Expedient and Monotone Learning Rules[1]

Tilman Börgers[2] Antonio J. Morales[3] Rajiv Sarin[4]

July 2001

[1] This is a revised and extended version of some sections of our earlier paper "Simple Behaviour Rules Which Lead To Expected Payoff Maximising Choices".

[2] Department of Economics and ELSE, University College London, Gower Street, London WC1E 6BT, United Kingdom, t.borgers@ucl.ac.uk.

[3] CentrA y Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, Plaza El Ejido s/n, 29013 Málaga, Spain, amorales@uma.es.

[4] Department of Economics, Texas A&M University, College Station, TX 77843, U.S.A., rsarin@econ.tamu.edu.

Abstract

This paper considers a learning rules for environments in which little prior information is available to the decision maker. Two properties of such learning rules, absolute expediency and monotonicity, are studied. The paper provides some necessary, and some sufficient conditions for these properties. A number of examples show that the there is a quite a large variety of learning rules which have the properties. It is also shown that all learning rules which have these properties are, in some sense, related to the replicator dynamics of evolutionary game theory.

1 Introduction

We arrive at most economic decisions of our lives through a learning process in which we adjust our behaviour in response to experience. For example, we learn in this way which consumer products we like, we learn how to invest our money, or we learn how to behave towards our colleagues at work. For economic theory it is therefore interesting to explore mathematical models of learning. There has been a large literature on this subject which has recently been surveyed, for example, by Fudenberg and Levine (1998).

A problem in this area is that there is a huge variety of learning models and that existing results are often specific to very particular models. This problem is particularly severe in learning environments in which the decision maker has very little prior information about the alternatives among which he chooses. If much prior information is available, then a Bayesian model of learning with a limited state space in which learning is simply updating of subjective beliefs over the state space seems plausible. But if the prior information is very incomplete, then the set of conceivable states of the world is often so large that a Bayesian model of learning seems highly unrealistic. Once one turns away from Bayesian learning models, however, it is difficult to see on which grounds to choose one learning model rather than another.

In this paper we want to contribute to a solution of this problem. Our approach is to begin with a very large and general class of learning models, and then to investigate which of these models have a particular, intuitively attractive property. We do not obtain a complete characterisation of all learning models with this property, but we present a number of results which together constitute a significant step towards such a characterisation.

Our paper is useful in two ways. Firstly, for theoretical work, we identify a particular class of learning models which have important properties in common, and which therefore provide an interesting object for further study. Our results also provide for theoretical research a shortcut which allows researchers to see easily whether any given learning rule belongs to this class. Secondly, for experimental work on learning, we identify some salient properties of learning behaviour which experimenters can seek to test on their data without having to postulate any particular functional forms.

The properties on which we focus are two *short run* properties, labeled "absolute expediency"[1] and "monotonicity"[2]. Roughly speaking, they say that the performance of the decision maker improves from each period to the next provided that the environment stays the same. The properties require this to be true in every environment in a very large class of environments. It is this universality of the requirement that gives it bite. The properties which we study appear attractive when the agent acts in an unknown environment which might change from time to time. Then "the best the decision maker can do" is to try to guarantee an upward trend in his performance.

Our main results provide some necessary and some sufficient conditions for absolutely expedient and monotone reinforcement learning procedures. Roughly speaking, we find that a necessary condition for absolute expediency or monotonicity is that local learning, i.e. learning which begins in a given and fixed current state of the decision maker, is *as if* the decision maker used a modified version of Cross' (1973) learning rule. Cross' learning rule is based on Bush and Mosteller's (1951, 1955) stochastic learning theory. It says that the decision maker adjusts his choice probabilities proportional to the payoff received. The modifications of this rule which are compatible with absolute expediency or monotonicity are learning rules which result if payoffs are subject to certain affine transformations before the rule is applied.

We know from earlier work (Börgers and Sarin (1997)) that there is a close connection between Cross' reinforcement learning model and the replicator dynamics of evolutionary game theory. Therefore, our results in this paper also imply a connection between absolutely expedient, and monotone learning rules and the replicator dynamics.

Note that our results are only *as if* results. We are not saying that the decision maker has to actually use Cross' rule. In fact, our characterisation encompasses a large variety of learning rules which, at first sight, look very different from Cross' rule. In particular, the characterisation applies to learning rules which have larger state spaces and therefore also larger memory than Cross' rule.

We also study sufficient conditions for absolute expediency or monotonicity. One interesting and, to us, surprising result is that there are absolutely

[1] This expression originates in the literature on machine learning (Lakshmivarahan and Thathachar (1973)). We discuss the connection between our work and that literature in Section 9 below.

[2] Monotonicity as defined in this paper is related to, but different from the monotonicity requirements studied in the evolutionary game theory literature. We discuss this connection in Section 3 below.

expedient learning rules which have in-built preconceptions about similarity relations among strategies. According to such rules the larger the payoff received when a strategy s_1 has been played, the larger is in the next round not only the probability of s_1 but also the probability of some other strategy s_2. Intuitively, the decision maker treats s_1 and s_2 as if they were “similar” strategies. It is surprising that learning rules of this type can be absolutely expedient because absolute expediency requires an improvement in *all* conceivable environments, including environments in which s_1 and s_2 are very different from each other. We shall show in an example that in-built similarity relations among strategies are compatible with absolute expediency if the payoff success of a strategy s_1 has a much stronger effect on the probability with which s_1 is chosen than on the probability with which a “similar” strategy s_2 is chosen.

Our paper has the flavor of an axiomatic approach to learning theory. However, we emphasize that our objective is *not* to isolate a *single* best learning rule. That would not be useful to either the theoretical researcher who wishes to obtain results for broader classes of learning rules nor for the experimental researcher who faces heterogeneity in the learning behaviour of subjects. This is one of several point which distinguishes our paper from papers such as Easley and Rustichini (1999) and Schlag (2001). We shall discuss these and other related papers in Section 9 below.

This paper is organised as follows. In Section 2 we introduce the simplest framework in which our analysis can be conducted. We call it a “local” model of learning because we shall only consider two periods, “today” and “tomorrow”, and because we shall take the decision maker’s behaviour today as given and fixed. One should think of the model in Section 2 as a “reduced” form model of learning. How such a model of learning can result from a much richer, fully specified model of learning is explained later in the paper. In Section 3 we formalize and motivate the notions of absolute expediency and monotonicity. In Section 4 we give Cross’ learning rule as an example of a simple learning rule which has these properties. Section 5, 6 and 7 contain our characterisation results. Section 8 extends our analysis to learning rules with very general state spaces, and provides some indication of the long run behaviour implied by absolute expediency and monotonicity. Section 9 discusses related literature. Section 10 concludes.

2 A Local Model of Learning

A decision maker chooses repeatedly from a finite set $S = \{s_1, s_2, ..., s_n\}$ of strategies which has at least two elements: $n \geq 2$. Every strategy s_i has a payoff distribution μ_i attached to it. We normalize payoffs so that they are between zero and one. The substantial assumption here is that there is *some* upper and *some* lower bound for payoffs. These may be arbitrarily large or small, respectively, and payoffs can then be normalized so that they are between zero and one. In the following definition an assignment of payoff distributions to strategies is called an environment.

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

For our analysis it does not matter where the payoff distributions come from. They could reflect randomness in nature, for example. It could also be that the decision maker is involved in a game, and that the payoff distributions reflect other players' behaviour.

In this section we shall study the decision maker's behaviour at two dates only, "today" and "tomorrow". The decision maker's behaviour today will be exogenous. We shall study learning rules which determine how the decision maker adjusts his behaviour tomorrow in response to his experiences today.

The decision maker knows the strategy set S, and he also knows that his strategy set tomorrow is the same set as it is today. But the decision maker does not know the environment E, and he doesn't know whether it will be the same tomorrow as it is today. He chooses a strategy from S today, and then observes the payoff realization. In response to this observation he has to choose a strategy from S tomorrow.

The decision maker's behaviour today is described by an exogenous probability distribution σ over S. This distribution describes how likely the decision maker is to choose each of his strategies today. Note that we do not assume that the decision maker is consciously randomizing. The distribution σ describes the likelihood of various strategies from the point of view of an outside observer.

We denote the probability which σ assigns to the strategy s_i by σ_i. Thus, σ is the vector: $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$. We make the following assumption.

