frac{d}{dM}\mu_{ih}(M) + \mu_{ih}(M)\frac{d}{dM} g_{ih}(\mu(M)) + (c_{ih} - \mu_{ih}(M)) \end{aligned}$$

and hence for $M = 0$,

$$\begin{aligned} 0 &= \frac{d}{dM}\varphi^M_{ih}(\mu(M))|_{M=0} \\ &= g_{ih}(\mu(0))\frac{d}{dM}\mu_{ih}(0) + \mu_{ih}(0)\frac{d}{dM} g_{ih}(\mu(0)) + (c_{ih} - \mu_{ih}(0)) \\ &= g_{ih}(x^*)\frac{d}{dM}\mu_{ih}(0) + \underbrace{x^*_{ih}}_{=0}\frac{d}{dM} g_{ih}(x^*) + (c_{ih} - \underbrace{x^*_{ih}}_{=0}) = g_{ih}(x^*)\frac{d}{dM}\mu_{ih}(0) + c_{ih} \\ &> g_{ih}(x^*)\frac{d}{dM}\mu_{ih}(0)\ . \end{aligned}$$

Thus, with x^* being a Nash equilibrium, we have $g_{ih}(x^*) \leq 0$ and therefore $\frac{d}{dM}\mu_{ih}(0) \geq 0$. Because of the strict inequality, we even have $g_{ih}(x^*) < 0$ and $\frac{d}{dM}\mu_{ih}(0) > 0$. Therefore, we can choose ε such that $\mu([0,\varepsilon)) \subset \Delta$ and a sequence of mutation equilibria converging to x^*. □

Remark. It should be noted that the proof of the above result shows that there is a continuously differentiable function mapping mutation rates to mutation equilibria and that this function is unique. In other words, given a $c \in \Delta^o$, the sequence approaches x^* in a unique manner.

4.4.2 Attracting mutation limits

Up to this point we have considered equilibria (or sets of equilibria) of (RD) such that for any $c \in \Delta^o$ and mutation rate $M > 0$ a mutation equilibrium of the respective (RMD) would be located arbitrarily close, depending on M. We have so far ignored the stability properties of the mutation equilibria arising nearby. If the mutation equilibrium arising nearby happens to be asymptotically stable for some mutation rate $M > 0$ and some $c \in \Delta^o$, then under suitable initial conditions the system will converge to a state close to the mutation limit. However, as with the notion of mutation equilibria, such behaviour of the system is mostly of interest if it does not depend on a lucky choice of c, in particular if nearby mutation equilibria turn out to be asymptotically stable for every choice of c. In this case, the mutation limit would be approximated arbitrarily close in all (RMD) only depending on $M > 0$. This idea motivates the following formal definition:

Definition 4.4.7 (Attracting Mutation Limit)**.** We call a mutation limit $X \subset \Delta$ *attracting* if for every $c \in \Delta^o$ and every sequence of mutation equilibria $(x^M)_{M>0}$ that converges to an element of X, there is $m > 0$ such that for all $M < m$, x^M is asymptotically stable. We call $x \in \Delta$ an *attracting* mutation limit point if the singleton set $\{x\}$ is an attracting mutation limit.

A sufficient condition for attracting mutation limits

It is known that if x^* is a strict Nash equilibrium, then $D\tilde{\varphi}(x^*)$ has only real, strictly negative eigenvalues, e.g., [81, lemma 1], and x^* is therefore regular and thus a mutation limit. Furthermore, we can show that x^* is an attracting mutation limit:

Lemma 4.4.8. *Let x^* be a strict Nash equilibrium. Then x^* is an attracting mutation limit.*

Proof. With the previous note, it is clear that x^* is a mutation limit. It remains to show that the mutation equilibria $(x^M)_{M>0}$ converging to x^* for any $c \in \Delta^o$ are asymptotically stable. Since all eigenvalues of the Jacobian at x^* have strictly negative real parts, and in fact are real, [81], we have that the eigenvalues of $D\tilde{\varphi}(x)$ have strictly negative real parts in a neighbourhood of x^*, as the roots of a polynomial vary continuously with its coefficients, e.g., [37], and $D\tilde{\varphi}$ is continuous. Therefore, in a neighbourhood of x^*, all eigenvalues of the Jacobian of $\tilde{\varphi}^M$, with $D\tilde{\varphi}^M(x) = D\tilde{\varphi}(x) - M \cdot I$, have strictly negative real parts for any $M \geq 0$, and thus the x^M are asymptotically stable, e.g., [78]. □

Remark 4.4.9. Since the strict Nash equilibria are exactly the asymptotically stable equilibria of (RD), this ensures that all asymptotically stable equilibria are also attracting mutation limits, including evolutionary stable equilibria.

The following example shows that attracting mutation limits are not necessarily strict Nash equilibria, and hence that the concept of attracting mutation limits is also weaker than evolutionary stability:

Example 4.4.10. Consider the 2-by-2 matching pennies game given by the payoffs:

$$\begin{pmatrix} (1,0) & (0,1) \\ (0,1) & (1,0) \end{pmatrix}$$

The strategy profile $((1/2, 1/2), (1/2, 1/2))$ is a Nash equilibrium but not strict and hence not asymptotically stable. However, it is an attracting mutation limit: The eigenvalues of the Jacobian $D\tilde{\varphi}$ are given by

$$\lambda_{1,2} = \pm\sqrt{(1-2x)^2(1-2y)^2 - 4x(1-x)y(1-y)}.$$

At $(1/2, 1/2)$, the radicand is negative and the eigenvalues purely imaginary. Hence, the radicand is negative in a neighbourhood and the eigenvalues purely imaginary. Then the eigenvalues of $D\tilde{\varphi}^M$ have real part $-M$ in that neighbourhood due to remark 4.3.7 and for M sufficiently small all mutation equilibria are asymptotically stable with corollary 3.6, and hence $((1/2, 1/2), (1/2, 1/2))$ is an attracting mutation limit point. This also holds for the general matching pennies game, which we prove in chapter 5.

A necessary condition for attracting mutation limits

The observation that not all Nash equilibria are attracting mutation limits relies on the following:

Lemma 4.4.11. *Let $x^* \in \Delta$ be an attracting mutation limit. Then all eigenvalues of the Jacobian $D\tilde{\varphi}(x^*)$ have nonpositive real parts.*

Proof. Suppose there is an eigenvalue of $D\tilde{\varphi}(x^*)$ with a strictly positive real part. Then there is $\varepsilon > 0$ and a neighbourhood U of x^* such that $D\tilde{\varphi}(x)$ has an eigenvalue λ with $\Re(\lambda) > \varepsilon$ for all $x \in U$. Let $(x^M)_{M>0}$ be a sequence of mutation equilibria converging to x^* for some $c \in \Delta^o$. Then there is ε' such that $x^M \in U$ for $M < \varepsilon'$. In particular, we can choose $\varepsilon' < \varepsilon$. Then the Jacobian $D\tilde{\varphi}^M(x^M)$, with $D\tilde{\varphi}^M(x^M) = D\tilde{\varphi}(x^M) - M \cdot I$, has an eigenvalue with strictly positive real part, and x^M is not asymptotically stable, as it is not even stable, e.g., [43]. Therefore, x^* is not an attracting mutation limit. □

This result, together with the following example, then demonstrates that not all Nash equilibria are attracting mutation limits:

Example 4.4.12. Consider the 2-by-2 coordination game given by:

$$\begin{pmatrix} (1,1) & (0,0) \\ (0,0) & (1,1) \end{pmatrix}$$

The strategy profile $((1/2, 1/2), (1/2, 1/2))$ is a Nash equilibrium, but its Jacobian has eigenvalues $1/2$ and $-1/2$ and therefore it is not an attracting mutation limit.

4.5 Discussion

We have shown that a very simple form of mutation leads to qualitative changes in the multi-population replicator dynamics. Furthermore, these changes do not depend on the specific choice of parameters but are of a general character. Not only do mutation limits exist for all continuously differentiable fitness functions, mutation can also cause the dynamics to approximate equilibria that would not be approximated without mutation, again independently of the choice of specific mutation parameters, which is due to asymptotically stable equilibria arising close to an original equilibrium, as in the matching pennies game. The closest results to our approach that we are aware of are presented in [81], and if considered as an approximation to (RD), certain aspects of (RMD) are clarified by those results, as indicated. The results presented here differ in that they show robustness in a system of families of approximations which are not related to perturbed normal-form game payoffs and in that they focus on the effects on the stability of equilibria, independent of the choice of the specific approximation.

With respect to periodic behaviour in biological populations it should be noted that the degree of stabilisation of RD depends on the mutation rate, resulting in a very slow approach of an asymptotically stable mutation equilibrium and seemingly periodic behaviour if mutation is low. In an empirical situation this can lead to difficulties in distinguishing dynamics with truly periodic behaviour from ones with only seemingly periodic behaviour if measuring on a (relatively) small time scale. Furthermore, in small populations stochastic effects will play a significant role. Therefore, under very low mutation, empirical findings of periodic fluctuations can be consistent with our results if measured in small populations on a small time scale, such that any stabilising effects of mutation will be more apparent in large populations on large time scales, or with sufficiently fast reproduction.

On the one hand, given the potential health impacts of even slight mutations on organisms and the fact that such mutations occur with a non-negligible probability, as mentioned earlier, and given further its role as a generator of variety on which evolutionary selection operates, it is clear that it is worth including mutation mechanisms in the study of populations, and one should expect results that deviate potentially significantly from models without mutation.

On the other hand, given that the multi-population replicator dynamics has been shown to be related to learning dynamics and that mutation-like terms have been

shown to arise in formulations of Q-learning algorithms, it is worth noting that our results show that replicator-mutator dynamics have more desirable convergence properties than the pure replicator dynamics, while remaining arbitrarily close to a Nash equilibrium. Therefore, attracting mutation limits resulting from a replicator-mutator dynamics can be considered a more suitable class of dynamic solution approaches for games than the pure multi-population replicator dynamics.

As shown, attracting mutation limits do not exist for all games, and the characterization of their existence is therefore an open problem. We will address this problem partially in forthcoming results on attracting mutation limits in the matching pennies game, which can be considered a model of antagonistic coevolution. Furthermore, we have considered a specific form of mutation, and therefore the question of which properties carry over to more complicated and more realistic mutation mechanisms remains.

4.6 Proofs of propositions 4.3.10 and 4.4.3

4.6.1 Proof of proposition 4.3.10

The proof of proposition 4.3.10 relies on the implicit function theorem, which we restate for convenience, e.g., as in [56, theorem 3.3.1]:

Theorem 4.6.1 (Implicit Function)**.** *Let $W \subset \mathbb{R}$, $X \subset \mathbb{R}^n$ be open and let $\rho : W \times X \to \mathbb{R}^n$, $(w,x) \mapsto \rho(w,x)$ be a continuously differentiable function. Let further $(w',x') \in W \times X$ be such that $\rho(w',x') = 0$ and the $n \times n$ matrix $\frac{\partial}{\partial x}\rho(w',x')$ be invertible.*

Then there exist an open neighbourhood $W_F \subset W$ of w', an open neighbourhood $X_F \subset X$ of x', and a continuously differentiable function $F : W_F \to X_F$ such that $\forall w \in W_F : \rho(w,F(w)) = 0$. Furthermore, for all $(w,x) \in W_F \times X_F$ we have that $\rho(w,x) = 0$ if and only if $x = F(w)$, i.e., F is unique.

For the proof of proposition 4.3.10 we will need a consequence of the implicit function theorem, based on the following statement that we can extend an implicitly defined function if the conditions of the implicit function theorem hold on the boundary of its domain:

Lemma 4.6.2. *Let $\rho : W \times X \to \mathbb{R}^n$ be as given in theorem 4.6.1 and let $R : W_R \to X_R$ be continuously differentiable, with open and convex $W_R \subset W$ and open $X_R \subset X$, such that:*

i) $\forall v \in W_R : \rho(v, R(v)) = 0$;

ii) $\forall (v,x) \in W_R \times X_R : \rho(v,x) = 0 \Leftrightarrow x = R(v)$.

If for some sequence $(v_n)_{n\in\mathbb{N}} \subset W_R$ *with* $v_n \to v' \in \partial W_R \cap W$ *and an accumulation point* $x' \in X$ *of* $(R(v_n))_{n\in\mathbb{N}}$, *the matrix* $\frac{\partial}{\partial x}\rho(v',x')$ *is invertible, then there is a unique continuously differentiable extension of* R *with the above properties whose domain is open and a proper superset of* W_R. *In particular,* $(R(v_n))_{n\in\mathbb{N}}$ *is convergent with limit* x'.

Proof. Let $(v_n)_{n\in\mathbb{N}} \subset W_R$ with $v_n \to v' \in \partial W_R \cap W$ and let $x' \in X$ be an accumulation point of $(R(v_n))_{n\in\mathbb{N}}$, such that the matrix $\frac{\partial}{\partial x}\rho(v',x')$ is invertible. Due to the continuity of ρ on $W \times X$, we have that $\rho(v',x') = 0$. With the implicit function theorem, there are open neighbourhoods $W' \subset W$ of v', where we can require W' to be convex, and $X' \subset X$ of x' and a unique continuously differentiable function $S : W' \to X'$ with the corresponding properties i) and ii).

We will show that there is N such that $(R(v_n))_{n\geq N} \subset X'$: As x' is an accumulation point of $(R(v_n))_{n\in\mathbb{N}}$, there are infinitely many $n \in \mathbb{N}$ with $R(v_n) \in X'$, in particular let $R(v_N) \in X'$. Note that we can assume $(v_n)_{n\geq N} \subset W'$ as $v' \in W'$ is the limit of that sequence. Assume that there is some $N' > N$ with $R(v_{N'}) \notin X'$ and let N' be minimal. W.l.o.g. let $N' = N+1$ and define $v : [0,1] \to W', t \mapsto (1-t)v_N + tv_{N'}$. Then $v([0,1]) \subset W'$ due to convexity. Consider that $R(v_N) \in X'$, with X' open. Therefore, there is some $\varepsilon > 0$ with $R(v([0,\varepsilon])) \subset X'$. However, with our assumption, $R(v(1)) = R(v_{N'}) \notin X'$. Then, with the complement of X' being closed, there is a minimal $\bar{t}$ such that $R(v(\bar{t})) \notin X'$. Then $R \circ v = S \circ v$ on $[0,\bar{t})$, but due to their continuity we then also have $R(v(\bar{t})) = S(v(\bar{t}))$ and thus $R(v(\bar{t})) \in X'$, in contradiction to $R(v(\bar{t})) \notin X'$. Thus, $R(v_{N'}) = R(v(1)) \in X'$, in contradiction to $R(v_{N'}) \notin X'$. Overall, we then have $(R(v_n))_{n\geq N} \subset X'$, and further $R([v_N, v')) \subset X'$ (assuming $v_N < v'$). This implies that $R = S$ on $W_R \cap W'$ and $T := R \cup S$ is a proper, continuously differentiable extension of R, satisfying properties i) and ii). In particular, due to $(R(v_n))_{n\geq N} = (T(v_n))_{n\geq N}$, $(R(v_n))_{n\in\mathbb{N}}$ is convergent with limit x'. □

The following lemma states that there is an implicitly defined function whose domain is such that the points at the boundary do not satisfy the conditions of the implicit function theorem:

Lemma 4.6.3. *Let $\rho : W \times X \to \mathbb{R}^n$ be as given in theorem 4.6.1 and $(w, x^w) \in W \times X$ such that $\rho(w, x^w) = 0$ and the matrix $\frac{\partial}{\partial x}\rho(w, x^w)$ is invertible. Then there exist open neighbourhoods $W^* \subset W$ of w, with W^* convex, and $X^* \subset X$ of x^w, and a continuously differentiable function $R^* : W^* \to X^*$ such that:*

i) $\forall v \in W^* : \rho(v, R^*(v)) = 0$;

ii) $\forall (v, x) \in W^* \times X^* : \rho(v, x) = 0 \Leftrightarrow x = R^*(v)$;

iii) for all $(v_n)_{n\in\mathbb{N}} \subset W^*$ with $v_n \to v' \in \partial W^* \cap W$ and every accumulation point $x' \in X$ of $(R^*(v_n))_{n\in\mathbb{N}}$, the matrix $\frac{\partial}{\partial x}\rho(v', x')$ is singular.

In particular, R^* is a maximally defined such function.

Proof. Let $\mathcal{R}$ be the set of all continuously differentiable functions $R_\alpha : W_\alpha \to X_\alpha$, with $W_\alpha \subset W$ convex and $X_\alpha \subset X$ being open neighbourhoods of w and x^w, respectively, such that R_α satisfies i) and ii). Due to ρ being continuously differentiable, $\frac{\partial}{\partial x}\rho$ is invertible in a convex, open neighbourhood of (w, x^w). With the implicit function theorem, $\mathcal{R}$ is not empty. We define a partial order on $\mathcal{R}$ by the set inclusion on the graphs of the functions $R_\alpha \in \mathcal{R}$.

Let $\mathcal{O}$ be a non-empty completely ordered chain in $\mathcal{R}$. Consider the function R' defined by the graph:

$$\Gamma(R') = \bigcup_{R_\alpha \in \mathcal{O}} \{(v, R_\alpha(v)) \mid v \in W_\alpha\}$$

Then $W' = \bigcup_{R_\alpha\in\mathcal{O}} W_\alpha \subset W$ and $X' = \bigcup_{R_\alpha\in\mathcal{O}} X_\alpha \subset X$ are open neighbourhoods of w and x^w and $R' : W' \to X'$ is a continuously differentiable function. Furthermore, $\{W_\alpha \mid R_\alpha \in \mathcal{O}\}$ is completely ordered by set inclusion as well and therefore, W' is convex. It is clear that R' satisfies i) as all R_α satisfy i). Let $(v, x) \in W' \times X'$. Then there is $R_\alpha \in \mathcal{O}$ with $v \in W_\alpha$, $x \in X_\alpha$, and $R'(v) = R_\alpha(v)$. Then, as R_α satisfies ii), we have $\rho(v, x) = 0 \Leftrightarrow x = R_\alpha(v) = R'(v)$, and thus R' satisfies ii). Therefore, $R' \in \mathcal{R}$, and with Zorn's Lemma, $\mathcal{R}$ contains a maximal element $R^* : W^* \to X^*$, such that R^* satisfies i) and ii).

For iii), let $(v_n)_{n\in\mathbb{N}} \subset W^*$ with $v_n \to v' \in \partial W^* \cap W$ and let $x' \in X$ be an accumulation point of $(R^*(v_n))_{n\in\mathbb{N}}$. Assume that the matrix $\frac{\partial}{\partial x}\rho(v', x')$ is invertible. With the previous lemma there is a proper extension of R^* and R^* is not maximal, a contradiction. Thus, $\frac{\partial}{\partial x}\rho(v', x')$ is singular. □

In order to apply the above lemma, for $M > 0$, we rewrite (RMD) as

$$\rho : \mathbb{R} \times X \to \mathbb{R}^S, (w,x) \mapsto w\varphi(x) + (c - x) \tag{4.6.1}$$

with $w = M^{-1}$. It is clear that $\rho(M^{-1},x) = M^{-1}\varphi^M(x)$ and therefore $\rho(M^{-1},x) = 0 \Leftrightarrow \varphi^M(x) = 0$ and that ρ is continuously differentiable on $\mathbb{R} \times X$ with some $X \supset \Delta$ open and bounded, depending on φ. Then we obtain lemma 4.3.9 as a corollary:

Corollary 4.6.4. *Let $c \in \Delta^o$ and $\underline{M}$ be as in lemma 4.3.8. Let x^M be a mutation equilibrium for some $M > \underline{M}$. Then there is a unique function $\mathscr{M} : (\underline{M}, \infty) \to \Delta$ such that $\mathscr{M}(M) = x^M$ and for all $m \in (\underline{M}, \infty)$, $\mathscr{M}(m)$ is a mutation equilibrium for m. In particular, $\mathscr{M}$ is continuously differentiable and $\mathscr{M}(m) \overset{m\to\infty}{\longrightarrow} c$.*

Proof. Consider that for $m > \underline{M}$, $D\varphi^m$ is invertible everywhere on Δ due to lemma 4.3.8, and that for $w = m^{-1}$ with ρ from (4.6.1), the matrix $\frac{\partial}{\partial x}\rho(w,x)$ is invertible whenever $D\varphi^m(x)$ is. Then let $\underline{w} = \underline{M}^{-1}$ and $w = M^{-1}$ for some $M > \underline{M}$. Then applying the previous lemma to w, x^M and ρ yields a continuously differentiable function $R : W \to \Delta$ with $W \subset \mathbb{R}$ and $w \in W$. Furthermore, the previous lemma guarantees that $[0,\underline{w}) \subset W$ because $\frac{\partial}{\partial x}\rho(v,x)$ is invertible $\forall v \in [0,\underline{w}), x \in \Delta$. Thus, $\mathscr{M} : (\underline{M}, \infty) \to \Delta$ with $m \mapsto R(m^{-1})$ is continuously differentiable and is as desired. □

With this we can prove proposition 4.3.10:

Proposition 4.6.5 (4.3.10)**.** *Let $c \in \Delta^o$ and $\underline{M}$ as in lemma 4.3.8. For all $M > \underline{M}$, the replicator-mutator dynamics* (RMD) *has a unique mutation equilibrium. The unique map $\mathscr{M} : M \mapsto x^M$ is continuously differentiable on $(\underline{M}, \infty)$.*

Proof. As φ is Lipschitz, let L_φ be the best Lipschitz constant for φ. Since φ is differentiable and Δ is convex, we further have that $L_\varphi = \|D\varphi\|_{\infty,\Delta} := \sup_{x\in\Delta} \|D\varphi(x)\| \geq \underline{M}$ with $\underline{M}$ from lemma 4.3.8. Choose $M' > L_\varphi$ and consider for $c \in \Delta^o$ and some $s > 0$ the function $F_{M',c} : \Delta \to \Delta$ with $[F_{M',c}(x)]_{ih} = x_{ih} + s\left(\varphi_{ih}(x) + M'(c_{ih} - x_{ih})\right)$. Then, we have that

$$[F_{M',c}(x)]_{ih} - [F_{M',c}(y)]_{ih} = (1 - sM')(x_{ih} - y_{ih}) + s\left(\varphi_{ih}(x) - \varphi_{ih}(y)\right)$$

and thus

$$\begin{aligned}\|F_{M',c}(x) - F_{M',c}(y)\| \leq& |1 - sM'|\|x - y\| + s\|\varphi(x) - \varphi(y)\| \\ \leq& |1 - sM'|\|x - y\| + sL_\varphi\|x - y\| = (|1 - sM'| + sL_\varphi)\|x - y\|.\end{aligned}$$

Choosing s such that $sM' \leq 1$, we have that:

$$|1 - sM'| + sL_\varphi = 1 - sM' + sL_\varphi = 1 + s(L_\varphi - M') < 1$$

Hence, $F_{M',c}$ is a contractive mapping and has a unique fixed point $x^{M'} \in \Delta^0$. Then every function $\mathscr{M}$ from corollary 4.6.4 satisfies $\mathscr{M}(M') = x^{M'}$ and thus all such functions are identical, yielding the uniqueness of mutation equilibria for all $M > \underline{M}$. □

4.6.2 Proof of proposition 4.4.3

In order to prove proposition 4.4.3, we need to extend (RMD) slightly, such that we can allow more general mutation to occur. Recall that $g_{ih}(x) = f_{ih}(x) - \bar{f}_i(x)$ and that then $\mathscr{E} = \{x \in \Delta \,|\, g(x) \leq 0\}$ is the set of Nash equilibria, where the inequality is component-wise. Then let $H = \mathscr{C}^1(\Delta, \mathbb{R}^S_{>0})$, and define for $c \in H$, $M > 0$:

$$[F_{M,c}(x)]_{ih} = x_{ih} + s\left(x_{ih} g_{ih}(x) + M\left(c_{ih}(x) - x_{ih} \textstyle\sum_{k \leq n_i} c_{ik}(x)\right)\right)$$

where $i \in I$, $h \in S_i$. Note that for all $s > 0$, the fixed points of $F_{M,c}$ are the stationary points of a suitably generalized (RMD). In particular, if $c \in H$ is constant on Δ, then the fixed points are exactly the mutation equilibria of (RMD) for a suitably chosen $\tilde{M}$. It is clear that for a choice of $c \in H$, we can choose $s > 0$ such that for all $M \in (0, \varepsilon_s)$, we have $F_{M,c}(\Delta) \subset \Delta$ and thus the set of fixed points is non-empty. Therefore, we assume a suitable choice of $s > 0$ (possibly depending on c). For convenience, let us denote by $\mathscr{F}(F_{M,c})$ the set of fixed points of $F_{M,c}$ for $c \in H$ and $M > 0$:

$$\mathscr{F}(F_{M,c}) = \{x \in \Delta \,|\, F_{M,c}(x) = x\}.$$

From the definition of a mutation limit, we extract the main property and say that a set $X \subset \Delta$ has the property (A) if

(A) for all $c \in \Delta^0$, there is a sequence of mutation equilibria $(x^M)_{M>0} \subset \Delta$ that converges to an element of X.

We extend this notion to $F_{M,c}$ and say that a set $X \subset \Delta$ has the property (A') if

(A') for all $c \in H$ and open $U \supset X$, there is $M > 0$ such that $\mathscr{F}(F_{M,c}) \cap U \neq \emptyset$.

Remark. It is clear that a set X has the property (A') if and only if for every $c \in H$ there is a sequence $(x^M)_{M>0} \subset \Delta$ such that $(x^M)_{M>0}$ converges to an element of X and every x^M in the sequence satisfies $x^M \in \mathscr{F}(F_{M,c})$. With this it is also clear that

a set has the property (A) if it has property (A'), due to $c \in \Delta^0$ being equivalent to a constant function in H.

The proof of proposition 4.4.3 will proceed as follows: We first show that $\mathscr{E}$ has the property (A'). Next, we show that a set with the property (A') contains a minimal set with that property, and that an analog but slightly modified result holds for the property (A). We then show that a minimal set with the property (A') is connected, based on a proof by Kinoshita [55]. Thus, we have that $\mathscr{E}$ contains a minimal set with the property (A'), which must be contained in a connected component of $\mathscr{E}$. Finally this set is connected and in particular has the property (A) and hence contains a minimal connected set with the property (A), proving proposition 4.4.3.

Existence. We show first that any minimal set with the property (A') must be contained in $\mathscr{E}$:

Lemma 4.6.6. *Let $X \subset \Delta$ be minimal with the property (A'). Then $X \subset \mathscr{E}$ and $\mathscr{E}$ has the property (A').*

Proof. Assume that $X \not\subset \mathscr{E}$. Let $c \in H$ and $(M_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$ be a null sequence, and $(x^{M_n})_{n\in\mathbb{N}} \subset \Delta$ convergent with limit x^* with $x^{M_n} \in \mathscr{F}(F_{M_n,c})$ for all $n \in \mathbb{N}$. From our earlier note on the possibility of a constant choice of $s > 0$ for all $n \in \mathbb{N}$, and from the continuity of g and c, we have that for all $i \in I$, $h \in S_i$, $x^*_{ih} g_{ih}(x^*) = 0$ holds.

We now show that $x^* \in \mathscr{E}$: If $x^* \in \Delta^0$, then for all $i \in I$, $h \in S_i$, $x^*_{ih} g_{ih}(x^*) = 0$ implies $g_{ih}(x^*) = 0$, i.e., $x^* \in \mathscr{E}$. If $x^* \in \partial\Delta$, then let some $(i,h) \in S$ be such that $x^*_{ih} = 0$, and let $\tilde{c}_i = \sup\left\{\sum_{k\le n_i} c_{ik}(x) \mid x \in \Delta\right\}$. Then $\tilde{c}_i < \infty$ and for $M > 0$:

$$\begin{aligned} x^M_{ih} = [F_{M,c}(x^M)]_{ih} &= x^M_{ih} + s\left(x^M_{ih} g_{ih}(x^M) + M\left(c_{ih}(x^M) - x^M_{ih}\sum\nolimits_{k\le n_i} c_{ik}(x^M)\right)\right) \\ &> x^M_{ih} + s\left(x^M_{ih} g_{ih}(x^M) - M x^M_{ih}\sum\nolimits_{k\le n_i} c_{ik}(x^M)\right) \ge x^M_{ih} + s x^M_{ih}\left(g_{ih}(x^M) - M\tilde{c}_i\right) \end{aligned}$$

Therefore, we have for all $M > 0$:

$$0 > s x^M_{ih}\left(g_{ih}(x^M) - M\tilde{c}_i\right) \quad\Leftrightarrow\quad 0 > g_{ih}(x^M) - M\tilde{c}_i \quad\Leftrightarrow\quad M\tilde{c}_i > g_{ih}(x^M)$$

Therefore, with $M \to 0$, we have $g_{ih}(x^*) \le 0$, and overall $x^* \in \mathscr{E}$. Thus $X \cap \mathscr{E}$ has the property (A') and X is not minimal, a contradiction. From $x^* \in \mathscr{E}$, it is clear that $\mathscr{E}$ has the property (A'). □

Minimality. We first show that the existence of a set with the property (A') implies the existence of a minimal such set, where the proof is fairly standard and adapted from [65, theorem 7.3]:

Lemma 4.6.7. *Let a compact set $X \subset \Delta$ have the property (A'). Then it contains a minimal compact set with the property (A').*

Proof. The proof uses Zorn's lemma. Let C be the set of compact subsets of X with the property (A'), i.e., $C = \left\{K \subset X \,\middle|\, K \neq \emptyset \text{ is compact and has the property } (A')\right\}$, and order C by reverse inclusion $\supset$. Let $O \subset C$ be completely ordered. Then O has the finite intersection property, as it is completely ordered by reverse inclusion and its elements are compact. Therefore, $K_\infty := \bigcap O \neq \emptyset$ and K_∞ is compact.

It remains to show that K_∞ has the property (A'): Assume K_∞ does not have the property (A'). Then there is a $c \in H$ and an open neighbourhood V of K_∞ such that no $F_{M,c}$ $(M > 0)$ has a fixed point in V. For $L \in O$, we have $L \not\subset V$ because L has the property (A'). Then $O' := \{L \backslash V : L \in O\}$ is a completely ordered collection of compact sets (L is compact and V is open) with the finite intersection property, inherited from the reverse inclusion ordering of O. Therefore, it has a nonempty intersection $K'_\infty \subset K_\infty \subset V$ but $K'_\infty \cap V = \emptyset$, which is a contradiction. Thus, K_∞ has the property (A') and therefore $K_\infty \in C$ is an upper bound of O. With Zorn's lemma then, C has a maximal element, which is a minimal compact subset of X with the property (A'). □

For the existence of a mutation limit we will have to make a similar step, however preserving connectedness:

Lemma 4.6.8. *Let a connected compact set $X \subset \Delta$ have the property (A). Then it contains a minimal connected compact set with the property (A).*

Proof. Let C be the set of all compact connected (non-empty) subsets of X with the property (A), partially ordered by $\supset$ and O a completely ordered chain in C. Then $K_\infty = \bigcap_{K \in O} K$ is non-empty, compact and has the property (A) by an argument completely analogous to the previous lemma.

