Econometrica, Vol. 72, No. 2 (March, 2004), 383–405

EXPEDIENT AND MONOTONE LEARNING RULES

BY TILMAN BÖRGERS, ANTONIO J. MORALES, AND RAJIV SARIN[1]

This paper considers learning rules for environments in which little prior and feedback information is available to the decision maker. Two properties of such learning rules are studied: absolute expediency and monotonicity. Both require that some aspect of the decision maker's performance improves from the current period to the next. The paper provides some necessary, and some sufficient conditions for these properties. It turns out that there is a large variety of learning rules that have the properties. However, all learning rules that have these properties are related to the replicator dynamics of evolutionary game theory. For the case in which there are only two actions, it is shown that one of the absolutely expedient learning rules dominates all others.

KEYWORDS: Learning, monotonicity, absolute expediency, replicator dynamics, bounded rationality.

1. INTRODUCTION

WE ARRIVE AT MOST ECONOMIC DECISIONS of our lives through a learning process in which we adjust our behavior in response to experience. For example, we learn in this way which consumption goods we like, how to invest our money, and how to behave towards colleagues at work. For economic theory it is therefore interesting to explore mathematical models of learning. Fudenberg and Levine (1998) have surveyed the large literature on this subject.

A problem is that a large variety of learning models exists, and that it is often not clear on which grounds to choose one model rather than another. If much prior and feedback information is available, then a Bayesian model with a limited state space might seem plausible. But if the prior and feedback information are very incomplete, then the set of conceivable states of the world is so large, and the basis on which the decision maker can update beliefs so small, that Bayesian models seem less plausible. Once one turns away from Bayesian learning models, it is difficult to see which learning models one should study.

This paper suggests that a useful way of proceeding is to investigate general properties of learning rules. We consider a large class of learning models that encompasses almost all learning models in which the only feedback information used by the decision maker is his own payoff. We investigate which of these models have a property labelled "absolute expediency." We also explore

[1]We are grateful to an editor, and to Jeff Ely, Drew Fudenberg, and several anonymous referees for very helpful comments. A first version of some parts of this paper was a chapter in Morales' Ph.D. thesis at University College London, and was circulated in 1998 in a paper entitled "Simple Behaviour Rules Which Lead to Expected Payoff Maximising Choices." The authors thank the following for financial support: Economic and Social Research Council (UK) through the Centre for Economic Learning and Social Evolution (ELSE) (Tilman Börgers); centrA and the Bank of Spain (Antonio J. Morales); the Program to Enhance Scholarly and Creative Activities and the Bush Program in Policy Research at Texas A&M University (Rajiv Sarin).

a closely related property, "monotonicity." Both properties refer to the decision maker's observable behavior only, not his beliefs or thoughts. Learning rules have these properties if the performance of a decision maker using the learning rules improves from the current period to the next, provided that the environment stays the same. When considering absolute expediency, the performance measure is expected payoffs. When considering monotonicity, the performance measure is the expected probability with which the strategy that maximizes expected payoffs is played. The properties require performance improvement in every environment in a very large class of environments.

Why are we interested in these properties? In everyday life, we often speak of the "learning curve" which describes the change in behavior when repeatedly facing a given task. Implicit is the idea that the learning decision maker gradually, but monotonically moves towards better choices. In environments that are subject to random shocks, one cannot expect that people learn monotonically with probability 1. A weaker property is that they learn monotonically in expected terms. We study learning algorithms with this feature, where we only focus on whether monotonic learning happens in expected terms from the first period to the next. This seems a simple and natural criterion for classifying learning schemes.

In some situations the properties that we study may be desirable. Suppose the decision maker has no information about the environment that prevails today, but he thinks that the environment is likely to stay unchanged in the short-run, though not in the long-run. It then seems plausible that the decision maker focuses on the short-run. Our assumption that the decision maker only thinks about the next period is an extreme form of myopia which we assume here for simplicity. The decision maker's focus on *improvement* is psychologically plausible. Introspection suggests that the present often serves as a status quo, and that a person's focus is on not letting it deteriorate. A decision maker who is genuinely uncertain about his environment might then seek improvement in his performance for *all* possible decision environments rather than trading off improvement in one environment against a reduction in performance in another.

We assume that the decision maker measures his performance either in terms of expected payoffs, or in terms of the expected probability of playing the strategy that maximizes expected payoffs. The focus on expected payoffs in our otherwise non-Bayesian model will appear surprising. To see why the non-Bayesian decision maker might be interested in expected payoffs, decompose the uncertainty facing the decision maker into two elements: (i) "What is the environment?" and (ii) "Which payoffs will the decision maker receive in *any* given environment?" In our paper, the decision maker is Bayesian with respect to the second, but not with respect to the first question, because the second question involves less complexity, and therefore a Bayesian treatment is less problematic, at least as a starting point for a study of boundedly rational learning schemes.

Our main results provide some necessary and some sufficient conditions for absolute expediency and monotonicity. A necessary condition for both absolute expediency and monotonicity is that the decision maker uses Cross' (1973) learning rule, or a modified version of this learning rule.[2] Cross' rule requires that the decision maker raise the probability of the strategy that he or she chose in proportion to the payoff received, and that all other choice probabilities be reduced proportionally. The modifications of this rule that are compatible with absolute expediency or monotonicity are learning rules in which payoffs are subjected to certain affine transformations before Cross' rule is applied. The coefficients of these transformations are allowed to depend on the decision maker's current mixed strategy, the strategy that he played, and the strategy whose probability he is updating.[3]

We know from earlier work (Börgers and Sarin (1997)) that there is a close connection between the expected movement of Cross' learning model and the replicator dynamics of evolutionary game theory. The necessary condition for absolute expediency and monotonicity that we find in this paper implies therefore an analogy between the expected movement of absolutely expedient or monotone learning rules and the replicator dynamics. In the case in which there are only two actions, the analogy is particularly tight: the expected movement of action probabilities equals the replicator dynamics, rescaled with some constant. The replicator dynamics and related evolutionary dynamics are often used in economic or social contexts. Our results strengthen the case of the use of replicator dynamics in contexts where learning is important.

Moving beyond necessary conditions, our next finding is that monotonicity is a more restrictive property than absolute expediency.[4] We show that all monotone learning rules are absolutely expedient, and we give an example of an absolutely expedient rule that is not monotone.

We have unfortunately not found a complete characterization of absolutely expedient learning rules, but we do have a complete characterization of monotone learning rules. We find that the most important property of monotone learning rules is that an increase in the payoff received with one particular action can never make any of the other actions more likely. By contrast, we show by means of examples that absolutely expedient learning rules can have the feature that the higher the payoff experienced with some action, the higher is the probability of playing one of the other actions in the next period. We interpret this as an implicit similarity relation between the concerned

[2]Cross' learning model is in the tradition of the mathematical learning theory developed by the psychologists Bush and Mosteller (1951).

[3]The effect of these transformations can be that the probability of the action played is lowered if the payoff received is low, which is, somewhat implausibly, ruled out by Cross' rule. An example of a rule with this feature is mentioned in Proposition 4.

[4]Provided that there are at least three actions. If there are only two actions, then the two properties are obviously equivalent.

actions. Absolutely expedient learning rules thus can embody an implicit similarity relation, but monotone learning rules cannot.

