Learning, Imitation and Economic Rationality

by

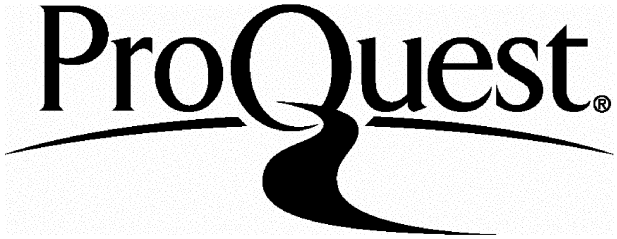
Antonio Jose Morales Siles

at the

University College of London

PhD Degree

1999



ABSTRACT

This thesis studies a population of agents facing repeatedly the same decision problem. Each agent knows the set of strategies available, but not the payoff distribution associated with each strategy. Agents follow simple behaviour rules which have no memory beyond what is encoded in the current "state" of the decision maker.

We consider two different frameworks: (i) Individual learning and (ii) imitation learning. We also distinguish rules where the "state space" of the decision maker is the set of pure strategies, and behavior rules where it is the set of mixed strategies. The results for these two cases differ dramatically.

In the case of individual learning, we say that a behaviour rule is maximising (approximately maximising) if asymptotically, for all underlying payoff distributions, the decision maker will play with probability one (close to one) the expected payoff maximising strategy. We show that no behaviour rule with pure strategy state space is (approximately) maximising. For the class of mixed strategy behaviour rules, we identify a property called monotonicity which implies approximate maximisation, provided learning proceeds in small steps. We characterise monotone learning rules, showing that they are closely related to the "replicator dynamics" of evolutionary game theory.

When considering imitation learning, we postulate that at each iteration agents have the opportunity of randomly sampling another agent, observing the strategy which this agent played and his payoff. We consider two different settings. In the first, the behaviour of the observed population is exogenously given and constant. We show that no pure strategy imitation rule is (approximately) maximising. For mixed strategy behaviour rules, we characterise the set of all monotone rules, showing that monotone rules involve imitation probabilities which are proportional to payoff differences. In the second setup all agents in the population are allowed to adjust their behaviour according to some imitation rule. We show that no pure strategy imitation rule is maximising i.e. there does not exist a rule such that, if every member of the population adopts it, asymptotically every agent plays the expected payoff maximising strategy with probability one, regardless of the true payoff distribution. We then define and analyse a weaker requirement, "equilibrium" imitation rules.

Contents

0.1 Acknowledgment 5

1 Motivation **6**

1.1 Introduction 6

1.2 On Evidence Regarding the Rationality Hypothesis in Economics 9

1.3 Learning and Rationality 15

1.4 Theoretical Literature on Learning 17

1.5 Learning and Bounded Rationality 21

2 Optimality of Learning and Imitation Rules: An Overview **23**

2.1 Introduction 23

2.2 Individual Learning 25

2.3 Individual Imitation 27

2.4 Imitation in Populations 29

3 Formal Framework **32**

4 Pure Strategy Rules for Individual Learning **34**

4.1 Introduction 34

4.2 Formal Framework 35

4.3 Definitions 35

4.4 Approximate Payoff Maximisation 37

4.5 Conclusion 41

5 Mixed Strategy Rules for Individual Learning **42**

5.1 Introduction 42

5.2 Formal Framework 44

5.3 Definitions 44
5.4 Monotone Behaviour Rules 46
5.5 Characterisation of Monotone Behaviour Rules 51
5.6 Proof of Proposition 7 55
5.7 Conclusion 59

6 Pure Strategy Imitation Rules **63**
6.1 Introduction 63
6.2 Formal Framework 64
6.3 Payoff Maximisation 65
6.4 Approximate Payoff Maximisation 67
6.5 Conclusion 72

7 Mixed Strategy Imitation Rules **74**
7.1 Introduction 74
7.2 Formal Framework 76
7.3 Definitions 76
7.4 Monotone Imitation Rules 78
7.5 Conclusion 85

8 Imitation and Equilibrium in Populations **87**
8.1 Introduction 87
8.2 Formal Framework 90
8.3 Definitions 91
8.4 Payoff Maximisation 92
8.5 Equilibrium Rules 95
8.6 The Binary Payoff 2x2 Decision Problem 99
8.7 Conclusion 109

9 Conclusions **110**

0.1 Acknowledgment

This thesis has been the most challenging intellectual adventure I have ever tried, and, like in all important adventures, there are some persons who has played an important role in its completion.

Ana Lozano, a former teacher of mine and currently a faculty mate in the Department of Economics at the University of Malaga, was responsible for having pushed me inside the academic career. Ana showed me the way towards the mountain.

After completing my Economics Degree at Malaga University, I spent two years in Madrid taking an Economics Postgraduate Course at Centro de Estudios Monetarios y Financieros. I met there Jorge Padilla. His lectures and above all my conversations with him, raised my interest in microeconomic and game theory. He suggested me the possibility of taking a PhD abroad. Jorge put me in the position to initiate the climbing of the mountain.

The mountain is placed in the Department of Economics of the University College of London. I arrived there in June 1994 to meet Tilman Börgers. After an English coffee in Cracks and a 3 hours meeting, he accepted to supervise my thesis. He introduced me into the field of learning theory, his own field of study, and was patient enough to forgive all my limitations. Despite of my persistent errors, he always provided me with more trials. In February 1997, I decided to go back to Malaga, where I became a teaching assistant at the Department of Economics. I really thank Tilman for letting me go. I believe Tilman has given to me much more than what an average supervisor usually offers. My most sincere gratitude goes to him. He has always been waiting for me at the top of the mountain.

Now the thesis is completed. After this five years, I am proud of saying that above all, the three mentioned persons have become very good friends of mine. I believe it to be my most important achievement.

There is still a fourth person who has been a necessary condition for me to reach the top of the mountain. Ana Moniche came with me to England in 1994 and I went with her to Malaga in February 1997. After these years, she has become more than a friend. I married her in April 1997. Ana, te quiero mucho.

Chapter 1

Motivation

1.1 Introduction

The assumption that economic agents' behaviour is rational is central to much of economic theory. In decisions under certainty, this assumption is usually interpreted to mean that agents maximize some preference ordering or some utility function. In decisions under uncertainty, the assumption means that agents maximize expected utility. Finally, in games rational behaviour is interpreted by economic theory as Nash equilibrium behaviour.

Recently, economic theorists have started to enquire into the foundations of the rationality hypothesis. In particular, it has been asked how agents might come to make rational choices. There are two reasons for this recent interest. One is that experimental data, and also some real world data, suggest that the rationality hypothesis has only a limited domain of validity, i.e. agents' choices are sometimes, but not always, rational. If one wants to understand when agents can be expected to be rational, and when rational behaviour should not be expected, one needs to have some understanding of the mechanisms which bring about rationality.

The second reason for the recent interest in the foundations of rationality is related to the fact that strategically interactive situations, i.e. games, often have many outcomes all of which are compatible with the rationality of all agents' decisions. Such outcomes are known as Nash equilibria. If games have many Nash equilibria, the question arises whether one can make statements about which equilibria are more likely to occur. This is known as the equilibrium refinement, or, if the goal is to single out

a unique equilibrium, as the equilibrium selection question. Progress on this question requires that one understands how agents find their way to an equilibrium.

The literature has considered two basic ways in which agents might find their way to equilibrium. One is through introspection, the other is through trial and error learning in real time. Both have some plausibility. If agents play important games in the real world, it can be expected that decision makers are willing to undertake a lot of strategic reasoning, and, in some cases, this might take them towards rational decision. In other cases, decisions might be relatively unimportant, but they have to be made very frequently. In such cases, initial decisions might well be flawed, but it can be hypothesized that later decisions will be rational or close to rational. In this thesis we shall only consider trial and error learning as a mechanism to steer agents towards rationality.

The hypothesis of trial and error learning has been formalized in a variety of ways. Different formalizations differ with respect to the information which agents are assumed to hold at the beginning of the learning process, with respect to the feedback information which agents are provided during the learning process, and with respect to the behaviour hypotheses concerning how agents respond to the information available to them, both in terms of their behaviour, and in terms of updating their beliefs.

It has become common to distinguish two broad classes of learning models. The first class are so-called belief-based learning models. In such models agents are aware of the basic structure of the situation which they are facing, i.e. they know their own and, possibly, others' strategy sets and they know their own payoff matrix. During the learning process agents observe other players' strategy choices, and, possibly, the state of nature. Agents form beliefs about their environment on this basis, and then maximize their expected payoff.

The second class of learning models are reinforcement learning models. Reinforcement learning models differ from belief-based learning models in all dimensions. Firstly, agents are assumed to hold less information at the outset. In particular, it is not assumed that agents know other players' strategy sets, or their own payoff matrix at the outset of the game. Secondly, during the learning process the only feedback which agents are assumed to receive is their own payoff. They don't necessarily observe other players' strategy choices, or the state of nature. Finally, agents' choices are assumed

to be instinctive responses to payoff experiences rather than the result of explicit maximisation of expected payoffs.

In the current thesis we shall only be concerned with models of reinforcement learning, not with belief-based learning processes, in single person decision problems, rather than in games. The reason for this is rather pragmatic. We restrict the scope of our investigation so as to make the investigation tractable and feasible.

We interpret the notion of reinforcement learning broadly. In particular, we allow for the possibility that the agent observes not only his/her own strategy choice and payoff, but also the strategy and payoff of some other agents who finds him/herself in the same situation. Note, however, that this other agent is not the agent with whom the game has been played. We thus include among reinforcement learning simple models of imitation. The reason for considering single agent learning models and imitation models together is that they share many features.

Much of previous work in the area of reinforcement learning has considered specific learning algorithms, and has investigated the predictions which can be derived from a given learning algorithm in single person decision problems and games. In this thesis, we take a different approach. We impose no particular functional forms, but allow for a large variety of learning algorithms. We then ask which algorithms in this class have the property that agents will learn rational choices in a variety of environments. We provide characterizations of learning algorithms with this property.

Our characterizations focus on two properties of a learning algorithm. The first such property is what we shall call the "state space" of the learning algorithm. The second is the functional form. The "state space" of a learning algorithm is simply the set of states in which the decision maker finds him/herself at any particular point in time. Our results will indicate that this set has to be, in a sense to be made precise below, "sufficiently big". Otherwise, the learning algorithm will not have sufficient memory to be able to deal with many different situations. As far as functional forms are concerned, we shall show that in our setting only learning algorithms which are linear in payoffs will achieve optimality in the long run. An intuitive reason for this is that expected payoffs themselves are linear functions of payoffs.

In a sense, this thesis presents an axiomatic approach to learning algorithms. How-

ever, our work should not be interpreted as aiming to single out "desirable" learning algorithms. Whether agents use learning algorithms which lead to rationality in a large variety of situations is an empirical question. Our results help to interpret empirical results regarding this question.

The remainder of this Chapter is organized as follows. In the next Section we elaborate in some more detail why economists have become interested in the foundations of the rationality hypothesis. In the third Section of the chapter, we briefly review some experimental findings on learning in decisions and games. Finally, in the fourth and last Section we review some of the theoretical literature on learning. A more detailed overview and discussion of our findings in this thesis is postponed to Chapter 2. The detailed discussion of the relations between our research and other authors' work is postponed until the main formal part of the thesis which begins with Chapter 3.

1.2 On Evidence Regarding the Rationality Hypothesis in Economics

One reason why economists have started to enquire into the foundations of the rationality hypothesis is that in some situations there is a mismatch between predictions based on the hypothesis and experimental and field data. The simplest such cases arise not in games, but in single person decision problems. In this section, we provide a partial review of some of the relevant evidence, first for single person decision problems, and then for games. The purpose of this Section is to support our earlier claim that the experimental evidence indicates the validity of the rationality hypothesis in some, but not in all circumstances. We therefore only quote a few selected experiments, and don't aim for completeness.

In decision problems under risk, experiments during the last 40 years have uncovered a number of *anomalies* all of which are deviations from the predictions of the theory of expected utility maximisation. In most of these studies, subjects are asked, possibly repeatedly, to choose one lottery from a pair of such lotteries. The idea of using pairs of lottery choices to elicit subject's preferences goes back to Maurice Allais [1].

Maurice Allais also found the most famous paradox in this area. Assume that there are three monetary prices: 2.5, 0.5 and 0 millions Euros. The subjects are confronted

with the following lottery choices. First, they have to choose either lottery $L_1 = (0, 1, 0)$ or lottery $L_1' = (.10, .89, .01)$ where the numbers indicate the probability of the three prizes in the order in which they were listed above. Next, subjects have to choose either lottery $L_2 = (0, .11, .89)$ or lottery $L_2' = (.10, 0, .90)$.

It is commonly observed that individuals strictly prefer L_1 over L_1', but that they also strictly prefer L_2' over L_2. But these choices are incompatible with expected utility theory. The choice of L_1 over L_1' implies that L_2 must be preferred over L_2' or the axioms of expected utility theory are violated.[1]

The *dissatisfaction* with the empirical accuracy of expected utility theory led to the formulation of a number of new theories of decision making under risk. Most of these theories have a larger number of free parameters than expected utility theory does, and therefore it is not surprising that they "explain" observed behaviour better than expected utility theory. How should one evaluate the status of expected utility theory in comparison to the new theories of decision making under risk? Two different responses are represented by recent articles by Hey and Orme [28] and Harless and Camerer [24].

Hey and Orme generate their own choice data, and then investigate whether expected utility, or alternative theories, are better explanations of the experimental data, once it is taken into account that the alternative theories have larger numbers of free parameters. They find that for 39% of their subjects, expected utility theory fits no worse than the other contenders. For the other 61% of the subjects, they conclude that the superiority of the alternative theories is not established. Therefore they argue that "... our study indicates that behaviour can be reasonably well modelled (...) as expected utility plus noise. Perhaps we should now spend some time on thinking about the noise, rather than about even more alternatives...".([28], p.1322).

A different viewpoint is argued by Harless and Camerer [24]. They develop a statistical test which can be use to aggregate results across studies. This aggregation is used to test the predictive utility (fit and parsimony) of the various theories of decision making. Furthermore, it can be used to test whether deviations from the expected utility theory are robust across studies. Although they find that there is no a single

[1] Denote by u_{25}, u_{05} and u_0 the utility values of the three monetary outcomes. Then the choice $L_1 \succ L_1'$ implies $u_{05} > .10u_{25} + .89u_{05} + .1u_0$. Adding $.89(u_0 - u_{05})$ to both sides we get $.11u_{05} + .89u_0 > .10u_{25} + .9u_0$ and therefore $L_2 \succ L_2'$.

winner among theories (everyone involves a trade-off between parsimony and fit), they argue that "... violations of expected utility theory are robust enough that modelling of aggregate economic behaviour based on alternatives to expected utility is well worth exploring."([24], p. 1286). Despite this negative view with respect to the performance of the expected utility theory in comparison to other theories of decision making, they recognize that their study dramatically confirms a conclusion previously observed in other experiments, i.e. that the expected utility theory predicts *well* when subjects choose between gambles which involve outcomes with positive probability, and predicts *poorly* when some outcomes have probability zero (see also Conslik [12] and Camerer [10]).

It is this second viewpoint which reinforces our argument in this thesis, that expected utility theory has not got an universal domain of validity, i.e. there are situations in which predictions based on that theory does not confirm with observed behaviour.

Within the realm of strategically interactive situations, rational behaviour is interpreted as Nash equilibrium behaviour. The question whether observed behaviour in games can always be rationalized has again a negative answer.

As an example, we shall consider experimentally observed behavior in auctions. More specifically, we shall focus on auctions with independent private values. In the auction literature, the independent private value model corresponds to the case in which the valuation of each bidder is privately known and bidders' valuations are drawn independently from each other.

We shall consider behaviour in two different auction formats: the *Vickrey* auction and the *English* auction. In the latter, the price is increased until one bidder remains. This bidder gets the object and pays his bid. The *Vickrey* auction is a second-price sealed bid auction. The highest bidder gets the object and pays the second-highest bid.

In the independent private value setup, bidders have a dominant strategy in both auctions. In the English auction, the dominant strategy is to bid up to one's true value. In the Vickrey auction, the dominant strategy is to bid one's true value. Moreover, in the independent private value setup, the two auctions are strategically equivalent, provided that bidders in the English auction follow threshold strategies of the form: stay in the auction until the bid reaches some particular boundary. If bidders in the English auction only consider strategies of this form, there is an isomorphism between

the strategy sets in the two auctions, and the payoff functions are identical with respect to this isomorphism.

The main conclusion from experiments with these auctions is that subjects do not behave in strategically equivalent ways. Kagel et al. [30] reports bidding *above* the dominant strategy price in second-price auctions, while in the English auction bidding follows the dominant strategy. This result has been replicated in numerous experiments so as to become an accepted fact.[2]

The second reason for the recent interest in the foundations of rationality is related to the fact that games often have many Nash equilibria. If games have many Nash equilibria, the question arises whether one can make statements about which equilibria are more likely to occur. This is known as the *equilibrium refinement problem*, or, if the goal is to single out exactly one equilibrium, as the *equilibrium selection problem*.

Subgame perfection is one of the most widely used refinements. However, predictions based on this concept often fail to capture observed behaviour. As an example, we shall consider experiments with sequential bargaining games. In the simplest sequential bargaining game, the *ultimatum game*, a pie has to be shared by two players. Player 1, the proposer, has the first move. Player 2 receives the offer and decide whether to accept it, in which case each player receives the proposed share, or to reject it, in which case each player receives nothing. The subgame perfect equilibrium prediction is that player 1 should ask for the whole pie and that player 2 should accept any proposal.

Experiments regarding this game have initially yielded contradictory results. Work by Ochs and Roth [43] settled the dispute by using a larger experimental design. Their results suggest that the subgame perfect equilibrium prediction fails as a predictor of observed behaviour: the mean offer was positive and a substantial proportion of positive offers were rejected.

On the other hand, Ochs and Roth's experiments indicate that the observed mean offers deviate from the perfect equilibrium prediction in a particular direction, the direction of equal division. This fact has since then been replicated so as to become an accepted fact: in the ultimatum game, there is a high concentration of equals divi-

[2]The breakdown of the strategic equivalence of the Vickrey and English auctions can be considered as an analogous to the *preference reversal* phenomenon (equivalent ways of eliciting preferences yield different revealed preferences). This phenomenon is described by psychologists as one of the most robust violations of EU in decision problems.

sion offers. This suggests that there exists an underlying behaviour mechanism which prompts the equal division offer in the ultimatum game and that subgame perfection is not able to capture it.

However, there are games which do not seem to be too different from the ultimatum game for which the subgame perfection prediction does match observed behaviour. An example is the *best-shot game.* In this game player 1 selects the quantity q_1 of a public good that he is willing to supply. Player 2 observes q_1 and selects another quantity q_2. The quantity of public good supplied is the maximum of q_1 and q_2. Payoffs are proportional to the quantity of the public good that is supplied minus costs of the quantity q_i.

Experimenters typically fix payoffs so that the unique subgame-perfect equilibrium involves the choice of $q_1 = 0$ for player 1, so that player 1 free rides on player 2. This equilibrium is similar to the subgame-perfect equilibrium of the ultimatum game in that the payoff distribution in equilibrium is very extreme, and in that the player who moves first exploits in some sense the player who moves second. Nonetheless, experiments have shown that in this case the subgame perfect prediction captures quite well the observed behaviour (see Prasnikar and Roth [45]).[3]

So far we have reviewed some selected work which evidences the inability of the rationality hypothesis to explain observed behaviour in experiments. We shall now review some real data analysis which also suggest that in real life situations people do not always behave as the rationality hypothesis predicts. The first concerns the decision on how to allocate today's money between consumption and saving, i.e. how much to consume today and how much to save it to finance future consumption. The theoretical foundations on intertemporal choice theory go back to the fifties, when the Permanent Income Theory [20] and the Life-Cycle Theory [37] were stated. These theories state that consumption is determined by the value of life-time resources, typically involving current financial and human wealth.

We shall focus on the Permanent Income Theory and specifically on one prediction whose consistency can be tested on real data, i.e. changes in consumption are related to unpredictable changes in income. This implies that changes in consumption are un-

[3] Other games for which the predictions based on subgame perfection match observed behaviour are the *market game* and the *impunity game.*

predictable, this statement being know as the Permanent Income Hypothesis. However, the seminal works by Flavin [21] using aggregate data and by Hall and Masking [23] using microeconomic data showed the failure of this hypothesis, i.e. consumption also responds to *predictable* changes in income. This excess sensitivity of consumption has since then being replicated in other studies so as to become an accepted fact.

The second real life situation we shall consider concerns field studies in common value auctions. In common value auctions, the value of the auctioned item is the same to all bidders. Although the bidders do not know the value of the item at the time they are bidding, they receive signal values. Note that although all bidders obtain unbiased estimates of the value of the item, they win in cases where they have the highest signal value, yielding below normal or even negative profits. The systematic failure to account for this adverse selection problem is known as the winner's curse. Note that this systematic failure violates the notion of economic rationality.

An example of real world common value auction is the oil lease auction. A number of field studies in these auctions have focused on the rate of returns for these leases. The seminal paper is [11] which claims that oil leases won by competitive bidding yield unexpectedly low rates of return, even less than the market rate of return on their investments, interpreting these results on a winner's curse basis. Evidence of the same phenomena in other kinds of auctions has since then been collected to support that the winner's curse is responsible for these low returns to winners. However the debate continues as these low rates of return might have alternative interpretations. The debate has now moved to laboratory experiments, where the winner's curse has been shown to be present.

In summary, we wish to argue that experimental economics and the analysis of field data have shown that there are situations in which observed behaviour is easy to rationalize, and that there are situations in which the observed behaviour is almost impossible to rationalize. If one wants to understand when rational behaviour should be expected and when agents cannot be expected to behave rationally, then it is necessary to have some understanding of the mechanisms which bring about rationality.

Experimental evidence also suggests that in games with multiple equilibria there some equilibria that are more likely to occur than others. However, existing refinements of Nash equilibrium do not seem to capture well what determines which equilibria will

occur. Again, a better understanding of this issue seems to require more insight into how equilibria come about.

1.3 Learning and Rationality

The literature has considered two basic mechanisms that bring about rationality: (i) rationality can emerge after a careful reasoning by the economic agent, and (ii) rationality can arise from trial and error learning, provided that the situation is encountered sufficiently often.

Rational reasoning can be expected to be a sensible mechanism if the agent is engaged in a situation in which actions can yield important consequences. In these circumstances, it might be expected that the agent devotes a great deal of reasoning to fully understand the strategic situation in which he is involved. In some cases, it might be that this strategic reasoning leads the agent to behave rationally.

But it can also be the case that the agent is facing a situation which is relatively unimportant but which calls for a decision very frequently. In this case, it may be that initial decisions are not rational but it can be hypothesized that through a trial and error process the agent can ultimately behave rationally or close to rational. Although both mechanisms have some plausibility, this thesis will be only concerned with the latter one.

The plausibility of the learning mechanism as generator of rationality can be assessed by turning our attention to experimental work and seeing how well "experienced" agents behave when the experiment involves repetitions of the same situation. In this section, we provide a partial review of some experimental evidence involving experienced agents in different setups. The purpose of this section is to support the claim that there are situations in which experienced agents find their way to rationality, but that there are also situations in which even experienced agents systematically deviate from rational behaviour. Furthermore, we wish to argue that the way in which experience affects agents' behaviour is different in different strategic environments.

The theory of learning has a long tradition in psychology. It was developed in the psychological literature as a way to explain both animal and human behaviour in simple decision problems and games. One of its main finding is an "irrational" behaviour so-

called *probability matching* behaviour (Estes and Straughan [19]).

This term refers to a decision maker facing a two-action decision problem, where each action yields binary payoffs. Suppose that strategy s yields one monetary unit with probability μ, and yields nothing with probability $1-\mu$. Suppose that strategy s' yields one monetary unit with probability $1-\mu$ and yields nothing with probability μ. It is said that the decision maker's behavior exhibits "probability matching" if the long run frequency with which strategy s is chosen is μ, and the long run frequency of strategy s' is $1-\mu$. This behaviour is "irrational" as long as $\mu \neq \frac{1}{2}$, because rationality would imply playing the unique highest expected payoff strategy with probability one.

In experiments involving repetitions of strategic situations, it is not always the case that agents find their way to rational behaviour as agents gain experience. Kagel et al. [30] reported failures of strategic equivalence in second-price and English auctions with private values. For the second-price auction, bids were above the dominant strategy. These results have since then been replicated in independent private values both with experienced and unexperienced agents (Harstad [25], Kagel and Levin [31]), so that these results are now widely accepted not to be attributed to bidders' inexperience.

In bargaining experiments, Binmore, Morgan, Shaked and Sutton [4] conducted a series of experiments with a version of the ultimatum game with optional breakdown. They report that in the case of experienced agents, the perfect equilibrium prediction fits worse than in the case of unexperienced agents. This can be interpreted as a case of "*unlearning*".

So far we have presented evidence on the failure of experienced agents to achieve rationality. We shall now review some experimental evidence to support the claim that experienced agents achieve rationality in some situations. The first example comes again from the literature of experimental bargaining games, more specifically from experiments using versions of the ultimatum game. Experimental work by Harrison and McCabe ([26] and [27]) shows that experience promotes convergence to the perfect equilibrium prediction.

The last set of examples refer to experiments involving repetitions of a strategic situation with more than one Nash equilibria. In this setting it is possible to study whether experience leads players to eventually coordinate on some equilibrium, and if equilibrium is observed, which are the properties of the selected equilibrium.

Van Huyck, Battalio and Beil [56] report experiments with *pure* coordination 15-player games. All these games have a number of Pareto ranked Nash Equilibria. By allowing a given group of subjects to play the same game repeatedly, they were able to observe whether individuals immediately coordinate or whether they eventually coordinate on one of them. Their results show that in all the sequences, there was no equilibria in the first period, but there was convergence towards an equilibrium as players gained experience.

Cooper, DeJong, Forsythe and Ross [13] report experiments considering two-person 3x3 symmetric games with two strict Pareto-ranked Nash equilibria, where the third strategy was strictly dominated. Although at early stages no equilibrium was reached, after players have accumulated experience, the play quickly converges to one of the pure strategy equilibria of the game.

In summary, we wish to argue that experimental economics have shown that there are situations in which experienced agents achieve rationality and that there are situations in which they do not. In order to understand when rational behaviour should be expected it is necessary to have some understanding of the mechanisms which bring about rationality.

1.4 Theoretical Literature on Learning

The hypothesis of trial and error learning has been formalized in different ways. Different formalizations differ with respect to the information the agents hold at the beginning of the learning process, with respect to the information they receive during the learning process and with respect to how the information available affects the agents' behaviour.

It has become common to distinguish two broad classes of learning models: belief-based learning models and reinforcement learning models. In belief-based learning models, agents are aware of the basic structure of the situation in which they are interacting, i.e. they know their own set of strategies and possibly their opponents' sets of strategies. They also know their own payoff matrix and possibly their opponents' payoff matrix. They form beliefs about this environment. During the learning process, they receive information about their opponents' strategy choices and possibly the realized state of nature. With the information available, they update their beliefs and choose

their strategies to maximize expected payoffs.

Reinforcement learning models differ from belief-based models in several ways. Firstly, it is not assumed that agents know their opponents' strategy sets or their own payoff matrix. Secondly, during the learning process agents only receive information about their own realized payoffs. They do not necessarily observed their opponents' choices, or the state of nature. Finally, agents' choices are assumed to be instinctive responses to payoff experiences rather than the result of any explicit maximization process.

To illustrate the distinction between these two different types of learning models, we shall give first some examples of belief-based learning models, and then some examples of reinforcement learning models. The *Cournot* adjustment model is an early example of a learning model of belief-based learning models. In this model, n firms repeatedly play a stage-game in discrete time. It is assumed that at the end of every stage, each firm observes the strategies used by its opponents in that stage. In the next period, each firm then chooses its strategy to be a best-response to the previous period's strategy profile.

Another example of a belief-based learning model which is widely used in game theory is the *fictitious play* model. In this model, players choose their actions in each period to maximize that period's expected payoff, where players' subjective beliefs about their opponents' behaviour in the next period equals the empirical distribution of actions which has been observed in the past periods.

A first example of reinforcement learning model is the *linear stochastic learning model* [8]. At every stage a subject is permitted to take n responses. The subject is described by a probability distribution over responses which indicates how likely she is to take any of the responses. After choosing a response, the subject receives a stimulus, which can be either a reward or a punishment. When rewarded, the probability of choosing the same response is increased in a linear fashion. If the response is not rewarded, then the new probability is decreased in a linear way.