It remains to show that K_∞ is connected: Assume that K_∞ is not connected. Then, there are open disjoint sets U_1, U_2, with $K_\infty \subset U_1 \cup U_2 =: U$ and $K_\infty \cap U_1 \neq \emptyset$, $K_\infty \cap U_2 \neq \emptyset$, and U open in X. X and all $K \in O$ are compact and, with X being Hausdorff, also closed. Thus $X \setminus K$ is open in X for $K \in O$. Then, with $\bigcup_{K \in O} X \setminus K = X \setminus \bigcap_{K \in O} K = X \setminus K_\infty$, we have that $\{U\} \cup \{X \setminus K | K \in O\}$ is an open cover of X, and

there is a finite subcover $\{U\} \cup \{X \setminus K_i | K_i \in O, 1 \leq i \leq n\}$, as X is compact. Thus $X = U \cup \bigcup_{1 \leq i \leq n} X \setminus K_i = U \cup X \setminus \bigcap_{1 \leq i \leq n} K_i$. As O is completely ordered by inclusion, we can assume that $K_i \supset K_n$ $(1 \leq i \leq n)$ and we have that $X = U \cup X \setminus K_n$. Thus $K_n \subset U = U_1 \cup U_2$, and hence K_n is not connected, a contradiction. Therefore, K_∞ is connected and $K_\infty \in C$. With Zorn's lemma the statement of the lemma follows. □

Connectedness. We gain connectedness as a necessary property of minimal sets with the property (A'), where the main idea of the proof is based on a proof by Kinoshita [55] and relies on the "convexity" of H:

Lemma 4.6.9. *If $K \subset \Delta$ has the property (A') and $K = (K_1 \cup \ldots \cup K_s)$ with the K_j disjoint and compact, then some K_j has the property (A'). If K is minimal with the property (A'), then K is connected.*

Proof. Let $K \subset \Delta$ have property (A') and $K = K_1 \cup \ldots \cup K_s$ with the K_j disjoint and compact. Assume that no K_j has the property (A'). Then there are $c_1, \ldots, c_s \in H$ and neighbourhoods $U_1, \ldots, U_s$ of $K_1, \ldots, K_s$ with disjoint closures such that for all $M > 0$, $\mathscr{F}(F_{M,c_j}) \cap U_j = \emptyset$. Let further $V_1, \ldots, V_s$ be strictly smaller neighbourhoods, i.e., $\overline{V_j} \subsetneq U_j$, and let U_0 be a neighbourhood of $\Delta \setminus (U_1 \cup \ldots \cup U_s)$ whose closure is disjoint from the $V_1, \ldots, V_s$, and c_0 any function in H. Then $\{U_0, U_1, \ldots, U_s\}$ is an open cover of Δ and with Δ being a compact subset of a topological vector space, there is a C^∞-partition of unity $\pi_0, \pi_1, \ldots, \pi_s$ such that $\pi_j(x) = 0$ $(\forall x \in \Delta \setminus U_j)$, and $\sum_{j=0}^{s} \pi_j(x) = 1$ $(\forall x \in \Delta)$, [65, theorem 6.2]. The convex combination, $\bar{c}$, with $\bar{c} : x \mapsto \sum_{j=0}^{s} \pi_j(x) c_j(x)$, is an element of H. Considering $F_{M,\bar{c}}$, we then have that $F_{M,\bar{c}}(x) = F_{M,c_j}(x)$ for $x \in V_j$. Thus $\mathscr{F}(F_{M,\bar{c}}) \cap V_j = \emptyset$ for $1 \leq j \leq s$ for all $M > 0$. Therefore, $F_{M,\bar{c}}$ has no fixed points in $(V_1 \cup \ldots \cup V_s) \supset K$ for any $M > 0$. This is a contradiction to the assumption that K has the property (A'). In particular, if K is minimal, then K is connected. □

Overall, this proves the following:

Proposition 4.6.10. *There is a mutation limit $X \subset \mathscr{E}$.*

Proof. With lemma 4.6.6, $\mathscr{E}$ has the property (A'). With $\mathscr{E}$ being compact due to $g \in \mathscr{C}(\Delta, \mathbb{R}^S)$ and $\mathscr{E} \subset \Delta$, and with lemma 4.6.7, there is a minimal compact set $X' \subset \mathscr{E}$ with the property (A'). Furthermore, with lemma 4.6.9, X' is connected. With the property (A'), X' also has the property (A). With lemma 4.6.8, X' contains

a minimal connected compact subset $X \subset X'$ with the property (A). By definition, X is a mutation limit. □

5 Mutation and asymptotic stability

In the multi-population replicator dynamics, interior equilibria cannot be asymptotically stable. In chapter 4, we have given a rigorous formulation of the idea of vanishing mutation and have linked the equilibria of systems with mutation to those without mutation. Furthermore, we have shown that, under the replicator dynamics, every game has at least one mutation limit, a connected component of equilibria that is robust under mutation: all systems with mutations have equilibria close to such a set. In particular, the specific mutation parameters are irrelevant.

One consequence of considering mutation is that interior equilibria can become asymptotically stable under arbitrarily low levels of mutation, and we have called such equilibria (or sets of equilibria) attracting mutation limits. There, we have also demonstrated that the sole equilibrium of the standard Matching Pennies game is such an attracting mutation limit. Here, we demonstrate that this phenomenon holds in much more general cases: Our first result shows that all Lyapunov stable equilibria are attracting mutation limits in all two-by-two systems. Our second result shows that in all two-player (rescaled) zero-sum games, all Lyapunov stable equilibria are attracting mutation limits. Overall, this indicates that mutation stabilizes evolution in these two classes of settings.

The Matching Pennies game and zero-sum games are of particular relevance in antagonistic biological settings such as host-parasite systems. Such a host-parasite system can be modelled as a system of two populations, where a parasite needs to match a host's trait in order to effectively infect the host, as considered, e.g., in [70, 88]. As shown in [99], under the assumption of constant host and parasite populations, the Lotka-Volterra dynamics in [88] is equivalent to a two population replicator dynamics with parasites and hosts playing a Matching Pennies type game.

5.1 Two population replicator-mutator dynamics

We first repeat the two-population replicator dynamics, proceeding to the replicator-mutator dynamics in a second step. In general, the populations are enumerated, but with only two populations, we will name the first population X and the second Y for convenience, with each population consisting of individuals of types from the type sets S_X and S_Y, respectively.[1] Let the frequencies of types in populations X and Y be given by $x = (x_h)_{h\in S_X} \in \Delta_X$ and $y = (y_h)_{h\in S_Y} \in \Delta_Y$, respectively, with $\Delta_i = \left\{ z \in \mathbb{R}_{\geq 0}^{|S_i|} \,\middle|\, \sum_{h\in S_i} z_h = 1 \right\}$ for $i \in \{X,Y\}$. For population X, let the fitness of a type $h \in S_X$ be given by $\tilde{f}_{X,h}(y)$ and analogously for population Y. If we consider that $x_{|S_X|} = 1 - \sum_{h\in S_X} x_h$ and $y_{|S_Y|} = 1 - \sum_{h\in S_Y} y_h$, then we can write the replicator dynamics as a reduced system ignoring $x_{|S_X|}$ and $y_{|S_Y|}$:

$$\begin{aligned} \dot{x}_h &= x_h \left(f_{X,h}(y) - \sum_{k<|S_X|} x_k f_{X,k}(y) \right) &&=: \varphi_{X,h}(x,y) \quad (h < |S_X|) \\ \dot{y}_j &= y_j \left(f_{Y,j}(x) - \sum_{k<|S_Y|} y_k f_{Y,k}(x) \right) &&=: \varphi_{Y,j}(x,y) \quad (j < |S_Y|) \end{aligned} \tag{RD}$$

where $f_{X,h}(y) = \tilde{f}_{X,h}(y) - \tilde{f}_{X,|S_X|}(y)$ for $h < |S_X|$ and analogously for $f_{Y,j}$ and $j < |S_Y|$. We denote by $J(x,y)$ the Jacobian of the system (RD) at some $(x,y) \in \Delta := \Delta_X \times \Delta_Y$.

Note that the continuous-time multi-population replicator dynamics (RD) can be considered as the limit case of weak-selection due to nearly infinite background fitness. Similarly, we obtain our replicator-mutator dynamics (RMD) as the weak-selection weak-mutation limit of a more general replicator-mutator equation, e.g. [77], as detailed in chapter 4. Under our assumption that mutation is memory-less, i.e., the probability $c_{i,h}$ that offspring will mutate to some type h is independent of the parent's type ($\forall i \in \{X,Y\}$, $h \in S_i$), we have that the mutation parameters c_i are from Δ_i ($i \in \{X,Y\}$) and can be interpreted as the population compositions favoured by mutation. As detailed in chapter 4, we can then extend (RD) as follows:

Given mutation parameters $c_i \in \Delta_i$ and mutation strength parameters $M_i > 0$ ($i \in \{X,Y\}$), the replicator-mutator dynamics is given by:

$$\begin{aligned} \dot{x}_h &= \varphi_{X,h}(x,y) - M_X(x_h - c_{X,h}) &&=: \varphi^M_{X,h}(x,y) \quad (h < |S_X|) \\ \dot{y}_j &= \varphi_{Y,h}(x,y) - M_Y(y_j - c_{Y,j}) &&=: \varphi^M_{Y,j}(x,y) \quad (j < |S_Y|) \end{aligned} \tag{RMD}$$

[1] We assume that we have chosen some ordering or some enumeration wherever we need to handle matrices or other ordered structures.

where $M = (M_X, M_Y)$. Again, we denote by $J^M(x,y)$ the Jacobian of (RMD) at some $(x,y) \in \Delta$. Note that the Jacobian of (RMD) satisfies

$$J^M(x,y) = J(x,y) - \mathfrak{M},$$

where $\mathfrak{M}$ is a block matrix of the form

$$\mathfrak{M} = \begin{pmatrix} M_X I_{|S_X|-1} & 0 \\ 0 & M_Y I_{|S_Y|-1} \end{pmatrix}$$

with I_n denoting the $n \times n$ identity matrix.

5.2 2×2 settings

As a special case, we consider settings with only two types in each population. In particular, this covers the classical case of 2×2 matrix games. In this setting, (RMD) depends only on x_1 and y_1 respectively and, dropping the indices for simplicity, we can write (RMD) as:

$$\begin{aligned} \dot{x} &= x(1-x)f_X(y) - M_X(x - c_X) \\ \dot{y} &= y(1-y)f_Y(x) - M_Y(y - c_Y) \end{aligned} \tag{5.2.1}$$

for $M = (M_X, M_Y)$ with $M_X, M_Y > 0$ and $c_X, c_Y \in (0,1)$, and the reduced fitness functions in (5.2.1) are obtained as $f_i := f_{i,1} - f_{i,2}$ for $i \in \{X, Y\}$. In this case, we have a two-dimensional dynamic system and we can completely characterize when regular (definition 4.4.5) interior Nash equilibria are attracting mutation limits for all, in particular non-linear, $\mathscr{C}^1$ functions f_X and f_Y. Further, the Jacobian J of (RD) at some (x,y) reduces to

$$J(x,y) := \begin{pmatrix} J_{Xx}(x,y) & J_{Xy}(x,y) \\ J_{Yx}(x,y) & J_{Yy}(x,y) \end{pmatrix} := \begin{pmatrix} (1-2x)f_X(y) & x(1-x)\frac{d}{dy}f_X(y) \\ y(1-y)\frac{d}{dx}f_Y(x) & (1-2y)f_Y(x) \end{pmatrix}$$

and consequently, the Jacobian J^M of (5.2.1) at (x,y) is given by

$$J^M(x,y) := J(x,y) - \begin{pmatrix} M_X & 0 \\ 0 & M_Y \end{pmatrix}.$$

Remark. Note that if $(x^*, y^*) \in \Delta$ is a Nash equilibrium, then we have that $f_X(y^*) = f_Y(x^*) = 0$ and hence the trace of the Jacobian satisfies $\mathrm{Tr}(J(x^*, y^*)) = 0$.

We approach the question of stability in two steps. We first consider the signs of the diagonal elements of J^M and then use this to obtain a result based on the determinant of J in a second step.

Lemma 5.2.1. *Let $M_X, M_Y \geq 0$ with $M_X > 0$ or $M_Y > 0$, and let (x^M, y^M) be an interior mutation equilibrium.*[2] *Then $J^M_{Xx}(x^M, y^M), J^M_{Yy}(x^M, y^M) \leq 0$ with at least one strictly negative and $\mathrm{Tr}(J^M) < 0$.*

Proof. Assume first that $M_X, M_Y > 0$. Consider that in an interior mutation equilibrium (x^M, y^M), we have $\varphi_X(x^M, y^M) = M_X(x^M - c_X)$ (and $\varphi_Y(x^M, y^M) = M_Y(y^M - c_Y)$) and thus

$$f_X(x^M, y^M) = M_X \frac{x^M - c_X}{x^M(1 - x^M)}$$

and further:

$$\begin{aligned} J^M_{Xx} = J_{Xx} - M_X = \frac{\partial}{\partial x}\varphi_X(x^M, y^M) - M_X = (1 - 2x^M)M_X \frac{x^M - c_X}{x^M(1 - x^M)} - M_X \\ = -M_X \frac{(x^M)^2 - 2x^M c_X + c_X + c_X^2 - c_X^2}{x^M(1 - x^M)} = -M_X \frac{(x^M - c_X)^2 + c_X(1 - c_X)}{x^M(1 - x^M)} < 0, \end{aligned}$$

and similarly $J^M_{Yy} = J_{Yy} - M_Y < 0$ and therefore $\mathrm{Tr}(J^M) < 0$. Now if $M_X = 0$, then $J^M_{Xx} = 0$, but with $M_Y > 0$ we still have $J^M_{Yy} < 0$, and similarly for $M_Y = 0$, which concludes the proof. □

With this we obtain the following result for Nash equilibria (x^*, y^*) with positive determinant of the Jacobian:

Proposition 5.2.2. *Let (x^*, y^*) be an interior Nash equilibrium such that the Jacobian $J(x^*, y^*)$ satisfies $|J(x^*, y^*)| > 0$. Then (x^*, y^*) is an attracting mutation limit.*

Proof. Let $|J(x^*, y^*)| > 0$. Then (x^*, y^*) is a regular equilibrium and hence a mutation limit with [8, lemma 4.6]. Let then $(x^M, y^M)_{(M_X, M_Y) \to 0}$ be a sequence of mutation equilibria converging to (x^*, y^*). Consider that the eigenvalues of the Jacobian $J^M(x^M, y^M)$ are given by:

$$\lambda_{1,2} = \frac{\mathrm{Tr}(J^M)}{2} \pm \sqrt{\left(\frac{\mathrm{Tr}(J^M)}{2}\right)^2 - |J^M|}$$

Now, for $M \to 0$, $\mathrm{Tr}(J^M(x^M, y^M)) \to \mathrm{Tr}(J(x^*, y^*)) = 0$, while $|J^M(x^M, y^M)| \to |J(x^*, y^*)| > 0$. Hence, there are $\overline{M}_x > 0$, $\overline{M}_y > 0$ such that, for all $M = (M_X, M_Y) \in (0, \overline{M}_x) \times (0, \overline{M}_y)$, the radicand is negative at (x^M, y^M). Together with lemma 5.2.1, we have that $\Re(\lambda_{1,2}) = \frac{1}{2}\mathrm{Tr}(J^M) < 0$, and with [78, theorem 1, p. 130], that (x^M, y^M) is asymptotically stable and thus (x^*, y^*) an attracting mutation limit. □

In particular, the previous statement applies to neutrally stable regular equilibria:

[2] Note that all mutation equilibria are interior points.

Proposition 5.2.3. *Let* (x^*, y^*) *be a regular interior Nash equilibrium. If* (x^*, y^*) *is neutrally stable, then it is an attracting mutation limit.*

Proof. Let (x^*, y^*) be a regular interior Nash equilibrium. Then $|J(x^*, y^*)| \neq 0$. Let further (x^*, y^*) be neutrally stable. Then with [78, theorem 2, p. 130] both eigenvalues of $J(x^*, y^*)$ have (the same) non-positive real part. With $\mathrm{Tr}(J(x^*, y^*)) = 0$, both eigenvalues are strictly imaginary and $|J(x^*, y^*)| \geq 0$ and hence $|J(x^*, y^*)| > 0$. With $|J(x^*, y^*)| > 0$ and proposition 5.2.2, (x^*, y^*) is an attracting mutation limit. □

For regular Nash equilibria the previous proposition yields a characterization of attracting mutation limits:

Corollary 5.2.4. *A regular interior Nash equilibrium,* (x^*, y^*)*, is an attracting mutation limit if and only if* $|J(x^*, y^*)| > 0$*.*

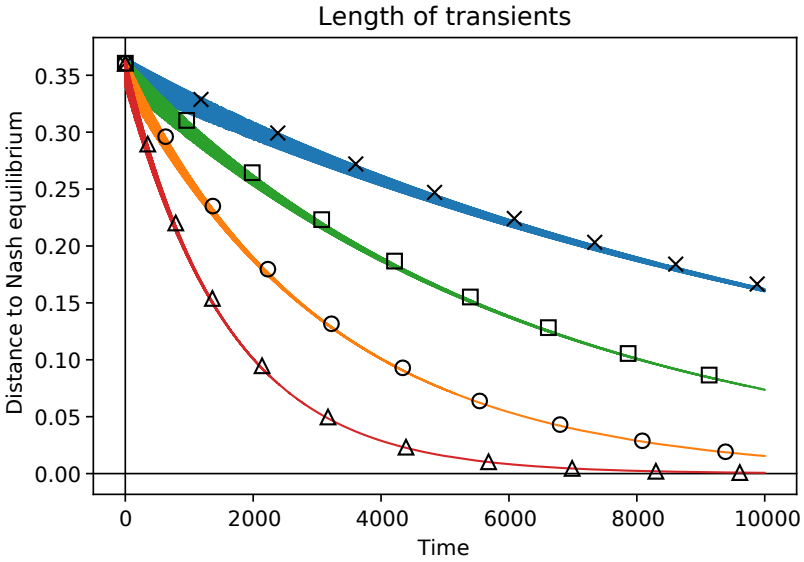
Proof. One direction follows directly from proposition 5.2.3. For the other direction let $|J(x^*, y^*)| < 0$. Then the eigenvalues must be real and with $\mathrm{Tr}(J(x^*, y^*)) = 0$, we have that one eigenvalue must be positive. Hence (x^*, y^*) is unstable and cannot be an attracting mutation limit owing to lemma 4.4.11. □

Example 5.2.5. Let us revisit the matching pennies game from example 4.4.10, with payoffs given as:

$$\begin{pmatrix} (1,0) & (0,1) \\ (0,1) & (1,0) \end{pmatrix}$$

It is clear that, after reducing to two dimensions, the unique Nash equilibrium (x^*, y^*) of the game is located at $(1/2,\ 1/2)$, where we give the strategy of the row player first. At the Nash equilibrium (x^*, y^*), the Jacobian has a positive determinant, specifically, $|J(x^*, y^*)| = 1/4$. It is then clear that (x^*, y^*) is an attracting mutation limit, i.e., mutation stabilises (x^*, y^*). However, the strength of the asymptotic stability depends on the mutation strength M, since M provides the bounds for the exponential stability of the linearisation. This implies that solutions take longer to approach the mutation limit for lower mutation strengths, returning to periodic orbits in the limit of $M \to 0$, as illustrated in figure 5.1.

Figure 5.1: Four solutions of (5.2.1) for the matching pennies game in example 5.2.5 with different mutation strengths $M_X = M_Y = M$. The coloured lines show the distance of the solution to the Nash equilibrium $(1/2, 1/2)$, with red, orange, green and blue lines corresponding to solutions for M-values of $2^{-4} \cdot 10^{-2}$, $2^{-5} \cdot 10^{-2}$, $2^{-6} \cdot 10^{-2}$ and $2^{-7} \cdot 10^{-2}$, respectively. Lines appear thicker at the start as the distance fluctuates stronger. The shapes '△', '○', '□' and '×' show the values of the exponential decay function $C_0 \cdot e^{-M \cdot t}$ with appropriate scaling constant C_0 for corresponding M-values of $2^{-4} \cdot 10^{-2}$, $2^{-5} \cdot 10^{-2}$, $2^{-6} \cdot 10^{-2}$ and $2^{-7} \cdot 10^{-2}$, respectively.



5.3 Antagonistic co-evolution: two-population zero-sum games

For convenience, in the situation of two-population zero-sum games, we use the more general notation, as introduced in chapter 4, i.e., with populations and types numbered. We consider the case of two populations where the game is a rescaled two-player zero-sum game, i.e., there is $\alpha > 0$ such that $\bar{f}_1(x) = -\alpha \bar{f}_2(x)$ for all $x \in \Delta$. We deal with this situation by considering the unreduced system of equations:

$$\dot{x}_{ih} = x_{ih}\left(f_{ih}(x) - \sum_{k \le n_i} x_{ik} f_{ik}(x)\right) =: \varphi_{ih}(x) \tag{RD}$$

Note that we set $\bar{f}_i(x) = \sum_{k \le n_i} x_{ik} f_{ik}(x)$ and hence $\partial_{ih} \bar{f}_i(x) = f_{ih}(x)$, which follows from $\partial_{ih} f_{ik}(x) = 0$ $(\forall i, h, k)$, where we write ∂_{ih} for $\frac{\partial}{\partial x_{ih}}$. For the Jacobian J_φ of φ, we have:

$$[J_\varphi(x)]_{ih,jk} = \partial_{jk}\varphi_{ih}(x) = x_{ih}\left(\partial_{jk} f_{ih}(x) - \partial_{jk}\bar{f}_i(x)\right) + \delta_{ij}\delta_{hk}\left(f_{ih}(x) - \bar{f}_i(x)\right),$$

where δ denotes the Kronecker delta.

The main objective of this section will be to show that a mutation equilibrium x^M, i.e., an equilibrium of (RMD), is asymptotically stable in antagonistic games. The

proofs will proceed as follows: The main lemma, lemma 5.3.1, will give an estimate for the real parts of the eigenvalues of the Jacobian of (RMD) at x^M, implying negative real parts, for all (non-zero) eigenvectors in the tangent space of Δ at $x^M \in \Delta^o$, i.e., for all eigenvectors in $T\Delta := \left\{\xi \in \mathbb{R}^S \;\middle|\; \forall i : \sum_{h \le n_i} \xi_{ih} = 0 \text{ and } \xi_i \neq 0\right\}$.

We then proceed to show that this property is sufficient for asymptotic stability, by considering an extension of (RMD) that is defined on a neighbourhood of Δ and coincides with (RMD) on Δ. For this extension, we prove that x^M is asymptotically stable by considering the eigenvalues of the Jacobian of that extension, where we prove some auxiliary lemmas to provide a rigorous proof. From these, the main proposition, proposition 5.3.8, of this section follows.

Lemma 5.3.1. *For all $\xi \in T\Delta$, we have*

$$\begin{aligned}&\xi^T(\mathfrak{A}(J_\varphi(x^M) - \mathfrak{M}) + (J_\varphi(x^M) - \mathfrak{M})^T\mathfrak{A}^T)\xi \\ &\quad = -2\left(M_1 \sum_{h \le n_1} (x^M_{1h})^{-2} c_{1h}\xi^2_{1h} + \alpha M_2 \sum_{h \le n_2} (x^M_{2h})^{-2} c_{2h}\xi^2_{2h}\right) < 0\,,\end{aligned}$$

where the diagonal matrix $\mathfrak{A}$ is defined as $[\mathfrak{A}]_{ih,jk} = \delta_{ij}\delta_{hk}(x_{ih})^{-1}\alpha^{\delta_{i2}}$, and the diagonal matrix $\mathfrak{M}$ is defined as $[\mathfrak{M}]_{ih,jk} = \delta_{ij}\delta_{hk}M_i$, with δ denoting the Kronecker delta.

Proof. For the proof we first consider $\xi^T(\mathfrak{A}J_\varphi(x^M) + J_\varphi(x^M)^T\mathfrak{A}^T)\xi$ by establishing the entries of $(\mathfrak{A}J_\varphi(x) + J_\varphi(x)^T\mathfrak{A}^T)$ for general $x \in \Delta^o$, taking into account its block structure, and subsequently considering x^M.

We start by noting that with $[\mathfrak{A}]_{ih,jk} = \delta_{ij}\delta_{hk}(x_{ih})^{-1}\alpha^{\delta_{i2}}$, we have that

$$[\mathfrak{A}J_\varphi(x)]_{ih,jk} = \alpha^{\delta_{i2}}(\partial_{jk}f_{ih}(x) - \partial_{jk}\bar{f}_i(x)) + \delta_{ij}\delta_{hk}(x_{ih})^{-1}\alpha^{\delta_{i2}}(f_{ih}(x) - \bar{f}_i(x))$$

and further

$$\begin{aligned}[\mathfrak{A}J_\varphi(x) + J^T_\varphi(x)\mathfrak{A}^T]_{ih,jk} &= [\mathfrak{A}J_\varphi(x)]_{ih,jk} + [\mathfrak{A}J_\varphi(x)]_{jk,ih} \\ &= \alpha^{\delta_{i2}}(\partial_{jk}f_{ih}(x) - \partial_{jk}\bar{f}_i(x)) + \delta_{ij}\delta_{hk}(x_{ih})^{-1}\alpha^{\delta_{i2}}(f_{ih}(x) - \bar{f}_i(x)) \\ &\quad + \alpha^{\delta_{j2}}(\partial_{ih}f_{jk}(x) - \partial_{ih}\bar{f}_j(x)) + \delta_{ij}\delta_{hk}(x_{jk})^{-1}\alpha^{\delta_{j2}}(f_{jk}(x) - \bar{f}_j(x)) \\ &= \alpha^{\delta_{i2}}(\partial_{jk}f_{ih}(x) - \partial_{jk}\bar{f}_i(x)) + \alpha^{\delta_{j2}}(\partial_{ih}f_{jk}(x) - \partial_{ih}\bar{f}_j(x)) \\ &\quad + 2\delta_{ij}\delta_{hk}(x_{ih})^{-1}\alpha^{\delta_{i2}}(f_{ih}(x) - \bar{f}_i(x))\,.\end{aligned}$$

The matrix $(\mathfrak{A}J_\varphi(x) + J^T_\varphi(x)\mathfrak{A}^T)$ has a block structure with two symmetric block matrices on the diagonal, where $i = j$, and a block matrix on the upper-right with

a transposed copy on the lower-left respectively, where $i \neq j$. We treat these blocks separately:

1) $i = j = 1$:

$$\begin{aligned}
[\mathfrak{A} J_\varphi(x) + J_\varphi^T(x)\mathfrak{A}^T]_{1h,1k} &= \alpha^{\delta_{12}}(\underbrace{\partial_{1k} f_{1h}(x)}_{=0} - \partial_{1k}\bar{f}_1(x)) \\
&\quad + \alpha^{\delta_{12}}(\underbrace{\partial_{1h} f_{1k}(x)}_{=0} - \partial_{1h}\bar{f}_1(x)) + 2\delta_{11}\delta_{hk}(x_{1h})^{-1}\alpha^{\delta_{12}}(f_{1h}(x) - \bar{f}_1(x)) \\
&= -\partial_{1k}\bar{f}_1(x) - \partial_{1h}\bar{f}_1(x) + 2\delta_{hk}(x_{1h})^{-1}(f_{1h}(x) - \bar{f}_1(x)) \\
&= -f_{1k}(x) - f_{1h}(x) + 2\delta_{hk}(x_{1h})^{-1}(f_{1h}(x) - \bar{f}_1(x))
\end{aligned}$$

2) $i = j = 2$:

$$\begin{aligned}
[\mathfrak{A} J_\varphi(x) + J_\varphi^T(x)\mathfrak{A}^T]_{2h,2k} &= \alpha^{\delta_{22}}(\underbrace{\partial_{2k} f_{2h}(x)}_{=0} - \partial_{2k}\bar{f}_2(x)) \\
&\quad + \alpha^{\delta_{22}}(\underbrace{\partial_{2h} f_{2k}(x)}_{=0} - \partial_{2h}\bar{f}_2(x)) + 2\delta_{22}\delta_{hk}(x_{2h})^{-1}\alpha^{\delta_{22}}(f_{2h}(x) - \bar{f}_2(x)) \\
&= \alpha(-\partial_{2k}\bar{f}_2(x) - \partial_{2h}\bar{f}_2(x) + 2\delta_{hk}(x_{2h})^{-1}(f_{2h}(x) - \bar{f}_2(x))) \\
&= \alpha(-f_{2k}(x) - f_{2h}(x) + 2\delta_{hk}(x_{2h})^{-1}(f_{2h}(x) - \bar{f}_2(x)))
\end{aligned}$$

3) $i = 1, j = 2$:

$$\begin{aligned}
&[\mathfrak{A} J_\varphi(x) + J_\varphi^T(x)\mathfrak{A}^T]_{1h,2k} \\
&\quad = \alpha^{\delta_{12}}(\partial_{2k} f_{1h}(x) - \partial_{2k}\bar{f}_1(x)) + \alpha^{\delta_{22}}(\partial_{1h} f_{2k}(x) - \partial_{1h}\bar{f}_2(x)) \\
&\quad = (\partial_{2k} f_{1h}(x) - \partial_{2k}\bar{f}_1(x)) + \alpha(\partial_{1h} f_{2k}(x) - \partial_{1h}\bar{f}_2(x)) \\
&\quad = (\partial_{2k}\partial_{1h}\bar{f}_1(x) - \partial_{2k}\underbrace{\bar{f}_1(x)}_{=-\alpha\bar{f}_2(x)}) + (\partial_{1h}\partial_{2k}\alpha\bar{f}_2(x) - \partial_{1h}\alpha\bar{f}_2(x)) \\
&\quad = (\partial_{2k}\partial_{1h}\bar{f}_1(x) + \alpha\partial_{2k}\bar{f}_2(x)) + (-\partial_{1h}\partial_{2k}\bar{f}_1(x) + \partial_{1h}\bar{f}_1(x)) \\
&\quad = \partial_{2k}\partial_{1h}\bar{f}_1(x) + \alpha\partial_{2k}\bar{f}_2(x) - \partial_{1h}\partial_{2k}\bar{f}_1(x) + \partial_{1h}\bar{f}_1(x) \\
&\quad = \alpha\partial_{2k}\bar{f}_2(x) + \partial_{1h}\bar{f}_1(x)
\end{aligned}$$

4) $i = 2, j = 1$:

$$\begin{aligned}
&[\mathfrak{A} J_\varphi(x) + J_\varphi^T(x)\mathfrak{A}^T]_{2h,1k} \\
&\quad = \alpha^{\delta_{22}}(\partial_{1k}\partial_{2h}\bar{f}_2(x) - \partial_{1k}\bar{f}_2(x)) + \alpha^{\delta_{12}}(\partial_{2h}\partial_{1k}\bar{f}_1(x) - \partial_{2h}\bar{f}_1(x)) \\
&\quad = \alpha(\partial_{1k}\partial_{2h}\bar{f}_2(x) - \partial_{1k}\bar{f}_2(x)) + (\partial_{2h}\partial_{1k}\bar{f}_1(x) - \partial_{2h}\bar{f}_1(x)) \\
&\quad = \alpha\partial_{1k}\partial_{2h}\bar{f}_2(x) + \partial_{1k}\bar{f}_1(x) - \alpha\partial_{2h}\partial_{1k}\bar{f}_2(x) + \alpha\partial_{2h}\bar{f}_2(x) \\
&\quad = \alpha\partial_{2h}\bar{f}_2(x) + \partial_{1k}\bar{f}_1(x)
\end{aligned}$$

Next, we consider that we are interested in the values at x^M where we have the additional property that $\varphi_{ih}^M(x^M) = 0$ and hence

$$f_{ih}(x^M) = (x_{ih}^M)^{-1} M_i (x_{ih}^M - c_{ih}) + \bar{f}_i(x^M) \quad (\forall i,h)\,.$$

Substituting this yields the following for the entries of the four block matrices:

1) For $i = j = 1$, we have:

$$\begin{aligned}[\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{1h,1k} &= -((x_{1k}^M)^{-1} M_1 (x_{1k}^M - c_{1k}) + \bar{f}_1(x^M)) \\ &- ((x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h}) + \bar{f}_1(x^M)) + 2\delta_{hk}(x_{1h}^M)^{-2} M_1 (x_{1h}^M - c_{1h})\end{aligned}$$

2) For $i = j = 2$, we have:

$$\begin{aligned}[\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{2h,2k} &= -\alpha((x_{2k}^M)^{-1} M_2 (x_{2k}^M - c_{2k}) + \bar{f}_2(x^M)) \\ &- \alpha((x_{2h}^M)^{-1} M_2 (x_{2h}^M - c_{2h}) + \bar{f}_2(x^M)) + \alpha 2\delta_{hk}(x_{2h}^M)^{-2} M_2 (x_{2h}^M - c_{2h})\end{aligned}$$

3) For $i = 1, j = 2$, we have:

$$\begin{aligned}[\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{1h,2k} &= \alpha \partial_{2k} \bar{f}_2(x^M) + \partial_{1h} \bar{f}_1(x^M) \\ &= \alpha f_{2k}(x^M) - \alpha \bar{f}_2(x^M) + \underbrace{\alpha \bar{f}_2(x^M)}_{=-\bar{f}_1(x^M)} + f_{1h}(x^M) \\ &= \alpha (x_{2k}^M)^{-1} M_2 (x_{2k}^M - c_{2k}) + (x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h})\end{aligned}$$

4) For $i = 2, j = 1$, we have similarly:

$$\begin{aligned}[\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{2h,1k} &= \alpha \partial_{2h} \bar{f}_2(x^M) + \partial_{1k} \bar{f}_1(x^M) \\ &= \alpha (x_{2h}^M)^{-1} M_2 (x_{2h}^M - c_{2h}) + (x_{1k}^M)^{-1} M_1 (x_{1k}^M - c_{1k})\end{aligned}$$