Assumption 1. *For every* $i = 1, 2, ..., n$ *the probability* σ_i *is strictly positive.*

The reason why this assumption is important to our analysis is that it implies that there is a positive probability that the decision maker will today play his optimal strategies. The problem which "learning" and the "learning rule" then have to solve is that they need to recognize a good strategy when it has been played, but they don't have to discover good strategies. By contrast, if some strategies are played with zero probability, then it may be that exactly those strategies are the "good" strategies, and that the learning process has to discover these strategies. This is a much harder task than the one which we study in this paper.

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

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

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 which he receives today (which is distributed according to μ_i), which mixed strategy the decision maker chooses tomorrow. Denote by $L(s_i, x)(s_j)$ the probability which 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 behaviour. If one combines the two steps of belief updating and behaviour adjustment one arrives at a learning rule in the sense of Definition 2, and our work below directly applies if the true learning process is of the type just described.

Our approach here is therefore much more general than an approach which focuses on learning rules in which the strategy simplex $\Delta(S)$ is the state space of the learning rule.[3] We shall give in Section 8 below examples of learning rules with much larger state spaces to which our analysis can nonetheless be applied. This is because for any given state of the decision maker a reduced form of the learning rule can be constructed which is of the form postulated in Definition 2.

An important assumption which we shall make is this:

[3] Our earlier paper, "Simple Behaviour Rules Which Lead to Expected Payoff Maximising Choices", took such an approach.

Assumption 2. *For all $s_i \in S$ the learning rule L is continuous in x.*

It seems to us reasonable that a decision maker with little information about his environment responds in a continuous way to payoff observations.

The perspective adopted in this section is in two senses "local". Firstly, we only focus on two days, today and tomorrow. We do not investigate the decision maker's behaviour over long time periods. Secondly, we take today's behaviour as exogenously given and fixed. We do not consider a variety of possible initial behaviours. We shall adopt a more "global" perspective in Section 8 below.

3 Absolute Expediency and Monotonicity

In this paper we study learning rules which guarantee improvements (in a sense to be made precise) in the decision maker's performance tomorrow in comparison to today. The scenario for which we check whether a learning rule achieves this improvement assumes that the payoffs tomorrow have the same distribution as the payoffs today, and that payoffs tomorrow are stochastically independent of the payoffs today. In other words: payoffs have to be i.i.d. Our justification for focusing on this scenario is derived from the agent's ignorance of his environment. If the agent knows very little about his environment, then it seems that "the best that he can do" is to aim for an improvement in his performance, assuming that payoffs tomorrow have the same distribution as payoffs today, and that tomorrow's payoff realisations are not in some way pre-determined by today's realisations.

Now let E be the environment which prevails today, and which will be assumed to prevail tomorrow as well. Because we assume that the decision maker does not know the environment E we shall ask that the decision maker's performance improves in any conceivable environment E. This rules out learning rules which trade off big improvements in some environments against deteriorations in the decision maker's performance in some other environments. Such learning rules seem to rely implicitly on the view that some environments are more important than others. By contrast, we want to study a decision maker who is entirely ignorant of his environment. Therefore, it seems reasonable to require improvement in performance for all environments E.

We now say more formally what we mean by improvement in the decision maker's performance. We first need more notation. 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 maximising strategies is denoted by S^*, that is $S^* = \{s_i \in S \,|\, \pi_i \geq \pi_j \text{ for all } j = 1, 2, ..., n\}$. Of course, π_i and S^* depend on E. But to keep our notation simple, we suppress the dependence on E in the notation.

Now fix not only the environment E but also some learning rule L. For every strategy s_i denote by $f(s_i)$ the expected change in the probability attached to s_i:

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

We extend this definition 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 environment E and the learning rule L, but, to keep things simple we suppress that dependence in our notation.

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 non-trivial environments expected payoffs are on average strictly higher tomorrow than they were today. An environment is "non-trivial" 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 non-optimal strategies.

We now provide a second formalization of the notion of improvement in the decision maker's performance which is somewhat different from Definition 3. Our second formalization will require that the probability assigned to the best actions increases in all non-trivial environments. This is not the main property studied in this paper because it is by definition payoffs, not action probabilities, which the decision maker cares about, and hence a property which refers to payoffs is intuitively more appealing. However, the second definition leads to simpler characterisations, and it is closely related to the first definition, as we shall show in this paper.

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 an issue which we study in detail 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.*

Monotonicity as defined above is closely related to properties of selection dynamics studied in evolutionary game theory.[4] Selection dynamics describe the evolution of the proportions of players playing different strategies in large populations. *Weak payoff-positivity* requires that *some* best reply has positive growth rate. It is weaker than monotonicity in the sense of Definition 4 because it does not require the set of *all* best replies to have positive growth rate. The more restrictive property of *payoff positivity* requires that all strategies which have more than average payoff have positive growth rates. This property is clearly more restrictive than monotonicity. *Payoff monotonicity* requires that the ordering of growth rates reflects the ordering of expected payoffs. That, too, is obviously more restrictive than monotonicity. The evolutionary literature does not contain characterisations of the functional form of selection dynamics with these properties, nor does it trace them back to the behaviour rules of individuals, like we do.

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 show that a selection dynamics satisfies aggregate monotonicity if and only if it is a equivalent to replicator dynamics with linearly transformed payoffs. Their work is related to ours as we, too, find below a connection between monotonicity and the replicator dynamics. The monotonicity requirement with which they work is more restrictive than ours. On the other hand, their result is obtained by considering a single environment only. By contrast, it is essential for our results that a learning rule must operate in multiple environments.

[4] All properties of selection dynamics mentioned in this paragraph are discussed in more detail in Section 5.5 of Weibull (1995).

4 An Example

The following example is taken from Cross (1973). We begin with this example because later we shall show that all learning rules which are absolutely expedient or monotone have some structural similarities with this example.

Example 1 *For all* $i, j \in \{1, 2, ..., n\}$ *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 shifts the probability of s_i into the direction of 1 where the size of the shift is proportional to x. If $x = 1$ then the decision maker goes all the way and 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 action which he actually played, even if the payoff was very low. For the moment this will not concern us. Although the example is important for our analysis, this particular feature will not be crucial.

We now show that Cross' learning rule is absolutely expedient and monotone. The expected movement of payoffs under Cross' learning rule is given by:

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

One can interpret the right hand side as the variance of expected payoffs 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.

The expected movement of the probability of any particular pure strategy s_i is under Cross' rule:

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

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 today of the expected payoff.[5] The condition $S^* \neq S$ and Assumption 1 imply that for strategies in S^* the difference between this strategy's expected payoff and the expected value of expected payoffs 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, i.e. the equation which describes in evolutionary game theory 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 in more detail in Börgers and Sarin (1997).

In the next section we shall show that all absolutely expedient or monotone learning rules have structural similarities with Cross' learning rule.

5 Unbiasedness

As a first step we study a relatively weak property which 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 $i = 1, 2, ..., n$.*

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 "favor" in some sense the

[5] The phrase "expected value of the expected payoff" seems at first sight strange, but recall our explanation in the previous paragraph of why expected payoffs are a random variable.

former strategies. This motivates why we refer to the property in Definition 5 as "unbiasedness".

The next lemma shows why this property is relevant to our analysis.

Lemma 1 *Every absolutely expedient and also every monotone learning rule is unbiased.*

Proof: Consider a biased learning rule L, i.e. suppose that we can find an environment E with $S = S^*$ such that for some strategy $s_i \in S$ we have: $f(s_i) < 0$. Now make a small change in the payoff distribution of s_i, so that the expected payoff of s_i increases, while leaving all other payoff distributions unchanged. In the new environment s_i is the *unique* expected payoff maximising strategy. But, because of the continuity of L in x (Assumption 2), for sufficiently small increase in $s_i^\prime$s payoff it will still be the case that $f(s_i) < 0$. This implies for the new environment: $g < 0$ and $f(s_i) < 0$. Thus we have obtained a contradiction to the assumption that L is absolutely expedient or monotone.

Q.E.D.

Our strategy is now to characterise first all unbiased learning rules, and then, building on this initial characterisation, to ask which additional conditions absolutely expedient or monotone learning rules have to satisfy.

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

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

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

and for every $i = 1, 2, ..., n$:

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

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

This result shows that 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 which he has played and on the strategy the probability of which he is adjusting. Conditions (3) and (4) restrict the coefficients of the linear transformation. They require that the coefficients of the affine transformation which are applied when s_i was played and s_i is updated are the expected values of the coefficients which are used when s_j was played and s is updated.

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

Before we prove Proposition 1 it is worthwhile to state the following observation which follows from elementary calculations.

