JOURNAL OF ECONOMIC THEORY 77, 1–14 (1997)
ARTICLE NO. ET972319

Learning Through Reinforcement and Replicator Dynamics*

Tilman Börgers

Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom

and

Rajiv Sarin

Department of Economics, College of Liberal Arts, Texas A & M University, College Station, Texas 77843-4228

Received November 26, 1994; revised April 1, 1997

This paper considers a version of R. R. Bush and F. Mosteller's (1951, *Psych. Rev.* **58**, 313–323; 1955, "Stochastic Models for Learning," Wiley, New York) stochastic learning theory in the context of games. We show that in a continuous time limit the learning model converges to the replicator dynamics of evolutionary game theory. Thus we provide a non-biological interpretation of evolutionary game theory. *Journal of Economic Literature* Classification Numbers: C72, D83.

1. INTRODUCTION[1]

The evolutionary approach to game theory attracts increasing attention. If the word "evolution" is used in a biological sense, then this approach is

* The material in this paper is taken from some sections of an earlier discussion paper of ours with the same title. We thank an associate editor, two referees, Murali Agastya, Ken Binmore, Vince Crawford, Drew Fudenberg, Nick Rau, Max Stinchcombe, and, especially, 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.

[1] The following papers, which were published after a first version of our paper was circulated, and of which we became aware only after their publication, contain results which are very similar to our Proposition 1: Sastry, Phansalkar and Thathachar [27] (their Theorem 3.1) and Phansalkar, Sastry, and Thathachar [23] (their Theorem 4.1). Unlike our paper, neither of these two papers mentions that the continuous time dynamics which they obtain is relevant to evolutionary game theory. The two papers also do not obtain a result analogous to our Proposition 2, concerning discrete time asymptotics.

0022-0531/97 $25.00

concerned with environments in which behavior is genetically determined, and strategy selection obtains because carriers of different genes differ in reproductive fitness. 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 different 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.

The purpose of this paper is to present a formalization of the analogy between learning at the individual level and biological evolution. We are interested in this case because it has received less attention,[2] and because, traditionally, game theory has been viewed as a theory of individual players rather than populations of players.

The model of biological evolution for which we provide learning foundations is very specific. It is the asymmetric, continuous time replicator dynamics.[3] This dynamic process has attracted much interest in the recent game theory literature.[4] 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

[2] The analogy between learning in populations and biological evolution has recently been formalized by, among others, Binmore and Samuelson [3], Cabrales [10] and Schlag [28].

[3] See Hofbauer and Sigmund [16], Taylor [32], and Taylor and Jonker [33].

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

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.[5]

The learning model which we consider in this paper is a special case of a more general model which we develop in Börgers and Sarin [4] and which is based on Bush and Mosteller's [7, 8] "stochastic learning theory." To place our work into context, we first outline the general model,[6] and then introduce the special case investigated here.

The general model considers several agents playing in discrete time repeatedly the same normal-form game. At each point in time, each player is characterized 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 affected by some unmodelled psychological factors.

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

We emphasize that payoffs in the 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, payoffs are simple parametrizations of players' responses to their experiences.

The players in the learning 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 too costly relative to the potential gains. The model thus seems most plausible if agents' behaviour is habitual, and not the result of careful reflection.

Consumers' choice of brands of everyday items are an example of an economic context in which the learning model might be applicable. Marketing research (see Meyer and Kahn (1991)) has indeed found empirical support for versions of Bush–Mosteller theory in this area. Experimental evidence for learning processes which are similar to the one in this paper has been

[5] Samuelson and Zhang [26] have shown that the (asymmetric) continuous time replicator dynamic, and certain multiples of it, are within a certain class of processes unique in that they satisfy "aggregate monotonicity." Implications of aggregate monotonicity have been investigated by Samuelson and Zhang [26] and Ritzberger and Weibull [24].

[6] We simplify the exposition here by omitting from our discussion the endogenous aspiration level postulated in Börgers and Sarin [4]. But see in this regard the comments in Section 6.

given by Suppes and Atkinson [31] and, more recently, by Mookherjee and Sopher [20] and Roth and Erev [25].