As there are relatively large sets of monotone, or absolutely expedient learning rules, one might ask: "Which of these rules is the best?" We shall call a learning rule "best monotone" if, in all environments, it leads to at least as large an expected increase in the probability of the best action as any other monotone rule. Similarly, we shall call a learning rule "best absolutely expedient" if, in all environments, it leads to a larger increase in expected payoffs than any other absolutely expedient rule. We show that for the case of two actions there is a unique rule that is both best monotone and best absolutely expedient, but that in the case of more than two actions there is no best monotone, and no best absolutely expedient rule.

This paper is organized as follows. Section 2 introduces the framework. Section 3 defines the main concepts, absolute expediency and monotonicity. Section 4 characterizes a property called "unbiasedness," which is necessary for both absolute expediency and monotonicity. Section 5 investigates which additional features unbiased learning rules have to have if they are to be absolutely expedient or monotone. Section 6 gives examples. Section 7 investigates whether some absolutely expedient, or monotone learning rule can be singled out as "best." Finally, Section 8 discusses related literature.

2. MODEL

A decision maker chooses from a finite set $S = \{s_1, s_2, \ldots, s_n\}$ of pure strategies that has at least two elements. Every strategy s_i gives payoffs according to a payoff distribution μ_i. We assume that there is *some* upper and *some* lower bound for payoffs. For our paper it is then without loss of generality to assume that the upper boundary is 1, and that the lower boundary is 0. In the following definition an assignment of payoff distributions to strategies is called an environment.

DEFINITION 1: An *environment* E is a collection $(\mu_i)_{i=1,2,\ldots,n}$ of probability measures, each of which has support in the interval $[0, 1]$.

We shall be concerned with the decision maker's behavior at two dates, "today" and "tomorrow." The environment is the same at these two dates. Payoffs today are stochastically independent of payoffs tomorrow.

The decision maker knows the strategy set S, the bounds for payoffs, and that his strategy set tomorrow is the same set as today. The decision maker does not know the environment E. He chooses a strategy from S today, and then observes the payoff realization. Tomorrow, he chooses a strategy from S again.

The decision maker's behavior today is described by a probability distribution $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ over S. Here, we denote by σ_i the probability assigned

to pure strategy s_i. The distribution σ describes how likely the decision maker is to choose each of his strategies today.

The decision maker's behavior today will be exogenous and fixed. Our analysis could form a building block for an analysis that includes a study of the optimal initial point for the learning process. Alternatively, our analysis could also be integrated into a study of learning algorithms that have the properties with which we are concerned at every interior initial point.[5]

The decision maker's behavior tomorrow is governed by a learning rule.

DEFINITION 2: A *learning rule* is a function $L : S \times [0, 1] \to \Delta(S)$.

A learning rule determines as a function of the pure strategy s_i, which the decision maker chooses today (and which is distributed according to σ), and of the payoff that he receives today (which is distributed according to μ_i), how likely each strategy is tomorrow. Denote by $L(s_i, x)(s_j)$ the probability that the decision maker's mixed strategy tomorrow assigns to the pure strategy s_j if the decision maker plays today the pure strategy s_i and receives the payoff x.

One should think of the learning rule in Definition 2 as a "reduced form" of the decision maker's true learning rule. The true learning rule may, for example, specify how the decision maker updates beliefs about the payoff distributions in response to his observations, and how these beliefs are translated into behavior. If one combines the two steps of belief updating and behavior adjustment one arrives at a learning rule in the sense of Definition 2. Our approach is therefore more general than an approach that focuses on learning rules in which the state space of the learning rule is the strategy simplex $\Delta(S)$.

Throughout this paper we will make the following assumption:

ASSUMPTION 1: *For every $i = 1, 2, \ldots, n$ the probability σ_i is strictly positive.*

If this assumption is violated, no learning rule can have the properties of absolute expediency and monotonicity that we study below. To see this suppose the environment were such that only strategies in some set $C \neq S$ were initially played with positive probability. Consider environments in which $i \in C$ implies that μ_i assigns probability 1 to payoff x, and $i \notin C$ implies that μ_i assigns probability 1 to payoff $y \neq x$. In such environments, for any learning rule L, the expected change in the probability of strategies is independent of the value of y. But absolute expediency and monotonicity require that the total expected

[5]We need to refer here to interior initial points because of Assumption 1. Some learning rules that we study in this paper have the property that extreme payoffs (0 or 1) in some period lead the decision maker to adopt in the next period a mixed strategy that is not interior. Such learning rules can then not always be applied repeatedly. However, such learning rules can be arbitrarily closely approximated by learning rules that never take the decision maker outside of the interior of the mixed strategy simplex. One simply has to multiply all changes in probabilities prescribed by the learning rule by the factor $1 - \varepsilon$ where $\varepsilon \in (0, 1)$ can be arbitrarily close to zero.

change in the probability of all strategies in C is negative if $y > x$, and zero if $x > y$.

3. ABSOLUTE EXPEDIENCY AND MONOTONICITY

Our focus is on learning rules that guarantee for given initial state an improvement in the decision maker's performance in every possible environment. To formalize this property, we fix some environment E. For any strategy $s_i \in S$ we denote the expected payoff of strategy s_i by π_i. That is, $\pi_i = \int_0^1 x \, d\mu_i$. The set of expected payoff maximizing strategies is denoted by S^*, that is, $S^* = \{s_i \in S \mid \pi_i \geq \pi_j \text{ for all } j = 1, 2, \ldots, n\}$. To keep our notation simple, we suppress the dependence of π_i and S^*, and of related variables below, on E.

Now fix a learning rule L. For every strategy s_i denote by $f(s_i)$ the expected change in the probability attached to s_i:

$$f(s_i) = \sum_{j=1}^{n} \sigma_j \left[\int_0^1 (L(s_j, x)(s_i) - \sigma_i) \, d\mu_j \right].$$

We extend the definition of f to subsets $\tilde{S}$ of S by setting: $f(\tilde{S}) = \sum_{s_i \in \tilde{S}} f(s_i)$. Finally, we define g to be the expected change in expected payoffs:

$$g = \sum_{i=1}^{n} f(s_i) \pi_i.$$

Of course, f and g depend on the learning rule L, but, to keep things simple, we suppress that dependence in our notation. Note that we also do not indicate the dependence of f and g on σ. This is because throughout the paper σ will be exogenous and fixed, as explained in Section 2.

We can now define the property of learning rules which is the focus of this paper.

DEFINITION 3: A learning rule L is *absolutely expedient* if for all environments E with $S^* \neq S$ we have $g > 0$.

In words, a learning rule is absolutely expedient if in all nontrivial environments expected payoffs are on average strictly higher tomorrow than today. An environment is "nontrivial" if $S^* \neq S$. If $S^* = S$, all strategies are optimal and nothing needs to be learned. If $S^* \neq S$, then there is scope for improvement in the decision maker's performance because, by Assumption 1, the decision maker assigns some positive probability to nonoptimal strategies.

A second formalization of the notion of "improvement" in the decision maker's performance requires that the probability assigned to the best actions increases in all nontrivial environments.

DEFINITION 4: A learning rule L is *monotone* if for all environments E with $S^* \neq S$ we have $f(S^*) > 0$.

The relation between monotonicity and absolute expediency will be studied below. However, the following observation is obvious.

REMARK 1: If $n = 2$, then a learning rule L is absolutely expedient if and only if it is monotone.

At this point we briefly remark on a subtle technical point.