The second example is the Cross' model [15]. It is a generalization of the previous one by allowing a variety of stimuli. A subject is facing a decision problem under risk. Suppose payoffs are normalized between zero and one. Then Cross' rule states that in each iteration the updated probability distribution is simply a weighted average of

the previous probability distribution and the unit vector placing all probability on the action just taken. The weight of the unit vector equals the payoff. Note that this rule has the feature that all payoffs have a reinforcing, and hence positive, effect on the probability of the action which the decision maker chose.

Regardless of the particular learning model considered, the trial and error learning theory considers dynamic systems for describing how the behaviour of the agents evolves over time. It usually assumes that there is a population of agents playing the same game repeatedly and makes assumptions about the size of the population, the matching procedure and the information gathered at the end of each round for each agent. Each agent is endowed with a learning model to adapt her behaviour. Different learning models give rise to different dynamic systems.

Given that the evolution of the behaviour of the population is described by a dynamic system, the central issue in the theory of learning is the convergence of the dynamic system. It can be the case that the dynamic system converges to some concept of equilibrium of the stage game, then it is said that the learning model succeeds in learning that equilibrium concept, i.e. players learn to play rationally. But it is not always the case that the dynamic system converges to some equilibrium of the stage game.

In the fictitious play model, for example, the key concept for studying the convergence of fictitious play is the empirical distribution of the opponents' play. If the empirical distribution converges, then the system converges to a Nash equilibrium. For some classes of games, fictitious play is known to converge to a Nash equilibrium. These include two player zero sum games, two player games in which each player has only two strategies, and potential games. However, convergence is not always ensured. The first example of a game for which empirical distributions do not converge is a game with two players in which each player has three strategies. The example is due to Shapley [52], who showed that fictitious play may lead to cycles in this example. Another example is due to Jordan [29]. Jordan showed that fictitious play leads to a cycle in a simple three-player matching pennies.

Thus, theoretical analysis of fictitious play has lead to predictions concerning the question in which games we should expect players to learn Nash equilibria, and which games this cannot be expected. Analysis of belief-based learning processes has, however,

also helped to make predictions about equilibrium refinement/selection. The most prominent examples of such an analysis are Young [57] and Kandori, Mailath and Rob [32]. Young [57] uses a belief-based learning model with random errors to derive the prediction that in two player coordination games in the long run the risk-dominant Nash equilibrium will prevail. Kandori, Mailath and Rob [32] obtain a similar conclusion using a belief-based learning model with stochastic adjustment in 2x2 symmetric games with two strict equilibria.

Belief-based learning models are typically not of great interest in single person decision problems, as the prediction will always be that agents will learn to make rational decisions. The situation is different when we consider reinforcement learning models. Consider, as an example, Cross's model. Contrary to initial conjectures by Cross, it has been shown that a decision maker who adopts Cross' learning rule will sometimes, but not always learn expected payoff maximising choices. Cross' model makes predictions, however, about the frequency with which this happens, which can be subjected to experimental tests. Similarly, Bush and Mosteller's linear model sometimes predicts "probability matching" behaviour in single person decisions under risk, but, in other decision problems, leads to optimization.

Much of the learning literature so far has focused on the investigation of very specific learning rules. The work reported in this thesis will deviate from this pattern, and attempt to consider very large classes of learning rules. There are, however, already instances in the learning literature where, at least at first sight, more general models are pursued. In particular, some might regard the *replicator dynamics* of evolutionary game theory as a more general dynamic system.

The replicator dynamics was first motivated in the field of biology [54], although it has received very much attention by economists in recent years.[4] It simply states that the population share of a strategy performing better than average grows in the population, the growth rate being proportional to the payoff difference with the mean. The standard motivation for replicator dynamics has a biological flavour, appealing to asexual reproductions, programmed strategies, inheritance of strategies, etc. When applied to economic contexts, it is implicitly assumed that there are some *learning*

[4] See special numbers in the *Journal of Economic Theory* (1992) and *Games and Economic Behaviour* (1993)

and/or *imitation* models at individual level which lead to this particular dynamics. In this sense, replicator dynamics is a reduced form of learning and imitation models.

However, replicator dynamics is a very specific model, as the learning and imitation models which has been shown to lead to this particular dynamics have a very specific functional form, i.e. they are quite "ad hoc" models (Börgers and Sarin [5] for individual learning and Schlag [51] for imitation). This specificity prompted several papers which develop more general dynamics of which replicator dynamics is a special case. The more relevant generalizations are the monotonicity concept (Samuelson and Zhang [48]) and the myopic adjustment (Swinkels [53]). A dynamic is *monotonic* if for pure strategy s getting higher payoff than strategy s' then it is true that the growth rate of the share of the population playing s is higher than the growth rate associated to s'. A process is myopic if holding the opponents' play fixed, the player's utility is not decreasing along

Our work in this thesis differs from these generalizations in one essential aspect. These papers investigate properties defined over population dynamics, without tracing back their results to learning rules for individual players. On the contrary, the current thesis focuses on learning rules and, without imposing any functional form, characterises classes of learning rules which lead the agent to behave rationally in the long run for every decision problem.

1.5 Learning and Bounded Rationality

Before we turn to a survey of our results, we wish to take up one further point regarding the interpretation of learning models in economic theory. Learning models of the type described above are often accused of being "ad hoc". This objection will be less forceful in the context of our own analysis, since we impose fewer restrictions on functional forms, but in a weakened form it will still apply, since there will be significant exogenous restrictions on the class of functions that we allow. We wish to argue in the following that one way of justifying these restrictions is to argue that they reflect the limits of the decision maker's abilities to conceptualize his environment, and to give full attention to the learning problem at hand. In this section, we elaborate this point, first for belief-based learning models and then for reinforcement learning models. It is important to discuss this point in some detail since it implicitly underlies much of the work reported

in this thesis.

Belief-based learning models admit an interpretation in terms of Bayesian learning. For example, fictitious play learning model can be interpreted as a Bayesian learning model when roughly speaking, agents believe that their opponents' play corresponds to a sequence of i.i.d. random variables with a fixed but unknown distribution. If the Bayesian agent perceives the environment as relatively complex in comparison to her ability to conceptualize it, she can prefer to use a simplification of the environment. If the simplification is that other agents' choices are i.i.d. multinomial random variables and that her prior beliefs over the unknown distribution take the form of a Dirichlet distribution[5], then Bayesian learning would reduce to fictitious play.

Reinforcement learning is harder to justify. In the classical multi-armed bandit problem, a fully rational player has infinite memory, holds a subjective prior over payoff distributions and choose her actions so as to maximise the expected discounted value of her return over an infinite period. By using reinforcement rules, we restrict attention to simpler behaviour rules; specifically we assume that the decision-maker forgets all information she acquired in any previous period. We could think of it in terms of a decision-maker who decides that the problem at hand is not important enough so as to pay full attention to it. She then could find easy to use the behavioural habit that is implicit in reinforcement learning models. An alternative explanation is that there is an exogenous limitation that directly affects the agent's memory capacity. In this situation, reinforcement learning could be a suitable learning model.

A decision-maker using some reinforcement learning rule can be therefore viewed as a boundedly rational agent, as he departs from the fully rational paradigm in having limited memory. Note that the current (possible stochastic) behavioural "habit" is the only variable that can provide the decision-maker with information about the past. This arises the question whether the memory constraint implicitly assumed in reinforcement models prevents the decision-maker to achieve optimality. In this thesis, we investigate the existence of reinforcement rules (bounded rational strategies) which lead the agent to behave optimally in the long run.

[5] A random vector p has a Dirichlet distribution with parameter vector α if its density is given by $f(p) = \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1-1}\cdots p_k^{\alpha_k-1}$ for all $p > 0$ such that $\sum_{z=1}^{k} p_z = 1$.

Chapter 2

Optimality of Learning and Imitation Rules: An Overview

2.1 Introduction

As has been explained in the first Chapter, the purpose of the current thesis is to investigate the ability of trial and error learning to lead agents to achieve rationality. Within the class of trial an error learning models proposed in the literature, we shall only be concerned with models of reinforcement learning. Furthermore, we shall only investigate the properties of reinforcement learning models in decision problems under risk, not in games. The reason for these restrictions is rather pragmatic: We restrict the theoretical framework in order to make it feasible to answer the first question addressed in the introduction: "When can rational behaviour be expected?"

Reinforcement models have been useful in explaining learning in a variety of experiments. Erev and Roth have used reinforcement learning theory to explain behaviour in experiments in which subjects played extensive form games [46] as well as simultaneous move games [17]. Mookherjee and Sopher have presented experimental support for reinforcement learning in Matching Pennies [38] and in other constant sum games [39]. These papers are related to an earlier tradition in psychology based on the mathematical learning theory initiated by Bush and Mosteller [7] and Estes [18]. Authors working in this tradition typically used reinforcement learning theory to interpret experimental data about behaviour in single person decision problems rather than games. The decision problems considered were versions of the multi-armed bandit problem in which

individuals repeatedly choose between strategies with unknown payoff distributions, observing at each iteration only the payoff realization. Some relevant experimental work is surveyed in the second half of Bush and Mosteller's 1955 book [8]. An important paper in this literature is Estes and Straughan [19].

The experimental references which we have listed above differ from each other in one very important respect: Some papers, such as Roth and Erev [46], use reinforcement learning models to explain experiments in which agents' behaviour, roughly speaking, approaches equilibrium.[1] By contrast, other papers, such as Estes and Straughan [19], use reinforcement learning models to explain experiments in which agents' long run behaviour is persistently non-rational. In fact, Estes and Straughan [19] use a reinforcement learning model to explain "probability matching".

In the current thesis, we are not concerned with evaluating this experimental evidence. Rather, we want to investigate how it can be that the same class of learning models explains such very diverse findings. One answer to this question is implied by Börgers and Sarin [5]. There, they show that a model of reinforcement learning due to Cross [15] is, in the continuous time limit, identical to the replicator dynamics of evolutionary game theory, and thus leads to long run expected payoff maximisation. A seemingly innocuous variation of the same model was shown in Börgers and Sarin [6] to predict probability matching. The model in Erev and Roth [46] is similar to the model in [5], whereas the model in Estes and Straughan [19] is similar to the model in [6].

The most obvious difference between the model in [5] and the model in [6] is that in [5] it is assumed that all payoffs have a positive, reinforcing effect on the agent, whereas in [6] it is allowed that some payoffs have a negative effect on the agent, making her less likely to choose the same action again. However, it turns out that it is not always the case that it is the sign of payoff effects that determines whether a model predicts optimization. In fact, in the current thesis we shall give an example of a learning process in which some payoffs do have a negative effect on the agent, and nevertheless the agent does learn to make expected payoff maximising choices in the long run.

Therefore, in the current thesis we address the question of which reinforcement

[1] Here, we are referring to Roth and Erev's [46] discussion of the *market game* and the *best shot game*. In these two games observed behaviour actually approached a *subgame-perfect* equilibrium. Roth and Erev also discuss experiments involving the *ultimatum bargaining game*. In these experiments observed behaviour did not approach a *subgame-perfect* equilibrium. In fact it isn't clear whether it approached any Nash equilibrium.

learning rules will lead the decision maker to play in the long run the expected payoff maximising strategy in every decision problem? In answering this question, we shall not impose any functional form on the reinforcement learning models considered. In this sense, our work takes a different approach from much of the existing work on reinforcement learning, where the usual practice is to study the performance of very specific learning models.

Furthermore, we shall also take a different approach by interpreting reinforcement learning broadly. The usual practice is to consider that the decision maker only receives information about her own strategy choice and payoff. We shall allow for the possibility that the decision maker also observes the strategy choice and payoff of some other agents who are facing the same decision problem. We impose a certain restriction on the updating rule in order to capture the essence of imitation. This will allow us to study the properties of simple imitation rules. We study learning and imitation models together because they share many features. We then investigate which rules lead the decision-maker to play the expected payoff maximising choice in the long run for every decision problem and every initial state. Although we do not provide a complete characterisation, we find a property that is shown to imply long run maximisation. We furthermore characterise rules which have this property.

2.2 Individual Learning

Individual learning refers to the case of a single decision maker facing a decision problem under risk. An agent chooses repeatedly among different actions. In each iteration, she receives some random payoff the distribution of which depends on her action, but not on time. Payoffs are stochastically independent between periods. The agent has no knowledge of the payoff distributions.

The literature has modelled reinforcement learning in two different fashions. It can be modelled using the set of pure strategies as the state space of the decision maker. The individual enters each period with a pure strategy which is the strategy she is currently inclined to play. She plays that strategy, receives a payoff, and then revises

her strategy. It is possible that this strategy revision is stochastic, but all information she is carried into the next period is the new pure strategy. These learning rules are called *pure strategy learning rules.*

A different model uses the set of all mixed strategies as the state space of the decision maker. The decision maker enters each period with a probability distribution over strategies which indicates how likely she is to take any of her actions. She then plays some pure strategy, receives a payoff, and then updates the probability distribution. The new distribution only depends on the previous distribution, on the action taken, and on the payoff received. It does not depend on any other aspect of history. The new distribution forms the state with which she enters the next period. These learning rules are called *mixed strategy learning rules.*

For these two classes of learning rules we first investigate which learning rules imply that the long run probability of the expected payoff maximising actions is one, independent of what the true distribution of payoffs is. We call such learning rules *maximising.* We show that no pure strategy learning rule is maximising. Unfortunately, we have not been able to settle the question of existence of maximising learning rules for the case that the state space is the set of all mixed strategies.

We then relax the maximisation property and consider a related property called *approximate maximisation.* It means that the long run probability of the expected payoff maximising actions can be made arbitrarily close to one, independent of what the true distribution of payoffs is. It is for this property that the distinction between pure and mixed strategy rules becomes crucial. We show that no pure strategy rule is approximately maximising. We then go on to partially characterize the set of approximately maximising mixed strategy learning rules.

The property of approximate maximisation was suggested by Börgers and Sarin [5]. They show that Cross' learning rule tracks in finite time the trajectory of the replicator equation provided that it moves *very slowly.* This is true because Cross' rule implies that at any point in time the expected change in the state variable of the agent is given by the replicator equation of evolutionary game theory; as replicator dynamics maximises expected fitness, an agent who adopts Cross' rule and adjusts her probability distribution only slowly, will choose in the long run with very high probability an action

which maximises expected payoffs. This is true for all true distributions of payoffs.[2]

In this thesis we define a property called *monotonicity* and show that monotone rules are approximately maximising provided that they move very slowly. Furthermore, we provide a complete characterization of monotone rules. Monotonicity means that the expected change in the probability of the expected payoff maximising action is positive. Cross' rule has this property because the replicator equation, which characterizes the expected movement in Cross' rule, has this feature. We show that a rule is monotone if and only if it is Cross' rule, whereby payoffs may be subjected to certain linear transformations. As a consequence, all monotone learning rules are linear in payoff -note that the Cross' rule is linear in payoff- and have the feature that their expected movement is given by some transformation of the replicator dynamics.

As we have seen, it makes a big difference for our results whether one takes the state space of the decision maker to be the set of pure strategies or the set of mixed strategies. The intuition is that the state space of the learning rule provides, in our framework, the only possibility for the decision maker to store - implicitly - information about her past experiences. The set of all pure strategies is too small to store the relevant information. By contrast, the set of all mixed strategies is sufficiently rich.

2.3 Individual Imitation

In the following, we shall interpret reinforcement learning in a broader sense than before. We shall allow the decision maker to adjust her state using additional information. At each stage she will have the opportunity to observe the strategy choice and the payoff of some other agents who face the same decision problem. This different setup will allow us to study the properties of simple imitation rules.

We shall introduce a population of agents who find themselves in the same situation, i.e. they all are facing an identical decision problem. (Note that these are not the agents with whom a game will be played.) When studying which imitation rules lead the decision maker to play rationally, it is obvious that the evolution of the population will play a crucial role. Therefore, whether or not the decision maker behaves rationally

[2] In [5], which was written in a game setting, continuous time approximations are only constructed for finite time horizons. We shall show in this thesis that the result can be extended to an infinite time horizon for single person decision setting.

will depend on two factors: (i) her own imitation rule and (ii) the evolution of the population.

There are therefore many ways in which one can try to assess whether a given imitation rule is "good". One might be to ask whether this imitation rule will lead to good decisions in an arbitrary environment, i.e. in an environment in which the other agents use arbitrary behaviour or imitation rules. Another might be to ask whether this imitation rule will lead to good decisions if it is used by everybody in the population.

Here, we shall deal with two cases. Initially, we shall focus on the case that the decision maker interacts with a population of other individuals whose behaviour is exogenously given and fixed. This case is our attempt at considering an environment in which the decision maker does not rely on the evolution of all other agents towards optimal actions, but in which also the task which the imitation rule has to solve is not too demanding. Later, we shall then move to an environment in which all population members use the same imitation rule.

Considering hence first the case in which the behaviour of the surrounding population is given and fixed, we shall impose a certain restriction on the updating rule in order to capture the essence of imitation. Imitation is the act of copying others' strategies. In the case of pure strategy reinforcement rules, this concept has a direct translation: the decision maker's next period pure strategy will be restricted to be either the own action or the sampled one. In the case of mixed strategy reinforcement rules, as the state of the decision maker is a probability distribution over the set of strategies, imitation can not be defined in such a direct way. We shall assume that the decision maker only updates the probabilities attached to the actions taken and sampled.

Having set these restrictions, we develop a similar exercise as in the case of individual learning, by investigating which imitation rules will lead the decision maker to play in the long run the expected payoff maximising strategy with probability arbitrarily close to one, independent of what the true payoff distribution is and regardless of the given and fixed population behaviour. We show that there are no approximately maximising pure strategy imitation rules. For the case of mixed strategy imitation rules, we characterise the set of monotone imitation rules, showing that its basic feature is the *proportional imitation*, i.e. the change of the relevant probabilities is proportional to

the payoff difference. This implies that the probability of the currently played strategy is increased if it gets a payoff greater than the sampled strategy's payoff. This can be interpreted as a reinforcement effect in which the decision maker considers the sampled strategy's payoff as an aspiration level. If the own strategy's payoff is above this aspiration level, the currently played strategy is reinforced. This feature of the monotone imitation rules makes it possible to converge to the expected payoff maximising strategy even in the case that the optimal strategy is absent in the population.

The intuition behind these results parallels the individual learning case. Even though in this framework the decision maker can observe different strategies' payoff in the same round, the pure strategy set is too small to carry enough information from round to round. However, the mixed strategy state space is sufficiently rich.

2.4 Imitation in Populations

The last chapter of the thesis takes a different approach to assess how "good" an imitation rule is. We shall consider a finite population of agents all of whom are facing the same two-strategy decision problem. Each agent is endowed with a *pure* strategy imitation rule to adapt her behaviour. Furthermore, each agent is endowed with a sampling rule, i.e. a probability distribution over the members of the population which indicates how likely it is that she meets any other member. In each iteration, after choosing a strategy and receiving a payoff, each agent samples, according to her sampling rule, other member of the population and observes her strategy and payoff. With this information, she adapts her behaviour according to her imitation rule.

We shall assess how "good" an imitation rule is by studying the evolution of the population when all members use this imitation rule. An imitation rule is called *maximising* if the population will converge in the long run to the expected payoff maximising strategy for every decision problem and regardless of the initial population distribution. We show that there are no maximising imitation rules.

The non-existence of maximising imitation rules means that for any fixed imitation rule, there are decision problems for which the probability of the population ending up playing the best strategy is *strictly* less than one. This feature might lead to an individual *dissatisfaction* with the performance of that imitation rule. Each individual

might wonder whether there is room for individual improvement, i.e. whether there is an alternative imitation rule she could use and which could make her "better off" given that all other members of the population are using the fixed imitation rule.

To formalize this issue, we shall define an appropriate payoff function. For given population and fixed imitation rule, the payoff function for any agent using any imitation rule is defined as the time average payoff that the agent receives along the path. This payoff function will be called the asymptotic payoff. Note that in the definition of the payoff function it is implicit the assumption that the agent does *not* discount the future, i.e. an infinitely patient agent. Although it is a very restrictive assumption, this will simplify the subsequent analysis. For given population and fixed imitation rule, we can therefore define the set of imitation rules which are best responses to the fixed imitation rule. An imitation rule will be called an *equilibrium* imitation rule if it belongs to the set of best responses to itself. This means there is no other imitation rule that an agent might use such that for every decision problem and initial distribution her asymptotic payoff is at least as good, and for at least one decision problem and one initial distribution, her asymptotic payoff is greater using the alternative rule.

Unfortunately, in the general framework the characterization of equilibrium rules has proven to be a major problem. We show that the rule "never imitate" is an equilibrium rule, though we have not made further progress. We shall then consider a restricted framework in order to gain further insight. In this new framework there are two agents and two strategies which yield binary payoffs. We identify a property which is of relevance to the problem at hand. An imitation rule is *unbiased* if the asymptotic payoff to all members of a population when all members use this imitation rule is closer to the expected payoff of the optimal strategy than to the expected payoff of the suboptimal strategy. We characterize the set of unbiased rules and show that biased imitation rules are not equilibrium ones. However we have not been able to prove that this property is a sufficient condition for the equilibrium property. We finally show three examples of unbiased rules which are equilibrium rules: the rules "never imitate", "always imitate" and "imitate if better".

The rest of the thesis is as follows: in the third chapter we introduce the general framework. In Chapters 4 and 5, we investigate pure and mixed strategy learning rules respectively. Chapters 6 and 7 consider pure and mixed strategy imitation rules

respectively. The analysis of equilibrium imitation rules is undertaken in Chapter 8. Finally, Chapter 9 concludes.

Chapter 3

Formal Framework

This Chapter lists those ingredients of our model which will be common to all subsequent Chapters.

Let W be a finite population (or set) of $\#W$ decision makers, with $\#W \geq 2$. Every decision maker repeatedly faces the same decision problem. Every decision maker has to choose from the same finite set of strategies S which has at least two elements. We assume that every decision maker knows S. We denote by $\Delta(S)$ the set of all probability distributions over S, and we denote by $\dot{\Delta}(S)$ the relative interior of $\Delta(S)$. Each strategy in S has a payoff distribution attached to it. We normalize payoffs to be between zero and one. This motivates the following definition:

Definition 1 *An environment E is a collection $(\mu_s)_{s \in S}$ of probability measures each having finite support in the interval $(0,1)$. For given environment E we define for every $s \in S$: $\pi_s \equiv \int_0^1 x d\mu_s$, i.e. π_s is the expected payoff associated with strategy s. We denote by S^* the set of expected payoff maximising strategies, i.e.: $S^* \equiv \{s \in S \mid \pi_s \geq \pi_{s'}$ for all $s' \in S\}$.*

As every member of the population is facing the same decision problem, a crucial issue is how payoffs across iterations and across different decision makers are correlated. Let E be the set of states of Nature. The payoff to decision maker w at a given iteration is then a function $\pi_w : S \times E \rightarrow (0,1)$. We shall assume that realizations of the state of Nature are independent across iterations. This motivates the following definitions:

Definition 2 Common Events Condition*: For each iteration $n \in \mathbb{N}_0$, the state of Nature is realized. This state of Nature is common to every decision maker.*

Definition 3 Independent Events Condition*: For each* $n \in \mathbb{N}_0$, *the state of Nature is independently realized across decision makers.*

Let W^A denote the set of individuals belonging to the population who are allowed to adjust their behaviour. Let W^D denote the set $W \setminus W^A$. Every member of W^D is programmed to play some pure strategy $s \in S$. Formally speaking, there is a function $C : W^D \to S$ which assigns a strategy to every member of W^D. Thus, individual $w' \in W^D$ is programmed to play pure strategy $C(w')$.

When available, between stages, every $w \in W$ samples another agent from the population W and observes both the strategy used and the payoff received by the sampled agent in that stage. For each individual $w \in W$ the sampling occurs following some sampling rule, i.e. some exogenously given probability distribution $e_w \in \Delta(W \setminus \{w\})$ where $e_w(w')$ is the probability that individual w samples individual w'.

Chapter 4

Pure Strategy Rules for Individual Learning

4.1 Introduction

Individual learning refers to a situation in which there is a single decision maker facing repeatedly a decision problem. In this Chapter, we investigate, without postulating any specific functional form, which properties characterise reinforcement learning rules which predict that in the long run the decision maker makes expected payoff maximising choices.

A decision maker chooses repeatedly among different actions. In each iteration, she receives some random payoff the distribution of which depends on her action, but not on time. Payoffs are stochastically independent between periods. The decision-maker has no knowledge of the underlying payoff distributions.

The reinforcement learning rules to be considered in this Chapter are modeled as follows: the decision maker enters each period with a *pure* strategy which is the strategy which she is currently inclined to play. She plays that strategy, receives some stochastic payoff and then revises her strategy. It is possible that the strategy revision is stochastic, but all the information which is carried into the next period is the new *pure* strategy. Note that the state space of these learning rules is the set of all *pure* strategies. They are therefore named as pure strategy learning rules. Karandikar et al.

[33] for example, propose a reinforcement learning model of this type.[1]

A learning rule which makes the decision maker play in the long run the expected payoff maximising strategy independent of what the underlying true payoff distribution is, is called *maximising.* A learning rule such that the asymptotic probability of choosing the expected payoff maximising choice is arbitrarily *close* to 1 is named *approximately maximising.* We show that no pure strategy learning rule is approximately maximising.

The intuition behind this result is that the state space of the learning rule provides the only possibility for the decision maker to implicitly store information about his past experiences. The set of all pure strategies is too small to store all relevant information.

The rest of the Chapter is organised as follows. Section 2 contains the formal framework and Section 3 states the main definitions. Section 4 shows that approximate payoff maximisation is impossible to achieve. Finally, Section 5 concludes.

4.2 Formal Framework

This Chapter refers to the case in which the set W^A is a singleton, i.e. $W^A = \{w\}$ and in which sampling is *not* allowed. Individual w is referred to as the decision maker.

4.3 Definitions

We begin formally defining learning rules which have the set of all pure strategies as its state space. In the following definition, and in subsequent definitions, we prefer, in fact, the neutral expression *behaviour rule* over the expression *learning rule.*

Definition 4 *A* pure strategy behaviour rule B *is a function:* $B : S \times (0,1) \times S \to [0,1]$ *such that for all* $s \in S, x \in (0,1)$*:* $\sum_{s' \in S} B(s,x,s') = 1$.

The intuitive interpretation of a pure strategy behaviour rule B is this: $B(s,x,s')$ is the probability with which strategy s' is chosen in iteration $n+1$ if strategy s was chosen in iteration n, and the payoff received was x.

[1] Karandikar et al. model [33] has, however, in addition an endogenous aspiration level which can take values in a continuum as a state variable.

Definition 5 *The* transition matrix T *corresponding to an environment* E *and a pure strategy behaviour rule* B *is a* $\#S \times \#S$ *matrix whereby the entry in the row corresponding to strategy* s*, and the column corresponding to strategy* s'*, is:*

$$t_{s,s'} = \int_0^1 B(s, x, s') d\mu_s.$$

Definition 6 *The* behaviour process *corresponding to an environment* E*, a behaviour rule* B*, and an initial probability distribution* $\delta_0 \in \Delta(S)$ *is the Markov chain* $\{s_n\}_{n \in \mathbb{N}_0}$ *with the initial distribution* δ_0 *and with the transition matrix* T *defined in Definition 5. For every* $n \in \mathbb{N}_0$ *we denote by* $\delta_n \in \Delta(S)$ *the marginal distribution of* s_n*.*

We can now introduce the first property which we analyze in this Chapter:

Definition 7 *A pure strategy behaviour rule* B *is* maximising *if for every environment* E *and every initial distribution* $\delta_0 \in \dot{\Delta}(S)$*:*

$$\lim_{n \to \infty} \delta_n(S^*) = 1.$$

Proposition 1 *No pure strategy behaviour rule is maximising.*

This result is almost obvious. Therefore, we do not provide a formal proof. Instead we briefly sketch the argument. Note first that any rule with $B(s, x, s) < 1$ for some $s \in S$ and $x \in (0, 1)$ cannot be maximising. This is because for every $s \in S$ and every $x \in (0, 1)$ there are environments E in which s is the only expected payoff maximising action, and s yields payoff x.[2] In such environments, if $B(s, x, s)$ were less than one, the agent would switch away from the expected payoff maximising action with positive probability, and therefore the asymptotic probability of playing that action could not be one. But now suppose that $B(s, x, s) = 1$ for all $s \in S$ and $x \in (0, 1)$. Then the decision maker sticks forever with the strategy which she chose initially. Clearly, this is not maximising.

[2] Recall that the set of possible payoffs is (0,1). Note that 0 is not included in this set.

