Learning Through Reinforcement and Replicator Dynamics ¤

Tilman Börgers
Department of Economics
University College London
Gower Street
London WC1E 6BT
U.K.

Rajiv Sarin
Department of Economics
College of Liberal Arts
Texas A&M University
College Station, TX 77843-4228
U.S.A.

First Version: October 1993
This Version: October 1995

¤ We are grateful to an associate editor, two referees, Murali Agastya, Ken Binmore, Vince Crawford, Drew Fudenberg, Nick Rau, Max Stinchcombe, and, especially, to Joel Sobel for helpful comments and discussions. Part of this research was undertaken while Rajiv Sarin was visiting the Economics Department of University College London. He thanks the department for hospitality and financial support. Tilman Börgers thanks the Economic and Social Research Council for financial support under research grant R000235526.

Abstract

This paper considers a version of Bush and Mosteller's ([5], [6]) stochastic learning theory in the context of games. We compare this model of learning to a model of biological evolution. The purpose is to investigate analogies between learning and evolution. We find that in the continuous time limit the biological model coincides with the deterministic, continuous time replicator process. We give conditions under which the same is true for the learning model. For the case that these conditions do not hold, we show that the replicator process continues to play an important role in characterising the continuous time limit of the learning model, but that a different effect ("Probability Matching") enters as well.

Journal of Economic Literature Classification Numbers: C72, D83.

Keywords: Games, Learning, Evolution.

1 Introduction

The evolutionary approach to game theory attracts increasing attention. If the word \evolution" is used in a biological sense, then this approach is concerned with environments in which behavior is genetically determined, and strategy selection obtains because carriers of di®erent genes di®er in reproductive ¯t-ness. However, often \evolution" is not intended to be understood biologically. Rather, \cultural evolution," i.e. a learning process, possibly in a population of interacting players, is meant. Implicit is the view that there is an analogy between biological evolution and learning.

There are two levels at which such an analogy can exist. First, it might exist at the level of the individual. Decision makers are usually not completely committed to just one set of ideas, or to just one way of behaving. Rather, several systems of ideas, or several possible ways of behaving are present in their minds simultaneously. Which of these predominate, and which are given less attention, depends on the experiences of the individual. The change which the \population of ideas" in the decision maker's mind undergoes may be analogous to biological evolution.

We can also imagine environments in which individual learning behavior is possibly di®erent from biological evolution (for example because individuals adjust too rapidly, as in the case of best response learning) but in which, at the population level, a process operates which is analogous to biological evolution. Decision makers observe and imitate each other. They talk to and convince each other. These processes may imply that the distribution of ideas and strategies in a population of agents changes over time in a way that is analogous to biological evolution.

In this paper we shall focus on the analogy between learning at the individual level and biological evolution. We are interested in this case because, traditionally, game theory has referred to individual players rather than to populations of players. Also, this analogy seems to have received less attention in the recent literature.[1] We shall construct discrete time models of individual learning and of biological evolution in games. We shall then show that these models, although di®erent in discrete time, exhibit identical, or related behavior, once a continuous time limit is constructed.

In the continuous time limit both models yield the (asymmetric) continuous time replicator dynamics (see [19], [34], [35]) or certain modi¯cations of it. This dynamic process has attracted much interest in the recent game

[1] References to papers which formalise the analogy at the population level are given at the end of this Introduction.

theory literature.[2] It postulates gradual movement from worse to better strategies. It thus contrasts with another important class of dynamic processes in game theory, best response dynamics, which involves instantaneous movement to best replies. The gradual movement postulated by replicator dynamics has often important implications. For example, in games such as the Battle of the Sexes, the quick movements of best response dynamics may prevent convergence to equilibrium while the gradual adjustment of replicator dynamics permits such convergence. On the other hand, if best response dynamics gradually slows down, as in "fictitious play," then there are examples such as Matching Pennies in which (continuous time, asymmetric) replicator dynamics cycles, but fictitious play converges.

When compared to other differentiable dynamic processes, i.e. processes in which the state variables are differentiable functions of time, [3] the continuous time replicator dynamics stands out because it is "aggregate monotonic" in the sense of Samuelson and Zhang [29]. Samuelson and Zhang show that the continuous time replicator dynamics, and certain multiples of it, are the only differentiable processes satisfying aggregate monotonicity. Samuelson and Zhang show that this property implies important facts about the dynamic process, for example, that it eliminates in the long run pure strategies which are strongly dominated by a mixed strategy. [4]

Given that replicator dynamics has a number of distinctive features it is important to investigate possible interpretations of it, i.e. to ask which models might give rise to this dynamics. If replicator dynamics is to be relevant to economics, it is particularly important to investigate interpretations of the replicator process as a learning process. Our paper provides one such interpretation. [5]

We begin the formal parts of this paper in the next section with a very stylized biological model. We consider finite normal-form games. For each player of the given game, there is a continuum size population of individuals. Each individual is genetically programmed to play a pure strategy. Time is discrete, and in each period all individuals are randomly matched in groups, where each group consists of one individual from each population. Each group then plays the game. The payoffs which the individuals receive determine their gross re-

[2] See, for example, the recent special issue (Volume 57 (1992)) of the Journal of Economic Theory.

[3] Note that the continuous time versions of best response dynamics are typically not differentiable.

[4] Further properties of aggregate monotonic dynamics are investigated in Ritzberger and Weibull [27].

[5] Alternative learning interpretations of the replicator dynamics have been obtained by authors who consider the interaction of many learning individuals in large populations. As was mentioned in footnote 1, the relevant work will be discussed at the end of this Introduction.

productive success. We employ a speci¯c assumption about deaths which we shall explain later. It is easily calculated that the evolution of the populations in our model can be described by a variant of the replicator equation in discrete time. If an appropriate continuous time limit is constructed, then the continuous time replicator process obtains.

We then turn to learning. The learning model which we consider is in the tradition of Bush and Mosteller's ([8], [9]) stochastic learning theory. The model concerns several agents playing in discrete time repeatedly the same normal-form game. At each point in time, each player is characterised by a probability distribution over her strategy set which indicates how likely she is to play any of her strategies. Players' choices are described as random because they are a®ected by some unmodelled psychological factors.

The probabilities adjust over time in response to experience. A player's experience consists ¯rstly of the fact that the player herself has chosen a particular strategy, and secondly of the payo® which she has received. Positive payo®s represent reinforcing experiences, which induce a player to increase the probability of the strategy just chosen. For given initial probabilities, a larger payo® induces a larger increase. Negative payo®s cause an analogous reduction in the probability with which a strategy is chosen.

Since Bush-Mosteller learning theory is likely to be less familiar to economists than other learning theories, some comments on the interpretation of the theory, and the motivation for considering this theory, are in order. We begin with interpretational issues, and emphasize ¯rst that payo®s in the Bush-Mosteller learning model are not to be interpreted as von Neumann-Morgenstern utilities, for which, of course, the distinction between positive and negative values is meaningless. Rather, payo®s are simple parametrizations of players' responses to their experiences.

Implicit in the learning model is the assumption that players' responses to their experiences are stable over time. This is not always plausible. Players might, for example, have an \aspiration level" to which they compare their experiences, and this aspiration level itself might adjust in response to players' experiences. We analyse a model which is similar to the model in this paper, but which includes a moving aspiration level, in [5]. The main e®ects which we describe in this paper remain present in the modi¯ed model.[6]

The players in the Bush-Mosteller model respond to very limited information only. This might be because no further information is available, or because the processing of any further information appears so costly relative to the potential gains that players prefer to ignore it. The model thus seems most plausible if agents' behaviour is habitual, and not the result of careful re°ection.

[6] The concluding section contains further details concerning moving aspiration levels.

In economics, the decision how much cash households hold, or the procedures adopted by firms to make routine decisions, might fall into this category. Another economic example of decisions to which Bush-Mosteller theory seems applicable is consumers' choice of brands of everyday items. Indeed, theoretical and empirical marketing research has sometimes modelled consumers' brand choice using Bush and Mosteller's learning theory. [7] This work lends some support to Bush and Mosteller's theory.

There are also experimental situations to which Bush and Mosteller's learning model might be applicable. We think primarily of situations in which subjects' information is very limited. Indeed, a remarkable book by Suppes and Atkinson [33] which concerns learning in game theoretic settings reports such experiments, and these experiments do provide some support for stochastic learning theory. In the more recent experimental literature in economics papers by Mookherjee and Sopher [24] and Roth and Erev [28] have shown that algorithms similar to Bush and Mosteller's learning processes may be successful in explaining learning behavior in economic experiments.

