Expedient and Monotone Learning Rules[1]

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

February 2001
INCOMPLETE DRAFT

[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.
[3]Department of Economics, University of Malaga.
[4]Department of Economics, Texas A&M University.

Abstract

This paper studies two properties of reinforcement learning rules: absolute expediency and monotonicity. These properties ensure that an agent whose environment stays constant between two periods improves her performance from one period to the next in terms of expected payoff (absolute expediency) or in terms of the probability with which the expected payoff maximising action is chosen (monotonicity). We give a simple characterisation of monotone learning rules and show that all monotone learning rules are absolutely expedient. It follows from this characterisation that there is a close link between such learning rules and the *replicator dynamics* of evolutionary game theory. Hence, we are able to show that they imply that in stationary environments in the long run the agent maximises expected payoff.

1 Introduction

We study models of reinforcement learning in which at any point in time the agent is described by a probability distribution over actions which indicates how likely she is to take any of her actions. The agent then takes a randomly determined action, receives a payoff, and then updates the probability distribution. The new distribution only depends on the previous distribution, on the action taken, and on the payoff received. The focus of the paper is on rules where the new distribution does not depend on any other aspect of history, nor does it depend on time.

Much of previous analytical work on reinforcement learning has focused on particular ad-hoc functional forms (see, for example, Börgers and Sarin [3], [4], Cross [11], Rustichini [25]). This use of ad hoc functional forms is also reflected in the experimental work on the subject. (see, for example, Barron and Erev (2000), Erev and Roth (1999), Mookherjee and Sopher (1998)). Our goal in this paper is to identify reinforcement learning rules which do not rely on such ad-hoc assumptions. Instead, we analyse all reinforcement learning rules which have particular properties which, in our opinion, are interesting.

We focus on near or intermediate term properties. We believe these properties are more interesting than asymptotic properties because the boundedly rational decision maker described by these models may not only be concerned with what happens in the very long run. This may be because the typical environment an agent faces may not be stationary for arbitrarily long periods of time. Furthermore, rules with nice near term properties may also have desirable long term properties. Although our work is theoretical, we hope that also help to guide experimental work by, for example, providing reinforcement learning rules which are analytically better understood and which could be tested for their descriptive ability with experimental subjects.

For simplicity, we shall restrict attention to single person decision problems rather than games. An agent chooses repeatedly among different actions. In each iteration, she receives some random payoff. The payoff distribution depends on the action which the agent takes, but the agent has no knowledge of the payoff distributions. Indeed, we shall assume that the agent does not even have any reason to believe that the environment is stationary. As the successive iterations of the problem take place, the only feedback information available to the agent is that she observes after each period the payoff which she received in that period. The agent does not observe the

payoff which she would have received had she chosen a different action.

We shall analyze two properties of reinforcement learning rules. The first property, which we call *monotonicity*, requires that whatever the decision problem the agent faces and for every state in which the decision maker chooses all actions with positive probability, the probability assigned to the expected payoff maximizing action strictly increases on average from one period to the next, provided the decision problem doesn't change in the intervening time. The second property which we think is probably more attractive we call *absolutely expediency*. It requires that whatever the decision problem the agent faces and for every state in which the decision maker chooses all actions with positive probability, her expected payoff increases from monotonically from one period to the next, provided that the decision problem doesn't change in the intervening time.

Both properties are appealing in the intermediate term. The agent's performance improves in expected term from period to period, as long as the environment doesn't change in the intervening time. Furthermore, this holds for any decision problem. This is attractive if agents have minimal information about the decision problem, and whether it might change provided that such a change is not too likely in the immediate future.

We characterize all monotonic reinforcement learning rules. We find that reinforcement learning rules are monotonic if and only if they are a modified Cross rule (Cross [11]). The key feature of Cross' rule is that the updated probabilities depend in a linear way on the payoff which the decision maker experienced. The more general class of learning rules which we identify in this paper shares this feature with Cross' rule. This feature is important because of the linearity of the expected payoff function. The modification to the Cross rule that is allowed is that payoffs may be subjected to certain linear transformations. The coefficients of the linear transformations are allowed to depend in a limited way on the current state of the decision maker. We know from earlier work (Börgers and Sarin [3]) that the expected movement of that Cross model is similar to the replicator dynamics of evolutionary game theory. In this paper we show that all monotone learning rules have the feature that their expected movement is given by some transformation of the replicator dynamics. We also show that all monotone reinforcement learning rules are absolutely expedient.