4.4 Approximate Payoff Maximisation

Proposition 1 leads us to seek behaviour rules which are *approximately maximising.* Roughly speaking, we mean by this behaviour rules for which $\lim_{n\to\infty} \delta_n(S^*)$ is *close* to 1 in all environments E and for all interior initial distributions δ_0. More specifically, we shall ask whether there is a family of behaviour rules, parametrized by some parameter ν, such that in all environments E the asymptotic probability of choosing an optimal action converges to 1 as ν tends to infinity. Thus, by choosing a sufficiently large ν the probability of choosing an optimal action can be made arbitrarily close to 1.

Definition 8 *A sequence of pure strategy behaviour rules* $\{B^\nu\}_{\nu\in\mathbb{N}}$ *is* approximately maximising *if for every environment E, every initial distribution $\delta_0 \in \dot{\Delta}(S)$ and every $\nu \in \mathbb{N}$ the limit* $\lim_{n\to\infty} \delta_n^\nu(S^*)$ *(where for every $n \in \mathbb{N}_0$ δ_n^ν is the marginal distribution of s_n if the initial distribution is δ_0 and if the behaviour rule is B^ν) exists, and we have:*

$$\lim_{\nu\to\infty} \lim_{n\to\infty} \delta_n^\nu(S^*) = 1.$$

The following proposition states that no pure strategy behaviour rule is approximately maximising.

Proposition 2 *No sequence of pure strategy behaviour rules is approximately maximising.*

Proof of Proposition 2. The proof is indirect. Suppose $\{B^\nu\}_{\nu\in\mathbb{N}}$ is a sequence of pure strategy behaviour rules which is approximately maximising.

Step 1: For all $s \in S$ and $x \in (0,1)$ there is some $\bar{\nu} \in \mathbb{N}$ such that for all $\nu \geq \bar{\nu}$:

$$0 < B^\nu(s,x,s) < 1$$

Proof: (i) We first show that there is some $\bar{\nu} \in \mathbb{N}$ such that $\nu \geq \bar{\nu}$ implies $B^\nu(s,x,s) > 0$. The proof is indirect. Suppose for every $\bar{\nu} \in \mathbb{N}$ there were some $\nu \geq \bar{\nu}$ such that $B^\nu(s,x,s) = 0$. Consider an environment such that $\mu_s(x) = 1$ and, for some $y < x$, $\mu_{s'}(y) = 1$ for all $s' \neq s$. For every ν such that $B^\nu(s,x,s) = 0$ we have: $\delta_n^\nu(s) \leq 1 - \delta_{n-1}^\nu(s)$ for all $n \in \mathbb{N}$. Therefore, if $\delta_n^\nu(s)$ converges for $n \to \infty$, its limit

cannot be more than 0.5. We have thus obtained a contradiction to the approximate maximisation property of $\{B^\nu\}_{\nu\in\mathbb{N}}$.

(ii) Next, we show that there is some $\bar{\nu} \in \mathbb{N}$ such that $\nu \geq \bar{\nu}$ implies $B^\nu(s,x,s) < 1$. The proof is indirect. Suppose for every $\bar{\nu}$ there were some $\nu \geq \bar{\nu}$ such that $B^\nu(s,x,s) = 1$. Consider an environment E such that $\mu_s(x) = 1$ and, for some $y > x$, $\mu_{s'}(y) = 1$ for all $s' \neq s$. If the decision maker begins with an initial distribution δ_0 which attaches positive probability to the strategy s, and adopts a behaviour rule B^ν such that $B^\nu(s,x,s) = 1$, then for all $n \in \mathbb{N}$: $\delta_n^\nu(S \setminus \{s\}) \leq 1 - \delta_0(s)$. Hence, if $\delta_n^\nu(S \setminus \{s\})$ converges for $n \to \infty$, its limit must be less than $1 - \delta_0(s) < 1$. We have thus obtained a contradiction to the approximate maximisation property $\{B^\nu\}_{\nu\in\mathbb{N}}$.

(iii) The result now holds if we take $\bar{\nu}$ to be the maximum of the two $\bar{\nu}$s referred to in parts (i) and (ii) of the proof.

Step 2: For any environment E and any initial distribution $\delta_0 \in \dot{\Delta}(S)$ there exists some $\bar{\nu}$ such that for $\nu \geq \bar{\nu}$ the behaviour process $\{s_n\}_{n\in\mathbb{N}}$ is an irreducible and aperiodic Markov chain.

Proof: Consider a given and fixed environment E and initial distribution δ_0. It follows immediately from Step 1 that the behaviour process $\{s_n\}_{n\in\mathbb{N}}$ is aperiodic for sufficiently large ν. It therefore only remains to show that it is for sufficiently large ν also irreducible.

We hence have to show there is some $\bar{\nu}$ such that for $\nu \geq \bar{\nu}$ the transition matrix T^ν has the property that all states communicate, i.e. for every pair of strategies s, s' there is some $N(s,s')$ such that the probability of moving from s to s' in $N(s,s')$ steps is positive.

The proof is indirect. Suppose there is some pair $s, s' \in S$ such that for every $\bar{\nu}$ there is some $\nu \geq \bar{\nu}$ such that the probability of moving from s to s' in any finite number of steps is zero. Step 1 shows that it must be that $s' \neq s$. Now, for all relevant ν and for all $n \in \mathbb{N}$: $\delta_n^\nu(s') \leq 1 - \delta_0(s)$. Hence, if $\delta_n^\nu(s')$ converges for $n \to \infty$, its limit will not be more than $1 - \delta_0(s)$. Observe that this is true independent of the payoff distribution attached to s'. In particular, it is true if s' is the unique expected utility maximising strategy. Thus, we have obtained a contradiction to the approximate maximisation property $\{B^\nu\}_{\nu\in\mathbb{N}}$.

Remark: By standard results in the theory of Markov chains[3], Step 2 implies that for any given environment E there is some $\bar{\nu} \in \mathbb{N}$ such that $\nu \geq \bar{\nu}$ implies that the sequence $\{\delta_n^\nu\}_{n\in\mathbb{N}}$ converges, and that its limit is a stationary distribution of the transition matrix T^ν and is independent of the initial value δ_0. In the following, we will denote this limit by δ_∞^ν.

Step 3: For all $s, s' \in S$ with $s \neq s'$ and for all $x, y \in (0,1)$ with $y > x$:

$$\lim_{\nu\geq\bar{\nu},\nu\to\infty} \frac{B^\nu(s,y,s')}{1-B^\nu(s',x,s')} = 0$$

where $\bar{\nu}$ is such that for all $\nu \geq \bar{\nu}$ we have: $B^\nu(s',x,s') < 1$.[4]

Proof: Consider an environment E such that $\mu_s(y) = 1$ and $\mu_{s'}(x) = 1$ for all $s' \in S$, $s' \neq s$. Choose $\bar{\nu} \in \mathbb{N}$ such that Step 1 applies to payoff x and all strategies $s' \neq s$, and such that Step 2 applies to the environment E. Because δ_∞^ν is a stationary distribution of T^ν, I have for all $s' \neq s$ and $\nu \geq \bar{\nu}$:

$$\delta_\infty^\nu(s') = \delta_\infty^\nu(s')B^\nu(s',x,s') + \delta_\infty^\nu(s)B^\nu(s,y,s') + \sum_{s''\neq s,s'} \delta_\infty^\nu(s'')B^\nu(s'',x,s')$$

$$\Rightarrow \delta_\infty^\nu(s') \geq \delta_\infty^\nu(s')B^\nu(s',x,s') + \delta_\infty^\nu(s)B^\nu(s,y,s')$$

$$\Leftrightarrow \frac{\delta_\infty^\nu(s')}{\delta_\infty^\nu(s)} \geq \frac{B^\nu(s,y,s')}{1-B^\nu(s',x,s')}$$

The approximate maximisation property of the sequence $\{B^\nu\}_{\nu\in\mathbb{N}}$ implies that the left hand side of the above inequality tends to zero as $\nu \to \infty$. Then also the right hand side of the above inequality must tend to zero as $\nu \to \infty$.

Step 4: Suppose $x, y, z \in (0,1)$ and $x < z < y$. Consider an environment E such that $\mu_s(x) > 0$ and $\mu_s(y) > 0$ for some $s \in S$, and $\mu_{s'}(z) = 1$ for all $s' \in S$ with $s' \neq s$. Then

$$\lim_{\nu\geq\bar{\nu},\nu\to\infty} \delta_\infty^\nu(s) = 0$$

where ν is such that for $\nu \geq \bar{\nu}$ Step 2 applies, and hence the distribution δ_∞^ν is well-defined.

[3] See, for example, p.214 of Grimmett and Stirzaker [22].

[4] Recall that by Step 1 such a $\bar{\nu}$ exists.

Proof: Choose $\bar{\nu} \in \mathbb{N}$ such that Step 1 applies to payoff x and strategy s, and such that Step 2 applies to the environment E. For any $\nu \geq \bar{\nu}$ the distribution δ^{ν}_{∞} is a stationary distribution of the transition matrix T^{ν}. This implies:

$$\delta^{\nu}_{\infty}(s) = \delta^{\nu}_{\infty}(s)(\mu_s(x)B^{\nu}(s,x,s) + \mu_s(y)B^{\nu}(s,y,s)) + \sum_{s' \neq s} \delta^{\nu}_{\infty}(s')B^{\nu}(s',z,s)$$

$$\Leftrightarrow \delta^{\nu}_{\infty}(s) = \frac{\sum_{s' \neq s} \delta^{\nu}_{\infty}(s')B^{\nu}(s',z,s)}{\mu_s(x)(1 - B^{\nu}(s,y,s)) + \mu_s(y)(1 - B^{\nu}(s,x,s))}$$

$$\Leftrightarrow \frac{\delta^{\nu}_{\infty}(s)}{1 - \delta^{\nu}_{\infty}(s)} = \frac{\sum_{s' \neq s} \frac{\delta^{\nu}_{\infty}(s')}{1 - \delta^{\nu}_{\infty}(s)} B^{\nu}(s',z,s)}{\mu_s(x)(1 - B^{\nu}(s,y,s)) + \mu_s(y)(1 - B^{\nu}(s,x,s))}$$

$$\Rightarrow \frac{\delta^{\nu}_{\infty}(s)}{1 - \delta^{\nu}_{\infty}(s)} \leq \frac{1}{\mu_s(y)} \sum_{s' \neq s} \frac{\delta^{\nu}_{\infty}(s')}{1 - \delta^{\nu}_{\infty}(s)} \frac{B^{\nu}(s',z,s)}{1 - B^{\nu}(s,x,s)}$$

$$\Rightarrow \frac{\delta^{\nu}_{\infty}(s)}{1 - \delta^{\nu}_{\infty}(s)} \leq \frac{1}{\mu_s(y)} \sum_{s' \neq s} \frac{B^{\nu}(s',z,s)}{1 - B^{\nu}(s,x,s)}$$

By Step 3 for every $s' \neq s$:

$$\lim_{\nu \geq \nu, \nu \to \infty} \frac{B^{\nu}(s',z,s)}{1 - B^{\nu}(s,x,s)} = 0$$

This implies:

$$\lim_{\nu \geq \nu, \nu \to \infty} \sum_{s' \neq s} \frac{B^{\nu}(s',z,s)}{1 - B^{\nu}(s,x,s)} = 0$$

Thus we can deduce:

$$\lim_{\nu \geq \nu, \nu \to \infty} \frac{\delta^{\nu}_{\infty}(s)}{1 - \delta^{\nu}_{\infty}(s)} = 0$$

This implies:

$$\lim_{\nu \geq \nu, \nu \to \infty} \delta^{\nu}_{\infty}(s) = 0$$

Step 5: The sequence $\{B^{\nu}\}_{\nu \in \mathbb{N}}$ is not approximately maximising.

Proof: Consider an environment E as described in Step 4 with the additional property that s maximises expected payoff. The result in Step 4 applies in this case. Hence $\lim_{\nu \geq \bar{\nu}, \nu \to \infty} \delta^{\nu}_{\infty}(s) = 0$. This contradicts the approximate maximisation property of the sequence $\{B^{\nu}\}_{\nu \in \mathbb{N}}$. ■

Proposition 2 shows that no pure strategy behaviour rule can achieve approximate payoff maximisation. The intuition behind this result is that the memory store provided

by a pure strategy state space is too small to deal with a large variety of environments.

4.5 Conclusion

The main result of this Chapter is the non-existence of approximately boundedly rational strategies for multi-armed bandit problems within the framework of pure strategy reinforcement rules. These behaviour rules have no memory beyond what is encoded in the state of the decision maker. In fact, the decision maker's state space provides the only possibility for the decision-maker to store information about her past experiences. But the memory store provided by a pure strategy state space is too small to deal with a large variety of environments, even if we ask only for *approximate* maximisation.

A number of papers in the economics literature have investigated pure strategy reinforcement rules, although mainly using aspiration-based models. Once the agent has played an action and received a payoff, she switches with positive probability from the action played if the achieved payoff falls below the aspiration level. Bendor, Mookherjee and Ray [3] and Karandikar, Mookherjee, Ray and Vega-Redondo [33] use aspiration-based models to study the behaviour in two-players games. In [3], the aspiration level is kept fixed whereas in [33] it evolves on the basis of the agent's own experienced payoffs. Other papers like Palomino and Vega-Redondo [44] apply these models to a large-population context with the common aspiration level evolving on the basis of social experience.

Chapter 5

Mixed Strategy Rules for Individual Learning

5.1 Introduction

Individual learning refers to a situation in which there is a single decision maker facing repeatedly a decision problem. We investigate, without postulating any specific functional form, which properties characterise reinforcement learning rules which predict that in the long run the decision maker makes expected payoff maximising choices.

A decision maker chooses repeatedly among different actions. In each iteration, she receives some random payoff the distribution of which depends on her action, but not on time. Payoffs are stochastically independent between periods. The decision maker has no knowledge of the payoff distributions.

In this Chapter we model reinforcement learning in a different fashion from the previous Chapter, i.e. at any point in time the decision maker is described by a probability distribution over actions which indicates how likely she is to take any of her actions. The decision maker then takes a randomly determined action, receives a payoff, and then updates the probability distribution. The new distribution only depends on the previous distribution, on the action taken, and on the payoff received. It does not depend on any other aspect of history. Reinforcement learning rules are therefore formulated using the set of all mixed strategies as the state space.

An example of a simple rule of this kind is the one considered by Cross [15] to which we referred in Chapter 1. Suppose payoffs are normalized between zero and one.

Then Cross' rule states that in each iteration the updated probability distribution is simply a weighted average of the previous probability distribution and the unit vector placing all probability on the action just taken. The weight of the unit vector equals the payoff. Note that this rule has the feature that all payoffs have a reinforcing, and hence positive, effect on the probability of the action which the decision maker chose.

A simple calculation shows that Cross' rule implies that at any point in time the expected change in the state variable of the decision maker is given by the replicator equation of evolutionary game theory if this equation is specialized to the case of a single person decision problem. In Börgers and Sarin [5] it was shown that this implies that Cross' learning rule tracks the trajectory of the replicator equation provided that it moves very slowly. Because evolution, as modelled by the replicator equation, maximises expected fitness, an agent who adopts Cross' learning rule and adjusts her probability distribution only slowly, will choose in the long run with very high probability an action which maximises expected payoffs. This is true for all true distributions of payoffs.[1]

In this Chapter we investigate which other reinforcement learning rules share with the slow moving Cross rule the feature that the long run probability of expected payoff maximising actions is close to one independent of what the true distribution of payoffs is. We call such rules *approximately maximising*. We do not obtain a complete characterization of such rules, but we define a property of learning rules called *monotonicity* which is of immediate relevance to our problem, and we obtain a complete characterization of monotone learning rules.

Monotonicity means that the expected change in the probability of the expected payoff maximising action is positive. Cross' rule has this property because the replicator equation, which characterizes the expected movement in Cross' rule, has this feature. Monotonicity matters because, as we show in this Chapter, all monotone learning rules are approximately maximising, provided that they move slowly.

We will show that a rule is monotone if and only if it is Cross' rule, whereby payoffs may be subjected to certain linear transformations. As a consequence, all monotone learning rules have the feature that their expected movement is given by

[1] In [5], which was written in a game setting, continuous time approximations are only constructed for finite time horizons. We shall show in this Chapter for the single person decision setting that the result can be extended to an infinite time horizon.

some transformation of the replicator dynamics.

The rest of the Chapter is organized as follows. Section 2 contains the formal framework. Section 3 sets up the framework for mixed strategy learning rules. Section 4 introduces the monotonicity property and its implications. Section 5 provides a characterization of monotone learning rules. Section 6 contains the proof of the main proposition of the Chapter: the monotonicity characterization.. And finally, Section 7 concludes.

5.2 Formal Framework

This chapter refers to the case in which the set W^A is a singleton, i.e. $W^A = \{w\}$ and in which sampling is not allowed. Individual w is referred to as the decision maker.

5.3 Definitions

We begin formally defining learning rules which have the set of all mixed strategies as its state space. In the following definition, and in subsequent definitions, we prefer, in fact, the neutral expression *behaviour rule* over the expression *learning rule.*

Definition 9 *A* mixed strategy behaviour rule B *is a function:* $B : \Delta(S) \times S \times (0,1) \to \Delta(S)$.[2]

The intuitive interpretation of a mixed strategy behaviour rule B is this: At each iteration n the decision maker's behaviour is described by a probability distribution $\sigma_n \in \Delta(S)$ which specifies for each pure strategy s how likely it is that the decision maker chooses s at iteration n. We shall also refer to σ_n as the *state of the decision maker at iteration n.* The distribution $B(\sigma_n, s, x)$ is then the state of the decision maker at iteration $n+1$ if her state at iteration n was σ_n, the pure strategy which she chose at iteration n was s, and the payoff which she received was x. For every $s' \in S$ we denote by $B(\sigma_n, s, x)(s')$ the probability which $B(\sigma_n, s, x)$ assigns to s'.

Throughout this Chapter we shall focus on behaviour rules which satisfy the following assumption:

[2]Note that we use, for simplicity, the same symbol B as in the previous Chapter to denote mixed strategy behaviour rules.

Assumption 1. *For any $s \in S$ the mixed strategy behaviour rule B is continuously differentiable in (σ_n, x) and the derivative of B with respect to (σ_n, x) is bounded from above and from below.*

This assumption allows us in Section 4 to appeal to well-known theorems regarding the approximation of slow moving stochastic processes by the solution of deterministic differential equations.

We denote by $\mathcal{B}(\Delta(S))$ the set of all Borel subsets of $\Delta(S)$.

Definition 10 *The* stochastic kernel K *corresponding to an environment E and a mixed strategy behaviour rule B is a function $K : \Delta(S) \times \mathcal{B}(\Delta(S)) \to [0,1]$ such that*

$$K(\sigma, \Omega) = \sum_{(s,x) \in \{s \in S, x \in (0,1) | B(\sigma,s,x) \in \Omega\}} \sigma(s) \cdot \mu_s(x)$$

for every $\sigma \in \Delta(S)$ and $\Omega \in \mathcal{B}(\Delta(S))$.

Intuitively, the *stochastic kernel* is the analog of a transition matrix for a Markov process with continuum size state space.

Definition 11 *The* behaviour process *corresponding to an environment E, a mixed strategy behaviour rule B, and an initial state $\sigma_0 \in \Delta(S)$ is the Markov process $\{\sigma_n\}_{n \in \mathbb{N}_0}$ with the initial distribution which assigns probability 1 to σ_0, and with the stochastic kernel K described in Definition 10.*

Definition 12 *A mixed strategy behaviour rule B is* maximising *if for every environment E and every initial state $\sigma_0 \in \dot{\Delta}(S)$ the probability of the event "$\sigma_n(S^*) \to 1$" is 1.*

We have not been able to settle the following, intriguing question:

Open Question *Do mixed strategy behaviour rules which are maximising exist?*

Although we have to leave this question unanswered, we do have interesting results concerning a class of *approximately maximising* behaviour rules. Here, the concept of *approximate maximisation* is defined in the same way as in Chapter 4.

Definition 13 *A sequence of mixed strategy behaviour rules $\{B^\nu\}_{\nu \in \mathbb{N}}$ is* approximately maximising *if for every environment E and every initial state $\sigma_0 \in \dot{\Delta}(S)$ the*

probability of the event "$\sigma_n^\nu(S^*) \to 1$" converges to 1 as $\nu \to \infty$. Here, $\{\sigma_n^\nu\}_{n \in \mathbb{N}_0}$ denotes the behaviour process corresponding to the behaviour rule B^ν and the initial state σ_0.

5.4 Monotone Behaviour Rules

For any mixed strategy behaviour rule B and environment E, we define a function f which assigns to every possible state of the decision maker σ, and every pure strategy s, the expected change in the probability attached to s if the current state is σ. Formally, $f : \Delta(S) \times S \to \mathbb{R}$ is defined by:

$$f(\sigma, s) = \sum_{s' \in S} \sigma(s') \int_0^1 B(\sigma, s', x)(s) - \sigma(s) d\mu_{s'}$$

for all $\sigma \in \Delta(S)$ and $s \in S$. For $\tilde{S} \subseteq S$, we define

$$f(\sigma, \tilde{S}) = \sum_{s \in \tilde{S}} f(\sigma, s).$$

Finally, we denote by $f(\sigma)$ the vector:

$$f(\sigma) = \{f(\sigma, s)\}_{s \in S}.$$

Definition 14 *A mixed strategy behaviour rule B is called* monotone *if*

(1) $\sigma \in \dot{\Delta}(S)$, $s \in S$ and $x \in (0,1)$ imply $B(\sigma, s, x) \in \dot{\Delta}(S)$

(2) for all environments E with $S^* \neq S$ and all states $\sigma \in \dot{\Delta}(S)$: $f(\sigma, S^*) > 0$.

Condition (1) guarantees that the behaviour process stays in the interior of the mixed strategy simplex provided that it starts in the interior. This makes it easier for us to appeal to continuous time approximations later in this Section. The main point is condition (2). It is this condition which motivates the label *monotonicity*. It says that, in expected terms, the decision maker will approach the expected payoff maximising choice in a monotonically increasing way.

In the following, we shall simplify terminology, and we shall call monotone mixed strategy behaviour rules simply "monotone behaviour rules". In the remainder of this

Section we present some propositions which motivate our interest in monotone behaviour rules.

Proposition 3 *Suppose that B is a monotone behaviour rule, and that E is an environment. Then for any initial state $\sigma_0 \in \dot{\Delta}(S)$ the event*

$$\text{“}\sigma_n(S^*) \to 0 \text{ or } \sigma_n(S^*) \to 1\text{”}$$

has probability 1.

Proof of Proposition 3. By the definition of monotone behaviour rules the stochastic process $\{\sigma_n(S^*)\}_{n\in\mathbb{N}_0}$ is a submartingale which is bounded from below by 0 and from above by 1. Therefore, by the *Martingale Convergence Theorem* (Grimmett and Stirzaker [22], p. 454), $\{\sigma_n(S^*)\}_{n\in\mathbb{N}_0}$ converges almost surely to a limit random variable σ_∞.

It remains to show that $\sigma_\infty(S^*) = 0$ or 1 with probability 1. This follows if we can show that for every pair $\alpha, \beta \in (0,1)$ with $\alpha < \beta$ the probability of $\sigma_\infty(S^*) \in [\alpha, \beta]$ is zero. Let a, b satisfy: $0 < a < \alpha < \beta < b < 1$. For every $\bar{n} \in \mathbb{N}$ and $\eta > 0$ let $\Phi^\eta_{\bar{n}}$ denote the event “$\sigma_n(S^*) \in [a,b]$ and $\mid \sigma_n(S^*) - \sigma_{n-1}(S^*) \mid \leq \eta$ for all $n \geq \bar{n}$”. Clearly, for every $\eta > 0$ the event $\sigma_\infty \in [\alpha, \beta]$ is contained in the event $\cup_{\bar{n}\in\mathbb{N}}\Phi^\eta_{\bar{n}}$, and therefore it suffices to show that for some $\eta > 0$ the probability of $\cup_{\bar{n}\in\mathbb{N}}\Phi^\eta_{\bar{n}}$ is zero. This follows if we show that for some $\eta > 0$ the probability of $\Phi^\eta_{\bar{n}}$ is zero for every $\bar{n} \in \mathbb{N}$.

Fix $\bar{n}$. For every $n \geq \bar{n}$ we denote by Ψ^η_n the event “$\sigma_n(S^*) \in [a,b]$ and $\mid \sigma_n(S^*) - \sigma_{n-1}(S^*) \mid \leq \eta$”. Write $\Pr(\Psi^\eta_n)$ for the probability of Ψ^η_n and write $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ for the probability of Ψ^η_{n+1} conditional on Ψ^η_n. Then the probability of $\Phi^\eta_{\bar{n}}$ can be written as: $\Pr(\Psi^\eta_{\bar{n}}) \cdot \Pr(\Psi^\eta_{\bar{n}+1} \mid \Psi^\eta_{\bar{n}}) \cdot \Pr(\Psi^\eta_{\bar{n}+2} \mid \Psi^\eta_{\bar{n}+1}) \cdot \Pr(\Psi^\eta_{\bar{n}+3} \mid \Psi^\eta_{\bar{n}+2}) \cdot \ldots$. Suppose we can show that there is some ζ with $0 < \zeta < 1$ such that for all $n \in \mathbb{N}$ the conditional probability $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ is bounded from above by ζ. Then the above infinite product converges to zero, and therefore the proof is complete.

Consider the expected value of $\sigma_{n+1}(S^*) - \sigma_n(S^*)$ conditional on Ψ^η_n. This is bounded from below by $\xi \equiv \min_{\sigma\in\Delta(S) \text{ and } \sigma(S^*)\in[a,b]} f(\sigma, S^*)$. Because B is monotone, $\xi > 0$. Now consider the probability that $\sigma_{n+1}(S^*) - \sigma_n(S^*) < \frac{\xi}{2}$, conditional on Ψ^η_n. Intuitively, for the *expected* change to be at least ξ, the probability that the *actual*

change is less than $\frac{\xi}{2}$ must not be too large. In fact, a trivial calculation shows that it cannot be more than $\frac{\xi}{2-\xi}$.

Now set $\eta = \frac{\xi}{2}$. Then the preceding paragraph implies that $\Pr(\Psi^{\eta}_{n+1} \mid \Psi^{\eta}_{n})$ is bounded from above by $\frac{\xi}{2-\xi}$. Thus we have found a uniform upper bound for $\Pr(\Psi^{\eta}_{n+1} \mid \Psi^{\eta}_{n})$ which is less than one, and the proof is complete. ■

This result is similar to Proposition 2 in Börgers and Sarin [5]. Börgers and Sarin's result refers, however, only to one special example of a monotone behaviour rule, namely Cross' rule which is Example 1 in the next Section. Börgers and Sarin prove their result using Theorem 2.3 in Norman [41] which applies to a larger class of learning models. However, Norman's result does not seem to apply in our context, and the proof of Proposition 3 does not rely on Norman's result, and is different from his proof.

For given behaviour rule B and given environment E we define a function g which assigns to every state $\sigma_0 \in \dot{\Delta}(S)$ the probability $g(\sigma_0)$ of the event $\sigma_n(S^*) \to 1$ if the initial state is σ_0.

Proposition 4 *Suppose that B is a monotone behaviour rule, and that E is an environment such that $S \neq S^*$. Then, for any $\sigma_0 \in \dot{\Delta}(S)$, $g(\sigma_0) > \sigma_0(S^*)$.*

Proof of Proposition 4. Because B is monotone, the unconditional expected values satisfy: $E(\sigma_n(S^*)) < E(\sigma_{n+1}(S^*))$ for all $n \in \mathbb{N}_0$. Hence: $\lim_{n\to\infty} E(\sigma_n(S^*)) > \sigma_0(S^*)$. Proposition 3 implies: $\lim_{n\to\infty} E(\sigma_n(S^*)) = g(\sigma_0)$. Thus, we can conclude: $g(\sigma_0) > \sigma_0(S^*)$. ■

This proposition describes a very weak, but certainly desirable property of monotone behaviour rules: in all non-trivial environments the probability with which the decision maker ends up playing an expected payoff maximising strategy is larger than the probability with which the decision maker played some such strategy initially.

Stronger results can be proved for monotone behaviour rules which move in small steps. We shall present two propositions which apply to this case. For any given monotone behaviour rule B we define for every $\varepsilon \in (0,1)$ a new behaviour rule B^{ε} by setting

$$B^{\varepsilon}(\sigma, s, x) - \sigma = \varepsilon(B(\sigma, s, x) - \sigma).$$

Intuitively, B^{ε} describes a behaviour process which moves into the same direction as B, but at speed ε. We are interested in limit properties of the behaviour process

corresponding to B^ε for fixed environment, and fixed initial state, where the limit which we wish to take is: $\varepsilon \to 0$.

