INTERNATIONAL ECONOMIC REVIEW
Vol. 41, No. 4, November 2000

NAIVE REINFORCEMENT LEARNING WITH ENDOGENOUS ASPIRATIONS*

BY TILMAN BÖRGERS AND RAJIV SARIN[1]

University College London, U.K., and
Texas A&M University, U.S.A.

This article considers a simple model of reinforcement learning. All behavior change derives from the reinforcing or deterring effect of instantaneous payoff experiences. Payoff experiences are reinforcing or deterring depending on whether the payoff exceeds an aspiration level or falls short of it. Over time, the aspiration level is adjusted toward the actually experienced payoffs. This article shows that aspiration level adjustments may improve the decision maker's long-run performance by preventing him or her from feeling dissatisfied with even the best available strategies. However, such movements also lead to persistent deviations from expected payoff maximization by creating "probability matching" effects.

1. INTRODUCTION

A simple and intuitively plausible principle for learning behavior in decision problems and games is as follows: Actions that yield payoffs above the decision maker's aspiration level are more likely to be chosen in the future, and actions that yield a payoff below the decision maker's aspiration level are less likely to be chosen in the future. Models of learning that directly formalize this idea, and which do not refer to any explicit optimization by the agent, will be referred to in the following as models of *reinforcement learning*. We distinguish such models from *belief-based learning models* such as fictitious play. These latter models attribute explicit subjective beliefs and the ability to maximize given these beliefs.

Economists recently have given some attention to reinforcement learning. One reason is that certain specifications of reinforcement learning models seem to hold promise in explaining experimental data. Examples of articles that come to this conclusion are those by Roth and Erev (1995), Mookherjee and Sopher (1997), and Erev and Roth (1998). In fact, some articles come to the conclusion that reinforcement learning models explain experimental data better than belief-based learning models,

* Manuscript received January 1998; revised May 1999

[1] We are grateful to Murali Agastya, Antonio Cabrales, George Mailath, two referees, and participants of the Second International Conference on Economic Theory: Learning in Games at Universidad Carlos III de Madrid for their comments on earlier versions of this article. Part of this research was undertaken while Tilman Börgers was visiting the Indian Statistical Institute in Delhi and the Institute of Advanced Studies in Vienna. He thanks both institutes for their hospitality. Tilman Börgers also thanks the Economic and Social Research Council for financial support under Research Grant R000235526.

namely, those by Camerer and Ho (1997), Chen and Tang (1998), and Mookherjee and Sopher (1994, 1997). Another reason for the recent interest in reinforcement learning among economists is that there is a close analogy between reinforcement learning and dynamic processes studied in evolutionary game theory (see Börgers and Sarin, 1997).

There is a long tradition of research on reinforcement learning in psychology. Early mathematical models of reinforcement learning in psychology are those of Bush and Mosteller (1951, 1955) and Estes (1950). Reinforcement theory continues to be one of the major approaches that psychologists use when studying learning. The prominence of reinforcement theories in current psychology of learning is evident from textbooks such as those of Lieberman (1993) and Walker (1995).

Previous analytical work on reinforcement learning models has focused on the case where the decision maker's aspiration level is exogenously given and fixed. One case that has received some attention is that the exogenously fixed aspiration level is below all conceivable payoff levels; see, for example, Arthur (1993), Börgers and Sarin (1997), and Cross (1973). A smaller branch of the literature has considered the case that there are only two possible payoffs values and that the aspiration level is exactly in the middle between these two values (see Bush and Mosteller, 1951, 1955; Schmalensee, 1975).

Experimental work and intuition suggest, however, that the aspiration level of an agent is endogenous and changes over time. For example, the article by Bereby-Meyer and Erev (1998) shows that reinforcement learning models with endogenous aspiration levels explain data better than models of learning with exogenous aspiration levels. How good a certain payoff "feels" depends on the past payoff experience of the agent. This article offers some first analytical results about the properties of reinforcement learning models when the aspiration level is endogenous. In addition, our model contains as a special case the case that the aspiration level is exogenous and fixed, and our article provides more general results for this case than have been available so far.

Our analysis is set in the context of a single-person decision problem under risk. Moreover, we shall postulate that the decision maker has only two choices. We make these assumptions for analytical simplicity. We shall argue in the last section of this article, however, that some of our results can be straightforwardly extended to the more general case in which the decision maker has more than two choices and in which he or she is involved in a game rather than a single-person decision problem.

We shall assume that the decision maker faces the same choice problem repeatedly. At any point in time, his or her behavior is given by a probability distribution over his or her two actions. The distribution should not be interpreted as conscious randomization. Rather, it indicates from the perspective of the outside observer how likely it the decision maker is to choose each of these actions. The decision maker also has an aspiration level. The decision maker chooses in each period some action, receives a payoff, and then compares the payoff to the aspiration level. If the payoff was above the aspiration level, then the decision maker enters the next period with a probability distribution that makes it more likely that he or she will choose the same action again. The increase in the probability of this action is proportional to the difference between the payoff and the aspiration level. The reverse occurs if

the payoff falls short of the aspiration level. The aspiration level itself is adjusted in the direction of the payoff realization.

To investigate our learning model, we introduce a continuous time approximation of the learning process. This is a technical device aimed at simplifying our work. The continuous time approximation is valid if, in each time interval, the decision maker plays very frequently and, after each iteration, responds to his or her experience with only very small adjustments to his or her choice probabilities. Whereas in discrete time the learning process is stochastic, in the continuous time limit it becomes deterministic, and the trajectories are characterized by simple differential equations.

We investigate these differential equations in detail in this article. We show that the equations reflect two forces that together determine the decision maker's behavior. First, there is a force that is similar to the force modeled by the replicator dynamics in evolutionary game theory. Roughly speaking, this force steers the process into the direction of expected payoff maximization.

A second force, however, draws the decision maker into the direction of *probability-matching* behavior. We briefly explain this term. Suppose the decision maker has to choose repeatedly one of two strategies s_1 and s_2. With probability μ, strategy s_1 yields one dollar, and strategy s_2 yields nothing. With probability $1 - \mu$, strategy s_2 yields one dollar, and strategy s_1 yields nothing. One says that the decision maker's behavior exhibits *probability matching* if the long-run frequency with which strategy s_1 is chosen is μ and the long-run frequency of strategy s_2 is $1 - \mu$. Probability matching is irrational, provided that $\mu \neq 0.5$, because rational behavior would require that one of the two actions is chosen with probability 1.

There is some empirical evidence of probability matching (see Siegel, 1960–1961; Winter, 1982). The phenomenon seems to arise more clearly if payoffs are small. The intuition why the reinforcement learning model predicts probability matching is that the decision maker in this model responds myopically to instantaneous payoff experiences. Since the optimal choice sometimes yields payoffs below the aspiration level, the decision maker is thrown back and forth between different choices.

Probability matching should be distinguished carefully from the *matching law* proposed by Herrnstein (Herrnstein, 1997; Herrnstein and Prelec, 1991). Herrnstein considers more complicated decision problems than we do. He assumes that the payoff distribution derived from a choice depends on the frequency with which this choice is made in some given finite time interval. Herrnstein's matching law asserts that choices are made such that the empirical average payoff for all choices is the same. Note that this will not be true for agents who probability match.

Because our learning model allows for more than two payoff levels, we introduce a generalized definition of probability matching. We then show that the replicator force and the probability-matching force together are the only forces that affect the decision maker's behavior. The replicator force is *the only* active force if all payoffs are above the aspiration level. If some payoffs are below the aspiration level, then the probability-matching force will be at work as well. The probability matching force is *the only* force present in the model if all payoffs deviate by the same amount from the aspiration level, but some are above and some below this level. Endogenous movements of the aspiration level affect the relative weight of the replicator force and the probability-matching force.

We next ask whether endogenous aspiration level movements are beneficial or harmful for the long-run performance of the decision maker. The answer depends on characteristics of the decision problem as well as the decision maker's initial aspiration level. If the decision maker's initial aspiration level is low, then, in most cases, endogenous aspiration level adjustments will be harmful for the decision maker. He or she would do better if he or she maintained a low aspiration level. The reason is that with a low aspiration level, the learning process acts like replicator dynamics and hence optimizes in the long run. Endogenous aspiration level movements will tend to raise the aspiration level and therefore will bring the probability-matching effect into play. This effect will prevent the decision maker from learning to play the optimal strategy.

If the decision maker's initial aspiration level is relatively high, then the issue is more complex. If the aspiration level is kept fixed, the probability-matching effect will prevent the decision maker from long-run optimization. Endogenous movements of aspiration level may help to alleviate this problem by making the decision maker more realistic. However, we shall show in this article that it is also possible that the endogenous aspiration level movements do additional harm to the decision maker.

An interesting implication of our results is that in the framework of this article, the only learning behavior that guarantees that the decision maker finds in the long run the expected payoff-maximizing strategy is learning behavior that starts with a very low initial aspiration level and which keeps this aspiration level constant over time. If the decision maker follows this rule, then his or her behavior will be determined by the replicator effect alone and hence will be optimal in the long run. Another way of putting this is that a reinforcement learner will find the optimal strategy if and only if he or she imitates the process of biologic evolution.

This article is organized as follows: Section 2 describes the decision problem that the decision maker faces and introduces the class of learning processes that we consider. Section 3 constructs differential equations that characterize the continuous time limit of the learning processes. We also explain how these differential equations reflect the two forces of replicator dynamics and probability matching. In Section 4 we present analytical and numerical results concerning the impact of endogenous aspiration level movements. Section 5 discusses related literature, and Section 6 considers some possible extensions of our research. Most of the proofs are in the Appendix.

2. THE MODEL

We consider a decision maker who has a choice between *two* strategies only: s_1 and s_2. We assume that the decision maker faces some risk. For simplicity, we postulate that the set of possible states of the world is finite. Each state has an *objective* probability of occurring. Payoffs depend on the strategy chosen and on the state of the world. We normalize payoffs to be between zero and one. We exclude the uninteresting case that the expected payoff of both strategies is the same. It is then without loss of generality to assume that s_1 has strictly higher expected payoff than s_2. This leads to the following definition.

DEFINITION 1. A *decision problem* $\mathscr{D}$ is a four-tuple $(S, \mathscr{E}, \mu, \pi)$ where

- $S \equiv \{s_1, s_2\}$ is the set of strategies.
- $\mathscr{E}$ is a nonempty, finite set of states of the world.
- μ is a probability measure on $\mathscr{E}$ such that $\mu(e) > 0$ for all $e \in \mathscr{E}$.
- $\pi : S \times \mathscr{E} \to (0, 1)$ is the decision maker's payoff function. It satisfies $\sum_{e \in \mathscr{E}} \mu(e)\pi(s_1, e) > \sum_{e \in \mathscr{E}} \mu(e)\pi(s_2, e)$.

The decision maker faces the same decision problem repeatedly. We denote the repetitions of the decision problem by n, where n takes values in $\mathbb{N} \cup \{0\}$. In each round, the decision maker first chooses a strategy, and then the state of the world is realized. For different n, the states of the world are independently and identically (according to μ) distributed. We assume that in each iteration the decision maker observes only his or her payoff. He or she does not observe the state of the world.