To relate the learning model to the replicator dynamics, we need to restrict attention to a special case, namely to the case that all payoffs are positive. Payoffs differ only in their size. Intuitively, this case might be interpreted as the case that all choices are "habit-forming," and that payoffs matter only in that they determine the extent to which habit formation occurs. This special case has previously been considered by Cross [11]. We therefore refer to the learning model in this paper also as "Cross' model."[7]

Our main result is that, in an appropriately constructed continuous time limit, Cross' learning model converges to the asymmetric, continuous time version of the replicator dynamics. We construct the continuous time limit by postulating that each time interval sees "many" iterations of the game, and that the adjustments which players make between two iterations of the game are "very small." If the continuous time limit is constructed in this way, a law of large numbers can be applied, and the (stochastic) learning process becomes in the limit deterministic. The limit process satisfies the differential equation which characterises the replicator dynamics.

It is important to note that the limit result refers only to arbitrary, *but finite* points in time. It is not true if *infinite* time is considered. More precisely, as will be shown in this paper, the asymptotic behavior for time tending to infinity of the discrete time learning processes may be quite different from the asymptotic behavior of the continuous time replicator dynamics.

Mathematically, this paper relies largely on results due to Norman [21, 22]. Mathematically related to our work is also Boylan's [6] investigation of the continuous time limit of a model in which large, but finite populations of agents interact in discrete time. 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 [5].

2. CROSS' LEARNING 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: $j \in J \equiv \{1, 2, ..., J\}$. The feasible strategies of C are: $k \in K \equiv \{1, 2, ..., K\}$. The payoff to player i when R plays j and C plays k is U^i_{jk}. We write U^i for the matrix with U^i_{jk} in row j and column k.

[7] Previous investigations of Cross' learning process are contained in Cross [11] and Schmalensee [29]. Other processes in the Bush–Mosteller class have been investigated in Bush and Mosteller [7, 8], and Lakshmivarahan and Narendra [17, 18].

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

Payoffs will be interpreted as "strengths of reinforcement." We shall assume that they satisfy: $0 < U^i_{jk} < 1$ for all i, j and k. We explained already in the Introduction that we focus 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 2 below.[9]

The players play the game repeatedly. 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 $P(n) \equiv (P_1(n), ..., P_J(n))$. For player C they are $Q(n) \equiv (Q_1(n), ..., Q_K(n))$. 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 $S(n)$ is 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, and we write $p_j(n)$ resp. $q_k(n)$ for the components of $p(n)$ and $q(n)$.

For any $L \in \mathbb{N}$, we denote by $\mathscr{S}^L$ the L-dimensional simplex. The set of all possible states for player R (resp. C) is then $\mathscr{S}^{J-1}$ (resp. $\mathscr{S}^{K-1}$). The set of all possible states of the game is $\mathscr{S} \equiv \mathscr{S}^{J-1} \times \mathscr{S}^{K-1}$.

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 nth 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

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

$$P_{j'}(n+1) = (1 - U^R_{jk})\, P_{j'}(n) \qquad \text{for all} \quad j' \neq j.$$

Player C updates $Q(n)$ in an analogous manner. Observe that the above formula is meaningful only if $U^R_{jk} \leqslant 1$. 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 nth 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

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

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 $E[\ldots \mid S(n) = s(n)]$ the expected value of the random variable $(\ldots)$ conditional on the state of the game in stage n being $s(n)$. We write e_r for the unit vector with a one in the rth row, and zeros elsewhere. For simplicity we drop the transpose symbols when referring to row vectors. Straightforward calculations show that for all $n \in \mathbb{N}$, $s \in \mathscr{S}$, $j \in J$ and $k \in K$

$$E[\Delta P_j(n) \mid S(n) = s(n)] = p_j(n)(e_j U^R q(n) - p(n)\ U^R q(n))$$
$$E[\Delta Q_k(n) \mid S(n) = s(n)] = q_k(n)(p(n)\ U^C e_k - p(n)\ U^C q(n)).$$

Note that the term on the right hand side of these equations is essentially the same as the term on the right hand side of the discrete time replicator equation of evolutionary game theory.[10] Of course, this is only the *expected movement* of the learning process. In the next section, we shall show, however, that, in an appropriately constructed continuous time limit, the learning process' *actual* movement will not differ from its *expected* movement.