Remark 2 *Let L satisfy the characterisation in Proposition 1, and let E be an environment. Then the expected movement of expected payoffs is given by:*

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

and for all $s_i \in S$ the expected movement of the probability of s is given by:

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

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 second formula, which once again is reminiscent of the replicator dynamics. The first 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.

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

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

By condition (4) in Proposition 1 the squared brackets equal zero, and thus $f(s_i) = 0$ for all $i = 1, 2, ..., n$.

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 any arbitrary strategy $s_i \in S$:

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

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

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 re-arranging one obtains:

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

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 actions 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) say.

Q.E.D.

We now ask which conditions the coefficients in Proposition 1 need to satisfy to ensure that an unbiased rule is either absolutely expedient or monotone. Notice that only the coefficients B_{ji} appear in the two formulas in Remark 2. Therefore, our investigation will focus on these coefficients.

Definition 6 *An unbiased learning rule L is* own-positive *if $B_{ii} > 0$ for all $i = 1, 2, ..., n$.*

This property means that the probability that the decision maker plays tomorrow the strategy which he played today increases in the payoff which the decision maker received today. It is a very plausible property. The following result shows that the learning rules which we want to characterise in this paper are own-positive.

Proposition 2 *An unbiased learning rule L which is absolutely expedient or monotone is own-positive.*

Proof: Let L be unbiased and absolutely expedient or monotone. Consider an environment in which all actions have the same expected payoff $x < 1$. By unbiasedness: $f(s_i) = 0$ for all $i = 1, 2, ..., n$. Now add some $\varepsilon > 0$ to the expected payoff of some action s_i. It is easy to calculate from the formulas in the proof of Proposition 1 that in this new environment $f(s_i) = \sigma_i(1-\sigma_i)B_{ii}\varepsilon$. The probability assigned to all the other strategies taken together changes by the negative of that amount. Now clearly $f(s_i)$ has to be positive for L to be absolutely expedient or monotone. This requires that $B_{ii} > 0$. This construction can be done for every strategy $s_i \in S$.

Q.E.D.

Unbiasedness and own-positivity are necessary, but in general not sufficient for absolute expediency. We shall demonstrate this with examples in Section 7 below. To proceed further we now introduce a further, more restrictive property.

Definition 7 *An unbiased learning rule L is* cross-negative *if*
(i) $B_{ji} \geq 0$ for all $i, j \in \{1, 2, ..., n\}$ with $i \neq j$; and
(ii) if C is a subset of S such that $C \neq \emptyset$ and $S \backslash C \neq \emptyset$ then there are strategies $s_i \in C$ and $s_j \in S \backslash C$ such that $B_{ji} > 0$.

Condition (i) in this definition means that if the decision maker played a strategy s_j today the probability that he plays a different strategy s_i tomorrow is non-increasing in the payoff which he received today. This rules out that the decision maker finds that s_j is "similar" to s_i, and that therefore a success today with s_j is encouraging news also for s_i. One can easily imagine circumstances in which this condition is not plausible, and restricts learning behaviour unduly. But recall that we are considering a decision maker who is ignorant about his environment. For such a decision maker this condition might seem plausible.

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.

Remark 3 *(i) If $n = 2$ then an unbiased learning rule is cross-negative if and only if it is own-positive.*

(ii) If $n \geq 3$ then an unbiased learning rule which is cross-negative is also own-positive, but not vice versa.

This remark follows directly from condition (4) of Proposition 1. The only exception is the comment "but not vice versa" in condition (ii). This comment will be proved by examples which we give in Section 7 below.

In the next two sections we shall now restrict attention to unbiased learning rule which are own-positive. We shall first discuss cross-negative rules and then we discuss other rules which are not cross-negative. In each case our focus will be on the question whether the rules which we are considering are absolutely expedient or monotone.

6 Cross-Negative Rules

The first result in this section shows that cross-negativity is sufficient for absolute expediency.

Proposition 3 *An unbiased learning rule L which is cross-negative is absolutely expedient.*

Proof: In this proof we shall find it convenient to work with the following formula for the expected change in a strategy's probability under unbiased learning rules. The formula is, of course, equivalent to the formula given in Remark 2, and follows from that formula by simple algebra.

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

Suppose that L is an unbiased, cross-negative learning rule. Let E be an environment such that $S^* \neq S$, i.e. $\pi_i \neq \pi_j$ for some $i, j \in \{1, 2, ..., n\}$. We need to prove that $g > 0$. We shall prove this by induction over the number of expected payoff values which are possible in E, i.e. by induction over $\sharp\{x \in [0,1] | \pi_i = x \text{ for some } i = 1, 2, ..., n\}$.

We begin with the case $\sharp\{x \in [0,1] | \pi_i = x \text{ for some } i = 1, 2, ..., n\} = 2$. Let the difference between the two payoff levels be $\varepsilon > 0$. Consider the expected change in the probability of the better strategies. Using formula (*) we can write this as:

$$f(S^*) = \sum_{s_i \in S^*} \sigma_i \sum_{s_j \in S \backslash S^*} \sigma_j B_{ji} \varepsilon$$

All expressions over which the sum is taken on the right hand side are non-negative. Moreover, condition (ii) in the definition of cross-negativity implies that some expressions have to be positive. Therefore, the total probability of S^* must increase and the probability of the set $S \backslash S^*$ must decrease. This implies $g > 0$.

Now suppose we had shown the assertion for all environments E with $\sharp\{x \in [0,1] | \pi_i = x \text{ for some } i = 1, 2, ..., n\} = \nu - 1$, and consider an environment E such that $\sharp\{x \in [0,1] | \pi_i = x \text{ for some } i = 1, 2, ..., n\} = \nu$. 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 $k \equiv \widehat{\pi} - \overline{\pi}$. 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 $s \in S$ by $f'(s)$ the expected change in the probability of strategy s 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 formula (*) 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 can't 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:

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

The first term in this difference is evidently strictly positive. The second term, which is subtracted, is non-negative because for every $s_j \in \overline{S}$ the expected change $f'(s_j)$ is non-positive. This is evident from formula (*) and the fact that the strategies in $\overline{S}$ are among the strategies with the lowest payoff in the modified environment. All the factors $(\pi_i - \pi_j)$ in formula (*) will be negative or zero. We can conclude that $g(\sigma) - g'(\sigma) > 0$, as required. Q.E.D.

If we consider monotonicity instead of absolute expediency then we find a stronger result than Proposition 3. Cross-negativity turns out to be both necessary and sufficient for monotonicity.

Proposition 4 *An unbiased learning rule is monotone if and only if it is cross-negative.*

Proof: Sufficiency: From (*) in the proof of Proposition 2 we have for every $s_i \in S$:

$$f(s_i) = \sigma_i \sum_{\substack{j=1 \\ j \neq i}}^{n} (\sigma_j B_{ji} (\pi_i - \pi_j)).$$

Consider any strategy $s_i \in S^*$. If L is cross-negative then all the expressions in the sum on the right hand side are non-negative, and hence the probability of s_i will not decrease in expected terms. If moreover, $S^* \neq S$ then we know by condition (ii) in Definition 7 that there exist $s_i \in S^*$ and $s_j \in S^* \backslash S$ such that $B_{ji} > 0$, and hence the expected change in the probability with which strategy s is played is strictly positive. Thus we can conclude that also $f(S^*) > 0$.

Necessity: Suppose that L is monotone. By Lemma 1 it is unbiased. We now prove that it also is cross-negative. We begin by proving that it has to satisfy condition (i) in the definition of cross-negativity.

Condition (i): Our proof is indirect. Suppose there were $j, i \in \{1, 2, ..., n\}$ 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, using (*) in the proof of Proposition 2:

$$\begin{aligned} f(s_i) &= \sigma_i \sum_{\substack{j=1 \\ j \neq i}}^{n} (\sigma_j B_{ji} (\pi_i - \pi_j)) \\ &= \sigma_i \left(B_{ji}\delta + \sum_{\substack{k=1 \\ k \neq ij}}^{n} (\sigma_k B_{ki} \varepsilon) \right) \end{aligned}$$

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

Condition (ii). The proof is indirect. Suppose there were some subset C of S such that $C \neq \emptyset$ and $S \backslash C \neq \emptyset$ and such that $B_{ji} = 0$ for all $s_i \in C$ and $s_j \in S \backslash C$. Consider an environment E such that all strategies in C yield payoff x with certainty, and all strategies in $S \backslash C$ 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.