We shall take the decision maker's choice at each iteration to be random. The interpretation of this assumption was discussed in the Introduction. The probability distribution over S at iteration n is denoted by p_n. The set of all such probability distributions, i.e., the one-dimensional simplex, will be denoted by Δ. By $p_n(s)$ we denote the probability with which strategy s is chosen at iteration n. At each iteration n, the decision maker also will have an *aspiration level* $a_n \in [0, 1]$. Roughly speaking, a_n indicates which payoff level the decision maker finds satisfactory at iteration n. The precise role of the aspiration level will become clear once we specify the learning rule.

We take p_0 and a_0 as exogenous. Our only assumption for p_0 and a_0 is that $p_0(s) \neq 0$ for both $s \in S$. We make this assumption to exclude the trivial case that a strategy is never played just because it does not have positive probability initially.

We specify the learning rule by describing how p_n and a_n change from one iteration to the next. Consider some fixed n, and suppose that the current state of the decision maker is (p_n, a_n). Assume also that in iteration n the decision maker chose strategy s, that the state of the world was e, and that the decision maker hence received the payoff $\pi(s, e)$.

If $\pi(s, e) \geq a_n$, we assume that the decision maker takes this as encouragement to play s again. Hence, in iteration $n + 1$, s will have a higher probability. The other strategy's probability decreases correspondingly. The size of the increase in the probability of s is proportional to the size of the difference $\pi(s, e) - a_n$. Formally, we assume that the new probability vector p_{n+1} is a convex combination of the old probability vector p_n and the unit vector that places all probability on s. The weight assigned to the unit vector is equal to $\pi(s, e) - a_n$.[2]

In addition to the probability vector p_n, the aspiration level a_n also is adjusted. We assume that the decision maker is "realistic" and adjusts a_n into the direction of $\pi(s, e)$. Formally, a_{n+1} is a convex combination of the old aspiration level a_n and the payoff $\pi(s, e)$ whereby the weight attached to $\pi(s, e)$ is a fixed parameter $\beta \in [0, 1)$ that measures the speed of adjustment of the aspiration level.[3]

[2] Notice that we can take this expression to be a weight because we assumed earlier that payoffs and aspiration level are between zero and one.

[3] Note that we allow β to be zero so that our model includes the case of a fixed exogenous aspiration level as a special case.

Formally, if we define $\alpha \equiv \pi(s, e) - a_n$, then the learning rule in the case $\pi(s, e) \geq a_n$ is

$$
\begin{aligned}
p_{n+1}(s) &= (1 - \alpha)p_n(s) + \alpha \\
p_{n+1}(\tilde{s}) &= (1 - \alpha)p_n(\tilde{s}) \qquad \text{for } \tilde{s} \neq s \\
a_{n+1} &= (1 - \beta)a_n + \beta\pi(s, e)
\end{aligned} \tag{1}
$$

If $\pi(s, e) \leq a_n$, we assume that the decision maker takes this as discouragement to play s. He or she shifts probability away from s. The probability of the other strategy is accordingly increased. The size of the decrease in the probability assigned to s is proportional to the size of the difference $a_n - \pi(s, e)$. The aspiration level is adjusted as before.

Formally, if we now define $\alpha \equiv \mid \pi(s, e) - a_n \mid$, then the learning rule in the case $\pi(s, e) \leq a_n$ is

$$
\begin{aligned}
p_{n+1}(s) &= (1 - \alpha)p_n(s) \\
p_{n+1}(\tilde{s}) &= (1 - \alpha)p_n(\tilde{s}) + \alpha \qquad \text{for } \tilde{s} \neq s \\
a_{n+1} &= (1 - \beta)a_n + \beta\pi(s, e)
\end{aligned} \tag{2}
$$

This completes the definition of the learning rule. For a given decision problem, there are three free parameters of the learning rule: the initial values p_0 and a_0 and the parameter β. Since we are interested in the formation of aspiration levels, and since this is determined by the parameters a_0 and β, we define the following shorthand terminology:

DEFINITION 2. An *aspiration formation rule* $\mathcal{A}$ is a pair $(a_0, \beta) \in [0, 1] \times [0, 1)$.

For given parameters p_0, a_0, and β, the learning rule implies that (p_n, a_n) $(n \in \mathbb{N} \cup \{0\})$ is a discrete time Markov process with state space $\Delta \times [0, 1]$. To proceed, we shall construct a continuous time approximation of this process.

3. THE CONTINUOUS TIME LIMIT

3.1. *Construction of the Continuous Time Limit*. We shall first define the continuous time model, and then we shall explain the sense in which it approximates the discrete time model. We denote time by $t \in \mathbb{R}_+$. At each point in time t the decision maker is described by a probability distribution over his or her strategies, $p_t \in \Delta$, and by an aspiration level, $a_t \in [0, 1]$. These variables will be differentiable functions of time t. The derivative of each variable with respect to t is equal to the *expected* movement of the stochastic learning process of the preceding section.

Formally, denote by $E[\cdots \mid \cdots]$ the expected value of the random variable indicated before the vertical line conditional on the event indicated after the vertical line.

Then we assume for both strategies $s \in S$

$$\frac{dp_t(s)}{dt} = E\big[p_{n+1}(s) - p_n(s) \mid p_n = p_t \text{ and } a_n = a_t\big] \tag{3}$$

and for the aspiration level a_t

$$\frac{da_t}{dt} = E\big(a_{n+1} - a_n \mid p_n = p_t \text{ and } a_n = a_t\big) \tag{4}$$

The first of these equations says that the derivative of $p_t(s)$ with respect to time is equal to the expected change in $p_n(s)$ that would occur in the discrete time model of Section 2 if p_n were equal to p_t and a_n were equal to a_t. The second equation contains an analogous statement for a_t. Here, expected values are taken *before* a (pure) strategy is actually chosen and a state of the world is realized.

We give explicit formulas for the expected values in the preceding equations in the next subsection. In the remainder of this subsection we discuss the relation between the preceding equations and the learning process. We only give an informal description. A precise result is stated in the context of a related model in our earlier article (Proposition 1 in Börgers and Sarin, 1997). The result given there is, in turn, based on a result due to Norman (Theorem 1.1 of Chapter 8 of Norman, 1972).

Suppose that in each time interval $[\tau, \tau + 1] \subset \mathbb{R}_+$ there are N independent trials, i.e., N opportunities to take a decision and to experience the payoff resulting from this decision. The amount of "real" time that passes between two trials is $1/N$. Suppose that after each trial the decision maker changes his or her strategy and his or her aspiration level by $1/N$ of the amount assumed in Equations (1) and (2). Now let N tend to infinity, keeping the initial values p_0 and a_0 fixed, and ask where the process is at a particular time $t \in \mathbb{R}_+$.[4] As N tends to infinity, the variance of strategy and aspiration level[5] at time t tends to zero, and the expected value tends to the solution of differential Equations (3) and (4), evaluated at time t. Thus, by solving the differential equations, we obtain for any finite t a good prediction of the state variables of our learning process in the case that N is very large.

Notice that in the preceding paragraph we did not refer to the asymptotic behavior for $t \to \infty$. As we explain in Börgers and Sarin (1997), the asymptotic behavior of the learning process in discrete time may be different from the asymptotic behavior of the solution of (3) and (4). In other words, if one takes *first* the limit for $t \to \infty$ and *then* the limit for $N \to \infty$, one may obtain results that are different from those which one obtains if one takes *first* the limit $N \to \infty$ and *then* the limit $t \to \infty$. In this article we focus on the second order of limits. The differential equations we study are frequently used to study the long-term behavior (e.g., Benveniste et al., 1990; Binmore et al., 1995) of the associated stochastic dynamic model.

[4] More precisely, consider the state of the process after $n \in \mathbb{N}$ iterations, whereby n depends on N and as N tends to infinity we have $n/N \to t$.

[5] Both are, of course, for any finite N, random variables.

3.2. *Interpreting the Differential Equation*. We shall now calculate the expected values on the right-hand sides of differential equations (3) and (4). We shall write the formulas in a way that leads to a simple and interesting interpretation. Recall that the expected values relate to what would happen in the discrete time model if, at iteration n, the current value of p_n were p_t and the current value of a_n were a_t. We need to introduce some new notation that relates to this hypothetical situation. For simplicity, we shall not reiterate explicitly, neither in the text nor in the notation, that all probabilities and all expected values to which we refer in this subsection are meant to be conditional on $p_n = p_t$ and $a_n = a_t$.

Consider some strategy $s \in S$. There are two events in the discrete time model that can lead to an increased probability for strategy s in iteration $n+1$. One is that s is played and that a payoff above the aspiration level is experienced. The other is that $\tilde{s} \neq s$ is played and that a payoff below the aspiration level is experienced. Call the total probability of these two events together $\sigma_t(s)$. We shall refer to this probability as the *probability of strategy s receiving a benefit*.

The extent to which the probability of s is increased in either of these two events depends, first, on the extent to which the payoff received deviates from the aspiration level and, second, on the probability with which s is currently played. We wish to measure the first of these two influences. Define $\alpha_t \equiv \mid \pi(s, e) - a_t \mid$. We denote by $E[\alpha_t(s)]$ the expected value of α_t conditional on the event that s receives a benefit, i.e., conditional on the event the probability of which we denoted earlier by $\sigma_t(s)$.[6] We shall refer to $E[\alpha_t(s)]$ also as the *expected benefit of strategy s*. Finally, we denote by $E(\alpha_t)$ the unconditional[7] expected value of α_t, and we denote by $E(\pi_t)$ the expected payoff.

To clarify these definitions, we give an example. Consider the decision problem in Figure 1. Here, rows correspond to strategies, and columns correspond to states of the world. At the top of each column we have indicated the probability with which the corresponding state of the world occurs. In the intersections of rows and columns we have indicated payoffs.

Suppose that the current probability of strategy s_1, $p_t(s_1)$, is $\frac{1}{3}$ and that the current aspiration level is $a_t = 0.4$. Then the variables defined above have the following values (where we restrict attention to strategy s_1):

$$\sigma_t(s_1) = \frac{1}{3} \cdot \frac{3}{4} + \frac{2}{3} \cdot \frac{1}{4} = \frac{5}{12}$$

$$E[\alpha_t(s_1)] = \frac{1}{\sigma_t(s_1)} \cdot \left(\frac{1}{3} \cdot \frac{3}{4} \cdot 0.4 + \frac{2}{3} \cdot \frac{1}{4} \cdot 0.1 \right) = \frac{7}{25}$$

[6] To simplify the notation, we do not indicate explicitly in the notation that we are conditioning on this event.

[7] Of course, we still condition on $p_n = p_t$ and $a_n = a_t$. We write *unconditional* only to indicate that we are not conditioning on the event that some particular strategy is successful.