We can now consider $(\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T)\xi$ for $\xi \in T\Delta$:

$$\begin{aligned}[(\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T)\xi]_{ih} &= \sum_j \sum_{k \le n_j} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{ih,jk}\xi_{jk} \\ &= \sum_{k \le n_1} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{ih,1k}\xi_{1k} + \sum_{k \le n_2} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{ih,2k}\xi_{2k}\end{aligned}$$

and considering $i = 1$ and $i = 2$ separately, we have for $i = 1$,

$$\begin{aligned}[(\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T)\xi]_{1h} &= \sum_j \sum_{k \le n_j} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{1h,jk}\xi_{jk} \\ &= \sum_{k \le n_1} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{1h,1k}\xi_{1k} + \sum_{k \le n_2} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M)\mathfrak{A}^T]_{1h,2k}\xi_{2k} \\ &= \sum_{k \le n_1} \Big(- \big((x_{1k}^M)^{-1} M_1 (x_{1k}^M - c_{1k}) + \bar{f}_1(x^M)\big)\, \xi_{1k}\end{aligned}$$

$$
\begin{aligned}
&- \left((x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h}) + \bar{f}_1(x^M) \right) \xi_{1k} + 2\delta_{hk} (x_{1h}^M)^{-2} M_1 (x_{1h}^M - c_{1h}) \xi_{1k} \Big) \\
&+ \sum_{k \le n_2} \left(\alpha (x_{2k}^M)^{-1} M_2 (x_{2k}^M - c_{2k}) \xi_{2k} + (x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h}) \xi_{2k} \right) \\
= &- M_1 \sum_{k \le n_1} \frac{x_{1k}^M - c_{1k}}{x_{1k}^M} \xi_{1k} + \bar{f}_1(x^M) \underbrace{\sum_{k \le n_1} \xi_{1k}}_{=0} \\
&- \Big((x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h}) + \bar{f}_1(x^M) \Big) \underbrace{\sum_{k \le n_1} \xi_{1k}}_{=0} \\
&+ 2 (x_{1h}^M)^{-2} M_1 (x_{1h}^M - c_{1h}) \xi_{1h} \\
&+ \alpha M_2 \sum_{k \le n_2} \frac{x_{2k}^M - c_{2k}}{x_{2k}^M} \xi_{2k} + \Big((x_{1h}^M)^{-1} M_1 (x_{1h}^M - c_{1h}) \Big) \underbrace{\sum_{k \le n_2} \xi_{2k}}_{=0} \\
= &\, 2 M_1 (x_{1h}^M)^{-2} (x_{1h}^M - c_{1h}) \xi_{1h} - M_1 \sum_{k \le n_1} \frac{x_{1k}^M - c_{1k}}{x_{1k}^M} \xi_{1k} + \alpha M_2 \sum_{k \le n_2} \frac{x_{2k}^M - c_{2k}}{x_{2k}^M} \xi_{2k}
\end{aligned}
$$

and in a similar manner for $i = 2$

$$
\begin{aligned}
[(\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M) \mathfrak{A}^T) \xi]_{2h} &= \sum_j \sum_{k \le n_j} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M) \mathfrak{A}^T]_{2h,jk} \xi_{jk} \\
= &\sum_{k \le n_1} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M) \mathfrak{A}^T]_{2h,1k} \xi_{1k} + \sum_{k \le n_2} [\mathfrak{A} J_\varphi(x^M) + J_\varphi^T(x^M) \mathfrak{A}^T]_{2h,2k} \xi_{2k} \\
= &\sum_{k \le n_1} \left(\alpha M_2 (x_{2h}^M)^{-1} (x_{2h}^M - c_{2h}) + M_1 (x_{1k}^M)^{-1} (x_{1k}^M - c_{1k}) \right) \xi_{1k} \\
&- \alpha \sum_{k \le n_2} \Big(\left((x_{2k}^M)^{-1} M_2 (x_{2k}^M - c_{2k}) + \bar{f}_2(x^M) \right) \\
&\qquad + \left((x_{2h}^M)^{-1} M_2 (x_{2h}^M - c_{2h}) + \bar{f}_2(x^M) \right) \Big) \xi_{2k} \\
&+ \alpha \sum_{k \le n_2} \left(2\delta_{hk} (x_{2h}^M)^{-2} M_2 (x_{2h}^M - c_{2h}) \right) \xi_{2k} \\
= &\, \alpha M_2 (x_{2h}^M)^{-1} (x_{2h}^M - c_{2h}) \underbrace{\sum_{k \le n_1} \xi_{1k}}_{=0} + M_1 \sum_{k \le n_1} \frac{x_{1k}^M - c_{1k}}{x_{1k}^M} \xi_{1k} \\
&- \alpha \Big(M_2 \sum_{k \le n_2} \frac{x_{2k}^M - c_{2k}}{x_{2k}^M} \xi_{2k} + \bar{f}_2(x^M) \underbrace{\sum_{k \le n_2} \xi_{2k}}_{=0} \Big) \\
&- \alpha \Big((x_{2h}^M)^{-1} M_2 (x_{2h}^M - c_{2h}) + \bar{f}_2(x^M) \Big) \underbrace{\sum_{k \le n_2} \xi_{2k}}_{=0} + 2\alpha M_2 (x_{2h}^M)^{-2} (x_{2h}^M - c_{2h}) \xi_{2h} \\
= &\, 2\alpha M_2 (x_{2h}^M)^{-2} (x_{2h}^M - c_{2h}) \xi_{2h} + M_1 \sum_{k \le n_1} \frac{x_{1k}^M - c_{1k}}{x_{1k}^M} \xi_{1k} - \alpha M_2 \sum_{k \le n_2} \frac{x_{2k}^M - c_{2k}}{x_{2k}^M} \xi_{2k}
\end{aligned}
$$

With the above considerations we can finally consider:

$$\xi^T(\mathfrak{A}J_\varphi(x^M)+J_\varphi^T(x^M)\mathfrak{A}^T)\xi=\sum_i\sum_{h\le n_i}\xi_{ih}[(\mathfrak{A}J_\varphi(x^M)+J_\varphi^T(x^M)\mathfrak{A}^T)\xi]_{ih}$$

$$=\sum_{h\le n_1}\xi_{1h}[(\mathfrak{A}J_\varphi(x^M)+J_\varphi^T(x^M)\mathfrak{A}^T)\xi]_{1h}+\sum_{h\le n_2}\xi_{2h}[(\mathfrak{A}J_\varphi(x^M)+J_\varphi^T(x^M)\mathfrak{A}^T)\xi]_{2h}$$

$$=\sum_{h\le n_1}\xi_{1h}\Big(2M_1(x^M_{1h})^{-2}(x^M_{1h}-c_{1h})\xi_{1h}$$

$$-M_1\sum_{k\le n_1}\frac{x^M_{1k}-c_{1k}}{x^M_{1k}}\xi_{1k}+\alpha M_2\sum_{k\le n_2}\frac{x^M_{2k}-c_{2k}}{x^M_{2k}}\xi_{2k}\Big)$$

$$+\sum_{h\le n_2}\xi_{2h}\Big(2\alpha M_2(x^M_{2h})^{-2}(x^M_{2h}-c_{2h})\xi_{2h}$$

$$+M_1\sum_{k\le n_1}\frac{x^M_{1k}-c_{1k}}{x^M_{1k}}\xi_{1k}-\alpha M_2\sum_{k\le n_2}\frac{x^M_{2k}-c_{2k}}{x^M_{2k}}\xi_{2k}\Big)$$

$$=\Big(2M_1\sum_{h\le n_1}\xi_{1h}(x^M_{1h})^{-2}(x^M_{1h}-c_{1h})\xi_{1h}$$

$$-M_1\sum_{k\le n_1}\frac{x^M_{1k}-c_{1k}}{x^M_{1k}}\xi_{1k}\underbrace{\sum_{h\le n_1}\xi_{1h}}_{=0}+\alpha M_2\sum_{k\le n_2}\frac{x^M_{2k}-c_{2k}}{x^M_{2k}}\xi_{2k}\underbrace{\sum_{h\le n_1}\xi_{1h}}_{=0}\Big)$$

$$+\Big(2\alpha M_2\sum_{h\le n_2}\xi_{2h}(x^M_{2h})^{-2}(x^M_{2h}-c_{2h})\xi_{2h}$$

$$+M_1\sum_{k\le n_1}\frac{x^M_{1k}-c_{1k}}{x^M_{1k}}\xi_{1k}\underbrace{\sum_{h\le n_2}\xi_{2h}}_{=0}-\alpha M_2\sum_{k\le n_2}\frac{x^M_{2k}-c_{2k}}{x^M_{2k}}\xi_{2k}\underbrace{\sum_{h\le n_2}\xi_{2h}}_{=0}\Big)$$

$$=2M_1\sum_{h\le n_1}(x^M_{1h})^{-2}(x^M_{1h}-c_{1h})\xi^2_{1h}+2\alpha M_2\sum_{h\le n_2}(x^M_{2h})^{-2}(x^M_{2h}-c_{2h})\xi^2_{2h}$$

Consider now for $\mathfrak{M}_{ih,jk}=\delta_{ij}\delta_{hk}M_i$ and ξ as above:

$$\xi^T\Big(\mathfrak{A}(J_\varphi(x^M)-\mathfrak{M})+(J_\varphi(x^M)-\mathfrak{M})^T\mathfrak{A}^T\Big)\xi$$

$$=\xi^T\Big(\mathfrak{A}J_\varphi(x^M)-\mathfrak{A}\mathfrak{M}+J_\varphi^T(x^M)\mathfrak{A}^T-\mathfrak{M}^T\mathfrak{A}^T\Big)\xi$$

$$=\xi^T\Big(\mathfrak{A}J_\varphi(x^M)+J_\varphi^T(x^M)\mathfrak{A}^T\Big)\xi-2\xi^T\mathfrak{A}\mathfrak{M}\xi$$

$$=2M_1\sum_{h\le n_1}(x^M_{1h})^{-2}(x^M_{1h}-c_{1h})\xi^2_{1h}+2\alpha M_2\sum_{h\le n_2}(x^M_{2h})^{-2}(x^M_{2h}-c_{2h})\xi^2_{2h}$$

$$-2M_1\sum_{h\le n_1}(x^M_{1h})^{-1}\xi^2_{1h}-2\alpha M_2\sum_{h\le n_2}(x^M_{2h})^{-1}\xi^2_{2h}$$

$$=-2\Big(M_1\sum_{h\le n_1}(x^M_{1h})^{-2}c_{1h}\xi^2_{1h}+\alpha M_2\sum_{h\le n_2}(x^M_{2h})^{-2}c_{2h}\xi^2_{2h}\Big)\quad<0$$

because at least for some i,h we have $\xi_{ih}\neq 0$ and $x^M,c\in\Delta^o$. □

For completeness' sake, we state the following simple, general lemma which, together with the previous lemma provides an estimate on the real parts of certain eigenvalues:

Lemma 5.3.2. *Let $\mathcal{S} \in \mathbb{R}^{n\times n}$ be a positive definite, symmetric matrix, and $A \in \mathbb{R}^{n\times n}$ a real matrix. If for all $x \in \mathbb{R}^n \setminus \{0\}$, $\langle \mathcal{S}Ax, x\rangle < 0$, then $\Re(\lambda) < 0$ for all $\lambda \in \sigma(A)$.*

Proof. Let $\lambda \in \sigma(A)$ be an eigenvalue of A and let $z \in \mathbb{C}^n \setminus \{0\}$ be a corresponding eigenvector with $z = x+iy$ for suitable $x, y \in \mathbb{R}^n$. Note that $\mathcal{S}$ induces an inner product $(x,y) \mapsto \langle x,y\rangle_{\mathcal{S}} := \langle \mathcal{S}^{\frac{1}{2}}x, \mathcal{S}^{\frac{1}{2}}y\rangle$ on $\mathbb{C}^n$ and that $\langle \mathcal{S}^{\frac{1}{2}}x, \mathcal{S}^{\frac{1}{2}}y\rangle = \langle \mathcal{S}x, y\rangle$, where $\mathcal{S}^{\frac{1}{2}}$ denotes the matrix such that $\mathcal{S}^{\frac{1}{2}}\mathcal{S}^{\frac{1}{2}} = \mathcal{S}$. W.l.o.g. we can assume $\langle z,z\rangle_{\mathcal{S}} = 1$. Then we have:

$$\begin{aligned}\Re(\lambda) &= \Re(\lambda\langle z,z\rangle_{\mathcal{S}}) = \Re(\langle \lambda z, z\rangle_{\mathcal{S}}) = \Re(\langle Az, z\rangle_{\mathcal{S}}) = \Re(\langle \mathcal{S}Az, z\rangle)\\ &= \Re(\langle \mathcal{S}Ax, x\rangle - i\langle \mathcal{S}Ax, y\rangle + i\langle \mathcal{S}Ay, x\rangle + \langle \mathcal{S}Ay, y\rangle)\\ &= \langle \mathcal{S}Ax, x\rangle + \langle \mathcal{S}Ay, y\rangle < 0\,.\end{aligned}$$

□

Before we consider the extended dynamical system, we state the following intuitively clear lemma stating that the tangent space is invariant under the Jacobian:

Lemma 5.3.3. *$T\Delta$ is invariant under $J_\varphi(x^M) - \mathfrak{M}$, i.e., for all $\eta \in T\Delta$ we have that $(J_\varphi(x^M) - \mathfrak{M})\eta \in T\Delta$.*

Proof. We consider $J_\varphi(x)\eta$ and $\mathfrak{M}\eta$ separately. Then for any i

$$\begin{aligned}\sum_{h\le n_i}[J_\varphi(x)\eta]_{ih} &= \sum_{h\le n_i}\sum_j\sum_{k\le n_j}\partial_{jk}\varphi_{ih}(x)\eta_{jk} = \sum_{h\le n_i}\nabla\varphi_{ih}^T(x)\eta\\ &= \sum_{h\le n_i}\lim_{t\to 0}\frac{\varphi_{ih}(x+t\eta)-\varphi_{ih}(x)}{t} = \lim_{t\to 0}\frac{1}{t}\Big(\underbrace{\sum_{h\le n_i}\varphi_{ih}(x+t\eta)}_{=0} - \underbrace{\sum_{h\le n_i}\varphi_{ih}(x)}_{=0}\Big) = 0\end{aligned}$$

and further

$$\sum_{h\le n_i}[\mathfrak{M}\eta]_{ih} = \sum_{h\le n_i}\sum_j\sum_{k\le n_j}\delta_{ij}\delta_{hk}M_i\eta_{jk} = M_i\sum_{h\le n_i}\eta_{ih} = 0.$$

From this the claim follows immediately. □

We now consider an extension of (RMD) to an open neighbourhood $U \supset \Delta$ of Δ. For i, define $s_i : x \in U \mapsto (\sum_{h\le n_i} x_{ih})^{-1} \in \mathbb{R}$, and $w_{ih} : x \in U \mapsto x_{ih}s_i(x)$ for $h \in S_i$. It is clear that there is such a neighbourhood U such that these functions are well-defined. Note that $s_i(\alpha x) = \alpha^{-1}s_i(x)$ and $w(\alpha x) = w(x)$, i.e., the functions are homogeneous

of degree -1 and 0 respectively, and in particular for $x \in \Delta$, we have $s_i(x) = 1$ and $w(x) = x$. We define now the following system for $(i,h) \in S$:

$$\dot{x}_{ih} = \rho_{ih}(x) := s_i(x)\varphi_{ih}(w(x)) + s_i(x)M_i(c_{ih} - x_{ih}) \tag{5.3.1}$$

It is clear that $\rho_{ih}(x) = \varphi^M_{ih}(x)$ for all $x \in \Delta$. It is further easy to see that the Jacobian J_ρ coincides with that of (RMD) on Δ^o:

Lemma 5.3.4. *For $x \in \Delta^o$ and $\xi \in T\Delta$, the Jacobian $J_\rho(x)$ satisfies $J_\rho(x)\xi = (J_\varphi(x) - \mathfrak{M})\xi$.*

Proof. Note that $x + t\xi \in \Delta$ for t sufficiently small. Hence, $\rho(x+t\xi) = \varphi^M(x+t\xi)$ and thus $J_\rho(x)\xi = (J_\varphi(x) - \mathfrak{M})\xi$. □

For $J_\rho(x^M)$ (and hence $(J_\varphi(x^M) - \mathfrak{M})$), we can provide explicitly the eigenvectors not contained in the tangent space of Δ^o at x^M:

Lemma 5.3.5. *For every i and a mutation equilibrium x^M, the vector $\mathfrak{z}^{M,i}$, where $[\mathfrak{z}^{M,i}]_{jk} = x^M_{jk}\delta_{ij}$ is an eigenvector of the Jacobian $J_\rho(x^M)$ of ρ, with eigenvalue $-M_i$.*

Proof. Let $\mathfrak{z}^{M,i}$ be as stated and $t > 0$. Note that

$$s_j(x^M + t\mathfrak{z}^{M,i}) = \left(\sum_{k\le n_j} x^M_{jk} + tx^M_{jk}\delta_{ij}\right)^{-1} = \left(1 + t\delta_{ij}\right)^{-1}\left(\sum_{k\le n_j} x^M_{jk}\right)^{-1} = \left(1 + t\delta_{ij}\right)^{-1}$$

$$\begin{aligned}\text{and} \qquad w_{jk}(x^M + t\mathfrak{z}^{M,i}) &= (x^M_{jk} + tx^M_{jk}\delta_{ij})s_j(x^M + t\mathfrak{z}^{M,i}) \\ &= x^M_{jk}(1 + t\delta_{ij})s_j(x^M + t\mathfrak{z}^{M,i}) = x^M_{jk}(1 + t\delta_{ij})(1 + t\delta_{ij})^{-1} = x^M_{jk}\end{aligned}$$

and hence $w(x^M + t\mathfrak{z}^{M,i}) = x^M$. Consider now:

$$\begin{aligned}&\nabla\rho_{jk}(x^M)^T\mathfrak{z}^{M,i} = \lim_{t\to 0}\frac{1}{t}\left(\rho_{jk}(x^M + t\mathfrak{z}^{M,i}) - \rho_{jk}(x^M)\right) \\ &= \lim_{t\to 0}\frac{1}{t}\Big(s_j(x^M + t\mathfrak{z}^{M,i})\varphi_{jk}(w(x^M + t\mathfrak{z}^{M,i})) \\ &\quad + s_j(x^M + t\mathfrak{z}^{M,i})M_j(c_{jk} - (x^M_{jk} + t[\mathfrak{z}^{M,i}]_{jk})) \\ &\quad - \left(s_j(x^M)\varphi_{jk}(w(x^M)) + s_j(x^M)M_j(c_{jk} - x^M_{jk})\right)\Big) \\ &= \lim_{t\to 0}\frac{1}{t}\Big((1 + t\delta_{ij})^{-1}\varphi_{jk}(x^M) + (1 + t\delta_{ij})^{-1}M_j(c_{jk} - x^M_{jk}(1 + t\delta_{ij})) \\ &\quad - (\varphi_{jk}(x^M) + M_j(c_{jk} - x^M_{jk}))\Big) \\ &= \lim_{t\to 0}\frac{1}{t}\left(\delta_{ij}\frac{1 - 1 - t}{1 + t}\varphi_{jk}(x^M) + \delta_{ij}\frac{1 - 1 - t}{1 + t}M_jc_{jk}\right) = \delta_{ij}\lim_{t\to 0}\frac{-1}{1 + t}\left(\varphi_{jk}(x^M) + M_jc_{jk}\right) \\ &= -\delta_{ij}(\varphi_{jk}(x^M) + M_jc_{jk}) = -M_j\delta_{ij}x^M_{jk} = -M_i[\mathfrak{z}^{M,i}]^M_{jk}\end{aligned}$$

Thus, we have $J_\rho(x^M)\mathfrak{z}^{M,i} = -M_i\mathfrak{z}^{M,i}$. □

The following lemma states that the vectors $\mathfrak{z}^{M,i}$ are orthogonal to $T\Delta$ under the operator $(\mathfrak{A}J_\rho(x^M) + (\mathfrak{A}J_\rho(x^M))^T)$, which will allow us to discard mixed terms in the proof of lemma 5.3.7:

Lemma 5.3.6. *Let $\mathfrak{z}^{M,i}$ be as above and $\xi \in T\Delta$. Then*

$$(\mathfrak{z}^{M,i})^T(\mathfrak{A}J_\rho(x^M) + (\mathfrak{A}J_\rho(x^M))^T)\xi = 0\,.$$

Proof. Note that:

$$(\mathfrak{z}^{M,i})^T(\mathfrak{A}J_\rho(x^M) + (\mathfrak{A}J_\rho(x^M))^T)\xi = (\mathfrak{z}^{M,i})^T(\mathfrak{A}J_\rho(x^M))\xi + (\mathfrak{A}J_\rho(x^M)\mathfrak{z}^{M,i})^T\xi$$

Now we have

$$\begin{aligned}(\mathfrak{A}J_\rho(x^M)\mathfrak{z}^{M,i})^T\xi &= -M_i(\mathfrak{A}\mathfrak{z}^{M,i})^T\xi = -M_i\sum_j\sum_{k\le n_j}\alpha^{\delta_{j2}}\frac{1}{x^M_{jk}}x^M_{jk}\delta_{ji}\xi_{jk}\\ &= -M_i\alpha^{\delta_{i2}}\sum_{k\le n_i}\xi_{ik} = 0\end{aligned}$$

due to the lemma 5.3.5, and

$$\begin{aligned}(\mathfrak{z}^{M,i})^T(\mathfrak{A}J_\rho(x^M))\xi &= \sum_j\sum_{k\le n_j}[\mathfrak{z}^{M,i}]_{jk}\frac{1}{x^M_{jk}}\alpha^{\delta_{j2}}[J_\rho(x^M)\xi]_{jk}\\ &= \sum_j\sum_{k\le n_j}\delta_{ji}\alpha^{\delta_{j2}}[J_\rho(x^M)\xi]_{jk} = \alpha^{\delta_{i2}}\sum_{k\le n_i}[(J_\varphi(x^M) - \mathfrak{M})\xi]_{ik} = 0\end{aligned}$$

where the second to last equality is due to lemma 5.3.4 and the last one due to lemma 5.3.3. This concludes the proof. □

We can now prove the main lemma:

Lemma 5.3.7. *For all $x \in \mathbb{R}^n \setminus \{0\}$, we have*

$$x^T(J_\rho(x^M) + J_\rho(x^M)^T)x < 0\,.$$

In particular, all eigenvalues of $J_\rho(x^M)$ have strictly negative real parts.

Proof. Let $x \in \mathbb{R}^n \setminus \{0\}$ and set $s_i = \sum_{h\le n_i} x_{ih}$. Then set $\xi_{ih} = x_{ih} - s_i x^M_{ih}$ such that $x_{ih} = s_i x^M_{ih} + \xi_{ih}$. Clearly, $\sum_{h\le n_i}\xi_{ih} = 0$ for all i. Overall, we have $x = \xi + \sum_i s_i\mathfrak{z}^{M,i}$, with $\mathfrak{z}^{M,i}$ as in lemma 5.3.5, and hence:

$$\begin{aligned}&x^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})x\\ &= \left(\xi + \sum_i s_i\mathfrak{z}^{M,i}\right)^T\left(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A}\right)\left(\xi + \sum_i s_i\mathfrak{z}^{M,i}\right)\end{aligned}$$

$$
\begin{aligned}
&= \xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\xi + 2\xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\left(\sum_i s_i \mathfrak{z}^{M,i}\right) \\
&\quad + \left(\sum_i s_i \mathfrak{z}^{M,i}\right)^T (\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\left(\sum_i s_i \mathfrak{z}^{M,i}\right) \\
&= \xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\xi + 2\sum_i s_i \underbrace{\xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\mathfrak{z}^{M,i}}_{=0 \text{ with lemma 5.3.6}} \\
&\quad + \sum_j \sum_i s_j s_i (\mathfrak{z}^{M,j})^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\mathfrak{z}^{M,i} \\
&= \xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\xi \\
&\quad + \sum_j \sum_i s_j s_i (\mathfrak{z}^{M,j})^T\mathfrak{A}J_\rho(x^M)\mathfrak{z}^{M,i} + s_j s_i (\mathfrak{z}^{M,j})^T J_\rho(x^M)^T\mathfrak{A}\mathfrak{z}^{M,i} \\
&= \xi^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})\xi + \sum_j \sum_i s_j s_i (\mathfrak{z}^{M,j})^T\mathfrak{A}\mathfrak{z}^{M,i}(-M_i - M_j) \\
&\overset{\text{Lemma } 5.3.4}{=} \xi^T(\mathfrak{A}(J_\varphi(x^M) - \mathfrak{M}) + (J_\varphi(x^M) - \mathfrak{M})^T\mathfrak{A})\xi - 2\sum_i M_i (s_i)^2 (\mathfrak{z}^{M,i})^T\mathfrak{A}\mathfrak{z}^{M,i}
\end{aligned}
$$

where the last equality follows from

$$
\begin{aligned}
(\mathfrak{z}^{M,j})^T\mathfrak{A}\mathfrak{z}^{M,i} &= \sum_l \sum_{p \le n_l} [\mathfrak{z}^{M,j}]_{lp} (x_{lp})^{-1} \alpha^{\delta_{l2}} [\mathfrak{z}^{M,i}]_{lp} \\
&= \sum_l \sum_{p \le n_l} x^M_{lp}\delta_{jl}(x^M_{lp})^{-1}\alpha^{\delta_{l2}}x^M_{lp}\delta_{il} = \alpha^{\delta_{i2}}\delta_{ij}\sum_{p \le n_i} x^M_{ip} = \alpha^{\delta_{i2}}\delta_{ij}
\end{aligned}
$$

Overall, we have together with lemma 5.3.1:

$$
\begin{aligned}
&x^T(\mathfrak{A}J_\rho(x^M) + J_\rho(x^M)^T\mathfrak{A})x \\
&\quad = \xi^T(\mathfrak{A}(J_\varphi(x^M) - \mathfrak{M}) + (J_\varphi(x^M) - \mathfrak{M})^T\mathfrak{A})\xi - 2\sum_i M_i (s_i)^2 (\mathfrak{z}^{M,i})^T\mathfrak{A}\mathfrak{z}^{M,i} \\
&\quad = \xi^T(\mathfrak{A}(J_\varphi(x^M) - \mathfrak{M}) + (J_\varphi(x^M) - \mathfrak{M})^T\mathfrak{A})\xi - 2\sum_i M_i (s_i)^2 \alpha^{\delta_{i2}} \\
&\quad = -2\Big(M_1 \sum_{h \le n_1} (x^M_{1h})^{-2} c_{1h}\xi^2_{1h} + \alpha M_2 \sum_{h \le n_2} (x^M_{2h})^{-2} c_{2h}\xi^2_{2h}\Big) \\
&\qquad - 2\sum_i M_i (s_i)^2 \alpha^{\delta_{i2}} < 0
\end{aligned}
$$

With lemma 5.3.2, we have that all eigenvalues of $J_\rho(x^M)$ have strictly negative real parts. □

From the above lemma, the following directly follows:

Proposition 5.3.8. *For all $M > 0$, every mutation equilibrium x^M is an asymptotically stable stationary point of* (RMD).

Proof. With lemma 5.3.7, the eigenvalues of the Jacobian $J_\rho(x^M)$ have strictly negative real parts. With, e.g., [78, theorem 1, p. 130], this implies that x^M is asymptotically stable under (5.3.1). With $x^M \in \Delta^o$ and the system (5.3.1) coinciding with (RMD) on Δ, we have that x^M is asymptotically stable under (RMD). □

Overall, this yields the following:

Proposition 5.3.9. *For every antagonistic two-population setting, all mutation limits are attracting and there is at least one attracting mutation limit.*

Proof. That all mutation limits are attracting, follows directly from proposition 5.3.8 together with the definition of attracting mutation limits. That there is at least one attracting mutation limit, then follows from the general existence of mutation limits (proposition 4.4.3). □

5.4 Discussion

We have investigated analytically the effect of mutation in the continuous time multi-population replicator dynamics. For 2×2 games we have shown that memoryless mutation always stabilises neutrally stable regular Nash equilibria. In particular, the results are general enough to include non-linear games, such as games with different interaction durations for different types, e.g., [103]. Effectively the slightest mutation probability affects the quality of antagonistic co-evolution in these situations and prevents cycling, with convergence being slower for lower mutation probabilities. For vanishing mutation, co-evolution converges extremely slowly to a Nash equilibrium, such that very weak mutation becomes almost indistinguishable from a pure replicator dynamics for short time intervals. The proofs relied on shifting the real parts of the Jacobian to the negative half-plane and in principle this approach might work for mutation matrices which are not diagonal. In our approach, we implicitly exploited the fact that our matrices did not affect the eigenspaces of the unperturbed Jacobian. With more general mutation, this would be an aspect requiring additional consideration.

For larger games, we have shown that in the antagonistic setting of rescaled zero-sum games, which includes constant-sum games, mutation again stabilises equilibria. Larger games are particularly interesting in that there we can consider the single-population replicator dynamics as a special case. Specifically, this is contained in the

setting where both populations have identical parameters and initial conditions. In this case, the dynamics evolves in a submanifold where populations remain identical, allowing us to identify the two with each other. Therefore, the case becomes identical to the setting where all interactions are with conspecifics—something we assume not to be the case for the multi-population setting. In this way, we can extend the notion of attracting mutation limits and the associated results to the single-population replicator dynamics. From this we can infer that the single-population RPS with mutation converges and that the mixed Nash equilibrium is an attracting mutation limit. However, although in this case mutation drives the system towards the mutation limit, we know that the mutation limit is not an ESS as it is not asymptotically stable. Thus, it is clear that the notions of ESS and attracting mutation limit are distinct and address different intuitions. It is further of interest whether the effects of mutation seen in the replicator dynamics extend to more complicated models, e.g., with variable population sizes as in [99], and whether the techniques employed here can be extended to those cases.

6 Replicator-mutator dynamics and mutation-bias learning

6.1 Introduction

Reinforcement learning algorithms have been employed in a wide range of problem settings with great success, e.g., [94], and for the single-agent case the conditions for convergence of, e.g., Q-learning have been clarified, [117]. However, for multi-agent reinforcement learning (MARL), questions of convergence are still very much open as a mathematically rigorous analysis becomes much more challenging. That even two-player settings can prove challenging for rigorous analysis is demonstrated by [89], which analyses the rich dynamics that can occur for the Rock-Paper-Scissors (RPS) game under the replicator dynamics. There is a vast array of different algorithms and an even greater array of problem settings, cf., [20]. In many cases, analysis beyond experimental evaluation is hardly possible. However, more general analysis is highly informative of why algorithms behave in a certain way and theoretical guarantees for at least the simplest of settings are highly desirable in order to assess the behaviour of MARL algorithms.

Building on the relation between the replicator dynamics and a simple form of reinforcement learning, called Cross learning [13, 27], we formulate two variants of a new reinforcement learning algorithm, called mutation-bias learning (MBL), and establish a relation between (RMD) introduced in chapter 4 and the stochastic processes induced by MBL. We demonstrate that the possibility of asymptotic stability of interior equilibria in (RMD) allows it to be used as an ordinary differential equation (ODE) approximation to MBL, in contrast to (RD), which diverges from its discrete approximations due to at most neutral stability of interior equilibria. While the process induced by Cross learning deviates from the solutions of (RD) for finite times in the case of neutral stability, globally asymptotically stable equilibria in (RMD) allow us to show that MBL processes revisit neighbourhoods of such equilibria infinitely often

almost surely, and hence the neighbourhoods of such mutation limits, showing MBL to be ε-rational with arbitrary certainty in such cases.

We illustrate these theoretical results with numerical experiments in a range of two-player games, as well as a three-player game, and compare the behaviours of the MBL variants to those of so-called Frequency-adjusted Q-learning (FAQ), [52], and Win-or-Learn-Fast Policy-Hill-Climbing (WoLF-PHC), [14], demonstrating the advantages of theoretical guarantees in the study of MARL algorithms. We further illustrate the trade-off between convergence and "rationality" caused by the mutation term of (RMD) and similar perturbations.

Game theory has proven to be a useful framework for the theoretical analysis of MARL algorithms. In particular, as MARL algorithms usually lead to (often stochastic) discrete time dynamic systems, the insights from the fields of learning dynamics in games and of evolutionary game theory have been particularly relevant. These fields further offer a wide range of approaches linking the stochastic discrete time dynamics of learning to a deterministic continuous time setting, e.g., [45, 87].

In particular, when learning algorithms are linked to a system of ODEs, questions of convergence can be addressed by considering the dynamic stability of equilibria in the ODE system, as e.g., in [24]. While Lyapunov stability and other properties of the continuous time case not always transfer to the discrete dynamics as is illustrated by the prominent example of the Rock-Paper-Scissors game, asymptotic stability in the continuous case can (under suitable conditions on algorithm parameters) imply the convergence of a MARL algorithm.

Evolutionary game theoretic approaches, and specifically relating (RD) to learning algorithms have informed a number of analyses of learning algorithms in multi-agent settings, e.g., [66, 74]. One of the earliest rigorous analyses of the relation between stochastic learning and (RD) is given in [13] and concerns Cross learning, [27]. A larger class of stochastic reinforcement learning rules is related to deterministic continuous time systems of (RD) type in [86]. Systems of (RD) type with additional perturbations have been related to various learning rules, including such with entropy related perturbation terms, [90], and exponential learning based on a logit model, [62]. Some analyses focus specifically on Q-learning based learning algorithms. For instance, [53] considers the stability and convergence properties of Q-learning in the two-player setting; however, the Q-values enter as expectations, not as random variables, and therefore the effects of stochasticity are not considered. A similar approach

is pursued in [109] with a corrected derivation given in [52]. However, both strands start from assumptions which have not been proved, and therefore no theoretical guarantees can be inferred.

An analysis of the convergence of multiple timescales algorithms, where Q-value estimates are learned quicker than policy changes occur, is given in [24]. Here, the convergence analysis relates to smoothed best-response dynamics. Furthermore, [22] gives conditions for convergence of an ε-greedy multi-agent Q-learning algorithm under stochastic payoffs. However, this algorithm operates on joint actions, which requires agents to be able to observe the actions chosen by all agents and therefore is distinct from the approaches mentioned above and from the algorithm we introduce.