REMARK 2: While it is without loss of generality to take the upper and lower boundaries on payoffs to be zero and one, in the light of Definitions 3 and 4 it is not quite without loss of generality to let the set of possible payoffs be the *closed* interval $[0, 1]$, as we did in Definition 1, rather than the *open* interval $(0, 1)$ (or a half-open interval). Our assumption that the set of payoffs is $[0, 1]$, in combination with Definitions 3 and 4, implies that when checking absolute expediency or monotonicity one needs to consider (among others) environments in which the upper or the lower boundary for payoffs are attained. If we had considered the (half-)open interval, then these environments would have been ruled out. Our proofs can easily be modified to cover the case in which the interval of possible payoffs is taken to be (half-)open.

We end this section with an example of a learning rule due to Cross (1973). In the next section we shall show that all absolutely expedient or monotone learning rules have a structure that is similar to the structure of Cross' learning rule.

EXAMPLE 1: For all $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$, and for all $x \in [0, 1]$,

$$L(s_i, x)(s_i) = \sigma_i + (1 - \sigma_i)x,$$

$$L(s_j, x)(s_i) = \sigma_i - \sigma_i x.$$

In words, if the decision maker plays strategy s_i and obtains payoff x, then he increases the probability of s_i, and the size of the increase is proportional to x. If $x = 1$, then the decision maker sets the probability of s_i equal to one. If $x = 0$, he leaves the probability of s_i unchanged. The probability of all other strategies is reduced so as to keep the sum of all probabilities equal to one, and to leave the ratios between the other probabilities unchanged. Notice that this learning rule has the somewhat counterintuitive feature that the decision maker *always* increases the probability of the strategy that he actually played, even if the payoff was very low. Not all absolutely expedient or monotone learning rules have this feature, as an example in Proposition 4 below shows.

We now show that Cross' learning rule is absolutely expedient and monotone. The expected movement of the probability of any particular pure strategy s_i under Cross' rule is

$$f(s_i) = \sigma_i \left[\pi_i - \sum_{j=1}^{n} (\sigma_j \pi_j) \right] \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

This equation shows that the expected change in the probability of any pure strategy s_i is proportional to the difference between that strategy's expected payoff, and the expected value of the expected payoff of the pure strategy played today. The condition $S^* \neq S$ and Assumption 1 imply that for strategies in S^* the difference between their expected payoff and the expected value of the expected payoff of the pure strategy played today is strictly positive. Thus the above equation shows that Cross' rule is monotone.

Note that the right-hand side of the equation for $f(s_i)$ is the same as the right-hand of the replicator equation in evolutionary game theory, which describes how proportions of different strategies in a population move if the population is subject to evolutionary selection. The connection between Cross' learning model and the replicator dynamics was explored further in Börgers and Sarin (1997).

The expected movement of payoffs under Cross' learning rule is given by

$$g = \sum_{i=1}^{n} \sigma_i \left[\pi_i - \sum_{j=1}^{n} (\sigma_j \pi_j) \right]^2.$$

The right-hand side is the variance of the expected payoff of the pure strategy chosen today. How can an expected value have a variance? The decision maker's pure strategy today is a random variable. Thus, also the expected payoff associated with that pure strategy is a random variable. The right-hand side is the variance of that random variable. Observe that $S^* \neq S$ and Assumption 1 imply that this variance is strictly positive. Thus we have shown that Cross' rule is absolutely expedient.

4. UNBIASEDNESS

In a first step we study a property that we call unbiasedness.

DEFINITION 5: A learning rule L is *unbiased* if for all environments E with $S^* = S$ we have $f(s_i) = 0$ for every $i = 1, 2, \ldots, n$.

In words this definition says that a learning rule is unbiased if the expected movement in all strategies' probabilities is zero provided that all strategies have

the same expected payoff. If in such an environment some strategies' probabilities increased in expected terms, and some other strategies' probabilities decreased, then the learning rule would implicitly "favor" the former strategies. This is why we refer to the property as "unbiasedness."

The next lemma shows that unbiasedness is necessary for absolute expediency and monotonicity.[6]

LEMMA 1: *Every absolutely expedient and every monotone learning rule is unbiased*.

PROOF: Let L be a biased learning rule. Consider an environment E such that $S = S^*$, and, for some strategy $s_i \in S$, we have: $f(s_i) < 0$. Now we shall construct a new environment by making a small change in the payoff distribution of s_i, leaving all other strategies' payoff distributions unchanged. We first consider the case that there is some x in the support of μ_i such that $x < 1$. We now reduce the probability that μ_i attaches to x by some $\varepsilon > 0$, and assign the probability ε instead to some payoff $x + \alpha$ where $0 < \alpha < 1 - x$. In the new environment, strategy s_i is the unique best strategy. The expected movement of the probability assigned to s_i is continuous in ε. For sufficiently small ε, therefore, the expected change in the probability of s_i is negative in the modified environment, as it was in the original environment. This contradicts absolute expediency and monotonicity. It remains to deal with the case that the support of s_i is the singleton 1. Because $S^* = S$, all other probability distributions μ_j must also assign probability 1 to the payoff 1. Because the expected movement in the probability of s_i is negative, there must be at least some other strategy s_j such that the expected movement in s_j's probability is positive. Replace the payoff distribution for that strategy by a distribution that assigns some positive probability $\varepsilon > 0$ to some payoff less than 1, instead of 1. If ε is sufficiently small, the expected movement in the probability of s_j will be positive. This contradicts absolute expediency and monotonicity. *Q.E.D.*

Our strategy is to characterize unbiased learning rules, and then to ask which additional conditions absolutely expedient or monotone learning rules have to satisfy. The proof of the following proposition is in the Appendix.

PROPOSITION 1: *A learning rule L is unbiased if and only if there are matrices $(A_{ij})_{i,j=1,2,\ldots,n}$ and $(B_{ij})_{i,j=1,2,\ldots,n}$ such that for every $(s_i, x) \in S \times [0, 1]$,*

$$L(s_i, x)(s_i) = \sigma_i + (1 - \sigma_i)(A_{ii} + B_{ii}x), \tag{1}$$

$$L(s_j, x)(s_i) = \sigma_i - \sigma_i(A_{ji} + B_{ji}x) \quad \text{for all} \quad j \neq i, \tag{2}$$

[6]In previous versions of this paper, we assumed that the learning rule was continuous in payoffs, and we used this assumption to prove Lemma 1. We are grateful to Jeff Ely for comments that induced us to reinvestigate whether we really needed the continuity assumption.

and for every $i = 1, 2, \ldots, n$,

$$A_{ii} = \sum_{j=1}^{n} (\sigma_j A_{ji}), \tag{3}$$

$$B_{ii} = \sum_{j=1}^{n} (\sigma_j B_{ji}). \tag{4}$$

Thus, a learning rule is unbiased if and only if the decision maker, after playing his action and receiving his payoff, first submits the payoff to an affine transformation and then applies Cross' rule. The coefficients of this affine transformation are allowed to depend on the strategy that he has played and on the strategy whose probability he is adjusting. Conditions (3) and (4) restrict the coefficients of the affine transformation. They require that the coefficients of the affine transformation that are applied when s_i was played and s_i is updated are the expected values (over j) of the coefficients that are used when s_j was played and s_i is updated.

The key feature of the learning rules in Proposition 1 is that they are linear in payoffs. Very informally speaking the intuition why linearity is necessary for unbiasedness is that expected payoffs are a linear function of payoffs. The linearity of the expected payoff function must be reflected in the linearity of an unbiased learning rule.