We also show that monotonicity is attractive in the long run provided that the environment which stays constant over a long time, and the agent does not adjust her behaviour too fast. In particular, we show that using

such rules ensures that in the long run the agent converges to choose an action which maximizes expected payoffs, with high probability. Roughly speaking, this happens because slowly updating behaviour ensures that the actual behavior keeps track of the expected adjustment of the rule. The expected adjustment, in turn, resembles the movement of the replicator dynamics when that is specialized to decision problems and which converges to the expected payoff maximizing actions in such problems.

The class of monotone reinforcement learning rules which we identify in this paper does not include the learning rule advocated by Erev and Roth ([13], [24]). Erev and Roth's learning rule is, in fact, excluded in our set-up in which the decision maker's state space is the set of all probability distributions over actions. In Erev and Roth's learning rule the state variable is the vector of accumulated payoffs associated with different actions. Mathematically speaking, this means that Erev and Roth's state space is unbounded, whereas ours is bounded. This implies that if all payoffs are positive, then in Erev and Roth's model the relative change in state variables converges to zero as the number of iterations goes to infinity. By contrast, in our model, the step size does not change as the number of iterations increases. We have excluded Erev and Roth's construction in our set-up because it seems more complex than ours, and because it implicitly assumes that the agent regards the environment as stable.

It is also noteworthy that Erev and Roth's learning rule does, in fact, not turn out to be monotonic nor even absolutely expedient if one generalises this notion in the natural way from our setup to theirs (see Sarin [27]). It is, however, asymptotically absolutely expedient. Rustichini [25] uses this to show that the rule finds the expected payoff maximising action asymptotically. Thus Erev and Roth's learning rule does not have the property which we identify as desirable in the near term, but it does perform well in the long run. We show that a simple modification of the Cross rule which makes it time dependent ensures that the time-dependent Cross rule is monotone and converges to the expected payoff maximizing action.

A literature which is closely related to our work is a branch of the machine learning literature which is concerned with the learning behaviour of stochastic automata. The concept of a stochastic automaton is similar to our concept of a learning rule. A useful overview of the literature on stochastic automata and learning has been provided by Narendra and Thathachar

[21].[1] In this literature, absolute expediency was originally defined by Lakshmivarahan and Thathachar [17]. Monotonicity is studied by Toyama and Kimura [30] 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 that of Narendra and Thathachar [21] (Theorem 6.1). Their result characterizes absolutely expedient learning rules assuming that the updating rule is linear in payoffs. The difference with our work is that we do not assume this linearity, but derive it. Moreover, the form of linearity which these authors assume is more restrictive than the form of linearity which we derive.[2] Also, unlike us, these authors do not draw any connections between the learning algorithms which they obtain and Cross' algorithm. Narendra and Thathachar also note that in their framework absolute expediency and monotonicity are equivalent (Comment 6.2). Toyama and Kimura [30] characterise monotone learning rules. They assume linearity of the learning rule in payoffs whereas we derive it. They do not present results on the relation between monotonicity and absolute expediency.[3]

Several authors writing in this literature have shown, as we do, that the asymptotic probability of the optimal action approaches one as the step size of a monotone or absolutely expedient learning rule is reduced. [4] We provide here an independent proof of this because our proof, unlike theirs, is built on the fact that the trajectories of the learning process will follow the replicator dynamics as the step size of the process is slowed down, which seems important for the intuition for the result.

In the economics literature a paper by Easley and Rustichini [12] is closely related to ours. Easley and Rustichini consider an individual who observes the payoff of each of her actions in the realized state of the world. Hence, their

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

[2] Narendra and Thathachar allow the coefficients in the linear transformation of payoffs to depend on the current state of the learning rule and on the action played. Our result shows that one can allow in addition dependence on the strategy whose probability is updated, and still maintain absolute expediency.