To demonstrate how the learning model is related to the replicator process, we consider first the case in which all payoffs are positive, i.e. all experiences are reinforcing. Experiences differ only in their strength of reinforcement. This case has previously been investigated by Cross ([13], see also [14]). If attention is restricted to this case, we therefore refer to the learning model also as \Cross' learning model."

An obvious difference between the biological process and Cross' learning process is that the learning process is stochastic whereas the biological process is deterministic. However, the processes are related in that the expected motion of the learning process, conditional on any state, is equal to the actual motion of the biological process, conditional on the same state. Since the biological process coincides with a version of the discrete time replicator process, this means, roughly speaking, that the learning process coincides in expected terms with the discrete time replicator process.

The difference between the two models disappears when the continuous time limit is taken. In this limit, also the learning model converges to the deterministic, continuous time replicator process. We prove this result by appealing to a mathematical result due to Norman [26]. The intuition is that, if the continuous time limit is taken, each time interval sees many iterations of the game, and the adjustments which players make between two iterations of the game are very small. Consequently, a law of large numbers applies, and the process becomes deterministic.

To develop further understanding of the relation between the biological

[7] Relevant work is surveyed in parts of Meyer and Kahn [23]. Among the empirical papers are [18] and [20].

model and Cross' learning model it is convenient to reinterpret Cross' learning model as a model of an agent who has simultaneously several contradictory ideas in mind, and who adjusts the weights given to these ideas in response to experience. We shall present such an interpretation in this paper. The interpretation will be based on ideas of Estes' [16] \stimulus sampling" theory of learning. Bush and Mosteller, in Chapter 2 of [9], have interpreted their general model in terms of Estes' theory, and our argument will be similar to Bush and Mosteller's. The reinterpretation of Cross' model is useful because it shows that the intuition for our result can be derived from an analogy between the \population of ideas" in an agent's mind and a population of genetically programmed individuals.

It is important to note that our result about the continuous time limit refers to arbitrary, but ¯nite points in time. It is no longer true if in¯nite time, i.e. the asymptotic behavior of the processes involved, is considered. We shall show that the asymptotic behavior of the biological process in discrete time, and the asymptotic behavior of the learning process in discrete time are quite di®erent from each other, and from the asymptotic behavior of the continuous time replicator model.

If payo®s are permitted to be negative, the continuous time limit of the learning model is characterised by a di®erential equation which is related to the replicator equation but di®erent from this equation. The right hand side of the di®erential equation for the learning process consists of two terms. One of these is of the \replicator type." The second term, however, re°ects an entirely di®erent force. If the second term alone were active, then players would equate the probability with which they choose a strategy with the probability with which this strategy is \successful," i.e. is reinforced. This behavior is often called \probability matching." There is some experimental evidence for behavior of this type (provided that payo®s are \small"; see Siegel [32] and the references quoted therein).

\Probability matching" is often irrational behavior. In decision problems, for example, maximization of expected payo®s requires agents typically to set the probability of one strategy equal to one, and to set the probability of all other strategies equal to zero.

It seems to have been known among psychologists that there is a relation between stochastic learning theory and probability matching. From this perspective, the contribution of our paper is to point out that stochastic learning theory is also related to replicator dynamics, and to show that, in the continuous time limit, the learning process can be decomposed into exactly these two forces.

Literature which is related to this paper includes the previous investigations of Cross' learning process in [13], [31]. Other processes in the Bush-Mosteller

class have been investigated in [8], [9], [21], [22]. All of these authors have focused on asymptotic properties of the process in discrete time. Some progress has been made, but knowledge of these properties is still very incomplete.

The continuous time limit of stochastic learning processes has previously been considered by Norman [26]. Our analysis relies heavily on his mathematical results. Norman used these results to study some special cases of the Bush and Mosteller's learning model which are different from the ones considered here. Also, he was concerned with different interpretational issues. Independent, and sometimes more general versions, of Norman's mathematical results concerning continuous time approximations have been developed in several contributions, for example in [2].

The continuous time limit of a model in which large, but finite populations of agents interact in discrete time has recently been constructed by Boylan [7]. His model differs both from our biological model (since he considers finite populations) and from our learning model (since he considers populations rather than individuals, and since individuals' transition from one \type" to another is deterministic rather than stochastic). However, the formal issues in his and our work are closely related. He employs mathematical techniques which are similar to those used in the references to which we appeal here, such as Norman [26]. Like us, Boylan emphasises the difference between results for finite points in time and asymptotic results. This latter issue is also one of the issues addressed in Boylan [6].

Other work concerning the analogy between learning and biological evolution is due to Binmore and Samuelson [4], Cabrales [11] and Schlag [30]. These papers show how imitation of better strategies in large populations of players can generate the replicator dynamics at the population level. Thus they formalise the second of the two main lines of argument concerning \social evolution" to which we referred at the beginning of this Introduction. This work is complementary to ours.

This paper is structured as follows: Section 2 describes the biological process and its continuous time limit. Section 3 explains the learning process in the case of positive payoffs, i.e. Cross' learning model, and derives its continuous time limit. Sections 2 and 3 together show that, in the continuous time limit, the two processes are identical. In Section 4 we give an intuitive explanation of this result by interpreting the learning model using ideas from Estes' [16] stimulus sampling theory of learning. Section 5 explains why our result does not extend to the infinite time horizon. In Section 6 we generalize the learning model and permit also negative payoffs. Section 7 concludes the paper.

2 The Biological Model

We consider a finite normal-form game with two players.[8] The two players will be indexed by i and will be called R (Row) and C (Column). The feasible strategies of R are: $j \in J \equiv \{1, 2, \ldots, J\}$. The feasible strategies of C are: $k \in K \equiv \{1, 2, \ldots, K\}$. The payoff to player i when R plays j and C plays k is U^i_{jk}. In this section payoffs will indicate the number of offsprings of a player. Hence $U^i_{jk} \in \mathbb{N} \cup \{0\}$ for all i, j and k. We write U^i for the matrix with U^i_{jk} in row j and column k.

There are two populations of players. Each population is of continuum size with total mass 1. Members of population 1 can fill the role of player R. Members of population 2 can fill the role of player C.

The game is played repeatedly. Repetitions are indexed by $n \in \mathbb{N}$. At the beginning of each round each player is characterised by the pure strategy which she is genetically programmed to play. Denote by $p_j(n)$ the proportion of players in population 1 programmed to play strategy j in stage n. Define $p(n) \equiv (p_1(n), \ldots, p_J(n))$. Similarly, denote by $q_k(n)$ the proportion of players in population 2 programmed to play the strategy k in stage n, and define $q(n) \equiv (q_1(n), \ldots, q_K(n))$. We then have $p(n) \in S^{J-1}$ and $q(n) \in S^{K-1}$, where, for any $L \in \mathbb{N}$, we denote by S^L the L-dimensional simplex. We define $S \equiv S^{J-1} \times S^{K-1}$.

In every stage n only a proportion θ (with $0 < \theta < 1$) of players in each population plays the game. All other players remain idle. The individuals who actually play the game are selected randomly from their respective populations. The two selected groups of players are then randomly matched in pairs to play the game.[9] Players play the pure strategies with which they are programmed. After playing, players reproduce. Individuals reproduce on their own, without a partner. The number of offsprings of any individual player is equal to the payoff that the player received when playing the game.

After reproduction, a proportion of each population dies. The number of deaths is such that the total size of each of the two populations remains constant. The individuals who die are randomly selected from all players who have not been born in the current period. Newborns cannot die.

The assumption described in the previous paragraph is logically consistent only if the number of newborns can never be greater than the number of

[8] We restrict attention to the case of just two players to simplify the presentation.

[9] Note that we implicitly assume that random matching schemes for continuum size populations exist. Although this implicit assumption is common in the literature, it is not obvious that it is justified. For countably infinite populations the issue has been investigated by Boylan [6] and Gilboa and Matsui [17], but we know of no corresponding work for continuum size populations.

existing players. To ensure this we assume that the product of ® and the maximal payo® is not greater than one: $®¢\max_{i;j;k}U^i_{jk} \cdot 1$.