3. THE CONTINUOUS TIME LIMIT

To construct a continuous time limit of the learning model, we imagine that the amount of "real" time that passes between two repetitions of the game is given by a number θ with $0 < \theta \leqslant 1$. After each repetition of the game, the players adjust their states by θ times what it was so far. The crucial assumption here is that players' adjustments of their states "slows down" at the same rate at which the time distance between two repetitions shrinks.

Formally, we replace the adjustment formulas given earlier by

$$P_j^\theta(n+1) = \theta U_{jk}^R + (1 - \theta U_{jk}^R)\ P_j^\theta(n)$$
$$P_{j'}^\theta(n+1) = (1 - \theta U_{jk}^R)\ P_{j'}^\theta(n), \qquad \text{for all} \quad j' \neq j,$$

where we introduce the upper index θ to indicate that we are now referring to a modified process. An analogous formula applies to $Q^\theta(n+1)$. We obtain a process $\{(P^\theta(n), Q^\theta(n))\}_{n \in \mathbb{N}}$, provided that we specify the initial random variables $(P^\theta(1), Q^\theta(1))$.

Since we imagine the time interval between repetitions to be θ, the random variable $S^\theta(n)$ describes the state of the process at time $n\theta$. We are interested in the limit $\theta \to 0$. We obtain the state of the limit process at

[10] Often authors give a version of the discrete time replicator dynamics where this term is divided by some denominator which does not, however, affect the sign of the expression. See, for example, p. 419 in Binmore [2].

some time $t \geqslant 0$ by investigating the limit of $S^\theta(n)$ for any sequence of θs and ns with the property that $\theta \to 0$ and $n\theta \to t$.

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

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

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

for all $t \geqslant 0$, $j \in J$ 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 [32].

PROPOSITION 1. *Suppose that for all* θ: $(P^\theta(1), Q^\theta(1)) = (\hat{p}(0), \hat{q}(0))$ *with probability* 1. *Consider some* t *with* $0 \leqslant t < \infty$ *and assume* $\theta \to 0$ *and* $n\theta \to t$. *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* $S^\theta(n)$ *converges in probability to* $(\hat{p}(t), \hat{q}(t))$.

In words, Proposition 1 says that, if θ is small, and if $n\theta$ is close to $t \geqslant 0$, then, with high probability, $S^\theta(n)$ will take a value that is close to the solution of the continuous time replicator equation at time t. Intuitively, frequent play and slow movement ensure that a law of large numbers applies, and therefore actual and expected movement of the learning process coincide.

Proof. We use Theorem 1.1 in Chapter 8 of Norman [22]. This theorem concerns the continuous time limit of discrete time Markov processes with infinite state spaces. The processes to which we apply this theorem are the processes $\{S^\theta(n)\}_{n \in \mathbb{N}}$. Our assertion follows immediately from parts (A) and (B) of Norman's theorem. Therefore, it is sufficient to verify that the assumptions of the theorem are satisfied. This is trivially true for Norman's assumptions (a.1)–(a.3).

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

$$v(p, q) \equiv E\left[\frac{\Delta S^\theta(n)}{\theta} \middle| S^\theta(n) = (p, q)\right].$$

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

$$w(p, q) \equiv Var\left[\frac{\Delta S^\theta(n)}{\theta} \middle| S^\theta(n) = (p, q)\right].$$

(Here, we denote by $Var[\ldots | S^\theta(n) = s]$ 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 $r: \mathscr{S} \to \mathbb{R}$ which is defined by

$$r(p, q) \equiv E\left[\left| \frac{\Delta S^\theta(n)}{\theta} \right|^3 \middle| S^\theta(n) = (p, q) \right].$$

(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 1 follows from the last equation in Section 2 where $v(p, q)$ was calculated. ∎

Remark 1. Using results in Norman [22] it can be shown that, under the assumptions of Proposition 4, for every $\varepsilon > 0$ and every $j \in J$ the probability $Pr(|P_j^\theta(n) - \hat{p}_j(t)| \geqslant \varepsilon)$ converges to zero at least as fast as θ. The analogous statement holds for every pure strategy of player C.