[3] In the Introduction to their paper (p.66), Toyama and Kimura do not really distinguish absolute expediency and monotonicity. As we argue in this paper, the equivalence of these two concepts in a framework with more than two actions is not obvious and needs to be shown.

[4] For absolutely expedient learning rules, Narendra and Thathachar provide the relevant result in Section 5.4.5. For monotone learning rules this is Toyama and Kimura's Theorem 3.

information setting provides more information than is typically assumed in reinforcement learning models. Our paper can be viewed as a counterpart of the Easley-Rustichini paper for a setting in which less information is assumed. In their setting, Easley and Rustichini provide axioms for an individual's learning process which ensure that she converges to choose the expected payoff maximizing action. The axioms which they require for their result are quite unrelated to the expediency and monotonicity conditions which we use. Finally, whereas their analysis provides no direct information on the form of reinforcement learning rule the agent uses, although their result reveals that their learning process moves according to some monotone transform of the replicator dynamic, we obtain direct results about the individual's learning process.

Absolute expediency of reinforcement learning rules was previously also investigated in unpublished work by Sarin [27] and Schlag [28]. Sarin combines absolute expediency with other axioms and obtains a learning process which moves like the Cross learning model. Schlag obtains a similar result assuming linearity of the learning rule in probabilities.

Our results are also related to characterizations of replicator dynamics in the literature on evolutionary game theory Schlag [29] considers population learning in a framework in which individuals' learning rules are explicit. Individuals can observe others' actions and payoffs, and thus their learning rules can contain an element of imitation. Schlag introduces a desirable property of learning rules which, roughly speaking, says that in all environments and all current states of the population the expected average payoff in the population is increasing from round to round. Schlag shows that the dynamics of a large, randomly matched population in which all individuals adopt a rule with this property can be approximated by the replicator dynamics. Our work is close in spirit to his, although our setting is, of course, quite different.

Samuelson and Zhang [26] show that a *selection dynamic* satisfies a condition called *aggregate monotonicity* if and only if it is a multiple of the replicator dynamics. Selection dynamics describe the evolution of a population of players. Samuelson and Zhang do not trace back their result to learning rules for individual players. Unlike *monotonicity* as defined in this paper, *aggregate monotonicity* concerns not just the frequency of the best action, but also of other actions. Samuelson and Zhang obtain their result considering just one single environment. By contrast, it is essential for our result that a learning rule must operate in multiple environments.

The remainder of this paper is organized as follows. Section 2 contains the formal framework. Section 3 defines absolute expediency and monotonicity. Section 4 contains a characterization of monotone learning rules. Section 5 describes the implications of monotonicity for long run learning. All proofs are collected in the Appendix.

2 Framework

A decision maker chooses repeatedly one strategy from a finite set S of strategies which has at least two elements. Throughout this paper we keep S fixed. We assume that the decision maker knows S. We index the repetitions of the decision problem by n where n takes values in $\mathbb{N}_0$ [5]. At each iteration, the decision maker first chooses a strategy s from S, and then observes the payoff which she received for that choice.

At each iteration each strategy in S has a payoff distribution attached to it. We normalize payoffs to be between zero and one. The substantial assumption behind this normalization is that the decision maker knows *some* upper and *some* lower bound for payoffs. These may be arbitrarily large or small, respectively. In the following definition an assignment of payoff distributions to strategies is called an environment:

Definition 1 *An environment E is a collection $(\mu_s)_{s\in S}$ of probability measures each having finite support in the interval $[0,1]$.*

At each iteration n there is an environment E_n. Once the decision maker has chosen her strategy at iteration n, the payoff is drawn randomly according to the payoff distribution which is attached to s in environment E_n. The decision maker does not know how E_n is determined. In particular, she doesn't know whether E_n is constant, or whether it is changing. Moreover, if it is changing, she doesn't know the manner in which it is changing.