The main contribution of the present approach is to demonstrate that the convergence problems of Cross learning can be addressed in a low complexity manner, namely mutation-bias learning with direct policy updates (MBL-DPU), that still allows a theoretical analysis while also showing behaviour similar to much more complex algorithms. For MBL-DPU, we prove converges to the replicator-mutator dynamics (RMD), given in chapter 4. There we have shown that (RMD) in principle allows non-strict Nash equilibria to be approximated by asymptotically stable mutation equilibria, which is particularly relevant for zero-sum games. Further, asymptotic stability in (RMD) allows MBL-DPU to reach ε-equilibria with arbitrary certainty. Thus, despite only slightly increasing the complexity of basic Cross learning, MBL-DPU allows for the solution of a much larger range of games.

Furthermore, the presented perturbative approach can be applied to a range of different algorithms, e.g., based on logistic choice. To this end, we formulate MBL-LC which follows the evolution of a policy under a logistic choice function (also known as Boltzmann policy or softmax policy). This formulation is parallel to the Q-learning based FAQ learning rule, [52].

6.2 Mutation-bias learning

6.2.1 Preliminaries

We consider the proposed algorithms in a basic game theoretical setting, where a stage game $(N, (\mathcal{A}_i)_{i\in N}, (r_i)_{i\in N})$ is repeatedly played by the players. However, the strategy space for the repeated game is not that of Markovian strategies, but simply the space of the mixed strategies of the stage game, the space of which we denote by

$\Delta = \times_{i \in N} \Delta_i$, as previously. In particular, we do not assume that players can react to past play in the current analysis.

As a reference frame for the MBL algorithm, we provide two MARL algorithms, specifically WoLF-PHC, [14], and FAQ learning, [52], both Q-learning based; the former, WoLF-PHC (algorithm 6.2.1), because it is a game theoretically informed algorithm, with a rigorous analysis available and based on direct policy search, thus making it one of the closest points of reference for MBL in the literature; the latter, FAQ learning (algorithm 6.2.2), because although no proofs are provided, it seems among the Q-learning based approaches the one most closely related to (RD) and to the ideas presented in [13] on which our approach is based.

The WoLF-PHC algorithm illustrates the potential of evolutionary game theory (EGT) to inform learning algorithms in so far as WoLF-PHC keeps track of the past average policy and compares the value of the current policy to that of the average policy. The time-average of a population has been shown to converge to a Nash equilibrium under the replicator dynamics in zero-sum games, e.g., [118, proposition 3.6, p.92] and comparing this to the current policy's payoff therefore makes sense from an EGT perspective.

While this might have been only a side-motivation for the formulation of WoLF-PHC, the FAQ learning algorithm is deliberately targeted at exploiting a relation between Q-learning and evolutionary dynamics. In [52], the authors claim that the FAQ learning algorithm generates trajectories which converge to the solutions of an (RD) type ODE system in probability.[1] Specifically, the system in question is

$$\dot{x}_{ih}(t) = \tau x_{ih} \left(\mathbb{E}[r_{ih}(t)] - \sum_{k \in \mathscr{A}_i} x_{ik}(t) \mathbb{E}[r_{ik}(t)] \right) + x_{ih} \left(\sum_{k \in \mathscr{A}_i} x_{ik} \ln(x_{ik}) - \ln(x_{ih}) \right) \tag{6.2.1}$$

for $i \in N, h \in \mathscr{A}_i$, such that the system consists of the usual (RD) part and a perturbative part that can be related to information entropy, elaborated upon in [52, 111]. The relative strengths of the replicator dynamics and the perturbative term are controlled by τ, such that τ^{-1} plays an analogous role to M in (RMD).

6.2.2 Mutation-bias learning algorithm

We can now formulate the stochastic learning rules and specify how they relate to the deterministic dynamics of (RMD). We consider two versions of MBL: one, based

[1] As the authors do not provide a proof for this claim, it should be taken with appropriate caution.

Algorithm 6.2.1 WoLF-PHC learning, [14], for generic player $i \in N$.

1. Let $\alpha \in (0,1]$, $\delta_l > \delta_w \in (0,1]$ be learning rates. Initialize $\forall h \in \mathscr{A}_i$

$$Q_{ih} \leftarrow 0, \quad x_{ih} \leftarrow \frac{1}{|\mathscr{A}_i|}, \quad C \leftarrow 0.$$

2. Repeat for each time t:

   a) Select action A_i according to mixed strategy x_i with suitable exploration.

   b) Observing the reward r_i resulting from action profile $(A_j)_{j\in N}$, set for $h = A_i$

$$Q_{ih} \leftarrow Q_{ih} + \alpha(r_i + \gamma \max_{h'} Q_{ih'} - Q_{ih}).$$

   c) Update estimate of average policy, $\bar{x}$,

$$\begin{aligned} C &\leftarrow C + 1, \\ \bar{x}_{ih'} &\leftarrow \bar{x}_{ih'} + \frac{1}{C}(x_{ih'} - \bar{x}_{ih'}) \quad (\forall h' \in \mathscr{A}_i). \end{aligned}$$

   d) Step x closer to the optimal policy w.r.t. Q.

$$x_{ih'} \leftarrow \begin{cases} x_{ih'} - \delta_{ih'} & \text{if } h' \neq \arg\max_k Q_{ik}, \\ x_{ih'} + \sum_{k\neq h'} \delta_{ik} & \text{otherwise,} \end{cases}$$

   with

$$\delta_{ik} = \min\left\{x_{ik}, \frac{\delta}{|\mathscr{A}_i| - 1}\right\} \text{ and } \delta = \begin{cases} \delta_w & \text{if } \sum_{h'} x_{ih'} Q_{ih'} > \sum_{h'} \bar{x}_{ih'} Q_{ih'}, \\ \delta_l & \text{otherwise.} \end{cases}$$

Algorithm 6.2.2 Frequency-adjusted Q-learning, [52], for generic player $i \in N$.

1. Set learning rate ϑ sufficiently small, choose initial $x_i \in \Delta_i$ and parameters $\beta > 0$, $\tau > 0$, $\gamma \geq 0$.

2. Repeat for each time t:

   a) Select action $A_i \in \mathscr{A}_i$ according to mixed strategy x_i.

   b) Observe reward r_i resulting from action profile $(A_j)_{j\in N}$.

   c) For $h = A_i$, set:

$$Q_{ih} \leftarrow Q_{ih} + \min\left\{\frac{\beta}{x_{ih}}, 1\right\} \vartheta \left(r_i + \gamma \max_{k\in\mathscr{A}_i} Q_{ik} - Q_{ih}\right).$$

   d) For all $h \in \mathscr{A}_i$, set: $x_{ih} \leftarrow \dfrac{e^{\tau Q_{ih}}}{\sum_{k\in\mathscr{A}_i} e^{\tau Q_{ik}}}$.

---

on direct policy updates, MBL-DPU (algorithm 6.2.3)—where the policy update corresponds to Cross learning, [27], with a mutation bias as a perturbation term; the other, based on logistic choice, MBL-LC (algorithm 6.2.4)—where the policy corresponds to logistic choice (also known as a Boltzmann distribution or as softmax policy) based on action-value estimates which are updated with a mutation bias perturbation.

MBL with direct policy update (MBL-DPU). We first consider the simplest version of MBL, which performs a simple direct policy update, MBL-DPU. We note that for $M_i = 0$ ($\forall i \in P$), MBL-DPU reduces to Cross learning [13, 27]. MBL-DPU is therefore an additive perturbation of Cross learning with perturbation term $\vartheta M_i\,(c_{ih} - x_{ih})$. We further note that the assumption in Cross learning, that rewards be restricted to $[0,1]$, is not necessary. It suffices that rewards are non-negative and bounded. In this case, ϑ can be chosen small enough to ensure well-definition of MBL-DPU. Note that this assumption is not restrictive for finite games, as boundedness is trivially satisfied for finite games and non-negativity can be ensured by adding a constant C_i to all payoffs r_i. This affects neither the position of Nash equilibria nor the dynamics in the deterministic limit—a straight-forward property of the replicator dynamics.

MBL with logistic choice (MBL-LC). The simple perturbation in MBL-DPU can be combined with a wide class of transformations on the payoffs without affecting the additive character of the perturbation. Additionally, we consider a more complex possibility to combine the mutation-like perturbation with a policy based on multinomial

---

Algorithm 6.2.3 MBL-DPU for generic player $i \in P$.

---

1. Set learning rate ϑ sufficiently small, choose initial $x_i \in \Delta_i$ and mutation parameters $M_i > 0$ and $c_i \in \Delta_i^{\circ}$.
2. Repeat for each time t:
   a) Select action A_i according to mixed strategy x_i.
   b) Observe reward r_i resulting from action profile $(A_j)_{j\in N}$.
   c) For all $h \in \mathscr{A}_i$, update x_{ih} according to

$$x_{ih} \leftarrow \begin{cases} x_{ih} + \vartheta(1 - x_{ih})r_i + \vartheta M_i\,(c_{ih} - x_{ih}) & \text{if } h = A_i, \\ x_{ih} - \vartheta x_{ih} r_i + \vartheta M_i\,(c_{ih} - x_{ih}) & \text{otherwise.} \end{cases}$$

---

logistic choice, as frequently encountered in Q-learning. In MBL-LC, the perturbation affects the Q-value updates instead of the policy. Hence, this version more closely resembles FAQ learning. In particular, restricting the frequency-adjustment in step 2.c) by applying a minimum is parallel to [52]. One can see that the logistic choice policy can still be expressed as a direct policy update with modified payoffs in the following way for a chosen action A_i:

$$x_{ih} \leftarrow \begin{cases} x_{ih} + (1 - x_{ih})\tilde{r}_i & \text{if } h = A_i \\ x_{ih} - x_{ih}\tilde{r}_i & \text{otherwise} \end{cases}, \qquad \text{where } \tilde{r}_i = \frac{x_{iA_i}(e^{\tau \Delta Q_{iA_i}} - 1)}{x_{iA_i}(e^{\tau \Delta Q_{iA_i}} - 1) + 1},$$

and ΔQ_{iA_i} denotes the update of the Q-value of the chosen action A_i. From this it is clear that an intermediate approach could be to use the simpler MBL-DPU combined with payoffs derived from Q-learning, which is equivalent to transforming payoffs accordingly.

Algorithm 6.2.4 MBL-LC for generic player $i \in N$.

---

1. Set learning rate ϑ sufficiently small, choose initial $x_i \in \Delta_i$ and mutation parameters $M_i > 0$ and $c_i \in \Delta_i^{\circ}$. Choose $\beta > 0$, $\tau > 0$.
2. Repeat for each time t:
   a) Select action A_i according to mixed strategy x_i.
   b) Observe reward r_i resulting from action profile $(A_j)_{j \in N}$.
   c) For $h = A_i$, set: $Q_{ih} \leftarrow Q_{ih} + \min\left\{\frac{\beta}{x_{ih}}, 1\right\} \vartheta \left(r_i + M_i \frac{c_{ih}}{x_{ih}}\right)$.
   d) For all $h \in \mathcal{A}_i$, set: $x_{ih} \leftarrow \frac{e^{\tau Q_{ih}}}{\sum_{k \in \mathcal{A}_i} e^{\tau Q_{ik}}}$.

6.2.3 Convergence of MBL

The question of convergence can be considered in two steps. First, one determines whether the stochastic process induced by the learning algorithm can be approximated by a deterministic dynamics. Second, one might transfer the convergence properties of the deterministic dynamics to the stochastic process. For MBL-DPU we have the following convergence result (proved in section 6.5 as proposition 6.5.3):

Proposition 6.2.1. *For every $T < \infty$, the family of stochastic processes $\{(X_{ih}^{\vartheta}(t))_{i,h}\}_{t\geq 0}$ induced by MBL-DPU converges to* (RMD) *in the sense that for all $\varepsilon > 0$:*

$$\sup_{x^0 \in \Delta} \Pr(\|X^{\vartheta}(n_{\vartheta}) - \Phi^M(x^0, T)\| > \varepsilon) \to 0 \quad as \quad \vartheta \to 0,$$

where $n_{\vartheta}\vartheta \xrightarrow{\vartheta \to 0} T$, x^0 is (almost surely) the initial state of the stochastic processes and $\Phi^M(x^0, \cdot)$ is the unique solution of (RMD) *with $\Phi^M(x^0, 0) = x^0$.*

Remark. Analogous to [13, 72], proposition 6.2.1 on its own does not yield an analysis of the asymptotic behaviour of the stochastic processes. If an equilibrium x^M of (RMD) is asymptotically stable and x^0 lies in the basin of attraction of x^M, then we have $\Phi^M(x^0, T) \to x^M$ as $T \to \infty$. However, with the asymptotic stability of x^M, we have that for T large enough, $\Phi^M(x^0, T)$ is arbitrarily close to x^M and together with proposition 6.2.1, any neighbourhood of x^M will be reached by the learning process with an arbitrary degree of certainty after finitely many steps. This, however, does not imply that the process must remain in this neighbourhood afterwards.

The utility of the mutation perturbation stems from the fact that learning without the mutation perturbation always leads to the boundary of Δ which is particularly unfortunate if the only Nash equilibrium is located in the interior.

In chapter 4 we have shown that every game has at least one connected equilibrium component that is approximated by mutation equilibria irrespective of the choice of the mutation parameter c as $M \to 0$. Furthermore, it was shown that for the Matching Pennies game, the Nash equilibrium is approximated by asymptotically stable mutation equilibria, warranting the name *attracting mutation limit*. In fact, for Matching Pennies those mutation equilibria are even globally asymptotically stable, i.e., trajectories converge to the mutation equilibrium for all initial states $x \in \Delta$. This is due to the fact that the system is planar and therefore the Poincaré-Bendixson theorem holds. More generally, we state this as the following proposition (a direct consequence of proposition 6.5.4 and corollary 6.5.5):

Proposition 6.2.2. *Let $x^* \in \Delta^0$ be an attracting mutation limit and U a neighbourhood of x^*. If the mutation equilibria approximating x^* are globally asymptotically stable, then for every mutation parameter $c \in \Delta^0$ there are $M > 0$ and $\vartheta > 0$ such that the stochastic process $\{X^{\vartheta}(t)\}_{t\in\mathbb{N}_0}$ induced by MBL-DPU visits U at a finite time almost surely, i.e., $X^{\vartheta}(S) \in U$ for some $S \in \mathbb{N}_0$ with probability 1. In fact, $\{X^{\vartheta}(t)\}_{t\in\mathbb{N}_0}$ visits U infinitely often with probability 1.*

In contrast to MBL-DPU, we do not have a proof of an analogous result for MBL-LC. Although it would be plausible for MBL-LC to behave similarly to MBL-DPU and to converge to (RMD) in the limit of $\vartheta \to 0$, the numerical results show that this is not as clear as it might seem.

6.2.4 Perturbation creates a trade-off between accuracy and speed

We note that neither MBL-DPU nor MBL-LC "converge" to a Nash equilibrium but only to an ε-equilibrium and in particular, that both stay away from the boundary of Δ. For MBL-DPU this is clear from the facts that the equilibria of (RMD) are not Nash equilibria and that the boundary of Δ is repelling. For MBL-LC this is also due to the exploration parameter τ. For the latter, it is further the case that τ cannot be let to approach ∞ as this collides with the $\vartheta \to 0$ limit and makes the time derivative of the policy unbounded. This results in a highly increased variance in the stochastic process, preventing effective learning of equilibria. This particular aspect applies also to other logistic choice based algorithms, particularly FAQ learning.

However, if MBL-LC and FAQ indeed converge to the corresponding ODE systems in the deterministic limit, then these include τ as a simple scaling parameter, cf. equation (6.2.1). Since constant positive rescalings do not change the trajectories, the systems can be rescaled by $1/\tau$ in such a way that τ effectively regulates the perturbation strength relative to the replicator dynamics. In the case of (RMD), $1/\tau$ can be absorbed by the mutation strength M. Thus an increase of τ has the same effect as a decrease of M which results in mutation equilibria moving closer to a Nash equilibrium, as desired. We can thus ignore τ in the policy function of MBL-LC. A similar reasoning can be applied to FAQ learning: A reduction in the perturbation term also results in a longer time to approach equilibria and this creates a trade-off between accuracy, i.e., the distance to the Nash equilibrium, and convergence speed.

6.3 Numerical results

We consider the behaviour of MBL-DPU, illustrating the theoretical results, and of MBL-LC, providing a first intuition, in a number of different game settings and compare these to FAQ-learning as an instance which is close to MBL-LC in its formulation and approach and to WoLF-PHC as a well-known MARL algorithm. We consider the much analysed Prisoner's Dilemma (PD) game as a case with a strict Nash

equilibrium, which illustrates the different behaviours when the Nash equilibrium is asymptotically stable and located on the boundary of Δ, more specifically at a vertex. As a second, significantly different case, we consider zero-sum games, or equivalently constant-sum, which have interior Nash equilibria and, as we have shown, are attracting mutation limits. This implies that the linearised system cannot be unstable at the Nash equilibrium. However, here (RD) does not converge and the differences of the algorithms compared to (RD) (and Cross learning) become more explicit.

6.3.1 Prisoner's dilemma

In PD, we have a strict Nash equilibrium. According to previous results, we have that strict Nash equilibria are asymptotically stable in (RD) and hence (RMD) has asymptotically stable equilibria nearby when mutation is sufficiently low. Therefore, MBL-DPU will also approach the mutation equilibrium and, depending on mutation strength, it will approximate the Nash equilibrium. However, as mutation equilibria (for generic $c \in \Delta^o$) differ from Nash equilibria, MBL-DPU cannot converge to the Nash equilibrium. In particular, MBL-DPU stays away from the boundary of Δ, while the Nash equilibrium is on the boundary.

MBL-DPU and MBL-LC. The experimental results (figures 6.1, 6.2) illustrate the behaviour of MBL-DPU and its convergence for different mutation strengths M. In accordance with intuition, convergence is quick for high mutation strength at the price of the mutation equilibrium being further away from the Nash equilibrium. For lower values of M, we have that the mutation equilibrium moves closer to the Nash equilibrium while convergence becomes slower.

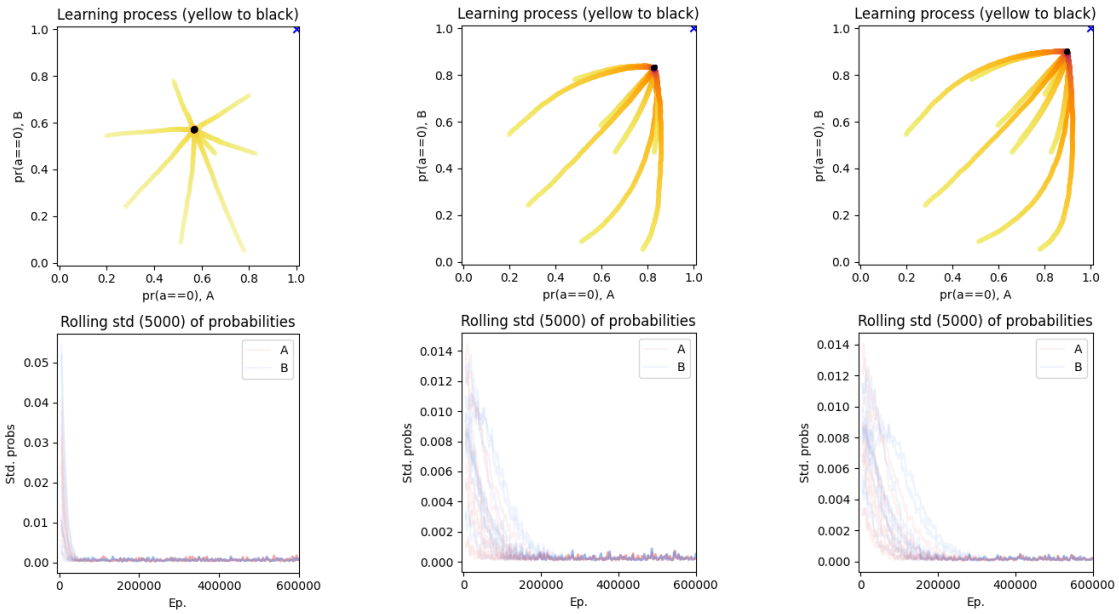
In comparison, MBL-LC (figures 6.3, 6.4) behaves similarly while converging much more quickly. An intuition for this is provided when considering that MBL-DPU can be viewed as a linear approximation to MBL-LC for small τ.

FAQ-learning. For FAQ-learning (figures 6.5, 6.6), the role of τ corresponds to that of M^{-1} in MBL. We have that, similarly to both MBL variants, with increasing values of τ (i.e., decreasing values of M), the dynamics approaches a region that lies closer to the Nash equilibrium. The intuition here is provided by the fact that the deterministic limit of FAQ is a replicator dynamics with a perturbative term whose effect depends on τ and which pulls the system towards the centre of Δ. Furthermore, convergence

is the slower the weaker the perturbative term is, much like in the two MBL variants. In contrast to the MBL variants, FAQ-learning defaults to the usual Q-learning when $x_{ih} \leq \beta$. This effectively neutralises the repelling dynamics at the boundary of Δ, which would otherwise result in very large (unbounded) changes in the Q-values for very low values of x_{ih}. Note that MBL-LC has x_{ih} occurring in the denominator twice and hence retains the repelling effect at the boundary of Δ.

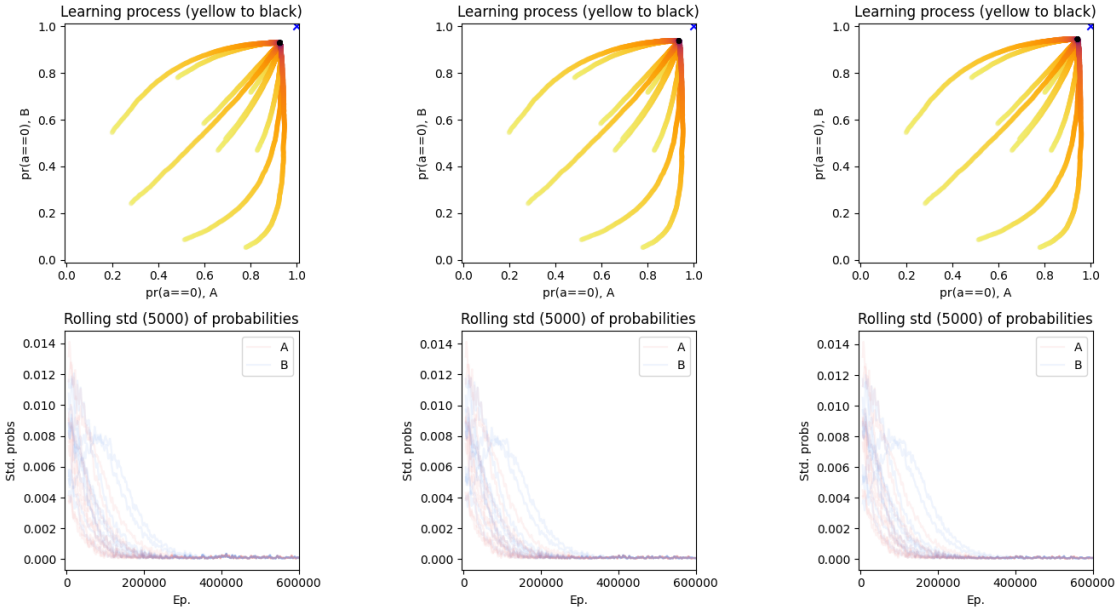
WoLF-PHC. In contrast to the other algorithms, WoLF-PHC (figure 6.7) follows a chosen direction for some time until it is replaced by a new direction, which results in a discrete sequence of directions and non-smooth trajectories. Convergence to the Nash equilibrium occurs much faster than for the other algorithms in the case of PD. However, strict Nash equilibria are also asymptotically stable in (RD) and thus PD is a base case which illustrates the different behaviours in a clear-cut situation, as opposed to more challenging and ambiguous situations without strict Nash equilibria.

Figure 6.1: MBL-DPU in self-play on the PD game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initial conditions. In each subfigure, the upper graph shows the ten trajectories in the projection on the first components of the players' strategies, in this case the 'defect' strategy, with the first player given on the horizontal axis and the second player on the vertical axis. Points coloured yellow correspond to earlier points in time, changing over orange and violet to black for later points in time. The position of the game's Nash equilibrium is marked with a blue cross in the projection plane. The lower graph shows the standard deviation of all components of the players' strategies for each point in time over the past 5000 time steps, for each of the ten initial conditions, coloured red and blue for the two players. Time is given on the horizontal axis. The standard deviation is computed with the usual Euclidean metric.



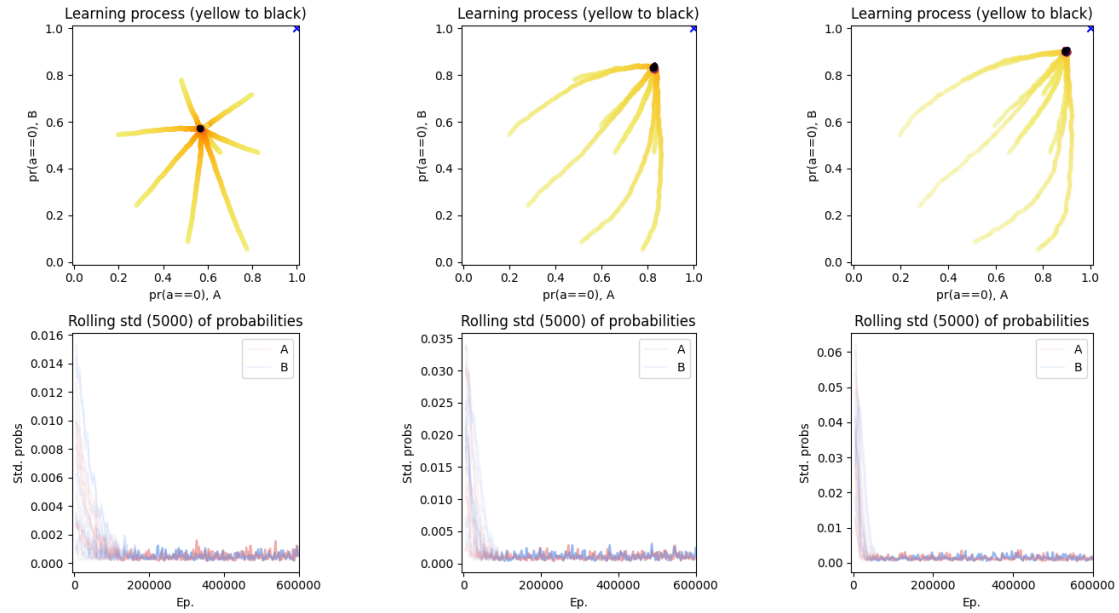
(a) $\tau = 1, M = 1$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.2: MBL-DPU in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



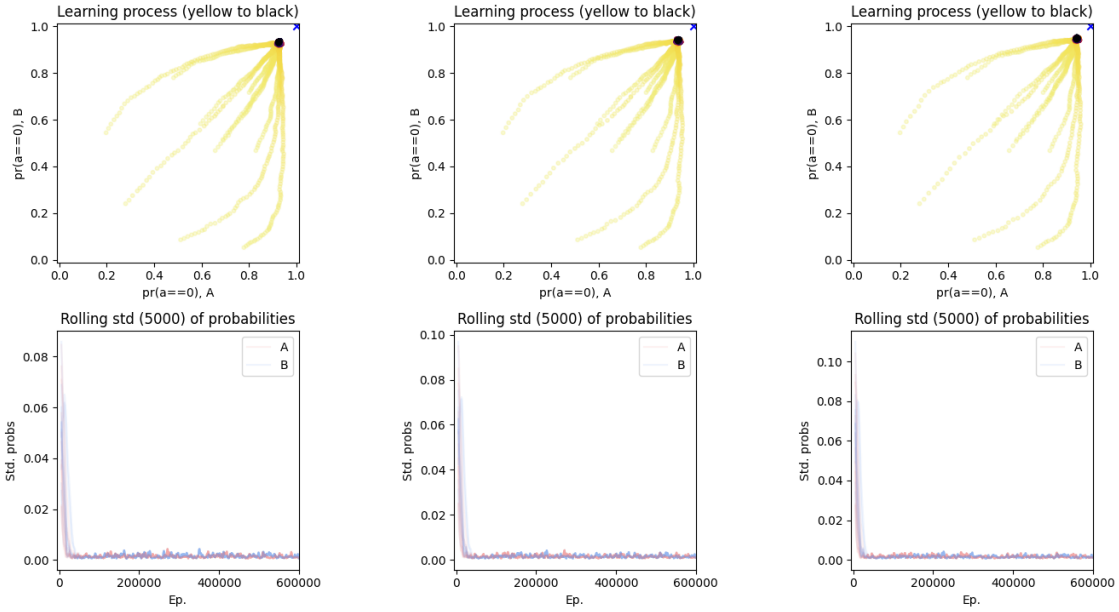
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.3: MBL-LC in self-play on the PD game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



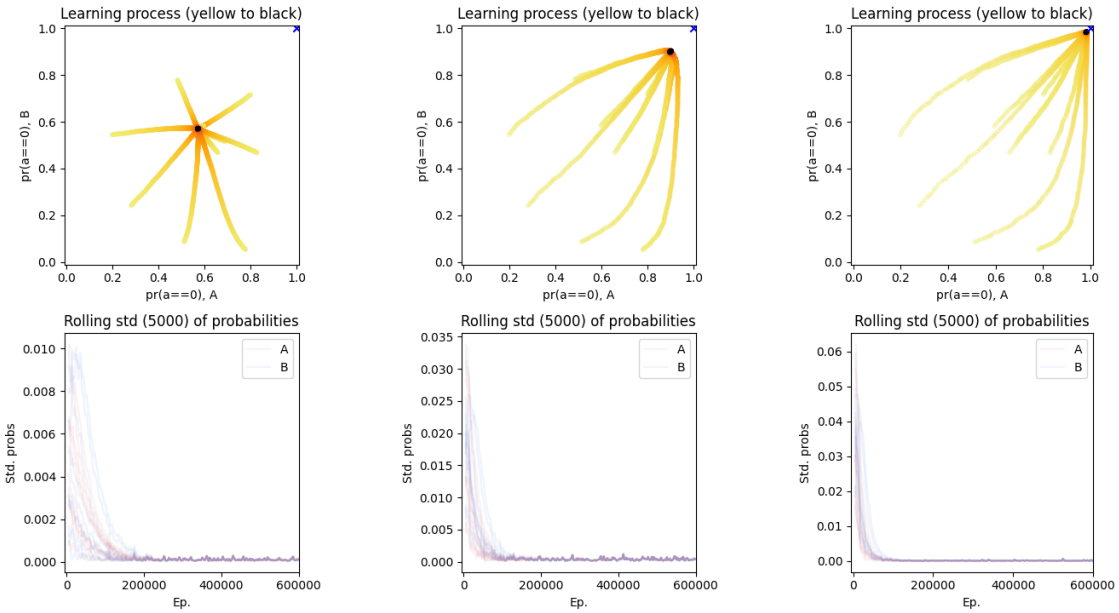
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.4: MBL-LC in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



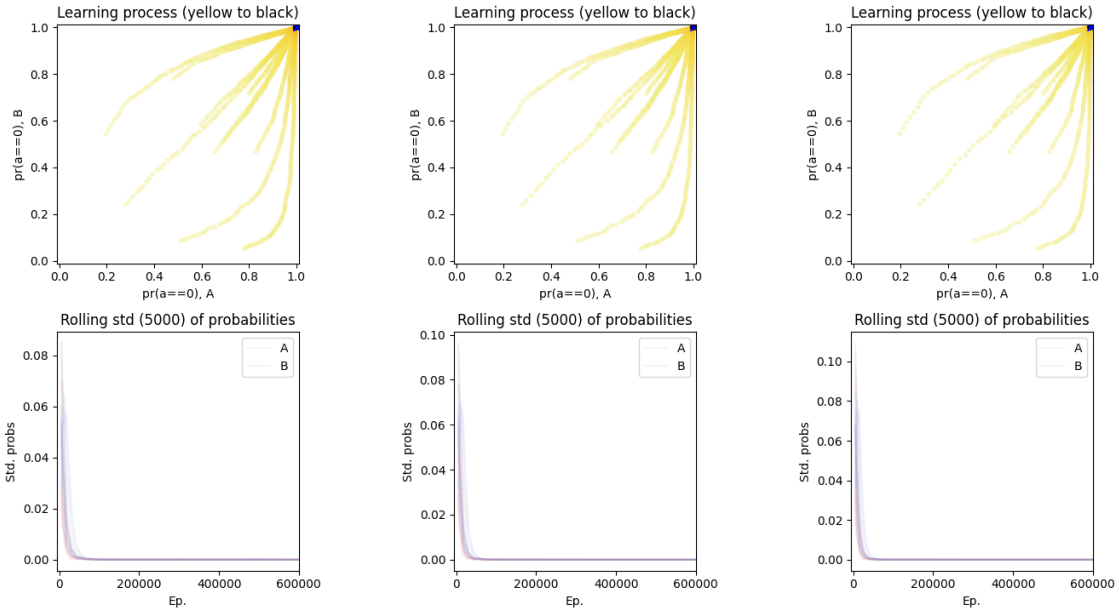
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.5: FAQ in self-play on the PD game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



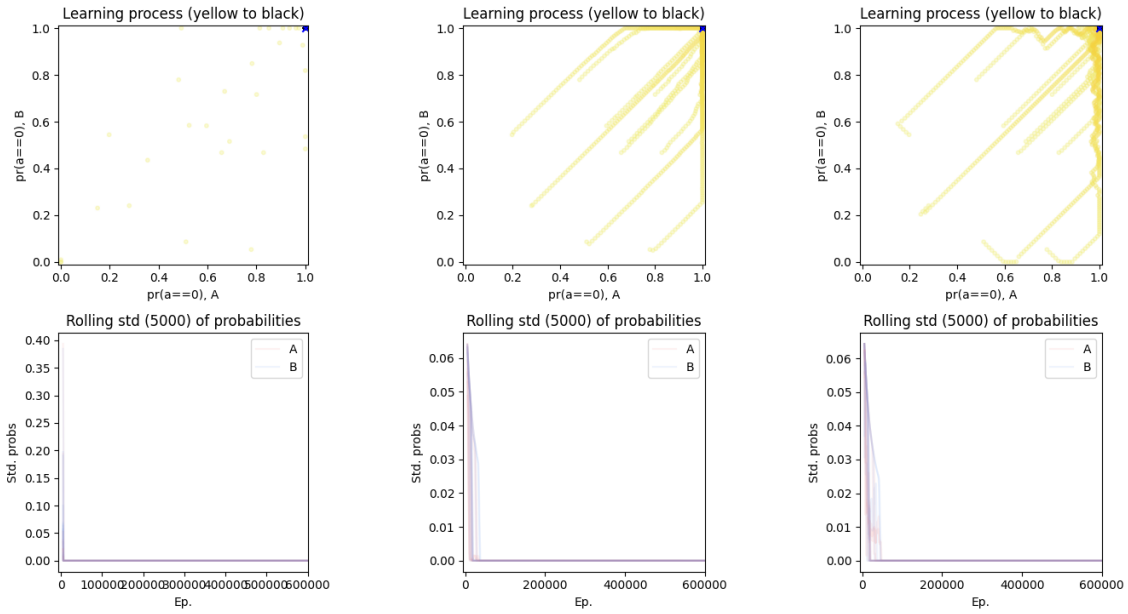
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.6: FAQ in self-play on the PD game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) $\tau = 30$, $M = 30^{-1}$ **(b)** $\tau = 35$, $M = 35^{-1}$ **(c)** $\tau = 40$, $M = 40^{-1}$

Figure 6.7: WoLF-PHC in self-play on the PD game with different learning schedules; for 10 different initialisations. Subgraph (a) has a high convergence speed such that only disconnected points can be seen. (See figure 6.1 for a detailed explanation of the graphs.)