Propositions 3 and 4 have the following implication.

Corollary 1 *Every monotone learning rule is absolutely expedient.*

Recalling Remark 3 we can also deduce from Propositions 2, 3 and 4:

Corollary 2 *If $n = 2$ then a learning rule is absolutely expedient if and only if it is monotone.*

An example of a learning rule which belongs to the class which we have discussed in this section is Cross' rule which we presented in Section 4. To indicate that the class of monotone rules is quite large, and encompasses many rules which are intuitively quite different from Cross' rule we now give a further example of a monotone learning rule. The example will describe learning based on an aspiration level. The example contrasts with the learning rule studied in Börgers and Sarin (2000) which is also based on an aspiration level, but which fails to be absolutely expedient or even unbiased. The rule in Börgers and Sarin (2000) is quite simple and intuitive. Example 2, by contrast, involves relatively complicated formulas.

Example 2 *Let any α with $0 \leq \alpha \leq 1$ be given. Using the notation of Proposition 1 we can then define a monotone learning rule by setting for all $i, j \in \{1, 2, ..., n\}$ with $i \neq j$:*

$$\begin{aligned}
A_{ii} &= -\sigma_i \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[(\sigma_j)^2 (1 - \sigma_j) \right] \alpha \\
B_{ii} &= +\sigma_i \sum_{\substack{j=1 \\ j \neq i}}^{n} \left[(\sigma_j)^2 (1 - \sigma_j) \right] \\
A_{ji} &= -(1 - \sigma_j)(1 - \sigma_i)\sigma_j \sigma_i \alpha \\
B_{ji} &= +(1 - \sigma_j)(1 - \sigma_i)\sigma_j \sigma_i
\end{aligned}$$

Note that for all i, j we have:

$$A_{ji} = -B_{ji}\alpha$$

which implies that

$$\begin{aligned}
L(s_i, x)(s_i) &= \sigma_i + (1 - \sigma_i) B_{ii} (x - \alpha) \\
L(s_j, x)(s_i) &= \sigma(s_i) - \sigma(s_i) B_{ji} (x - \alpha) \text{ for all } j \neq i.
\end{aligned}$$

Therefore, according to this learning rule, if strategy s_i was played in iteration n, the decision maker increases (resp. decreases) in period $n+1$ the probability assigned to s_i if the payoff x which the decision maker received in iteration n was above (resp. below) α. Intuitively, α thus plays the role of an "aspiration level." If the probability assigned to s_i is increased (resp. decreased), the probability of all other strategies is decreased (resp. increased).

It is obvious that the learning rule in Example 2 is cross-negative. Therefore we only show that it is unbiased, i.e. satisfies conditions (3) and (4) in Proposition 1. We write the proof only for condition (4) because it is completely analogous for condition (3). We need to show that for every $i = 1, 2, ..., n$:

$$\begin{aligned} B_{ii} &= \sum_{j=1}^{n} \sigma_j B_{ji} \Leftrightarrow \\ (1-\sigma_i)B_{ii} &= \sum_{\substack{j=1 \\ j\neq i}}^{n} \sigma_j B_{ji} \\ (1-\sigma_i)\sigma_i \sum_{\substack{j=1 \\ j\neq i}}^{n} \left[(\sigma_j)^2(1-\sigma_j)\right] &= \sum_{\substack{j=1 \\ j\neq i}}^{n} (\sigma_j(1-\sigma_j)(1-\sigma_i)\sigma_j\sigma_i) \end{aligned}$$

which is obviously true.

7 Cross-Positive Rules

We call an unbiased learning rule "cross-positive" if it is not cross-negative. Are there unbiased cross-positive learning rules which are absolutely expedient? In this section we show that the answer is yes. We show this by giving an example of such a rule. We also give an example which proves that not all own-positive and cross-positive rules are absolutely expedient. We do not have a general charcterisation of those own-positive rules which are absolutely expedient, and our examples suggest that such a characterisation may have to be quite complicated.

Intuitively, it is surprising that there are examples of cross-positive rules which are absolutely expedient. Cross-positivity means intuitively, as we suggested earlier, that some notion of similarity of strategies is built into the learning rule. On the other hand absolute expediency requires that the learning rule is able to improve performance in any environment, including environments in which the strategies which the learning rule treats as similar are, in reality, not similar at all.

The results of the previous sections show that cross-positive rules which are absolutely expedient can only exist in the case $n \geq 3$. Therefore we restrict attention to this case.

Example 3 *Suppose $n = 3$ and the current state is: $\sigma_1 = \sigma_2 = \sigma_3 = \frac{1}{3}$. (We show in the appendix how to generalize this example to an arbitrary number of strategies and arbitrary initial state). Define:*

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

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

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

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

Thus, if strategy s_1 is played and strategy s_2 is updated, or vice versa, then this rule adjusts the behaviour as if a positive payoff had not only been received for strategy s_1 but also for strategy s_2, or vice versa. In this sense the rule treats strategies s_1 and s_2 as similar. We now show that nonetheless the rule is absolutely expedient in *all* environments, and hence even in environments in which s_1 and s_2 are, in fact, very dissimilar.

Using the formula in Remark 2 the expected change in expected payoffs under this rule can be calculated as:

$$\begin{aligned} g &= \frac{1}{3} \cdot \frac{1}{10} \cdot (\pi_1^2 + \pi_2^2 + 3\pi_3^2) \\ &\quad -\frac{1}{9} \cdot \frac{1}{10}(\pi_1^2 - \pi_1\pi_2 + 3\pi_1\pi_3) \\ &\quad -\frac{1}{9} \cdot \frac{1}{10}(-\pi_2\pi_1 - \pi_2^2 + 3\pi_2\pi_3) \\ &\quad -\frac{1}{9} \cdot \frac{1}{10}(3\pi_3\pi_1 + 3\pi_3\pi_2 + 3\pi_3^2) \\ &= \frac{1}{9} \cdot \frac{1}{10} \cdot 2(\pi_1^2 + \pi_2^2 + 3\pi_3^2 + \pi_1\pi_2 - 3\pi_1\pi_3 - 3\pi_2\pi_3) \end{aligned}$$

We are going to show that the term in brackets is strictly positive except when all expected payoffs are identical: $\pi_1 = \pi_2 = \pi_3$, in which case it is obviously zero. We can write the term in brackets it as:

$$\pi_1^2 + \pi_2^2 + \pi_1\pi_2 - 3\pi_3(\pi_1 + \pi_2 - \pi_3)$$

The term which gets subtracted here is largest if:

$$\pi_3 = \frac{\pi_1 + \pi_2}{2}$$

If we substitute this for π_3 we thus get a lower boundary:

$$\begin{aligned} & \pi_1^2 + \pi_2^2 + \pi_1\pi_2 - 3\frac{\pi_1 + \pi_2}{2}(\pi_1 + \pi_2 - \frac{\pi_1 + \pi_2}{2}) \\ = \; & \frac{1}{4}(\pi_1 - \pi_2)^2 \end{aligned}$$

Clearly, this will always be non-negative. Moreover, it will be zero only if $\pi_1 = \pi_2$. Now recall that this is a lower bound, and that for fixed π_1 and π_2 this lower bound will be attained only if $\pi_3 = \frac{\pi_1+\pi_2}{2}$. Thus, we have found that the expression which we are investigating is strictly positive except if $\pi_1 = \pi_2$ and $\pi_3 = \frac{\pi_1+\pi_2}{2}$. In other words, it will be strictly positive except if all expected payoffs are identical. This is what we wanted to prove.

We now give an example which proves that not all unbiased and own-positive rules are absolutely expedient.

Example 4 *Suppose $n \geq 3$, 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 leads to the following formulas:*

$$A_{ji} = 0 \textit{ for all } j, i \in \{1, 2, ..., n\}.$$

$$B_{ii} = \frac{\sigma_i(1 - \sigma_1 - \sigma_2)}{(\sigma_1 + \sigma_2)(1 - \sigma_i)} \textit{ for } i = 1, 2.$$

$$B_{12} = B_{21} = -\frac{1 - \sigma_1 - \sigma_2}{\sigma_1 + \sigma_2}.$$

$$B_{ji} = 1 \textit{ if } j \notin \{1, 2\} \textit{ or } i \notin \{1, 2\}.$$

The expected movement of payoffs under this rule is the same as in the modified environment in which strategies s_1 and s_2 are replaced by a new strategy s which has expected payoff $\sigma_1\pi_1 + \sigma_2\pi_2$. The same is true for the probabilities of all strategies $s \neq s_1, s_2$. The expected movement of strategies s_1 and s_2 is the same as the expected movement of the new strategy which replaces s_1 and s_2 multiplied by $\frac{\sigma_1}{\sigma_1+\sigma_2}$ resp. $\frac{\sigma_2}{\sigma_1+\sigma_2}$.