At each iteration n the decision maker's behaviour is described by a probability distribution $\sigma_n \in \Delta(S)$ where $\Delta(S)$ is the simplex of mixed strategies. The distribution σ_n specifies for each pure strategy s how likely it is that the decision maker chooses s at iteration n. We shall also refer to σ_n as the *state of the decision maker at iteration n*. We denote the relative interior of $\Delta(S)$ by $\dot{\Delta}(S)$.

[5] Throughout this paper we shall denote by $\mathbb{N}_0$ the set $\mathbb{N} \cup \{0\}$.

The decision maker updates σ_n in response to the strategy which she chose at iteration n, and the payoff which she received.

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

The distribution $L(\sigma_n, s, x)$ is thus the state of the decision maker at iteration $n+1$ if her state at iteration n was σ_n, the pure strategy which she chose at iteration n was s, and the payoff which she received was x. For every $s' \in S$ we denote by $L(\sigma_n, s, x)(s')$ the probability which $L(\sigma_n, s, x)$ assigns to s'.

Throughout this paper we shall focus on learning rules which satisfy the following assumption:

Assumption 1. *For any $s \in S$ the learning rule L is continuously differentiable in (σ_n, x).*

This assumption allows us in Section 4 to appeal to well-known theorems regarding the approximation of slow moving stochastic processes by the solution of deterministic differential equations. Continuity of the learning rule in payoff, however, will also play a fundamental role in the characterization of monotone and expedient learning rules.

We finish the section with the formal definition of the learning process generated by a given learning rule L in a stationary environment E. We denote by $\mathcal{B}(\Delta(S))$ the set of all Borel subsets of $\Delta(S)$.

Definition 3 *The* stochastic kernel[6] *K corresponding to an environment E and a learning rule L is a function $K : \Delta(S) \times \mathcal{B}(\Delta(S)) \to [0,1]$ such that*

$$K(\sigma, \Omega) = \sum_{(s,x) \in \{s \in S, x \in [0,1] | L(\sigma,s,x) \in \Omega\}} \sigma(s) \cdot \mu_s(x)$$

for every $\sigma \in \Delta(S)$ and $\Omega \in \mathcal{B}(\Delta(S))$.

Definition 4 *The* learning process *corresponding to a stationary environment E, a learning rule L, and an initial state $\sigma_0 \in \Delta(S)$ is the Markov process $\{\sigma_n\}_{n \in I\!N_0}$ with the initial distribution which assigns probability 1 to σ_0, and with the stochastic kernel K described in Definition 3.*

[6]Intuitively, the *stochastic kernel* is the analog of a transition matrix for a Markov process with continuum size state space.

3 Absolute Expediency and Monotonicity

We next formalize the two properties of learning rules in which we are interested in this paper. Both properties are *short run* properties. They refer to the expected changes in the decision maker's behaviour between a period n and the immediately succeeding period $n+1$, assuming that the environment stays the same between these two periods. Both properties require that, in expected terms, the decision maker's performance improves from period n to period $n+1$.

Given any environment E and any strategy $s \in S$, we denote the expected payoff of strategy s by π_s. That is,

$$\pi_s = \int_0^1 x d\mu_s$$

The set of expected payoff maximising strategies is denoted by S^*. That is,

$$S^* = \{s \in S \mid \pi_s \geq \pi_{s'} \text{ for all } s' \in S\}$$

We suppress in our notation that π_s and S^* depend on E, since, whenever we use this notation, it will always be evident to which environment we are referring.

For any learning rule L and environment E we define a function f which assigns to every possible state of the decision maker σ, and every pure strategy s, the expected change in the probability attached to s if the current state is σ. Formally, $f : \Delta(S) \times S \to \mathbb{R}$ is defined by:

$$f(\sigma, s) = \sum_{s' \in S} \sigma(s') \int_0^1 L(\sigma, s', x)(s) - \sigma(s) d\mu_{s'}$$

for all $\sigma \in \Delta(S)$ and $s \in S$. For $\tilde{S} \subseteq S$, we define

$$f(\sigma, \tilde{S}) = \sum_{s \in \tilde{S}} f(\sigma, s).$$

We denote by $f(\sigma)$ the vector:

$$f(\sigma) = \{f(\sigma, s)\}_{s \in S}.$$

Finally, we define:

$$g(\sigma) = \sum_{s \in S} f(\sigma, s)\pi_s.$$

Definition 5 *A learning rule L is* monotone *if for all environments E with $S^* \neq S$ and all states $\sigma \in \dot{\Delta}(S)$: $f(\sigma, S^*) > 0$.*

Intuitively, a learning rule is monotone, if in every interior state σ, for every environment E, the expected probability of playing an expected payoff maximising strategy increases from iteration n to iteration $n+1$, provided that the environment doesn't change between these two periods.