(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}.

(b) Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

(c) Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

6.3.2 Zero-sum games

For two-player zero-sum games, we have shown that the Nash equilibrium is an attracting mutation limit. Therefore, while (RD) (and Cross learning) would not converge to interior equilibria (with Cross learning eventually approaching the boundary), (RMD) converges to the mutation equilibrium for every choice of mutation probabilities, $c \in \Delta^o$ and $M > 0$, and so does MBL-DPU. Here, stability is induced by the perturbative terms and their varying strengths have two effects which have to be weighed against each other. We demonstrate the general idea in the simple situation of the Matching Pennies (MP) game. Further, we illustrate the changing behaviour when we grow the strategy space by considering different versions of the Rock-Paper-Scissors game, RPS-n, with $n = 3, 5, 9$, where n denotes the number of strategies available to each player.

Matching Pennies

The MP game is a particularly simple case of a zero-sum game and hence provides an informative perspective on the basic characteristics of the different algorithms. In general, we see that the location of the mutation equilibrium depends on the mutation strength M, while convergence is slower for lower values of M creating a trade-off between these, as discussed in more detail below.

MBL-DPU and MBL-LC. Comparing MBL-DPU and MBL-LC, we see again that the LC-variant (figures 6.10, 6.11) approaches the mutation equilibrium more quickly than the DPU-variant (figures 6.8, 6.9). However, we see that the DPU-variant exhibits a much smaller variance, more precisely standard deviation, in the vicinity of the mutation equilibrium due to its slower change, with both variants roughly differing by a factor between 5 and 10 (for $M = 40^{-1}$). This illustrates the stronger effect that single larger payoffs have on the LC-variant, producing a larger variance near the mutation equilibrium.

FAQ-learning. For FAQ-learning (figures 6.12, 6.13) we see a similar behaviour as MBL-LC, however with a smaller variance near the equilibrium for weaker perturbation (figure 6.13). As with the MBL variants, FAQ exhibits slower convergence for weaker perturbation with larger variance near its (apparently asymptotically stable) equilibrium. However, we also observe that with FAQ, solutions can get trapped near

the boundary (note the trapped solution in the upper left corner in figure 6.13), which we do not observe for the MBL variants and have proved not to be the case for MBL-DPU.

WoLF-PHC. Similar to the other algorithms, WoLF-PHC (figure 6.14) follows spiral-like trajectories towards a region close to the Nash equilibrium. It also shows a lower variance near the (apparently asymptotically stable) equilibrium. However, WoLF-PHC employs a learning rate schedule which reduces the learning rate over time and thus reduces variance.[2] One should note that WoLF-PHC has a considerably higher complexity as it relies on a reliable way to estimate action-values as well as a long-term population average. It is clear that for WoLF-PHC to be implemented by an evolutionary dynamics in a population, further biological mechanisms would be required to reflect these additional quantities, potentially in the form of e.g., cross-generational effects or some other age-structure in the population and corresponding mechanisms.

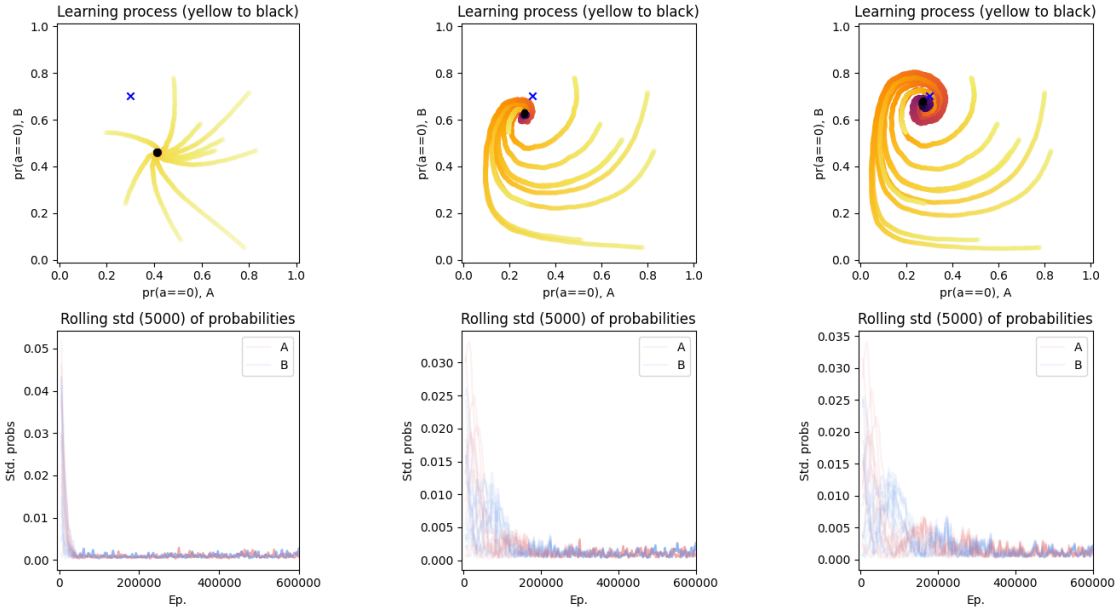
Zero-sum games with larger action spaces

While MP is an informative illustration of the different behaviours, it should be noted that MP reduces to a planar dynamical system, which does not allow many complex behaviours, as exemplified by the Poincaré-Bendixson theorem, e.g., [107, theorem 7.16] holding for planar systems. Hence, higher-dimensional zero-sum games allow a further understanding of the differences between the algorithms and shed light on the effect of larger state spaces while preserving the neutral stability of interior equilibria. We consider here the Rock-Paper-Scissors game of different sizes (3, 5 and 9 actions).

MBL-DPU and MBL-LC. In RPS-3, MBL-DPU (figures 6.15, 6.16) shows a similar behaviour to MP with a marked dependence of the behaviour of the variance on the value of M. In contrast, MBL-LC (figures 6.17, 6.18) shows a much quicker convergence, with the variance dropping after similar numbers of episodes (around 10^5) for all values of M. As with MBL-DPU, the residual variance increases with weaker mutation. This is in accordance with the neutral stability of the Nash equilibrium, allowing for larger fluctuations.

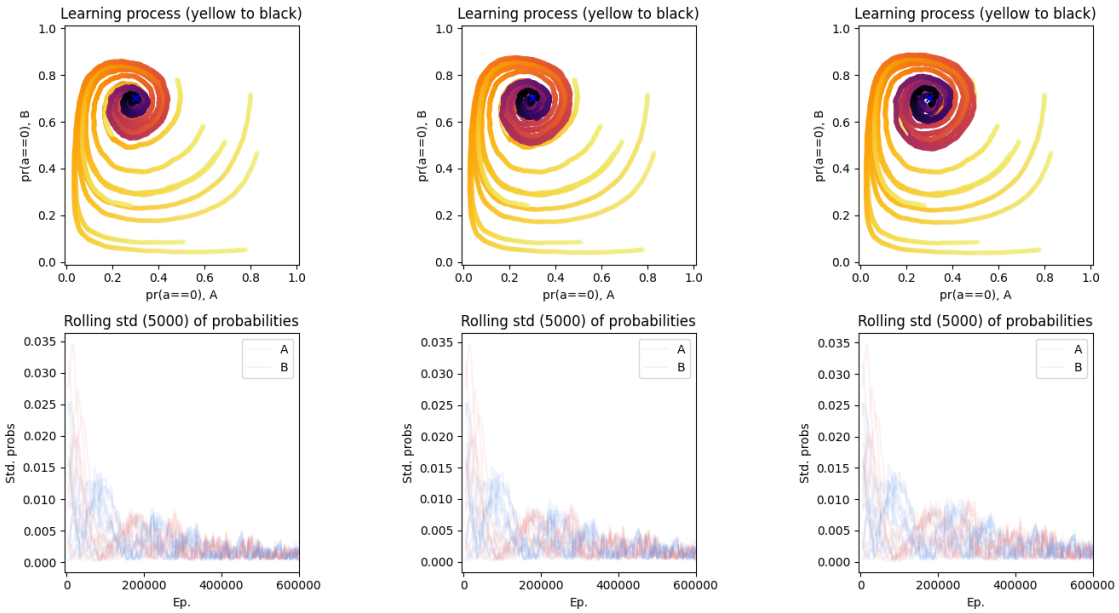
[2] It would be possible to evaluate WoLF-PHC with a fixed learning rate or use a reduction schedule for the other algorithms. However, the former would be a deviation from the canonical formulation of WoLF-PHC while the latter would not be based on a principled approach. Hence, this heterogeneous situation is an appropriate base scenario.

Figure 6.8: MBL-DPU in self-play on the MP game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



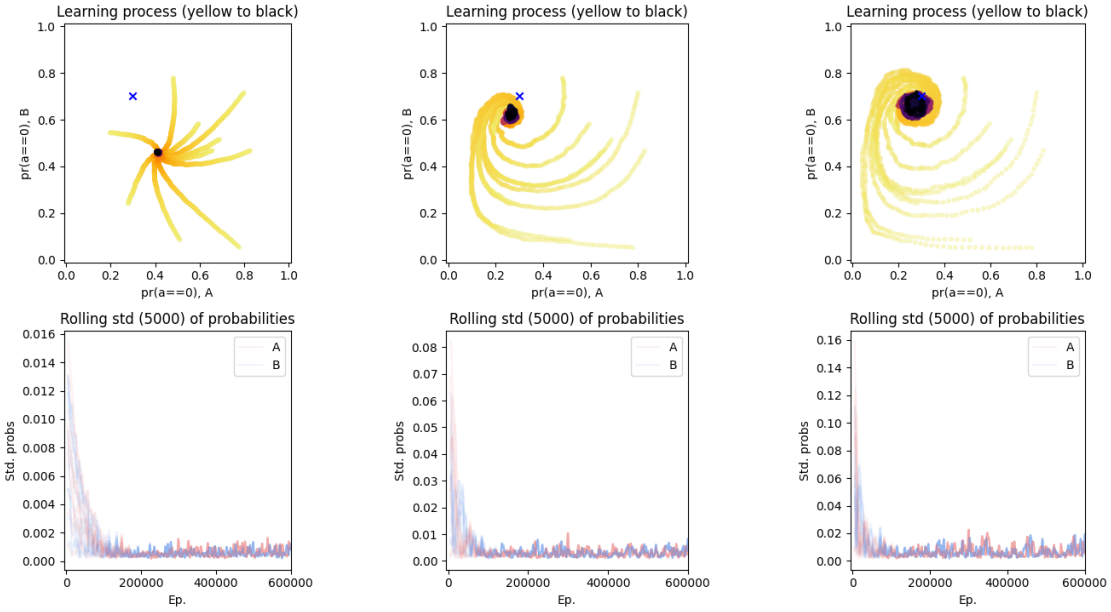
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.9: MBL-DPU in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



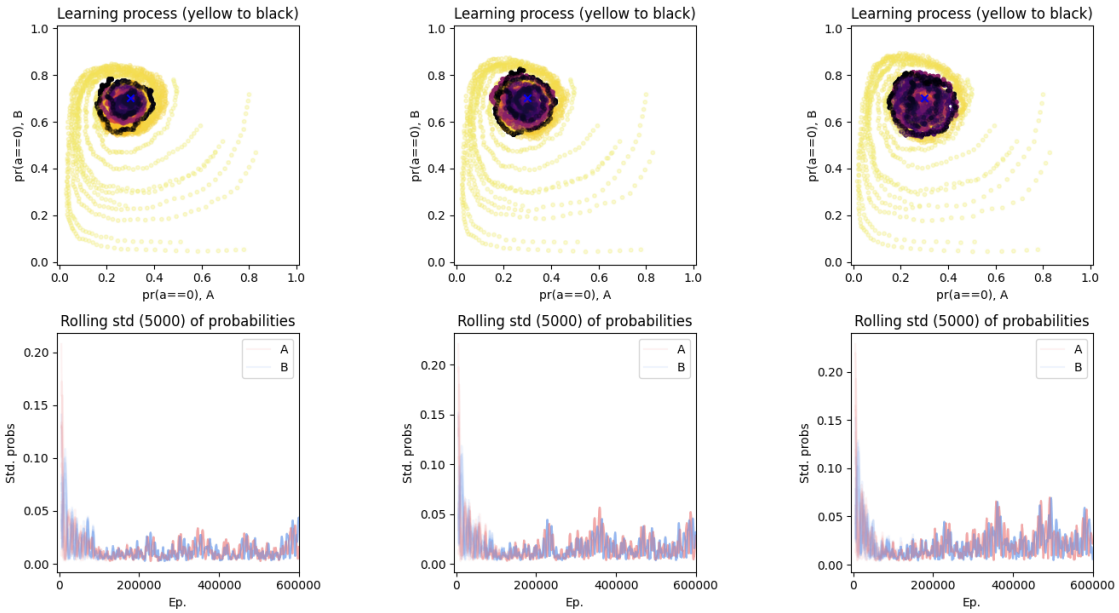
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.10: MBL-LC in self-play on the MP game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



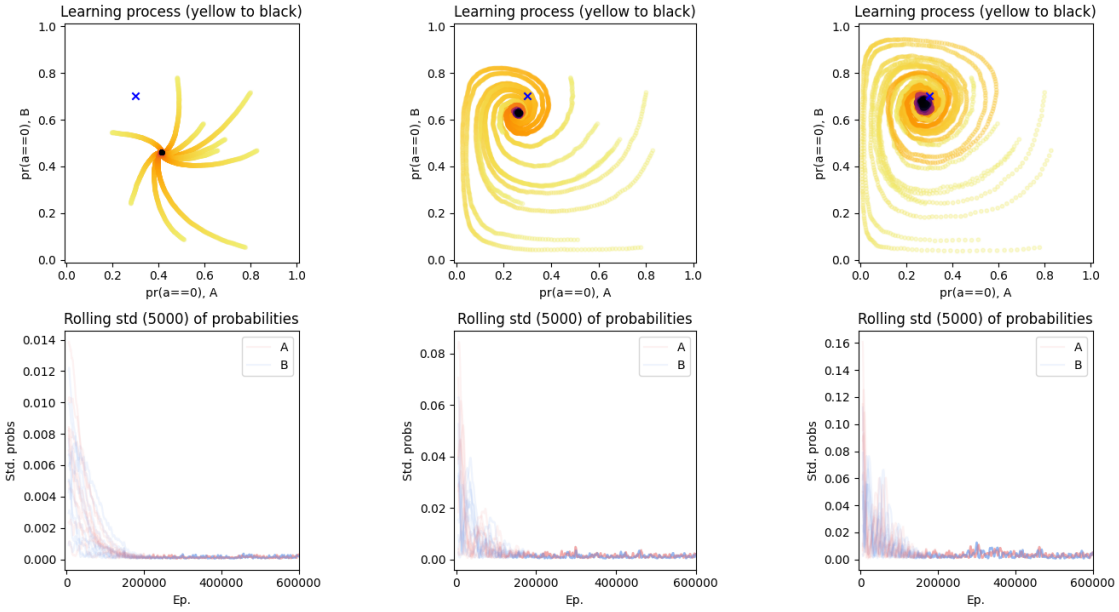
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.11: MBL-LC in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



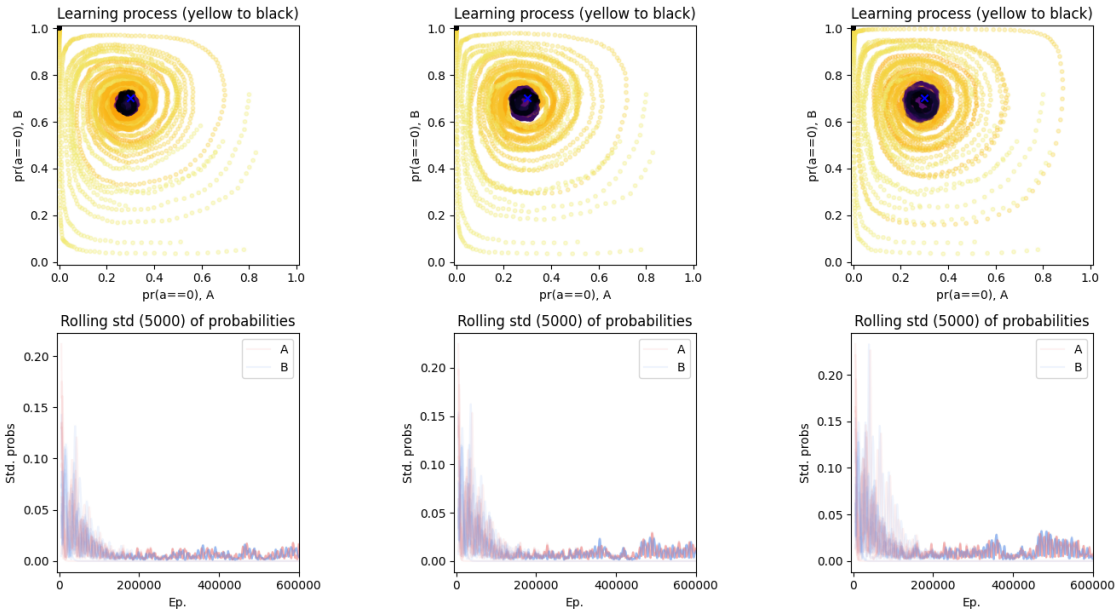
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.12: FAQ in self-play on the MP game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



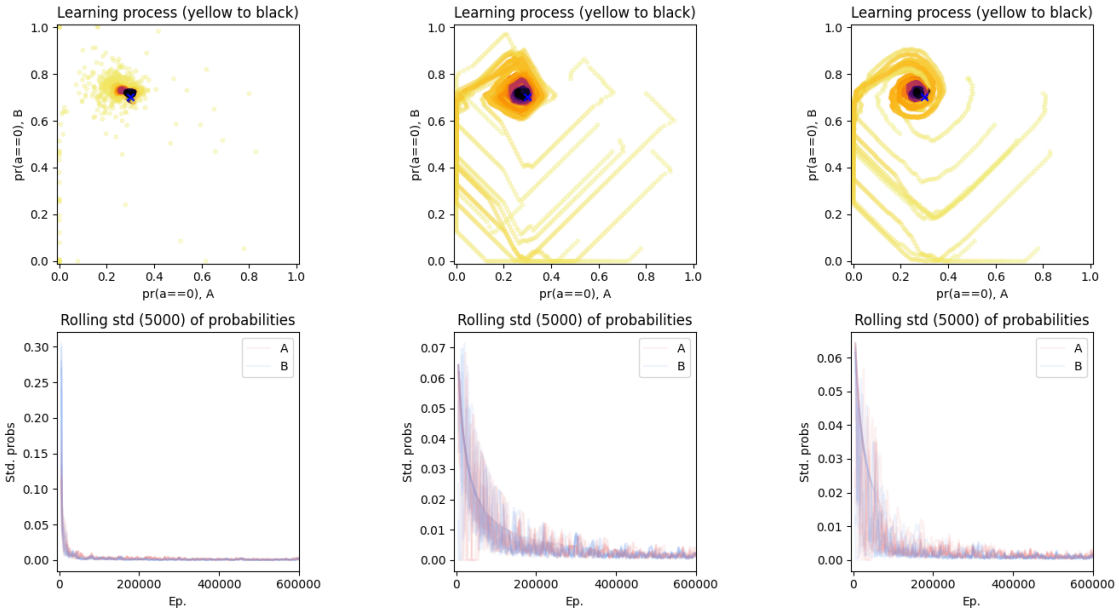
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.13: FAQ in self-play on the MP game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.14: WoLF-PHC in self-play on the MP game with different learning schedules; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}.

(b) Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

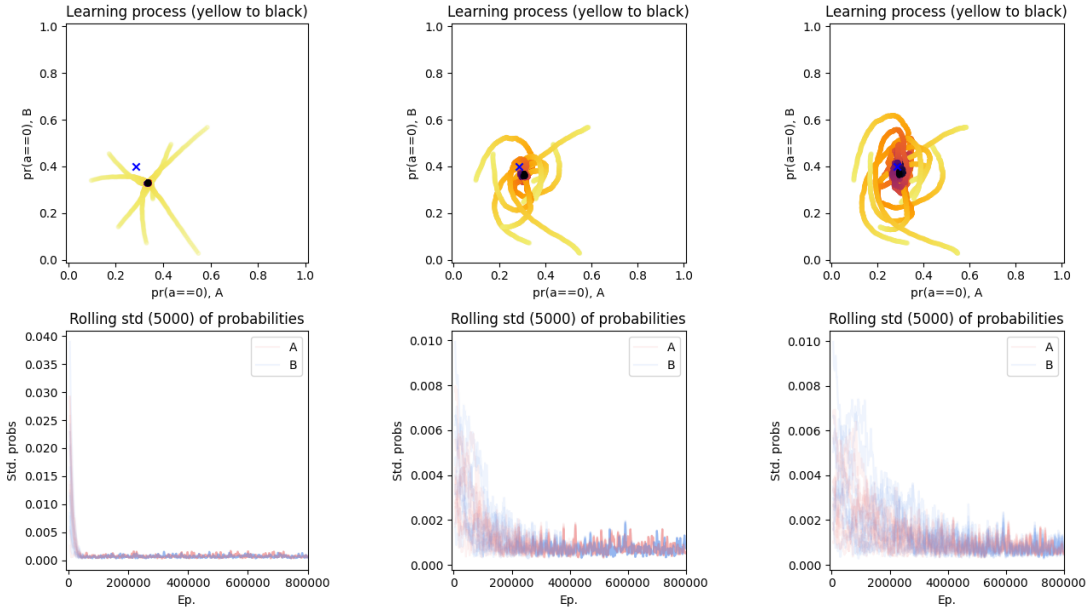
(c) Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

In RPS-5, both MBL variants (figures 6.22, 6.23 for MBL-DPU and figures 6.24, 6.25 for MBL-LC) show behaviours similar to their RPS-3 counterparts. In RPS-9, MBL-DPU (figures 6.29, 6.30) again shows similar behaviour, with slower convergence compared to its RPS-3 and RPS-5 counterparts. Interestingly, MBL-LC (figures 6.31, 6.32) seems to have two distinct regions to which trajectories evolve, suggesting a qualitatively different behaviour from MBL-DPU or a potentially stronger sensitivity to the choice of ϑ.

FAQ-learning. Like for MP, we see a quicker convergence for FAQ in RPS-3 (figures 6.19, 6.20) compared to the MBL variants, but with trajectories similar to those of MBL-LC when considering low values of M, in which case the replicator dynamics makes a stronger contribution to the trajectories. Similar to MBL-LC, but already in RPS-5, FAQ shows two distinct regions to which trajectories evolve when perturbation is weak (figures 6.26, 6.27), whereas the former does not show such a split for RPS-5. In RPS-9, FAQ shows such a split for stronger perturbation levels already and shows even three distinct such regions for weaker perturbation (figures 6.33, 6.34).

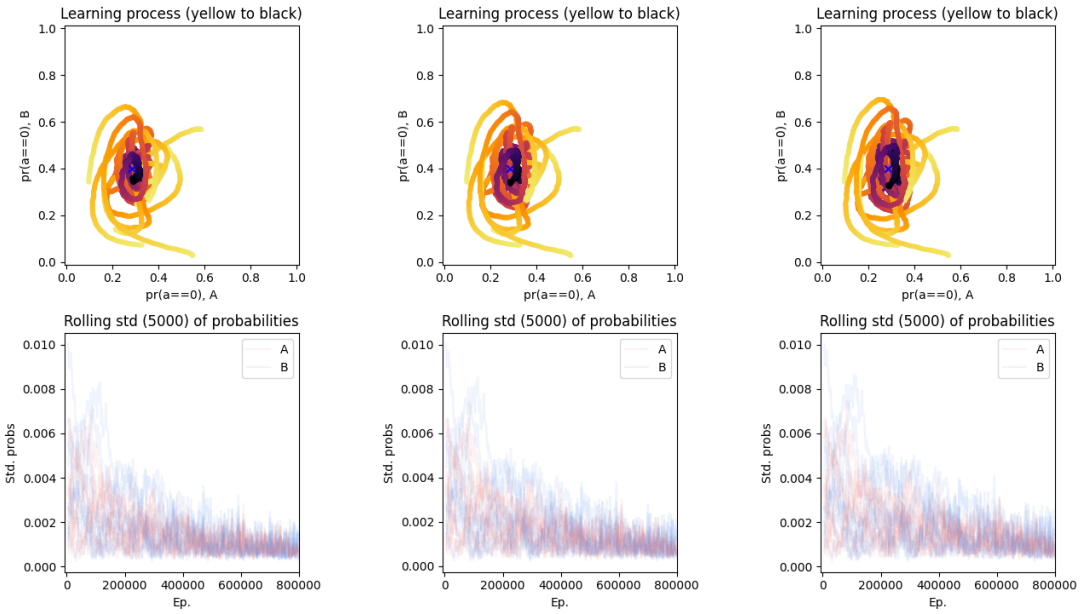
WoLF-PHC. For WoLF-PHC, we see a still quicker convergence in RPS-3 (figure 6.21) than for the other algorithms, similar to the MP case. However, the behaviour is much less clear in RPS-5 (figure 6.28). Here, trajectories do not consistently approach a specific region. It is possible that the reduction schedules for the learning rates, which force each trajectory to converge, lead to trajectories stalling prematurely. This becomes even more pronounced in RPS-9 (figure 6.35), where WoLF-PHC seems to initially move away from the Nash equilibrium and to get stuck along the boundaries of Δ.

Figure 6.15: MBL-DPU in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



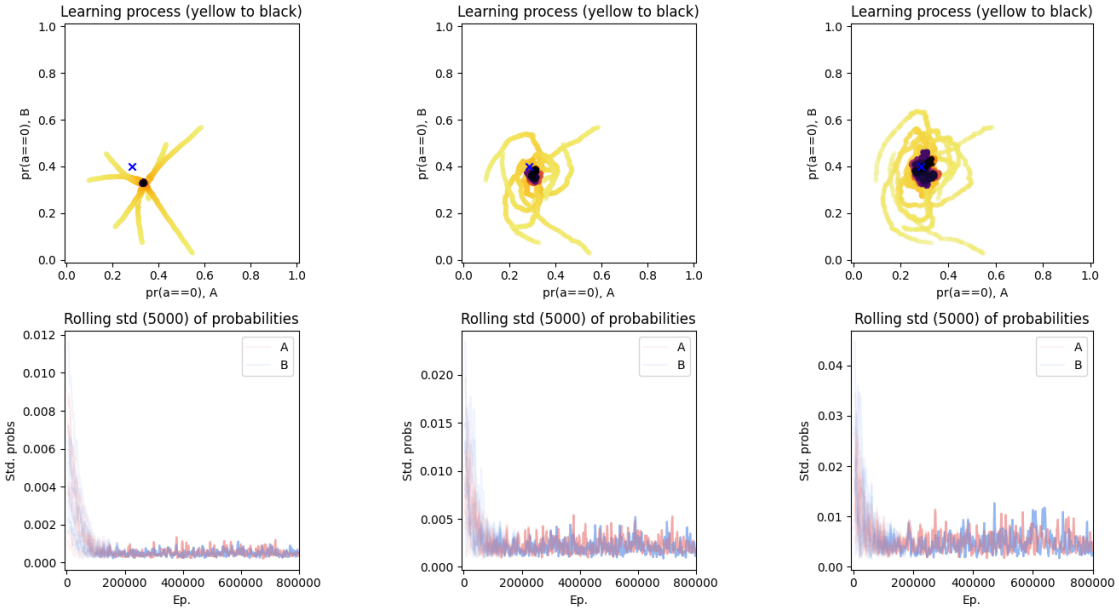
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.16: MBL-DPU in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



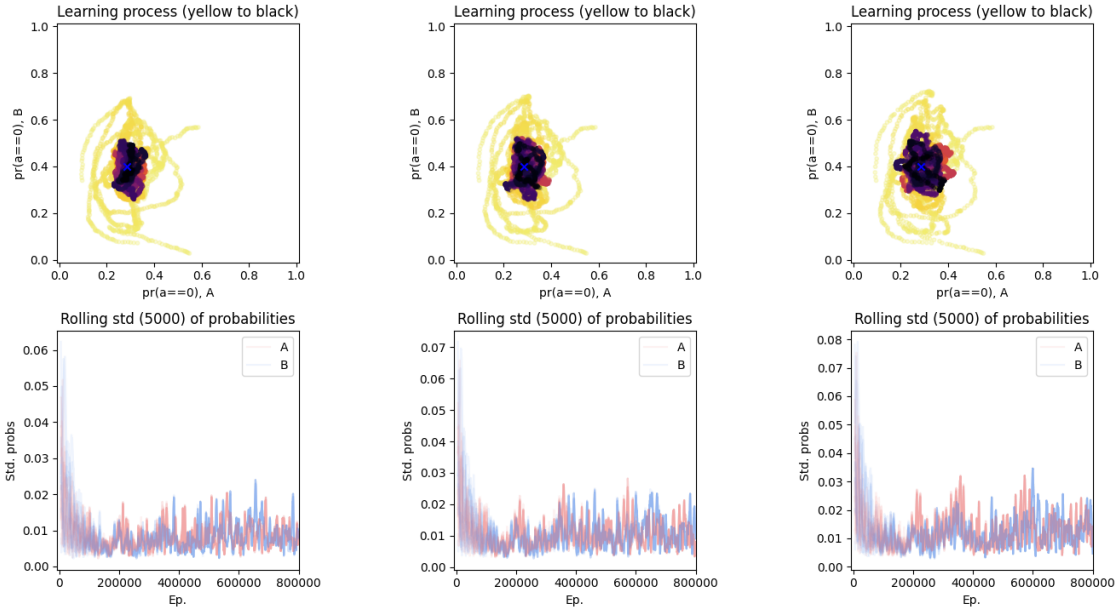
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.17: MBL-LC in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



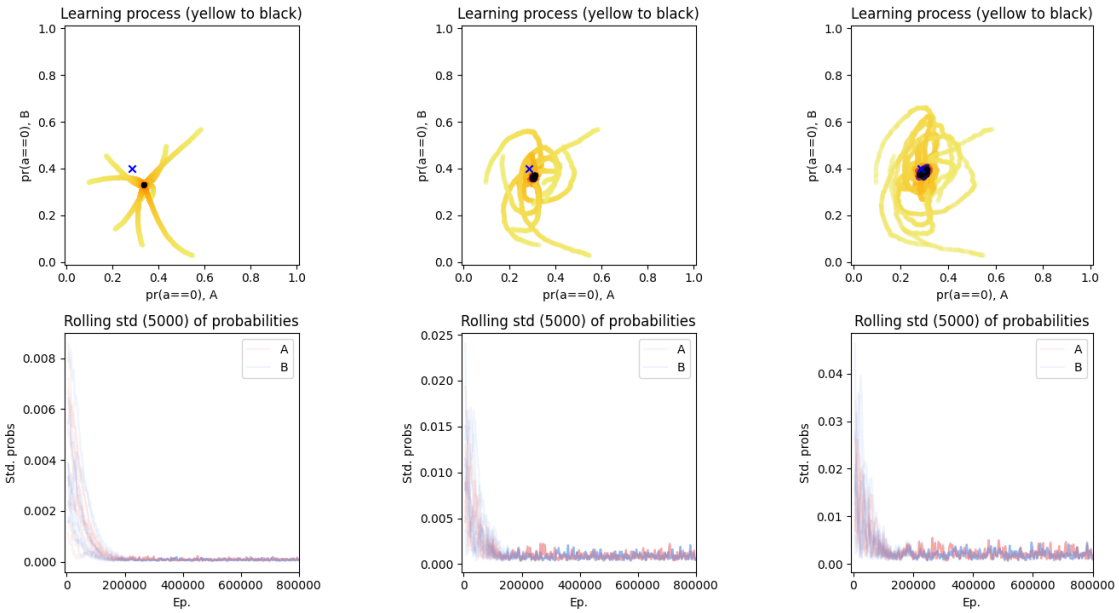
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.18: MBL-LC in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



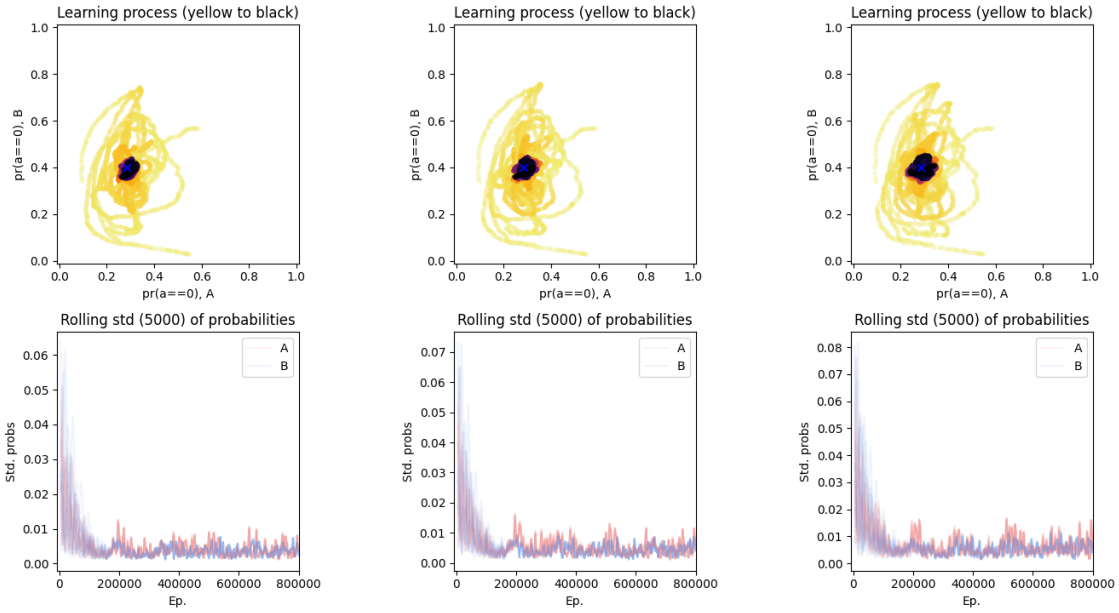
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.19: FAQ in self-play on the RPS-3 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.20: FAQ in self-play on the RPS-3 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations.