	0.75	0.25
s_1	0.8	0.1
s_2	0.6	0.3

FIGURE 1

$$E(\alpha_t) = \frac{1}{3} \cdot \frac{3}{4} \cdot 0.4 + \frac{1}{3} \cdot \frac{1}{4} \cdot 0.3 + \frac{2}{3} \cdot \frac{3}{4} \cdot 0.2 + \frac{2}{3} \cdot \frac{1}{4} \cdot 0.1 = \frac{29}{120}$$

$$E(\pi_t) = \frac{1}{3} \cdot \frac{3}{4} \cdot 0.8 + \frac{1}{3} \cdot \frac{1}{4} \cdot 0.1 + \frac{2}{3} \cdot \frac{3}{4} \cdot 0.6 + \frac{2}{3} \cdot \frac{1}{4} \cdot 0.3 = \frac{67}{120}$$

Using the notation introduced so far, we can now rewrite the expected values on the right-hand sides of differential Equations (3) and (4). Since the two probabilities $p_t(s_1)$ and $p_t(s_2)$ add up to one, it suffices to write just one equation for the probabilities. The following equations result from straightforward calculations, and therefore, we omit their proof.

$$\frac{dp_t(s_1)}{dt} = p_t(s_1)\{E[\alpha_t(s_1)] - E(\alpha_t)\} + E[\alpha_t(s_1)][\sigma_t(s_1) - p_t(s_1)] \tag{5}$$

and

$$\frac{da(t)}{dt} = \beta[E(\pi_t) - a_t] \tag{6}$$

Consider the two summands on the right-hand side of Equation (5). The first term has the form of the standard replicator equation from evolutionary biology, with the exception that "payoffs" are replaced by "benefits." To understand the structure of this term, suppose for the moment the second term were zero. If $p_t(s_1) \neq 0$, we can divide both sides of Equation (5) by $p_t(s_1)$, and we find that the *relative* change in $p_t(s_1)$ is equal to the difference between the expected benefit of strategy s_1 and the expected benefit of all strategies. This is what also happens in replicator dynamics, with the exception that in the replicator dynamics it is "payoffs" rather than "benefits" that matter. In our learning model it is clear that benefits rather than payoffs determine a strategy's success.

Consider now the second term on the right-hand side of Equation (5). Suppose for the moment the first term were zero. The sign of the second term is the same as the sign of $\sigma_t(s_1) - p_t(s_1)$. As a consequence, if $\sigma_t(s_1) \geq p_t(s_1)$, then $p_t(s_1)$ will increase, and if $\sigma_t(s_1) \leq p_t(s_1)$, then $p_t(s_1)$ will decrease. If this term alone were active, and if $\sigma_t(s_1)$ converged for $t \to \infty$, then it would have to be the case that $p_t(s_1)$ also converged and that $\lim_{t\to\infty} p_t(s_1) = \lim_{t\to\infty} \sigma_t(s_1)$. Hence, asymptotically, the decision maker would equate the probability with which s_1 is chosen and the probability with which s_1 receives a benefit. If we think of the event that s_1 receives a benefit as the event that s_1 is "successful," then this amounts to probability matching in the sense explained in the Introduction. We can hence say that the second term of the preceding differential equation pulls the decision maker into the direction of probability matching.

Thus we find that the differential equation for $p_t(s_1)$ contains exactly two terms, the first of which reflects a version of replicator dynamics and the second of which reflects a version of probability matching. There are no other forces active in this differential equation, and these two forces enter additively.

Consider now the differential equation for a_t. The sign of the right-hand side is identical to the sign of $E(\pi_t) - a_t$. Hence a_t moves into the direction of the expected payoff. This reflects the "realism" in the decision maker's aspiration level that we assumed in Section 2.

3.3. *Two Extreme Cases*. To develop further intuition for differential Equations (5) and (6), we consider in this subsection two extreme cases. In the first case only the replicator force will be present, whereas in the second case only the probability-matching force will be present. In both cases we assume that $\beta = 0$, and hence we abstract from movements in the aspiration level. The aspiration level therefore will remain for all t at its exogenous initial level, a_0.

The first case is that the initial aspiration level is below all feasible payoffs; i.e., $a_0 \leq \pi(s, e)$ for all $s \in S$ and $e \in \mathscr{E}$. In this case, the decision maker experiences all outcomes as pleasant and reinforcing. He or she lives in a heavenly world. His or her behavior nevertheless evolves because outcomes differ in reinforcement *strength*. The differential equation for $p_t(s_1)$ reduces in this case to the standard replicator equation:

$$\frac{dp_t(s_1)}{dt} = p_t(s_1)\{E[\pi(s_1)] - E(\pi_t)\} \tag{7}$$

Here we write $E[\pi(s_1)]$ for the expected payoff of strategy s_1.

To see that this equation is correct, notice first that in the case that we are considering the probability matching effect equals zero. This is so because the only way in which strategy s_1 can receive a benefit is by being played. Hence the probability with which action s_1 receives a benefit, $\sigma_t(s_1)$, will equal the probability with which s_1 is played, $p_t(s_1)$, for all t. As a consequence, the probability-matching term will always equal zero.

This leaves the replicator term. In general, the replicator term in our model refers to "benefits," whereas the replicator equation conventionally refers to "payoffs." However, in the case that we are considering, this distinction does not matter. This is so because in this case benefits are equal to payoffs received minus the (constant) aspiration level. Hence *differences* of benefits, as they appear in the replicator term, are equal to *differences* of payoffs. Therefore, learning Equation (5) is exactly the same as the replicator equation.

It is well known that in the replicator process the weight attached to strategies that maximize the expected payoff converges to one as time tends to infinity.[8] Hence the first extreme case considered here is one in which the learning process finds the optimal strategy.[9]

[8] Recall that we have assumed that both strategies have initially positive weight.

[9] In this special case of low and fixed aspirations in which all payoffs are positive, our result can be shown to extend (by the results in Börgers and Sarin, 1997) to the situation in which the agent has a finite number of strategies.

	μ	1-μ
s_1	x	y
s_2	y	x

FIGURE 2

In the second case, Equation (5) will reduce to pure probability matching. We shall hence eliminate the replicator term. For this we assume that there are only two possible values of payoffs and that these are exactly symmetric on either side of the aspiration level. In other words, the decision maker experiences either a "success" or a "failure," and the "size" of these two experiences is exactly identical. Formally, this is the requirement that $|\pi(s, e) - a_0| = c$ for all $s \in S$ and $e \in \mathscr{E}$ and for some constant $c > 0$. Under this assumption, the expected benefit of each of the two strategies is equal to c. Therefore, the replicator term of Equation (5) equals zero, and we are left with the probability-matching term:

$$\frac{dp_t(s_1)}{dt} = c[\sigma_t(s_1) - p_t(s_1)] \tag{8}$$

We mentioned already in the preceding subsection that this implies $\lim_{t\to\infty} p_t(s_1) = \lim_{t\to\infty} \sigma_t(s_1)$, provided that $\sigma_t(s_1)$ converges for $t \to \infty$. Unfortunately, it is in general not immediate that $\sigma_t(s_1)$ converges, since $\sigma_t(s_1)$ may depend on $p_t(s_1)$. A case in which convergence of $\sigma_t(s_1)$ is obvious is the case in which $\sigma_t(s_1)$ does not depend on $p_t(s_1)$. Figure 2 represents such a case. Here, we assume that $\mu \in (0, 1)$, that $0 < y < x < 1$, and that $a_0 = (x + y)/2$. In this case, Equation (8) reduces to

$$\frac{p_t(s_1)}{dt} = c[\mu - p_t(s_1)] \tag{9}$$

and it is clear that $p_t(s_1) \to \mu$ for $t \to \infty$. Thus we have a simple case of asymptotic probability matching.

4. ASYMPTOTIC OPTIMIZATION

4.1. *Necessary and Sufficient Conditions*. In this section we investigate whether, in the long run, the decision maker benefits from having an endogenous aspiration level. We use the continuous time approximation developed in the preceding section. We focus on the limit $t \to \infty$.

In the continuous time approximation, if the decision maker's behavior converges for $t \to \infty$, it converges to a rest point of differential Equations (3) and (4). We therefore begin with the following definition:

DEFINITION 3. Consider a given decision problem $\mathscr{D}$ and a given aspiration formation rule $\mathscr{A}$. A *rest point* of differential Equations (3) and (4) is a pair $(p^*, a^*) \in \Delta \times [0, 1]$ for which the right-hand sides of Equations (3) and (4) equal zero.

Of course, our concern is not only with the existence of certain rest points but also with the dynamic stability of these rest points. Therefore, we introduce the following definition:

DEFINITION 4. Consider a given decision problem $\mathcal{D}$ and a given aspiration formation rule $\mathcal{A}$. A rest point (p^*, a^*) of differential Equations (3) and (4) is *globally asymptotically stable* if the solution of differential Equations (3) and (4) converges for $t \to \infty$ to this rest point from all initial points $p_0 \in \Delta$ that satisfy $p_0(s) \neq 0$ for both $s \in S$.

We can now define *optimality* of an aspiration formation rule:

DEFINITION 5. An aspiration formation rule $\mathcal{A}$ is *optimal* in the decision problem $\mathcal{D}$ if differential Equations (3) and (4) have a rest point (p^*, a^*) with $p^*(s_1) = 1$ and this rest point is globally asymptotically stable.

In this subsection we provide necessary and sufficient conditions for an aspiration formation rule to be optimal. In the next subsection we shall supplement the analytical results of this subsection with some numerical simulations.

As a benchmark case we consider first the case that the aspiration level is exogenous ($\beta = 0$).

PROPOSITION 1. *For any decision problem $\mathcal{D}$ there is an $\bar{a} \in (0, 1)$ such that an aspiration formation rule $\mathcal{A}$ which satisfies $\beta = 0$ is optimal in the decision problem $\mathcal{D}$ if and only if $a_0 \leq \bar{a}$.*

In words, this result says that with an exogenous and fixed aspiration level, the decision maker optimizes asymptotically if and only if the aspiration level is below some threshold $\bar{a}$. The value of $\bar{a}$ may depend on the decision problem at hand.

The formal proof of Proposition 1 is in the Appendix. It is easy to obtain some intuition for the result. If the exogenous aspiration level a_0 is smaller than the payoff $\pi(s, e)$ for all $s \in S$ and $e \in \mathcal{E}$, then the learning process with fixed aspirations is in the continuous time limit equivalent to replicator dynamics, and it is well known that replicator dynamics asymptotically optimize in decision problems. On the other hand, if the exogenous aspiration level a_0 is larger than the minimum payoff that is possible when strategy s_1 is played, then the probability-matching effect makes it impossible that strategy s_1 is played with probability 1, since sometimes strategy s_1's payoff will be below the aspiration level, and hence strategy s_2 will have a positive probability of success. Probability matching will then imply that the decision maker plays strategy s_2 asymptotically with positive probability.

The preceding arguments refer only to extreme values of a_0. Proposition 1 deals, in addition, with intermediate values of a_0, and asserts that there is a unique threshold that separates those aspiration values which induce asymptotically optimal choices from those that do not. Showing this constitutes the main formal difficulty in the proof. Readers of the proof will notice that the proof also provides a simple method for calculating the threshold $\bar{a}$ for any given decision problem $\mathcal{D}$.