Definition 6 *A learning rule L is* absolutely expedient *if for all environments E with $S^* \neq S$ and all states $\sigma \in \dot{\Delta}(S)$: $g(\sigma) > 0$.*

In words, a learning rule is absolutely expedient if in every interior state σ, for every non-trivial environment E, expected payoffs increase on average from iteration n to iteration $n+1$, provided that the environment doesn't change between these two periods.

4 Expedient and Monotone Learning Rules

In this section we characterize all monotone learning rules and give a number of necessary conditions for Absolute Expediency. First, we provide an example. Then we show that all monotone and all expedient learning rules share certain features of this example.

Example 1 *(Cross [11].) For all $\sigma \in \Delta(S)$, $s, s' \in S$ with $s \neq s'$, and $x \in [0,1]$:*

$$L(\sigma, s, x)(s) = \sigma(s) + (1 - \sigma(s))x$$

$$L(\sigma, s', x)(s) = \sigma(s) - \sigma(s)x$$

Notice that this learning rule has the somewhat counterintuitive feature that the decision maker *always* increases the probability of the action which she actually played, even if the payoff was very low.

The expected movement for Cross' learning rule is given by:

$$f(\sigma, s) = \sigma(s)[\pi_s - \sum_{s' \in S} (\sigma(s')\pi_{s'})]$$

for all $\sigma \in \Delta(S)$ and all $s \in S$. Notice that the expression on the right hand side appears also on the right hand side of the replicator equation of

evolutionary game theory. It is clear from this expression that Cross' rule is both monotone and absolutely expedient. It is also clear that the expected movement for Cross' rule is zero whenever all actions yield the same expected payoff. As a matter of fact, we turn this observation into a property and show that it happens to be a necessary condition for both expediency and monotonicity properties.

Definition 7 *A learning rule L is* unbiased *if for all environments E with $S^* = S$ and all states $\sigma \in \dot{\Delta}(S)$: $f(\sigma, s) = 0$ for every $s \in S$.*

The next proposition characterises all unbiased learning rules.

Proposition 1 *A learning rule L is* unbiased *if and only if there are functions $\mathcal{A} : \Delta(S) \times S \times S \to \mathbb{R}$ and $\mathcal{B} : \Delta(S) \times S \times S \to \mathbb{R}$ such that for every $(\sigma, s, x) \in \Delta(S) \times S \times [0, 1]$:*

(1) $L(\sigma, s, x)(s) = \sigma(s) + (1 - \sigma(s)) [\mathcal{A}(\sigma, s, s) + \mathcal{B}(\sigma, s, s)x]$

(2) $L(\sigma, s', x)(s) = \sigma(s) - \sigma(s) [\mathcal{A}(\sigma, s', s) + \mathcal{B}(\sigma, s', s)x]$ for all $s' \neq s$

and, for every $\sigma \in \dot{\Delta}(S)$ and $s \in S$:

(3) $\mathcal{A}(\sigma, s, s) = \sum_{s' \in S} \sigma(s')\mathcal{A}(\sigma, s', s)$

(4) $\mathcal{B}(\sigma, s, s) = \sum_{s' \in S} \sigma(s')\mathcal{B}(\sigma, s', s)$

The above proposition shows that a learning rule is *unbiased* if and only if the decision maker, after playing her action and receiving her payoff, submits first the payoff to a *linear* transformation and then applies Cross' rule. The coefficients of this linear transformation are allowed to depend on the current state of the decision maker, on the strategy which she has played, and on the strategy the probability of which she is adjusting.[7] Conditions (3) and (4) restrict the coefficients of the linear transformation.