It is then clear that the fact that Cross' rule is unbiased implies that this new rule is unbiased, too. However, it is not absolutely expedient. Consider an environment in which the expected payoff of strategies s_1 and s_2 taken together equals the expected payoff of all other strategies: $\sigma_1\pi_1 + \sigma_2\pi_2 = \pi_i$ for all $i \neq 1, 2$, but in which $\pi_1 > \pi_2$. Then in expected terms no strategy's probability will change, and therefore also the expected utility will stay the same. However, absolute expediency requires it to increase.

8 A Global Model of Learning

The model that we have considered so far was a "local" model in two senses. Firstly, we took the decision maker's initial behaviour as exogenous, and didn't consider learning rules which work for a variety of different initial behaviours. Secondly, we only looked at two periods: "today" and "tomorrow", and didn't trace the decision maker's behaviour of longer time spans. In this section we shall discuss how both restrictions can be relaxed. The motivation for relaxing the first restriction is that decision maker might not know how to choose his initial action distribution, and might therefore desire a learning rule which improves his performance independent of where he starts. The motivation for relaxing the second condition is that we might be interested in understanding for particular environments where the learning process leads in the long run. We relax both restrictions simultaneously by introducing a model of learning which is well-defined for any initial position, and which the decision maker can thus apply repeatedly. We then keep track of his behaviour over longer time spans. The learning model of this section is, in this sense, "global".

The decision maker's learning process now has a state space V which is some arbitrary subset of a finite-dimensional Euclidean space. The state determines behaviour via a function $b : V \to \Delta(S)$. If the decision maker is at some particular point in time in state $v \in V$ then he chooses his strategy according to the distribution $b(v)$. Separating state space of the learning rule and behaviour in this way allows the decision maker to have a larger memory. For example, it allows us to consider the case that the decision maker keeps track of how many times he has already played the same situation.

How the decision maker's state is adjusted in response to his experiences is described by a learning rule which we now denote by Λ.

Definition 8 *A* global learning rule *is a function* $\Lambda : V \times S \times [0,1] \to V$*.*

If the decision maker enters a period in state v, chooses in that period the strategy s_i (distributed according to $b(v)$), and receives payoff x (distributed according to μ_i), then in the next period his state is $\Lambda(v, s_i, x)$.

We want to make our earlier analysis applicable to global learning rules. In analogy to Assumption 2 we therefore introduce the following assumption.

Assumption 3: The behaviour function b is continuous in v, and the global learning function Λ is continuous in (v, x)..

Consider a global learning rule Λ, and focus on some state v. The global learning rule then implies a "local" learning rule of the type we discussed in earlier sections. We shall call it the local learning rule L_v at v, and we define it by: $L_v(s_i, x) = b(\Lambda(v, s_i, x))$ for every $x \in [0,1]$. The local learning rule L_ν is, in a sense, the reduced form of Λ at the state ν.

We can now extend the definitions of absolute expediency and monotonicity to global learning rules.

Definition 9 *A global learning rule Λ is absolutely expedient (resp. monotone) if for every $v \in V$ such that $b(v)$ assigns positive probability to all $s_i \in S$ we have that the local learning rule L_v at v is absolutely expedient (resp. monotone).*

We can now apply the results of the previous sections to global learning rules which are absolutely expedient resp. monotone. At every state at which $b(\nu)$ assigns positive probability to all $s_i \in S$ the local learning rule will satisfy all the result of the earlier sections.

In the remainder of this section we discuss examples of global learning rules which are absolutely expedient or monotone. The simplest class of examples can be obtained by setting the state space equal to the interior of the strategy simplex, $V = \Delta(S)$, letting b be the identity, and defining the global learning rule by applying the local learning rules which we considered in our earlier examples at all possible states. In this way we can construct absolutely expedient global learning rules based on Examples 1, 2 and 3, and the first two of these examples will also be monotone.

For the case of Cross' rule, i.e. Example 1, the corresponding global learning rule was studied in Börgers and Sarin (1997). There it was shown that Cross' rule with very small step size stays with high probability close to trajectories of the replicator dynamics. This was shown for game environments and for finite time horizons, but the result can be extended to decisions with i.i.d payoffs and to infinite time horizons. Hence the rule picks in such environments in the long run with high probability expected payoff maximising actions. This extension, and generalizations of these results to all monotone learning rules, were proved in earlier versions of this paper. Interestingly, mathematical difficulties prevented us from extending these results to all absolutely expedient rules.

A second class of examples can be obtained by extending the state space to include the number of the current period: $V = \Delta(S) \times \mathbb{N}$ with typical element (σ, t), letting b again be the identity, and defining the updating rule by applying the learning rules of Examples 1, 2 or 3, except that the step size in round $t \in \mathbb{N}$ is equal to $\frac{1}{t}$. That is, the vector of movement is the same as in these examples, but all of its components are multiplied by $\frac{1}{t}$. The idea is that in later iterations the decision maker adjusts his behaviour by less because he is aware that he has accumulated experience. Multiplication of the vector of movement by a positive number which does not depend on the strategy played or the payoffs received does not affect the properties of monotonicity or absolute expediency. In i.i.d environments rules of the type described in this paragraph can be expected to pick with probability 1 an expected payoff maximising action. We conjecture that this can be shown using similar methods as in Rustichini (1999).

Our final example, a learning rule due to Erev and Roth (1995, 1998), has a significantly richer state space than the examples which we have discussed so far. The example illustrates how the framework of this section encompasses a large variety of learning rules.

Example 5 *Erev and Roth (1995, 1998). The state space is $V = \mathbb{R}^n_{>0}$. If $v = (v_i)_{i=1,2,\ldots,n} \in V$ then v_i is interpreted as the decision maker's "propensity" to play s_i.*

The decision maker chooses strategies with probabilities which are propor-

tional to the propensities:

$$b(v_i) = \left(\frac{v_i}{\Sigma_{j=1}^n v_j}\right)_{i=1,2,...,n}$$

The global learning rule Λ *is defined by:*

$$\begin{aligned} \Lambda(v, x, s_i) &= (\hat{v}_i)_{j=1,2,...,n} \qquad \textit{where} \\ \hat{v}_i &= v_i + x \qquad \textit{and} \\ \hat{v}_j &= v_j \qquad \textit{if } j \neq i. \end{aligned}$$

Thus only the propensity of the strategy actually played gets updated. The new propensity to play that strategy equals the old propensity plus the received payoff.

It is interesting to investigate the relation between Erev and Roth's rule and Cross' learning rule. This connection becomes clear if one calculates the local learning rule L_v induced by Erev and Roth's learning rule for any given state ν. It is:

$$\begin{aligned} L_v\left(s_i, x\right)\left(s_i\right) &= \sigma_i + \frac{1}{\Sigma_{k=1}^n v_k + x}\left(1 - \sigma_i\right)x \\ L_v\left(s_j, x\right)\left(s_i\right) &= \sigma_i - \frac{1}{\Sigma_{k=1}^n v_k + x}\sigma_i x \qquad \text{for all } j \neq i. \end{aligned}$$

This induced local learning rule is Cross' rule, except that the direction of movement is multiplied by $\frac{1}{\Sigma_{k=1}^n v_k + x}$. This factor is stochastic. However, with probability 1 it converges to zero as long as there is a positive probability of strictly positive payoffs. Erev and Roth's learning rule is thus Cross' learning rule with a stochastically declining step size.

Note that the learning rule induced by the Erev-Roth global learning rule is not linear in payoffs. Therefore we can deduce, using Proposition 1, that it is not unbiased, and Lemma 1 then implies that it is neither absolutely expedient nor monotone. We now give an example to demonstrate this.[6] Let $S = \{s_1, s_2\}$. Suppose that the environment is such that s_1 yields 0.1 and 0.9 with probability $\frac{1}{2}$ each, and s' yields 0.4 for sure. Consider the state v such that: $v_1 = 0.02$ and $v_2 = 0.05$. Thus the probability of s_1 today is $\frac{2}{7}$. The

[6] A similar example appeared first in Sarin (1995).

expected probability of s_1 tomorrow is: $\frac{2}{7}(\frac{1}{2} \cdot \frac{0.12}{0.19} + \frac{1}{2}\frac{0.92}{0.99}) + \frac{5}{7} \cdot \frac{0.02}{0.49} \approx 0.252 < \frac{2}{7} \approx 0.286$. This implies that in this environment, for this particular state v, the implied local learning rule at state v is not monotone and not absolutely expedient.