We now construct the equation which describes the evolution of the two populations over time. We de¯ne: $¢p_j(n) ´ p_j(n+1) ¡ p_j(n)$ and $¢q_k(n) ´ q_k(n+1) ¡ q_k(n)$. Also, we write e_r for the unit vector with a one in the r-th row, and zeros elsewhere. In the following proposition, and also later in similar contexts, we drop for notational simplicity the required transpose symbols. Proposition 1 follows from straightforward calculations.

Proposition 1 For every $n \in \mathbb{N}$, $j \in J$ and $k \in K$:

$$¢p_j(n) = ®p_j(n)\left[e_j U^R q(n) ¡ p(n)U^R q(n)\right]$$
$$¢q_k(n) = ®q_k(n)\left[p(n)U^C e_k ¡ p(n)U^C q(n)\right]$$

Proposition 1 shows that the proportion of individuals playing a particular strategy grows if and only if this strategy yields higher than average expected payo®. The percentage increase or decrease in the proportion of players playing a particular strategy is equal to a proportion ® of the di®erence between that strategy's expected payo® and the average expected payo®.

The model that we have presented is very similar to one in Chapter 9 of Binmore [3]. An important di®erence is that we have changed the assumption about deaths made in [3].[10] In [3] it is assumed that all individuals, including the newborns, can die. With this assumption, the right hand sides of the formulas in Proposition 1 have to be divided by some denominator.[11] The equations which include this denominator are often called the \discrete time replicator equations" [12]. We have altered the assumption about deaths because this will facilitate the comparison between the model of this section and the learning model of the next section.

As was explained in the Introduction our focus in this paper will be on continuous time limits. To construct the continuous time limit of the biological model of this section we conduct a thought experiment in which, in each \real" time interval, the game is played very often, but the proportion of players who are selected in each round to play is very small. Speci¯cally, we assume that the time interval between two successive stages is of length $0 < \mu \cdot 1$, and that the proportion of active players in each stage is $\mu®$ (where μ is the same constant in both assumptions). We denote the resulting process by $\{(p^\mu(n); q^\mu(n))\}_{n \in \mathbb{N}}$. This process satis¯es Proposition 1 if we replace ® by $\mu®$.

[10] The two other di®erences are that Binmore considers the case of symmetric games, whereas we deal with potentially asymmetric games, and that he assumes that in each round all individuals play, whereas we assume that only a fraction plays.

[11] See p.419 in [3].

[12] for asymmetric two player games.

Since we imagine that the time interval between two repetitions is of length μ, the variables $(p^\mu(n); q^\mu(n))$ describe the state of the process at time μn. We are now interested in the continuous time limit, i.e. in the limit $\mu \to 0$. To obtain the state of the limit process at some time $t \geq 0$ we consider the limit of $(p^\mu(n); q^\mu(n))$ for a sequence of μs and ns with the property that $\mu \to 0$ and $\mu n \to t$.

To describe this limit we need to introduce the "continuous time replicator equation." Let $\hat{p}(t) \in S^{J-1}$ and $\hat{q}(t) \in S^{K-1}$ for all $t \geq 0$. Suppose that $\hat{p}$ and $\hat{q}$ are differentiable functions, and that they satisfy:

$$\frac{d\hat{p}_j(t)}{dt} = \hat{p}_j(t)\left[e_j U^R \hat{q}(t) - \hat{p}(t) U^R \hat{q}(t)\right]$$

$$\frac{d\hat{q}_k(t)}{dt} = \hat{q}_k(t)\left[\hat{p}(t) U^C e_k - \hat{p}(t) U^C \hat{q}(t)\right]$$

for all $t \geq 0$, $j \in J$ and $k \in K$. Then we call $\hat{p}$ and $\hat{q}$ the "solution of the continuous time replicator equation" for initial values $\hat{p}(0)$ and $\hat{q}(0)$. The continuous time replicator equation in the form just described is due to Taylor [34]. [13]

The following proposition says that for $\mu \to 0$ the process constructed in this section is characterised by the continuous time replicator equation.

Proposition 2 Suppose that for all μ: $(p^\mu(1); q^\mu(1)) = (\hat{p}(0); \hat{q}(0))$. Consider some t with $0 \leq t < 1$ and assume $\mu \to 0$ and $n\mu \to t$. Let $\hat{p}$ and $\hat{q}$ be the solution of the continuous time replicator equation for initial values $\hat{p}(0)$ and $\hat{q}(0)$. Then $(p^\mu(n); q^\mu(n)) \to (\hat{p}(t); \hat{q}(t))$.[14]

Proof: This follows from a theorem that is well-known in numerical mathematics because it underlies "Euler's method" for the numerical solution of ordinary differential equations.[15] The theorem is stated as Theorem 203A in [10]. The theorem uses an assumption which refers to the function $v : S \to \mathbb{R}^{J+K}$ which is defined by:

$$v(p; q) = \left(\frac{\Delta p^\mu(n)}{\mu}; \frac{\Delta q^\mu(n)}{\mu}\right)$$

[13] The continuous time replicator equation was first introduced in [35] for symmetric two player games. The version that we use here was introduced later for asymmetric two player games.

[14] The reference quoted in the proof of Proposition 2 also shows that under the assumptions of this result $|p_j^\mu(n) - \hat{p}_j(t)|$ converges to zero at least as fast as μ.

[15] Euler's method solves ordinary differential equations by discretizing them.

where it is assumed that $p^\mu(n) = p$ and $q^\mu(n) = q$. The theorem requires that this function is Lipschitz. Since v is polynomial on a compact domain this is satisfied.

The conclusion of the theorem is that, in the continuous time limit, (p,q) converges to the solution of the differential equation $(dp/dt, dq/dt) = v(p,q)$ with initial value $(\hat{p}(0), \hat{q}(0))$, evaluated at time t. Thus the assertion follows from Proposition 1, where $v(p,q)$ was calculated.

Q.E.D.

3 Cross' Learning Model

The game that we consider in this section has the same set of players and the same sets of strategies as before. However, payoffs now play a different role, and hence we introduce a new notation for them. We write Π^i_{jk} for the payoff to player i when R plays j and C plays k. In this section, payoffs will be interpreted as "strengths of reinforcement." We shall assume that they satisfy: $0 < \Pi^i_{jk} < 1$ for all i, j and k. We explained already in the Introduction that we focus in this section and in the following two sections on the case that all payoffs are non-negative, i.e. that there is no deterrence. It will become clear below why we need, in addition, that payoffs are not greater than one. Without this assumption we would not be able to give payoffs the interpretation used below. The fact that the two inequalities are strict rather than weak will only be used in the proof of Proposition 5 below.[16] We write Π^i for the matrix with Π^i_{jk} in row j and column k.

The ultimate purpose of this section is to show that, in the continuous time limit, the model of this section and the model of the previous section coincide. For this we shall need a relation between the payoffs in the two models. We shall make throughout the following assumption: $\alpha U^i_{jk} = \Pi^i_{jk}$ for all i, j and k. Once it is noted that, in the model of the previous section, the "effective" payoffs were αU^i_{jk}, it is clear that this is the relation that we need.

In contrast to Section 2, we shall now assume that the game is played not by two populations but by two individual players: $i = R, C$. These players play the game repeatedly, and, as before, the iterations of the game are indexed by $n \in \mathbb{N}$. At the beginning of stage n each player i is characterised by the probability with which she plays each of her strategies. For player R these probabilities are $P(n) \equiv (P_1(n), \ldots, P_J(n))$. For player C they are $Q(n) \equiv (Q_1(n), \ldots, Q_K(n))$. We call $P(n)$ (resp. $Q(n)$) the "state" of player R (resp. C) at stage n. We define $S(n) \equiv (P(n), Q(n))$. Thus $S(n)$ can be called

[16] In the verification of Norman's condition (H8).

the \state of the game" at stage n. In our model P(n), Q(n), and S(n) will be random variables. We write p(n), q(n), and s(n) for realisations of these variables.

The set of all possible states for player R (resp. C) is S^{J-1} (resp. S^{K-1}). The set of all possible states of the game is $S \times S^{J-1} \times S^{K-1}$. To simplify notation we identify the element of player i's state space that allocates all probability to one of i's strategies with that strategy itself. In other words, the sets of vertices of the two players' state spaces are identified with J and K.