We now turn to the case of an endogenous aspiration level, i.e., $\beta > 0$. To be able to state our result for this case, we need some additional terminology:

DEFINITION 6. In a given decision problem $\mathcal{D}$, the strategy s_1 is called

- *Safe* if $\pi(s_1, e) = \pi(s_1, \tilde{e})$ for all $e, \tilde{e} \in \mathcal{E}$.
- *Dominant* if $\pi(s_1, e) \geq \pi(s_2, e)$ for all $e \in \mathcal{E}$.

PROPOSITION 2. (*i*) *Consider a decision problem $\mathcal{D}$ in which s_1 is safe and dominant. Then any aspiration formation rule $\mathcal{A}$ that satisfies $\beta > 0$ is optimal in $\mathcal{D}$.*

(*ii*) *Consider a decision problem $\mathcal{D}$ in which s_1 is not safe or not dominant. Then no aspiration formation rule $\mathcal{A}$ that satisfies $\beta > 0$ is optimal in $\mathcal{D}$.*

We give the formal proof of Proposition 2 in the Appendix. Here we only discuss the intuition behind the result. First, it is relatively easy to show that an aspiration formation rule that lets the aspiration level move endogenously is indeed optimal if the expected payoff-maximizing strategy is safe and dominant. The more difficult part of the proof is the proof of the second part of the proposition. Suppose first that the optimal strategy were not safe. If $p_t(s_1)$ were to converge for $t \to \infty$ to 1, then the aspiration level would have to converge to the expected payoff achieved by s_1. This is an immediate implication of the differential equation for a_t. Since s_1 is not safe, this would imply that in the long run there would be a positive probability of s_1's payoff falling below aspiration level and s_2 being successful. As in the context of Proposition 1, probability matching would then induce the decision maker to choose s_2 with positive probability and hence would make asymptotic optimization impossible.

The case that s_1 is safe but not dominant is more difficult. In this case, if s_1 is played with probability of almost one in the discrete time model, all possible changes in the probability of s_1 will either be very small or will occur with very low probability only. However, the negative effects outweigh the positive effects in order of magnitude, and hence $dp(s_1)/ds_1 < 0$ if $p_t(s_1)$ is close to one. This is what the formal argument in the Appendix demonstrates. It is the main formal difficulty in the proof of Proposition 2.

We now summarize our results in a diagram. Consider a given decision problem $\mathcal{D}$ and a given aspiration formation rule $\mathcal{A}$. Call the initial aspiration level a_0 high if it is above the threshold $\bar{a}$ of Proposition 1. Otherwise, call it low. Figure 3 indicates in which cases the aspiration formation rule is optimal. In each box of the figure there is a cross ($\times$) if an aspiration formation rule with exogenous aspiration level optimizes, and there is a circle ($\circ$) if an aspiration formation rule with endogenous aspiration level optimizes.

Figure 3 suggests a simple extension of our results. So far we have asked for a given decision problem $\mathcal{D}$ and a given aspiration formation rule $\mathcal{A}$ whether the aspiration formation rule is optimal in $\mathcal{D}$. In reality, however, learning rules have to deal with a large set of decision problems, not just with a single-decision problem. It is therefore natural to ask which aspiration formation rules are optimal for a large set of decision problems. A simple corollary of Propositions 1 and 2 is

	s_1 is safe and and dominant	s_1 is not safe or not dominant
high initial aspirations	◯	
low initial aspirations	⊗	×

FIGURE 3

COROLLARY 1. *An aspiration formation rule $\mathcal{A}$ is optimal in all decision problems $\mathcal{D}$ if and only if $a_0 = 0$ and $\beta = 0$.*

Corollary 1 shows that among the aspiration formation rules that we consider here, only those are optimal in a variety of decision problems which lead to learning behavior that imitates, in a sense, evolution. We have referred to related results in the Introduction.

The proof of Corollary 1 is obvious from Figure 3. If $\beta > 0$, the aspiration formation rule will not be optimal in decision problems in which the strategy s_1 is not safe or not dominant. If $\beta = 0$ but $a_0 > 0$, the aspiration formation rule will not be optimal in decision problems in which $a_0 < \pi(s, e)$ for some $e \in \mathcal{E}$. On the other hand, if $a_0 = 0$ and $\beta = 0$, then the aspiration formation rule will lead to learning behavior that, in the continuous time limit, is in *all* decision problems the same as replicator dynamics and hence asymptotically optimizes.

4.2. *Simulations*. The results summarized in Figure 3 show that there are two cases in which the comparison between learning with *exogenous* aspiration level and learning with *endogenous* aspiration level is straightforward. First, if the optimal strategy is safe and dominant, and if the initial aspiration level is too high, then it is better to have an endogenous aspiration level. Second, if the optimal strategy is not safe or not dominant, and if the initial aspiration level is sufficiently low, then it is better to keep the aspiration level fixed and not to adjust it endogenously. We begin this subsection with two simulations that illustrate these two cases.

The first simulation concerns a decision problem under certainty, i.e., a decision problem in which the set $\mathcal{E}$ has only one element. This is the simplest case of a decision problem in which the expected payoff-maximizing action is both safe and dominant. The decision problem that we consider is displayed in Figure 4. Figure 5 shows a numerically obtained[10] phase diagram for this decision problem. This phase diagram refers to the case that the aspiration level is endogenous. For the simulation, we have set $\beta = 0.1$. The phase diagram shows the simultaneous movements of the probability $p_t(s_1)$ of playing the better strategy and of the aspiration level a_t.

All trajectories in Figure 5 converge to the rest point in which $p^*(s_1) = 1$ and $a^* = 0.6$. The aspiration formation rule is optimal, as Proposition 2 asserts. Notice that it

[10] To construct the numerical phase diagrams in this article, we used MATHEMATICA.

s_1	0.6
s_2	0.3

FIGURE 4

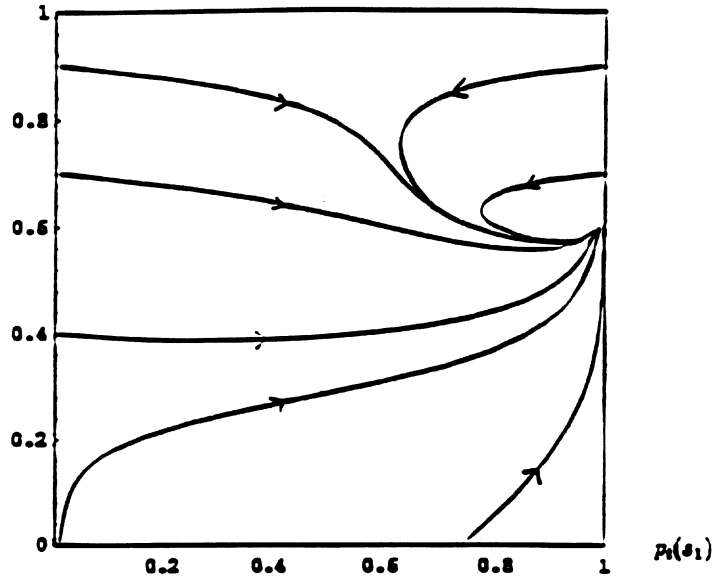

FIGURE 5

is obvious from analytical considerations, though not from Figure 5, that the learning process has an additional rest point at $p^*(s_1) = 0$ and $a^* = 0.3$. This rest point's basin of attraction is, however, of measure zero. Only those trajectories which start with initial values satisfying $p_0(s_1) = 0$ and $a_0 \leq 0.3$ converge to this rest point.

Of particular interest in Figure 5 are those trajectories which begin with a "too high" aspiration level, say, an aspiration level above 0.6. In these cases, the decision maker would not asymptotically optimize if the aspiration level were kept fixed. By contrast, with an endogenously moving aspiration level, the decision maker does optimize asymptotically.

To explain how endogenous movements in the aspiration level bring about asymptotic optimization, we consider as an example the trajectory that begins in the top right corner of the state space. The initial values for this trajectory are $p_0(s_1) = 0.99$ and $a_0 = 0.9$. Hence the decision maker chooses the payoff-maximizing strategy s_1 with an initial probability close to 1. However, his or her aspiration level is far too high. Therefore, he or she is disappointed by the payoff which he or she receives when playing s_1 and hence shifts probability to the alternative strategy s_2. At the same time, he or she adjusts his or her aspiration level into the direction of the experienced payoffs, i.e., downward. Thus the trajectory points into the interior of the state space.

As the state variables move along this trajectory, two effects take place. First, the decision maker gathers experience with the strategy s_2 and is disappointed by this strategy as well. Second, the aspiration level is gradually reduced. As the aspiration level approaches 0.6, the payoff associated with strategy s_1, the size of the decision maker's disappointment with s_1 tends to zero. These two effects lead to a reversal in the downward trend of the probability with which s_1 is played. In the long run, as $t \to \infty$, the decision maker returns to playing s_1 with high probability, but he or she now holds a more realistic aspiration level, and hence the situation becomes stable.

Next, we give an example in which the expected payoff-maximizing strategy is not safe. Hence in this example an aspiration formation rule with fixed and sufficiently low aspiration level would be optimal; however, an aspiration formation rule with endogenous aspiration level is not optimal. The example is shown in Figure 6, and the corresponding phase diagram of the process with moving aspiration level ($\beta = 0.1$) is shown in Figure 7.

Figure 7 suggests that the learning process with endogenous aspiration level has a globally asymptotically stable rest point that is in the interior of the state space. Hence the asymptotic probability of the expected payoff-maximizing strategy is not equal to one, and the aspiration formation rule is not optimal. This confirms Proposition 2.

It is particularly interesting to trace trajectories that start with a low aspiration level, say, an aspiration level below 0.3. If the decision maker kept the aspiration

	0.5	0.5
s_1	0.7	0.5
s_2	0.3	0.3

FIGURE 6

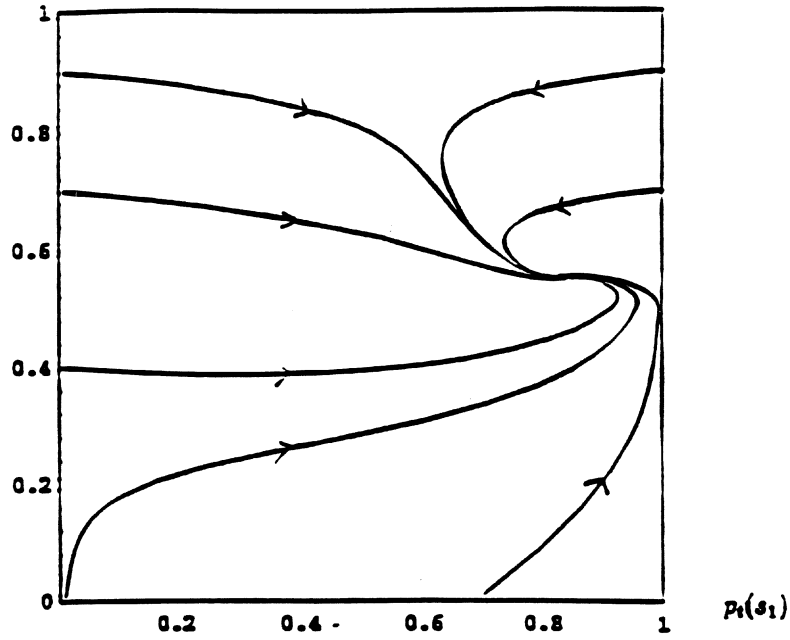

FIGURE 7

level fixed, then he or she ultimately would play the optimal strategy with probability one. The endogenous increase in the aspiration level prevents this from happening.