For our first result we introduce a continuous time variable $t \geq 0$, and we adapt the behaviour process introduced in Definition 11 so that it is a continuous time behaviour process. If the behaviour rule is B^ε, we assume that the amount of "real time" which passes between two iterations of the decision problem equals ε. In the time interval which passes between two iterations of the decision problem the state of the decision maker remains constant. This motivates the following definition:

Definition 15 *The* continuous time behaviour process *corresponding to an environment E, a mixed strategy behaviour rule B^ε, and an initial state $\sigma_0 \in \dot{\Delta}(S)$ is the stochastic process $\{\tilde{\sigma}_t^\varepsilon\}_{t\geq 0}$ whose initial distribution assigns probability 1 to σ_0, and which satisfies for any $t \geq 0$:*[3]

$$\tilde{\sigma}_t^\varepsilon = \sigma^\varepsilon_{[\frac{t}{\varepsilon}]}$$

where $\{\sigma_n^\varepsilon\}_{n\in\mathbb{N}_0}$ is the (discrete time) behaviour process corresponding to B^ε.

This definition of the continuous time behaviour process has the following desirable feature. If one investigates the behaviour process in the case that ε is close to zero, but fix some time interval $[0, t]$, then, as ε is reduced, the number of iterations over which we keep track of the decision maker's behaviour is correspondingly increased. If one didn't increase the number of iterations, but instead kept it fixed, then, if ε were close to zero, almost no change in the decision maker's behaviour would be observed.

To characterise the limit as $\varepsilon \to 0$ we introduce a deterministic dynamic process which starts in σ_0, and which moves into the direction of the *expected movement* of B^ε. Formally, we define for the behaviour process introduced in Definition 15 a corresponding deterministic continuous time process $\{\hat{\sigma}_t^\varepsilon\}_{t\geq 0}$ by setting $\hat{\sigma}_0 = \sigma_0$, and, for every $t > 0$, $\hat{\sigma}_t^\varepsilon = \hat{\sigma}^\varepsilon_{[\frac{t}{\varepsilon}]}$ if $\frac{t}{\varepsilon}$ is not an integer, and $\hat{\sigma}_t^\varepsilon = \hat{\sigma}^\varepsilon_{[\frac{t}{\varepsilon}]-1} + f^\varepsilon(\sigma^\varepsilon_{[\frac{t}{\varepsilon}]-1})$ otherwise. Here, f^ε is the function that describes the expected movement of the behaviour rule B^ε.

Proposition 5 *Suppose that B is a monotone behaviour rule, that E is an environment, and that $\sigma_0 \in \dot{\Delta}(S)$. Then for any $t > 0$, $\delta > 0$ and $p \in [0, 1)$ there is an $\bar{\varepsilon} > 0$*

[3] For $x \in \mathbb{R}$ we denote by $[x]$ the largest integer smaller or equal to x.

such that $\varepsilon \leq \bar{\varepsilon}$ *implies that the probability of the event*

$$\text{“} \max_{0 \leq \tau \leq t} \|\tilde{\sigma}_\tau^\varepsilon - \hat{\sigma}_\tau^\varepsilon\| \leq \delta \text{”}$$

is at least p.

This proposition shows that over any finite time horizon the stochastic behaviour process $\{\tilde{\sigma}_t^\varepsilon\}_{t \geq 0}$ stays with high probability close to the deterministic process $\{\hat{\sigma}_t^\varepsilon\}_{t \geq 0}$, provided that ε is close to zero. Standard results in numerical mathematics show moreover that over any finite time horizon the deterministic process $\{\hat{\sigma}_t^\varepsilon\}_{t \geq 0}$ stays close to the solution of the differential equation

$$\frac{d\bar{\sigma}_t}{dt} = f(\bar{\sigma}_t),$$

provided that ε is close to zero.[4] Thus we can conclude that over finite time intervals and for small ε the solution of the differential equation constitutes a good approximation to the behaviour process.

We omit the proof of Proposition 5. The result is very closely related to Proposition 1 in Börgers and Sarin [5]. Börgers and Sarin's result applies only to the special case of Cross' rule. However, the general case is not different. In Börgers and Sarin, but also in general, the result can be proved using Theorem 1.1 in Chapter 8 of Norman [42]. A technical point is that Norman's result applies only to particular points in time, t, whereas in Proposition 5 we refer to a time interval $[0, t]$. The extension to time intervals can be derived along the lines indicated in Corradi and Sarin [14].

Notice that the deterministic process to which Proposition 5 refers has the property:

$$\lim_{t \to \infty} \hat{\sigma}_t^\varepsilon(S^*) = 1$$

provided that the initial state is interior. Thus, Proposition 5 comes very close to asserting that for small ε the behaviour rule B^ε is approximately maximising. However, Proposition 5 considers only *finite* time intervals $[0, t]$. Therefore, we provide a further result which concerns the *asymptotics* for $t \to \infty$ of the decision maker's behaviour.

[4] See, for example, Theorem 203A in Butcher [9].

Proposition 6 *Suppose that B is a monotone behaviour rule, and that E is an environment. Then for all $\sigma_0 \in \dot{\Delta}(S)$:*

$$\lim_{\varepsilon \to 0} g^{\varepsilon}(\sigma_0) = 1.$$

Here, the function g^{ε} assigns to every initial state σ_0 the probability of the event "$\sigma_n^{\varepsilon}(S^*) \to 1$" if the behaviour rule is B^{ε}.

Proof of Proposition 6. Consider a given and fixed initial state $\sigma_0 \in \dot{\Delta}(S)$. Recall that this implies: $\lim_{t \to \infty} \hat{\sigma}_t^{\varepsilon}(S^*) = 1$. Therefore for every $\delta > 0$ there will be a $t > 0$ such that $\hat{\sigma}_t^{\varepsilon}(S^*) \geq 1 - \delta$. By Proposition 5 there will then be for every $\delta > 0$ and $p \in [0, 1)$ a $t > 0$ and an $\bar{\varepsilon} > 0$ such that $\varepsilon \leq \bar{\varepsilon}$ implies that the probability of the event "$\tilde{\sigma}_t^{\varepsilon}(S^*) \geq 1 - \delta$" is at least p. Now recall from Proposition 4 that, conditional on "$\tilde{\sigma}_t^{\varepsilon}(S^*) \geq 1 - \delta$" the probability of "$\tilde{\sigma}_t^{\varepsilon} \to 1$" is at least $1 - \delta$. Thus we can conclude that for every $\delta > 0$ and $p \in [0, 1)$ there will be an $\bar{\varepsilon} > 0$ such that $\varepsilon \leq \bar{\varepsilon}$ implies that the probability of the event "$\tilde{\sigma}_t^{\varepsilon} \to 1$" is at least $p(1 - \delta)$. This implies the claim. ■

5.5 Characterisation of Monotone Behaviour Rules

In this Section we characterise all monotone behaviour rules. First, we provide an example. Then we show that all monotone behaviour rules share certain features of this example.

Example 1 *(Cross [15].) For all $\sigma \in \Delta(S)$, $s, s' \in S$ with $s \neq s'$, and $x \in (0, 1)$:*

$$B(\sigma, s, x)(s) = \sigma(s) + (1 - \sigma(s))x$$

$$B(\sigma, s', x)(s) = \sigma(s) - \sigma(s)x$$

Notice that this behaviour rule has the somewhat counterintuitive feature that the decision maker *always* increases the probability of the action which he actually played, even if the payoff was very low.

It is obvious that Cross' rule satisfies condition (1) in Definition 14. The expected movement for Cross' behaviour rule is given by:

$$f(\sigma, s) = \sigma(s)[\pi_s - \sum_{s' \in S}(\sigma(s')\pi_{s'})]$$

for all $\sigma \in \Delta(S)$ and all $s \in S$. Notice that the expression on the right hand side appears also on the right hand side of the replicator equation of evolutionary game theory. It is clear from this expression that Cross' rule has the property required by part (2) of Definition 14.

The next result shows that a behaviour rule is monotone if and only if the decision maker first submits her payoff to a linear and increasing transformation (where the coefficients may depend on the current state of the decision maker, on the strategy which she has played, and on the strategy the probability of which she is adjusting), and then applies Cross' rule:

Proposition 7 *A mixed strategy behaviour rule B is monotone if and only if there are functions $\mathcal{A} : \dot{\Delta}(S) \times S \times S \to \mathbb{R}$ and $\mathcal{B} : \dot{\Delta}(S) \times S \times S \to \mathbb{R}_{>0}$ such that for every $(\sigma, s, x) \in \dot{\Delta}(S) \times S \times (0,1)$:*

(1) $B(\sigma, s, x)(s) = \sigma(s) + (1 - \sigma(s))\Big(\mathcal{A}(\sigma, s, s) + \mathcal{B}(\sigma, s, s)x\Big)$

(2) $B(\sigma, s', x)(s) = \sigma(s) - \sigma(s)\Big(\mathcal{A}(\sigma, s', s) + \mathcal{B}(\sigma, s', s)x\Big)$ for all $s' \neq s$

and, for every $\sigma \in \dot{\Delta}(S)$ and $s \in S$:

(3) $\mathcal{A}(\sigma, s, s) = \sum_{s' \in S} \sigma(s')\mathcal{A}(\sigma, s', s)$

(4) $\mathcal{B}(\sigma, s, s) = \sum_{s' \in S} \sigma(s')\mathcal{B}(\sigma, s', s)$

Conditions (1) and (2) in Proposition 7 show that all monotone rules are like Cross' rule with linearly transformed payoffs. Conditions (3) and (4) place an unbiasedness condition on the coefficients of the linear payoff transformation. They say that the coefficients have to be such that expected motion is zero whenever all strategies yield the same expected payoff. The proof of Proposition 7 shows why this condition is a necessary condition for monotonicity, and why (3) and (4) ensure that it is met.

The following remark shows that the formula for the expected movement of monotone behaviour rules has the same structure as the formula for the expected movement of Cross' rule. Of course, allowance must be made for the fact that payoffs may be submitted to linear increasing transformations. But, once this is taken care of, expected movement of any monotone behaviour rule is the same as the movement of evolutionary replicator dynamics. By Proposition 6 this implies in particular that slow moving monotone behaviour rules will stay with high probability close to the deterministic trajectory of replicator dynamics.

Remark 1 *Let B be a monotone behaviour rule, and let E be an environment. Then for every $\sigma \in \dot{\Delta}(S)$ and every $s \in S$ the expected movement of the probability of s is given by:*

$$f(\sigma, s) = \sigma(s)[\mathcal{B}(\sigma, s, s)\pi_s - \sum_{s' \in S}(\sigma(s')\mathcal{B}(\sigma, s', s)\pi_{s'})]$$

We conclude this Section with two further examples of monotone behaviour rules. The proofs of monotonicity for these two behaviour rules are straightforward and therefore omitted.

Example 2 *Let any α with $0 \leq \alpha \leq 1$ be given. Using the notation of Proposition 7 we can then define a monotone behaviour rule by setting for all $\sigma \in \Delta(S), s, s' \in S$ with $s \neq s'$:*

$$\mathcal{A}(\sigma, s, s) = -\sigma(s) \sum_{s' \neq s}[(\sigma(s'))^2(1 - \sigma(s'))]\alpha$$

$$\mathcal{B}(\sigma, s, s) = +\sigma(s) \sum_{s' \neq s}[(\sigma(s'))^2(1 - \sigma(s'))]$$

$$\mathcal{A}(\sigma, s', s) = -(1 - \sigma(s'))(1 - \sigma(s))\sigma(s')\sigma(s)\alpha$$

$$\mathcal{B}(\sigma, s', s) = +(1 - \sigma(s'))(1 - \sigma(s))\sigma(s')\sigma(s)$$

According to this behaviour rule, if strategy s was played in iteration n, the decision maker increases (resp. decreases) in period $n + 1$ the probability assigned to s if the payoff x which the decision maker received in iteration n was above (resp. below) α. Intuitively, α thus plays the role of an "aspiration level." If the probability assigned

to s is increased (resp. decreased), the probability of all other strategies is decreased (resp. increased).

Remark 2 *An alternative model of reinforcement learning with an aspiration level was investigated in Börgers and Sarin [6]. The model of that paper postulates that a payoff which is an amount of x below the aspiration level leads to the probability of the action just played to be multiplied by x and all other probabilities to be increased proportionally. If the payoff is above the aspiration level, then the Cross rule is applied.*

This model appears plausible, however, it fails to satisfy conditions (3) and (4) of Proposition 7. These conditions ensure that expected movement in probabilities is zero when all actions have identical expected payoffs. If the above rule is applied, and some particular action has probability close to one, and has a positive probability of receiving negative payoff, the expected change in this actions probability is negative, independent of the expected payoff of all other actions.

Example 3 *Suppose that $S = \{1, 2, ..., \#S\}$. For any given strategy $s \in S$ we define two strategies $s \oplus 1$ and $s \ominus 1$ both of which are also contained in S. In general, $s \oplus 1 \equiv s + 1$ and $s \ominus 1 = s - 1$. But there are two cases in which this is not well-defined, and in these cases we set:$\#S \oplus 1 \equiv 1$, and $1 \ominus 1 \equiv \#S$. Using the notation of Proposition 7 we can then define a monotone behaviour rule by setting for all $\sigma \in \Delta(S)$ and $s \in S$:*

$$\mathcal{A}(\sigma, s, s \oplus 1) = \prod_{s' \neq s, s \oplus 1} \sigma(s')$$

$$\mathcal{B}(\sigma, s, s \oplus 1) = 1 - \mathcal{A}(\sigma, s, s \oplus 1)$$

$$\mathcal{A}(\sigma, s, s \ominus 1) = -\prod_{s' \neq s, s \ominus 1} \sigma(s')$$

$$\mathcal{B}(\sigma, s, s \ominus 1) = 1 + \mathcal{A}(\sigma, s, s \ominus 1)$$

and, for $s' \neq s \ominus 1, s \oplus 1$:

$$\mathcal{A}(\sigma, s, s') = 0$$

$$\mathcal{B}(\sigma, s, s') = 1$$

Hence, if strategy s is played in iteration n, the probability of this action is increased in the same way as it is increased in Cross' behaviour rule. The probability of strategy $s \ominus 1$, however, is reduced by more than it would be the case in Cross' rule, and the probability of strategy $s \oplus 1$ is reduced by less than in Cross' rule. In fact, if the payoff x was very low, the probability of $s \oplus 1$ may be increased rather than decreased. The probability of actions $s' \notin \{s \ominus 1, s, s \oplus 1\}$ is reduced in the same way as in Cross' rule. A behaviour rule of this type might capture the intuitive idea that strategy $s \oplus 1$ is "the opposite" of s, whereas strategy $s \ominus 1$ is "similar" to s. Notice, however, that these relations are not transitive in the example.

5.6 Proof of Proposition 7

To see that every behaviour rule which has the properties listed in Proposition 7 is monotone note first that condition (1) in the definition of monotonicity is trivially satisfied. Moreover, as noted in Remark 1, for every $\sigma \in \dot{\Delta}(S)$ and every $s \in S$ the expected movement of the probability of s is given by:

$$f(\sigma, s) = \sigma(s)[\mathcal{B}(\sigma, s, s)\pi_s - \sum_{s' \in S}(\sigma(s')\mathcal{B}(\sigma, s', s)\pi_{s'})]$$

Using condition (4) in Proposition 7 we can re-write this as:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s}(\sigma(s')\mathcal{B}(\sigma, s', s)(\pi_s - \pi_{s'}))$$

If $\pi_s \geq \pi_{s'}$ for all $s' \neq s$, this is non-negative, and if the inequality is strict for some s' then this expression is positive. This implies that B is monotone.

In the remainder of the proof we consider some given monotone behaviour rule B, and we show that B has to have the properties listed in Proposition 7. We proceed in five steps.

Step 1: Consider an environment in which $S^* = S$. Then for every $\sigma \in \dot{\Delta}(S)$ and every $s \in S$: $f(\sigma, s) = 0$.

Proof: Suppose there were an environment with $S^* = S$, a $\sigma \in \dot{\Delta}(S)$, and an $s \in S$ such that $f(\sigma, s) \neq 0$. Then there has to be some $s \in S$ such that $f(\sigma, s) < 0$. Now suppose that we change payoffs slightly, so that s becomes the *unique* expected

payoff maximising action. Because of the continuity of the behaviour rule, the expected movement in the probability of s will remain negative, contradicting monotonicity.

Step 2: There are functions $\tilde{A} : \Delta(S) \times S \times S \to \mathbb{R}$ and $\tilde{B} : \Delta(S) \times S \times S \to \mathbb{R}$ such that for every $(\sigma, s, x) \in \dot{\Delta}(S) \times S \times (0,1)$

(1) $B(\sigma, s, x)(s) = \tilde{A}(\sigma, s, s) + \tilde{B}(\sigma, s, s)x$

(2) $B(\sigma, s, x)(s') = \tilde{A}(\sigma, s, s') - \tilde{B}(\sigma, s, s')x$

Proof: Let $a, b, c \in (0,1)$, and suppose $a < b < c$. Let $\hat{s} \in S$, and consider two environments, E and $\tilde{E}$, both of which have the property that all strategies have the same expected payoff, i.e. in both environments $S^* = S$. Suppose also that the payoff distributions of any strategy $s \in S$ with $s \neq \hat{s}$ is the same in E and in $\tilde{E}$. Finally, suppose that in environment E strategy $\hat{s}$ yields payoff a with probability p and payoff c with probability $1 - p$, whereas in environment $\tilde{E}$ strategy $\hat{s}$ yields payoff b with certainty. Here, p is given by: $p = \frac{c-b}{c-a}$. This ensures that the expected payoff of $\hat{s}$ is the same in the two environments.

Denote by $f(\sigma, s)$ (resp. $\tilde{f}(\sigma, s)$) the expected change in the probability of any strategy $s \in S$ in the environment E (resp. $\tilde{E}$). Step 1 implies for all $s \in S$:

$$f(\sigma, s) = \tilde{f}(\sigma, s) = 0 \Rightarrow$$

$$f(\sigma, s) - \tilde{f}(\sigma, s) = 0 \Leftrightarrow$$

$$pB(\sigma, \hat{s}, a)(s) + (1-p)B(\sigma, \hat{s}, c)(s) = B(\sigma, \hat{s}, b)(s)$$

Replacing p by $\frac{c-b}{c-a}$ and re-arranging yields:

$$(c-a)(B(\sigma, \hat{s}, b)(s) - B(\sigma, \hat{s}, a)(s)) = (b-a)(B(\sigma, \hat{s}, c)(s) - B(\sigma, \hat{s}, a)(s))$$

This implies that either

$$B(\sigma, \hat{s}, a)(s) = B(\sigma, \hat{s}, b)(s) = B(\sigma, \hat{s}, c)(s)$$

or

$$\frac{B(\sigma, \hat{s}, b)(s) - B(\sigma, \hat{s}, a)(s)}{B(\sigma, \hat{s}, c)(s) - B(\sigma, \hat{s}, a)(s)} = \frac{b-a}{c-a}$$

As this must be true for all a, b, c with $a < b < c$, it follows that $B(\sigma, \hat{s}, x)(s)$ must be linear in x, as asserted.

Step 3: For every $\sigma \in \dot{\Delta}(S)$ and all $s, s' \in S$:

$$\tilde{B}(\sigma, s, s') > 0$$

Proof: Consider any $\sigma \in \dot{\Delta}(S)$. We prove the claim first for the case $s \neq s'$. The proof is indirect. Suppose there were $s, s' \in S$ with $s \neq s'$ such that $\tilde{B}(\sigma, s, s') < 0$. Consider an environment E such that $\mu_{s'}(x) = 1, \mu_s(x - \delta) = 1$ and $\mu_{s''}(x - \varepsilon) = 1$ for all $s'' \neq s, s'$. Suppose $\delta, \varepsilon > 0$. Then:

$$\begin{aligned}
f(\sigma, s') &= \sigma(s')(\tilde{A}(\sigma, s', s') + \tilde{B}(\sigma, s', s')x) \\
&\quad +\sigma(s)(\tilde{A}(\sigma, s, s') - \tilde{B}(\sigma, s, s')(x - \delta)) \\
&\quad + \sum_{s'' \neq s, s'} \sigma(s'')(\tilde{A}(\sigma, s'', s') - \tilde{B}(\sigma, s'', s')(x - \varepsilon)) \\
&\quad -\sigma(s') \\
&= \sigma(s')(\tilde{A}(\sigma, s', s') + \tilde{B}(\sigma, s', s')x) \\
&\quad + \sum_{\hat{s} \neq s} \sigma(\hat{s})(\tilde{A}(\sigma, \hat{s}, s') - \tilde{B}(\sigma, \hat{s}, s')x) \\
&\quad -\sigma(s') \\
&\quad +\sigma(s)B(\sigma, s, s')\delta + \sum_{s'' \neq s, s'} \sigma(s'')B(\sigma, s'', s')\varepsilon
\end{aligned}$$

By Step 1, the first three lines of this sum add up to zero. Therefore:

$$f(\sigma, s') = \sigma(s)B(\sigma, s, s')\delta + \sum_{s'' \neq s, s'} \sigma(s'')B(\sigma, s'', s')\varepsilon$$

If $B(\sigma, s, s') < 0$, then this term becomes negative for sufficiently small ε, contradicting monotonicity.

It remains to prove the claim for the case $s = s'$. Since for every $(\sigma, s, x) \in \dot{\Delta}(S) \times S \times (0, 1)$ $B(\sigma, s, x)$ is a probability vector, we must have:

$$\sum_{s' \in S} B(\sigma, s, x)(s') = 1 \Leftrightarrow$$

$$\tilde{A}(\sigma,s,s)+\tilde{B}(\sigma,s,s)x+\sum_{s'\neq s}\tilde{A}(\sigma,s,s')-\sum_{s'\in S}\tilde{B}(\sigma,s,s')x=1$$

This can be true for every $x\in(0,1)$ only if:

$$\tilde{B}(\sigma,s,s)-\sum_{s'\neq s}\tilde{B}(\sigma,s,s')=0$$

$$\tilde{B}(\sigma,s,s)=\sum_{s'\neq s}\tilde{B}(\sigma,s,s')$$

We have already shown that all the terms on the right hand side are positive. Therefore, the left hand side has to be positive as well.

Step 4: For every $\sigma\in\dot{\Delta}(S)$ and $s\in S$:

(1) $\sum_{s'\in S}\sigma(s')\tilde{A}(\sigma,s',s)=\sigma(s)$

(2) $\sum_{s'\neq s}\sigma(s')\tilde{B}(\sigma,s',s)=\sigma(s)\tilde{B}(\sigma,s,s)$

Proof: Consider an environment E such that all actions give the same, deterministic payoff, i.e. for some x: $\mu_s(x)=1$ for all $s\in S$. By Step 1, for every $s\in S$ and $\sigma\in\dot{\Delta}(S)$, $f(\sigma,s)=0$. By Step 2:

$$f(\sigma,s)=\sum_{s'\in S}\sigma(s')\tilde{A}(\sigma,s',s)-\sigma(s)+x[\sigma(s)\tilde{B}(\sigma,s,s)-\sum_{s'\neq s}\sigma(s')\tilde{B}(\sigma,s',s)]=0$$

This can be true for all x only if Step 4 is true.

Step 5: To complete the proof we define the functions $\mathcal{A}$ and $\mathcal{B}$ by setting for all $\sigma\in\dot{\Delta}(S)$ and $s,s'\in S$ with $s\neq s'$:

$$\mathcal{A}(\sigma,s,s)=\frac{\tilde{A}(\sigma,s,s)-\sigma(s)}{1-\sigma(s)}$$

$$\mathcal{A}(\sigma,s',s)=\frac{\sigma(s)-\tilde{A}(\sigma,s',s)}{\sigma(s)}$$

$$\mathcal{B}(\sigma,s,s)=\frac{\tilde{B}(\sigma,s,s)}{1-\sigma(s)}$$

$$\mathcal{B}(\sigma,s',s)=\frac{\tilde{B}(\sigma,s',s)}{\sigma(s)}$$

By Step 3 $\mathcal{B}(\sigma,s,s)>0$ and $\mathcal{B}(\sigma,s,s')>0$. Step 2 implies that with these definitions

$\mathcal{A}$ and $\mathcal{B}$ satisfy conditions (1) and (2) of Proposition 7. Finally, Step 4 implies that $\mathcal{A}$ and $\mathcal{B}$ also satisfy conditions (3) and (4) of Proposition 7. ■

5.7 Conclusion

This Chapter has investigated the existence of mixed strategy reinforcement learning rules which are approximately rational strategies for multi-armed bandit problems. Although the goal was the characterisation of all learning rules that are approximately maximising, i.e. rules that lead the decision-maker to play in the long run the expected payoff maximising choice with probability arbitrarily close to one, the complexity of the problem has prevented us from fulfilling it. However, we have defined a property called monotonicity and shown that monotone behaviour rules are approximately maximising. Monotonicity means that the expected change in the probability of the expected payoff maximising action is always positive for every state of the decision-maker and every decision problem. We have characterised all monotone behaviour rules, showing that a rule is monotone if and only if it is Cross' rule, whereby payoffs may be subjected to certain linear transformation. Furthermore we have shown that there is a close link between monotonicity and replicator dynamics, as monotone rules keep track of the replicator equation as long as the learning is slow.

The continuous differentiability assumption has played an important role as it has allowed us to approximate slow moving stochastic processes by the solution of deterministic differential equations. Beside this technical use, it has restricted the extent to which the decision maker can use her state space as a device for encoding information about her past payoff experiences. The decision maker could achieve approximate maximisation by keeping a record of her payoff experiences by manipulating appropriately those digits in the decimal expansion of her strategy which are far behind the decimal point. The earlier digits, which are more relevant to the actual behaviour, could then be used to induce initially an appropriate amount of experimentation, and later a choice which maximises the observed average payoff. But this sort of rules would not been continuous and therefore have been rule out. A formal analysis of this sort of construction seems interesting but would have taken us far from the main purpose of this thesis, i.e. the use of *simple* learning rules as bounded rationality strategies for

multi-armed bandit problems.

Another important issue to be investigated in future research is the relationship between monotonicity and absolute expediency. The latter property seems more appealing as it is concerned with expected payoffs. However the results in the literature about absolutely expedient rules are less general than our results on monotone rules. Moreover, the examples of monotone rules provided in Section 5 happen to be absolutely expedient rules. We have been unable to draw the exact relationship between these two properties.

There is a large and diverse set of references which address issues related to those covered in this Chapter. There are several related results in the economics literature. The first of these is in Samuelson and Zhang [48], who show that a *selection dynamic* satisfies a condition called *aggregate monotonicity* if and only if it is a multiple of the replicator dynamics. Selection dynamics describe the evolution of a population of players. Samuelson and Zhang do not trace back their result to behaviour rules for individual players. Unlike *monotonicity* as defined in this Chapter, *aggregate monotonicity* concerns not just the frequency of the best action, but also of other actions. Samuelson and Zhang obtain their result considering just one single environment. By contrast, it is essential for our result that a behaviour rule must operate in multiple environments.

More closely related to our work are two recent papers by Rustichini [47] and Easley and Rustichini [16]. Both papers, like this Chapter, consider a repeated individual decision under risk, and then axiomatize behaviour rules which move like replicator dynamics. Easley and Rustichini's [16] work differs, however, in two important ways from our study. Firstly, they take only the decision maker's ordinal ranking of outcomes, but not his von Neumann Morgenstern payoff as exogenously given. Moreover, they have a different informational assumption: In each period, the decision maker observes the outcome not only for the action which she took, but for all actions. Easley and Rustichini then axiomatize behaviour rules which move as if the decision maker had a given and fixed von Neumann Morgenstern utility function, and adjusted her choice probabilities in the style of replicator dynamics.

Our analysis differs from that of Easley and Rustichini in that it operates with only one axiom, monotonicity, and correspondingly obtains a more general functional form for the learning process.

Rustichini [47] covers both the informational assumption which we made in this Chapter, and the informational assumption which is made in Easley and Rustichini [16]. Rustichini introduces axioms concerning the expected motion of a learning process which then yield a replicator type process. He shows how stochastic approximation results might lead to a long run identity of expected and actual motion. However, he does not translate his results concerning expected motion into results concerning the individual's behaviour rule. Moreover, he invokes a number of axioms such as symmetry, and linearity in payoffs, which have no analog in this Chapter.

The literature which is most closely related to the analysis in Sections 4 and 5 is a branch of the machine learning literature which is concerned with the learning behaviour of stochastic automata. The concept of a stochastic automaton is similar to our concept of a mixed strategy behaviour rule. A useful overview of the literature on stochastic automata and learning has been provided by Narendra and Thathachar [40].