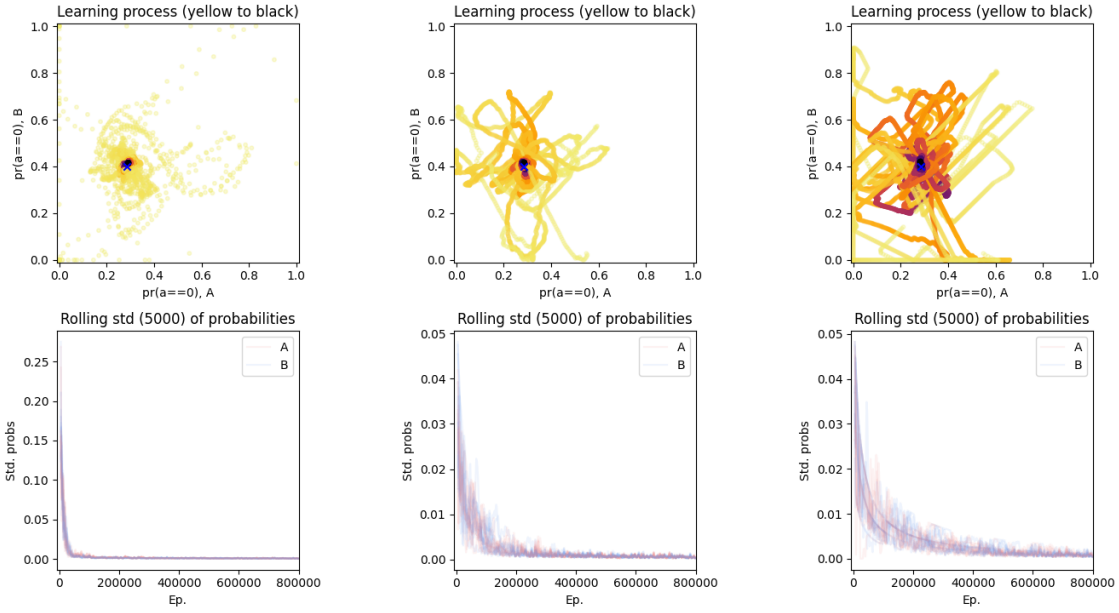


(a) $\tau = 30$, $M = 30^{-1}$

(b) $\tau = 35$, $M = 35^{-1}$

(c) $\tau = 40$, $M = 40^{-1}$

Figure 6.21: WoLF-PHC in self-play on the RPS-3 game with different learning schedules; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)

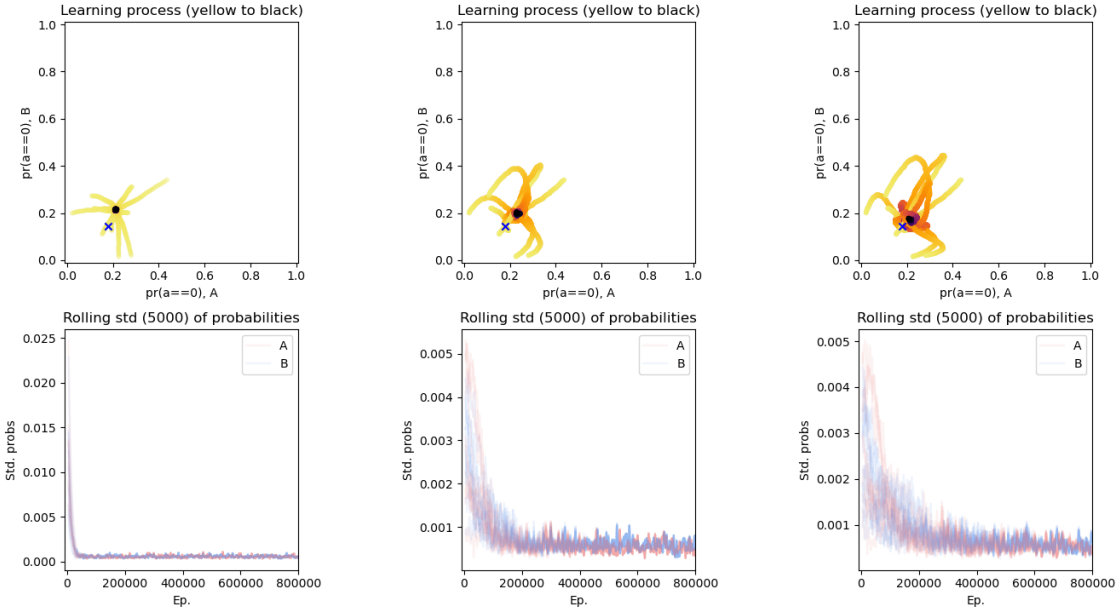


(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}.

(b) Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

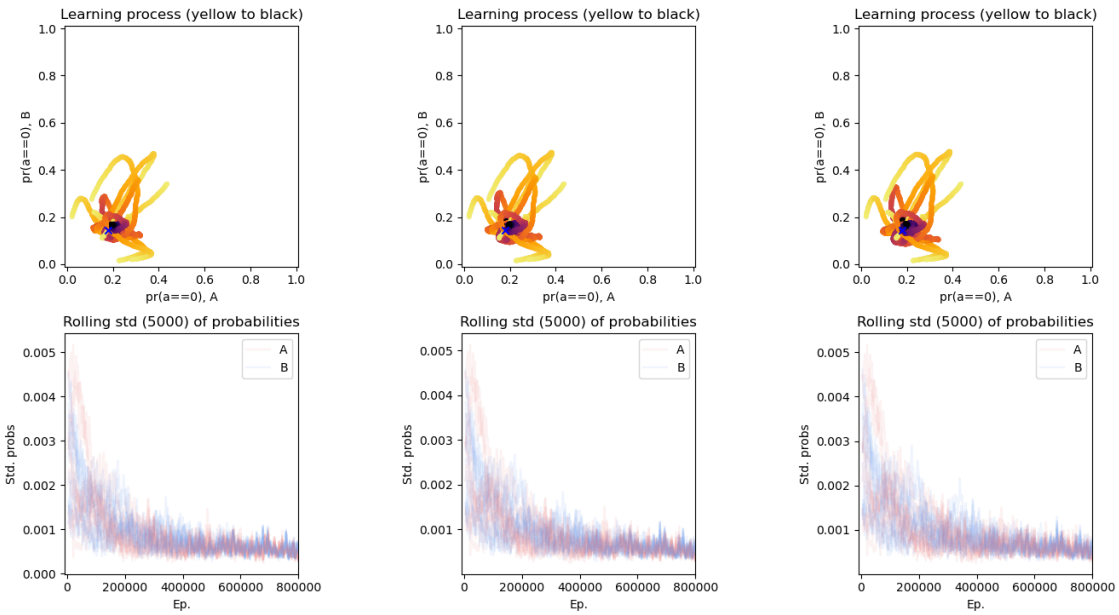
(c) Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

Figure 6.22: MBL-DPU in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



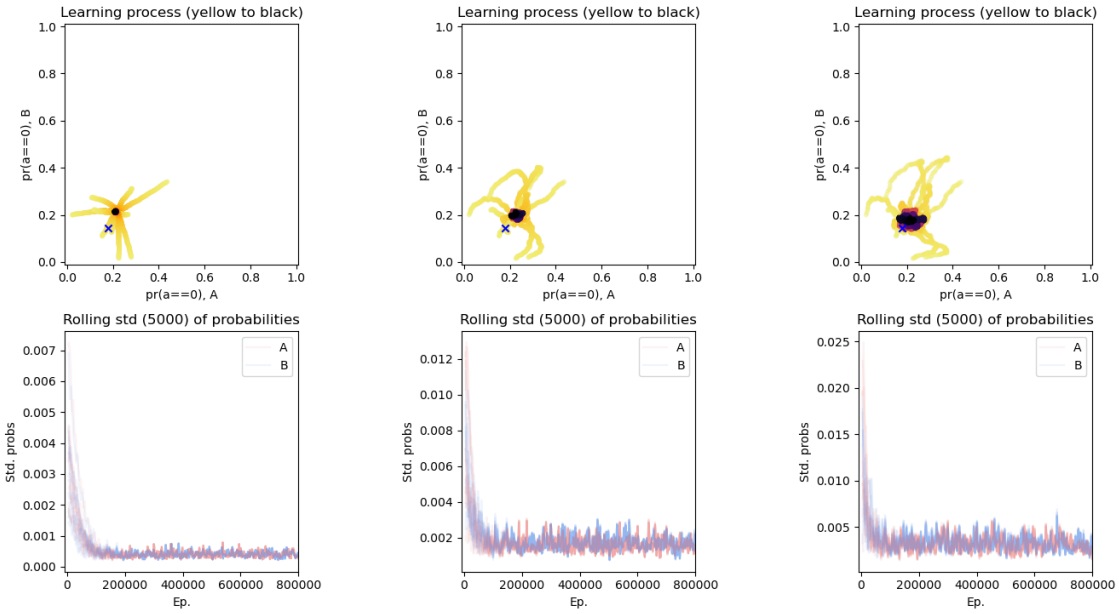
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.23: MBL-DPU in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations.



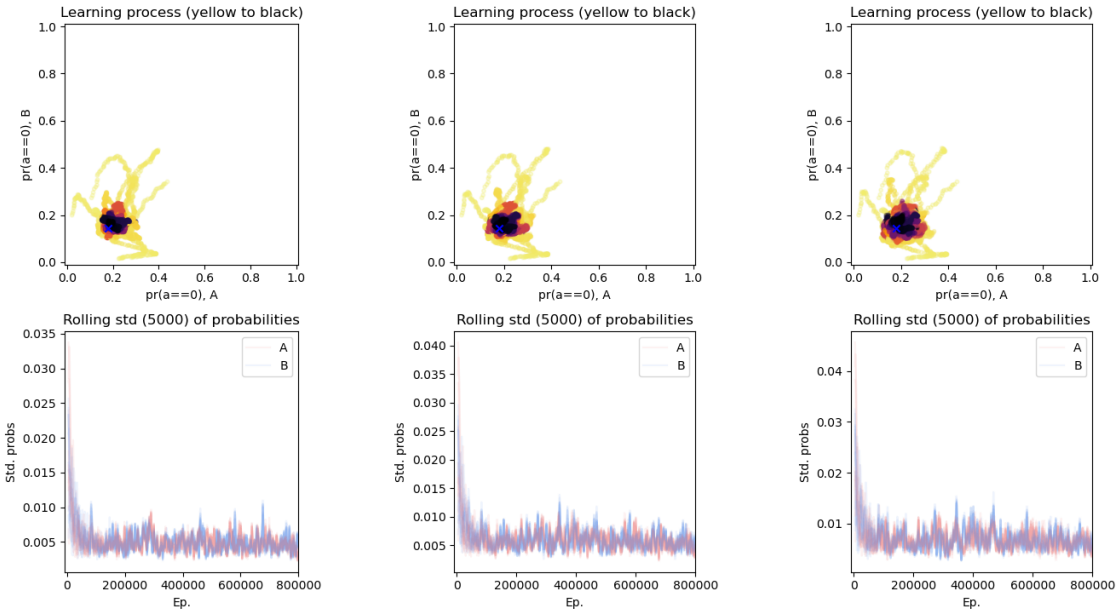
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.24: MBL-LC in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



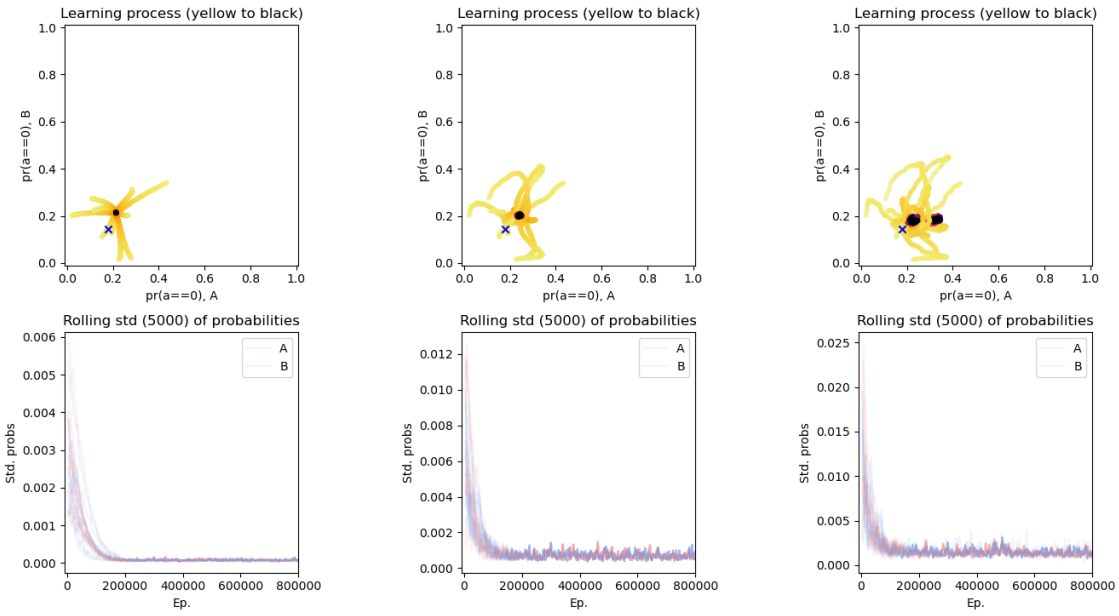
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.25: MBL-LC in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



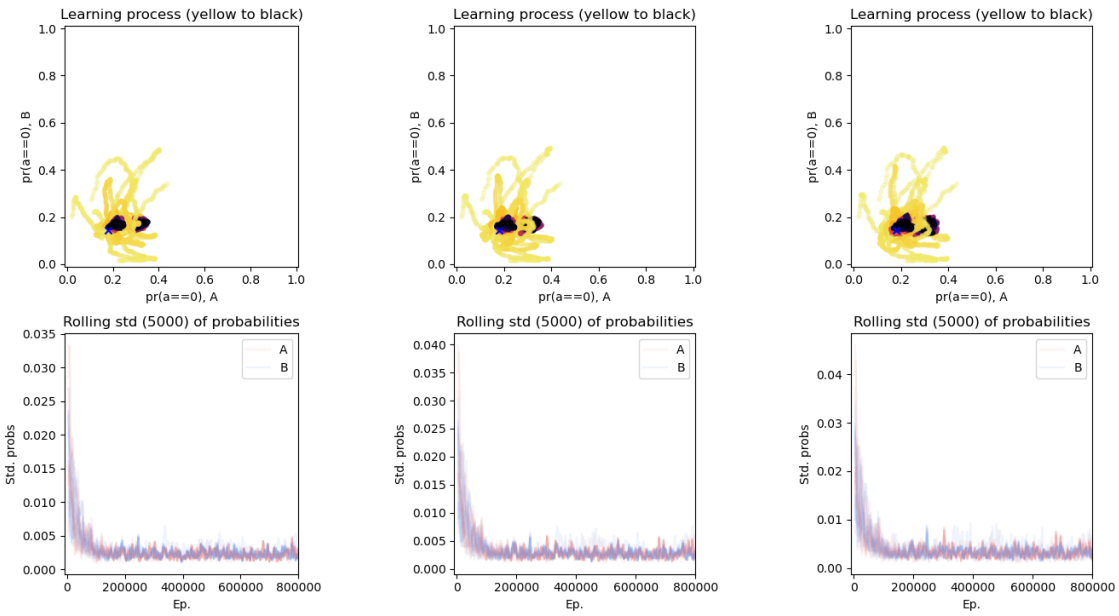
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.26: FAQ in self-play on the RPS-5 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations.



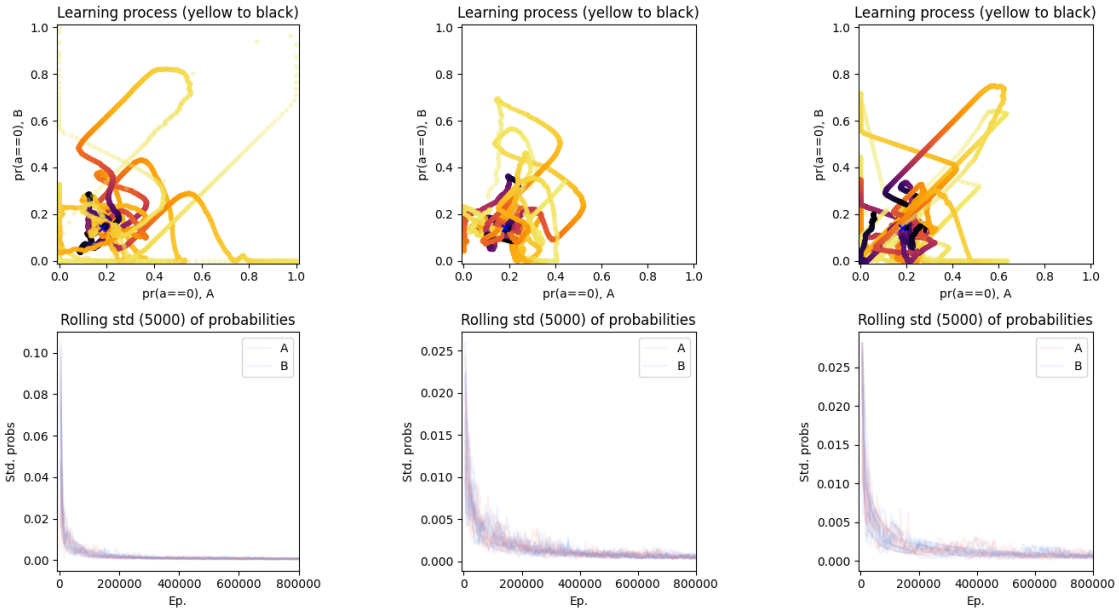
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.27: FAQ in self-play on the RPS-5 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.28: WoLF-PHC in self-play on the RPS-5 game with different learning schedules; for 10 different initialisations.

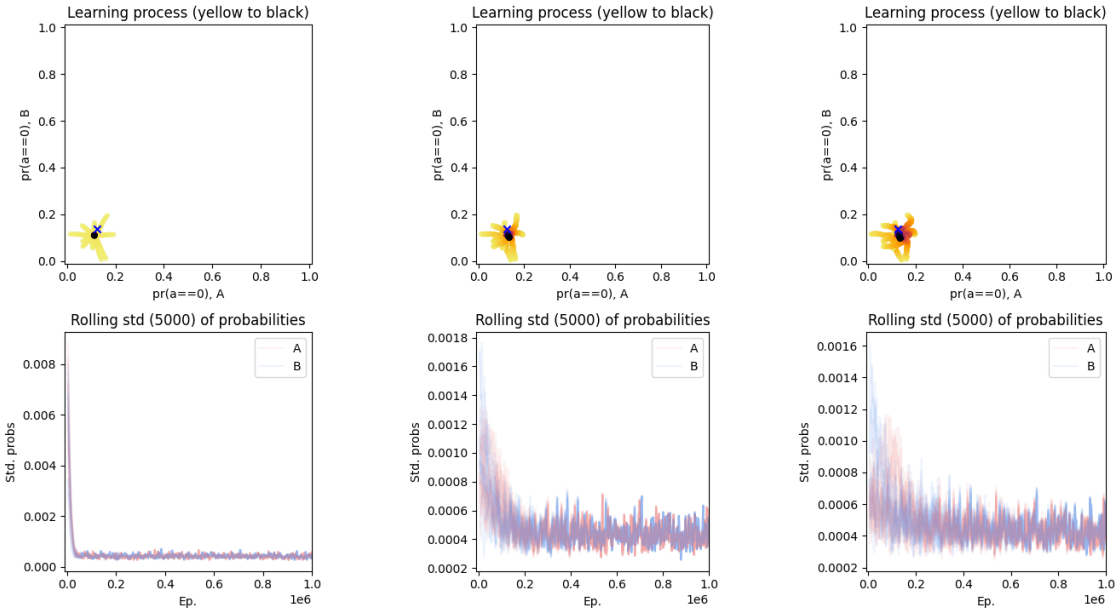


(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}.

(b) Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

(c) Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

Figure 6.29: MBL-DPU in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)

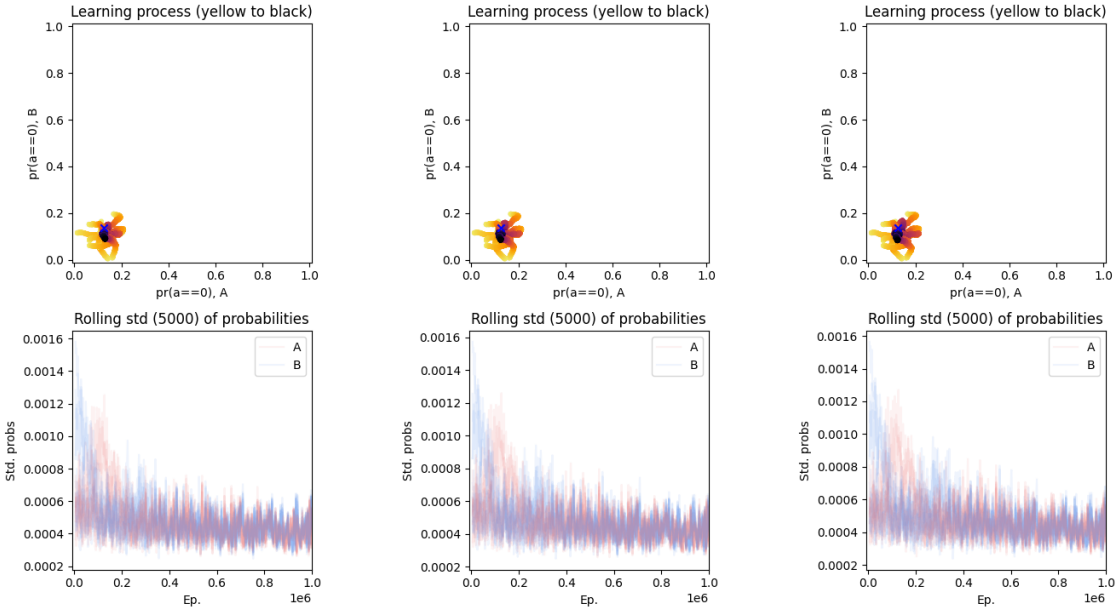


(a) $\tau = 1, M = 1^{-1}$

(b) $\tau = 10, M = 10^{-1}$

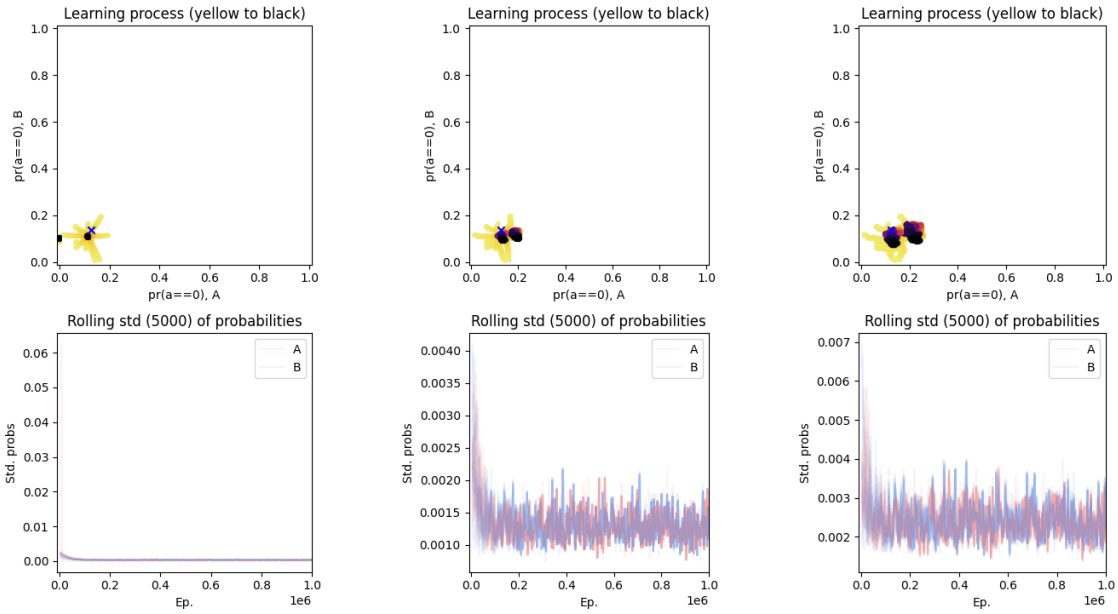
(c) $\tau = 20, M = 20^{-1}$

Figure 6.30: MBL-DPU in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations.



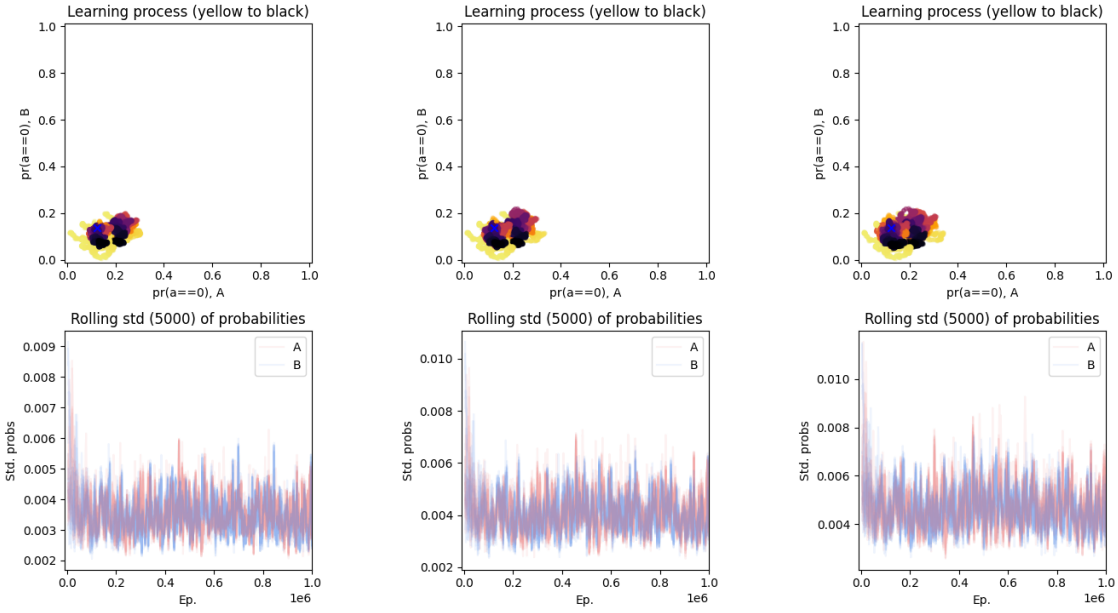
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.31: MBL-LC in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



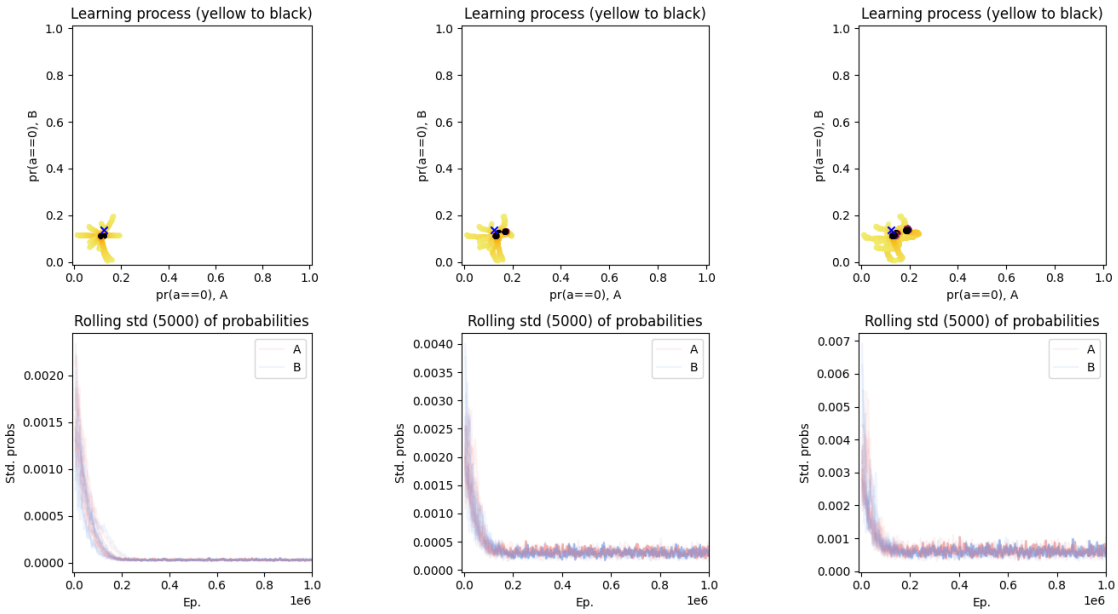
(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.32: MBL-LC in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations.