We assume that, at each stage, a player observes only the strategy that she plays, and the payoff that she receives. Players hence don't observe the other players' strategies. After making their observations, players update their states. If player R played strategy j in the n-th repetition of the game, and if she received payoff U^R_{jk}, then she updates her state by taking a weighted average of the old state, and of the unit vector which puts all probability on strategy j. The weight that is put on the unit vector is equal to the payoff U^R_{jk}. Formally, this means:

$$\begin{aligned} P_j(n+1) &= U^R_{jk} + (1 - U^R_{jk})\, P_j(n) \\ P_{j'}(n+1) &= (1 - U^R_{jk})\, P_{j'}(n) \qquad \text{for all } j' \neq j \end{aligned}$$

Player C updates Q(n) in an analogous manner. Observe that the above formula is meaningful only if $U^R_{jk} \leq 1$. This is why we introduced this assumption earlier.

For given initial random variables (P(1); Q(1)) the above equations define a stochastic process $\{P(n); Q(n)\}_{n \in \mathbb{N}}$. We refer to this process as \Cross' learning process."

Suppose that players have reached the n-th repetition of the game, and that the current state of the game is s. Conditional on this, the state in period n+1 is still a random variable. We want to describe the expected movement of the state. We define: $\Delta P_j(n) \equiv P_j(n+1) - P_j(n)$ and $\Delta Q_k(n) \equiv Q_k(n+1) - Q_k(n)$. We denote by E[::: | S(n) = s] the expected value of the random variable (...) conditional on the state of the game in stage n being s. The following result follows from straightforward calculations.

Proposition 3 For all $n \in \mathbb{N}$, $s \in S$, $j \in J$ and $k \in K$:

$$\begin{aligned} E[\Delta P_j(n) \mid S(n) = s] &= p_j \left[e_j U^R q - p U^R q \right] \\ E[\Delta Q_k(n) \mid S(n) = s] &= q_k \left[p U^C e_k - p U^C q \right] \end{aligned}$$

Propositions 1 and 3 together show that for given current state the expected movement of the two players in the learning model is exactly the same as the actual (deterministic) movement of the two populations in the biological model. The two processes di®er only in that the learning process is stochastic whereas the biological process is deterministic.

Next, we construct the continuous time limit of the learning process and show that in this limit expected motion and actual motion of the players' states coincide. We construct the continuous time limit in a way that is analogous to the previous section. We hence imagine again that the amount of \real" time that passes between two repetitions of the game is given by a number μ with $0 < \mu \leq 1$. After each repetition of the game, the players adjust their states by μ times what it was so far. Formally, we replace the adjustment formulas given earlier by:

$$\begin{aligned} P_j^\mu(n+1) &= \mu U_{jk}^R + (1 - \mu U_{jk}^R)\, P_j^\mu(n) \\ P_{j'}^\mu(n+1) &= (1 - \mu U_{jk}^R)\, P_{j'}^\mu(n) \qquad \text{for all } j' \neq j \end{aligned}$$

where we introduce the upper index μ to indicate that we are now referring to a modi¯ed process. An analogous formula applies to $Q^\mu(n+1)$. We obtain a process $\{(P^\mu(n), Q^\mu(n))\}_{n \in \mathbb{N}}$, provided that we specify the initial random variables $(P^\mu(1), Q^\mu(1))$. This process satis¯es Proposition 3 if one multiplies the right hand sides of the equations in Proposition 3 by μ.

Since we imagine the time interval between repetitions to be μ the random variable $S^\mu(n)$ describes the state of the process at time $n\mu$. As in Section 2 we are interested in the limit $\mu \to 0$. We obtain the state of the limit process at some time $t \geq 0$ by investigating the limit of $S^\mu(n)$ for any sequence of μs and ns with the property that $\mu \to 0$ and $n\mu \to t$.

Proposition 4 Suppose that for all μ: $(P^\mu(1), Q^\mu(1)) = (\hat{p}(0), \hat{q}(0))$ with probability 1. Consider some t with $0 \leq t < 1$ and assume $\mu \to 0$ and $n\mu \to t$. Let $\hat{p}$ and $\hat{q}$ be the solution of the continuous time replicator equation for initial values $\hat{p}(0)$ and $\hat{q}(0)$. Then $S^\mu(n)$ converges in probability to $(\hat{p}(t), \hat{q}(t))$.

Proof: We use Theorem 1.1 in Chapter 8 of Norman [26]. This theorem concerns the continuous time limit of discrete time Markov processes with in¯nite state spaces. The processes to which we apply this theorem are the processes $\{S^\mu(n)\}_{n \in \mathbb{N}}$. Our assertion follows immediately from parts (A) and (B) of Norman's theorem. Therefore, it is su±cient to verify that the assumptions of the theorem are satis¯ed. This is trivially true for Norman's assumptions (a.1)-(a.3).

Norman's assumptions (b.1)-(b.3) refer to the function $v : S \to \mathbb{R}^{J+K}$ which is de¯ned by:

$$v(p;q) \equiv E\left[\frac{\Delta S^{\mu}(n)}{\mu} \mid S^{\mu}(n) = (p;q)\right]$$

Norman's assumption (b.4) refers to the function $w : S \to \mathbb{R}^{(J+K)^2}$ which is defined by:

$$w(p;q) \equiv Var\left[\frac{\Delta S^{\mu}(n)}{\mu} \mid S^{\mu}(n) = (p;q)\right]$$

(Here, we denote by $Var[... \mid S^{\mu}(n) = s]$ the variance-covariance matrix of the random variable (...) conditional on the event that the state of the game in stage n is s.) Norman's assumption (c) refers to the function $r : S \to \mathbb{R}$ which is defined by:

$$r(p;q) \equiv E\left[\left(\frac{\Delta S^{\mu}(n)}{\mu}\right)^3 \mid S^{\mu}(n) = (p;q)\right]$$

(Here, if $x \in \mathbb{R}^{J+K}$, we define: $|x|^3 = \sum_{i=1}^{J+K} |x_i|^3$.)

Norman's condition (b.1) requires v to be differentiable, condition (b.2) requires the derivative of v to be bounded, and condition (b.3) requires the derivative of v to be Lipschitz. Condition (b.4) requires w to be Lipschitz. Condition (c) requires r to be bounded from above. In our case, all functions involved are obviously polynomial (in the case of r: piecewise polynomial and continuous) functions with compact domains, and hence all of Norman's assumptions are satisfied.

The conclusion of Norman's Theorem is that in the continuous time limit the state variable S converges in probability to the solution of the differential equation $ds/dt = v(p;q)$ with initial value $(\hat{p}(0);\hat{q}(0))$, evaluated at time t. Thus, the assertion of Proposition 4 follows from Proposition 3 where $v(p;q)$ was calculated.

Q.E.D.

In words, Proposition 4 says that, if μ is small, and if $n\mu$ is close to $t \geq 0$, then, with high probability, $S^{\mu}(n)$ will take a value that is close to the solution of the continuous time replicator equation at time t.[17] [18] The intuition for

[17] Using results in [26] it can also be shown that, under the assumptions of Proposition 4, for every $\varepsilon > 0$ and every $j \in J$ the probability $Pr(|P_j^{\mu}(n) - \hat{p}_j(t)| \geq \varepsilon)$ converges to zero at least as fast as μ. The analogous statement holds for every pure strategy of player C.

[18] A stronger version of Proposition 4 would assert that, as μ tends to zero, the distribution of the polygonal curve connecting the points $(n\mu; S^{\mu}(n))$ (where $n\mu \leq t$) converges weakly to the probability distribution which gives probability one to the solution of the replicator equation. Although we believe this result to be true, we don't deal with it here since its statement and proof would involve additional complications.

this result is that frequent play and slow movement ensure that a law of large numbers applies, and therefore actual and expected movement of the learning process coincide. Propositions 2 and 4 together show that the learning model and the biological model coincide when the continuous time limit is considered, and attention is restricted to some ¯nite point in time. Thus, they demonstrate an analogy between learning and biological evolution.

4 Stimulus Sampling

The learning model of Section 3 postulates a particular behavior without giving a description of the internal structure of players that gives rise to this behavior. This is also true for Bush and Mosteller's general theory of learning, of which the model in Section 3 is a special instance. Proceeding like this has the advantage that the formal framework admits several di®erent interpretations. On the other hand, the general theory is too abstract to suggest intuitions. For this reason Bush and Mosteller presented in Chapter 2 of [9] a speci¯c interpretation of their model. It was based on ideas from Estes' [16] stimulus sampling theory of learning. In this section we give a similar interpretation that applies to our context.[19] Then we use this interpretation to develop intuition for the results of the previous sections.