Particularly closely related to our work is Toyama and Kimura [55]. They, too, investigate monotone behaviour rules. They show that monotonicity implies approximate maximisation if the process moves in small increments (their Corollary 1), and they provide a characterization of monotone behaviour rules (their Theorem 1). They allow a larger class of environments than we do, by allowing payoffs which are not i.i.d., but they restrict attention to a smaller class of algorithms, by assuming linearity in payoffs. They do not consider pure strategy rules. They pay less attention than we do to technical assumptions some of which, like differentiability, we think to be of conceptual importance as well.

Other papers in this literature have investigated a property which has some similarity to monotonicity, and which is called *absolute expediency.* This property was originally defined by Lakshmivarahan and Thathachar [34]. Absolute expediency requires that in all environments, conditional on all possible states of the automaton, the expected payoff in the next period is larger than it is in the current period. This property thus differs from the monotonicity property investigated in this Chapter in that it refers to payoffs rather than the probability of the expected payoff maximising actions.

The machine learning literature has found a number of characterizations of absolute expediency if additional restrictions are imposed on either the set of possible environ-

ments or the set of possible behaviour rules. Under these restrictions it has also found that absolute expediency implies approximate maximisation[5] if the process moves in small increments. A selection of relevant papers is: Aso and Kimura [2], Meybodi and Lakshmivarahan [36], and Lakshmivarahan and Thathachar [35]. However, to our knowledge, no result has been proved in this literature that would hold at the same level of generality as our result in this Chapter.

Absolute expediency has also been investigated by Sarin [49] and Schlag [50]. Sarin combines absolute expediency with other axioms and obtains a learning process which moves like replicator dynamics. Schlag obtains a similar result assuming linearity of the learning rule in probabilities.

[5] In this literature, the approximate maximisation property is referred to as *ε-optimality*.

Chapter 6

Pure Strategy Imitation Rules

6.1 Introduction

Imitation is one of the mechanisms by which patterns of behaviour spread throughout a population. Imitation processes are similar to evolutionary processes, and the analysis of imitation helps to link evolutionary theory to economic contexts. In this Chapter we develop a model of reinforcement learning by imitation. To this end, besides the individual decision maker, we introduce a population of agents who find themselves in the same situation, i.e. they all are facing an identical decision problem. In each iteration, the decision maker has to choose from a finite set of strategies, each of which yields an uncertain payoff. After playing a strategy and receiving a payoff, the decision maker has the opportunity to observe the action and the payoff of a member of a population. Then she revises her strategy.

In this Chapter we shall focus on the case that the decision maker is endowed with a pure strategy learning rule. The strategy revision is allowed to be stochastic, but all the information which is carried into the next period is the new *pure* strategy. Furthermore, to capture the essence of imitation, the decision maker's next period new pure strategy is restricted to be either the own action or the sampled one. Pure strategy learning rules which satisfy this condition will be called *pure strategy imitation rules.*

We shall investigate whether pure strategy imitation rules are able to generate optimality in the long run. Quite obviously, the ability of imitation rules to lead the decision maker towards optimal actions will depend on the evolution of the population behaviour. Therefore, there are many ways in which we can try to assess whether a

given imitation rule is "good". One possibility is to investigate the performance of this imitation in an arbitrary environment, i.e. in an environment in which the population behaviour is given and fixed. Another possibility is to investigate the performance of this imitation rule when all other agents use this rule. In this Chapter we shall investigate the first possibility. The second possibility will be investigated in Chapter 8. The case considered in this Chapter is our attempt at studying an environment in which the decision maker does not rely on the evolution of all other agents towards optimal actions.

Similarly as in Chapter 4, we are lead to investigate which pure strategy imitation rules are *approximately maximising*, i.e. lead the decision maker to play in the long run the expected payoff maximising strategy with probability arbitrarily close to one, independent of what the true payoff distribution is and regardless of the given and fixed population behaviour. We show that no approximately maximising imitation rule exists. This negative result parallels that of Chapter 4. Note that in comparison to that situation, in this framework the decision maker has more knowledge of the environment - because she observes other agents - but is constrained in the way she can adjust his behaviour - because she is restricted to either stick to her own action or to imitate the other agent's action. The intuition for our result in this Chapter is similar to the intuition for the analogous result in Chapter 4: a pure strategy state space is not rich enough to store all the relevant information to achieve approximate payoff maximisation.

The rest of the chapter is as follows. Section 2 states the formal framework and Section 3 contains the definitions. Section 4 studies approximate payoff maximisation and section 5 concludes.

6.2 Formal Framework

This Chapter refers to the case in which the set W^A is a singleton, i.e. $W^A = \{w\}$ and in which sampling *is* allowed. Individual w is referred to as the decision maker.

Each member of the population W^D is programmed to play a pure strategy $s \in S$. Formally speaking, there is a function $C : W^D \to S$ which assigns a strategy to every member of W^D. Thus, individual $w' \in W^D$ is programmed to play pure strategy $C(w')$.

A sampling rule for the decision maker is a probability distribution $e \in \Delta(W^D)$, where $e(w')$ is the probability that the decision maker samples individual w'.[1]

6.3 Payoff Maximisation

A pure strategy learning rule is formally defined as follows:

Definition 16 *A pure strategy learning rule B is a function*[2] $B : S \times (0,1) \times S \times (0,1) \times S \to [0,1]$, *such that for all $s, s' \in S$, and $x, y \in (0,1)$: $\sum_{s'' \in S} B(s, x, s', y, s'') = 1$*

The interpretation of a pure strategy learning rule is as follows: $B(s, x, s', y, s'')$ is the probability of choosing strategy s'' at iteration $n+1$ after choosing strategy s at iteration n, getting a payoff x and sampling an individual choosing strategy s' and receiving a payoff y.

Throughout this Chapter we focus on learning rules which satisfy the following assumption.

Assumption 2: $B(s, x, s', y, s'') = 0$ for all $s'' \neq s, s'$ and all $x, y \in (0,1)$.

Assumption 2 is intended to capture the essence of imitation learning. The learning rules which satisfy this assumption are therefore called *imitation rules.*

The function C and the decision maker's sampling rule e induce a probability distribution $y(\cdot)$ over S, where $y(s)$ is the probability that the decision-maker samples strategy $s \in S$. It is formally given by

$$y(s) = \sum_{w' \in W^D} e(w') I(C(w'), s)$$

where $I(C(w'), s)$ is an indicator function which takes the value 1 if $C(w') = s$.

Definition 17 *The* transition matrix *T corresponding to an environment E, a function C, a sampling rule e and a pure strategy imitation rule B is a $\#S \times \#S$ matrix whereby the entry in the row corresponding to strategy s, and the column corresponding*

[1] We do not specify the sampling rules for $w' \in W^D$ because they are not allowed to revise their behaviour and therefore their sampling rules are irrelevant.

[2] Note that we use, for simplicity, the same symbol B as in previous Chapters to denote pure strategy imitation rules.

to strategy s', is:[3]

$$t_{s,s'} = \int_0^1 \int_0^1 y(s')B(s,x,s',y,s')dF(\mu_s,\mu_{s'}).$$

Definition 18 *The* imitation process *corresponding to an environment E, a function C, a sampling rule e, an imitation rule B, and an initial probability distribution $\delta_0 \in \Delta(S)$ is the Markov chain $\{s_n\}_{n\in\mathbb{N}_0}$ with the initial distribution δ_0 and with the transition matrix T defined in previous definition. For every $n \in \mathbb{N}_0$ we denote by $\delta_n \in \Delta(S)$ the marginal distribution of s_n.*

We now formally define a first property of pure strategy imitation rules. It refers to a rule which leads the decision maker to learn the expected payoff maximising strategy.

Definition 19 *A pure strategy imitation rule B is* maximising *if for every environment E, every function C, every sampling rule e and every initial distribution $\delta_0 \in \Delta(S)$ such that $\delta_0(s^*) > 0$ for at least one $s^* \in S^*$ and $y(s^*) > 0$ for at least one $s^* \in S^*$ the following is true:*

$$\lim_{n\to\infty} \delta_n(S^*) = 1.$$

Note the difference between this definition and definition 7. This definition is weaker that definition 7, in the sense that it only requires long run optimality when the probability of sampling an expected payoff maximising strategy is strictly positive.

Despite this qualification, the next proposition shows the non existence of maximising pure strategy imitation rules.

Proposition 8 *No pure strategy imitation rule is maximising.*

This result is almost obvious. The sketch of the proof is as follows. Note that an imitation rule with $B(s,x,s',y,s') > 0$ for some $s,s' \in S$ and $x,y \in (0,1)$ cannot be maximising. Take some function C, some sampling rule e, and some environment E such that $\mu_s(x) > 0$ and $\mu_{s'}(y) > 0$, $S^* = \{s\}$ and $y(s') > 0$. Then $B(s,x,s',y,s') > 0$ implies that the decision maker would switch away from the expected payoff maximising action with strictly positive probability. Therefore, the asymptotic probability of

[3]This is a slight abuse of notation. It is writen in this way to encompass common as well as independent events condition.

playing the expected payoff maximising strategy would be less than one. Thus maximisation requires $B(s,x,s',y,s')=0$ for all $s,s' \in S$ and $x,y \in (0,1)$. But clearly that is not maximising either.

Proposition 8 leads us to seek imitation rules which are *approximately maximising.* Roughly speaking, we mean by this imitation rules for which $\lim_{n\to\infty} \delta_n(S^*)$ is *close* to 1 for all environments E, all functions C, all sampling rules e and all initial distributions δ_0 which satisfy the conditions of Proposition 8. More specifically, we shall ask whether there is a family of imitation rules, parametrized by some parameter ν, such that in all environments the asymptotic probability of choosing an optimal action converges to 1 as ν tends to infinity, provided that the conditions of Proposition 8 hold. If this is true, then, by choosing a sufficiently large ν, the probability of choosing an optimal action can be made arbitrarily close to 1.

6.4 Approximate Payoff Maximisation

In this Section, we look for imitation rules which generate approximately rational behaviour in the long run. It is shown that no pure strategy imitation rule is approximately maximising.

Definition 20 *A sequence of pure strategy imitation rules* $\{B^\nu\}_{\nu\in\mathbb{N}}$ *is* approximately maximising *if for every environment E, every function C and every sampling rule $e(\cdot)$ such that $y(s^*)>0$ for at least one $s^* \in S^*$, every initial distribution $\delta_0 \in \Delta(S)$ such that $\delta_0(s^*)>0$ for at least one $s^* \in S^*$, and every $\nu \in \mathbb{N}$ the limit $\lim_{n\to\infty} \delta_n^\nu(S^*)$ (where for every $n \in \mathbb{N}_0$ δ_n^ν is the marginal distribution of s_n if the initial distribution is δ_0 and if the imitation rule is B^ν) exists, and we have:*

$$\lim_{\nu\to\infty} \lim_{n\to\infty} \delta_n^\nu(S^*) = 1.$$

Proposition 9 *No sequence of pure strategy imitation rules is approximately maximising.*

Proof of Proposition 9. The proof is indirect. Suppose that $\{B^\upsilon\}_{\upsilon\in N}$ is a sequence of pure strategy imitation rules which is approximately maximising.

Step 1: For all $s, s' \in S$ and $x, y \in (0,1)$ with $x < y$, there is some $\overline{\upsilon} \in N$ such that for all $\upsilon \geq \overline{\upsilon}$:

$$B^{\upsilon}\left(s, x, s', y, s'\right) > 0.$$

Remark: Step 1 says that for υ large enough, the imitation rule has to imitate better payoff strategies.

Proof: The proof is indirect. Suppose for every $\overline{\upsilon} \in N$ there were some $\upsilon \geq \overline{\upsilon}$ such that $B^{\upsilon}\left(s, x, s', y, s'\right) = 0$. Consider an environment E such that $\mu_s(x) = 1$ and, for some $y > x$, $\mu_{s'}(y) = 1$ for all $s' \neq s$. Consider some function C and some sampling rule e such that $y(s'') = 0$ for all $s'' \neq s, s'$. If the decision maker begins with an initial distribution δ_0 which attaches positive probability to the strategy s, and adopts an imitation rule B^{υ} such that $B^{\upsilon}\left(s, x, s', y, s'\right) = 0$, then for all $n \in N$: $\delta_n^{\nu}(S \setminus \{s\}) \leq 1 - \delta_0$. Hence if $\delta_n^{\nu}(S \setminus \{s\})$ converges for $n \to \infty$, its limit must be less than $1 - \delta_0 < 1$. We have thus obtained a contradiction to the approximate maximisation property of $\{B^{\upsilon}\}_{\upsilon \in N}$.

Step 2: For all $s, s' \in S$ and $x, y \in (0,1)$ with $x > y$, there is some $\overline{\upsilon} \in N$ such that for all $\upsilon \geq \overline{\upsilon}$:

$$B^{\upsilon}\left(s, x, s', y, s'\right) < 1.$$

Remark: Step 2 says that for υ large enough, the imitation rule can not imitate a lower payoff strategy with probability one.

Proof: The proof is indirect. Suppose for every $\overline{\upsilon} \in N$ there were some $\upsilon \geq \overline{\upsilon}$ such that $B^{\upsilon}\left(s, x, s', y, s'\right) = 1$. Consider an environment E such that $\mu_s(x) = 1$ and, for some $y < x$, $\mu_{s'}(y) = 1$ for all $s' \neq s$. Consider some function C and some sampling rule e such that $y(s) = y(s') = \frac{1}{2}$ and $y(s'') = 0$ for all $s'' \neq s, s'$. For every ν such that $B^{\nu}\left(s, x, s', y, s'\right) = 1$ we have: $\delta_n^{\nu}(s) \leq \frac{1}{2}\delta_{n-1}^{\nu}(s) + \frac{1}{2}\left(1 - \delta_{n-1}^{\nu}(s)\right)$ for all $n \in N$. Here, the first term on the right hand side corresponds to the case that the decision maker played s in period $n-1$, and sampled strategy s. The second term corresponds to the case that the decision maker played any $s' \neq s$ in period $n-1$, and sampled strategy s. Obviously, the right hand simplifies to $\frac{1}{2}$. Therefore, if $\delta_n^{\nu}(s)$ converges for $n \to \infty$, its limit can not be more than $\frac{1}{2}$. We have thus obtained a contradiction to the approximate maximisation property of $\{B^{\upsilon}\}_{\upsilon \in N}$.

Step 3: Suppose $x, \epsilon \in (0,1)$ with $1-\epsilon > x > 2\epsilon > 0$. Consider an environment E such that $1 > \mu_s(x+\epsilon) > 0$ and $1 > \mu_s(x-\epsilon) > 0$ for some $s \in S$, $\mu_{s'}(x) = 1$ for some $s' \in S$ with $s' \neq s$ and $\mu_{s''}(x-2\epsilon) = 1$ for $s'' \neq s, s'$. Consider some function C and some sampling rule e such that $y(s), y(s') > 0$ and $y(s'') = 0$ for $s'' \neq s, s'$. Then there is some $\overline{v} \in N$ such that for $v \geq \overline{v}$ the imitation process $\{s_n\}_{n \in N}$ is a Markov chain with the unique irreducible class $\{s, s'\}$. Moreover, this class is aperiodic.

Proof: Choose $\overline{v}$ such that the result of Step 1 is true for all strategies in S. We first show that $\{s, s'\}$ is an irreducible and aperiodic class and then we show that is the unique irreducible class.

(i) By step 1, states s and s' *intercommunicate*; moreover, the probability of moving away from the set $\{s, s'\}$ is zero because $y(s'') = 0$ for $s'' \neq s, s'$. This implies that $\{s, s'\}$ is an irreducible class. It is aperiodic because $0 < y(s), y(s') < 1$.

(ii) Note that by step 1, state s is *accessible* from any state $s'' \neq s, s'$, because $y(s') > 0$, i.e. the probability of moving from s'' to s in one step is positive. But state s'' is not accessible from $\{s, s'\}$. This implies that states $s'' \neq s, s'$ are *transient.*

Remark: By standard results in the theory of Markov chains, Step 3 implies that for environment E and $v \geq \overline{v}$, the investigation of the properties of the imitation process $\{s_n\}_{n \in N}$ reduces to the investigation of the Markov chain whose states are $\{s, s'\}$. Let T^{ν} denote the transition matrix of the latter chain. As this chain is irreducible and aperiodic, the sequence $\{\delta_{\infty}^{\nu}\}_{n \in N}$ converges and its limit is a stationary distribution of the transition matrix T^{ν} and is independent of the initial distribution δ_0. In the following, we shall denote this limit by δ_{∞}^{ν}.

Step 4: Suppose $x, y \in (0,1)$ with $x < y$. Consider an environment E such that $\mu_s(y) = 1$, $\mu_{s'}(x) = 1$ $\mu_{s''}(\frac{x}{2}) = 1$ for $s'' \neq s, s'$. Consider some function C and some sampling rule e such that $y(s), y(s') = \frac{1}{2}$ and $y(s'') = 0$ for $s'' \neq s, s'$. Then there exists some $\overline{v} \in N$ such that:

$$\lim_{\substack{v \to \infty \\ v \geq \overline{v}}} \frac{B^v(s, y, s', x, s')}{B^v(s', x, s, y, s)} = 0$$

Proof: The proof is indirect. Suppose it is not true. Choose $\overline{v}$ such that the result of Step 1 is true for all strategies in S. Then it follows that for finite ν

$B^v(s, y, s', x, s') > 0$. This implies, as in step 3, that the imitation process $\{s_n\}_{n \in N}$ is a Markov chain with the unique irreducible class $\{s, s'\}$. Moreover, this class is

aperiodic. Then it follows that the sequence $\{\delta_\infty^\nu\}_{n\in N}$ converges and its limit is a stationary distribution of the transition matrix T^ν and is independent of the initial distribution δ_0. Let δ_∞^ν denote this limit. Then it is true that

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \lim_{\substack{\upsilon\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{B^\upsilon(s,y,s',x,s')}{B^\upsilon(s',x,s,y,s)} > 0$$

and we have found a contradiction to the approximate maximisation property of $\{B^\upsilon\}_{\upsilon\in N}$.

Step 5: Consider an environment E as described in Step 3. Let $\mu_s(x+\epsilon)=\mu$. Let $y(s)=y(s')=\frac{1}{2}$. Then it is true that

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \frac{\mu}{1-\mu} \lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{B^\nu(s',x,s,x+\epsilon,s)}{B^\nu(s,x-\epsilon,s',x,s')}$$

Take $\overline{\upsilon}$ to be the maximum of the two $\overline{\upsilon}$s referred to in Steps 2 and 3. Because δ_∞^ν is a stationary distribution of T^ν, we have that $\delta_\infty^\nu(s)$ equals

$$\begin{aligned}
\delta_\infty^\nu(s) &= \delta_\infty^\nu(s)\left[\begin{array}{c} \frac{1}{2}1+ \\ \frac{1}{2}\left[\mu B^\nu(s,x+\epsilon,s',x,s)+(1-\mu)B^\nu(s,x-\epsilon,s',x,s)\right] \end{array}\right] + \\
&\quad \delta_\infty^\nu(s')\left[\begin{array}{c} \frac{1}{2}0 \\ +\frac{1}{2}\left[\mu B^\nu(s',x,s,x+\epsilon,s)+(1-\mu)B^\nu(s',x,s,x-\epsilon,s)\right] \end{array}\right] \\
&\Leftrightarrow \\
\frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} &= \frac{\mu B^\nu(s',x,s,x+\epsilon,s)+(1-\mu)B^\nu(s',x,s,x-\epsilon,s)}{\mu B^\nu(s,x+\epsilon,s',x,s)+(1-\mu)B^\nu(s,x-\epsilon,s',x,s')}
\end{aligned}$$

To simplify the exposition, we change notation

$$\begin{aligned}
A &= B^\nu(s',x,s,x+\epsilon,s) \\
B &= B^\nu(s',x,s,x-\epsilon,s) \\
C &= B^\nu(s,x+\epsilon,s',x,s) \\
D &= B^\nu(s,x-\epsilon,s',x,s')
\end{aligned}$$

Then the above expression is re-written

$$\frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \frac{\mu A+(1-\mu)B}{\mu C+(1-\mu)D}$$

By rearranging we get:

$$\frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \frac{1}{\frac{C}{A} + \frac{1-\mu}{\mu}\frac{D}{A}} + \frac{1}{\frac{\mu}{1-\mu}\frac{C}{B} + \frac{D}{B}}$$

Then it follows that

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \frac{1}{\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{C}{A} + \frac{1-\mu}{\mu}\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{D}{A}} + \\ + \frac{1}{\frac{\mu}{1-\mu}\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{C}{B} + \lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{D}{B}}$$

By step 4, we know that $\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{C}{A} = 0$ and $\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{B}{D} = 0$. Then

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \frac{\mu}{1-\mu}\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{A}{D} = \frac{\mu}{1-\mu}\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{B^\nu(s', x, s, x+\epsilon, s)}{B^\nu(s, x-\epsilon, s', x, s')}$$

as asserted.

Step 6: The sequence $\{B^\upsilon\}_{\upsilon\in N}$ is not approximately maximising.

Proof: The proof is divided in two parts.

(i) Consider the environment of step 5 with $0 < \mu < \frac{1}{2}$. For this environment, $S^* = \{s'\}$. Therefore it has to be that

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = 0$$

which implies

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{B^\nu(s', x, s, x+\epsilon, s)}{B^\nu(s, x-\epsilon, s', x, s')} = 0$$

(ii) Consider the environment of step 5 with $1 > \mu > \frac{1}{2}$. For this environment, $S^* = \{s\}$. But part (i) implies

$$\lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{\delta_\infty^\nu(s')} = \lim_{\substack{\nu\to\infty \\ \upsilon\geq\overline{\upsilon}}} \frac{\delta_\infty^\nu(s)}{1-\delta_\infty^\nu(s)} = 0$$

Hence $\lim_{\nu\geq\overline{\nu},\nu\to\infty} \delta_\infty^\nu(s) = 0$. This contradicts the approximate maximisation property of the sequence $\{B^\nu\}_{\nu\in\mathbb{N}}$. ■

Note that although proposition 2 and proposition 9 yield qualitatively similar results, the frameworks are quite different. In this Chapter, the decision maker is provided with *more* information about the environment -because she observes others agents- although she is constrained in the way she can adjust her behaviour -because she is assume to either stick or to switch to the observed action-. However, the intuition is similar to that of Proposition 2: the memory store provided by a pure strategy state space is too small to deal with a large variety of environments.

6.5 Conclusion

In this Chapter we have investigated the ability of pure strategy reinforcement imitation rules for leading the decision maker towards optimality without relying on the evolution of the population behaviour. It is shown that no pure strategy imitation rule can achieve approximately payoff maximisation. The intuition behind this result is that a pure strategy state space is too small to store all the relevant information about the environment.

This result helps to interpret the relationship between our analysis in this Chapter and a paper by Schlag [51]. In this article, Schlag considers social learning in a framework in which individuals can observe others' actions and payoffs, and thus their learning rules can contain an element of imitation. Schlag introduces a desirable property of learning rules which, roughly speaking, says that in all environments and all current states of the population the expected average payoff in the *population* is increasing from round to round. Schlag shows that the dynamics of a large, randomly matched population in which all individuals adopt a rule with this property can be approximated by the replicator dynamics. This means that this rule will lead the entire population to achieve rationality in the long because, as replicator dynamics maximises expected fitness, the population will choose in the long run with very high probability the action which maximises expected payoffs.

Thus the key for Schlag's result is the fact that the population behaviour is evolving over time. Whereas according to our Propositions 8 and 9 an isolated individual adopting a pure strategy imitation rule cannot achieve long run payoff maximisation for given and *fixed* population behaviour, a large group of individuals adjusting their

behaviour can. The intuition is that the group composition provides an additional state space in which information about the environment can be accumulated.

Recall that Chapter 4 showed that a pure strategy state space is too small to achieve (approximately) payoff maximisation when the decision maker adjusts her behaviour according to her *own* experience. In this Chapter, a similar analysis to Chapter 4 has been done although in a different framework as we have provided the decision maker with additional information about the environment by allowing her to observe other agents who find themselves in the same situation. As in Chapter 4, we have shown the non-existence of (approximately) maximising pure strategy *imitation* rules (as defined in Assumption 2). One might then conclude that a pure strategy state space is too small to deal with a variety of environments even if the decision maker can use others' experience to adjust her behaviour. We do not wish to draw this conclusion. It might be the case that the non-existence of maximising rules is due to the presence of Assumption 2. Further research should investigate the existence of (approximately) maximising behaviour rules in this framework but without imposing Assumption 2.

Chapter 7

Mixed Strategy Imitation Rules

7.1 Introduction

In this Chapter we continue exploring the approach we took in Chapter 6 to assess "how good" an imitation rule is, although now we shall assume that the decision maker adjusts her behaviour using a mixed strategy imitation rule instead of a pure strategy rule. In this approach, the decision maker has the opportunity to observe a member of a population of agents all of whom are facing the same decision problem. We evaluate the performance of an imitation rule by assuming that the behaviour of this population is exogenously given and fixed. We hence look for mixed strategy imitation rules which lead the decision maker to play optimally without relying on the evolution of the behaviour of a population .

The decision maker is characterized by a probability distribution which shows how likely it is that she chooses any of her actions. After choosing an action and receiving a stochastic payoff, she samples a member of a population who is facing a similar decision problem. After observing the action and the payoff received by the sampled member, the decision maker updates her probability distribution. The new distribution only depends on the previous distribution, on the actions taken and sampled and on the payoffs received and sampled. It does not depend on any other aspect of the history. Note that this sort of learning rules is a generalization of the mixed strategy behaviour rules considered in Chapter 5.

In the framework of mixed strategy rules, we shall use the following formalization of what it means to imitate: We shall assume that the decision maker only updates

the probabilities associated to the actions taken and sampled, all others probabilities are left unchanged. Learning rules which satisfy this condition will be called *mixed strategy imitation rules.* As in the previous Chapter, our goal is not to compare the performance of mixed strategy imitation rules versus mixed strategy behaviour rules, but to investigate if imitative learning is able to generate optimality in the long run. We shall follow a similar structure as in Chapter 5.

To this end, we define a property called *maximisation.* An imitation rule is maximising if the long run probability of the expected payoff maximising action is one independent of what the true distribution of payoffs is. We have not been able to settle the question of the existence of maximising imitation rules.

As in Chapter 5, we do not provide a complete characterisation of approximately maximising rules. We define a property called monotonicity and we show this property implies approximate maximisation. Monotonicity means that the expected change in the probability of the expected payoff maximising action is positive. We obtain a complete characterization of monotone rules for the common events condition and for the independent events condition. We show that both characterisations are the same.

The basic feature of monotone imitation rules is *proportional* imitation, meaning that the change in the probability attached to the taken action is proportional to the payoff difference. Note that this implies that the probability associated to the taken action is increased if its payoff is greater than the sampled one. This can be interpreted as a reinforcement effect in which the decision maker considers the sampled strategy's payoff as an aspiration level, i.e. if the own strategy's payoff is above this aspiration level, the currently played strategy is reinforced. This feature of the monotone imitation rules makes it possible to converge to the expected payoff maximising strategy even in the case that the optimal strategy is absent in the population.

The rest of the Chapter is as follows. Section 2 introduces the formal framework and Section 3 states the main definitions. Section 4 characterizes monotone imitation rules. Finally Section 5 concludes.

7.2 Formal Framework

This Chapter refers to the case in which the set W^A is a singleton, i.e. $W^A = \{w\}$ and in which sampling *is* allowed. Individual w is referred to as the decision maker.

Each agent of the population W^D is programmed to play a pure strategy $s \in S$. Formally speaking, there is a function $C : W^D \to S$ which assigns a strategy to every member of W^D. Thus, individual $w' \in W^D$ is programmed to play pure strategy $C(w')$.

A sampling rule for the decision maker is a probability distribution $e \in \Delta(W^D)$, where $e(w')$ is the probability that the decision maker samples agent w'.[1]

7.3 Definitions

We next define reinforcement rules with mixed strategy state space.

Definition 21 *A* mixed strategy learning rule B *is a function*

$$B : \Delta(S) \times S \times (0,1) \times S \times (0,1) \to \Delta(S)$$

The interpretation of a mixed strategy learning rule is as follows: At each iteration n the decision maker behaviour is described by a probability distribution $\sigma_n \in \Delta(S)$ which specifies for each pure strategy s how likely it is that she chooses strategy s at iteration n. We shall refer to σ_n as the state of the decision maker at iteration n. The distribution $B(\sigma_n, s, x, s', y)$ is then the state of the decision maker at iteration $n+1$ if her state at iteration n was σ_n, the pure strategy which she chose at iteration n was s, the payoff received was x, the sampled member's pure strategy was s' and the sampled member's payoff received was y. For every $s'' \in S$, we denote by $B(\sigma_n, s, x, s', y)(s'')$ the probability which $B(\sigma_n, s, x, s', y)$ assigns to s''.