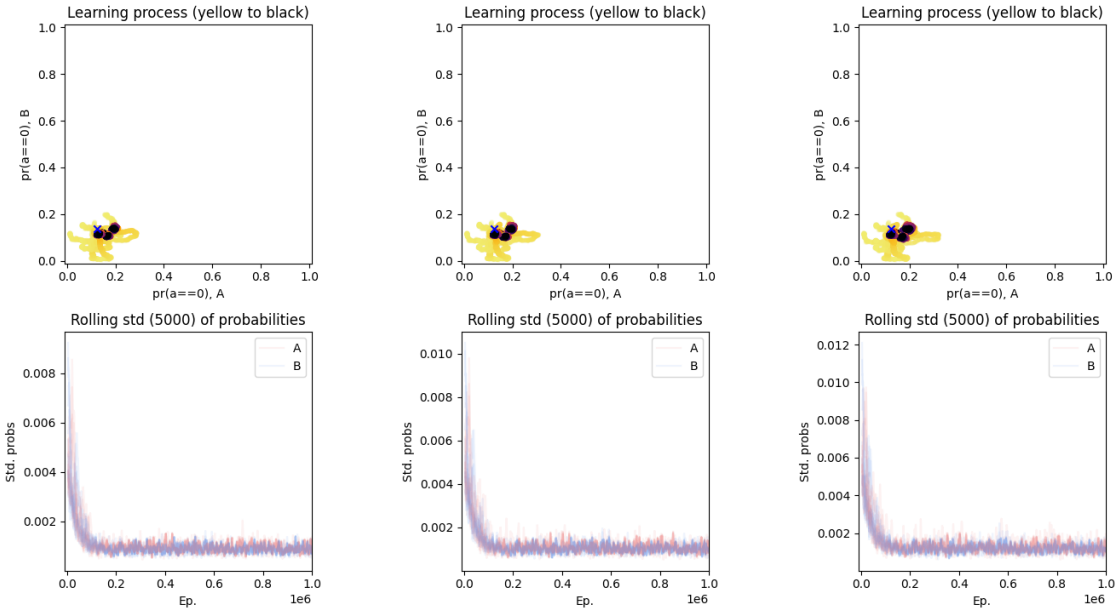
(a) $\tau = 30, M = 30^{-1}$ **(b)** $\tau = 35, M = 35^{-1}$ **(c)** $\tau = 40, M = 40^{-1}$

Figure 6.33: FAQ in self-play on the RPS-9 game with different values for τ (1, 10, 20) or M (1, 10^{-1}, 20^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) $\tau = 1, M = 1^{-1}$ **(b)** $\tau = 10, M = 10^{-1}$ **(c)** $\tau = 20, M = 20^{-1}$

Figure 6.34: FAQ in self-play on the RPS-9 game with different values for τ (30, 35, 40) or M (30^{-1}, 35^{-1}, 40^{-1}) equivalently; $\vartheta = 5 \cdot 10^{-3}$; for 10 different initialisations.

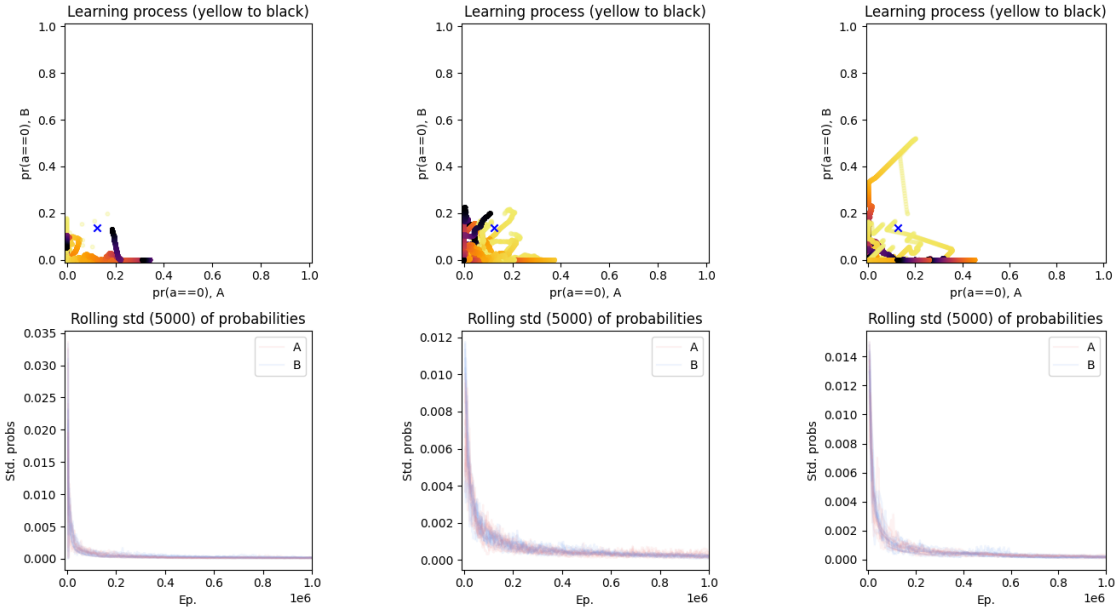


(a) $\tau = 30$, $M = 30^{-1}$

(b) $\tau = 35$, $M = 35^{-1}$

(c) $\tau = 40$, $M = 40^{-1}$

Figure 6.35: WoLF-PHC in self-play on the RPS-9 game with different learning schedules; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}.

(b) Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

(c) Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

6.3.3 Three-player matching pennies

Further, we consider the behaviour of the MBL variants in comparison to FAQ learning and WoLF-PHC in a three-player Matching Pennies (3MP) game introduced in [51], with payoffs as given in table 6.1. The similarity to the standard MP game becomes clear when one considers that the payoff structure reflects the following idea: The first player wants to match the second player's action. The second player wants to match the third player's action. However, the third player does not want to match the first player's action. The unique Nash equilibrium for 3MP is located at the centre of Δ. Note that, as initially proposed, 3MP is not a zero-sum game.

Table 6.1: Payoff tuples for the three-player Matching Pennies (3MP) game with the first player's action determining the row, the second player's action the column, and the third player's action the table.

	H	T
H	$(1,1,-1)$	$(-1,-1,-1)$
T	$(-1,1,1)$	$(1,-1,1)$

(a) Payoffs when the third player chooses 'H'.

	H	T
H	$(1,-1,1)$	$(-1,1,1)$
T	$(-1,-1,-1)$	$(1,1,-1)$

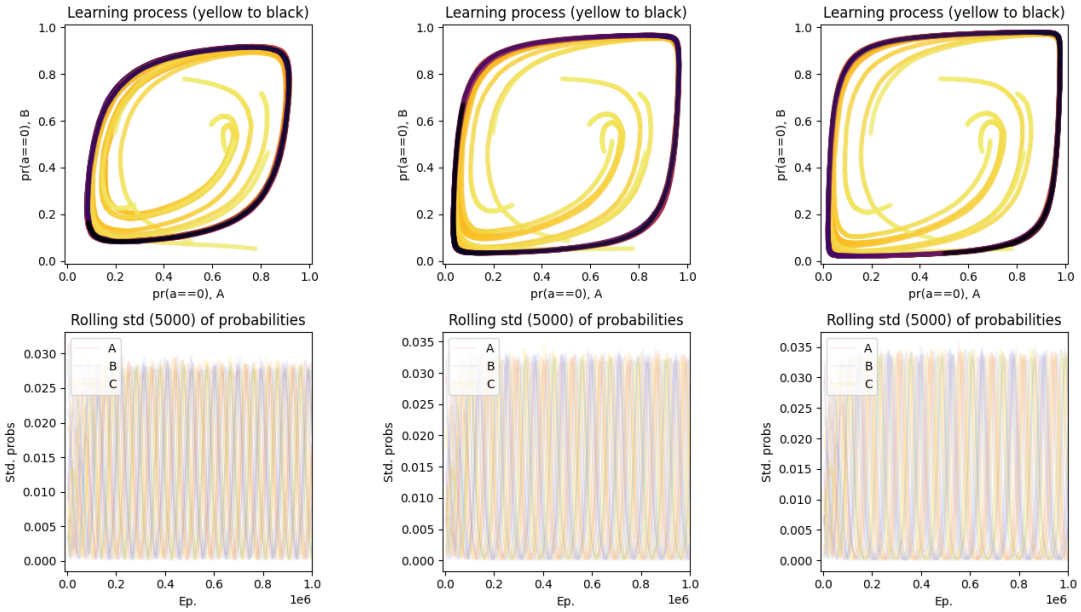
(b) Payoffs when the third player chooses 'T'.

In 3MP, both MBL variants (figures 6.36, 6.37) show apparently asymptotically stable periodic limit behaviours, which approach the boundary of Δ as mutation diminishes. MBL-DPU not approaching the Nash equilibrium is to be expected, since the Jacobian of (RD) has eigenvalues with positive real parts at the Nash equilibrium and hence it is not an attracting mutation limit due to lemma 4.4.11.[3] We further see a very similar behaviour for FAQ (figure 6.38) with τ^{-1} showing an analogous effect to M in MBL, quite similar to the two-player settings. Likewise, WoLF-PHC (figure 6.39) exhibits apparently asymptotically stable trajectories, at least in the projection onto the first actions of the first two players. Again, WoLF-PHC shows a reduction of variance over time, presumably due to diminishing learning rates. In [14], the authors show that WoLF-PHC converges to the Nash equilibrium when $\delta_l/\delta_w = 3$ (as opposed to $\delta_l/\delta_w = 2$). Since there is no established ODE approximation of WoLF-PHC that we are aware of, the reasons for this remain unclear as acknowledged in [14]. One should also note that we have made sure that the Nash equilibrium is not located at the centre of Δ in the two-player games because the perturbation term in FAQ has its equilibrium there and convergence might easily have been coincidental. For 3MP, we have not made any such adaptations and some behaviours might change

[3] The eigenvalues in question are easily calculated as -1 and $1/2(1 \pm i\sqrt{3})$.

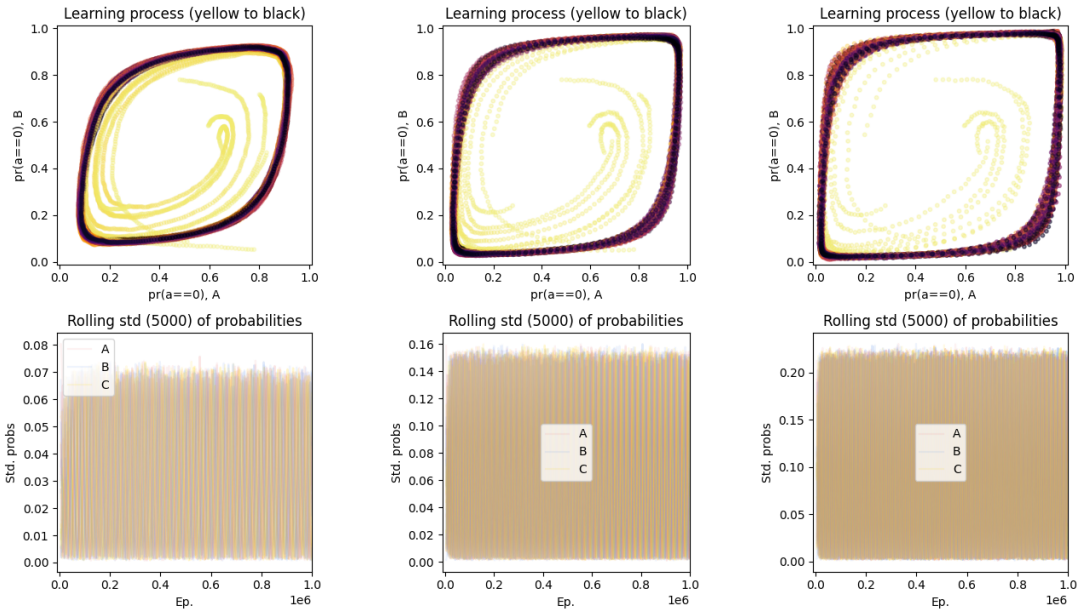
when the Nash equilibrium is moved away from the centre.

Figure 6.36: MBL-DPU in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1}, 20^{-1}, 30^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



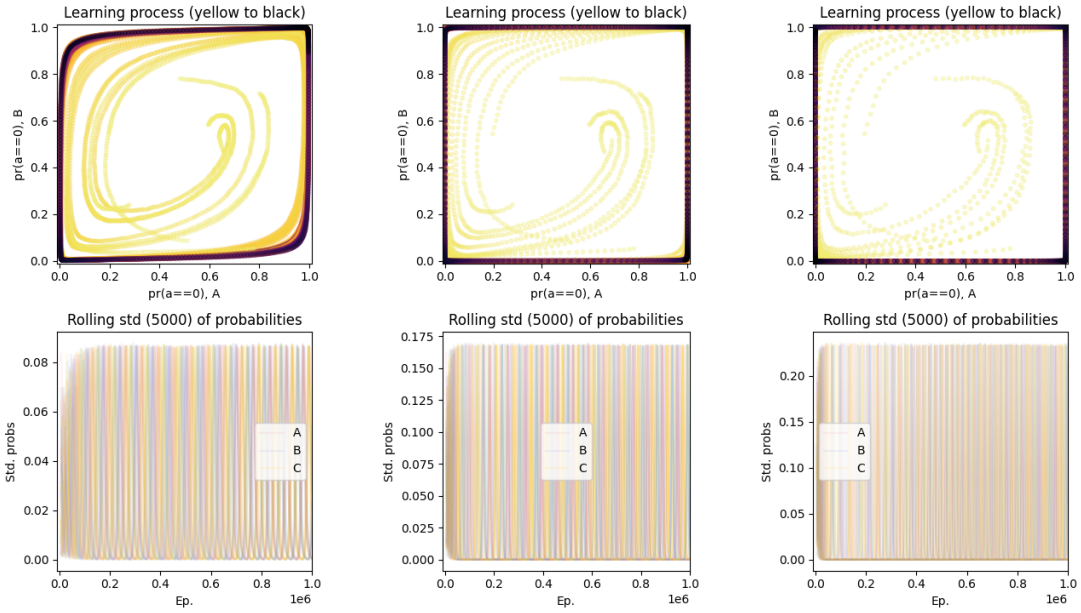
(a) $\tau = 10, M = 10^{-1}$ **(b)** $\tau = 20, M = 20^{-1}$ **(c)** $\tau = 30, M = 30^{-1}$

Figure 6.37: MBL-LC in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1}, 20^{-1}, 30^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



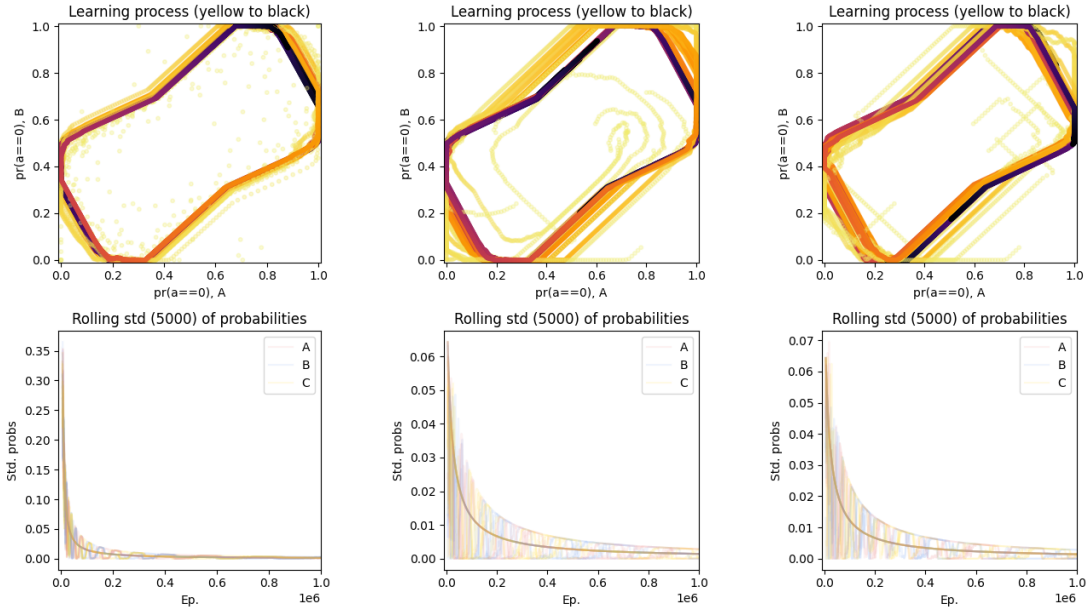
(a) $\tau = 10, M = 10^{-1}$ **(b)** $\tau = 20, M = 20^{-1}$ **(c)** $\tau = 30, M = 30^{-1}$

Figure 6.38: FAQ in self-play on the 3MP game with different values for τ (10, 20, 30) or M (10^{-1}, 20^{-1}, 30^{-1}) equivalently; $\vartheta = 10^{-4}$; for 10 different initialisations.



(a) $\tau = 10, M = 10^{-1}$ **(b)** $\tau = 20, M = 20^{-1}$ **(c)** $\tau = 30, M = 30^{-1}$

Figure 6.39: WoLF-PHC in self-play on the 3MP game with different learning schedules; for 10 different initialisations. (See figure 6.1 for a detailed explanation of the graphs.)



(a) Initial learning rate 10^{-1} for Q. Win learning rate 10^{-2}. **(b)** Initial learning rate 10^{-1} for Q. Win learning rate $1/2 \cdot 10^{-4}$. **(c)** Initial learning rate 10^{-2} for Q. Win learning rate $1/2 \cdot 10^{-4}$.

6.4 Discussion

The results on MBL established here relate both variants to the broad literature on mutli-agent reinforcement learning, in particular on Q-learning based variants. In particular, these results address the important question of whether and, if so, when Nash equilibria can be learned in a multi-agent setting. Although MBL-DPU shows slower convergence in simple settings compared to MBL-LC, FAQ and WoLF-PHC, it is also clear that it is among the simplest approaches and only slightly more complex than Cross learning. Similar to FAQ and WoLF-PHC, the MBL variants require agents neither to be able to observe and process others' actions nor payoffs, a significant distinction from, e.g., joint-action learning algorithms. Furthermore, its simplicity allows us to precisely specify the relation to (RMD), which in turn allows a deeper understanding of the behaviour of MBL-DPU in different settings and allows us to employ methods from evolutionary game theory in the study of MARL algorithms in a mathematically rigorous way. For instance, it is clear that hyperbolic equilibria of (RD) will remain hyperbolic in (RMD) for small M while neutrally stable equilibria in zero-sum games tend to be stabilised by mutation, such that the behaviour of MBL-DPU can be anticipated in specific situations.

This qualitative difference in understanding is illustrated by the results for the RPS variants where the more complex algorithms are initially well-behaved, i.e., approach regions close to the Nash equilibrium, but for larger actions sets start exhibiting more complex trajectories which do not approach the Nash equilibrium, while MBL-DPU shows almost no qualitative changes in the larger RPS variants. In particular, due to the relation to (RMD), we can check whether MBL-DPU would not converge for any choice of parameters, whereas it is difficult to decide, whether the loss of convergence to the Nash equilibrium for MBL-LC, FAQ and WoLF-PHC is a matter of principle or whether there are parameters to recover learning. This is illustrated in the 3MP game, where all algorithms fail to learn for the chosen parameters. For concrete games it is however possible, e.g., to check the eigenvalues of the equilibria in order to determine convergence of MBL-DPU, as opposed to the more complex algorithms. For algorithms based on Q-learning, it is further important that at each time point the Q-values are a reliable estimate for the current situation, which quickly draws in questions from stochastic approximation theory addressed in [58].

In this sense, the results present an answer to the question of what the conceptu-

ally simplest approach to MARL algorithms should contain in order for convergence towards sensible equilibria to be rigorously guaranteed at least in some classes of settings. This consideration of a conceptual lower bound, loosely speaking, is warranted if one is interested in seeking a mathematically rigorous foundation and a general understanding beyond experimental benchmark simulations.

6.5 Proofs of propositions 6.2.1 and 6.2.2

The proofs employ a result proved in [72, p. 118], which we state in the following and then proceed to prove propositions 6.2.1 and 6.2.2.

6.5.1 A theorem on learning with small steps

The result from [72] we employ is phrased in the following situation: Let $J \subset \mathbb{R}_{>0}$ be a parameter set with $\inf J = 0$ and $N \in \mathbb{N}$, such that for every $\vartheta \in J$, $\{X_n^\vartheta\}_{n\geq 0} \subset I_\vartheta$ is a Markov process with stationary probabilities and $I_\vartheta \subset \mathbb{R}^N$. We denote by $\mathbb{E}_x[X_n^\vartheta]$ the expected value of X_n^ϑ given $X_0^\vartheta = x$. Let further I be the minimal closed convex set with $\bigcup_\vartheta I_\vartheta \subset I$. Define

$$H_n^\vartheta = \Delta X_n^\vartheta / \vartheta$$

and let $w(x,\vartheta)$, $S(x,\vartheta)$, $s(x,\vartheta)$ and $r(x,\vartheta)$ for $(x,\vartheta) \in I \times J$ be given as:

$$\begin{aligned}
w(x,\vartheta) &= \mathbb{E}[H_n^\vartheta | X_n^\vartheta = x] \in \mathbb{R}^N \\
S(x,\vartheta) &= \mathbb{E}[\,(H_n^\vartheta)^2 | X_n^\vartheta = x] \in \mathbb{R}^{N\times N} \\
s(x,\vartheta) &= \mathbb{E}[\,(H_n^\vartheta - w(x,\vartheta))^2 | X_n^\vartheta = x] = S(x,\vartheta) - w^2(x,\vartheta) \in \mathbb{R}^{N\times N} \\
r(x,\vartheta) &= \mathbb{E}[\|H_n^\vartheta\|^3 | X_n^\vartheta = x] \in \mathbb{R}\,.
\end{aligned}$$

where $x^2 = xx^T$ and $\|x\| = \sqrt{x^T x}$ for $x \in \mathbb{R}^N$.

We can now state theorem 8.1.1 from [72, p. 118] (omitting part (C)):

Theorem 6.5.1 (Norman)**.** *In the above situation, let the following conditions be satisfied:*

The family of sets $(I_\vartheta)_\vartheta$ *satisfies*

$$\forall x \in I \,:\, \lim_{\vartheta\to 0} \inf_{y\in I_\vartheta} \|x - y\| = 0\,. \tag{a.1}$$

There are functions w and s on I such that:

$$\sup_{x\in I_\vartheta} \|w(x,\vartheta) - w(x)\| \in \mathcal{O}(\vartheta)\,, \tag{a.2}$$

$$\sup_{x\in I_\vartheta} \|s(x,\vartheta) - s(x)\| \to 0 \text{ for } \vartheta \to 0\,, \tag{a.3}$$

where $\mathcal{O}$ refers to the Bachmann–Landau notation.

The function w is differentiable, i.e., there is a function w' such that for all $x \in I$:

$$\lim_{\substack{y\to x\\ y\in I}} \frac{\|w(y) - w(x) - w'(x)(y-x)\|}{\|y-x\|} = 0\,. \tag{b.1}$$

The function w' is bounded:

$$\sup_{x\in I} \|w'(x)\| < \infty\,. \tag{b.2}$$

The functions w' and s satisfy the Lipschitz condition:

$$\sup_{x,y\in I, x\neq y} \frac{\|w'(x) - w'(y)\|}{\|x-y\|} < \infty\,, \tag{b.3}$$

$$\sup_{x,y\in I, x\neq y} \frac{\|s(x) - s(y)\|}{\|x-y\|} < \infty\,. \tag{b.4}$$

The function r is bounded:

$$\sup_{\vartheta\in J, x\in I_\vartheta} r(x,\vartheta) < \infty\,. \tag{c}$$

Let further for $\vartheta \in J$ and $x \in I_\vartheta$, $\mu_n(\vartheta,x) = \mathbb{E}_x[X_n^\vartheta]$ and $\omega_n(\vartheta,x) = \mathbb{E}_x[\|X_n^\vartheta - \mu_n(\vartheta,x)\|^2]$.

In this case, the following hold:

(A) $\omega_n(\vartheta,x) \in \mathcal{O}(\vartheta)$ uniformly in $x \in I_\vartheta$ and $n\vartheta \le T$ for any $T < \infty$.

(B) For any $x \in I$, the differential equation

$$f'(t) = w(f(t))$$

has a unique solution $f(t) = f(t,x)$ with $f(0) = x$. For all $t \ge 0$, we have $f(t) \in I$, and

$$\mu_n(\vartheta,x) - f(n\vartheta,x) \in \mathcal{O}(\vartheta)$$

uniformly in $x \in I_\vartheta$ and $n\vartheta \le T$.

Remark 6.5.2. We note that parts (A) and (B) imply that for all $\varepsilon > 0$,

$$\Pr(\|X_n^\vartheta - f(T,x)\| > \varepsilon) \to 0$$

for $n\vartheta \to T$, $\vartheta \to 0$, and given that $X_0^\vartheta = x$ almost certainly for all ϑ.

6.5.2 Convergence of MBL-DPU

We restate the simple reinforcement-mutation rule of MBL-DPU, denoting the mixed strategies with an upper-case X to underscore that this is a random variable and denoting the dependence on a parameter ϑ, denoting the whole family of stochastic processes as $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0}$. Let $U(x) = (U_{ih}(x))_{i,h}$ be a discrete, non-negative random variable whose probability distribution depends on $x \in I$ with its support being independent of x, and let $M < \overline{M}$ for some upper bound $\overline{M} < \infty$.

For an agent i and a chosen strategy h, the update rule then is given as follows:

$$\begin{aligned} X_{ih}^{\vartheta}(n+1) &= X_{ih}^{\vartheta}(n) + \vartheta\left((1 - X_{ih}^{\vartheta}(n))U_{ih}(X^{\vartheta}(n))\right) + \vartheta M\left(c_{ih} - X_{ih}^{\vartheta}(n)\right) \\ X_{ik}^{\vartheta}(n+1) &= X_{ik}^{\vartheta}(n) + \vartheta\left((-X_{ik}^{\vartheta}(n))U_{ih}(X^{\vartheta}(n))\right) + \vartheta M\left(c_{ik} - X_{ik}^{\vartheta}(n)\right) \quad \text{for } k \neq h\,. \end{aligned} \tag{6.5.1}$$

We can now show that this rule indeed approximates the replicator-mutator dynamics for $\vartheta \to 0$ in the sense of remark 6.5.2:

Proposition 6.5.3. *There is J such that the stochastic processes $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0}$ given by* (6.5.1) *approximates the replicator-mutator dynamics for $\vartheta \to 0$ in the sense of remark 6.5.2 if $X^{\vartheta}(0) \in I$ for all $\vartheta \in J$.*

Proof. The proof proceeds by showing that $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0}$ satisfies the conditions of theorem 6.5.1. For an agent i and a chosen strategy h, we have:

$$\begin{aligned} H_{ih}^{\vartheta}(n+1) &= \Delta X_{ih}^{\vartheta}(n+1)/\vartheta = (1 - X_{ih}^{\vartheta}(n))U_{ih}(X^{\vartheta}(n)) + M(c_{ih} - X_{ih}^{\vartheta}(n)) \\ H_{ik}^{\vartheta}(n+1) &= \Delta X_{ik}^{\vartheta}(n+1)/\vartheta = -X_{ik}^{\vartheta}(n)U_{ih}(X^{\vartheta}(n)) + M(c_{ik} - X_{ik}^{\vartheta}(n)) \quad \text{for } k \neq h \end{aligned}$$

Note that in this case, $H_{ih}^{\vartheta}(n+1)$ is independent of ϑ if $X^{\vartheta}(n)$ is given, which simplifies the analysis. Let us set $u_{ih}(x) = \mathbb{E}[U_{ih}(X^{\vartheta}(n))|X^{\vartheta}(n) = x]$, where it is clear that there is no dependence on n. Note that u is polynomial in the components of x and hence smooth.

Condition (a.1): In our case, I is given as the polyhedron $\bigtimes_i \Delta_i$ and $I_{\vartheta} = I$ for all ϑ and thus condition (a.1) is satisfied. It remains to show that $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0} \subset I$. Note that U_{ih} is a discrete non-negative random variable and thus bounded by some $C < \infty$. For $\vartheta < (C + \overline{M})^{-1}$, we have $\vartheta M < 1$. Assume that $X_{ih}^{\vartheta}(n) = x \in I$, then for an agent i and a chosen strategy h we have

$$\begin{aligned} X_{ih}^{\vartheta}(n+1) &= x_{ih} + \vartheta\left((1 - x_{ih})U_{ih}(n+1) + M(c_{ih} - x_{ih})\right) \\ &= x_{ih}(1 - \vartheta M) + \vartheta(1 - x_{ih})U_{ih}(n+1) + \vartheta M c_{ih} \geq 0 \end{aligned}$$

and for every other strategy $k \neq h$, we have

$$\begin{aligned} X_{ik}^{\vartheta}(n+1) &= x_{ik} + \vartheta\left((-x_{ik})U_{ih}(n+1) + M(c_{ik} - x_{ik})\right) \\ &= x_{ik}(1 - \underbrace{\vartheta(U_{ih}(n+1) + M)}_{<1}) + \vartheta M c_{ik} \geq 0\,. \end{aligned}$$

A simple calculation shows that $\sum_k X_{ik}^{\vartheta}(n+1) = 1$ if $x \in I$. Thus we have that $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0} \subset I$ if $X^{\vartheta}(0) \in I$ for all ϑ and we can choose $J = (0, (C + \overline{M})^{-1})$.

Conditions (a.2) & (a.3): Consider first the function w:

$$\begin{aligned} w_{ih}(x,\vartheta) &= E[H^{\vartheta}(n)|X^{\vartheta}(n) = x] \\ &= x_{ih}(1 - x_{ih})E[U_{ih}(n+1)|X^{\vartheta}(n) = x] + x_{ih}M(c_{ih} - x_{ih}) \\ &\quad + \sum_{k\neq h} x_{ik}(-x_{ih})E[U_{ik}(n+1)|X^{\vartheta}(n) = x] + x_{ik}M(c_{ih} - x_{ih}) \\ &= x_{ih}\left(u_{ih}(x) - \sum_k x_{ik}u_{ik}(x)\right) + M(c_{ih} - x_{ih}) \end{aligned}$$

It is clear that w does not depend on ϑ and that condition (a.2) is trivially satisfied. Similarly, $S(x,\vartheta)$ and $s(x,\vartheta)$ do not depend on ϑ and condition (a.3) is trivially satisfied.

Conditions (b.1)–(b.4): Since the function u is smooth, so is w. In particular, we have that $\sup_{x\in I}\|w'(x)\| < \infty$ because I is compact and w' is continuously differentiable, from which follows that w' satisfies the Lipschitz-condition (b.3) on I. Similarly, s is smooth and satisfies (b.4).

Condition (c): Again, r does not depend on ϑ, and is smooth on I, which is compact. Thus it is bounded on I and condition (c) is satisfied.

As a consequence, we can apply theorem 6.5.1 to the family $\{X^{\vartheta}(n)\}_{n\geq 0}$ and with remark 6.5.2 we have that for all $\varepsilon > 0$,

$$\Pr(\|X^{\vartheta}(n) - f(T,x)\| > \varepsilon) \to 0$$

for $n\vartheta \to T$, $\vartheta \to 0$, and given that $X^{\vartheta}(0) = x$ for all ϑ, where for all i and h, f is the unique solution to the differential equations

$$f'_{ih}(t) = w_{ih}(f(t)) = f_{ih}(t)\left(u_{ih}(f(t)) - \sum_k f_{ik}(t)u_{ik}(f(t))\right) + M(c_{ih} - f_{ih}(t))$$

with $f(0) = x$. □

Proposition 6.5.4. *Let x^M be an equilibrium of* (RMD) *and U an open neighbourhood of x^M. If x^M is globally asymptotically stable, then there is $\vartheta > 0$ such that the stochastic*

process $\{(X_{ih}^{\vartheta}(n))_{i,h}\}_{n\geq 0}$ *defined in* (6.5.1) *visits* U *almost surely after finitely many steps.*

Proof. Let $\Phi^M(x,\cdot) : \mathbb{R}_{\geq 0} \to \Delta$ satisfy (RMD) with $\Phi^M(x,0) = x$ for all $x \in \Delta$. Let further $U' \subset U$ such that $\bigcup_{x\in U'} B_\delta(x)$ for some $\delta > 0$, where $B_\delta(x)$ denotes an open ball with radius δ around x. As x^M is globally asymptotically stable, there is for each $x \in \Delta$ a $t' < \infty$ such that for all $t > t'$: $\Phi^M(x,t) \in U'$.

This is because there is a neighbourhood $V \subset U'$ of x^M such that $\forall x^0 \in V, t > 0 : \Phi^M(x^0,t) \in U'$ due to the Lyapunov stability of x^M. Since x^M is asymptotically stable, for every x there is a $t > 0$ such that $\Phi^M(x,t) \in V$ and hence the solution will remain in U' afterwards.

Therefore, define $\tau : \Delta \to \mathbb{R}$ such that:

$$\tau(x) = \inf\{T > 0 : \ \Phi^M(x,T) \in V\}$$

Since the RHS of (RMD) is continuously differentiable by assumption, it is also Lipschitz continuous. Thus, Φ is continuous in the first argument and so is τ as the following argument shows:

Let $x \in \Delta$ and $\varepsilon_1 > 0$. Then there is $t > \tau(x)$ such that $\Phi^M(x,s) \in V$ for $s \in (\tau(x),t]$. Choose $s \in (\tau(x),t]$ such that $|\tau(x) - s| < \varepsilon_1$. Then $\Phi^M(x,t) \in V$ and there is a neighbourhood U_x of x such that for all $y \in U_x$, $\Phi^M(y,s) \in V$. Hence $\tau(y) < s < \tau(x) + \varepsilon_1$.