The following remarks follow from Proposition 1 through elementary calculations.

REMARK 3: Let L satisfy the characterization in Proposition 1, and let E be an environment. Then for all $s_i \in S$ the expected change of the probability of s_i is given by

$$f(s_i) = \sigma_i \left[B_{ii}\pi_i - \sum_{j=1}^{n} (\sigma_j B_{ji}\pi_j) \right].$$

The expected movement of expected payoffs is given by

$$g = \sum_{i=1}^{n} (\sigma_i B_{ii} {\pi_i}^2) - \sum_{i=1}^{n} \sum_{j=1}^{n} (\sigma_i \sigma_j B_{ij} \pi_i \pi_j).$$

These two formulas reduce to the analogous formulas for the Cross model in the previous section if all the coefficients B_{ij} equal one. This is evident for the first formula, which is reminiscent of the replicator dynamics. The second formula reduces in the case that all the coefficients equal one to the difference between the expected value of the square of π_i and the square of the expected value of π_i, which is, of course, the variance.

REMARK 4: Suppose $n = 2$. Then the conditions in Proposition 1 can be satisfied only if there are constants A and B such that $A_{ij} = A$ and $B_{ij} = B$ for all $i, j = 1, 2$. This follows from straightforward calculations. Substituting this into the formulas in Remark 3, we find that for $n = 2$ the expected movement of an unbiased learning process is exactly equal to the replicator dynamics, multiplied by the factor B.

5. OWN AND CROSS EFFECTS

Next, we ask which additional conditions, beyond those in Proposition 1, learning rules satisfy if they are absolutely expedient or monotonic. Notice that Remark 3 indicates that it is the coefficients $(B_{ji})_{i,j=1,2,\ldots,n}$ that matter for the expected movement of the probability of expected payoffs and of the probability of playing one of the best strategies. Therefore, our investigation will focus on these coefficients.

We first note that if there are only two actions it is immediate from Remark 4 how we can characterize absolutely expedient or monotone rules.

REMARK 5: Suppose $n = 2$. Then a learning rule L is absolutely expedient (equivalently: monotone) if and only if $B_{ij} > 0$ for $i, j = 1, 2$.

We now turn to the general case of two or more actions.

DEFINITION 6: A learning rule L is *own-positive* if $B_{ii} > 0$ for all $i = 1, 2, \ldots, n$.

This property means that the probability that the decision maker plays tomorrow the strategy that he played today increases in the payoff that the decision maker received today. The following result shows that the learning rules that we study in this paper are own-positive.

PROPOSITION 2: *Every absolutely expedient or monotone learning rule is own-positive*.

PROOF: Let L be absolutely expedient or monotone. Consider an environment in which all actions have the same expected payoff $x < 1$. By Proposition 1, $f(s_i) = 0$ for all $i = 1, 2, \ldots, n$. Now add some $\varepsilon > 0$ to the expected payoff of some strategy s_i. It is easy to calculate from the formulas in the proof of Proposition 1 that in this new environment $f(s_i) = \sigma_i(1 - \sigma_i)B_{ii}\varepsilon$. Clearly $f(s_i)$ has to be positive if L is absolutely expedient or monotone. This requires that $B_{ii} > 0$. This holds for all i. *Q.E.D.*

The above proposition shows that own-positivity is necessary for absolute expediency or monotonicity. However, it turns out that it is not sufficient. We introduce a further, more restrictive property.

DEFINITION 7: A learning rule L is *cross-negative* if:

(i) $B_{ji} \geq 0$ for all $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$; and

(ii) if C is a subset of S such that $C \neq \emptyset$, and $S \setminus C \neq \emptyset$, then there are strategies $s_i \in C$, and $s_j \in S \setminus C$ such that $B_{ji} > 0$.

Condition (i) in this definition means that if the decision maker played a strategy s_j today, then the probability that he plays a different strategy s_i tomorrow is nonincreasing in the payoff that he received today. This rules out that the decision maker regards s_i as "similar" to s_j, and therefore treats a success today with s_j as encouraging news also for s_i.

Cross-negativity allows for the possibility that some cross effects are null, i.e. that the size of the payoff received today has no impact on the probability with which some other strategy is played tomorrow. However, not *all* cross-effects can be null. This is implied by condition (ii). Condition (ii) means that whenever one partitions S into two subsets, then one can find a pair of strategies, one from each subset, such that the cross effect is strictly negative.

A simple inspection of condition (4) in Proposition 1 shows that cross-negativity implies own-positivity but not vice versa (except when the number of actions is 2).

It may seem plausible that cross-negativity is necessary for absolute expediency or monotonicity. Our decision maker is ignorant about his environment, and thus one might think that a learning rule must not have built in similarity relations. It turns out that this intuition is only partially correct.

PROPOSITION 3: (i) *A learning rule is monotone if and only if it is cross-negative*. (ii) *Every cross-negative rule is absolutely expedient*.

The proof of part (i) is simple and transparent, but the proof of part (ii) is more involved. Therefore, part (ii) is proved in the Appendix.

PROOF: We will find it convenient to work with the following expression for the expected change in the probability attached to any action s_i. This expression can be obtained by inserting condition (4) of Proposition 1 into the formula of Remark 3.

$$f(s_i) = \sigma_i \sum_{\substack{j=1 \\ j \neq i}}^{n} \sigma_j B_{ji}(\pi_i - \pi_j) \quad \text{for all} \quad i = 1, 2, \ldots, n. \tag{5}$$

Sufficiency proof for part (i): Consider an environment E with $S^* \neq S$ and any strategy $s_i \in S^*$. If L is cross-negative then all the expressions in the sum on the right-hand side of equation (5) are nonnegative. Moreover, condition (ii) in the definition of cross-negativity ensures that there exist $s_i \in S^*$ and $s_j \in S^* \setminus S$

such that $B_{ji} > 0$, and hence the expected change in the probability with which strategy s_i is played is strictly positive. Thus we can conclude that $f(S^*) > 0$.

Necessity proof for part (i): Suppose that L is monotone. We begin by proving that it has to satisfy condition (i) in the definition of cross-negativity. Our proof is indirect. Suppose there were $j, i \in \{1, 2, \ldots, n\}$ with $j \neq i$ such that $B_{ji} < 0$. Consider an environment E such that s_i yields payoff x with probability 1, s_j yields payoff $x - \delta$ with probability 1, and all other strategies s_k (if any) yield payoff $x - \varepsilon$ with probability 1. Here we assume $\delta, \varepsilon > 0$. Then, equation (5) implies

$$f(s_i) = \sigma_i \sum_{k=1, k\neq i}^{n} (\sigma_k B_{ki}(\pi_i - \pi_k))$$

$$= \sigma_i \left(\sigma_j B_{ji} \delta + \sum_{k=1, k\neq i, j}^{n} (\sigma_k B_{ki} \varepsilon) \right).$$

If $B_{ji} < 0$, then this expression becomes negative when ε is sufficiently close to zero, which contradicts monotonicity.