Throughout this Chapter, we focus on mixed strategy learning rules which satisfy the following assumption.

Assumption 3. $B(\sigma_n, s, x, s', y)(s'') = \sigma_n(s'')$ for all $s'' \neq s, s'$ and all $x, y \in (0,1)$.

This assumption is intended to capture the essence of imitative learning. It states that the probabilities attached to non-observed strategies are not updated. The learning

[1] We do not specify the sampling rules for $w' \in W^D$ because they are not allowed to revise their behaviour and therefore their sampling rules are irrelevant.

rules satisfying assumption 3 are called *imitation rules.*

Throughout this Chapter we focus on imitation rules which satisfy the following assumption:

Assumption 4. For any $s \in S$ the mixed strategy imitation rule B is continuously differentiable in (σ_n, x, y) and the derivative of B with respect to (σ_n, x, y) is bounded from above and from below.

This is a technical assumption which will allow us to appeal to well-known theorems regarding the approximation of slow moving stochastic processes by the solution of deterministic differential equations.

The function C and the decision maker's sampling rule e induce a probability distribution $y(\cdot)$ over S, where $y(s)$ is the probability that the decision maker samples strategy $s \in S$. This probability distribution will be called *strategy sampling rule.* It is formally given by

$$y(s) = \sum_{w' \in W^D} e(w') I(C(w'), s)$$

where $I(C(w'), s)$ is an indicator function which takes the value 1 if $C(w') = s$.

We denote by $\mathbf{B}(\Delta(S))$ the set of all Borel subsets of $\Delta(S)$.

Definition 22 *The* stochastic kernel K *corresponding to an environment E, a function C, a sampling rule e and a mixed strategy imitation rule B is a function $K : \Delta(S) \times \mathcal{B}(\Delta(S)) \to [0,1]$ such that*

$$K(\sigma, \Omega) = \sum_{(s,x,s',y) \in \{s,s' \in S, x,y \in (0,1) | B(\sigma,s,x,s',y) \in \Omega\}} \sigma(s) \cdot y(s') \cdot \mu_s(x) \cdot \mu_{s'}(y)$$

for every $\sigma \in \Delta(S)$ and $\Omega \in \mathcal{B}(\Delta(S))$.

Intuitively, the *stochastic kernel* is the analog of a transition matrix for a Markov process with continuum size state space.

Definition 23 *The* imitation process *corresponding to an environment E, a function C, a sampling rule e, a mixed strategy imitation rule B and an initial state $\sigma_0 \in \Delta(S)$ is the Markov process $\{\sigma_n\}_{n \in N}$ with the initial distribution which assigns probability 1 to σ_0 and with the stochastic kernel $\mathbf{K}$ described in previous definition.*

Definition 24 *A mixed strategy imitation rule B is* maximising *if for every environment E, every function C, every sampling rule e and every initial state $\sigma_0 \in \dot{\Delta}(S)$ the probability of the event "$\sigma_n(S^*) \to 1$" is 1.*

We have not been able to settle the following question:

Open Question *Do mixed strategy imitation rules which are maximising exist?*

However, we do have interesting results concerning a class of *approximately maximising* imitation rules. Here, the concept of *approximate maximisation* is defined in the same way as in Chapter 5.

Definition 25 *A sequence of mixed strategy imitation rules $\{B^\nu\}_{\nu \in N}$ is* approximately maximising *if for every environment E, every function C, every sampling rule e and every initial state $\sigma_0 \in \dot{\Delta}(S)$ the probability of the event "$\sigma_n^\nu(S^x) \to 1$" converges to 1 as $\nu \to \infty$. Here, $\{\sigma_n^\nu\}_{n \in N}$ denotes the imitation process corresponding to the imitation rule B^ν and the initial state σ_0.*

The next section will explore a particular class of approximately maximising mixed strategy imitation rules.

7.4 Monotone Imitation Rules

For any mixed strategy imitation rule B, environment E, function C and sampling rule e, we define a function f which assigns to every possible state of the decision maker σ, and every pure strategy s, the expected change in the probability attached to s if the current state is σ. Formally, $f : \Delta(S) \times S \to R$ is defined by:

$$f(\sigma, s) = \sum_{s' \in S} \sigma(s') \sum_{s'' \in S} y(s'') \int_0^1 \int_0^1 \left[B(\sigma, s', x, s'', y)(s) - \sigma(s) \right] dF(\mu_{s'}, \mu_{s''})$$

for all $\sigma \in \Delta(s)$ and $s \in S$. For $\overline{S} \subseteq S$, we define

$$f(\sigma, \overline{S}) = \sum_{s \in \overline{S}} f(\sigma, s)$$

Finally, we denote by $f(\sigma)$ the vector

$$f(\sigma) = \{f(\sigma, s)\}_{s \in S}$$

Definition 26 *A mixed strategy imitation rule B is* monotone *if*

1. $\sigma \in \dot{\Delta}(S), s \in S, s' \in S$ and $x, y \in (0,1)$ imply $B(\sigma, s, x, s', y) \in \dot{\Delta}(S)$
2. for all environments E with $S^* \neq S$, all functions C, all sampling rules e and all states $\sigma \in \dot{\Delta}(S)$: $f(\sigma, S^*) > 0$

The second condition of the above definition is what motivates the label monotonicity. It says that the probability attached to the expected payoff maximising strategy increases in expected terms from round to round.

The following proposition fully characterizes monotone imitation rules under common events condition as well as independent events condition.

Proposition 10 *A mixed strategy imitation rule is* monotone *if and only if there exists a function $\widetilde{B} : \Delta(S) \times S \times S \to R_{>0}$ such that for every $(\sigma, s, x, s', y) \in \dot{\Delta}(S) \times S \times (0,1) \times S \times (0,1)$ with $s' \neq s$:*

1. $B(\sigma, s, x, s', y)(s) = \sigma(s) + \widetilde{B}(\sigma, s, s')(x - y)$

Proof of Proposition 10. We divide the proof in two parts. Part (i) deals with the common events condition and part (ii) deals with the independent events condition.
(i) For the Common Events Condition: To see that every imitation rule which has property (1) in proposition 10 is monotone note that condition (1) in the definition of monotonicity is trivially satisfied. Moreover, for every $\sigma \in \dot{\Delta}(S)$ and every $s \in S$, the expected movement of the probability of s is given by:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} y(s') \int_0^1 [B(\sigma, s, x, s', y)(s) - \sigma(s)] \, d\mu_s +$$
$$+ y(s) \sum_{s' \neq s} \sigma(s') \int_0^1 [B(\sigma, s', x, s, y)(s) - \sigma(s)] \, d\mu_{s'}$$

By using property (1) we can rewrite this as

$$f(\sigma, s) = \sum_{s' \neq s} \left(\pi^s - \pi^{s'}\right) \left[y(s)\sigma(s')\widetilde{B}(\sigma, s', s) + y(s')\sigma(s)\widetilde{B}(\sigma, s, s')\right]$$

If $\pi^s \geq \pi^{s'}$ for all $s' \neq s$, this expression is non-negative, and if the inequality is strict for some s', then it is positive. This implies that the imitation rule is monotone.

In the remainder of the proof, we consider some given monotone imitation rule and we show that the imitation rule has to have property (1) of proposition 10. We proceed in five steps.

Step 1: Consider an environment E in which $S^* = S$. Then, for every function C, every sampling rule e, every $\sigma \in \dot{\Delta}(S)$ and every $s \in S$: $f(\sigma, s) = 0$.

Proof: Suppose there were an environment E in which $S^* = S$, a function C, a sampling rule e, a $\sigma \in \dot{\Delta}(S)$ and a $s \in S$ such that $f(\sigma, s) \neq 0$. Then there has to be some $s \in S$ such that $f(\sigma, s) < 0$. Now suppose that we change payoffs slightly, so that $S^* = \{s\}$. Because of the continuity of the imitation rule, the expected movement in the probability of s will remain negative, contradicting monotonicity.

Step 2: $B(\sigma, s, x, s', x)(s) = \sigma(s)$ for all $x \in (0,1), \sigma \in \dot{\Delta}(S)$.

Proof: Consider environment E such that strategies s and s' yield payoff b with certainty. Consider some function C and sampling rule e. For this environment $S^* = S$. The expected movement of strategy s is given by:

$$\begin{aligned} f(\sigma, s) \;=\; & y(s)\left[\sigma(s)\left[B(\sigma, s, b, s, b)(s) - \sigma(s)\right] + \sigma(s')\left[B(\sigma, s', b, s, b)(s) - \sigma(s)\right]\right] + \\ & y(s')\left[\sigma(s)\left[B(\sigma, s, b, s', b)(s) - \sigma(s)\right] + \sigma(s')\left[B(\sigma, s', b, s', b)(s) - \sigma(s)\right]\right] \end{aligned}$$

Step 1 implies $f(\sigma, s) = 0$ for every function C and every sampling rule e, i.e. for every strategy sampling rule $y(\cdot)$. Then it follows that:

$$\sigma(s)\left[B(\sigma, s, b, s', b)(s) - \sigma(s)\right] + \sigma(s')\left[B(\sigma, s', b, s', b)(s) - \sigma(s)\right] = 0$$

Recall that Assumption 3 states that $B(\sigma, s', b, s', b)(s) = \sigma(s)$. Then it follows that $B(\sigma, s, b, s', b)(s) = \sigma(s)$ as asserted.

Step 3: There is function $\widetilde{B} : \Delta(S) \times S \times S \to R$ such that for all $s' \neq s$:

$$B\left(\sigma, s, x, s', y\right)(s) = \sigma(s) + \widetilde{B}\left(\sigma, s, s'\right)(x - y)$$

Proof: Let $a, b, c \in (0,1)$, and suppose $a < b < c$. Consider environment E such that strategy s yields payoff a with probability p and payoff c with probability $1 - p$, whereas strategy $s' \neq s$ yields payoff b with certainty. Let $\widehat{p}$ denote the value of p such that $S^* = S$, i.e. $\widehat{p}$ is given by $\frac{c-b}{c-a}$. Let $\widehat{E}$ denote the environment E such that $p = \widehat{p}$.

The expected movement of the probability of s is given by:

$$f(\sigma,s)=\sigma(s)\sum_{s'\neq s}y(s')\int_0^1\left[B(\sigma,s,x,s',y)(s)-\sigma(s)\right]d\mu_s+$$
$$+y(s)\sum_{s'\neq s}\sigma(s')\int_0^1\left[B(\sigma,s',x,s,y)(s)-\sigma(s)\right]d\mu_{s'}$$

Step 1 implies $\int_0^1\left[B(\sigma,s,x,s',y)(s)-\sigma(s)\right]d\mu_s=0$ for every environment such that $S=S^*$. In particular, for environment $\widehat{E}$ this condition reduces to:

$$\widehat{p}\left[B(\sigma,s,a,s',b)(s)-\sigma(s)\right]+(1-\widehat{p})\left[B(\sigma,s,c,s',b)(s)-\sigma(s)\right]=0$$

Replacing $\widehat{p}$ by $\frac{c-b}{c-a}$ and rearranging terms yields:

$$\frac{B(\sigma,s,c,s',b)(s)-\sigma(s)}{B(\sigma,s,a,s',b)(s)-\sigma(s)}=\frac{c-b}{a-b}$$

At this must be true for all a,b,c with $a<b<c$ it follows that $B(\sigma,s,c,s',b)(s)-\sigma(s)$ must be proportional to payoff difference, as asserted.

Step 4: For every $\sigma\in\dot{\Delta}(S)$, $\widetilde{B}(\sigma,s,s')>0$.

Proof: Consider any $\sigma\in\dot{\Delta}(S)$. The proof is indirect. Suppose there were $s,s'\in S$ with $s'\neq s$ such that $\widetilde{B}(\sigma,s,s')<0$. Consider environment E defined by $\mu_s(a)=1$, $\mu_{s'}(b)=1$ for all $s'\neq s$ and a strategy sampling rule such that $y(s')=1$, where $a,b\in(0,1)$ with $a>b$. Note that $S^*=\{s\}$. Then

$$f(\sigma,s)=\sigma(s)\widetilde{B}(\sigma,s,s')(a-b)$$

But $\widetilde{B}(\sigma,s,s')<0$ implies $f(\sigma,s)<0$ contradicting monotonicity.

(ii) For the Independent Events Condition: To see that every imitation rule which has property (1) in proposition 10 is monotone, note that condition (1) in the definition of monotonicity is trivially satisfied. Moreover, for every $\sigma\in\dot{\Delta}(S)$ and every $s\in S$, the expected movement of the probability of s is given by:

$$f(\sigma,s)=\sum_{s'\in S}\sigma(s')\sum_{s''\in S}y(s'')\int_0^1\int_0^1\left[B(\sigma,s',x,s'',y)(s)-\sigma(s)\right]d\mu_{s'}d\mu_{s''}$$

We can decompose this expression in four terms:

$$\begin{aligned} f(\sigma,s) &= \sigma(s)\sum_{s'\neq s} y(s')\int_0^1\int_0^1 \left[B(\sigma,s,x,s',y)(s)-\sigma(s)\right]d\mu_s d\mu_{s'} + \\ &+y(s)\sum_{s'\neq s}\sigma(s')\int_0^1\int_0^1 \left[B(\sigma,s',x,s,y)(s)-\sigma(s)\right]d\mu_s d\mu_{s'} + \\ &+\sum_{s'\in S} y(s')\sigma(s')\int_0^1\int_0^1 \left[B(\sigma,s',x,s',y)(s)-\sigma(s)\right]d\mu_{s'} d\mu_{s'} + \\ &+\sum_{s'\neq s}\sum_{s''\neq s,s'}\sigma(s')y(s'')\int_0^1\int_0^1 \left[B(\sigma,s',x,s'',y)(s)-\sigma(s)\right]d\mu_{s'} d\mu_{s''} \end{aligned}$$

Note than assumption 3 implies that both the third and the fourth line equal zero. Therefore, the above expression reduces to:

$$\begin{aligned} f(\sigma,s) &= \sigma(s)\sum_{s'\neq s} y(s')\int_0^1\int_0^1 B(\sigma,s,x,s',y)(s)-\sigma(s)d\mu_s d\mu_{s'} + \\ &+y(s)\sum_{s'\neq s}\sigma(s')\int_0^1\int_0^1 B(\sigma,s',x,s,y)(s)-\sigma(s)d\mu_s d\mu_{s'}+ \end{aligned}$$

Using property (1) the above expression can be rewritten:

$$\begin{aligned} f(\sigma,s) &= \sigma(s)\sum_{s'\neq s} y(s')\widetilde{B}(\sigma,s,s')\left[\int_0^1 x d\mu_s - \int_0^1 y d\mu_{s'}\right] - \\ &-y(s)\sum_{s'\neq s}\sigma(s')\widetilde{B}(\sigma,s',s)\left[\int_0^1 x d\mu_{s'} - \int_0^1 y d\mu_s\right] \end{aligned}$$

Rearranging terms:

$$f(\sigma,s) = \sum_{s'\neq s}\left(\pi^s-\pi^{s'}\right)\left[y(s)\sigma(s')\widetilde{B}(\sigma,s',s)+y(s')\sigma(s)\widetilde{B}(\sigma,s,s')\right]$$

If $\pi^s \geq \pi^{s'}$ for all $s' \neq s$, this expression is non-negative, and if the inequality is strict for some s', then it is positive. This implies that the imitation rule is monotone.

In the remainder of the proof, we consider some given monotone imitation rule and we show that the imitation rule has to have property (1) of proposition 10. We proceed in 4 steps.

Step 1: Consider an environment E in which $S^* = S$. Then, for every $\sigma \in \dot{\Delta}(S)$

and every $s \in S$: $f(\sigma, s) = 0$.

Proof: As step 1 of part (i).

Step 2: $B(\sigma, s, x, s', x)(s) = 0$ for all $x \in (0,1), \sigma \in \dot{\Delta}(S)$ and $s' \neq s$.

Proof: As step 2 of part (i).

Step 3: There is function $\widetilde{B} : \Delta(S) \times S \times S \to R$ such that for all $s, s' \in S$ with $s' \neq s$:

1. $B(\sigma, s, x, s', y)(s) = \sigma(s) + \widetilde{B}(\sigma, s, s')(x - y)$

Proof: Let $a, b, c \in (0,1)$, and suppose $a < b < c$. Consider environment E such that strategy s yields payoff a with probability p and payoff c with probability $1 - p$, whereas strategy $s' \neq s$ yields payoff b with certainty. Let $\widehat{p}$ denote the value of p such that $S^* = S$, i.e. $\widehat{p}$ is given by $\frac{c-b}{c-a}$. Let $\widehat{E}$ denote the environment E such that $p = \widehat{p}$.

The expected movement of the probability of s in environment E is given by:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} y(s') \int_0^1 \int_0^1 [B(\sigma, s, x, s', y)(s) - \sigma(s)] \, d\mu_s d\mu_{s'} +$$
$$+ y(s) \sum_{s' \neq s} \sigma(s') \int_0^1 \int_0^1 [B(\sigma, s', x, s, y)(s) - \sigma(s)] \, d\mu_{s'} d\mu_s$$

As any strategy $s' \neq s$ yields payoff b with certainty, the above expression can be written as:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} y(s') \int_0^1 [B(\sigma, s, x, s', b)(s) - \sigma(s)] \, d\mu_s +$$
$$+ y(s) \sum_{s' \neq s} \sigma(s') \int_0^1 [B(\sigma, s', b, s, y)(s) - \sigma(s)] \, d\mu_s$$

Step 1 implies $\int_0^1 [B(\sigma, s, x, s', b)(s) - \sigma(s)] \, d\mu_s$ in every environment such that $S = S^*$. In particular for environment $\widehat{E}$ this equation reduces to:

$$\widehat{p}\,[B(\sigma, s, a; s', b)(s) - \sigma(s)] + (1 - \widehat{p})\,[B(\sigma, s, c; s', b)(s) - \sigma(s)] = 0$$

At this must be true for all a, b, c with $a < b < c$ it follows that $B(\sigma, s, c, s', b)(s) - \sigma(s)$ must be proportional to payoff difference, as asserted.

Step 4: For every $\sigma \in \dot{\Delta}(S)$, $\widetilde{B}(\sigma, s, s') > 0$.

Proof: As step 4 of part (i).

■

This proposition shows that *monotonicity* implies that the change in the relevant probabilities is proportional to the payoff difference. Note that this means that monotonicity incorporates a reinforcement component, the sampled payoff being an aspiration level. A realized payoff bigger than the sampled payoff reinforces the strategy taken, otherwise its probability is decreased. This reinforcement effect makes it possible to learn the expected payoff maximizing strategy even though this strategy is not present in the population.

Note that condition (1) fully characterizes monotonicity under both the common and the independent events condition. This is a direct consequence of assumption 3. Note that this assumption implies no change in the decision maker's state as long as the sampled and taken actions are the same to each other, regardless of payoff realizations.

We now outline the relationship between monotonicity and approximate maximisation. We do not provide a formal analysis as it is similar to that on Chapter 5. The key result for linking these two properties is the fact that for a monotone imitation rule, the probability attached to the expected payoff maximising strategy follows a stochastic process which is a *submartingale*. This allows us to conclude that the probability attached to the event "$\sigma_n(S^*) \to 1$" is always higher than $\sigma_0(S^*)$. This property just says that the probability of being trapped into the expected payoff maximising action is always greater than the probability with which it is initially played. Note that this is a lower bound for this asymptotic probability.

To study approximate maximisation, we need a sequence of imitation rules derived from the imitation rule B. The starting point is therefore to generate a sequence of imitation rules indexed by a parameter $\varepsilon \in (0,1)$. Any member of this family is defined as follows

$$B^{\varepsilon}(\sigma, s, x, s', y) - \sigma = \varepsilon \left[B(\sigma, s, x, s', y) - \sigma \right]$$

Note that B^{ε} describes a behaviour process which moves into the same direction as B, but at speed ε. Notice also that if B is monotone, every member of the sequence is also monotone. We are interested in limit properties of the behaviour process corresponding to B^{ε} for fixed environment, fixed function C, fixed sampling procedure e, and fixed initial state, where the limit which we wish to take is: $\varepsilon \to 0$.

After introducing a continuous time variable $t \geq 0$ and following similar steps as in Chapter 5, the behaviour process B^{ε} when $\varepsilon \to 0$ can be approximated by the solution

of the following differential equation

$$\frac{d\overline{\sigma}_t}{dt} = f(\overline{\sigma}_t)$$

where $f(\cdot)$ describes the expected movement of imitation rule B. Note that the solution of the differential equation satisfies $\lim_{t\to\infty}\overline{\sigma}_t = 1$.

But recall that this is a "good" approximation only for *finite* time intervals and that we are interested in the asymptotics of the process. To overcome this difficulty we use the fact that for monotone imitation rules, $\sigma_0(S^*)$ is a lower bound for the probability of the event "$\sigma_n(S^*) \to 1$". This is enough because for large enough finite time intervals, the behaviour process corresponding to $\sigma_t^\varepsilon(S^*)$ when $\varepsilon \to 0$ is close to 1, which is to say (using that lower bound) that the probability of the event "$\sigma_n^\varepsilon(S^*) \to 1$" is close to 1.

7.5 Conclusion

In this Chapter we have focused on imitative behaviour, specifically on mixed strategy imitation rules. Although a complete analysis of the imitation rules which lead the agent to behave optimally in the long run has not been provided, we have stated a property called monotonicity which implies that any imitation rule satisfying this property can achieve optimality provided they evolve in small steps.

x

We have furthermore characterized all monotone imitation rules. Its basic component is that the change in the decision maker's state is proportional to the payoff difference. A related proportional imitation component is found in Schlag [51] although in a quite different framework. In [51] Schlag considers *pure* strategy imitation rules in an evolving population whereas our characterization concerns *mixed* strategy imitation rules concerning one single decision maker. In addition, Schlag axiomatizes *strictly improving* rules, a property concerned with the evolution of the whole population, whereas in our setting we focus on a property concerned with the behaviour of a single individual while the behaviour of the population remain fixed. Improving rules imitate higher payoff strategies with a probability which is proportional to the payoff difference. In our setting monotonicity implies that the change in the decision maker's state is pro-

portional to the payoff difference, although monotonicity incorporates a reinforcement component, the sampled payoff being an aspiration level, i.e. the probability attached to the own action is increased if it gets a higher payoff than the observed action, otherwise is decreased. This component makes it possible to achieve optimality even if the expected payoff maximising strategy is not present in the population.

Chapter 8

Imitation and Equilibrium in Populations

8.1 Introduction

This Chapter differs from the previous ones in one basic aspect: It focuses on the behaviour of an *entire* population of agents all of whom change their behaviour using some imitation rule, whereas in the previous two Chapters only one individual changed her behaviour through imitation, with all other individuals' behaviour remaining fixed. We shall study the evolution of a population of agents, all of whom face the same two-strategy decision problem. As in the two preceding Chapters, note that the agents of the population are not playing any game against each other. All members of the population are endowed with a pure strategy imitation rule to adapt their behaviour and have the opportunity to observe the strategy and the payoff of one other member of the population. To define this sampling procedure, all members of the population are endowed with a (possibly different) sampling rule, i.e. a probability distribution defined over the population (except herself) which indicates how likely it is that an agent observes any other agent.

In each iteration, after choosing a strategy and receiving a payoff, each agent samples, according to her sampling rule, some other member of the population and observes her strategy and payoff. With this information, she adapts her behaviour according to her imitation rule.

The first property which we investigate refers to the evolution of a population

when all members use the *same* imitation rule. An imitation rule is *maximising* if all members of the population will end up playing the expected payoff maximising strategy independent of what the true payoff distribution is and regardless of the initial state of the population. We show that there are no such maximising imitation rules.

If a maximising rule existed, and if all agents cared only about their asymptotic payoff, then it would be obvious that it would be in their interest to adopt such a rule. In the absence of a maximising rule, individual agents' incentives to adopt any particular rule need to be investigated more carefully. To formalize this issue, we shall begin by defining an appropriate payoff function. For given population and fixed imitation rule, the payoff function for any agent using any imitation rule is defined as the time average payoff that the agent receives along the path. This payoff function will be called the asymptotic payoff. Note that in the definition of the payoff function the assumption is implicit that the agent does *not* discount the future, i.e. we deal with infinitely patient agents. This is a very restrictive assumption. It will simplify the subsequent analysis.

For given population and fixed imitation rule, we can then define the set of imitation rules which are best responses to the fixed imitation rule. An imitation rule will be called an *equilibrium* imitation rule if it belongs to the set of best responses to itself. This means that there is no other imitation rule that an agent might use such that for every decision problem and initial distribution her asymptotic payoff is at least as good, and for at least some decision problem and some initial distribution, her asymptotic payoff is greater using the alternative rule.

Unfortunately, the investigation of equilibrium rules in the general framework has turned out to be very complex. We have only been able to find an example of an equilibrium imitation rule. Somewhat paradoxically, it is "never imitate". This can be understood in light of the result of Chapter 6. If all members of the population use the rule "never imitate", then an agent wondering about using an alternative rule finds herself in the situation analyzed in Chapter 6, where the behaviour of the rest of the population is fixed. Note that the rule "never imitate" yields, for every environment, the maximum asymptotic payoff to those agents who starts playing the optimal strategy. Then, by definition, a better reply should also yield this maximum asymptotic payoff to those agents. But note that any alternative rule should prescribe switching to the sampled strategy for at least one payoff realization. Then for any alternative rule, we

can always find an environment in which this alternative rule implies switching with positive probability from strategy s to strategy s' and also from strategy s' to strategy s. But note that this rule makes an agent who plays the optimal strategy initially switch away from it with positive probability as long as she samples an agent playing the suboptimal strategy. Note that even if the agent eventually switches back to the optimal strategy, this sampling event *always* happens because the population behaviour is fixed. Therefore, the asymptotic payoff to that agent is necessarily lower than the maximum asymptotic payoff.

In order to gain further insights, we shall consider a more restricted domain of analysis. The new framework will consider a two-strategy decision problem with binary payoffs, and a population of only two agents. We identify a property which is of relevance to the problem at hand. An imitation rule is *unbiased* if the asymptotic payoff to all members of a population when all members use this imitation rule is closer to the expected payoff of the optimal strategy than to the expected payoff of the suboptimal strategy. We characterize the set of unbiased rules and show that biased imitation rules are not equilibrium ones. A rule is biased if the probability with which it switches to lower payoff strategies is strictly higher than the probability with which it switches to higher payoff strategies. Unfortunately, we have not been able to prove that this property is also a sufficient condition for the equilibrium property. However, we show two examples of unbiased rules which are equilibrium rules: "always imitate" and "imitate if better". Note that in this setting, the proportional imitation rule, investigated by Schlag [51], would reduce to "imitate if better", and therefore it is very tempting to reach the conclusion that the proportional imitation rule is an equilibrium rule. However, this conclusion is too premature and should not be inferred from this Chapter.

Schlag's paper uses a similar framework to ours. He introduces a property of learning rules which says that in all environments and all current states of the population, the expected average payoff of the *population* increases from round to round. He shows that the proportional imitation rule satisfies this property and that the dynamics of a large, randomly matched population in which all members of the population adopt this rule can be approximated by the replicator dynamics. This means that with high probability, this rule will yield optimality in the long run. But this also means that

for fixed population size, there are environments in which there is a strictly positive probability that the population will choose in the long run a strategy which does not maximize expected payoff. This means that the proportional imitation rule is not a maximising rule and, as we have argued before, that this fact calls for a careful analysis of the incentives that individual members of the population have to stick to the proportional imitation rule.

Schlag's attempt to analyze this is to interpret his property from an individual-oriented perspective. He assumes that if in every round, some members of the population are replaced by new born members who do not know ex-ante which particular members they are to replace. It is then in the interest of the new members to adopt the proportional imitation rule as long as they have a uniform prior. This is so because in this case, the change of their expected payoff will be positive for all environments and for all states of the population.

This argumentation does not seem to us very convincing. It has a population-oriented evolutionary flavour that is linked to the individual's point of view by assuming new-born agents replacing existing ones. We really believe that the basic analysis piece should be the individual rather than the population evolution; moreover we prefer to consider that the individuals in the population are to face the decision problem forever. Agents are modelled as boundedly rational by assuming that they use simple behaviour rules with limited memory to adjust their behaviour. Given the rules used by the rest of the population, an agent chooses the rule that maximises her asymptotic payoff.