We also have $\tau(y) > \tau(x) - \varepsilon_1$ due to the following: Consider $d := \inf\{\|\Phi^M(x,\tau(x) - \varepsilon_1) - v\| : v \in V\} > 0$. Note that the Lipschitz condition implies that there is $L > 0$ such that for all $t > 0$ and all $y \in \Delta$

$$\|\Phi^M(x,t) - \Phi^M(y,t)\| \leq \|x - y\| e^{Lt}$$

and for all $t \in [0, \tau(x) - \varepsilon_1]$

$$\|\Phi^M(x,t) - \Phi^M(y,t)\| \leq \|x - y\| e^{L(\tau(x)-\varepsilon_1)}$$

and w.l.o.g. we can assume that $\forall y \in U_x$, we have $\|x - y\| e^{L(\tau(x)-\varepsilon_1)} < \frac{d}{2}$. Thus we have for all $v \in V$

$$\begin{aligned} 0 < d \leq \|\Phi^M(x,t) - v\| &= \|\Phi^M(x,t) - \Phi^M(y,t) + \Phi^M(y,t) - v\| \\ &\leq \|\Phi^M(x,t) - \Phi^M(y,t)\| + \|\Phi^M(y,t) - v\| \\ &\leq \|x - y\| e^{L(\tau(x)-\varepsilon_1)} + \|\Phi^M(y,t) - v\| < \frac{d}{2} + \|\Phi^M(y,t) - v\| \end{aligned}$$

and so for all $y \in U_x$, we have $\inf\{\|\Phi^M(y,t)-v\| : v \in V, t \in [0,\tau(x)-\varepsilon_1]\} \geq \frac{d}{2} > 0$ and thus $\tau(y) > \tau(x) - \varepsilon_1$. So τ is continuous on Δ. Let then $T := \sup_{x\in\Delta} \tau(x) < \infty$. Note that for all $x \in \Delta$ we have that for all $t > T$, $\Phi^M(x,t) \in U'$ and $B_\delta(\Phi^M(x,t)) \subset U$.

Let further $\eta > 0$. Then with proposition 6.5.3, there are $\vartheta > 0$, $n_\vartheta \in \mathbb{N}$ such that for all $x \in \Delta$,

$$\Pr(X^\vartheta(n_\vartheta) \in B_\delta(\Phi^M(x,T)) \subset U | X^\vartheta(0) = x) > \eta$$

and so

$$\Pr(X^\vartheta(n_\vartheta) \in U) > \eta.$$

From here it is easy to see that the first hit time of U for $\{X^\vartheta(t)\}_{t\in\mathbb{N}_0}$ is almost surely finite, i.e., the earliest time t for which $X^\vartheta(t) \in U$: Let $Z(k) := X^\vartheta(kn_\vartheta)$ for $k \in \mathbb{N}_0$ and let S be the first hit time of U for $\{Z(k)\}_{k\in\mathbb{N}_0}$, such that S is a random variable with values in $\mathbb{N}_0 \cup \{\infty\}$. Clearly the first hit time of U for $\{X^\vartheta(t)\}_{t\in\mathbb{N}_0}$ is smaller than for $\{Z(k)\}_{k\in\mathbb{N}_0}$.

We have that for all $z \in \Delta$ and all $k \in \mathbb{N}$:

$$\Pr(Z_{k+1} \in B_\delta(\Phi^M(z,T)) \subset U | Z_k = z) > \eta$$

and hence

$$\Pr(Z_{k+1} \in U) > \eta.$$

Then we have for S,

$$\Pr(S \leq k+1) = \Pr(S \leq k) + (1 - \Pr(S \leq k)) \Pr(Z_{k+1} \in U) > \Pr(S \leq k)(1-\eta) + \eta$$

and a quick induction argument yields

$$\Pr(S \leq k+1) > 1 - (1-\eta)^k (1 - (1-\eta)\Pr(S=0))$$

The probability of a finite hitting time is then:

$$\Pr(S \in \mathbb{N}_0) = \lim_{k\to\infty} \Pr(S \leq k+1) \geq 1 - \lim_{k\to\infty} (1-\eta)^k (1 - (1-\eta)\Pr(S=0)) = 1$$

In particular, the hitting time of U for $\{X^\vartheta(t)\}_{t\in\mathbb{N}_0}$ is finite almost surely. □

It immediately follows that:

Corollary 6.5.5. *If x^M is a globally asymptotically stable equilibrium of* (RMD) *and U an open neighbourhood of x^M, then there is $\vartheta > 0$ such that the stochastic process $\{X^\vartheta(n)\}_{n\geq 0}$ defined in* (6.5.1) *visits U infinitely often almost surely.*

Proof. Consider for any finite $t' \in \mathbb{N}_0$ the probability that $\{X^{\vartheta}(n)\}_{n\geq 0}$ will not visit U afterwards. This is clearly the same as the probability that the process $\{Z^{\vartheta}(n)\}_{n\geq 0}$ induced by (6.5.1) and starting in $X^{\vartheta}(t')$, i.e., $Z^{\vartheta}(0) = X^{\vartheta}(t')$ almost surely, will not visit U at all. The previous proposition shows that this probability is 0, which concludes the proof. □

7 Discussion

The present thesis has striven to formulate a common perspective on evolutionary and learning dynamics in multi-agent settings and to demonstrate that a common perspective can help in furthering the understanding of the parallels and the differences of these. Besides the formal correspondence between the evolution of interacting populations and the evolution of players' mixed strategies, we have shown that evolutionary dynamics can indeed provide a sensible foundation for learning algorithms. Similar ideas have been explored with respect to Cross learning and the unperturbed multi-population (RD) in [13]. However, we are not aware of attempts to provide rigorous results regarding the stability of multi-population (RD) in order to enable the study of convergence in corresponding multi-agent reinforcement learning algorithms nor of attempts to formulate learning in artificial neural networks in the language of evolutionary game dynamics. Understanding the stability properties and limitations of the unperturbed (RD) also helps to understand the inevitable difficulties which Cross learning encounters in simple games already, e.g., when the only Nash equilibrium is in the interior, in which case there can be no asymptotic stability of the equilibrium. The proof ideas for this as well as the general ideas of evolution have pointed towards the significant effect mutation has in this respect. An often used approach in evolutionary game theory is to assume that we can make a separation of time-scales between mutation and selection, where we assume that selection happens faster than mutation such that individuals are perfect replicators. We have considered how relaxing this assumption affects the stability properties of the dynamics. In particular, we have derived from the general replicator-mutator equation [77] an infinite background-fitness formulation in which selection and mutation become additive. Our further assumption of memoryless mutation has led to mutation becoming a linear perturbation on the multi-population replicator dynamics. These assumptions have allowed us to study the dependence of the dynamics on the choice of mutation parameters and to prove that certain properties are independent of this choice, i.e., to prove the existence of mutation limits. Further, we have formulated the concept

of attracting mutation limits relating the equilibria of the unperturbed (RD) to the stability properties of the mutation equilibria under a range of mutation parameters, i.e., any choice of $c \in \Delta^o$ and for every positive M in some open interval around 0. We were further able to demonstrate that this setting allows us to prove the existence of attracting mutation limits for all two-player zero-sum games and for all 2×2 games, including such with non-linear fitness functions.

The explicit consideration of mutation has enabled us to strengthen the relation between evolutionary dynamics and learning by exploiting the relation between (RD) and Cross learning and formulating the MBL algorithm as a parallel to (RMD) and we proved that MBL-DPU is describable by (RMD) in a precise sense. This and the analytical results regarding stability allowed us to compare MBL-DPU in numerical experiments to other well-known MARL algorithms, which perform well in certain cases but lack the theoretical guarantees we have established for MBL-DPU. The significance of these guarantees was demonstrated by the deterioration of the performance of the other algorithms in the RPS-5 and RPS-9 settings, in which MBL-DPU continued to show the theoretically predicted behaviour. Furthermore, a simple eigenvalue analysis of the Nash equilibrium of the three-player MP game in (RD) would show that the Jacobian of (RD) has eigenvalues with positive real parts at the Nash equilibrium and hence is not an attracting mutation limit. Therefore, convergence of MBL-DPU towards the Nash equilibrium should not be expected and does indeed not occur. This systematic relation between a mathematically tractable ODE system and the learning algorithm is not given for FAQ or WoLF-PHC and hence it is not clear to what to attribute their behaviour and to anticipate how these would behave in a range of situations.

Overall, we have shown that evolutionary processes can be qualitatively sensitive to assumptions about mutation—and we have imposed some assumptions ourselves where the sensitivity of results on these assumptions is still an open question. Given that evolutionary game theoretical approaches are used to study a large variety of phenomena including the emergence of cooperation, conflict or multi-cellularity, it seems important to point out that the models employed can yield qualitatively different outcomes when strong assumptions are relaxed only slightly. Of course, the present results were obtained under a number of assumptions, e.g., frequency dependent selection, specific mutation mechanisms, a focus on interspecific interactions which are independent of intraspecific effects and asexual reproduction among oth-

ers. Where such assumptions are appropriate for biological populations, the present results imply that even simple mutation can have a significant effect on the long-term evolution of populations. In such settings, the deterministic fitness value of a trait might be thought of as the expected reproductive output of individuals with that trait, resulting from stochastic discrete interactions given the current population compositions. A more complex population structure, e.g., accounting for life stages, would require a suitable extension of the current formulation and the results should not be expected to translate to such a setting without considerable effort, if at all, due to the fact that we assume no intraspecific effects. This would probably be violated by additional life stage transition dynamics resulting in a different dynamics, e.g., as derived in [3]. If life stages remain simple and interactions happen only at one stage, there is hope of simplifying the dynamics. However, this should still be expected to result in a system of delay-differential equations structurally sufficiently distinct from the system considered here. Although, depending on the specific structure, some results or techniques might be translatable.

We have further illustrated the benefits of considering evolutionary processes and learning processes side-by-side and establishing formally rigorous relations between idealised ODE models and discrete stochastic processes as they inevitably arise in computational settings. This side-by-side consideration illustrates the parallels between the role of mutation in evolutionary processes and that of exploration in dynamic programming algorithms, such as reinforcement learning, showing a similar trade-off between accuracy and convergence. Overall, these results demonstrate that such a common perspective can indeed be fruitful and beneficial for both areas of inquiry. It should be noted, of course, that this common perspective is an abstract one and does not imply that an actual *biological* process of natural selection occurs when individual organisms learn. It is rather highlighting that both, reinforcement learning and natural selection, can be modelled by differential increases of the probability masses of game theoretic strategies resulting from interaction outcomes.

Additionally, analysis of the replicator-mutator dynamics has shown that there are at least formal parallels with so-called interior-point methods in constrained optimization, also known as barrier methods. In particular, the mutation strength M resembles the duality gap in interior-point methods (IPMs). It would therefore be sensible for future research to specify the relation between (RMD) and MBL-DPU on the one hand and IPMs on the other hand. In particular, IPMs have a well established

theory of when they converge and well-established schedules for the reduction of the duality gap. As we have seen, reducing M leads to the mutation equilibrium moving closer to the Nash equilibrium, but reducing it too quickly slows down convergence unnecessarily. In this respect, the results and methods related to duality gap reduction schedules in IPMs can provide a guideline for sensible reduction schedules for M in MBL-DPU.

Finally, a main motivation for the present analyses is the study of the evolution of interaction networks and in particular the evolution of network topologies induced by population strategies. Therefore, the explicit consideration of more complex strategies, in particular with state dependencies, is a main area of further development of the present approach. We have preliminarily considered interacting populations at the beginning of this thesis. Although we have not further specified such interactions in the formulation of the replicator-mutator dynamics, we have taken care to keep the formulation general enough to include multiple interacting populations and to include potential dependencies of strategies on observed actions of other players. Any future analysis including an explicit dependence of strategies will significantly benefit from the results we have obtained, allowing us to approach the study of evolving topologies. A main question in specifying an appropriate setting for this analysis is which formulation of such dependencies will indeed result in a replicator or replicator-mutator dynamics. Regarding the applicability of this approach to learning in artificial neural networks, it will be of particular interest to exploit the properties of potential games, since it is to be expected that there is only very limited antagonism between neurons in neural networks. This will provide a contrast to the purely antagonistic settings we have considered in the present thesis as a kind of worst-case setting.

Bibliography

[1] Allen, B. and Nowak, M. A. “Games on Graphs”. In: *EMS Surveys in Mathematical Sciences* 1.1 (2014), pp. 113–151.

[2] Allen, B. and Rosenbloom, D. I. S. “Mutation Rate Evolution in Replicator Dynamics”. In: *Bulletin of Mathematical Biology* 74 (2012), pp. 2650–2675.

[3] Argasinski, K. and Broom, M. “Towards a Replicator Dynamics Model of Age Structured Populations”. In: *Journal of Mathematical Biology* 82.5 (2021), p. 44.

[4] Argasinski, K. “The Dynamics of Sex Ratio Evolution: Dynamics of Global Population Parameters”. In: *Journal of Theoretical Biology* 309 (2012), pp. 134–146.

[5] Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. “Understanding Deep Neural Networks with Rectified Linear Units”. In: *arXiv e-prints* (2016), arXiv: 1611.01491.

[6] Axelrod, R. and Hamilton, W. D. “The Evolution of Cooperation”. In: *Science* 211.4489 (1981), pp. 1390–1396.

[7] Bauer, J., Alonso, E., and Broom, M. “Mutation Stabilises Neutrally Stable Equilibria in Antagonistic Co-Evolution and Almost All 2x2 Matrix Games”. In Preparation.

[8] Bauer, J., Broom, M., and Alonso, E. “The Stabilization of Equilibria in Evolutionary Game Dynamics through Mutation: Mutation Limits in Evolutionary Games”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2231 (2019), p. 20190355.

[9] Bauer, J., West, S., Alonso, E., and Broom, M. “Mutation-Bias Learning: Mutation Stabilizes Simple Reinforcement Learning in Zero-Sum Games”. In Preparation.

[10] Bergstrom, C. T. and Lachmann, M. “The Red King Effect: When the Slowest Runner Wins the Coevolutionary Race”. In: *Proceedings of the National Academy of Sciences* 100.2 (2003), pp. 593–598.

[11] Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. “Evolutionary Dynamics of Multi-Agent Learning: A Survey”. In: *Journal of Artificial Intelligence Research* 53.1 (2015), pp. 659–697.

[12] Bomze, I. M. and Bürger, R. “Stability by Mutation in Evolutionary Games”. In: *Games and Economic Behavior* 11.2 (1995), pp. 146–172.

[13] Börgers, T. and Sarin, R. “Learning Through Reinforcement and Replicator Dynamics”. In: *Journal of Economic Theory* 77.1 (1997), pp. 1–14.

[14] Bowling, M. and Veloso, M. “Multiagent Learning Using a Variable Learning Rate”. In: *Artificial Intelligence* 136.2 (2002), pp. 215–250.

[15] Boylan, R. T. “Evolutionary Equilibria Resistant to Mutation”. In: *Games and Economic Behavior* 7.1 (1994), pp. 10–34.

[16] Broom, M. and Rychtář, J. “A General Framework for Analysing Multiplayer Games in Networks Using Territorial Interactions as a Case Study”. In: *Journal of Theoretical Biology* 302 (2012), pp. 70–80.

[17] Broom, M. and Rychtář, J. *Game-Theoretical Models in Biology*. Boca Raton, FL: CRC Press, 2013.

[18] Bürger, R. “Mutation-Selection Models in Population Genetics and Evolutionary Game Theory”. In: *Acta Applicandae Mathematicae* 14.1-2 (1989), pp. 75–89.

[19] Bürger, R. “Mathematical Properties of Mutation-Selection Models”. In: *Genetica* 102 (1998), pp. 279–298.

[20] Buşoniu, L., Babuška, R., and De Schutter, B. “A Comprehensive Survey of Multiagent Reinforcement Learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2 (2008), pp. 156–172.

[21] Chang, J. C. and Kan, Y. W. “Beta 0 Thalassemia, a Nonsense Mutation in Man”. In: *Proceedings of the National Academy of Sciences* 76.6 (1979), pp. 2886–2889.

[22] Chapman, A. C., Leslie, D. S., Rogers, A., and Jennings, N. R. “Convergent Learning Algorithms for Unknown Reward Games”. In: *SIAM Journal on Control and Optimization* 51.4 (2013), pp. 3154–3180.

[23] Collins, D. W. and Jukes, T. H. “Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence”. In: *Genomics* 20.3 (1994), pp. 386–396.

[24] Collins, E. J. and Leslie, D. S. “Convergent Multiple-Timescales Reinforcement Learning Algorithms in Normal Form Games”. In: *The Annals of Applied Probability* 13.4 (2003), pp. 1231–1251.

[25] Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L., and Wallace, R. B. “Detection of Sickle Cell Beta S-Globin Allele by Hybridization with Synthetic Oligonucleotides.” In: *Proceedings of the National Academy of Sciences* 80.1 (1983), pp. 278–282.

[26] Cressman, R. *The Stability Concept of Evolutionary Game Theory*. Heidelberg: Springer, 1992.

[27] Cross, J. G. “A Stochastic Learning Model of Economic Behavior”. In: *The Quarterly Journal of Economics* 87.2 (1973), pp. 239–266.

[28] Crow, J. F. “The High Spontaneous Mutation Rate: Is It a Health Risk?” In: *Proceedings of the National Academy of Sciences* 94.16 (1997), pp. 8380–8386.

[29] Davies, H. et al. “Mutations of the BRAF Gene in Human Cancer”. In: *Nature* 417.6892 (2002), pp. 949–954.

[30] Dvoretzky, A. “On Stochastic Approximation”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley: University of California Press, 1956, pp. 39–55.

[31] Erovenko, I. V., Bauer, J., Broom, M., Pattni, K., and Rychtář, J. “The Effect of Network Topology on Optimal Exploration Strategies and the Evolution of Cooperation in a Mobile Population”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2230 (2019), p. 20190399.

[32] Fairbank, M. and Alonso, E. “The Divergence of Reinforcement Learning Algorithms with Value-Iteration and Function Approximation”. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. 2012, pp. 1–8.

[33] Fudenberg, D. and Levine, D. “Limit Games and Limit Equilibria”. In: *Journal of Economic Theory* 38.2 (1986), pp. 261–279.

[34] Fudenberg, D. and Tirole, J. *Game Theory*. Cambridge, Mass: MIT Press, 1991.

[35] Glicksberg, I. L. “A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points”. In: *Proceedings of the American Mathematical Society* 3.1 (1952), p. 170.

[36] Hamosh, A., King, T. M., Rosenstein, B. J., Corey, M., Levison, H., Durie, P., Tsui, L. C., McIntosh, I., Keston, M., and Brock, D. J. “Cystic Fibrosis Patients Bearing Both the Common Missense Mutation Gly—-Asp at Codon 551 and the Delta F508 Mutation Are Clinically Indistinguishable from Delta F508 Homozygotes, except for Decreased Risk of Meconium Ileus”. In: *American Journal of Human Genetics* 51.2 (1992), pp. 245–250.

[37] Harris, G. and Martin, C. “Shorter Notes: The Roots of a Polynomial Vary Continuously as a Function of the Coefficients”. In: *Proceedings of the American Mathematical Society* 100.2 (1987), pp. 390–392.

[38] Harsanyi, J. C. “Oddness of the Number of Equilibrium Points: A New Proof”. In: *International Journal of Game Theory* 2.1 (1973), pp. 235–250.

[39] Hart, S. and Mas-Colell, A. “Uncoupled Dynamics Do Not Lead to Nash Equilibrium”. In: *American Economic Review* 93.5 (2003), pp. 1830–1836.

[40] Hashimoto, T. and Nishibe, M. “Theoretical Model of Institutional Ecosystems and Its Economic Implications”. In: *Evolutionary and Institutional Economics Review* 14.1 (2017), pp. 1–27.

[41] Hauert, C. and Schuster, H. G. “Effects of Increasing the Number of Players and Memory Size in the Iterated Prisoner’s Dilemma: A Numerical Approach”. In: *Proceedings of the Royal Society B: Biological Sciences* 264.1381 (1997), pp. 513–519.

[42] Hilbe, C., Nowak, M. A., and Sigmund, K. “Evolution of Extortion in Iterated Prisoner’s Dilemma Games”. In: *Proceedings of the National Academy of Sciences* 110.17 (2013), pp. 6913–6918.

[43] Hirsch, M. W. and Smale, S. *Differential Equations, Dynamical Systems, and Linear Algebra*. New York: Academic Press, 1974.

[44] Hofbauer, J. “The Selection Mutation Equation”. In: *Journal of Mathematical Biology* 23.1 (1985), pp. 41–53.

[45] Hofbauer, J. and Sigmund, K. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press, 1998.

[46] Hofbauer, J. and Sigmund, K. “Evolutionary Game Dynamics”. In: *Bulletin of the American Mathematical Society* 40.4 (2003), pp. 479–519.

[47] Hoffman, M., Suetens, S., Gneezy, U., and Nowak, M. A. “An Experimental Investigation of Evolutionary Dynamics in the Rock-Paper-Scissors Game”. In: *Scientific Reports* 5.1 (2015).

[48] Imhof, L. A., Fudenberg, D., and Nowak, M. A. “Evolutionary Cycles of Cooperation and Defection”. In: *Proceedings of the National Academy of Sciences* 102.31 (2005), pp. 10797–10800.

[49] Izquierdo, S. S. and Izquierdo, L. R. “Strictly Dominated Strategies in the Replicator-Mutator Dynamics”. In: *Games* 2.3 (2011), pp. 355–364.

[50] Jaakkola, T., Jordan, M. I., and Singh, S. P. “On the Convergence of Stochastic Iterative Dynamic Programming Algorithms”. In: *Neural Computation* 6 (1994), pp. 1185–1201.

[51] Jordan, J. “Three Problems in Learning Mixed-Strategy Nash Equilibria”. In: *Games and Economic Behavior* 5.3 (1993), pp. 368–386.

[52] Kaisers, M. and Tuyls, K. “Frequency Adjusted Multi-Agent Q-Learning”. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*. AAMAS ’10. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 309–316.

[53] Kianercy, A. and Galstyan, A. “Dynamics of Boltzmann Q Learning in Two-Player Two-Action Games”. In: *Physical Review E* 85.4 (2012), p. 041145.

[54] Kingman, J. F. C. “A Simple Model for the Balance between Selection and Mutation”. In: *Journal of Applied Probability* 15.1 (1978), pp. 1–12.

[55] Kinoshita, S. “On Essential Components of the Set of Fixed Points”. In: *Osaka Mathematical Journal* 4.1 (1952), pp. 19–22.

[56] Krantz, S. G. and Parks, H. R. *The Implicit Function Theorem: History, Theory, and Applications*. New York: Springer, 2013.

[57] LeCun, Y., Bengio, Y., and Hinton, G. “Deep Learning”. In: *Nature* 521.7553 (2015), pp. 436–444.

[58] Leslie, D. S. “Reinforcement Learning in Games”. PhD thesis. University of Bristol, 2004.

[59] Littman, M. L. “Markov Games as a Framework for Multi-Agent Reinforcement Learning”. In: *Machine Learning Proceedings 1994*. Ed. by W. W. Cohen and H. Hirsh. San Francisco: Morgan Kaufmann, 1994, pp. 157–163.

[60] Luzzatto, L. “Sickle Cell Anaemia and Malaria”. In: *Mediterranean Journal of Hematology and Infectious Diseases* 4.1 (2012), e2012065.

[61] Marotta, C. A., Wilson, J. T., Forget, B. G., and Weissman, S. M. “Human Beta-Globin Messenger RNA. III. Nucleotide Sequences Derived from Complementary DNA”. In: *Journal of Biological Chemistry* 252.14 (1977), pp. 5040–5053.

[62] Marsili, M., Challet, D., and Zecchina, R. “Exact Solution of a Modified El Farol’s Bar Problem: Efficiency and the Role of Market Impact”. In: *Physica A: Statistical Mechanics and its Applications* 280.3-4 (2000), pp. 522–553.

[63] Maynard Smith, J. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

[64] Maynard Smith, J. and Price, G. R. “The Logic of Animal Conflict”. In: *Nature* 246.5427 (1973), pp. 15–18.

[65] McLennan, A. *Advanced Fixed Point Theory for Economics*. Singapore: Springer, 2018.

[66] Mertikopoulos, P. and Sandholm, W. H. “Learning in Games via Reinforcement and Regularization”. In: *Mathematics of Operations Research* 41.4 (2016), pp. 1297–1324.

[67] Minde, D. P., Anvarian, Z., Rüdiger, S. G., and Maurice, M. M. “Messing up Disorder: How Do Missense Mutations in the Tumor Suppressor Protein APC Lead to Cancer?” In: *Molecular Cancer* 10.1 (2011), p. 101.

[68] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. “Playing Atari with Deep Reinforcement Learning”. In: *arXiv e-prints* (2013), arXiv:1312.5602. arXiv: 1312.5602.

[69] Nash, J. “Non-Cooperative Games”. In: *Annals of Mathematics* 54.2 (1951), pp. 286–295.

[70] Nee, S. “Antagonistic Co-Evolution and the Evolution of Genotypic Randomization”. In: *Journal of Theoretical Biology* 140.4 (1989), pp. 499–518.

[71] Neel, J. V. “Mutation and Disease in Man”. In: *Canadian Journal of Genetics and Cytology* 20.3 (1978), pp. 295–306.

[72] Norman, M. F. *Markov Processes and Learning Models*. Mathematics in Science and Engineering v. 84. New York: Academic Press, 1972.

[73] Nowak, M. A., Sasaki, A., Taylor, C., and Fudenberg, D. “Emergence of Cooperation and Evolutionary Stability in Finite Populations”. In: *Nature* 428.6983 (2004), pp. 646–650.

[74] Omidshafiei, S., Papadimitriou, C., Piliouras, G., Tuyls, K., Rowland, M., Lespiau, J.-B., Czarnecki, W. M., Lanctot, M., Perolat, J., and Munos, R. “α-Rank: Multi-Agent Evaluation by Evolution”. In: *Scientific Reports* 9.1 (2019).

[75] Osborne, M. J. and Rubinstein, A. *A Course in Game Theory*. Cambridge, Mass: MIT Press, 1994.

[76] Pacheco, J. M., Santos, F. C., Souza, M. O., and Skyrms, B. “Evolutionary Dynamics of Collective Action in N-Person Stag Hunt Dilemmas”. In: *Proceedings of the Royal Society B* 276.1655 (2009), pp. 315–321.

[77] Page, K. M. and Nowak, M. A. “Unifying Evolutionary Dynamics”. In: *Journal of Theoretical Biology* 219.1 (2002), pp. 93–98.

[78] Perko, L. *Differential Equations and Dynamical Systems*. Third. New York: Springer, 2001.

[79] Press, W. H. and Dyson, F. J. “Iterated Prisoner’s Dilemma Contains Strategies That Dominate Any Evolutionary Opponent”. In: *Proceedings of the National Academy of Sciences* 109.26 (2012), pp. 10409–10413.

[80] Puterman, M. L. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. Hoboken: John Wiley & Sons, 1994.

[81] Ritzberger, K. “The Theory of Normal Form Games from the Differentiable Viewpoint”. In: *International Journal of Game Theory* 23.3 (1994), pp. 207–236.

[82] Ritzberger, K. and Weibull, J. W. “Evolutionary Selection in Normal-Form Games”. In: *Econometrica* 63.6 (1995), pp. 1371–1399.

[83] Robbins, H. and Monro, S. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[84] Rowland, B. A., Maida, A. S., and Berkeley, I. S. N. “Synaptic Noise as a Means of Implementing Weight-Perturbation Learning”. In: *Connection Science* 18.1 (2006), pp. 69–79.

[85] Rummery, G. A. and Niranjan, M. *On-Line Q-Learning Using Connectionist Systems*. Technical Report CUED/F-INFENG/TR 166. Cambridge University Engineering Department, 1994.

[86] Rustichini, A. “Optimal Properties of Stimulus—Response Learning Models”. In: *Games and Economic Behavior* 29.1-2 (1999), pp. 244–273.

[87] Sandholm, W. H. *Population Games and Evolutionary Dynamics*. Economic Learning and Social Evolution. Cambridge, Mass.: MIT Press, 2010.

[88] Sardanyés, J. and Solé, R. V. “Matching Allele Dynamics and Coevolution in a Minimal Predator–Prey Replicator Model”. In: *Physics Letters A* 372.4 (2008), pp. 341–350.

[89] Sato, Y., Akiyama, E., and Farmer, J. D. “Chaos in Learning a Simple Two-Person Game”. In: *Proceedings of the National Academy of Sciences* 99.7 (2002), pp. 4748–4751.

[90] Sato, Y. and Crutchfield, J. P. “Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems”. In: *Physical Review E* 67.1 (2003).

[91] Sawyer, S. A., Parsch, J., Zhang, Z., and Hartl, D. L. “Prevalence of Positive Selection among Nearly Neutral Amino Acid Replacements in Drosophila”. In: *Proceedings of the National Academy of Sciences* 104.16 (2007), pp. 6504–6510.

[92] Shapley, L. S. “Stochastic Games”. In: *Proceedings of the National Academy of Sciences* 39.10 (1953), pp. 1095–1100.

[93] Sidore, C. et al. “Genome Sequencing Elucidates Sardinian Genetic Architecture and Augments Association Analyses for Lipid and Blood Inflammatory Markers”. In: *Nature Genetics* 47.11 (2015), pp. 1272–1281.

[94] Silver, D. et al. “Mastering the Game of Go without Human Knowledge”. In: *Nature* 550.7676 (2017), pp. 354–359.

[95] Sinervo, B. and Lively, C. M. “The Rock–Paper–Scissors Game and the Evolution of Alternative Male Strategies”. In: *Nature* 380.6571 (1996), pp. 240–243.

[96] Singh, S., Jaakkola, T., Littman, M. L., and Szepesvári, C. “Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms”. In: *Machine Learning* 38.3 (2000), pp. 287–308.

[97] Skyrms, B. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press, 2003.

[98] Smead, R. and Forber, P. “The Coevolution of Recognition and Social Behavior”. In: *Scientific Reports* 6.1 (2016).

[99] Song, Y., Gokhale, C. S., Papkou, A., Schulenburg, H., and Traulsen, A. “Host-Parasite Coevolution in Populations of Constant and Variable Size”. In: *BMC Evolutionary Biology* 15.1 (2015), p. 212.

[100] Sutton, R. S. “Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding”. In: *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press, 1996, pp. 1038–1044.

[101] Sutton, R. S. “Learning to Predict by the Methods of Temporal Differences”. In: *Machine Learning* 3.1 (1988), pp. 9–44.

[102] Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. Cambridge, Mass.: MIT Press, 1998.

[103] Taylor, C. and Nowak, M. A. “Evolutionary Game Dynamics with Non-Uniform Interaction Rates”. In: *Theoretical Population Biology* 69.3 (2006), pp. 243–252.

[104] Taylor, P. D. “Evolutionarily Stable Strategies with Two Types of Player”. In: *Journal of Applied Probability* 16.01 (1979), pp. 76–83.

[105] Taylor, P. D. and Jonker, L. B. “Evolutionary Stable Strategies and Game Dynamics”. In: *Mathematical Biosciences* 40.1 (1978), pp. 145–156.

[106] Tesauro, G. “TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play”. In: *Neural Computation* 6.2 (1994), pp. 215–219.

[107] Teschl, G. *Ordinary Differential Equations and Dynamical Systems*. Providence, RI: American Mathematical Society, 2012.

[108] Tsitsiklis, J. N. “Asynchronous Stochastic Approximation and Q-Learning”. In: *Machine Learning* 16.3 (1994), pp. 185–202.

[109] Tuyls, K., ’t Hoen, P. J., and Vanschoenwinkel, B. “An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games”. In: *Autonomous Agents and Multi-Agent Systems* 12.1 (2006), pp. 115–153.

[110] Tuyls, K. and Nowé, A. “Evolutionary Game Theory and Multi-Agent Reinforcement Learning”. In: *The Knowledge Engineering Review* 20.1 (2005), pp. 63–90.

[111] Tuyls, K., Verbeeck, K., and Lenaerts, T. “A Selection-Mutation Model for Q-Learning in Multi-Agent Systems”. In: *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. AAMAS ’03. New York: ACM, 2003, pp. 693–700.

[112] van Damme, E. *Stability and Perfection of Nash Equilibria*. Heidelberg: Springer, 1991.

[113] Veller, C. and Hayward, L. K. “Finite-Population Evolution with Rare Mutations in Asymmetric Games”. In: *Journal of Economic Theory* 162 (2016), pp. 93–113.

[114] von Neumann, J. “Zur Theorie Der Gesellschaftsspiele”. In: *Mathematische Annalen* 100.1 (1928), pp. 295–320.

[115] von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. 60th anniversary ed. Princeton: Princeton University Press, 2007.

[116] Watkins, C. J. “Learning from Delayed Rewards”. PhD Thesis. Cambridge: University of Cambridge, 1989.

[117] Watkins, C. J. and Dayan, P. “Q-Learning”. In: *Machine Learning* 8 (1992), pp. 279–292.

[118] Weibull, J. W. *Evolutionary Game Theory*. Cambridge, Mass.: MIT Press, 1995.

[119] Weiss, G., ed. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. Cambridge, Mass: MIT Press, 1999.