9 Related Literature

Papers which characterise all learning rules which possess certain properties exist both in the economics literature and in the literature on machine learning. In this section we first review papers from the economics literature, and then papers from the literature on machine learning.

Schlag (1994) and Sarin (1995) are two early unpublished works on absolute expediency of global learning rules.[7] Both papers study global learning rules with state space $\Delta(S)$. For the case of two actions Schlag provides a characterisation of Cross' learning rule. He assumes that the rule is affine in payoffs where the coefficients of the transformation of payoffs are only allowed to depend on the current mixed strategy, and not on the pure strategies which has been played and which is being updated. Schlag shows that the Cross rule is the absolutely expedient, and that it maximises expected payoff gain among all absolutely expedient rules. Sarin (1995) also deals with the case of two strategies. He does not assume that the learning rule is affine in payoffs. He characterises the Cross' rule and some multiples of it using further axioms. Notice the difference in spirit between these two papers and ours. The aim of our paper is *not* to axiomatize Cross' or any other learning rule. Our aim is to describe a large and flexible class of learning rules all of which have certain properties in common.

The last comment in the previous paragraph describes also what distinguishes our paper from a recent new paper by Schlag (2001). Schlag is there again concerned with the case of two actions. He focuses on global learning rules with finite or countably infinite state space. He assumes that the global learning rule Λ is affine in payoffs where the coefficients of the payoff transformation are allowed to depend on the current state. The main innovation of Schlag (2001) is to introduce a farsighted player whose focus is on long run discounted payoffs. We justified the myopia postulated in our paper by suggesting that the decision maker might not know whether the environment

[7] Both papers work in game rather than single person decision settings. However, it is easy to adapt their results to the case of single person decisions.

is stationary, and might therefore not be able to evaluate long run payoffs. Schlag's approach seems more natural when the decision maker is certain to live in a stationary environment. Schlag's focus is on finding axioms which select unique rules.[8]

Easley and Rustichini (1999) consider an individual who faces a repeated decision problem under risk, and who observes in every round not only the payoff of the strategy which he has chosen but also the payoff which other strategies would have received had they been chosen.[9] Hence the agent receives more feedback information than in our model. The state space V of their learning rule is the set of preferences over mixed strategies which satisfy the von Neumann Morgenstern axioms. The elements of the state space can be viewed as "weights" for the decision maker's strategies. Easley and Rustichini study a variety of axioms for an individual's global learning rule and show that these axioms imply that the individual's asymptotically chooses the expected payoff maximizing action, and that weights are updated according to a transformation of replicator dynamics. Their axioms are unrelated to the absolute expediency and monotonicity conditions which we discuss here.

A large set of papers related to ours can be found in the literature on machine learning, and specifically in that part which is concerned with the learning behaviour of stochastic automata. The concept of a stochastic automaton is similar to our concept of a global learning rule Λ with state space $V = \Delta(s)$. A useful overview of the literature on stochastic automata and learning has been provided by Narendra and Thathachar (1989).[10] In this literature, absolute expediency was originally defined by Lakshmivarahan and Thathachar (1973). Monotonicity is studied by Toyama and Kimura (1977) who refer to this property as *absolute adaptability.*

The most general characterisation 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. By contrast, we do not assume this but derive it (Proposition 1). The form of linearity which Narendra and Thathachar assume is more restrictive than the form of linearity which we derive in that Narendra and Thathachar allow the coefficients

[8] At the time of writing, Schlag's paper is still incomplete.

[9] Formally, that is expressed in their model by the condition that the individual observes in each round the "state of the world".

[10] For our setup their Chapter 6 is the relevant chapter.

in the affine transformation of payoffs to depend only on the current state of the learning rule and on the action played whereas our results show that one can allow in addition dependence on the strategy whose probability is updated, and still maintain absolute expediency. Narendra and Thathachar note that in their framework absolute expediency and monotonicity are equivalent (Comment 6.2). This follows also from our results because Narendra and Thathachar's assumptions about the form of the learning rule imply that every unbiased rule which is of this form must be cross-negative. Thus, Propositions 3 and 4 show the equivalence of absolute expediency and monotonicity in Narendra and Thathachar's framework.

Toyama and Kimura (1977) characterise 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 which has been played nor on the action which is updated. Thus, again, in their framework monotonicity and absolute expediency are actually equivalent.

10 Conclusion

Within a very large class of learning rule we have provided results about all learning rules which have certain desirable properties: absolute expediency and monotonicity. Our analysis leads to interesting lines of further theoretical and experimental research. Further theoretical research could investigate the long run implications of learning behaviour that adopts one of the learning rules studied in this paper. We indicated some conjectures in Section 8. Experimental research could investigate whether subjects' choice probabilities are affine functions of payoff experiences, as required in Proposition 1, and whether the coefficients of these affine functions satisfy the constraints which we discussed in Sections 5, 6 and 7.

11 Appendix: Generalizing Example 3

In this appendix we generalize Example 3 to the case of arbitrarily many strategies and arbitrary initial state. Suppose that $n \geq 3$. Suppose that the current state is some probability distribution σ on S such that $\sigma_i > 0$ for all $i = 1, 2, ..., n$. We wish to define a learning rule which treats two designated elements of S as "similar". We shall denote these two elements by s_1 and s_2. The learning rule will involve two constants k and ℓ. We assume that they are both elements of the open interval $(0,1)$. We shall also assume that they are both "sufficiently small". What we mean by this will later be made more precise.

The idea of the learning rule is that if s_1 is played and payoff x is received, then this payoff is attributed not just to s_1 but also to s_2. Both strategies' probabilities are then updated as in Cross' rule. An important point is that the probability of s_1 is updated as if the payoff had been kx, whereas the probability of s_2 is updated as if the payoff had been ℓx. The proof below will show that the learning rule is absolutely expedient if ℓ is sufficiently small *relative to k*. The interpretation of this condition is that, although payoffs received when one strategy is played can be attributed to the other strategy as well, this effect must be weak in relative terms.

We begin by giving the formal definition of our learning rule. The formulas are complicated because we need to fit them into the framework of Proposition 1. The formulas reflect the ideas described in the previous paragraph and also the assumption that all strategies other than s_1 and s_2 are treated symmetrically.

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

$$B_{11} = B_{22} = k$$

$$B_{12} = -\ell \frac{1-\sigma_2}{\sigma_2}$$

$$B_{21} = -\ell \frac{1-\sigma_1}{\sigma_1}$$

$$\begin{aligned} B_{1i} &= \frac{k(1-\sigma_1)+\ell(1-\sigma_2)}{1-\sigma_1-\sigma_2} \\ \text{for all } i &= 1, 2, ..., n \text{ with } i \neq 1, 2 \end{aligned}$$

$$\begin{aligned} B_{2i} &= \frac{k(1-\sigma_2)+\ell(1-\sigma_1)}{1-\sigma_1-\sigma_2} \\ \text{for all } i &= 1, 2, ..., n \text{ with } i \neq 1, 2. \end{aligned}$$

$$\begin{aligned} B_{i1} &= \frac{k(1-\sigma_1)+\ell\frac{\sigma_2}{\sigma_1}(1-\sigma_1)}{1-\sigma_1-\sigma_2} \\ \text{for all } i &= 1, 2, ..., n \text{ with } i \neq 1, 2. \end{aligned}$$

$$\begin{aligned} B_{i2} &= \frac{k(1-\sigma_2)+\ell\frac{\sigma_1}{\sigma_2}(1-\sigma_2)}{1-\sigma_1-\sigma_2} \\ \text{for all } i &= 1,2,...,n \text{ with } i \neq 1,2. \end{aligned}$$

$$\begin{aligned} B_{ij} &= \frac{k\sigma_1(1-\sigma_1)+k\sigma_2(1-\sigma_2)+\ell\sigma_1(1-\sigma_2)+\ell\sigma_2(1-\sigma_1)}{(\sigma_1+\sigma_2)(1-\sigma_1-\sigma_2)} \\ \text{for all } i,j &\in \{1,2,...,n\}\backslash\{1,2\}. \end{aligned}$$

We can now say more precisely what we mean by the assumption that k and ℓ are "sufficiently small". We mean by this that all positive entries of the above matrix need to be less than one. It is obvious that for fixed current state σ it is possible to choose k and ℓ so that this is true. On the other hand one can also easily see that it is not possible to choose the same k and ℓ for all current states σ.. For any given and fixed k and ℓ there will always be some states σ for which the condition is violated. Thus, the constants k and ℓ in the above formula will change with the state σ, and are "constant" only in as far as they don't depend on the payoff x.