The rest of the Chapter is as follows. Sections 2 and 3 will state the formal framework and the main definitions, respectively. Section 4 explores the existence of maximising rules. The equilibrium issue is taken up in Sections 5 and 6. Section 7 concludes.

8.2 Formal Framework

This Chapter refers to the case in which the decision problem has two strategies, i.e. $S = \{s_1, s_2\}$, every agent of the population is allowed to adjust her behaviour, i.e. $W^A = W$, and in which sampling *is* allowed. For each individual $w \in W$ the sampling occurs following some sampling rule, i.e. some exogenously given probability distribution $e_w \in \Delta(W \setminus \{w\})$ where $e_w(w')$ is the probability that individual w samples individual

$w' \neq w$.

8.3 Definitions

Every agent $w \in W$ is characterized by a *pure strategy imitation rule*, which is formally stated in the following definition.

Definition 27 *A pure strategy imitation rule for individual $w \in W$ is a function A_w :* $S \times (0,1) \times S \times (0,1) \to [0,1]$.

The interpretation of a pure strategy imitation rule is the following: $A_w(s,x,s',y)$ is the probability that individual w chooses strategy s' at iteration $n+1$ after choosing strategy s at iteration n, getting a payoff x and sampling an agent choosing strategy s' and receiving a payoff y.

Throughout this Chapter, we focus on learning rules which satisfy the following assumption.

Assumption 5. For every individual $w \in W$, $A_w(s_1,x,s_2,y) = A_w(s_2,x,s_1,y)$ for all $x,y \in (0,1)$.

This assumption means that the imitation rule only depends on payoffs and not on the identity of particular strategies. This assumption motivates the following definition which simplifies terminology.

Definition 28 *A* switching function *for individual $w \in W$ is a function $F_w : (0,1) \times (0,1) \to [0,1]$ such that $F_w(x,y) = A_w(s,x,s',y)$ for all $x,y \in (0,1)$.*

Let $\mathcal{F}$ denote the set of all possible switching functions.

Because, unlike in the previous two Chapters, in this Chapter all members of the population are allowed to adjust their behaviour through imitation, we now have to introduce for all members w of the population a sampling rule, $e_w : W \to [0,1]$, which indicates how likely is that this agent meets the other agents. Obviously, we shall assume that $e_w(w) = 0$ for all $w \in W$.

Definition 29 *A sampling rule for individual w is* complete *if*

$$e_w(w') > 0 \text{ for all } w' \in W \setminus \{w\}$$

Throughout this Chapter, we focus on sampling rules which satisfy the following assumption.

Assumption 6. For every individual $w \in W$, e_w is complete.

This assumption means that every agent is sampled by each other agent with strictly positive probability.

A population is thus described by a 3-tuple $\left(\#W, \{F_w\}_{w\in W}, \{e_w\}_{w\in W}\right)$.

8.4 Payoff Maximisation

This Chapter investigates the performance of a given imitation rule F by focusing on the behaviour of a population of agents, all of whom use this imitation rule. We will refer to these population as F-monomorphic. For given imitation rule F, the behaviour of the monomorphic population $\left(\#W, F, \{e_w\}_{w\in W}\right)$ will evolve over time. We will next define formally the population process.

Consider the $\#W$-dimensional vector which has a 0 in the entry corresponding to individual w_t if this individual has chosen strategy s_1 at iteration n and which has a 1 in that entry otherwise. Read this state as the binary expansion of some number. We shall denote that number as θ_n. We refer to θ_n as the state of the population at iteration n. Let $\Theta = \left\{0, 1, 2, ..., \sum_{k=0}^{\#W} 2^k\right\}$ denote the set of all possible states of the population. For the sake of clarity, rename states 0 and $\sum_{k=0}^{N} 2^k$ as states $S1$ and $S2$ respectively. Let $\overline{\Theta}$ be defined as $\Theta \backslash \{S1, S2\}$. Let Θ^* denote the state in which all agents play the expected payoff maximising strategy.

Definition 30 *The* transition matrix P *corresponding to a monomorphic population* $\left(\#W, F, \{e_w\}_{w\in W}\right)$ *and an environment* E *is a* $\#\Theta \times \#\Theta$ *matrix where the entry in the row corresponding to state* i*, and the column corresponding to state* j*,*

$$p_{ij} = prob(\theta(n+1) = j \mid \theta(n) = i)$$

is determined by the matching and imitation process described in the text.

Definition 31 *The* population process *corresponding to a monomorphic population* $\left(\#W, F, \{e_w\}_{w\in W}\right)$*, an environment* E *and an initial distribution* $\delta_0 \in \Delta(\Theta)$ *is the Markov chain* $\{\theta_n\}_{n\in N_0}$ *with the initial distribution* δ_0 *and with the transition matrix*

P defined in definition 30. For every $n \in \mathbb{N}_0$ we denote by $\delta_n \in \Delta(\Theta)$ the marginal distribution of θ_n.

We can now introduce the first property which we analyze in this Chapter.

Definition 32 *An imitation rule is* maximising *if for every monomorphic population $(\#W, F, \{e_w\}_{w \in W})$, every environment E and every initial distribution $\delta_0 \in \Delta(\overline{\Theta})$:*

$$\lim_{n \to \infty} \delta_n(\Theta^*) = 1$$

This property means that all members of the population will end up playing the expected payoff maximising strategy. Note that this definition only refers to initial distributions which place probability zero on states $S1$ and $S2$. This is so because as the agents are restricted to play either the actual strategy or the sampled strategy, these two states are absorbing. The next proposition shows that there are no maximising equilibrium rules.

Proposition 11 *No imitation rule is maximising.*

Proof of Proposition 11. The proof is indirect. Let F be a maximising rule. We divide the proof in two parts. The first part deals with a population of two members and the second part deals with a population of more than two members.

Part (ii). Fix $\#W = 2$. Note that in this case, the sampling rules are trivial. Note that the initial state of the population is one agent playing strategy s_1 and the other agent playing strategy s_2.

Step 1. Maximisation implies "Imitate a lower payoff strategy with probability strictly less than 1".

Proof: The proof is indirect. Let F be a maximising rule such that for some $x_0, y_0 \in (0,1)$ with $x_0 > y_0$, $F(x_0, y_0) = 1$. Consider environment $\overline{E}$ defined by $\mu_{s_1}(x_0) = 1$ and $\mu_{s_2}(y_0) = 1$. For this environment, $\Theta^* = S1$. Note that in this environment, the agent playing the optimal strategy will sample with probability one strategy s_2 and will switch away with probability one. This means that the probability of the population getting absorbed in state Θ^* is zero. We have then found a contradiction.

Step 2. Maximisation implies "Imitate a higher payoff strategy with strictly positive probability".

Proof: The proof is indirect. Let F be a maximising rule such that for some $x_0, y_0 \in (0,1)$ with $x_0 < y_0$, $F(x_0, y_0) = 0$. Consider environment $\overline{E}$ defined by $\mu_{s_1}(x_0) = 1$ and $\mu_{s_2}(y_0) = 1$. For this environment, $\Theta^* = S2$. Note that in this environment, the agent playing strategy s_1 will sample with probability one strategy s_2 and will stick with probability one. This means that the probability of the population getting absorbed in state Θ^* is zero. We have then found a contradiction.

Step 3. Consider environment $\widetilde{E}$ defined by $\mu_{s_1}(x_0) = \mu_{s_2}(y_0) = \mu$ and $\mu_{s_1}(y_0) = \mu_{s_2}(x_0) = 1 - \mu$ with $x_0 > y_0$ and $\mu \in (0,1)$. For this environment, with positive probability, the agent playing strategy s_1 will get payoff x_0 and the agent playing strategy s_2 will get payoff y_0. With positive probability, the agent playing s_1 sticks (step 1) and the agent playing s_2 switches (Step 2). This means that with positive probability the population gets absorbed in state $S1$. But this event is independent of the particular value of μ. Take environment $\widetilde{E}$ with $\mu > \frac{1}{2}$. This implies $\Theta^* = S2$. We have then found a contradiction to the payoff maximisation of F.

Part (ii). Fix $\#W > 2$ and fix $\{e_w\}_{w \in W}$.

Step 1. Maximisation implies "never switch to a lower payoff strategy".

Proof: The proof is indirect. Let F be a maximising rule such that for some $x_0, y_0 \in (0,1)$ with $x_0 > y_0$, $F(x_0, y_0) > 0$. Consider environment $\overline{E}$ defined by $\mu_{s_1}(x_0) = 1$ and $\mu_{s_2}(y_0) = 1$. For this environment, $\Theta^* = S1$. Moreover as $F(x_0, y_0) > 0$, an agent will switch away with positive probability from the optimal strategy when she samples strategy s_2. If one can show that for this environment there is at least one state $\theta \in \overline{\Theta}$ such that $p^{(1)}_{\theta,S2} > 0$, this would imply that the rule is not maximising, as with positive probability the population is trapped in state $S2$.

For state $\theta \in \overline{\Theta}$, let $\Phi_s(\theta)$ denote the subset of the population W playing strategy $s \in S$ when the population state is θ. Consider any state $\theta \in \overline{\Theta}$ such that $\#\Phi_{s_2}(\theta) > 1$. With positive probability, every $w \in \Phi_1(\theta)$ samples a member of $\Phi_2(\theta)$ and with positive probability they will switch to strategy s_2. Furthermore, with positive probability every $w \in \Phi_2(\theta)$ samples a member of $\Phi_2(\theta)$ and therefore they will continue playing strategy s_2. This means that there is a positive probability of transition from θ to $S2$ in one step, i.e. $p^{(1)}_{\theta,S2} > 0$. We have thus found a contradiction to the maximisation property of F.

Step 2. Maximisation implies "switch with positive probability to a better payoff strat-

egy".

Proof. The proof is indirect. Let F be a maximising rule such that for some $x_0, y_0 \in (0,1)$ with $x_0 > y_0$, $F(y_0, x_0) = 0$. Consider environment $\overline{E}$ defined by $\mu_{s_1}(x_0) = 1$ and $\mu_{s_2}(y_0) = 1$. For this environment, $\Theta^* = S1$. By Step 1, we know that $F(x_0, y_0) = 0$. This implies that for this environment, the imitation rule F becomes "never imitate". But this is clearly non-maximising.

Step 3. Consider environment $\widetilde{E}$ defined by $\mu_{s_1}(x_0) = \mu_{s_2}(y_0) = \mu$ and $\mu_{s_1}(y_0) = \mu_{s_2}(x_0) = 1 - \mu$ with $x_0 > y_0$ and $\mu \in (0,1)$. For this environment, an agent playing strategy s_i and sampling strategy s_j will switch with positive probability. Then it is true that for any state $\theta \in \overline{\Theta}$ such that $\#\Phi_i(\theta) > 1$, $p^{(1)}_{\theta \, \mathrm{Si}} > 0$. We have then found a contradiction to the maximisation property of F. ■

Proposition 11 shows that there are no imitation rules which lead to optimality for every environment and every initial distribution. The next section will investigate whether one individual might improve by using an alternative imitation rule.

8.5 Equilibrium Rules

The last section has suggested the possibility of individual improvement via the use of an alternative imitation rule due to the fact that there are no imitation rules which lead to optimality in all situations. In this Section we develop a framework in which to address this issue. The starting point is a population of agents described by a 3-tuple $\left(\#W, \{F_w\}_{w \in W}, \{e_w\}_{w \in W}\right)$. For given environment E, given initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$ and given sampling rules $\{e_w\}_{w \in W}$, the evolution of the population is described by a Markov chain where the transition probabilities depend upon the imitation rules used by the agents of the population. Note that without loss of generality we can assume that the initial distribution puts probability zero on states $S1$ and $S2$. Recall that these states are absorbing independent of what the imitation rules are and therefore the possible improvement is not to take place when the population starts out in one of these states.

Our first step is to define a payoff function

$$\pi_w : \mathcal{F} \times \mathcal{F} \to R$$

such that $\pi_w(f, F) \in R$ is the payoff to agent w when he uses imitation rule f and all other members of the population use the rule F. For given state $\theta \in \overline{\Theta}$, let $\Psi(w, \theta)$ be the strategy that agent w plays at state θ. Let $\overline{\Theta}_{w,s_1}$ and $\overline{\Theta}_{w,s_2}$define a partition of $\overline{\Theta}$, where $\overline{\Theta}_{w,s} = \left\{\theta \in \overline{\Theta} \mid \Psi(w, \theta) = s\right\}$. Starting out at state θ, let $V_{\theta,\theta'}(n)$ be the number of visits to state θ' before round n, i.e. $V_{\theta,\theta'}(n) = \sum_{k=0}^{n-1} I_{\{\theta_n=\theta'\}}$, where $I_{\{\theta_n=\theta'\}}$ is an indicator function which takes the value 1 when $\theta_n = \theta'$ and 0 otherwise. Obviously, $V_{\theta,\theta'}(n)$ is a random variable. By standard results about finite Markov chains, it is true that $\frac{V_{\theta,\theta'}(n)}{n}$ converges with probability one to some limit, i.e.

$$Prob\left[\lim_{n\to\infty} \frac{V_{\theta,\theta'}(n)}{n} \text{ exists}\right] = 1$$

Denote this limit by $\tilde{c}_{\theta,\theta'}$, and note that it is a random variable. Denote its expected value by $c_{\theta,\theta'}$.

For given environment E, given initial state $\theta \in \bar{\Theta}$, and given sampling rules $\{e_w\}_{w\in W}$ the asymptotic payoff to agent w is:

$$\pi_{s_1} \sum_{\theta' \in \overline{\Theta}_{w,s_1}} \tilde{c}_{\theta,\theta'} + \pi_{s_2} \sum_{\theta' \in \overline{\Theta}_{w,s_2}} \tilde{c}_{\theta,\theta'}$$

The expected value of this random variable is:

$$\pi_{s_1} \sum_{\theta' \in \overline{\Theta}_{w,s_1}} c_{\theta,\theta'} + \pi_{s_2} \sum_{\theta' \in \overline{\Theta}_{w,s_2}} c_{\theta,\theta'}$$

Note that in the definition of the asymptotic payoff, we have made use of the Law of Large Numbers for payoffs. Conditional to the visits to a particular state $\theta' \in \overline{\Theta}_{w,s_i}$, the decision maker receives a time average payoff which converges with probability one to the expected payoff π_{s_i} by the Law of Large Numbers.[1]

If the initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$ is non-trivial, we still have to take expected

[1]Recall that payoff realizations are assumed to be stochastically independent across iterations and that in a given iteration, each strategy s has a payoff distribution attached to it that does not change over time.

values over initial states. Thus, we arrive at this formula for the asymptotic payoff:

$$\pi_w(f,F) = \sum_{\theta \in \overline{\Theta}} \delta_0(\theta) \left(\pi_{s_1} \sum_{\theta' \in \overline{\Theta}_{w,s_1}} c_{\theta,\theta'} + \pi_{s_2} \sum_{\theta' \in \overline{\Theta}_{w,s_2}} c_{\theta,\theta'} \right)$$

We can now state the main property of this section.

Definition 33 *Let W be a monomorphic population with sampling rules $\{e_w\}_{w \in W}$. An imitation rule f is a* better response *than an imitation rule g for individual w to the population rule F if*

1. For every environment E and every initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$, $\pi_w(f,F) \geq \pi_w(g,F)$
2. There exists environment E and initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$ with $\pi_w(f,F) > \pi_w(g,F)$

Definition 34 *An imitation rule f is a* best response *to a population rule F if there is no better response. Let $BR_w(F)$ denote the set of best responses to the population rule F for individual $w \in W$.*

Definition 35 *An imitation rule F is an equilibrium rule if for every individual $w \in W$, $F \in BR_w(F)$.*

Note that a monomorphic population using an equilibrium rule can be said to be stable, i.e. in this class of monomorphic populations there are no individual incentives to deviate and use an alternative rule.

Unfortunately, we have not been able to characterise the set of equilibrium rules for given population W and given sampling procedure. We have only been able to produce a single example of equilibrium rule that paradoxically is "never imitate".

Proposition 12 *"Never imitate" is an equilibrium rule.*

Proof of Proposition 12. Let F be the rule "never imitate", i.e. $F(x,y) = 0$ for all $x, y \in (0,1)$. The asymptotic payoff to agent w using the rule F is given by

$$\pi_w(F,F) = \sum_{\theta \in \overline{\Theta}_{w,s_1}} \delta_0(\theta)\pi_{s_1} + \sum_{\theta \in \overline{\Theta}_{w,s_2}} \delta_0(\theta)\pi_{s_2}$$

Note that in this case, every state of the Markov chain is an absorbing state. Let f be an alternative imitation rule. We shall prove that for every alternative rule f there is at least one environment and one initial state for which $\pi_w(f,F) < \pi_w(F,F)$. Note that this would mean that there is no better reply than F to the population rule F and therefore that F is an imitation rule.

Given that $f \neq F$, it follows that there exists at least one pair (x,y) such that $f(x,y) > 0$. Consider an environment $\widetilde{E}$ defined by $\mu_{s_1}(x) = \mu_{s_2}(y) = \mu$ and $\mu_{s_1}(y) = \mu_{s_2}(x) = 1-\mu$ with $\mu \in (0,1)$. Note that agent w will switch away from both strategy s_1 and s_2 with strictly positive probability because $f(x,y) > 0$. Consider that the initial distribution puts probability one in one particular state, and that agent w is the only agent playing the optimal strategy in that state. Without loss of generality, let s_1 be the expected payoff maximising strategy. By using the rule F, agent w gets an asymptotic payoff π^{s_1}. By using the alternative rule f, there is a positive probability that agent w switches away from strategy s_1 to strategy s_2.[2] This means that there is probability one that the population is trapped in the absorbing state $S2$, yielding an asymptotic payoff $\pi^{s_2} < \pi^{s_1}$. ■

The intuition of this result comes from the fact that the population behaviour is fixed. Note that it is always possible to find an environment in which an alternative rule to "never imitate" makes an agent who plays the optimal strategy to switch away from it with positive probability by starting the process in an state in which there is at least an agent playing the suboptimal strategy. Note that even if the agent eventually switches back to the optimal strategy, the event of switching away from the optimal strategy *always* happens because the population behaviour is fixed. This is what motivates that "never imitate" is an equilibrium rule.[3]

In order to get more insights into the features of equilibrium rules, the next Section will consider a more restricted framework.

[2]Note than then all the members of the population will be playing strategy s_2.

[3]Note that the equilibrium property of the rule "never imitate" is robust to the introduction of a discount factor in the definition of payoffs.

8.6 The Binary Payoff 2x2 Decision Problem

In this section we shall restrict the domain of our analysis to get further insight into the equilibrium property. We shall consider the case of a population with two agents, each of whom is facing the same two-strategy decision problem with binary payoff, i.e. with probability μ strategy s_1 yields payoff one and strategy s_2 yields payoff zero; and with probability $1-\mu$ strategy s_1 yields payoff zero and strategy s_2 yields payoff one. Let $\pi_{s_1} = \mu$ denote the expected payoff of strategy s_1 and let $\pi_{s_2} = 1-\mu$ denote the expected payoff of strategy s_2.

Note that for a population of two agents, the sampling rules are trivial, i.e. $e_w(w') = 1$ for $w \neq w'$,and $w, w' = 1, 2$. In this restricted environment, an imitation rule for agent w is a collection of four switching probabilities, i.e. $F_w(0,0), F_w(0,1), F_w(1,0)$ and $F_w(1,1)$. In order to further simplify the setup, we shall also assume that whenever a strategy yields payoff one, the other strategy yields payoff 0. This restriction will allow us to just consider two switching probabilities, $F_w(0,1)$ and $F_w(1,0)$. To simplify notation, let $\alpha_w = F_w(1,0)$ and $\beta_w = F_w(0,1)$. An imitation rule F will be denoted by the pair (α, β).

The evolution of the population is a 4-state Markov chain with transition matrix:

	(s_1,s_1)	(s_1,s_2)	(s_2,s_1)	(s_2,s_2)
(s_1,s_1)	1	0	0	0
(s_1,s_2)	p_{21}	p_{22}	p_{23}	p_{24}
(s_2,s_1)	p_{31}	p_{32}	p_{33}	p_{34}
(s_2,s_2)	0	0	0	1

where

$$
\begin{aligned}
p_{21} &= \mu(1-\alpha_1)\beta_2 + (1-\mu)(1-\beta_1)\alpha_2 \\
p_{22} &= \mu(1-\alpha_1)(1-\beta_2) + (1-\mu)(1-\beta_1)(1-\alpha_2) \\
p_{23} &= \mu\alpha_1\beta_2 + (1-\mu)\beta_1\alpha_2 \\
p_{24} &= \mu\alpha_1(1-\beta_2) + (1-\mu)\beta_1(1-\alpha_2) \\
p_{31} &= \mu(1-\alpha_2)\beta_1 + (1-\mu)(1-\beta_2)\alpha_1 \\
p_{32} &= \mu\alpha_2\beta_1 + (1-\mu)\beta_2\alpha_1 \\
p_{33} &= \mu(1-\alpha_2)(1-\beta_1) + (1-\mu)(1-\beta_2)(1-\alpha_1) \\
p_{34} &= \mu\alpha_2(1-\beta_1) + (1-\mu)\beta_2(1-\alpha_1)
\end{aligned}
$$

and where the state (s_i, s_j) is interpreted as agent 1 playing strategy s_i and agent 2 playing strategy s_j.

Note that in this restricted framework, the set $\overline{\Theta}$ has only two elements, i.e. $\overline{\Theta} = \{(s_1, s_2), (s_2, s_1)\}$. Hence, the population process will heavily depend on the characterization of these two states. The next proposition classifies them in terms of the imitation rules used by the agents.

Proposition 13 *Fix $\mu \in (0,1)$. Then the state space of the Markov chain can be classified according to the following characterization:*

1. *If there exists at least some α_i or $\beta_i \in (0,1)$, then (s_1, s_2) and (s_2, s_1) are transient states.*

2. *For α_i, $\beta_i \in \{0,1\}$ then*

   (a) *If $\alpha_i = \beta_i = 0$ for $i = 1, 2$, then (s_1, s_2) and (s_2, s_1) are absorbing states.*

   (b) *If at least some α_i or β_i equals 1 and*

      i. *$\alpha_i = \beta_j$ for $i \neq j$, $i, j = 1, 2$ then $\{(s_1, s_2), (s_2, s_1)\}$ is a positive recurrent closed class.*

      ii. *$\alpha_i \neq \beta_j$ for some $i \neq j$, $i, j = 1, 2$, then (s_1, s_2) and (s_2, s_1) are transient states.*

Proof of Proposition 13. The following straightforward equivalencies will be useful:

$$(i)\ p_{21} > 0 \Leftrightarrow p_{34} > 0$$

$$(ii)\ p_{24} > 0 \Leftrightarrow p_{31} > 0$$

$$(iii)\ p_{21} = 0 \Leftrightarrow p_{34} = 0$$

$$(iv)\ p_{24} = 0 \Leftrightarrow p_{31} = 0$$

$$(v)\ p_{23} > 0 \Leftrightarrow p_{32} > 0$$

Part 1. Note that as long as the above probabilities are positive, states (s_1, s_2) and (s_2, s_1) will be transient states because each of them would communicate with an absorbing state. First, let $\alpha_1 \in (0,1)$. If $\beta_2 = 0$ then (ii) applies. If $\beta_2 = 1$ then (i) applies. Second, let $\beta_1 \in (0,1)$. If $\alpha_2 = 0$ then (ii) applies. If $\alpha_2 = 1$ then (i) applies. Third, let $\alpha_2 \in (0,1)$. If $\beta_1 = 0$ then (i) applies. If $\beta_1 = 1$ then (ii) applies. And finally, let $\beta_2 \in (0,1)$. If $\alpha_1 = 0$ then (i) applies. If $\alpha_1 = 1$ then (ii) applies.

Part 2.a. Note that if α_i, $\beta_i = 0$ for $i = 1, 2$, then $p_{22} = p_{33} = 1$ and therefore states (s_1, s_2) and (s_2, s_1) are absorbing states.

Part 2.b.i. Firstly, let $\alpha_1 = \beta_2 = 1$ and let either $\alpha_2 = \beta_1 = 0$ or $\alpha_2 = \beta_1 = 1$. Then it is true that $p_{21} = p_{24} = 0$ and by (iii) and (iv), $p_{31} = p_{34} = 0$, and it is true that $p_{23} > 0$ that implies $p_{32} > 0$ by (v). This means that states (s_1, s_2) and (s_2, s_1) intercommunicate and do not communicate to either states (s_1, s_1) and (s_2, s_2), i.e. states (s_1, s_2) and (s_2, s_1) form a closed recurrent class. The proof for the case $\alpha_2 = \beta_1 = 1$ is similar.

Part 2.b.ii. First, consider the case $\alpha_1 \neq \beta_2$. There are two subcases, either $\alpha_1 = 1$ and $\beta_2 = 0$ or $\alpha_1 = 0$ and $\beta_2 = 1$. In the first subcase, it is the case that $p_{24} > 0$ which implies $p_{31} > 0$ by (ii) and the conclusion holds. In the second subcase, it is the case that $p_{21} > 0$ which implies $p_{34} > 0$ by (i) and the conclusion holds. Second, consider the case $\alpha_2 \neq \beta_1$. There are two subcases, either $\alpha_2 = 1$ and $\beta_1 = 0$ or $\alpha_2 = 0$ and $\beta_1 = 1$. In the first subcase, it is the case that $p_{21} > 0$ which implies $p_{34} > 0$ by (i) and the conclusion holds. In the second subcase, it is the case that $p_{24} > 0$ which implies $p_{31} > 0$ by (ii) and the conclusion holds. ■

Remark 3 *For the cases in which $\mu = 0$ or $\mu = 1$ a similar analysis can be done. It will be done if necessary.*

For given imitation rules F_w and $F_{w'}$ if states (s_1, s_2) and (s_2, s_1) happen to be transient then the long run behaviour of the Markov chain is described by the following

asymptotic distribution. Let X denote the probability of the population being trapped in state (s_1, s_1) conditional on being at state (s_1, s_2). Let Y denote the probability of the population being trapped in state (s_1, s_1) conditional on being at state (s_2, s_1). Then it follows that:

$$\begin{aligned} X &= p_{21} + p_{22}X + p_{23}Y \\ Y &= p_{31} + p_{32}X + p_{33}Y \end{aligned}$$

The solution of this equations system is:

$$X = \frac{p_{21}(1 - p_{33}) + p_{23}p_{31}}{(1 - p_{33})(1 - p_{22}) - p_{32}p_{23}}$$

$$Y = \frac{p_{31}(1 - p_{22}) + p_{32}p_{21}}{(1 - p_{33})(1 - p_{22}) - p_{32}p_{23}}$$

Note that in this case, the expected values of the random variables $\widetilde{c}_{\theta,\theta'}$ are the following:

$$\begin{aligned} c_{(s_1,s_2),(s_1,s_1)} &= 1 \times X + 0 \times (1 - X) = X \\ c_{(s_1,s_2),(s_2,s_2)} &= 0 \times X + 1 \times (1 - X) = 1 - X \\ c_{(s_2,s_1),(s_1,s_1)} &= 1 \times Y + 0 \times (1 - Y) = Y \\ c_{(s_2,s_1),(s_2,s_2)} &= 0 \times Y + 1 \times (1 - Y) = 1 - Y \end{aligned}$$

Thus the asymptotic payoff to agent 1 is given by:

$$\pi_1(F_w, F_{w'}) = \delta_0(s_1, s_2)\left[X\pi_{s_1} + (1 - X)\pi_{s_2}\right] + \delta_0(s_2, s_1)\left[Y\pi_{s_1} + (1 - Y)\pi_{s_2}\right]$$

and the asymptotic payoff to agent 2 is given by:

$$\pi_2(F_{w'}, F_w) = \delta_0(s_1, s_2)\left[Y\pi_{s_1} + (1 - Y)\pi_{s_2}\right] + \delta_0(s_2, s_1)\left[X\pi_{s_1} + (1 - X)\pi_{s_2}\right]$$

Note that for a monomorphic population, i.e. $F_w = F_{w'}$, it is true that $p_{21} = p_{31}$, $p_{22} = p_{33}$, $p_{23} = p_{32}$ and $p_{24} = p_{34}$.This implies that $X = Y$ and therefore

$$\pi_1(F_w, F_{w'}) = \pi_2(F_{w'}, F_w) = X\pi_{s_1} + (1 - X)\pi_{s_2}$$

for every initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$.

The following definition will be relevant for the analysis of equilibrium rules.