Next we prove that L has to satisfy condition (ii) in the definition of cross-negativity. The proof is indirect. Suppose there were some subset C of S such that $C \neq \emptyset$ and $S \setminus C \neq \emptyset$ and such that $B_{ji} = 0$ for all $s_i \in C$ and $s_j \in S \setminus C$. Consider an environment E such that all strategies in C yield payoff x with certainty, and all strategies in $S \setminus C$ yield payoff $y < x$ with certainty. Using the same formula as before it is immediate that $f(s) = 0$ for all strategies in C, and hence that the rule is not monotone. *Q.E.D.*

The above proposition leaves the question open whether absolutely expedient rules exist that are not monotone. Such rules must include at least one positive cross-effect. This means that a notion of similarity of two strategies is built into the learning rule. But, in the true environment, these strategies might not be similar at all. Even in such environments the rule must improve expected payoffs. In the next section we shall give an example of such a rule.

6. EXAMPLES

We begin with an example of an absolutely expedient rule that is not monotone. In this rule $n = 3$, and the current mixed strategy is the uniform distribution. Intuitively, the rule treats actions 1 and 2 as similar. In an earlier version of the paper we have shown how this example can be extended to the case $n > 3$, and to the case of arbitrary initial state. The details are available from the authors. We omit the straightforward calculation which shows that this rule is absolutely expedient.

EXAMPLE 2: Suppose $n = 3$ and the current state is: $\sigma_1 = \sigma_2 = \sigma_3 = \frac{1}{3}$. Define:

$$A_{ji} = 0 \quad \text{for all} \quad j, i \in \{1, 2, 3\},$$

$$B_{ii} = \frac{1}{10} \quad \text{for all} \quad i \in \{1, 2, 3\},$$

$$B_{12} = B_{21} = -\frac{1}{10},$$

$$B_{i3} = B_{3i} = \frac{3}{10} \quad \text{for all} \quad i \in \{1, 2, 3\}.$$

At this stage one may wonder whether all own-positive rules are absolutely expedient. This is not the case. Suppose that there are $n = 3$ actions, and that the decision maker applies Cross' rule with the following modification: If s_1 or s_2 have been played and a payoff x has been received, then the decision maker applies Cross' rule to the joint probability of s_1 and s_2, and moreover keeps the relative probabilities of these two strategies unchanged. This rule is own-positive. Now consider an environment in which the expected payoff of strategies s_1 and s_2 taken together equals the expected payoff of s_3: $\sigma_1\pi_1 + \sigma_2\pi_2 = \pi_3$, but in which $\pi_1 > \pi_2$. Then in expected terms no strategy's probability will change, and therefore also the expected payoff will stay the same. However, absolute expediency requires it to increase.

Our final example shows how the results can be used to assess whether a learning rule is monotone or absolutely expedient. The rule that we consider is due to Roth and Erev (1995) and Erev and Roth (1998). Their learning rule has the state space $V = R^n_{>0}$ with generic element $v = (v_1, v_2, \ldots, v_n)$. The vector v describes the decision maker's "inclination" to play any of his n strategies. The decision maker's mixed strategy is proportional to v. After playing strategy s_i and receiving payoff x, the decision maker adds x to the inclination of playing s_i, leaving all other inclinations unchanged. The following formulae describe the implied change in the strategy probabilities.

EXAMPLE 3: The Roth–Erev learning rule is given by

$$L_v(s_i, x)(s_i) = \sigma_i + \frac{1}{\sum_{k=1}^n v_k + x}(1 - \sigma_i)x,$$

$$L_v(s_j, x)(s_i) = \sigma_i - \frac{1}{\sum_{k=1}^n v_k + x}\sigma_i x \quad \text{for all} \quad j \neq i.$$

Note that this learning rule is Cross' Rule, except that the direction of the movement is multiplied by $1/(\sum_{k=1}^n v_k + x)$. The learning rule is not linear in

payoffs because x appears in the denominator. Therefore, according to Proposition 1, it is not unbiased and, according to Lemma 1, it is neither monotone nor absolutely expedient.

7. BEST LEARNING RULES

We have found a large set of monotone learning rules, and a larger set of absolutely expedient learning rules. Is any of these learning rules "best"? A natural definition of "best" in the context of monotonicity is that the expected increase in the probability of playing the best actions is maximized in all environments. A natural definition of "best" in the context of absolute expediency is that the increase in expected payoffs is maximized in all environments.

DEFINITION 8: Given an initial state σ a learning rule L is called *best monotone for* σ if it is monotone, and if for every other monotone learning rule L' and for every environment E we have: $f(S^*) \geq f'(S^*)$, where $f(S^*)$ is the expected change in the probability of the expected payoff maximizing actions if L is used, and $f'(S^*)$ is the expected change in the probability of the expected payoff maximizing actions if L' is used.

A learning rule L is called *best absolutely expedient for* σ if it is absolutely expedient, and if for every other absolutely expedient learning rule L' and for every environment E we have: $g \geq g'$, where g is the expected change in expected payoffs if L is used, and g' is the expected change in expected payoffs if L' is used.

In the case that there are two actions only, a learning rule is obviously best absolutely expedient if and only if it is best monotone, and it is sufficient to focus on best monotone rules. The following proposition characterizes for this case the best monotone rule.

PROPOSITION 4: *Let* $n = 2$, *and consider a fixed initial state* (σ_1, σ_2). *Then there is a unique best monotone learning rule for that initial state. It is given by*

$$A_{ij} = -\frac{\min\{\sigma_1, \sigma_2\}}{\max\{\sigma_1, \sigma_2\}} \quad \textit{for} \quad i, j \in \{1, 2\} \quad \textit{and}$$

$$B_{ij} = \frac{1}{\max\{\sigma_1, \sigma_2\}} \quad \textit{for} \quad i, j \in \{1, 2\}.$$

PROOF: Recall from Remark 4 that in the case $n = 2$ conditions (3) and (4) of Proposition 1 imply that all coefficients in the A matrix in Proposition 1 have to be identical, and all coefficients in the B matrix have to be identical: $A_{ij} = A$ for $i, j \in \{1, 2\}$ and $B_{ij} = B$ for $i, j \in \{1, 2\}$. The expected change in the probability of strategy s_i is

$$f(s_i) = B\sigma_i[\pi_i - (\sigma_1\pi_1 + \sigma_2\pi_2)].$$

Thus, the best monotone learning rule is the one for which B is largest. The admissible values for A and B are those for which, for any payoff value x, the formula for updating the probability of strategies yields a value in $[0, 1]$. A simple calculation shows that among all admissible values of A and B the one indicated in Proposition 4 has the highest value of B. *Q.E.D.*

REMARK 6: Note that the best monotone learning rule incorporates an *endogenous* aspiration level. To see this note that the probability of playing the same action tomorrow as was played today is given by

$$L(s_i, x)(s_i) = \sigma_i + (1 - \sigma_i)\frac{1}{\max\{\sigma_1, \sigma_2\}}(x - \min\{\sigma_1, \sigma_2\}).$$

Thus, the probability $\min\{\sigma_1, \sigma_2\}$ serves as an aspiration level. If the payoff received is below this probability, then the probability of playing the action is reduced. Otherwise, it is increased. The aspiration level is the higher the closer together the probabilities of the two strategies.[7]

REMARK 7: Several rules singled out by Schlag (2002) as having good properties induce the same behavior as the rule that Proposition 4 identifies as the best monotone rule for uniform initial state $(\frac{1}{2}, \frac{1}{2})$. Schlag's work is restricted to the case of two actions and initial state $(\frac{1}{2}, \frac{1}{2})$. Proposition 9 of Schlag (2002) lists properties of rules that are "closest to ideal" (i.e. minimize some measure of regret) among all ex-ante improving rules (see Section 8 for an explanation of this term). One property that is listed is that the strategy that was played in period 1 is repeated in period 2 with a probability that is equal to x, the payoff received. This is exactly the same as the best monotone learning rule that we have identified for uniform initial state. For uniform initial state $(\frac{1}{2}, \frac{1}{2})$, the best monotone rule chooses $A = -1$ and $B = 2$. This implies $L(s_i, x)(s_i) = x$.