Our proof of absolute expediency of this learning rule will proceed in two steps. First, we prove it for the case that $n = 3$. Then we prove it for the case $n > 3$. A key idea in the proof is to reduce the case $n > 3$ to the case $n = 3$.

Proof of Absolute Expediency in the Case $n = 3$: Using the formula in Remark 2 we can calculate g as follows, where we first collect all terms which do not involve π_3, and then all terms which do involve π_3.

$$\begin{aligned} g = {} & \sigma_1(1-\sigma_1)k\pi_1^2+\ell\sigma_2(1-\sigma_1)\pi_1\pi_2 \\ & +\ell\sigma_1(1-\sigma_2)\pi_1\pi_2+\sigma_2(1-\sigma_2)k\pi_2^2 \\ & -\{k\sigma_1(1-\sigma_1)\pi_1+\ell\sigma_1(1-\sigma_2)\pi_1 \\ & +k\sigma_2(1-\sigma_2)\pi_2+\ell\sigma_2(1-\sigma_1)\pi_2 \\ & +k\sigma_1(1-\sigma_1)\pi_1+\ell\sigma_2(1-\sigma_1)\pi_1 \\ & +k\sigma_2(1-\sigma_2)\pi_2+\ell\sigma_1(1-\sigma_2)\pi_2\}\pi_3 \\ & +(k\sigma_1(1-\sigma_1)+k\sigma_2(1-\sigma_2)+\ell\sigma_1(1-\sigma_2)+\ell\sigma_2(1-\sigma_1))\pi_3^2 \end{aligned}$$

To simplify this formula we use the following notation:

$$\begin{aligned} \alpha &\equiv k\sigma_1(1-\sigma_1) \\ \beta &\equiv k\sigma_2(1-\sigma_2) \\ \gamma &\equiv \ell(\sigma_1(1-\sigma_2)+\sigma_2(1-\sigma_1)) \end{aligned}$$

Then we can write g as follows:

$$\begin{aligned} g \;=\; & \alpha\pi_1^2+\beta\pi_2^2+\gamma\pi_1\pi_2 \\ & -\{2\alpha\pi_1+2\beta\pi_2+\gamma(\pi_1+\pi_2)\}\pi_3 \\ & +(\alpha+\beta+\gamma)\pi_3^2 \end{aligned}$$

Now observe that for fixed π_1 and π_2 the expected change in expected payoffs is a quadratic function of π_3 where the coefficient in front of the square is positive. Thus, we can use calculus methods to find that value of π_3 for which the increase in g is smallest. It is:

$$\pi_3=\frac{2\alpha\pi_1+2\beta\pi_2+\gamma(\pi_1+\pi_2)}{2(\alpha+\beta+\gamma)}$$

We see that this is a weighted average of π_1 and π_2. Now substitute this value of π_3 in the equation above. We obtain a lower boundary for g:

$$\begin{aligned} g \;\geq\; & \alpha\pi_1^2+\beta\pi_2^2+\gamma\pi_1\pi_2 \\ & -\frac{(2\alpha\pi_1+2\beta\pi_2+\gamma(\pi_1+\pi_2))^2}{4(\alpha+\beta+\gamma)} \end{aligned}$$

We multiply the right hand side of this equation by $4(\alpha+\beta+\gamma)$, which is clearly sign-preserving, and call the expression which we get $\widehat{g}$. We obtain:

$$\widehat{g}=4(\alpha+\beta+\gamma)(\alpha\pi_1^2+\beta\pi_2^2+\gamma\pi_1\pi_2)-(2\alpha\pi_1+2\beta\pi_2+\gamma(\pi_1+\pi_2))^2$$

This simplifies to:

$$\widehat{g}=(4\alpha\beta-\gamma^2)(\pi_1-\pi_2)^2$$

Suppose we can show that the first bracket in this expression is strictly positive. Then obviously whenever $\pi_1\neq\pi_2$ the whole expression is strictly positive and we are done. Now consider the case that $\pi_1=\pi_2$ and that $\pi_3\neq\pi_1$. Recall that $\widehat{g}$ is only the lower boundary for g. It has been calculated for the worst case value of π_3. The worst case is one in which π_3 is a weighted average of π_1 and π_2. Thus, if $\pi_1=\pi_2$ then the worst case is that $\pi_3=\pi_1=\pi_2$. In that case obviously the above expression is zero. But because this is the unique minimizer of expected change in expected payoff in all other cases g will be strictly positive.

Thus it is sufficient to show that the first bracket in the above expression is strictly positive:

$$4\alpha\beta>\gamma^2$$

Now we recall the definition of α, β, and γ, and substitute back:

$$4k^2\sigma_1\sigma_2(1-\sigma_1)(1-\sigma_2) > \ell^2(\sigma_1(1-\sigma_2)+\sigma_2(1-\sigma_1))^2$$

This will be true if ℓ is sufficiently small relative to k. Note that the appropriate choice of ℓ depends on the current state, but not on the payoffs.

Proof of Absolute Expediency in the Case $n > 3$: To simplify formulas, be shall denote B_{ij} where $i, j \in \{1, 2, ..., n\}\backslash\{1, 2\}$ simply by "B". That is:

$$B \equiv \frac{k\sigma_1(1-\sigma_1)+k\sigma_2(1-\sigma_2)+\ell\sigma_1(1-\sigma_2)+\ell\sigma_2(1-\sigma_1)}{(\sigma_1+\sigma_2)(1-\sigma_1-\sigma_2)}.$$

To check the absolute expediency of this learning rule we now calculate the expected change in expected payoffs using the formula for g in Remark 2:

$$\begin{aligned} g &= \sigma_1(1-\sigma_1)k\pi_1^2+\ell\sigma_2(1-\sigma_1)\pi_1\pi_2 \\ &+\ell\sigma_1(1-\sigma_2)\pi_1\pi_2+\sigma_2(1-\sigma_2)k\pi_2^2 \\ &-\sum_{\substack{i=1\\ i\neq 1,2}}^{n}\sigma_1\sigma_i\frac{k(1-\sigma_1)+\ell(1-\sigma_2)}{1-\sigma_1-\sigma_2}\pi_1\pi_i \\ &-\sum_{\substack{i=1\\ i\neq 1,2}}^{n}\sigma_2\sigma_i\frac{k(1-\sigma_2)+\ell(1-\sigma_1)}{1-\sigma_1-\sigma_2}\pi_2\pi_i \\ &-\sum_{\substack{i=1\\ i\neq 1,2}}^{n}\sigma_i\frac{\sigma_1k(1-\sigma_1)+\ell\sigma_2(1-\sigma_1)}{1-\sigma_1-\sigma_2}\pi_1\pi_i \\ &-\sum_{\substack{j=1\\ j\neq 1,2}}^{n}\sigma_j\frac{\sigma_2k(1-\sigma_2)+\ell\sigma_1(1-\sigma_2)}{1-\sigma_1-\sigma_2}\pi_2\pi_j \\ &+\sum_{\substack{i=1\\ i\neq 1,2}}^{n}\sigma_i(1-\sigma_i)B(\pi_i)^2-\sum_{\substack{i,j\in\{1,2,\ldots,n\}\backslash\{1,2\}\\ i\neq j}}\sigma_i\sigma_jB\pi_i\pi_j. \end{aligned}$$

The proof strategy is to reduce the case $n > 3$ to the case $n = 3$ with which we dealt in the first step. We shall re-write the above sum so that one part of sum equals expected payoff change in the case that of only three strategies

s_1, s_2 and s_3, where the third action's expected payoff is the expected payoff of the decision maker in his current state conditional on him not playing s_1 or s_2. It is useful to introduce notation for the relevant conditional expected payoff:

$$\widehat{\pi} \equiv \sum_{\substack{i=1\\i\neq 1,2}}^{n} \left(\frac{\sigma_i}{1-\sigma_1-\sigma_2}\pi_i \right).$$

The second term in our decomposition of g will correspond to the interaction among strategies other than s_1 and s_2. Because the learning dynamics is the Cross dynamics if only these strategies are considered, the expected change in expected payoffs will take the form of a variance. It will therefore be useful to have notation available for the expected value of the squared expected profits, conditional on the decision maker not playing s_1 or s_2. Define:

$$\widehat{\widehat{\pi}} \equiv \sum_{\substack{i=1\\i\neq 1,2}}^{n} \left(\frac{\sigma_i}{1-\sigma_1-\sigma_2}\pi_i^2 \right).$$

Thus, the variance of expected payoffs, conditional on the decision maker not playing s_1 or s_2, is: $\widehat{\widehat{\pi}} - (\widehat{\pi})^2$, and this will be non-negative, and positive whenever the strategies other than s_1 and s_2 don't all have the same expected payoff.