Definition 36 *An imitation rule F is* unbiased *if for every environment E and every initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$, $\pi_w(F,F) \geq \frac{\pi_{s_1}+\pi_{s_2}}{2}$ for $w = 1,2$.*

This property means that every agent in a monomorphic population gets an asymptotic payoff closer to the maximising expected payoff. The next proposition characterizes the set of unbiased imitation rules

Proposition 14 *An imitation rule $F = (\alpha, \beta) \neq (0,0)$ is unbiased if and only if $\alpha \leq \beta$.*

Proof of Proposition 14. The proof is divided in two steps. The first step deals with the imitation rule $F = (1,1)$. The second step deals with imitation rules $F \neq (1,1)$.
Step 1. Consider the imitation rule $F = (1,1)$. In this case, the set $\{(s_1, s_2), (s_2, s_1)\}$ is a recurrent periodic closed class of states. This means that for any environment and every initial distribution, the variables $\widetilde{c}_{\theta,\theta'}$ are no longer random. With probability one, these variables take the value $c_{\theta,\theta'}$ which can be interpreted as the long run proportion of time spent by the Markov chain in each of these two states. For this case, it is clear that they equal $\frac{1}{2}$, yielding a payoff $\frac{\pi_{s_1}+\pi_{s_2}}{2}$ and the claim follows.
Step 2. Let $F \neq (1,1)$. Then by Proposition 13, states (s_1, s_2) and (s_2, s_1) are transient. The asymptotic payoffs are given by the asymptotic distribution. Recall that for any monomorphic population $X = Y$. Therefore, for every environment E and every initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$, the asymptotic payoff to individual w is:

$$\pi_w(F,F) = [X\pi_{s_1} + (1-X)\pi_{s_2}]$$

Then it follows that:

$$\pi_w(F,F) - \frac{\pi_{s_1}+\pi_{s_2}}{2} = \left(X - \frac{1}{2}\right)(\pi_{s_1} - \pi_{s_2})$$

Straightforward calculations show that:

$$X - \frac{1}{2} = \frac{(\beta-\alpha)(2\mu-1)}{2[\alpha(1-\beta)+\beta(1-\alpha)]}$$

Noting that $\pi_{s_1} - \pi_{s_2} = 2\mu - 1$, we get:

$$\pi_w(F,F) - \frac{\pi_{s_1} + \pi_{s_2}}{2} = \frac{(2\mu-1)^2}{2\left[\alpha(1-\beta)+\beta(1-\alpha)\right]}(\beta-\alpha)$$

and therefore our claim follows because[4]

$$sign\left(\pi_w(F,F) - \frac{\pi_{s_1} + \pi_{s_2}}{2}\right) = sign(\beta-\alpha)$$

■

Note that the rule $F = (0,0)$ is not included in the definition of unbiased imitation rules. The reason is that it is the only rule with $\alpha = \beta$ which is biased.

The next proposition shows that an imitation rule $F = (\alpha, \beta)$ with $\alpha > \beta$ is not a best reply to itself, i.e. it is not an equilibrium rule.

Proposition 15 *Let $F = (\alpha, \beta)$ be an imitation rule such that $\alpha > \beta$. Then the rule $f = (\beta, \alpha)$ is a better response than the rule F for individual w to the population rule F.*

Proof of Proposition 15. We first deal with the rule $F = (1,0)$. In this case, the asymptotic payoff to each agent is given by $\pi_w(F,F) = (1-\mu)\pi_{s_1} + \mu\pi_{s_2}$. If individual 1 uses the alternative rule $f = (0,1)$, then states (s_1, s_2) and (s_2, s_1) form a recurrent closed class of states with stationary distribution $(\mu, 1-\mu)$. Then the asymptotic payoff to individual one is $\pi_w(f,F) = \mu\pi_{s_1} + (1-\mu)\pi_{s_2}$and therefore it follows that $\pi_w(f,F) \geq \pi_w(F,F)$, with strict inequality for $\mu \neq \frac{1}{2}$. The analysis for individual 2 is similar.

We now turn to the case in which $F \neq (1,0)$. We first prove that for these rules, the asymptotic payoff to every agent for every initial distribution and every environment is given by $\pi_w(f,F) = \frac{\pi_{s_1}+\pi_{s_2}}{2}$. Note that $F \neq (1,0)$ implies that states (s_1, s_2) and (s_2, s_1) are transient and therefore that the asymptotic payoff to agent w will be characterized by the asymptotic distribution. Straightforward calculations show that $X = Y = \frac{1}{2}$. Then the asymptotic payoff to each agent is given by $\frac{\pi_{s_1}+\pi_{s_2}}{2}$.

Hence, for every $F = (\alpha, \beta)$ and $f = (\beta, \alpha)$, it is true that $\pi_w(f,F) = \frac{\pi_{s_1}+\pi_{s_2}}{2}$ for every agent w, every environment and every initial distribution.

[4]Note that the denominator is strictly positive.

We shall next prove that f is better response than F to the population rule F. To this end, note that:

$$\pi_w(F,F) - \pi_w(f,F) = \pi_w(F,F) - \frac{\pi_{s_1} + \pi_{s_{21}}}{2}$$

But by Proposition 14, we know that

$$\pi_w(F,F) - \frac{\pi_{s_1} + \pi_{s_{21}}}{2} = \frac{(2\mu - 1)^2}{2\left[\alpha(1-\beta) + \beta(1-\alpha)\right]}(\beta - \alpha)$$

Therefore for environment $\mu \neq \frac{1}{2}$ it follows that $\pi_w(f,F) > \pi_w(F,F)$ and for environment $\mu = \frac{1}{2}$ it follows that $\pi_w(f,F) = \pi_w(F,F)$. ■

Proposition 15 leads us to look for imitation rules among the set of unbiased rules. We have been unable to prove that *every* unbiased rule is an equilibrium one. In the remainder of this section, we shall show some examples of unbiased equilibrium rules.

Proposition 16 *The rule "*Always imitate*" is an equilibrium rule.*

Proof of Proposition 16: This rule is represented by the pair $F = (1,1)$. If both agents use this rule, then the set $\{(s_1,s_2),(s_2,s_1)\}$ is a recurrent closed class of states. This means that for every initial distribution $\delta_0 \in \Delta\left(\overline{\Theta}\right)$, the asymptotic payoff is given by:

$$\pi_w(F,F) = X_F \pi_{s_1} + (1 - X_F)\pi_{s_2} \text{ for } w = 1,2.$$

where $X_F = \frac{1}{2}$.

We shall prove that there are no better responses than F to F by focusing on the asymptotic payoff to agent 1 for initial distribution $\delta_0(s_1,s_2) = 1$.

We shall divide the set of alternative imitation rules in two groups. The first group is composed of the rule $f = (0,0)$. If agent 1 uses this rule, then her asymptotic payoff is given by $\pi_1(f,F) = \pi_{s_1}$. But it is clear that $\pi_1(f,F) < \pi_1(F,F)$ for environments in which s_2 is the expected payoff maximizing strategy. This implies that the rule $f = (0,0)$ is not a better response that F to the population rule F.

We shall investigate now alternative rules $f = (\alpha,\beta) \neq (0,0)$. In this case, from Proposition 13, states $\{(s_1,s_2),(s_2,s_1)\}$ are transient and the asymptotic payoff to agent 1 is given by:

$$\pi_1(f,F) = X_f \pi_{s_1} + (1 - X_f)\pi_{s_2}$$

where

$$X_f = \frac{\mu(1-\alpha)+(1-\mu)(1-\beta)}{1-[\mu\beta+\alpha(1-\mu)][(1-\mu)\beta+\alpha\mu]}$$

Note that X_f is a function of μ. To emphasize this, we change notation and write $X_f(\mu)$. We shall now prove that $X_f\left(\frac{1}{2}\right) \neq X_F\left(\frac{1}{2}\right)$ implies that f is not better response than F to the population rule F. We shall prove it by finding an environment in which $\pi_1(f,F) - \pi_1(F,F) < 0$.

First, consider the case $X_f\left(\frac{1}{2}\right) > X_F\left(\frac{1}{2}\right)$. Consider a new environment defined by $\mu = \frac{1}{2} - \varepsilon$, where $\varepsilon > 0$. By continuity it is true that $X_f\left(\frac{1}{2}-\varepsilon\right) > X_F\left(\frac{1}{2}-\varepsilon\right)$. But this implies that

$$\begin{aligned}
\pi_1(F,F) - \pi_1(f,F) &= \left[X_F\left(\frac{1}{2}-\varepsilon\right) - X_f\left(\frac{1}{2}-\varepsilon\right)\right](\pi_{s_1} - \pi_{s_2}) \Rightarrow \\
\pi_1(F,F) - \pi_1(f,F) &= \left[X_F\left(\frac{1}{2}-\varepsilon\right) - X_f\left(\frac{1}{2}-\varepsilon\right)\right](-2\varepsilon) \Rightarrow \\
\pi_1(F,F) - \pi_1(f,F) &> 0
\end{aligned}$$

Then it follows that the rule f is not a better response than the rule F to the population rule F. On the other hand, for the case $X_f\left(\frac{1}{2}\right) < X_F\left(\frac{1}{2}-\varepsilon\right)$ a similar argument runs for the environment defined by $\mu = \frac{1}{2} + \varepsilon$.

We are therefore left with those rules f such that $X_f\left(\frac{1}{2}\right) = \frac{1}{2}$. But we shall show that there are no such rules. Note that this condition reduces to

$$(\alpha+\beta)^2 + 4(1-\alpha-\beta) = 0$$

Let $G(\alpha,\beta) = (\alpha+\beta)^2 + 4(1-\alpha-\beta)$. We shall show that for every f with $\alpha,\beta \in [0,1]$, $G(\alpha,\beta) > 0$. To prove this, note that $\frac{\partial G}{\partial \alpha} = \frac{\partial G}{\partial \beta} = 2(\alpha+\beta) - 4 < 0$ for every $\alpha,\beta \in [0,1)$. As $G(1,1) = 0$, it follows that $G(\alpha,\beta) > 0$ for every pair $\alpha,\beta \in [0,1)$. ■

We shall conclude this section by showing that the rule "*Imitate if better*" is also an equilibrium rule.

Proposition 17 *The rule* "Imitate if better" *is an equilibrium rule.*

Proof of Proposition 17: This rule is the pair $F = (0,1)$. If agent 1 uses this rule,

then her asymptotic payoff for every initial distribution is given by:

$$\pi_1(F,F) = X_F \pi_{s_1} + (1 - X_F)\, \pi_{s_2}$$

where $X_F = \mu$.

In the following, we shall focus on the asymptotic payoff to individual 1 for initial distribution $\delta_0(s_1, s_2) = 1$.

We shall divide the set of alternative rules f in two groups. The first group is composed of the imitation rule "Imitate of worse", i.e. $f = (1,0)$. Consider environment defined by $\mu = 0$. Then $\pi_1(f,F) = 0 < 1 = \pi_1(F,F)$. It follows that f is not a better response than F to the population rule F.

The second group is composed of the imitation rules $f = (\alpha, \beta) \neq (1,0)$. By Proposition 13, states $\{(s_1, s_2), (s_2, s_1)\}$ are transient and the asymptotic payoff to agent 1 is given by:

$$\pi_1(f,F) = X_f \pi_{s_1} + (1 - X_f)\pi_{s_2}$$

where

$$X_f = \frac{\mu(1-\alpha)[1-\mu(1-\beta)] + \alpha\mu\beta\mu}{[1-\mu(1-\beta)][1-(1-\mu)(1-\beta)] - \alpha(1-\mu)\alpha\mu}$$

Using a similar argument as in the proof of Proposition 16, we only need to focus on rules such that $X_f\left(\mu = \frac{1}{2}\right) = X_F\left(\mu = \frac{1}{2}\right)$. This equation reduces to

$$\frac{1+\beta-\alpha}{(1+\beta)^2 - \alpha^2} = \frac{1}{2}$$

The solution to this equation is $\alpha + \beta = 1$.

We have then found a continuum of rules $f = (\alpha, \beta)$ which are candidates to be better responses than $F = (0,1)$ to the population rule $F = (0,1)$. We shall show that all the rules characterized by $\alpha + \beta = 1$ are not better responses than F to the population rule F. We shall do it by showing that $\pi_w(F,F) = \pi_w(f,F)$ for every environment, every initial distribution and $w = 1, 2$.

For any rule $f = (\alpha, \beta)$ with $\alpha + \beta = 1$ it is true that:

$$X_f = \frac{\mu(1-\alpha)(1-\mu\alpha) + \alpha(1-\alpha)\mu^2}{[1-(1-\mu)\alpha][1-\mu\alpha] - \alpha^2(1-\mu)\mu}$$

By rearranging we get:

$$X_f = \frac{\mu(1-\alpha)}{1-\alpha} = \mu$$

Furthermore, for any rule $f = (\alpha, \beta)$ with $\alpha + \beta = 1$ it is true that:

$$Y_f = \frac{\mu(1-\alpha)\left(1-(1-\mu)\alpha\right) + \alpha(1-\alpha)\mu(1-\mu)}{\left[1-(1-\mu)\alpha\right]\left[1-\mu\alpha\right] - \alpha^2(1-\mu)\mu}$$

By rearranging we get:

$$Y_f = \frac{\mu(1-\alpha)}{1-\alpha} = \mu$$

Then it follows that $X_f = Y_f$, that implies that the asymptotic payoff to agent w is given by

$$\pi_w(f, F) = X_f \pi_{s_1} + (1 - X_f)\pi_{s_2}$$

where $X_f = \mu$.

Recall that for any monomorphic population it is true that $X = Y$. Therefore, the asymptotic payoff to individual w for any environment and every initial distribution is given by

$$\pi_w(F, F) = X_F \pi_{s_1} + (1 - X_F)\pi_{s_2}$$

where $X_F = \mu$.

We have then found that $X_f = X_F$. As this is true for every environment and every initial distribution, it follows that $\pi_w(F, F) = \pi_w(f, F)$. ■

The above proof shows a counterintuitive fact. If an agent is facing another individual who uses the rule "Imitate is better", then the agent is indifferent between using the rule "Imitate is better" or using another rule $f = (\alpha, \beta)$ with $\alpha + \beta = 1$ and $\alpha > 0$. Note that this includes unbiased rules such that for example $f = (.0, .1)$. The indifference between these two rules seems striking to us. We have no intuition for this result.

Note that if one applies Schlag's proportional imitation rule [51] to this setting, it reduces to "imitate if better". Hence it is very tempting to conclude that proportional imitation is an equilibrium rule. However we are aware that it is not a sensible conclusion because this setting is very restrictive. Additional work needs to be done to assess whether there are no individual incentives to deviate from the proportional imitation rule. Note that a result of this sort would reinforce the replicator dynamics

axiomatization based on this rule in [51].

8.7 Conclusion

In this Chapter we have investigated the ability of pure strategy imitation rules to lead an entire population of agents towards rationality when all members of the population use the *same* imitation rule. We have shown that there are no imitation rules such that all members of the population will end up playing the expected payoff maximising strategy independent of what the true payoff distribution is and regardless of the initial state of the population. We then have investigated the existence of equilibrium rules once an appropriate payoff function is defined. However this question has shown to be complex. We have shown a single example of imitation rule which paradoxically is "never imitate".

The restrictive setting of two-strategy binary payoff decision problems with a population of two agents has allowed us to get some further insights into this question. We have shown that some well-known imitation rules like "always imitate" and "imitate if better" are equilibrium rules. Note that the latter is the translation of the proportional imitation rule [51] into this setting. Thus it is very tempting to reach the conclusion that the proportional imitation rule is an equilibrium rule. This would reinforce the replicator dynamics axiomatization based on this rule. We leave an open door for further research in this direction.

The equilibrium analysis undertaken in this Chapter is our attempt to investigate individual incentives to adhere to a particular imitation rule in population contexts. As the basic analysis piece should be the individual, and given that agents are not replaced, individual members of the population evaluate the performance of different behaviour rules given the rules used by the remaining members of the population and choose the rule which maximises her asymptotic payoff. Once every single member of the population uses an equilibrium rule profile, no agent will have incentives to deviate.

Chapter 9

Conclusions

In recent years, economic theorists have started to enquire into the foundations of the rationality hypothesis. The current thesis has joined this line of thought as it has tried to answer to following question: when should rational behaviour be expected. Within the different mechanisms proposed by the economics literature to bring about rationality, this thesis has focused on reinforcement learning as a mechanism to bring about rationality in decision problems.

Reinforcement learning models describe decision makers as simple stimulus-response machines, without attributing any beliefs to them, in sharp contrast with belief-based learning models, such as fictitious play, with which economists are typically more familiar. In this thesis, reinforcement learning has been interpreted in a broad sense, by allowing the decision maker to observe not only her own strategy choice and payoff, but also the strategy and payoff of some other agents who finds themselves in the same situation. In this way we have been able to also consider learning by imitation.

In this framework, we have studied the ability of reinforcement learning to lead the decision maker to play in the long run the expected payoff maximising strategy. In investigating this ability, we have considered a large class of reinforcement models, without postulating any functional form. This approach has allowed us to identify some features which characterize those reinforcement rules which steer the agent towards rationality in all decision problems.

Our characterizations focus on two properties of a learning model. The first property concerns the state space of the decision maker, i.e. the set of states in which the decision maker finds herself at any particular point in time. It has been shown that if the state

space of the learning model is the set of pure strategies, then there are no reinforcement learning rules which lead to optimality. However, if the state space is taken to be the set of all mixed strategies, then there are rules which lead the decision maker to learn the optimal strategy. This difference arises because the state space provides the only possibility for the decision maker to store information about her past experiences and the memory store provided by a pure strategy state space is too small to deal with a large variety of environments. However, the set of all mixed strategies is sufficiently large. For this case, we have identified a property called *monotonicity* which has been shown to achieve optimality provided the learning rule evolves in small steps. A rule is monotone if the expected change in the probability of the expected payoff maximising action is positive.

The second property refers to the functional form which characterizes monotone learning rules: linearity in payoffs. An intuitive reason for this is that expected payoffs themselves are linear functions of payoffs. Obviously, this property has different implications for the different frameworks in which reinforcement learning has been considered in this thesis. Monotone learning rules are linear in payoffs and the expected movement of the decision maker's state is closely related to replicator dynamics. Monotone imitation rules are linear in payoff difference, displaying a reinforcement component with the observed payoff being an aspiration level.

The final Chapter of the thesis should be viewed as a different exercise. In a population context in which the members of the population are allowed to adjust their behaviour using simple imitation rules, we have investigated individual incentives to adhere to a particular imitation rule by using an equilibrium analysis. Each individual chooses her imitation rule so as to maximise her asymptotic payoff taken as given and fixed the rules used by the remaining members of the population. Once the non-existence of individuals incentives to deviate is assured the analysis of the aggregate evolution of the population can be safely done.

In a sense, this thesis is an axiomatic approach to learning models. However, our approach is not free of drawbacks. We shall point out two of them which are related to our interpretation of *bounded rational* decision-makers.

First, we have characterised a class of learning rules that lead the decision-maker to play in the long run the expected payoff maximising choice in any circumstance. Thus a fully rational decision-maker who does not discount the future would find these decision rules optimal. This seems a serious limitation. We model boundedly rational decision makers as being infinitely patient. We do not wish to argue that this is a realistic assumption. A more satisfactory analysis should incorporate a discount factor strictly less than one, and focus on transitional payoffs rather than exclusively focusing on asymptotic payoffs. Our naive assumption, which leads to a simpler analysis, should be seen as a starting point to a complete analysis of reinforcement learning rules as describing boundedly rational decision makers.

And second, a fully rational decision maker could find many strategies other than those discussed in this thesis which achieve asymptotic expected payoff maximisation. In this thesis we have focused on decision rules which have no memory. Thus, the current (possibly stochastic) behavioral "habit" is the only variable which can provide the decision-maker with information about the past. Note that this memory constraint is *implicitly* included in the reinforcement learning rules, i.e. there is no explicit consideration of it. It therefore seems worthwhile to build the analysis on a more explicit model of memory constraints.

And finally, as a final word, let us say that our work should not be interpreted as aiming to single out "desirable" learning models. Whether agents use learning models which lead to optimality in a large variety of situations is an empirical question. We hope our characterisation of monotone behaviour rules will help to interpret empirical results regarding this question.

Bibliography

[1] Allais, M., Le comportement de l'homme rationel devant le risque, critique des postulates et axiomes de l'ecole Americaine, *Econometrica* 21, 503-546, 1953.

[2] Aso, H. and M. Kimura, Absolute Expediency of Learning Automata, *Information Sciences* 17 (1979), 91-112.

[3] Bendor, J., D. Mookherjee and D. Ray, Aspirations, adaptive learning and cooperation in repeated games, Discussion Paper, Planning Unit, Indian Statistical Institute, New Delhi, 1992. Revised version, June 1995.

[4] Binmore, K., P. Morgan, A. Shaked and J. Sutton, Do people exploit their bargaining power?: An experimental study, *Games and Economic Behaviour* 3 (1991), 295-322.

[5] Börgers, T. and R. Sarin, Learning through reinforcement and replicator dynamics, *Journal of Economic Theory* 77 (1997), 1-14.

[6] Börgers, T. and R. Sarin, Naive reinforcement learning with endogenous aspirations,forthcoming, *International Economic Review.*

[7] Bush R.R, and F. Mosteller, A Mathematical model for simple learning, *Psychological Review* 58 (1951), 313-333.

[8] Bush R.R, and F. Mosteller, Stochastic Models of Learning, New York: John Wiley & Sons, 1955.

[9] Butcher, J.C., *The Numerical Analysis of Ordinary Differential Equations*, Chichester, etc.: John Wiley & Sons, 1987.

[10] Camerer, C., Recent Tests of Generalizations of EU Theories, in *Utility: Theories, Measurement, and Applications*, ed. by W. Edwards. Dordretch: Klower, 1992.

[11] Capen, E.C., R.V. Clapp, and W.M. Campbell, Competitive bidding in high-risk situations, *Journal of Petroleum Technology* 23 (1971), 641-653.

[12] Conlisk, J., Three Variants of the Allais Example, *American Economic Review* 79 (1989), 392-407.

[13] Cooper, R.W., Douglas DeJong, Robert Forsythe, and Thomas Ross, Selection criteria in coordination games: Some experimental results, *American Economic Review* 80 (1990), 218-233.

[14] Corradi, V. and R. Sarin, Continuous Approximations of Stochastic Evolutionary Game Dynamics, mimeo., University of Pennsylvania and Texas A&M University, June 1996.

[15] Cross, J.G., A stochastic learning model of economics behaviour, *Quarterly Journal of Economics* 87 (1973), 239-266.

[16] Easley, D. and A. Rustichini, Choice Without Beliefs, mimeo., Cornell University and Tilburg University, 1997.

[17] Erev, I., and A. Roth, Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria, *American Economic Review* 88 (1998), 848-881.

[18] Estes, W.K., Towards a statistical theory of learning, *Psychological Review* 57 (1950), 94-107.

[19] Estes, W. and J. Straughan, Analysis of a Verbal Conditioning Situation in Terms of Statistical Learning Theory, *Journal of Experimental Psychology*, 47 (1954), 225-234.

[20] Friedman, M., *A Theory of the consumption function*, Princeton, Princeton University Press, 1957.

[21] Flavin, M., The adjustment of consumption to changing expectations about future income, *Journal of Political Economy*, 89 (1981), 974-1009.

[22] Grimmett, G. and D. Stirzaker, *Probability and Random Processes* (second edition), Oxford: Clarendon Press, 1992.

[23] Hall, R. and S. Maskin, The sensitivity of consumption to transitory income: estimates from panel data on households, *Econometrica* 50 (1982), 461-481.

[24] Harless, D. and C. Camerer, The Predictive Utility of Generalized Expected Utility Theories, *Econometrica* 62 (1994), 1251-1290.

[25] Harstad, R.M., Dominant strategy adoption, efficiency, and bidders' experience with pricing rules, mimeo. Virginia Commonwealth U., 1990.

[26] Harrison, G. and K.A. McCabe, Testing non-cooperative bargaining theory in experiments, *Research in Experimental Economics*, R. Mark Isaac, editor, Greenwich, Conn., JAI Press, 1992, 137-169.

[27] Harrison, G. and K.A. McCabe, Expectations and fairness in a simple bargaining experiment. Working Paper B-92-10. September, University of South Carolina, 1992.

[28] Hey, J. and C. Orme, Investigating Generalizations of Expected Utility Theory Using Experimental Data, *Econometrica* 62 (1994), 1291-1326.

[29] Jordan, J., Three problems in learning mixed-strategy equilibria, *Games and Economic Behavior* 5 (1993), 368-386.

[30] Kagel, J.H., R.M. Harstad, and D. Levin, Information impact and allocation rules in auctions with affiliated private values: A laboratory study, *Econometrica* 55 (1987), 1275-1304.

[31] Kagel, J.H., and D. Levin, Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders, *Economic Journal* 103 (1993), 868-879.

[32] Kandori, M., G. Mailath, and R. Rob, Learning, mutation and long run equilibria in games, *Econometrica* 61 (1993), 27-56.

[33] Karandikar, R., D. Mookherjee, D. Ray and F. Vega-Redondo, Evolving Aspirations and Cooperation, *Journal of Economic Theory* 80 (1998), 292-331.

[34] Lakshmivarahan, S., and M.A.L. Thathachar, Absolutely Expedient Learning Algorithms for Stochastic Automata, *IEEE Transactions on Systems, Man and Cybernetics* 3 (1973), 281-286.

[35] Lakshmivarahan, S., and MA.L. Thathachar, Absolute Expediency of Q- and S-Models, *IEEE Transactions on Systems, Man and Cybernetics* 6 (1976), 222-226.

[36] Meybodi, M. and S. Lakshmivarahan, ε-Optimality of a General Class of Absorbing Barriers Learning Algorithms, *Information Sciences* 28 (1982), 1-20.

[37] Modigliani, F. and R. Brumberg, Utility analysis and the consumption function: an interpretation of cross-section data, in Kenneth K. Kurihara (ed.), *Postkeynesian Economics*, New Brunswick, Rutgers University Press (1954), 388-436.

[38] Mookherjee, D., and B. Sopher, Learning Behavior in an Experimental Matching Pennies Game, *Games and Economic Behavior* 7 (1994), 62-91.

[39] Mookherjee, D. and B. Sopher, Learning and Decision Costs in Experimental Constant Sum Games, *Games and Economic Behavior* 19 (1997), 97-132.

[40] Narendra, K. and M. Thathachar, *Learning Automata: An Introduction.* Englewood Cliffs: Prentice Hall, 1989.

[41] Norman, M.F., Some convergence theorems of stochastic learning models with distance diminishing operators, *Journal of Mathematical Psychology* 5 (1968), 61-101.

[42] Norman, M.F., *Markov Processes and Learning Models*, New York and London: Academic Press, 1972.

[43] Ochs, J. and A. Roth, An Experimental Study of Sequential Bargaining, *American Economic Review* 79 (1989), 355-384.

[44] Palomino, F. and Fernando Vega-Redondo, Convergence of Aspirations and (Partial) Cooperation in the Prisoner's Dilemma, forthcoming, *International Journal of Game Theory.*

[45] Prasnikar , V. and A. Roth, Considerations of fairness amd strategy: Experimental data from sequential games, *Quarterly Journal of Economics* (1992), August, 865-888.

[46] Roth, A., and I. Erev, Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term, *Games and Economic Behavior* 8 (1995), 164-212.

[47] Rustichini, A., Optimal Properties of Stimulus-Response Learning Models, mimeo., Tilburg University, 1997.

[48] Samuelson L. and J. Zhang, Evolutionary stability in asymmetric games, *Journal of Economic Theory* 57 (1992), 363-391.

[49] Sarin, R., Learning Through Reinforcement: The Cross Model, mimeo., Texas A&M University, 1995.

[50] Schlag, K., A Note on Efficient Learning Rules, mimeo., University of Bonn, 1994.

[51] Schlag, K., Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits, *Journal of Economic Theory* 78 (1998),130-156.

[52] Shapley, L., Some topics in two-person games. In *Advances in Game Theory*, ed. by M. Drescher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press,1964.

[53] Swinkels, J., Adjustment dynamics and rational play in games, *Games and Economic Behaviour* 5 (1993), 455-484.

[54] Taylor, P. and L. B. Jonker, Evolutionary stable strategies and game dynamics, *Mathematical Biosciences* (1978) 40, 455-484.

[55] Toyama, Y. and M. Kimura, On Learning Automata in Nonstationary Random Environments, *Systems, Computers, Controls* 8 (1977), No.6, 66-73.

[56] van Huyck, B., R. Battalio, and R. Beil, Tacit coordination games, strategic uncertainty, and coordination failure, *American Economic Review* 80 (1990), 234-248.

[57] Young, P., The evolutions of conventions, *Econometrica* 61 (1993), 57-84.