We now move to the case of more than two actions. We show that in the simplest possible circumstances, three actions and uniform initial state, there is no best monotone learning rule, and also no best absolutely expedient learning rule. We have not generalized this result to more than three actions, or other initial states. But our result suggests that the chances of finding best learning rules in general are slim. The proof of the following result is in the Appendix.

PROPOSITION 5: *Let $n = 3$, and consider the fixed initial state $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. No best monotone learning rule and no best absolutely expedient learning rule exists for this initial state.*

[7]A different learning rule with endogenous aspiration level was studied in Börgers and Sarin (2000). The rule studied there is not absolutely expedient.

Even though there is no best rule in the case $n \geq 3$, some rules might achieve in all environments a larger increase in expected payoff or the probability of the best strategies than other rules. We leave it to future research to investigate such "dominance" relations among learning rules.

8. RELATED LITERATURE

Schlag (1994) and Sarin (1995) study axioms for learning rules, among them absolute expediency. Because they add other axioms and assumptions to absolute expediency, they characterize a smaller class of learning rules than our paper. Schlag (1994) assumes that the rule is affine in payoffs and that the coefficients of the transformation of payoffs do not depend on the current mixed strategy. Sarin (1995) assumes that the rule by which the probability of an unchosen action is updated depends only on the payoff received, not on the action chosen. He also assumes a form of multiplicative separability of the learning rule.

A more recent paper by Schlag (2002) considers the case of two actions only. Schlag assumes that payoffs are identically and independently distributed in all time periods, and that the decision maker uses the same learning rule throughout. He calls learning rules "ex ante improving" if expected payoffs are monotonically increasing from each period to the next, where, expected values are taken unconditionally, i.e. before period 1 begins. Contrast this with absolute expediency in our paper. If a decision maker uses repeatedly an absolutely expedient rule,[8] then expected payoffs increase from each period to the next, not just in ex ante terms, but also in interim terms, i.e. if expected change in expected payoffs is calculated conditional on the mixed strategy at the beginning of each period. Schlag does not aim for a complete characterization of "ex ante improving" rules, but he selects among the "ex ante improving" rules those that are "best" according to further criteria. The rule that he then obtains is the same as the rule that we obtain as the "best" absolutely expedient rule in the case of two actions and uniform initial state.[9]

Schlag (2002) also considers absolute expediency as defined in this paper. He indicates that this property is in conflict with other desirable long-run properties, if attention is restricted to learning rules with small finite state space.[10]

A large set of papers related to ours can be found in the literature on machine learning, and specifically in the part that is concerned with the learning

[8]Recall from footnote 5 that some of the absolutely expedient learning rules considered in this paper cannot be used repeatedly, but that such rules can be closely approximated by rules that can be iterated.

[9]See Remark 7 in Section 7.

[10]See part (iii) of Proposition 5 (where the state space of the learning rule is assumed to be of cardinality 2) and part (ii) of Proposition 7 in Schlag (2002) (where the state space of the learning rule is assumed to be of cardinality 4).

behavior of stochastic automata.[11] In this literature, absolute expediency was originally defined by Lakshmivarahan and Thathachar (1973). Monotonicity is studied by Toyama and Kimura (1977) who refer to it as *absolute adaptability*.

The most general characterization of absolutely expedient learning rules in this literature of which we are aware is Theorem 6.1 in Narendra and Thathachar (1989). This result characterizes absolutely expedient learning rules assuming that the updating rule is affine in payoffs, and that the coefficients in the affine transformation of payoffs depend only on the action played, but not on the strategy whose probability is updated. Narendra and Thathachar also show that in their framework absolute expediency and monotonicity are equivalent.[12]

Toyama and Kimura (1977) characterize monotone learning rules. Like Narendra and Thathachar they assume linearity of the learning rule in payoffs whereas we derive it. They allow the coefficients of the payoff transformation to depend on the current state, but neither on the action that has been played nor on the action that is updated. Their results are implied by ours.[13]

Absolute expediency and monotonicity are also closely related to properties of "selection dynamics" studied in evolutionary game theory. These dynamics describe the evolution of the proportions of players playing different strategies in large populations. The analogue of absolute expediency in the evolutionary literature is *weak compatibility* as defined by Friedman (1991). Weak compatibility requires that the average population payoff increase over time. Friedman studies implications of weak compatibility but does not provide a characterization of weakly compatible evolutionary dynamics. It may be possible to adapt our results to an evolutionary setting, but we have not pursued this.

The closest analogue of monotonicity in the evolutionary literature is *payoff monotonicity*, which requires that the ordering of growth rates of the proportions of a population playing different strategies be the same as the ordering of expected payoffs. The evolutionary literature does not contain characterizations of the functional form of selection dynamics with these properties. Samuelson and Zhang's (1992) *aggregate monotonicity* is more restrictive than payoff monotonicity in that the requirement applies not only to pure but also to mixed strategies. Samuelson and Zhang, like us, find a connection between monotonicity and the replicator dynamics. They show that a selection dynamics satisfies aggregate monotonicity if and only if it is equivalent to replicator dynamics with linearly transformed payoffs. Their result is obtained by con-

[11] A useful overview of the literature on stochastic automata and learning has been provided by Narendra and Thathachar (1989), in particular Chapter 6.

[12] Narendra and Thathachar's assumptions about the form of the learning rule imply that every unbiased rule that is of this form must be cross-negative (using the terminology of this paper that is introduced below). Thus, our Propositions 3 and 4 imply the equivalence of absolute expediency and monotonicity in Narendra and Thathachar's framework.

[13] Note that our results imply that in their framework monotonicity and absolute expediency are actually equivalent.

sidering a single environment only, while it is essential for our results that a learning rule must operate in multiple environments.

Our work is also related to Schlag's (1998) work on imitation. He considers decision makers who observe the choices and payoffs of other decision makers facing the same environment. For the case of two actions Schlag characterizes imitation rules that ensure an increase in expected payoffs, averaged across the population. He finds that the imitation probability is proportional to payoffs, and that the resulting population dynamics is a rescaled version of the replicator dynamics.

Dept. of Economics and ELSE, University College London, Gower Street, London WC1E 6BT, United Kingdom; t.borgers@ucl.ac.uk; http://www.ucl.ac.uk/uctpa01/borgers.htm,

Departamento de Teoria e Historia Economica, Facultad de Ciencias Economicas y Empresariales, Universidad de Malaga, Plaza El-Ejido s/n, 29013 Malaga, Spain; amorales@uma.es; http://webdeptos.uma.es/theconomica/wpmoralesant.htm,

and

Department of Economics, Texas A&M University, College Station, TX 77843-4228, U.S.A.; rsarin@econ.tamu.edu; http://econweb.tamu.edu/rsarin/.

Manuscript received August, 2001; final revision received June, 2003.