To begin our calculations we note that we can re-write the above expression as:

$$\begin{gathered}
g = \sigma_1(1-\sigma_1)k\pi_1^2 + \ell\sigma_2(1-\sigma_1)\pi_1\pi_2 \\
+\ell\sigma_1(1-\sigma_2)\pi_1\pi_2 + \sigma_2(1-\sigma_2)k\pi_2^2 \\
-\sigma_1(k(1-\sigma_1)+\ell(1-\sigma_2))\pi_1\widehat{\pi} \\
-\sigma_2(k(1-\sigma_2)+\ell(1-\sigma_1))\pi_2\widehat{\pi} \\
-(\sigma_1 k(1-\sigma_1)+\ell\sigma_2(1-\sigma_1))\pi_1\widehat{\pi} \\
-(\sigma_2 k(1-\sigma_2)+\ell\sigma_1(1-\sigma_2))\pi_2\widehat{\pi} \\
+\sum_{\substack{i=1\\i\neq 1,2}}^{n} \sigma_i(1-\sigma_i)B\pi_i^2 - \sum_{\substack{i,j\in\{1,2,\ldots,n\}\setminus\{1,2\}\\i\neq j}} \sigma_i\sigma_j B\pi_i\pi_j.
\end{gathered}$$

We next analyse the last two terms of this expression:

$$\sum_{\substack{i=1\\ i\neq 1,2}}^{n} \sigma_i(1-\sigma_i)B\pi_i^2 - \sum_{\substack{i,j\in\{1,2,\dots,n\}\setminus\{1,2\}\\ i\neq j}} \sigma_i\sigma_j B\pi_i\pi_j$$

$$= \sum_{\substack{i=1\\ i\neq 1,2}}^{n} \sigma_i B\pi_i^2 - \sum_{i,j\in\{1,2,\dots,n\}\setminus\{1,2\}} \sigma_i\sigma_j B\pi_i\pi_j$$

$$\begin{aligned}
&= (1-\sigma_1-\sigma_2)B\widehat{\widehat{\pi}} - (1-\sigma_1-\sigma_2)^2 B\widehat{\pi}^2 \\
&= (1-\sigma_1-\sigma_2)B\widehat{\widehat{\pi}} - (1-\sigma_1-\sigma_2)B\widehat{\pi}^2 + (\sigma_1+\sigma_2)(1-\sigma_1-\sigma_2)B\widehat{\pi}^2
\end{aligned}$$

Substituting for B in the last expression we get:

$$\begin{aligned}
= \;& (1-\sigma_1-\sigma_2)B(\widehat{\widehat{\pi}} - \widehat{\pi}^2) \\
& +(k\sigma_1(1-\sigma_1) + k\sigma_2(1-\sigma_2) + \ell\sigma_2(1-\sigma_1) + \ell\sigma_1(1-\sigma_2))\widehat{\pi}^2
\end{aligned}$$

Now we are going to substitute this back into our expression for g, whereby we write the second of the above two lines before we write the first line:

$$\begin{gathered}
g = \sigma_1(1-\sigma_1)k\pi_1^2 + \ell\sigma_2(1-\sigma_1)\pi_1\pi_2 \\
+\ell\sigma_1(1-\sigma_2)\pi_1\pi_2 + \sigma_2(1-\sigma_2)k\pi_2^2 \\
-\sigma_1(k(1-\sigma_1) + \ell(1-\sigma_2))\pi_1\widehat{\pi} \\
-\sigma_2(k(1-\sigma_2) + \ell(1-\sigma_1))\pi_2\widehat{\pi} \\
-(\sigma_1 k(1-\sigma_1) + \ell\sigma_2(1-\sigma_1))\pi_1\widehat{\pi} \\
-(\sigma_2 k(1-\sigma_2) + \ell\sigma_1(1-\sigma_2))\pi_2\widehat{\pi} \\
+(k\sigma_1(1-\sigma_1) + k\sigma_2(1-\sigma_2) + \ell\sigma_2(1-\sigma_1) + \ell\sigma_1(1-\sigma_2))\widehat{\pi}^2 \\
+(1-\sigma_1-\sigma_2)B(\widehat{\widehat{\pi}} - \widehat{\pi}^2)
\end{gathered}$$

Now observe that the first seven lines of this expression are exactly the expression for expected movement of expected payoffs if there is a third strategy with expected payoff $\widehat{\pi}$. Thus we know from Step 1that this expression is non-negative. The last line is the product of three non-negative factors where the third factor is the variance of expected payoffs conditional on not playing s_1 and s_2.

We still have to make sure that the above expression is strictly positive as long as not all strategies have the same expected payoff. From the case $n = 3$ we know that whenever $\pi_1 \neq \pi_2$ or whenever $\pi_1 = \pi_2 \neq \widehat{\pi}$ the sum of the first nine lines of the above expression will be strictly positive. Consider the case that $\pi_1 = \pi_2 = \widehat{\pi}$, but that not all strategies other than s_1 and s_2 have the same payoff. Then the last line in the above expression will be strictly positive because the variance will be strictly positive. This concludes the proof.

References

[1] Börgers, Tilman, and Rajiv Sarin, Learning Through Reinforcement and Replicator Dynamics, *Journal of Economic Theory* 77 (1997), 1-14.

[2] Börgers, Tilman and Rajiv Sarin, Naive Reinforcement Learning With Endogenous Aspirations, *International Economic Review* 41 (2000), 921-950.

[3] Bush, Robert and Frederick Mosteller, A Mathematical Model For Simple Learning, *Psychological Review* 58 (1951), 313-323.

[4] Bush, Robert, and Frederick Mosteller, *Stochastic Models for Learning*, New York: John Wiley & Sons, 1955.

[5] Cross, John, A Stochastic Learning Model of Economic Behavior, *Quarterly Journal of Economics* 87 (1973), 239-266.

[6] Easley, David and Aldo Rustichini, Choice Without Beliefs, *Econometrica* 67 (1999), 1157-1184.

[7] Erev, Ido, and Alvin Roth, Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria, *American Economic Review* 88 (1998), 848-881.

[8] Fudenberg, Drew and David Levine, *The Theory of Learning in Games*, Cambridge and London: MIT Press, 1998.

[9] Lakshmivarahan, S., and Mandayam Thathachar, Absolutely Expedient Learning Algorithms for Stochastic Automata, *IEEE Transactions on Systems, Man and Cybernetics* 3 (1973), 281-286.

[10] Narendra, Kumpati and Mandayam Thathachar, *Learning Automata: An Introduction.* Englewood Cliffs: Prentice Hall, 1989.

[11] Roth, Alvin, and Ido Erev, Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term, *Games and Economic Behavior* 8 (1995), 164-212.

[12] Rustichini, Aldo, Optimal Properties of Stimulus-Response Learning Models*, Games and Economic Behavior* 29 (1999), 244-273.

[13] Samuelson, Larry, and Jianbo Zhang, Evolutionary Stability in Asymmetric Games, *Journal of Economic Theory* 57 (1992), 363-391.

[14] Sarin, Rajiv, Learning Through Reinforcement: The Cross Model, mimeo., Texas A&M University, 1995.

[15] Schlag, Karl, A Note on Efficient Learning Rules, mimeo., University of Bonn, 1994.

[16] Schlag, Karl, How to Choose - A Boundedly Rational Approach to Repeated Decision Making, mimeo., European University Institute, Florence, 2001.

[17] Toyama, Yoshihito and Masayuki Kimura, On Learning Automata in Nonstationary Random Environments, *Systems, Computers, Controls* 8 (1977), No.6, 66-73.

[18] Weibull, Jörgen, *Evolutionary Game Theory*, Cambridge and London: The MIT Press, 1995.