Consider as an example the trajectory that begins in the point $p_0(s_1) = 0.7$ and $a_0 = 0.01$. If the decision maker starts in this point, the probability of s_1, and also the aspiration level, will increase initially. This continues until the aspiration level reaches, roughly, 0.5, the minimum payoff possible under strategy s_1. When the aspiration level reaches this value, the probability of s_1 has already almost reached 1. The endogenous aspiration level adjustment forces the aspiration level to move further, since the expected payoff is larger than 0.5. But once the aspiration level exceeds 0.5, the probability-matching effect starts to affect the decision maker's behavior. He or she becomes disappointed by the strategy s_1 and tries again the alternative strategy s_2. The probability $p_t(s_1)$ therefore decreases. This continues until a rest point is reached.

So far we have focused on examples in which the results of the preceding subsection allow an unambiguous comparison of learning with and without an endogenous aspiration level. We now turn to cases in which such a comparison is not possible on the basis of the results of the preceding subsection.

Consider first cases in which the optimal strategy is safe and dominant and in which the initial aspiration level is sufficiently low. In such cases, the decision maker will learn to play the optimal strategy independent of whether he or she adjusts his or her aspiration level or not. As long as we focus on the asymptotics of the decision maker's behavior, nothing additional can be said about this case.

Consider next decision problems in which the optimal strategy is not safe or not dominant and in which the initial aspiration level is too high. In such cases, the decision maker will *not* learn to play the optimal strategy independent of whether he or she adjusts his or her aspiration level or not. However, in such cases, it is conceivable that under one of the two types of learning rules the decision maker's asymptotic performance is "less bad" than under the other. We illustrate this with the example in Figure 8, which is a special case of the example in Figure 3. Figure 9 shows the phase diagram of the learning rule with endogenous aspiration level ($\beta = 0.1$) for this example.

If the decision maker's initial aspiration level in this example is exactly in the middle of the two possible payoff values, i.e., if $a_0 = 0.5$, the learning rule with fixed aspiration level will lead to "pure" probability matching; i.e., the strategy s_1 will be chosen with probability 0.8. This follows from the calculations in Subsection 3.3.

	0.8	0.2
s_1	0.8	0.2
s_2	0.2	0.8

FIGURE 8

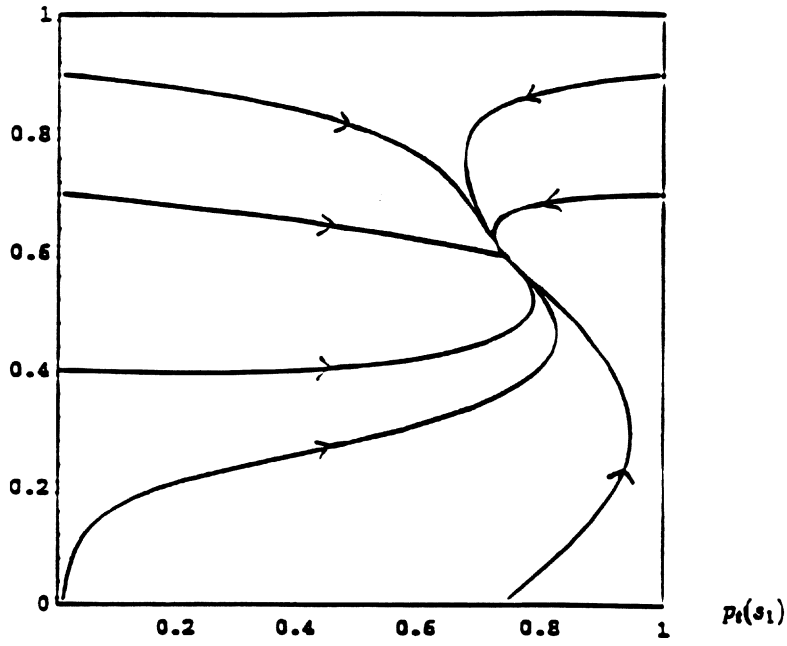

FIGURE 9

For the case that the aspiration level is endogenous, Figure 9 suggests that the learning process is globally asymptotically stable. An interesting question to ask is whether in the unique rest point in Figure 9 the decision maker does better or worse than with pure probability matching. The somewhat surprising answer is that the decision maker does *worse* if he or she adjusts his or her aspiration level. The asymptotic probability of choosing the strategy s_1 turns out to be less than 0.8.

To explain the intuition for this, we show in Figure 10 a trajectory that starts in the point of pure probability matching: $p_0(s_1) = 0.8$ and $a_0 = 0.5$. Starting from this point, there will be a tendency for a_t to increase. The reason is that in the initial point, a_0 is below the current expected payoff. If the decision maker played both strategies with equal probability, his or her expected payoff would exactly equal a_0. However, in the initial point he or she plays the strategy with higher payoff more often, and hence a_0 is smaller than the current expected payoff.

In the initial point there will be no tendency for $p_t(s_1)$ to change, and hence the trajectory points vertically upward in the phase diagram. However, once the aspiration level has increased, there also will be pressure on $p_t(s_1)$ to change. To see why this pressure works *against* s_1, notice first that an increase in the aspiration level will reduce the size of successes but increase the size of failures. Therefore, those strategies which are mainly sustained by the failure of other strategies will benefit. Now consider the point of pure probability matching. In this point the probability of success of strategy s_1 is 0.64, and the probability of failure of strategy s_2 is 0.16. Hence s_1 is mainly sustained by successes. By contrast, the probability of success of strategy s_2 is 0.04, and the probability of failure of strategy s_1 is 0.16. Hence strategy s_2 is mainly sustained by failures of s_1. It is for this reason that an increase in the

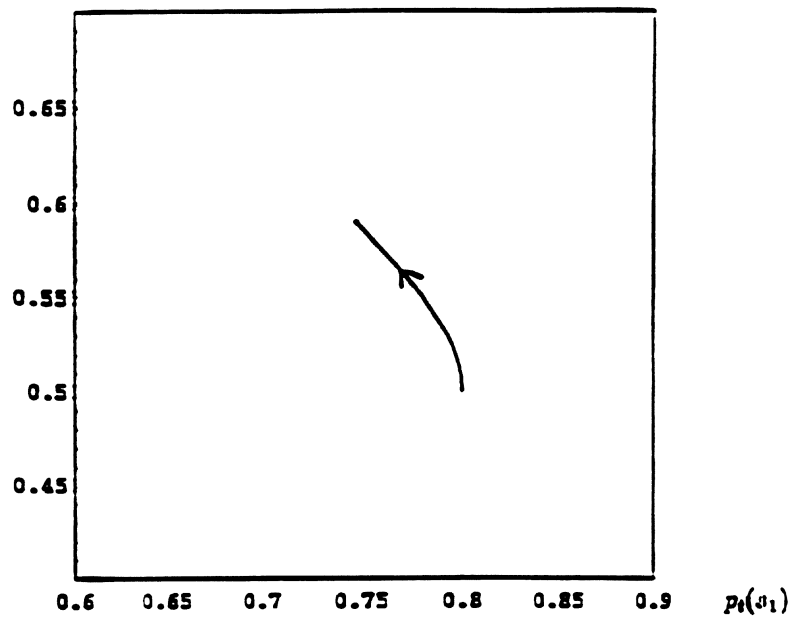

FIGURE 10

aspiration level reduces the probability with which s_1 is played and increases the probability with which s_2 is played.

We now generalize the preceding observation. We consider the class of examples given by Figure 2. We adopt the assumptions concerning x, y, and μ that were introduced in the context of Figure 2. We then have the following result:

PROPOSITION 3. *If the decision problem $\mathcal{D}$ is given by Figure* 2, *and if the aspiration formation rule $\mathcal{A}$ satisfies $\beta > 0$, then there is a unique rest point (p^*, a^*) of differential Equations* (3) *and* (4). *This rest point satisfies* $0.5 < p^*(s_1) < \mu$.

The formal proof of this result is in the Appendix. The intuition behind the result is the same as the intuition that we explained earlier in the context of Figure 10. Observe that Proposition 4 does not make any assertion concerning the asymptotic stability of the rest point. Our simulations suggest that it is globally asymptotically stable, but we have not been able to prove this.

The formal and numerical results of this section suggest the following conjecture: If the asymptotic aspiration level of the decision maker is above the initial aspiration level, the aspiration level adjustment cannot improve the decision maker's performance. In the opposite case, the aspiration level adjustment cannot worsen the decision maker's performance. Unfortunately, we have been unable to prove this conjecture.

5. RELATED LITERATURE

The idea that reinforcement learning procedures will "behave well" in decisions under risk only if they imitate evolution has been formalized previously in articles by Sarin (1995) and Schlag (1994). Both articles consider relatively large classes of learning procedures, introduce certain axioms, and then show that the only learning processes that satisfy these axioms are those which are, in some way, equivalent to replicator dynamics. Neither of these two articles, however, allows for an endogenous aspiration level.

A related recent study that investigates the consequences of endogenous movements of the aspiration level is that of Gilboa and Schmeidler (1996). They consider the same type of decision problem as we do and study the following learning rule: In each period the decision maker assesses the past performance of each strategy by looking back at all those previous periods in which this strategy was chosen and summing the differences between the payoffs received in those periods and his or her (current) aspiration level. The decision maker chooses that strategy for which this sum is largest. The aspiration level in the next period is a weighted average of the current aspiration level and the maximum average performance of any strategy in the past. Thus the state space of Gilboa and Schmeidler's learning rule is larger than the state space of the decision maker in our model. Moreover, Gilboa and Schmeidler's decision maker performs explicit maximizations. We think that our model is of interest in this context because it describes a less sophisticated decision maker who is still capable of achieving optimal decision making in the long run.

Gilboa and Schmeidler's main result is that if the initial aspiration level is sufficiently high, or if the decision maker arbitrarily raises it from time to time, then the decision maker will choose in the long run an expected payoff-maximizing action. The benefit of high aspiration levels in Gilboa and Schmeidler's framework is that they induce the decision maker to experiment sufficiently frequently with all actions. In our framework, because choice is stochastic rather than deterministic, as in Gilboa and Schmeidler, and because we construct a continuous time limit with "many" choices in each time interval, all actions automatically will be chosen sufficiently frequently. The argument that shows that high aspiration levels do not induce some form of probability matching in Gilboa and Schmeidler's framework is subtle. Roughly speaking, if all actions initially are chosen sufficiently frequently, then in the long run the performance measure for the optimal action will be so much better than the performance measure for nonoptimal actions that occasional bad payoff experiences cannot overcome this difference and induce the decision maker to switch actions.[11] Endogenous aspirations are also considered in Bereby-Meyer and Erev (1998) in the context of decision problems; however, the focus of that article is experimental, and no analytical results regarding the long-run behavior of the learning dynamics are provided.