APPENDIX

PROOF OF PROPOSITION 1: *Sufficiency*: If $S^* = S$, i.e. if there is some x such that $\pi_i = x$ for all $i = 1, 2, \ldots, n$, then the formula for $f(s_i)$ in Remark 3 becomes

$$f(s_i) = \sigma_i x \left(B_{ii} - \sum_{j=1}^{n} (\sigma_j B_{ji}) \right) \quad \text{for} \quad i = 1, 2, \ldots, n.$$

By condition (4) in Proposition 1 the term in big brackets equals zero, and thus $f(s_i) = 0$ for all $i = 1, 2, \ldots, n$.

Necessity: We proceed in three steps.

Step 1: If L is unbiased, then for all $s_j, s_i \in S$ the function $L(s_j, x)(s_i)$ is affine in x.

PROOF: Let L be an unbiased learning rule, and consider two environments, E and $\widetilde{E}$. In environment E all strategies receive some payoff x with $0 < x \leq 1$ with certainty. In environment $\widetilde{E}$ some strategy $s_j \in S$ receives payoff 1 with probability x, and payoff 0 with probability $1 - x$. All other strategies receive again payoff x with certainty. Both environments are then such that all strategies have the same expected payoff. Therefore, unbiasedness requires that in both environments the expected change in the probability assigned to any strategy s_i is zero. Denoting by $f(s_i)$ expected changes in probabilities in environment E, and by $\tilde{f}(s_i)$ expected changes in

probabilities in environment $\widetilde{E}$, we obtain thus for arbitrary strategy $s_i \in S$:

$$f(s_i) = \sigma_j L(s_j, x)(s_i) + \sum_{k=1, k\neq j}^{n} \sigma_k L(s_k, x)(s_i) - \sigma_i = 0,$$

$$\widetilde{f}(s_i) = \sigma_j x L(s_j, 1)(s_i) + \sigma_j (1-x) L(s_j, 0)(s_i) + \sum_{k=1, k\neq j}^{n} \sigma_k L(s_k, x)(s_i) - \sigma_i = 0.$$

Subtracting these two equations from each other yields

$$\sigma_j L(s_j, x)(s_i) - \sigma_j x L(s_j, 1)(s_i) - \sigma_j (1-x) L(s_j, 0)(s_i) = 0.$$

Dividing by σ_j and rearranging one obtains

$$L(s_j, x)(s_i) = L(s_j, 0)(s_i) + \big(L(s_j, 1)(s_i) - L(s_j, 0)(s_i)\big)x.$$

Thus we have concluded that $L(s_j, x)(s_i)$ is an affine function of x. Note that our argument is true for arbitrary pairs of strategies s_j and s_i.

Step 2: If the function $L(s_j, x)(s_i)$ is affine in x, then it can be written in the form asserted in Proposition 1.

PROOF: Consider first the case $j = i$. We can write the formula for $L(s_i, x)(s_i)$ in Proposition 1 as: $\sigma_i + (1-\sigma_i)A_{ii} + (1-\sigma_i)B_{ii}x$. Now recall the last equation in Step 1. Clearly, we can choose A_{ii} such that $\sigma_i + (1-\sigma_i)A_{ii} = L(s_i, 0)(s_i)$, and we can choose B_{ii} such that $(1-\sigma_i)B_{ii} = (L(s_i, 1)(s_i) - L(s_i, 0)(s_i))$. The last equation in Step 1 then shows that with these definitions $L(s_i, x)(s_i)$ has the form asserted in Proposition 1. For $L(s_j, x)(s_i)$ where $j \neq i$ we can proceed analogously.

Step 3: The coefficients have to satisfy the restrictions (3) and (4).

PROOF: Suppose that all actions give the same deterministic payoff x. Then the expected change in the probability of strategy s_i can be calculated using formulas (1) and (2) in Proposition 1. One obtains

$$f(s_i) = \sigma_i \left[\left(A_{ii} - \sum_{j=1}^{n} \sigma_j A_{ji} \right) + \left(B_{ii} - \sum_{j=1}^{n} \sigma_j B_{ji} \right) x \right].$$

This expression has to be zero for all $x \in [0, 1]$. This can only be true if both expressions in big round brackets equal zero. This is what conditions (3) and (4) require. *Q.E.D.*

PROOF OF PART (ii) OF PROPOSITION 3: Let L be a monotone learning rule. We will prove the assertion by induction over the number of different expected payoffs available in the environment, i.e. over $\sharp\{x \in [0, 1] | \pi_i = x \text{ for some } i = 1, 2, \ldots, n\}$. We will begin with the case that this number is 2, i.e. there are two different payoffs, π^+ and π^-, with $\pi^+ > \pi^-$. Then

$$g = \pi^+ \sum_{s_i \in S^*} f(s_i) + \pi^- \sum_{s_i \notin S^*} f(s_i) = (\pi^+ - \pi^-) f(S^*) > 0.$$

Now suppose we had shown the assertion for all environments E with $\sharp\{x \in [0, 1] | \pi_i = x \text{ for some } i = 1, 2, \ldots, n\} = \nu - 1$, and consider an environment E such that $\sharp\{x \in [0, 1] | \pi_i = x \text{ for some } i = 1, 2, \ldots, n\} = \nu$. Denote the set of all strategies with the lowest expected payoff level

by $\overline{S}$. Denote the corresponding expected payoff level by $\overline{\pi}$. Denote the set of all strategies with the second lowest expected payoff level by $\widehat{S}$. Denote the corresponding expected payoff level by $\widehat{\pi}$. Define $k \equiv \widehat{\pi} - \overline{\pi}$ and note that $k > 0$. Consider a modified environment in which the expected payoff of all strategies in $\overline{S}$ is raised to $\widehat{\pi}$. Denote the expected change of payoffs in this modified environment by g'. By the inductive assumption we know that $g' > 0$. We shall now show that $g - g' > 0$. This then obviously implies the claim.

To calculate $g - g'$ we denote for every $s_i \in S$ by $f'(s_i)$ the expected change in the probability of strategy s_i in the modified environment. Then:

$$\begin{aligned} g - g' &= \sum_{s_i \notin \overline{S}} f(s_i)\pi_i + \sum_{s_j \in \overline{S}} f(s_j)\overline{\pi} \\ &\quad - \sum_{s_i \notin \overline{S}} f'(s_i)\pi_i - \sum_{s_j \in \overline{S}} f'(s_j)(\overline{\pi} + k) \\ &= \sum_{s_i \notin \overline{S}} (f(s_i) - f'(s_i))\pi_i \\ &\quad + \sum_{s_j \in \overline{S}} (f(s_j) - f'(s_j))\overline{\pi} - \sum_{s_j \in \overline{S}} f'(s_j)k. \end{aligned}$$

Using equation (5) we have for strategies $s_i \notin \overline{S}$

$$f(s_i) - f'(s_i) = \sigma_i \sum_{s_j \in \overline{S}} \sigma_j B_{ji} k.$$

Because the sum of the probabilities cannot change, we can conclude that

$$\begin{aligned} \sum_{s_j \in \overline{S}} (f(s_j) - f'(s_j)) &= -\sum_{s_i \notin \overline{S}} (f(s_i) - f'(s_i)) \\ &= -\sum_{s_i \notin \overline{S}} \sum_{s_j \in \overline{S}} \sigma_i \sigma_j B_{ji} k. \end{aligned}$$