Suppose that each player when making a choice is subject to many stimuli. Speci¯cally, for each player there is a continuum of such stimuli. The total mass of this continuum is one. Each stimulus is programmed to suggest one particular choice to the player, but di®erent stimuli may suggest di®erent choices. The player chooses a strategy by selecting randomly one of these stimuli.

Once a player has chosen a strategy, and experienced a payo®, some randomly selected stimuli are re-programmed to suggest the particular strategy that the player has just taken. The measure of the set of re-programmed stimuli is equal to the payo® which the player experienced.

A straightforward calculation shows that this model of players' behavior generates exactly the process that we described in Section 3. Thus, the model provides one possible interpretation of the framework of Section 3.

We can re-phrase this interpretation of Cross' learning model in biological language to obtain a biological model that is exactly equivalent to the learning model. For this we identify the two continua of stimuli in°uencing each of the two players with two continuum size populations of agents with genetically inherited strategies. The randomly selected stimulus which each player follows

[19] We shall make some simpli¯cations in comparison to Bush and Mosteller's argument.

can then be interpreted as a randomly selected individual who is playing the game \on behalf of her population."

After the two individuals have interacted, they reproduce. Each of the two representative individuals has o®springs which are of positive measure in comparison to the population from which they come, and this measure is equal to the payo® received in the game. Deaths occur in the way which was also postulated in Section 2. It is now evident that this pseudo-biological model is exactly equivalent to the learning model.

It is also clear how this biological model di®ers from the biological model of Section 2. Whereas in Section 2 we assumed that the proportion of each population that plays and reproduces is of positive measure, and hence of continuum size, in the model that we have just constructed this is done by two randomly selected, representative individuals. Also, in the model of Section 2, the o®springs of any particular individual are of measure zero, whereas in the model just described the two representative individuals have sets of o®springs of positive measure.

As a consequence, the pseudo-biological model is a stochastic version of the biological model. In expected terms the two models are identical, as was shown in Propositions 1 and 3. By Propositions 2 and 4, the di®erence between the models disappears, and both models become deterministic, if a continuous time limit is taken, and if attention is restricted to a ¯nite point in time.

5 Asymptotic Analysis

The convergence results of Propositions 2 and 4 apply to any point in time $t < 1$. They have no implications for the asymptotic behavior, for $t \to 1$, of the discrete and continuous time processes. In fact, the asymptotic behavior of the discrete time processes may be very di®erent from that of the continuous time process. Moreover, the asymptotic behavior may be di®erent for the two discrete time processes that we consider. These di®erences may arise even for arbitrarily low values of μ.

To show these points, we ¯rst state a result concerning the asymptotic behavior of the discrete time learning process. The result says that, with probability 1, the learning process will converge to a situation in which both players play some pure strategy with probability 1. This result holds for all possible speeds of learning.

Proposition 5 For all $\mu > 0$ and for all initial variables $(P^\mu(1); Q^\mu(1))$ with probability 1 the sequence $f(P^\mu(n); Q^\mu(n))g_{n2IN}$ converges, and its limit is in

$J \times K$.[20]

Proof: Taking μ to be given and fixed, we use Theorem 2.3 of Norman [25]. The first sentence of that theorem says that under certain assumptions a stochastic process will converge with probability one to one of its absorbing states. In our learning model it is clear that the set of absorbing states is $J \times K$. Thus, our assertion follows if the assumptions of Norman's theorem are satisfied. The conditions which Norman labels (H1)-(H6) are merely technical conditions which are easily verified.

Condition (H7) requires in our context the following: Consider any period n. Let $s(n)$ and $s'(n)$ be two possible states of the two players at the beginning of period n. Consider also some fixed strategy pair (j,k). Denote by $s(n+1)$ the state that is reached if the initial state was $s(n)$ and (j,k) was played in period n, and let $s'(n+1)$ be the state that is reached if the initial state was $s'(n)$ and (j,k) was played in period n.[21] Then $d(s(n+1), s'(n+1)) \leq d(s(n), s'(n))$, where d denotes Euclidean distance. In words the requirement is hence that, with probability 1, the updating process acts as a contraction. A straightforward calculation shows that this requirement is satisfied in our model.

Note that the inequality in the above requirement is weak. Norman's assumption (H8) requires that in certain cases the inequality is strict. However, in our model, the inequality is always strict, so that also (H8) is satisfied.

Norman's assumption (H9) is not required for the result that we are applying here. Assumption (H10) can be phrased as follows: For any initial state s, the closure of the set of states that can be reached from s with positive probability within finite time, contains at least one of the absorbing states. To see that this is true, notice that for any initial state s and for every player i there is a strategy of i such that the probability that this strategy is played m times is positive for all $m \in \mathbb{N}$. Playing the same strategy any finite number of times will, however, generate a sequence of states that converges to an absorbing state.

Q.E.D.

We now compare the asymptotic behavior described in Proposition 5 to the asymptotic behavior of replicator dynamics. It is well known that there are many games in which replicator dynamics does not converge to a pure strategy outcome (see Section 17 of [19]). Thus, the asymptotics of the learning

[20] One can also prove that, for any completely mixed starting point, every element of $J \times K$ has a positive probability of being the limit of $(P^{\mu}(n), Q^{\mu}(n))_{n \in \mathbb{N}}$. This can be shown using the methods of Section 7.2 of [9]. We are grateful to Nick Rau for this observation.

[21] Of course, the probability with which (j,k) is played will depend on the state at the beginning of period n. However, this does not matter for the following argument.

process may be very different from the asymptotics of replicator dynamics. Mathematically speaking, the learning process converges pointwise, but not uniformly on the complete real line, to continuous time replicator dynamics.

We give an example that illustrates this point, and that also shows that the asymptotics of discrete time replicator dynamics may be different both from the asymptotics of continuous time replicator dynamics, and from the asymptotics of discrete time learning. The example is a version of "Matching Pennies." [22]

	L	R
T	0.2,0.8	0.8,0.2
B	0.8,0.2	0.2,0.8

Example 1

Denote by p_1 (resp. q_1) the probability with which player R (resp. player C) chooses "T" (resp. "L"). It is well-known that in this game the continuous time replicator process cycles along the level curves of $p_1(1-p_1)q_1(1-q_1)$. The phase diagram of the process is described by the unbroken lines in Figure 1.

By Proposition 5 the learning process will converge with probability one to one of the corners of the unit square. The phase diagram in Figure 1 thus illustrates the difference between the asymptotic properties of the learning process and of continuous time replicator dynamics.

For the discrete time replicator process Proposition 1 implies that, at any point in the phase diagram, the direction of movement of the process in discrete time is the same as the direction of movement in continuous time. Thus, every step of the discrete time process goes into a direction that is tangential to the trajectory of the continuous time process. This is illustrated by arrows in Figure 1. The arrows show that at each step of the discrete time replicator process the value of $p_1(1-p_1)q_1(1-q_1)$ decreases, provided that we don't start in the equilibrium point $(0.5; 0.5)$. It can also be shown that the discrete time replicator process in this game will not converge, unless it starts in $(0.5; 0.5)$.[23] We can conclude that any trajectory of the discrete time process which does not start in $(0.5; 0.5)$ will asymptotically approach the boundaries of the unit square without converging to any point on this boundary. Thus, in this example, the discrete time replicator process behaves asymptotically

[22] The payoffs in this example can be interpreted as either the values πU^i_{jk} of Section 2, or as the values U^i_{jk} of Section 3.

[23] We don't give a formal proof. But, intuitively, it is straightforward to see that there cannot be any limit point in which any of the two probabilities is interior. This leaves the corners of the unit square as possible limit points. But in a neighbourhood of a corner, the movement of the replicator process is always away from the corner point.