Karandikar et al. (1998) also investigate learning with an endogenous aspiration level. They consider a two-player game in which each player has two strategies. They stipulate the following learning rule: If the decision maker receives a payoff above his or her aspiration level, then he or she sticks with his or her current strategy. Otherwise, he or she switches with some probability to the alternative strategy. The probability of switching is an increasing function of the difference between aspiration level and payoff. In each period, with a high probability, the aspiration level is adjusted toward the current payoff. With a small but positive probability, the aspiration level is perturbed and takes some random value. Karandikar et al.'s most striking results refer to the "Prisoner's dilemma." Karandikar et al. focus on the case that the trembles in the aspiration level are very unlikely and that the regular adjustment in the aspiration level occurs sufficiently slowly. They find that, in the long run, both players will cooperate with a probability close to one.

Relating Karandikar et al.'s results to ours is difficult because we consider a single-person decision problem, whereas they consider a game, and the endogenous changes in the payoff distributions that result from this play a key role in the rather subtle intuition for their result.[12] One point worth noting is that in Karandikar et al.'s model the "state space," i.e., the set of states in which the decision makers might find themselves at any given time, is the set of all pairs of a pure strategy and an aspiration level. In our model, by contrast, players are described by mixed strategies and an aspiration level. Our larger state space and our focus on the case in which the mixed strategies adjust slowly ensure that a large amount of experimentation is built into our learning model. By contrast, in Karandikar et al.'s model, experimentation occurs only if the aspiration level makes a large upward jump. Thus experimentation occurs very infrequently and is associated with high aspiration levels. This makes it easier in

[11] Kim (1995) and Pazgal (1997) extend Gilboa and Schmeidler's model to games and find the possibility of perpetual cooperation in the "Prisoner's dilemma."

[12] The intuition for their result is explained in detail in Section 4 of their article.

Karandikar et al's model for the decision maker to get trapped in a state in which he or she plays a dominated strategy.

Another model with endogenous aspiration levels recently has been analyzed by Palomino and Vega-Redondo (1998). They consider a continuum-size population. In each period, players are randomly matched to play the "Prisoner's dilemma." The learning rule is the same as in Karandikar et al. except that the aspiration adjustment is modeled differently. Each individual's aspiration level is adjusted in the direction of some statistic of play in the whole population, such as the average payoff. There are no random fluctuations in aspirations. Palomino and Vega-Redondo obtain asymptotically a mixed population in which some but not all players cooperate. The intuition for their result is related to the intuition for probability matching in our model. For each individual player, payoffs are asymptotically stochastic; they are sometimes above and sometimes below the aspiration level. Therefore, the strictly dominant strategy is not being learned. Specifically, in the "prisoner's dilemma," defecting players are disappointed by their payoffs if they meet other defectors, and they will switch to cooperation. The presence of cooperators implies that the aspiration level is indeed above the (defect-defect) payoff. Thus cooperation becomes self-sustaining.

A recent article by Dixon (1998) is similar to that of Palomino and Vega-Redondo (1998). One main difference is that there is no random matching in Dixon's model. Each pair of players stays together forever. Dixon finds that all players will cooperate in the long run.

6. EXTENSIONS

One of the assumptions of our model is that for given payoff and aspiration level, the size of behavior adjustments made by the decision maker is the same over time. One might argue that at later periods the decision maker will be more reluctant to change his or her behavior. This feature is present, for example, in the Erev and Roth (1998) model of reinforcement learning. A formal analysis of Erev and Roth's model, or related models, is beyond the scope of this article. One feature of their model that complicates its analysis is that the decrease in the step size of the decision maker's behavior adjustments is endogenous. The step size depends on the decision maker's experiences, which, in turn, depend on his or her choices. It becomes much easier to extend the results of this article if the decrease in the step size is made exogenous.[13] The simplest way of doing this is to postulate that at iteration n the change in the decision maker's choice probabilities is exactly $1/n$ of what it is in the model introduced in Section 2. It is well known from the stochastic approximation literature (Benveniste et al., 1990) that the asymptotic behavior of the resulting learning process is closely linked to that of the deterministic differential equation that we constructed in Section 2 as a continuous time approximation. In particular, if the differential equation has a unique, globally stable rest point, then the stochastic learning process will converge to this rest point with probability 1.

[13] We know of no analysis that would indicate whether an exogenous or an endogenous reduction in step size is a better model of real learning behaviour.

Our simulations suggest that the learning process with endogenous aspiration level typically has a unique globally stable rest point that is interior.[14] If this is correct, then the analysis of this article applies to the asymptotic behavior of the modified learning process as well. For the case that the aspiration level is exogenous and sufficiently low, the analysis is complicated by the fact that the approximating differential equation, the replicator equation, has multiple fixed points, namely, all pure strategies. However, since all fixed points except the optimal one have no basin of attraction, it should be easy to extend stochastic approximation results and to show convergence to the optimal action with probability 1.

Another important extension of our work would be to consider the case in which the decision maker has more than two actions. This extension raises technical problems. In the most natural extension of our model to the case of more than two actions, probability that is taken away from some action is distributed among the remaining actions in proportion to their current probabilities. Such a proportional rule, however, implies that expected motion is not a continuous function of the current state. This makes it impossible to appeal to standard theorems when taking continuous time limits. Moreover, if there are more than two actions, an even more sophisticated definition of probability matching will be needed. We expect that these problems can be resolved, but we have not yet done so. We expect the broad picture to remain the same as in the case of two actions.

Finally, it seems highly desirable to extend our work to the case that the decision maker faces a game rather than a single-person decision problem. The results of this article suggest that we should expect that learning rules with either an exogenous but relatively high aspiration level or an endogenously moving aspiration level have in many games interior rest points in which all players' behavior involves some combination of optimization and probability matching. In such rest points, one would expect randomization to be self-enforcing: Each decision maker's payoff is stochastic because the other players randomize. Moreover, as a result of the probability-matching effect, this randomness in payoffs induces each decision maker to randomize himself or herself. We conjecture that the only learning rules for which no such rest points exist are those for which the aspiration levels are exogenous and lower than all conceivable payoffs. The learning dynamics in this case become analogous to the replicator dynamics of evolutionary game theory. The details of this case are in Börgers and Sarin (1997).

APPENDIX

PROOF OF PROPOSITION 1. We consider a given and fixed decision problem $\mathcal{D}$. We proceed in three steps. The first step contains a preliminary observation. In the remaining two steps we construct the threshold $\bar{a}$.

STEP 1. In this step we prove that an aspiration formation rule $\mathcal{A}$ with $\beta = 0$ is not optimal if $a_0 > \min_{e \in \mathcal{E}} \pi(s_1, e)$. The proof is indirect. Suppose that the inequality

[14] In simple examples, it is also easy to prove this formally. We do not, however, have a general proof.

holds and that the aspiration formation rule is optimal. Then, for any initial value $p_0(s_1) \in (0, 1)$, we must have $p_t(s_1) \to 1$. Hence $p^*(s_1) = 1$ and $a^* = a_0$ must be a rest point of differential Equations (3) and (4). But this is impossible. To show that this is impossible, we show that $p^*(s_1) = 1$ and $a^* = a_0$ imply $dp_t(s_1)/dt < 0$. We calculate $dp_t(s_1)/dt < 0$ by considering the discrete time model of Section 2 and calculating the expected change in $p_n(s_1)$ conditional on $p_n(s_1) = 1$ and $a_n = a_0$. When the decision maker plays s_1, the payoff that he or she experiences will sometimes be below a_0 [because we assumed $a_0 > \min_{e \in \mathscr{E}} \pi(s_1, e)$]. If such payoffs are experienced, $p_n(s_1)$ will be reduced. It also may happen that the payoff is equal to or above a_0. In the latter case, $p_n(s_1)$ cannot be increased because it is already equal to 1. In expectations, therefore, $p_n(s_1)$ will decrease. Thus $dp_t(s_1)/dt < 0$. Hence $p^*(s_1) = 1$ and $a^* = a_0$ are not a rest point.

STEP 2. Suppose that $\min_{e \in \mathscr{E}} \pi(s_1, e) \leq \min_{e \in \mathscr{E}} \pi(s_2, e)$. Then we can set $\bar{a} = \min_{e \in \mathscr{E}} \pi(s_1, e)$. This is so because, for $a_0 \leq \bar{a}$, according to the argument in Subsection 3.3, differential Equation (3) becomes the replicator equation. Therefore, $p_t(s_1) \to 1$, provided that $p_0(s_1) \neq 0$, and hence the aspiration formation rule is optimal. If, on the other hand, $a_0 > \bar{a}$, then Step 1 shows that the aspiration formation rule is not optimal.

STEP 3. Suppose that $\min_{e \in \mathscr{E}} \pi(s_1, e) > \min_{e \in \mathscr{E}} \pi(s_2, e)$. Clearly, if $a_0 \leq \min_{e \in \mathscr{E}} \pi(s_2, e)$, for the same reason as in Step 2, the aspiration formation rule is optimal. Also, if $a_0 > \min_{e \in \mathscr{E}} \pi(s_1, e)$, by Step 1, the aspiration formation rule is not optimal. We must hence seek the threshold $\bar{a}$ in the interval $[\min_{e \in \mathscr{E}} \pi(s_2, e), \min_{e \in \mathscr{E}} \pi(s_1, e)]$.

Consider any $a_0 \in (\min_{e \in \mathscr{E}} \pi(s_2, e), \min_{e \in \mathscr{E}} \pi(s_1, e)]$. Consider also a fixed $p_t(s_1) \in (0, 1)$. We wish to calculate the right-hand side of Equation (3). We denote by $\Pi_1(a_0)$ the difference between the expected payoff of strategy s_1 and a_0. Next, we partition the set $\mathscr{E}$ into two subsets, $\mathscr{E}_2^-(a_0) \equiv \{e \in \mathscr{E} : \pi(s_2, e) < a_0\}$, and $\mathscr{E}_2^+(a_0) \equiv \{e \in \mathscr{E} : \pi(s_2, e) \geq a_0\}$.[15] We denote by $\mu_2^-(a_0)$ the probability of $\mathscr{E}_2^-(a_0)$ and by $\mu_2^+(a_0)$ the probability of $\mathscr{E}_2^+(a_0)$. Finally, we denote by $\Pi_2^-(a_0)$ the expected value of the difference between a_0 and s_2's payoff, conditional on $\mathscr{E}_2^-(a_0)$, and by $\Pi_2^+(a_0)$ the expected value of the difference between s_2's payoff and a_0, conditional on $\mathscr{E}_2^+(a_0)$.