Remark 2. A stronger version of Proposition 4 would assert that, as θ tends to zero, the distribution of the polygonal curve connecting the points $(n^\theta, S^\theta(n))$ (where $n \in \mathbb{N}$ and $n\theta \leqslant t$) 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.

4. DISCUSSION

The model of Section 2 postulates a particular learning 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 2 is a special instance. Proceeding like this has the advantage that the formal framework admits several different 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 [8] a specific interpretation of their model. It was based on ideas from Estes' [14] stimulus sampling theory of learning. In this section we give a similar interpretation that applies to our context.[11] Then we use this interpretation to develop intuition for the relation between learning and evolution as formalized in Proposition 1.

Suppose that each player when making a choice is subject to many stimuli. Specifically, 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 strategy to the player, but different stimuli may suggest different strategies. The player chooses a strategy by selecting randomly one of these stimuli and adopting the strategy corresponding to this stimulus. Once a player has chosen a strategy, and experienced a payoff, 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 payoff which the player experienced.

A straightforward calculation shows that this model of players' behavior generates exactly the process that we described in Section 2. Thus, the model provides one possible interpretation of the framework of Section 2.

To develop intuition for the relation between learning and evolution which is described by Proposition 1, it is useful to compare the stimulus sampling model just described to the biological model which underlies the replicator dynamics. We therefore now briefly review this biological model. We describe a version of the model which is particularly suitable for our purposes.

The biological model postulates that the game is played not by two individual players, but by two populations of players, with each population being of continuum size. Players' behaviour in the game is determined by their genes, which are fixed during a player's lifetime.

The game is played repeatedly. In every stage a proportion of players in each population is selected randomly to play the game. The two selected groups of players are then randomly matched in pairs.[12] Players play the pure strategies with which they are programmed.

After playing the game 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. Offsprings inherit the behaviour of their (single) parents. After reproduction,

[11] We shall make some simplifications in comparison to Bush and Mosteller's argument.

[12] 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 [5] and Gilboa and Matsui [15], but we know of no corresponding work for continuum size populations.

a proportion of the members 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.[13]

It is easily calculated that the *actual* change from one iteration to the next in the proportion of players playing a particular strategy is equal to the *expected* change in the probability of playing some particular strategy in the learning model of Section 2. Thus, the learning model of Section 2 and the biological model just described coincide in their *expected* movement.

Intuitively, the analogy between the two models is obvious. The sets of stimuli in the learning model correspond to the populations of players in the biological model. The re-programming of stimuli in the learning model is the analog of the reproduction and death processes in the biological model.

However, the learning model evolves stochastically whereas the biological model is deterministic. Two factors account for this difference. Firstly, whereas in the biological model in each iteration the proportion of individuals playing is a random sample of continuum size, and hence by the law of large numbers composed as the original population, in the learning model at each point in time only *one* strategy is played. Secondly, whereas in the biological model the outcome of any single interaction cannot have any impact on the population composition (by the continuum assumption), in the learning model the outcome of a single interaction occurring affects the probabilities with which strategies are played in the next period.

If a continuous time limit is constructed, the difference between the two models disappears, and both models converge to the same deterministic continuous time limit. The randomness in the learning model disappears because at each iteration only a very small proportion of stimuli is re-programmed. Any finite time interval will see many iterations of the game. In the continuous time limit the total effect of the re-programming of stimuli is determined by a law of large numbers, and is deterministic.

5. ASYMPTOTIC ANALYSIS

Proposition 1 applies to any *finite* point in time $t < \infty$. It does not, however, refer to the asymptotic behavior, for $t \to \infty$, of the discrete and continuous time processes. In fact, the asymptotic behavior of the learning process may

[13] To make these assumptions consistent one needs to assume that the proportion of players who play the game is sufficiently small. Otherwise it could happen that the population size can stay constant only if also newborns can die.

be very different from that of the continuous time replicator dynamics. This is even true for arbitrarily low values of θ.