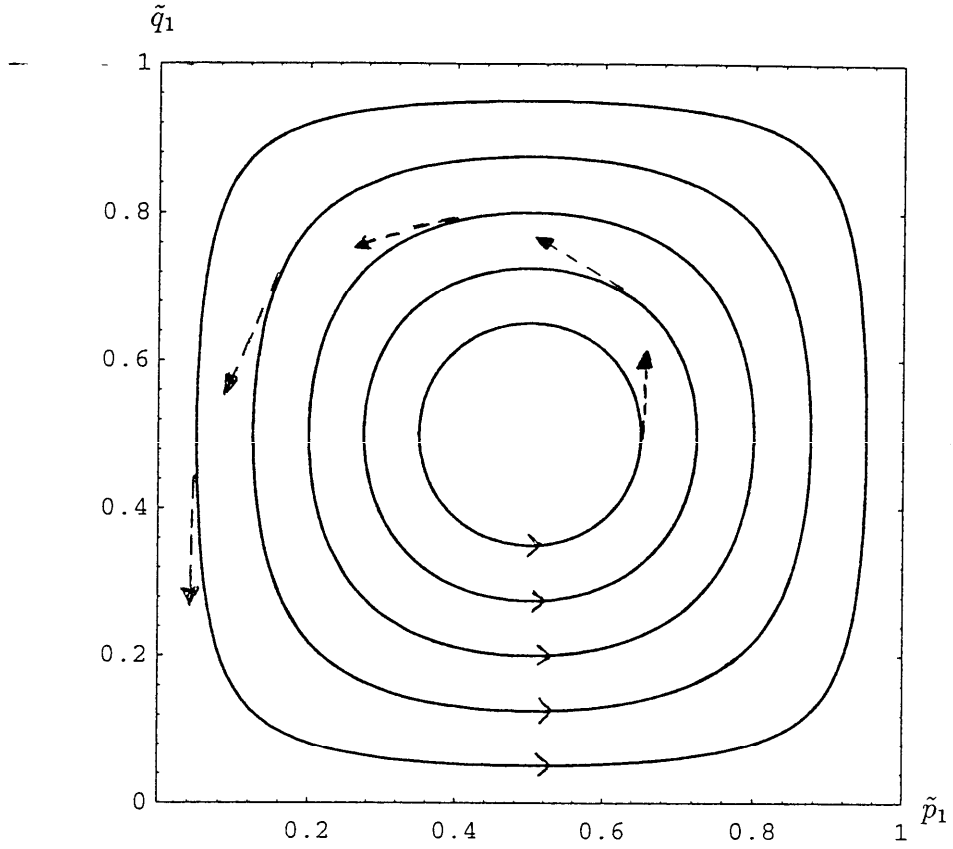

Figure 1

quite differently from the continuous time replicator process. This example also shows how the learning process and the biological process may have different asymptotics. Note that Example 1 is robust under perturbations of payoffs. This can be seen from the discussion of examples of this type in Section 17 of Hofbauer and Sigmund [19].

That discrete and continuous time replicator dynamics may have different asymptotics has been noted by a number of authors (most of whom use a slightly different version of discrete time replicator dynamics): [12], [15], [29] or [36]. The idea to illustrate this possibility geometrically as in Figure 1 appears first in [1]. The possibility that a discrete time, stochastic version of the replicator process is absorbed in a randomly selected corner of state space while the continuous time replicator process cycles along closed curves appears first in Section 2 of [6] where an example similar to our Example 1 is presented. Boylan's discrete time process differs from ours in that it describes stochastic evolution in a finite population rather than learning.[24]

[24]The possibility of differing asymptotics of stochastic discrete time models and related deterministic continuous time models is also one of the issues addressed in [7].

6 Negative Payoffs

So far, our analysis has relied on the assumption that all payoffs are positive. We now extend the analysis to the case that some payoffs are negative. As before we maintain the assumption that the absolute value of payoffs is greater than zero and less than one. If player R, say, chooses a strategy j and receives a payoff $U^R_{jk} > 0$, then she updates her strategy as before. If she receives a payoff $U^R_{jk} < 0$, then she takes probability away from strategy j and shifts it to other strategies.

For reasons which will become clear below, we shall discuss two different specifications of how, in the case of negative payoffs, probability which is taken away from one strategy is re-allocated to the other strategies. The first specification is that the probability is re-distributed among the remaining strategies in proportion to their old probabilities. We shall call this specification "proportional updating." Formally, this is defined by:[25]

$$P_j(n+1) = (1 - |U^R_{jk}|)\, P_j(n)$$

$$P_{j'}(n+1) = P_{j'}(n) + |U^R_{jk}|\, P_j(n) \frac{P_{j'}(n)}{1 - P_j(n)} \quad \text{for } j' \neq j$$

Player C updates Q(n) in an analogous manner.

Observe that, although the formula which defines proportional updating looks different from the formula which applies in the positive payoff case, proportional updating is actually exactly symmetric to the updating behaviour with positive payoffs. If payoffs are positive, probability is added to the strategy just played, and it is taken away from all others in proportion to their current probabilities. If payoffs are negative, probability is taken away from the strategy just played, and is added to the other strategies in proportion to their current probabilities. Moreover, the amount of probability added resp. taken away depends in both cases in exactly the same way on the absolute value of the payoff.

Besides "proportional updating" we shall also consider "random updating." Random updating differs from all updating rules considered so far in that P(n + 1) is a random variable even if one conditions on P(n), the strategy j and the payoff U^R_{jk}. With random updating the probability which player R takes away from strategy j is assigned to a single alternative strategy j'. This strategy j' is randomly selected, whereby each strategy j' has a probability of being selected which is proportional to the probability with which it is currently played. Formally, random updating is defined by the assumption that for every strategy $j' \neq j$ there is a probability $P_{j'}(n)/(1 - P_j(n))$ that the new state of player R is:

[25] We use the notation of Section 3.

$$
\begin{aligned}
P_j(n+1) &= (1 - \sum_{k \neq j} \sigma^R_{jk}) P_j(n) \\
P_{j'}(n+1) &= P_{j'}(n) + \sigma^R_{jk} P_j(n) \quad \text{for } j' \neq j \\
P_k(n+1) &= P_k(n) \quad \text{for } k \neq j, j'
\end{aligned}
$$

Player C updates Q(n) in an analogous manner.

Notice that proportional updating and random updating differ only if a player has more than two strategies. Notice also that, even if a player has more than two strategies, the expected change in this player's strategy, conditional on any state s(n), is the same under the two updating rules. Therefore, the two updating rules give rise to the same continuous time limit.

Because proportional updating is symmetric to the case of positive payoffs, it might appear to be the specification which we should prefer. We have introduced random updating nevertheless because it facilitates our interpretation of the continuous time differential equation below.

To characterise the continuous time limit of the two models, one can no longer apply the theorem of Norman quoted in Section 3, because the functions v, w and r referred to in that theorem need no longer have the regularity properties required for the theorem. If a player has at least three pure strategies, then these functions may have discontinuities in those states in which this player plays a pure strategy. The discontinuities result from the assumption that probability which is taken away from one strategy is redistributed among the remaining strategies [26] in proportion to their current probability. If all other strategies currently have very small probabilities, then even a small change in current probabilities may lead to a large change in the expected updated probabilities.

Fortunately, the discontinuities occur only on the boundary of the state space. The functions v, w and r are well-behaved on any compact subset of the interior of the state space. Moreover, if the learning process starts in the interior of the state space, if only a finite time interval is considered, and if the process is close to the continuous time limit, then the process will stay with high probability within a compact subset of the interior of the state space. We believe that we can show that this is sufficient for the continuous time limit to have the properties asserted by Norman. We omit the formal proof of this, though, since it would make this paper much longer, and would change the emphasis of the paper.

The differential equation which we obtain in the continuous time limit is related to, but not identical to the replicator equation. Without loss of

[26] Deterministically or stochastically.

generality we state it only for the probability of some strategy j of player R. We first need additional terminology and notation. This terminology and notation will refer to the model with random rather than proportional updating.

We shall say that strategy j receives a "benefit" if an event occurs which leads to an increase in j's probability. In the random updating model strategy j receives a benefit if either j is played and a positive payoff is received, or if some other strategy is played, a negative payoff is received, and strategy j is selected as the strategy to which probability is re-allocated. If strategy j receives a benefit, we shall also say that it is "successful". Also, we shall call the absolute value of the payoff received in this event the "size of the benefit for strategy j."

Define $|U^R|$ to be the matrix of absolute values of player R's payoffs. Hence this matrix represents the size of all potential "benefits." Define for every $j \in J$ the matrix U_j^{R*} to be the matrix which is obtained from U^R if in row j all negative entries are replaced by zeros whereas all positive entries are left unchanged, and in rows $j' \neq j$ all positive entries are replaced by zeros whereas all negative entries are replaced by their absolute values. U_j^{R*} thus describes the size of potential benefits for strategy j.

Next, we introduce some notation which refers to the case that the players are in some particular state s. For simplicity, we suppress in the notation the dependence on s. We write p_j^* for the J-dimensional vector the j-th entry of which is p_j and, for $j' \neq j$, the j'-th entry of which is $p_{j'}(p_j/(1 - p_{j'}))$. Roughly speaking, this vector indicates the probability with which playing any particular strategy in J leads to a benefit for strategy j, provided that a positive (if j is played) resp. a negative (if $j' \neq j$ is played) payoff is received.