A simple calculation shows that Equation (3) becomes

$$\frac{dp_t(s_1)}{dt} = (1 - p_t)\{p_t[\Pi_1(a_0) - \mu_2^+(a_0)\Pi_2^+(a_0)] + (1 - p_t)\mu_2^-(a_0)\Pi_2^-(a_0)\} \tag{A.1}$$

For any $p_t(s_1) \in (0, 1)$, the sign of the expression on the right-hand side is determined by the sign of the expression in braces. We therefore focus on this expression. Notice that for $p_t \to 1$, this expression tends to $A(a_0)$, which we define as follows:

$$A(a_0) \equiv \Pi_1(a_0) - \mu_2^+(a_0)\Pi_2^+(a_0) \tag{A.2}$$

[15] Note that $\mathscr{E}_2^+$ (but not $\mathscr{E}_2^-$) may be empty. In this case, the argument that follows needs to be modified. However, the required modifications are obvious, and therefore, we ignore this case.

Now suppose that $A(a_0) \geq 0$. Then the expression in braces in Equation (A.1) will be positive for all $p_t(s_1) \in (0, 1)$. This is so because it is a convex combination of the nonnegative expression $A(a_0)$ and of the expression $\mu_2^-(a_0)\Pi_2^-(a_0)$, which, because of $a_0 > \min_{e\in\mathscr{E}} \pi(s_2, e)$, is positive. If the right-hand side of Equation (A.1) is positive for all $p_t(s_1) \in (0, 1)$, then $p_t(s_1)$ must be monotonically increasing and hence converging. The limit must be a rest point. The only rest point that can be the limit is $p^*(s_1) = 1$. We can conclude that $A(a_0) \geq 0$ implies that the aspiration formation rule is optimal.

On the other hand, if $A(a_0) < 0$, then the expression in braces in Equation (A.1) will be negative if $p_t(s_1)$ is sufficiently close to one. Therefore, $p_t(s_1)$ cannot converge to one. Thus the aspiration formation rule is optimal if and only if $A(a_0) \geq 0$ holds.

Next, we observe that $a_0 \to \min_{e\in\mathscr{E}} \pi(s_2, e)$ implies that $A(a_0) > 0$ because in this limit $A(a_0)$ tends to the difference of the expected payoffs of s_1 and s_2, which is by assumption positive. We also note that $A(a_0)$ is piecewise linear and strictly decreasing on the interval $(\min_{e\in\mathscr{E}} \pi(s_2, e), \min_{e\in\mathscr{E}} \pi(s_1, e)]$. This is so because for all a_0 for which there is no $e \in \mathscr{E}$ with $a_0 = \pi_2(s_2, e)$, the function A is differentiable in a_0, its derivative is $-[1 - \mu_2^+(a_0)]$, and locally $\mu_2^+(a_0)$ does not depend on a_0. Because $a_0 > \min_{e\in\mathscr{E}} \pi_2(s_2, e)$, the probability $\mu_2^+(a_0)$ is less than 1, and hence the derivative is negative.

We now distinguish two cases. The first case is that $A(a_0)$ is everywhere nonnegative on $(\min_{e\in\mathscr{E}} \pi(s_2, e), \min_{e\in\mathscr{E}} \pi(s_1, e)]$. In this case we can set $\bar{a} = \min_{e\in\mathscr{E}} \pi(s_1, e)$. With this definition, if $a_0 \leq \bar{a}$, and if a_0 is in the domain of the function A, we have $A(a_0) \geq 0$, and as we saw above, this implies that the aspiration formation rule is optimal. If a_0 is outside the domain of the function A, the arguments given at the beginning of this part of the proof show that the aspiration formation rule is optimal (respectively not optimal) as required.

The second case is that $A(a_0)$ does become negative on $(\min_{e\in\mathscr{E}} \pi(s_2, e), \min_{e\in\mathscr{E}} \times \pi(s_1, e)]$. Because $A(a_0)$ is strictly decreasing, there is then a unique value of a_0 for which $A(a_0) = 0$. We set $\bar{a}$ equal to this value. If $a_0 \leq \bar{a}$, and if a_0 is in the domain of the function A, then $A(a_0) \geq 0$, and the aspiration formation rule is optimal. If $a_0 > \bar{a}$, and if a_0 is in the domain of the function A, then $A(a_0) < 0$, and the aspiration formation rule is not optimal. Finally, the case that a_0 is outside the domain of the function A can again be dealt with as in the first paragraph of this part of the proof. □

PROOF OF PROPOSITION 2. (i) Throughout this part of the proof, we denote by π_1 the payoff that strategy s_1 yields with certainty. We also write Π_2 for the expected payoff of strategy 2. We shall consider a solution to differential Equations (3) and (4) that starts with interior initial probabilities. We shall show that for any such solution it is true that $\lim_{t\to\infty} p_t(s_1) = 1$.

The proof will consider separately two cases. The first case is that $a_t \leq \pi_1$ for some $t \geq 0$. If this is the case, there also needs to be some $t \geq 0$ with $a_t < \pi_1$. This is so because the expected payoff is always less than π_1 [since $p_t(s_1) < 1$ for all $t \geq 0$], and hence $a_t = \pi_1$ implies that $da_t/dt < 0$. Therefore, if the aspiration level reaches π_1, it also has to fall below π_1. As we focus on the limit for $t \to \infty$, there is now no

loss of generality in defining the first case by the condition $a_t < \pi_1$ for all $t \geq 0$. The second case will be the case that $a_t > \pi_1$ for all $t \geq 0$.

To deal with the first case, we shall show first that in this case we have $dp_t(s_1)/dt > 0$ for all $t \geq 0$. This will be sufficient for the proof because it then follows that $p_t(s_1)$ is monotonically increasing and convergent. Then also a_t must be convergent. The limit must be a rest point (p^*, a^*). It is straightforward to see that the only rest point with $p^*(s_1) > 0$ is given by $p^*(s_1) = 1$ and $a^* = \pi_1$. Hence this must be the limit of (p_t, a_t) for $t \to \infty$. Thus we can conclude that the aspiration formation rule is optimal.

So assume case 1 and let some $t \geq 0$ be given. We partition the set $\mathscr{E}$ into two subsets: $\mathscr{E}_2^- \equiv \{e \in \mathscr{E} : \pi(s_2, e) < a_t\}$ and $\mathscr{E}_2^+ \equiv \{e \in \mathscr{E} : \pi(s_2, e) \geq a_t\}$. For the moment, we assume that none of these sets is empty. Later, we shall explain how to modify the proof if this is not true.

Consider the expected change in $p_n(s_1)$ conditional on $p_n = p_t$ and $a_n = a_t$ and conditional on the event that the state of the world is contained in $\mathscr{E}_2^-$. If any state of the world in $\mathscr{E}_2^-$ occurs, the probability $p_n(s_1)$ will be increased, independent of whether s_1 or s_2 has been played. Hence the expected change in $p_n(s_1)$ conditional on the above-mentioned events is positive.

Consider next the expected change in $p_t(s_1)$ conditional on $p_n(s_1) = p_t(s_1)$ and $a_n = a_t$ and conditional on the event that the state of the world is contained in $\mathscr{E}_2^+$. Since the aspiration level is not greater than any of the relevant payoffs, the argument of Subsection 3.3 implies that the expected change conditional on these events is given by the replicator equation. Moreover, since s_1 is dominant, its expected payoff is at least as large as that of s_2, even if we condition on $\mathscr{E}_2^+$. Hence the expected change is nonnegative. The total expected change in $p_n(s_1)$, conditional on $p_n = p_t$ and $a_n = a_t$, is a convex combination of the expected change conditional on these events and conditional on $\mathscr{E}_2^+$ and the expected change conditional on the preceding events and on $\mathscr{E}_2^-$. We can conclude that it is positive.

Obviously, the argument continues to hold if $\mathscr{E}_2^+$ is empty. Moreover, if $\mathscr{E}_2^-$ is empty, then the fact that s_1 has higher expected payoff than s_2 implies that the expected change conditional on $\mathscr{E}_2^+$ must be positive. This completes the proof for the first case.

We now consider the case that $a_t > \pi_1$ for all $t \geq 0$. For this case, we shall show that for all $\bar{p} \in (0, 1)$ there exist a $\tau \geq 0$ and a $\delta \in (0, \pi_1 - \Pi_2)$ such that $t \geq \tau$ and $p_t(s_1) \leq \bar{p}$ imply that $dp_t(s_1)/dt \geq \delta$. This obviously implies that, after time τ, p_t can remain below $\bar{p}$ for only a finite amount of time. This is true for $\bar{p}$ arbitrarily close to one, and hence we can conclude that $p_t(s_1)$ converges for $t \to \infty$ to one. It follows that a_t converges to π.

Consider a given and fixed $\bar{p}$. We shall now construct the corresponding τ and δ. We begin with the observation that $a_t > \pi_1$ implies that a_t is strictly decreasing. In fact, it is straightforward to see that for every $\varepsilon > 0$ there is some $\tau(\varepsilon) > 0$ such that $t \geq \tau(\varepsilon)$ implies that $a_t \leq \pi_1 + \varepsilon$. We shall argue that we can set the τ that we need to construct equal to $\tau(\varepsilon)$ for some $\varepsilon > 0$.

Suppose that $t \geq \tau(\varepsilon)$ for some $\varepsilon > 0$. A straightforward calculation, which uses the fact that $a_t > \pi_1$ implies that *all* payoffs are smaller than the aspiration level,

shows that Equation (3) becomes

$$\frac{dp_t(s_1)}{dt} = [1 - p_t(s_1)]^2(a_t - \Pi_2) - p_t(s_1)^2(\pi_1 - a_t)$$

$$\geq [1 - p_t(s_1)]^2(\pi_1 - \Pi_2) - p_t(s_1)^2\varepsilon \tag{A.3}$$

If $p_t(s_1) \leq \bar{p}$, this is greater than or equal to

$$B(\varepsilon) \equiv (1 - \bar{p})^2(\pi_1 - \Pi_2) - \bar{p}^2\varepsilon \tag{A.4}$$

Since $\pi_1 > \Pi_2$, this expression is positive if ε is sufficiently close to zero. In fact, we can find for any $\delta \in (0, \pi_1 - \Pi_2)$ an $\varepsilon > 0$ such that $B(\varepsilon) = \delta$. If we then set $\tau = \tau(\varepsilon)$, then the claim that we made at the beginning of the discussion of this case is true.

(ii) Suppose that s_1 is not safe. If the aspiration formation rule were optimal, there would have to be a rest point (p^*, a^*) in which $p^*(s_1) = 1$. In such a rest point, a^* would have to be equal to the expected payoff of s_1. Because s_1 is not safe, the fact that a^* is equal to the expected payoff of strategy s_1 implies that $\pi(s_1, e) < a^*$ for some $e \in \mathscr{E}$. We can then use the argument used in Step 1 of the proof of Proposition 1 to show that this cannot be a rest point. We omit the details.