To show this we now 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 limit in which both players play some pure strategy. This result holds for all possible values of θ. It is well-known that in games such as "Matching Pennies" the replicator dynamics will *not* converge to a pure strategy outcome, but will cycle in the interior of the state space.[14] Therefore, the following result indicates a difference in the asymptotic behaviour of the learning process and of replicator dynamics.

PROPOSITION 2. *For all* $\theta > 0$ *and for all initial variables* $(P^\theta(1), Q^\theta(1))$ *with probability* 1 *the sequence* $\{(P^\theta(n), Q^\theta(n))\}_{n \in \mathbb{N}}$ *converges, and its limit is in* $J \times K$.[15]

Proof. Taking θ to be given and fixed, we use Theorem 2.3 of Norman [21]. 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 $\tilde{s}(n)$ and $\tilde{s}'(n)$ 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.[16] Then $d(s(n+1), s'(n+1)) \leqslant d(s(n), s'(n))$, 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.

[14] See Section 17 of Hofbauer and Sigmund [16].

[15] We adopt here a standard abuse of notation, and denote by $J \times K$ those elements of $\mathscr{S}^{J-1} \times \mathscr{S}^{K-1}$ which place all probability on one pure strategy combination.

[16] 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.

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. ∎

Remark 3. One can also prove that, for any $\theta > 0$, if the initial value is completely mixed, every element of $J \times K$ has a positive probability of being the limit of $\{(P^\theta(n), Q^\theta(n))\}_{n \in \mathbb{N}}$. This can be shown using the methods of Section 7.2 of Bush and Mosteller [8].[17]

Remark 4. The asymptotic behaviour of the learning process which is described in Proposition 2 differs not only from that of *continuous time* replicator dynamics, but also from that of *discrete time* replicator dynamics.[18] In "Matching Pennies," for example, discrete time replicator dynamics will cycle along expanding trajectories,[19] but will not get absorbed by any pure strategy outcome.

6. CONCLUSION

Two assumptions of the learning model considered in this paper may appear rather implausible. The first is that *all* experiences are "habit forming," i.e., make the decision maker more likely to choose the same action again. Intuition suggests that some experiences are "deterring," and induce the decision maker to *reduce* the probability of the action which lead to the experience. A second aspect of the model which might be criticized as implausible is that the impact of payoff experiences on the decision maker's behaviour is exogenous and fixed over time. How good a certain payoff feels to an agent should itself depend on the agent's past payoff experience. In practice, it seems that decision makers often form an "aspiration" based on their experiences with different actions, that payoff rewards are reinforcing or deterring depending on how they compare to the aspiration level, and that decision makers adjust their aspirations as they gain experience. In the

[17] We are grateful to Nick Rau for this observation.

[18] It is well-known that the two versions of replicator dynamics can have different asymptotics. See Cabrales and Sobel [9], Dekel and Scotchmer [13], Samuelson and Zhang [26] or Weissing [34].

[19] See Akin and Losert [1].

current paper the aspiration level was, by contrast, implicitly taken to be exogenous and fixed at zero.

In Börgers and Sarin [4] we develop a model which takes both of these points into account. We then construct a continuous time limit, as in the current paper, and find that the right hand side of the limit differential equation consists of two terms, one of which is similar to the right hand side of the replicator equation. The other term, however, has the effect that the decision maker exhibits a particular form of behaviour which does not maximise expected payoffs: so-called "probability matching" (see, for example, Siegel [30]). For a more detailed analysis of this model the reader is referred to Börgers and Sarin [4].