We define moreover q_j^* to be the J-dimensional vector the j'-th entry of which is the probability with which player C plays a strategy which leads to a non-zero entry in the j'-th row of matrix U_j^{R*}, if such an entry exists, and which has zeros elsewhere. The total probability with which strategy j is successful is hence π_j^*, the vector product of p_j^* and q_j^*. Conditional on this event occurring, the expected size of the benefit is $U_j^{R*} \cdot (p_j^* U_j^{R*} q)/\pi_j^*$.

With this notation, the continuous time limit of the learning process is given by:

$$\frac{dp_j}{dt} = p_j(U_j^{R*} - p \cdot |U^R| \cdot q) + U_j^{R*}(\pi_j^* - p_j)$$

The proof is a simple calculation which we omit.

On the right hand side of the above differential equation, the first term is analogous to the right hand side of the replicator equation. However, notice that this term now refers to "benefits" rather than "payoffs." Clearly, when negative payoffs are allowed, it is "benefits" not "payoffs" which matter.

The sign of the second term is the same as the sign of $(\pi_j^* - p_j)$. Hence it is positive if the probability of strategy j being successful is bigger than the probability of strategy j being played, and it is negative otherwise. If this term alone were active, strategy j would hence be played with exactly the same probability with which it is successful. In more special contexts[27] behaviour that equates the probability with which a strategy is played and the probability with which it is successful has been called "probability matching" by psychologists (see, for example, [32], and the references quoted there). Therefore, we say that the second term in our differential equation represents the "probability matching force."

"Probability matching" is in most cases irrational behavior. Expected payoff maximisation usually requires one strategy to be chosen with probability 1, even if that strategy's probability of success is not equal to 1. Thus, in the case that payoffs may be negative we find that players' behavior is partly irrational.

Notice that, if we multiply out the products on the right hand side of the above equation, the first term cancels against the fourth term. Nevertheless it is more natural to write the equation in the above form, because this form reveals more clearly the two forces present in the dynamic process.

We emphasised earlier that the definitions of the variables entering the above differential equation are derived from the random updating model, not from the proportional updating model. The main reason why this matters is that only in the random updating model the probabilities π_j^* add up to one. In the proportional updating model several strategies may be "successful" at the same time, and hence the sum of the success probabilities may be larger than one. It then no longer makes intuitive sense to say that agents are trying to match choice and success probabilities.

To obtain further insight into the above equation, we shall now describe two extreme cases. The first case will be such that the probability matching term in our differential equation vanishes and only the replicator term remains. In the second case the reverse will be true.

The first case is simply the case which we considered in the previous sections, i.e. the case in which all payoffs are positive. In that case a strategy is successful if and only if it is played. Therefore the two probabilities π_j^* and p_j are identical. Thus the probability matching term vanishes, and only the replicator term remains. Moreover, the distinction between "benefits" and "payoffs" becomes void. Hence the replicator term in the above equation is just the same as the conventional replicator term.

The second case is the case in which all payoffs are of equal absolute value, but some are positive and some are negative. In this case the expected benefit,

[27] See Example 2 below.

conditional on some strategy j being successful, and the expected bene¯t of all strategies, are identical and equal to the absolute value of payo®s. Therefore the replicator term vanishes. Behaviour in the continuous time limit is hence determined by the probability matching term only. Psychologists, when investigating probability matching, have typically referred to situations of this type. A typical example is Example 2.

	α_1	$1-\alpha_1$
T	0.5	-0.5
B	-0.5	0.5

Example 2

Example 2 is a one agent decision problem rather than a game. Player R chooses between T and B. The columns in the middle and on the right denote states of nature which occur with probabilities α_1 and $1 - \alpha_1$ respectively. If the ¯rst state occurs, strategy T is successful. Otherwise, B is successful. [28]

Denote by p_1 the probability with which player R chooses T. The continuous time equation for p_1 specializes to:

$$\frac{dp_1}{dt} = 0.5(\alpha_1 - p_1)$$

Obviously, for all initial values, the solution of this equation will converge for $t \to \infty$ to α_1. The model thus predicts in the long run pure probability matching by player R.

Next, we give an example of a 2-player game which is of the same type as Example 2 in that only the probability matching term, but not the replicator term matter. The example is a version of "Matching Pennies". We give this example because it is interesting to compare it with our earlier version of "Matching Pennies", Example 1. By comparison to Example 1, the following game is a more conventional version of matching pennies.

[28] Formally the example ¯ts into our 2 player framework if one supposes that player C chooses the "state of nature", also that C chooses among these states with initial probabilities identical to those of "nature" in the decision problem, and that all payo®s of player C equal zero. Player C will then stick forever to her initial choice probabilities. Player C thus acts as "nature."

	L	R
T	-0.5,0.5	0.5,-0.5
B	0.5,-0.5	-0.5,0.5

Example 3

If we denote by $\tilde{p}_1$ the probability with which player R chooses T, and by $\tilde{q}_1$ the probability with which player C chooses L, then the continuous time equations are:

$$\frac{d\tilde{p}_1}{dt} = 0.5(1 - \tilde{q}_1 - \tilde{p}_1)$$

$$\frac{d\tilde{q}_1}{dt} = 0.5(\tilde{p}_1 - \tilde{q}_1)$$

Since all payoffs are of equal absolute value, these equations contain only probability matching expressions. Figure 2 shows the phase diagram for these two equations. Unlike in the case of replicator dynamics in Figure 1, there are now no cycles and the mixed strategy Nash equilibrium is globally asymptotically stable.

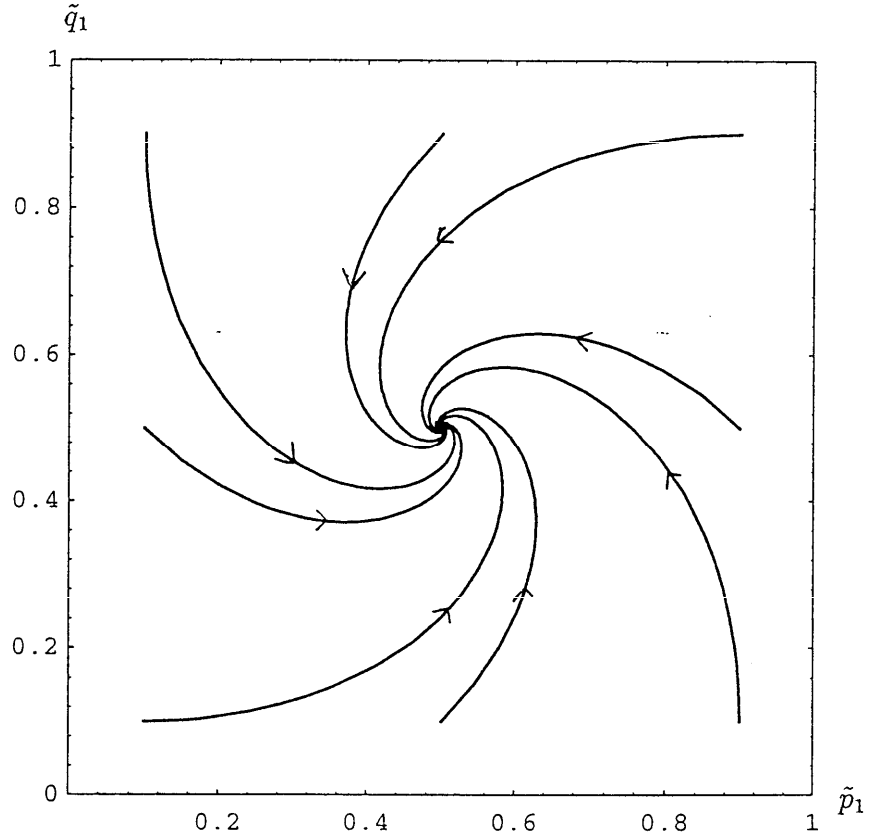

Figure 2

We should emphasise that it is accidental that the limit point of the process in Figure 2 is a Nash equilibrium. In general, probability matching is irrational, and therefore the limit points of our learning process will not be Nash equilibria. In Example 3 it happens that in the Nash equilibrium probability matching and expected payo® maximisation coincide.