Next, suppose that s_1 is safe but not dominant. Denote by π_1 the payoff that is received with certainty if s_1 is played. Suppose that the aspiration rule were optimal. Then we would have for $t \to \infty$ that $(p_t(s_1), a_t) \to (1, \pi_1)$. In the following we shall show, however, that there are $\bar{p} \in (0, 1)$ and $\bar{a} \in (0, 1)$ with $\bar{a} < \pi_1$ such that $dp_t(s_1)/dt < 0$ whenever $(p_t(s_1), a_t) \in (\bar{p}, 1) \times (\bar{a}, \pi_1)$. This will imply that the aspiration formation rule cannot optimize. Suppose that the learning process started with a probability $p_0(s_1) \leq \bar{p}$ and some aspiration level $a_0 < \pi_1$. Then $a_t < \pi_1$ for all $t \geq 0$. Hence, to converge to $(1, \pi_1)$, the process would have to enter the rectangle $(\bar{p}, 1) \times (\bar{a}, \pi_1)$. Since $dp_t(s_1)/dt < 0$ in this rectangle, this is impossible.

To construct the threshold values $\bar{a}$ and $\bar{p}$, we begin by partitioning the set $\mathscr{E}$ into two subsets: $\mathscr{E}_2^- \equiv \{e \in \mathscr{E} : \pi(s_2, e) < \pi_1\}$ and $\mathscr{E}_2^+ \equiv \{e \in \mathscr{E} : \pi(s_2, e) \geq \pi_1\}$. Notice that the fact that s_1 yields higher expected payoff but is not dominant implies that both these sets are nonempty. Denote by μ_2^- and μ_2^+ their respective probabilities. Finally, denote by Π_2^- and Π_2^+ the expected payoff of s_2 conditional on $\mathscr{E}_2^-$ and $\mathscr{E}_2^+$, respectively.

Suppose that the aspiration level a_t satisfies $\max_{e \in \mathscr{E}_2^-} \pi(s_2, e) < a_t < \pi_1$. A simple calculation then shows that differential Equation (3) becomes

$$\frac{dp_t(s_1)}{dt} = [1 - p_t(s_1)]\{p_t(s_1)[(\pi_1 - a_t) - \mu_2^+(\Pi_2^+ - a_t)]$$

$$+[1 - p_t(s_1)]\mu_2^-(a_t - \Pi_2^-)\} \tag{A.5}$$

Now suppose that a_t is so close to π_1 that the first expression in the braces is negative. Then the right-hand side of Equation (A.5) will be negative if and only if

$$p_t(s_1) > \frac{\mu_2^-(a_t - \Pi_2^-)}{\mu_2^+(\Pi_2^+ - a_t) - (\pi_1 - a_t) + \mu_2^-(a_t - \Pi_2^-)} \tag{A.6}$$

We now choose $\bar{a} \in (0, \pi_1)$ such that all $a_t \in (\bar{a}, \pi_1)$ satisfy the assumptions with respect to a_t that we have made so far in the proof and such that the right-hand side of Equation (A.6) has a supremum strictly smaller than 1 as a_t varies in $(\bar{a}, \pi_1)$. Inspection of Equation (A.6) shows that we can find such an $\bar{a}$. We choose $\bar{p}$ to be larger than the supremum of the right-hand side of Equation (A.6) but smaller than 1. Then $\bar{a}$ and $\bar{p}$ have the required properties. □

PROOF OF PROPOSITION 3. Using the same argument as in the first paragraph of part (ii) of the proof of Proposition 2, one can show that in any rest point $p^*(s_1)$ must be interior; i.e., $p^*(s_1) \in (0, 1)$. This implies that in any rest point $a^* \in (x, y)$. Therefore, the conditions that define a rest point can be written as

$$\frac{p^*(s_1)}{1-p^*(s_1)} = \frac{\mu}{1-\mu} \cdot \frac{p^*(s_1)(x-a^*) + [1-p^*(s_1)](a^*-y)}{p^*(s_1)(a^*-y) + [1-p^*(s_1)](x-a^*)} \tag{A.7}$$

and

$$\begin{aligned} a^* = {} & p^*(s_1)\mu x + [1-p^*(s_1)]\mu y \\ & + p^*(s_1)(1-\mu)y + [1-p^*(s_1)](1-\mu)x \end{aligned} \tag{A.8}$$

We next replace a^* in the first equation by the right-hand side of the second equation. After simplification, we obtain

$$\frac{p^*(s_1)}{1-p^*(s_1)} = \frac{\mu}{1-\mu} \cdot \frac{1-\mu+2p^*(s_1)[1-p^*(s_1)](2\mu-1)}{\mu-2p^*(s_1)[1-p^*(s_1)](2\mu-1)} \tag{A.9}$$

Any solution $p^*(s_1)$ of this equation corresponds to a rest point if one sets the aspiration level equal to the expected payoff that results if s_1 is played with probability $p^*(s_1)$. Thus, by finding the solutions of the preceding equation, we find all rest points of our example.[16]

Denote the left-hand side of the preceding equation by $A[p(s_1)]$ and the right-hand side by $B[p(s_1)]$.[17] Note that $A(0) = 0$, A is continuous and monotonically increasing in $p(s_1)$, and for $p(s_1) \to 1$, we have $A[p(s_1)] \to \infty$. As regards B, note first that $B(0) = B(1) = 1$. Next, it is straightforward from the formula for B that B is continuous, symmetric around 0.5 and that B is monotonically increasing for all $p(s_1) < 0.5$ and decreasing for all $p(s_1) > 0.5$. To see this, observe that B depends on $p(s_1)$ only through the term $2p(s_1)[1-p(s_1)](2\mu-1)$ that appears both in the numerator and in the denominator. Since this expression has the properties ascribed to B, and since this expression is added in the numerator and subtracted in the denominator, the assertion follows.

Finally, we observe that $A(0.5) = 1 < B(0.5) = \mu/(1-\mu)$ and that

$$A(\mu) = \frac{\mu}{1-\mu} > B(\mu) = \frac{\mu}{1-\mu} \cdot \frac{1-\mu+2\mu(1-\mu)(2\mu-1)}{\mu-2\mu(1-\mu)(2\mu-1)} \tag{A.10}$$

These observations together imply the proposition. □

[16] It is interesting that the equation no longer contains x and y.

[17] We denote the argument of these functions by $p(s_1)$ rather than $p^*(s_1)$ because we want these functions to be defined for all $p(s_1) \in (0, 1)$ and not just for the equilibrium values $p^*(s_1)$.

REFERENCES

ARTHUR, B., "On Designing Economic Agents that Behave Like Human Agents," *Journal of Evolutionary Economics* 3 (1993), 1–22.

BENVENISTE, A., M. MÉTIVIER, AND P. PRIOURET, *Adaptive Algorithms and Stochastic Approximations* (Berlin: Springer-Verlag, 1990).

BEREBY-MEYER, Y., AND I. EREV, "On Learning to Become a Successful Loser: A Comparison of Alternative Abstractions of Learning Processes in the Loss Domain," *Journal of Mathematical Psychology* 42 (1998), 266–86.

BINMORE, K., L. SAMUELSON, AND R. VAUGHN, "Musical Chairs: Modeling Noisy Evolution," *Games and Economic Behavior* 11 (1995), 1–35.

BÖRGERS, T., AND R. SARIN, "Learning Through Reinforcement and Replicator Dynamics," *Journal of Economic Theory* 77 (1997), 1–14.

BUSH, R.R., AND F. MOSTELLER, "A Mathematical Model for Simple Learning," *Psychological Review* 58 (1951), 313–23.

—— AND ——, *Stochastic Models for Learning* (New York: Wiley, 1955).

CAMERER, C., AND T. HO, "Experience-Weighted Attraction Learning in Games," *Econometrica* 67 (1999), 827–874.

CHEN, Y., AND F. TANG, "Learning and Incentive Compatible Mechanisms for Public Goods Provision: An Experimental Study," *Journal of Political Economy*, 106 (1998), 633–62.

CROSS, J.G., "A Stochastic Learning Model of Economic Behavior," *Quarterly Journal of Economics* 87 (1973), 239–66.

——, *A Theory of Adaptive Economic Behavior* (Cambridge, England: Cambridge University Press, 1983).

DIXON, H., "Keeping up with the Joneses: Competition and Evolution of Collusion in an Oligopolistic Economy," CEPR discussion paper 1810, 1998.

EREV, I., AND A. ROTH, "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," *American Economic Review* 88 (1998), 848–81.

ESTES, W.K., "Toward a Statistical Theory of Learning," *Psychological Review* 57 (1950), 94–107.

GILBOA, I., AND D. SCHMEIDLER, "Case-Based Optimisation," *Games and Economic Behavior* 15 (1996), 1–26.

HERRNSTEIN, R.J., *The Matching Law* (Cambridge, MA: Harvard University Press, 1997).

—— AND D. PRELEC, "Melioration: A Theory of Distributed Choice," *Journal of Economic Perspectives* 5 (1991), 137–56.

KARANDIKAR, R., D. MOOKHERJEE, D. RAY, AND F. VEGA-REDONDO, "Evolving Aspirations and Cooperation," *Journal of Economic Theory* 80 (1998), 292–331.

KIM, Y., "Satisficing, Cooperation and Coordination," mimeo, Queen Mary and Westfield College, University of London, 1995.

LIEBERMAN, D., *Learning: Behavior and Cognition* (Pacific Grove, CA, Brooks/Cole Publishing Company, 1993).

MOOKHERJEE, D., AND B. SOPHER, "Learning in an Experimental Matching Pennies Game," *Games and Economic Behavior* 7 (1994), 62–91.

—— AND ——, "Learning and Decision Costs in Experimental Constant Sum Games," *Games and Economic Behavior* 19 (1997), 97–132.

NORMAN, M.F., *Markov Processes and Learning Models* (New York: Academic Press, 1972).

PALOMINO, F., AND F. VEGA-REDONDO, "Convergence of Aspirations and (Partial) Cooperation in the Prisoner's Dilemma," mimeo, University of Alicante, 1998.

PAZGAL, A., "Satisficing Leads to Cooperation in Mutual Interest Games," *International Journal of Game Theory* 26 (1997), 439–53.

ROTH, A., AND I. EREV, "Learning in Extensive Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Run," *Games and Economic Behavior* 6 (1995), 164–212.

SARIN, R., "Learning Through Reinforcement: The Cross Model," mimeo, Texas A&M University, 1995.

SCHLAG, K., "A Note on Efficient Linear Learning Rules," mimeo, University of Bonn, 1994.

SCHMALENSEE, R., "Altermative Models of Bandit Selection," *Journal of Economic Theory* 10 (1975), 333–42.

SIEGEL, S., "Decision Making and Learning under Varying Conditions of Reinforcement," *Annals of the New York Academy of Sciences* 89 (1960–1961), 766–83.

SUPPES, P., AND R. ATKINSON, *Markov Learning Models for Multiperson Interaction*, (Stanford, CA: Stanford University Press, 1960).

WALKER, J., *The Psychology of Learning: Principles and Process* (Upper Saddle River, NJ: Prentice-Hall, 1995).

WINTER, S., "Binary Choice and the Supply of Memory," *Journal of Economic Behavior and Organization* 3 (1982), 277–321.