REFERENCES

1. E. Akin and V. Losert, Evolutionary dynamics of zero-sum games, *J. Math. Biol.* **20** (1984), 231–258.
2. K. Binmore, "Fun and Games," Heath, Lexington, 1992.
3. K. Binmore and L. Samuelson, Muddling through: noisy equilibrium selection, mimeo, University College London, 1993.
4. T. Börgers and R. Sarin, Naive reinforcement learning with endogenous aspirations, mimeo, University College London and Texas A & M University, 1995.
5. R. Boylan, Laws of large numbers for dynamical systems with randomly matched individuals, *J. Econ. Theory* **57** (1992), 473–504.
6. R. Boylan, Continuous approximation and dynamical systems with randomly matched individuals, *J. Econ. Theory* **66** (1995), 615–625.
7. R. R. Bush and F. Mosteller, A mathematical model for simple learning, *Psych. Rev.* **58** (1951), 313–323.
8. R. R. Bush and F. Mosteller, "Stochastic Models for Learning," Wiley, New York, 1955.
9. A. Cabrales and J. Sobel, On the limit points of discrete selection dynamics, *J. Econ. Theory* **57** (1992), 392–407.
10. A. Cabrales, Stochastic replicator dynamics, mimeo, University of California, San Diego, 1993.
11. J. G. Cross, A stochastic learning model of economic behavior, *Quart. J. Econ.* **87** (1973), 239–266.
12. J. G. Cross, "A Theory of Adaptive Economic Behavior," Cambridge Univ. Press, Cambridge, 1983.
13. E. Dekel and S. Scotchmer, On the evolution of optimizing behavior, *J. Econ. Theory* **57** (1992), 392–407.
14. W. K. Estes, Toward a statistical theory of learning, *Psych. Rev.* **57** (1950), 94–107.
15. I. Gilboa and A. Matsui, A model of random matching, *J. Math. Econ.* **21** (1992), 185–197.
16. J. Hofbauer and K. Sigmund, "The Theory of Evolution and Dynamical Systems," Cambridge Univ. Press, Cambridge, 1988.
17. S. Lakshmivarahan and K. S. Narendra, Learning algorithms for two-person zero-sum games of incomplete information, *Math. Operations Res.* **6** (1981), 379–386.
18. S. Lakshmivarahan and K. S. Narendra, Learning algorithms for two-person zero-sum stochastic games with incomplete information: A unified approach, *Siam J. Control Optimization* **20** (1982), 541–552.

19. R. J. Meyer and B. E. Kahn, Probabilistic models of consumer choice behavior, *in* "Handbook of Consumer Behavior" (T. S. Robertson and H. H. Kassarjian, Eds.), Prentice-Hall, Englewood Cliffs, 1991.
20. D. Mookherjee and B. Sopher, Learning behavior in an experimental matching pennies game, *Games Econ. Behav.* **7** (1994), 62–91.
21. M. F. Norman, Some convergence theorems for stochastic learning models with distance diminishing operators, *J. Math. Psych.* **5** (1968), 61–101.
22. M. F. Norman, "Markov Processes and Learning Models," Academic Press, New York/London, 1972.
23. V. V. Phansalkar, P. S. Sastry, and M. A. L. Thathachar, Absolutely expedient algorithms for learning Nash equilibria, *Proc. Indian Acad. Sci. Math. Sci.* **104** (1994), 279–294.
24. K. Ritzberger and J. Weibull, Evolutionary selection in normal form games, *Econometrica* **63** (1995), 1371–1399.
25. A. Roth and I. Erev, Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term, *Games Econ. Behav.* **8** (1995), 164–212.
26. L. Samuelson and J. Zhang, Evolutionary stability in asymmetric games, *J. Econ. Theory* **57** (1992), 363–392.
27. P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information, *IEEE Trans. Systems, Man Cybernetics* **24** (1994), 769–777.
28. K. Schlag, Why imitate, and if so, how?, mimeo, University of Bonn, 1994.
29. R. Schmalensee, Alternative models of bandit selection, *J. Econ. Theory* **10** (1975), 333–342.
30. S. Siegel, Decision making and learning under varying conditions of reinforcement, *Annals New York Acad. Sci.* **89** (1960–1961), 766–783.
31. P. Suppes and R. Atkinson, "Markov Learning Models for Multiperson Interaction," Stanford Univ. Press, Stanford, 1960.
32. P. D. Taylor, Evolutionarily stable strategies with two types of players, *J. Appl. Probability* **16** (1979), 76–83.
33. P. D. Taylor and L. B. Jonker, Evolutionarily stable strategies and game dynamics, *Math. Biosci.* **40** (1978), 145–156.
34. F. J. Weissing, Evolutionary stability and dynamic stability in a class of evolutionary normal form games, *in* "Game Equilibrium Models I: Evolution and Game Dynamics" (R. Selten, Ed.), Springer-Verlag, Berlin, 1991.