We ¯nally give an example in which there are three strategies. We give this example in order to illustrate the role of \random updating" in our theory. The example is, like Example 2, a one player decision problem under risk, not a game.

	α_1	$1-\alpha_1$
T	0.5	-0.25
C	0.5	0.25
B	-0.5	0.25

Example 4

We shall denote by $p_1; p_2$ and p_3 the probabilities of the strategies T, C and B respectively. The continuous time di®erential equation of p_1 is:

$$\frac{dp_1}{dt} = p_1(0{:}5 - (\alpha_1 \cdot 0{:}5 + (1 - \alpha_1) \cdot 0{:}25)) + 0{:}5\left(\left(p_1 + p_3\frac{p_1}{p_1 + p_2}\right)^{\alpha_1} - p_1\right)$$

We shall explain how to construct the second term, i.e. the probability matching term, in this equation. We need to compute the probability with which strategy T is successful. There are two events in which strategy T is successful. First, T may be played and receive a positive payo®. The probability of this event is $p_1^{\alpha_1}$. Alternatively, B may be played, receive a negative payo®, and T may be chosen to receive the re-assigned probability. The probability of this event is $p_3(p_1/(p_1 + p_2))^{\alpha_1}$. We hence obtain as the total probability with which T is successful: $(p_1 + p_3(p_1/(p_1 + p_2)))^{\alpha_1}$. The expected bene¯t of the top strategy, conditional on this event, is 0.5. This explains the probability matching term in the above equation.

7 Issues for Further Research

We conclude the paper by listing some issues for further research. In the previous section we derived a system of differential equations which characterises the continuous time limit of the learning process in the general case, but we did not investigate general properties of this system of equations. One issue for further research is hence a general study of these differential equations.

In Section 5 we emphasised that our results apply only to finite points in time, and not to the asymptotics for time tending to infinity. The discrete time asymptotics of the learning process are of particular interest to us, and we hope to deal with these in the future.

In this paper it was exogenous whether an experience is reinforcing or deterring. It seems more plausible that this is endogenous. Specifically, suppose that agents compare their experiences to an aspiration level, and that an experience is reinforcing or deterring depending on whether the payoff received is above or below the aspiration level. In this paper, we have implicitly assumed that the aspiration level is fixed over time and equal to zero. It seems more plausible to assume that the aspiration level adjusts over time in response to agents' experiences.

We investigate a model which includes this assumption in [5]. In that model, agents are "realistic" and adjust in each iteration their aspiration level towards the actually experienced payoff. In the continuous time limit this implies that the aspiration level moves towards the actual expected payoff. The adjustment of strategies is as in this paper, and hence, in the continuous time limit, strategy adjustment is governed by a replicator and a probability matching force. The endogeneity of the aspiration level then makes it in most cases unavoidable that there is an element of probability matching in the continuous time limit. Specifically, suppose that for every strategy there is some positive variance of payoffs. Once the endogenous aspiration level is sufficiently close to the expected payoff, the actual payoff will sometimes be below the aspiration level. Hence, "negative" payoffs become unavoidable, and probability matching will affect behavior. The endogenous adjustment of the aspiration level thus creates an element of irrationality. We describe the details of this effect in [5].

A final and important issue for further research is the extent to which the results in this paper depend on the particular functional forms of strategy adjustment which was postulated.

References

[1] Akin, E. and V. Losert, Evolutionary Dynamics of Zero-Sum Games, Journal of Mathematical Biology 20 (1984), 231-258.

[2] Benveniste, A., M. Métivier and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Berlin etc.: Springer Verlag, 1990.

[3] Binmore, K., Fun and Games, Lexington: D.C. Heath and Company, 1992.

[4] Binmore, K. and L. Samuelson, Muddling Through: Noisy Equilibrium Selection, mimeo., University College London, 1993.

[5] Börgers, T. and R. Sarin, Naive Reinforcement Learning With Endogenous Aspirations, mimeo., University College London and Texas A & M University, 1995.

[6] Boylan, R., Laws of Large Numbers for Dynamical Systems with Randomly Matched Individuals, Journal of Economic Theory 57 (1992), 473-504.

[7] Boylan, R., Continuous Approximation of Dynamical Systems with Randomly Matched Individuals, Journal of Economic Theory 66 (1995), 615-625.

[8] Bush, R.R. and F. Mosteller, A Mathematical Model for Simple Learning, Psychological Review 58 (1951), 313-323.

[9] Bush, R.R. and F. Mosteller, Stochastic Models for Learning, New York: Wiley, 1955.

[10] Butcher, J.C., The Numerical Analysis of Ordinary Differential Equations, Chichester etc.: Wiley, 1987.

[11] Cabrales, A., Stochastic Replicator Dynamics, mimeo., University of California, San Diego, 1993.

[12] Cabrales, A. and J. Sobel, On the Limit Points of Discrete Selection Dynamics, Journal of Economic Theory 57 (1992), 392-407.

[13] Cross, J.G., A Stochastic Learning Model of Economic Behavior, Quarterly Journal of Economics 87 (1973), 239-266.

[14] Cross, J.G., A Theory of Adaptive Economic Behavior, Cambridge: Cambridge University Press, 1983.

[15] Dekel, E. and S. Scotchmer, On the Evolution of Optimizing Behavior, Journal of Economic Theory 57 (1992), 392-407.

[16] Estes, W.K., Toward a Statistical Theory of Learning, Psychological Review 57 (1950), 94-107.

[17] Gilboa, I. and A. Matsui, A Model of Random Matching, Journal of Mathematical Economics 21 (1992), 185-197.

[18] Givon, M., and D. Horsky, Application of a Composite Stochastic Model of Brand Choice, Journal of Marketing Research 16 (1979), 258-267.

[19] Hofbauer, J. and K. Sigmund, The Theory of Evolution and Dynamical Systems, Cambridge: Cambridge University Press, 1988.

[20] Kuehn, A., Consumer Brand Choice as a Learning Process, Journal of Advertising Research 2 (1962), 10-17.

[21] Lakshmivarahan, S. and K.S. Narendra, Learning Algorithms for Two-Person Zero-Sum Games of Incomplete Information, Mathematics of Operations Research 6 (1981), 379-386.

[22] Lakshmivarahan, S. and K.S. Narendra, Learning Algorithms for Two-Person Zero-Sum Stochastic Games with Incomplete Information: A Uni¯ed Approach, Siam Journal of Control and Optimization 20 (1982), 541-552.

[23] Meyer, R.J. and B.E. Kahn, Probabilistic Models of Consumer Choice Behavior, in: T.S. Robertson and H.H. Kassarjian (editors), Handbook of Consumer Behavior, Englewood Cli®s: Prentice-Hall, 1991.

[24] Mookherjee, D. and B. Sopher, Learning Behavior in an Experimental Matching Pennies Game, Games and Economic Behavior 7 (1994), 62-91.

[25] Norman, M.F., Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators, Journal of Mathematical Psychology 5 (1968), 61-101.

[26] Norman, M.F., Markov Processes and Learning Models, New York and London: Academic Press, 1972.

[27] Ritzberger, K. and J. Weibull, Evolutionary Selection in Normal Form Games, Econometrica, forthcoming.

[28] 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.

[29] Samuelson, L. and J. Zhang, Evolutionary Stability in Asymmetric Games, Journal of Economic Theory 57 (1992), 363-392.

[30] Schlag, K., Why Imitate, and If So, How ?, mimeo., University of Bonn, 1994.

[31] Schmalensee, R., Alternative Models of Bandit Selection, Journal of Economic Theory 10 (1975), 333-342.

[32] Siegel, S., Decision Making and Learning under Varying Conditions of Reinforcement, Annals of the New York Academy of Sciences 89 (1960-1961), 766-783.

[33] Suppes, P. and R. Atkinson, Markov Learning Models for Multiperson Interaction, Stanford: Stanford University Press, 1960.

[34] Taylor, P.D., Evolutionarily Stable Strategies With Two Types of Players, Journal of Applied Probability 16 (1979), 76-83.

[35] Taylor, P.D. and L.B. Jonker, Evolutionarily Stable Strategies and Game Dynamics, Mathematical Biosciences 40 (1978), 145-156.

[36] Weissing, F.J., Evolutionary Stability and Dynamic Stability in a Class of Evolutionary Normal Form Games, in: R. Selten (ed.), Game Equilibrium Models I: Evolution and Game Dynamics, 29-97, Berlin: Springer-Verlag, 1991.