Using these formulas, we can rewrite our earlier equation as

$$\begin{aligned} g - g' &= \sum_{s_i \notin \overline{S}} \sum_{s_j \in \overline{S}} \sigma_i \sigma_j B_{ji} k \pi_i \\ &\quad - \sum_{s_i \notin \overline{S}} \sum_{s_j \in \overline{S}} \sigma_i \sigma_j B_{ji} k \overline{\pi} - \sum_{s_j \in \overline{S}} f'(s_j)k \\ &= k \sum_{s_i \notin \overline{S}} \sum_{s_j \in \overline{S}} \sigma_i \sigma_j B_{ji} (\pi_i - \overline{\pi}) - k \sum_{s_j \in \overline{S}} f'(s_j). \end{aligned}$$

We will prove that the above expression is positive. The first term in this difference is evidently strictly positive, because L is monotone (i.e. $B_{ji} \geq 0$), $\pi_i > \overline{\pi}$ and $k > 0$. It remains to prove that $\sum_{s_j \in \overline{S}} f'(s_j) < 0$. But this is true because cross-negativity implies that the expected change in the probability of the set of worst strategies is strictly negative. The proof is analogous to the sufficiency proof of part (i) of Proposition 3. We conclude that $g - g' > 0$, as required. *Q.E.D.*

PROOF OF PROPOSITION 5: *No best monotone rule exists*: Our proof is indirect. Let L^* be a best monotone learning rule for initial state $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. It has to attain the largest expected gain in the probability of the best actions for every environment. In particular, this has to be true for environments with $\pi_i > \pi_j = \pi_k$, for $i, j, k = 1, 2, 3$ and $i \neq j, k$. For these environments, the expected change in the probability of strategy s_i is

$$f(s_i) = \frac{2}{9} B_{ii}(\pi_i - \pi_k).$$

This is largest if B_{ii} is largest. As in the proof of Proposition 4, it is easy to verify that the set of admissible values for B_{ii} has a strictly positive upper bound. Let B denote this upper bound. As the argument applies to arbitrary i, we conclude that $B_{ii} = B$ for every $i \in \{1, 2, 3\}$.

This result, together with condition (4) of Proposition 1, and with the restrictions for the (B_{ij})-matrix implied by the fact that updated learning probabilities have to add up to one, implies that the matrix of B_{ij}-coefficients must be of the following form (where we denote the coefficient B_{12} by b):

$$(B_{ij}) = \begin{pmatrix} B & 2B-b & b \\ b & B & 2B-b \\ 2B-b & b & B \end{pmatrix}.$$

The expected change in the probability with which strategy s_1 is played is then

$$\frac{1}{3}\left[B\pi_1 - \left(\frac{1}{3}B\pi_1 + \frac{1}{3}b\pi_2 + \frac{1}{3}(2B-b)\pi_3\right)\right].$$

Now suppose $\pi_1 > \pi_2 > \pi_3$. Then the above expression becomes larger as b gets larger. On the other hand, if $\pi_1 > \pi_3 > \pi_2$, then the above expression becomes larger as b gets smaller. Thus, no value of b maximizes the above expression in all environments. This contradicts the existence of a best monotone learning rule L^*.

No best absolutely expedient rule exists: Like the proof in the first part, also this proof is indirect. Using the same arguments as in the first part, one shows that the matrix (B_{ij}) has to have the form derived in the first part.

If the matrix of B_{ij}-coefficients is of this form, then the expected movement of expected payoffs is

$$g = \frac{B}{3}\sum_{i=1}^{3}{\pi_i}^2 - \frac{2B}{9}\sum_{i=1}^{3}\sum_{j=i}^{3}\pi_i\pi_j.$$

Note that this is independent of b. Thus, if there is any best absolutely expedient learning rule, then all learning rules with a (B_{ij})-matrix that is of the form derived above will be best absolutely expedient. One possible choice for b is: $b = B$. This is the choice on which we focus. With this choice of b it follows that $B_{ij} = B$ for all $i, j = 1, 2, 3$.

Now let δ satisfy $0 < \delta < \frac{B}{2}$, and consider an alternative rule L' with the following matrix of B_{ij}-coefficients:

$$(B'_{ij}) = \begin{pmatrix} B & B-2\delta & B+2\delta \\ B & B-\delta & B-2\delta \\ B & B & B \end{pmatrix}.$$

This matrix satisfies the restrictions of Proposition 1. All entries of this matrix are strictly positive. Therefore, this rule is monotone and hence absolutely expedient. With this rule, the expected

change in expected payoffs is

$$g' = \frac{1}{3}\left(B\pi_1^2 + (B-\delta)\pi_2^2 + B\pi_3^2\right) - \frac{1}{9}\left(B\pi_1^2 + (B-\delta)\pi_2^2 + B\pi_3^2 + 2(B-\delta)\pi_1\pi_2 + 2(B+\delta)\pi_1\pi_3 + 2(B-\delta)\pi_2\pi_3\right).$$

Differentiating this with respect to δ yields

$$\frac{\partial g'}{\partial \delta} = \frac{2}{9}(\pi_1 - \pi_2)(\pi_2 - \pi_3).$$

Clearly, if $\pi_1 \neq \pi_2$ and $\pi_2 \neq \pi_3$, this derivative is not equal to zero. Thus, either by raising δ, or by lowering it, a higher value of the expected change in expected payoffs can be achieved. This contradicts the assumption that the learning rule that we are considering, which corresponds to the case $\delta = 0$, is best absolutely expedient. *Q.E.D.*

REFERENCES

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

——— (2000): "Naive Reinforcement Learning with Endogenous Aspirations," *International Economic Review*, 41, 921–950.

BUSH, R., AND F. MOSTELLER (1951): "A Mathematical Model for Simple Learning," *Psychological Review*, 58, 313–323.

CROSS, J. (1973): "A Stochastic Learning Model of Economic Behavior," *Quarterly Journal of Economics*, 87, 239–266.

EREV, I., AND A. ROTH (1998): "Predicting How People Play Games: Reinforcement Learning in Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, 88, 848–881.

FRIEDMAN, D. (1991): "Evolutionary Games in Economics," *Econometrica*, 59, 637–666.

FUDENBERG, D., AND D. LEVINE (1998): *The Theory of Learning in Games*. Cambridge and London: MIT Press.

LAKSHMIVARAHAN, S., AND M. THATHACHAR (1973): "Absolutely Expedient Learning Algorithms for Stochastic Automata," *IEEE Transactions on Systems, Man and Cybernetics*, 3, 281–286.

NARENDRA, K., AND M. THATHACHAR (1989): *Learning Automata: An Introduction.* Englewood Cliffs: Prentice-Hall.

ROTH, A., AND I. EREV (1995): "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, 8, 164–212.

SAMUELSON, L., AND J. ZHANG (1992): "Evolutionary Stability in Asymmetric Games," *Journal of Economic Theory*, 57, 363–391.

SARIN, R. (1995): "Learning Through Reinforcement: The Cross Model," Unpublished Manuscript, Texas A&M University.

SCHLAG, K. (1994): "A Note on Efficient Learning Rules," Unpublished Manuscript, University of Bonn.

——— (1998): "Why Imitate, and if so How? A Bounded Rational Approach to Multi-Armed Bandits," *Journal of Economic Theory*, 78, 130–156.

——— (2002): "How to Choose—A Boundedly Rational Approach to Repeated Decision Making," Unpublished Manuscript, European University Institute.

TOYAMA, Y., AND M. KIMURA (1977): "On Learning Automata in Nonstationary Random Environments," *Systems, Computers, Controls*, 8, 66–73.