It is worth noticing that the linearity of the learning rule in payoff, the most striking feature of unbiasedness, comes from the fact the property,

[7]Note that this is more general than transforming the decision problem by submitting every payoff to a linear transformation and then using the Cross' rule in the transformed decision problem.

stated in expected terms, is defined to hold in a variety of environments which contains trivial environments.[8]

The following remark shows that the formula for the expected movement of unbiased learning rules has the same structure as the formula for the expected movement of Cross' rule. Of course, allowance must be made for the fact that payoffs may be submitted to linear transformations. But, once this is taken care of, expected movement of any unbiased learning rule is the same as the movement of evolutionary replicator dynamics.

Remark 1 *Let L be a* unbiased *learning rule, and let E be an environment. Then for every $\sigma \in \Delta(S)$ and every $s \in S$ the expected movement of the probability of s is given by:*

$$f(\sigma, s) = \sigma(s) \left[\mathcal{B}(\sigma, s, s)\pi_s - \sum_{s' \in S} \sigma(s')\mathcal{B}(\sigma, s', s)\pi_{s'} \right]$$

Note also that using condition (4) of Proposition 1, this expression can be written as

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} \sigma(s')\mathcal{B}(\sigma, s', s) \left(\pi_s - \pi_{s'}\right)$$

We next show why this property is relevant to our analysis.

Lemma 1 *Every* monotone *learning rule is* unbiased.

Lemma 2 *Every* absolutely expedient *learning is* unbiased.

The sketch of the proofs is as follows. Consider any biased rule. By definition we can find a trivial environment and a state σ for which there exists some action s with $f(\sigma, s) < 0$. Increase slightly the payoff of s so that it becomes the *unique* expected payoff maximising action. By continuity, it is still the case that $f(\sigma, s) < 0$.. The proof for absolute expediency is similar, although a bit more unraveled. Note again the role played by the fact that the properties are defined to hold in a variety of environments containing trivial environments and the role played by the continuity assumption in payoff.

In order to fully characterise these properties, additional conditions other than unbiasedness are needed. But note that the basis structure of these

[8] An environment E is called trivial if all strategies yield the same expected payoff.

properties is already known by lemmas 1 and 2. The following definitions, which deal with the sign of the coefficients of the linear transformation, will be useful.

Definition 8 *A learning rule L is* own-positive *if $\mathcal{B}(\sigma, s, s) > 0$ for all $s \in S$ and every $\sigma \in \dot{\Delta}(S)$.*

Definition 9 *A learning rule L is* cross-negative *if for all $\sigma \in \dot{\Delta}(S)$:*

(i) $\mathcal{B}(\sigma, s', s) \geq 0$ for all $s', s \in S, s' \neq s$

(ii) There do not exist non-empty subsets $C \subset S$ such that $\mathcal{B}(\sigma, s', s) = 0$ for $s' \in C$ and $s \notin C$.

A learning rule is own-positive if the greater the payoff action gets, the more its probability is (strictly) increased. A learning rule is cross-negative if the greater the payoff an action gets, the greater the probability of the other strategies is (weakly) decreased.[9] Note that cross-negativity allows for the possibility that some cross-effects are null, i.e. $\mathcal{B}(\sigma, s', s) = 0$ for some $\sigma \in \dot{\Delta}(S)$ and $s, s' \in S, s \neq s'$. Condition (ii) restricts these null cross-effects.

Proposition 2 *A learning rule L is* monotone *if and only if L is* unbiased *and* cross-negative.

This proposition shows that the distinctive feature of monotonicity is the cross-negativity property.[10] A quick look at the formula for the expected movement of the probability attached to the optimal strategy reveals why is sufficient.[11] To grasp the intuition for the necessity, we better re-interpret

[9] The reader should not be confused by the terminology "cross–negative" and the fact that it requires $\mathcal{B}(\sigma, s', s) \geq 0$. Look at condition (2) of Proposition 1, and look for the sign *minus* in the formula for $L(\sigma, s', x)(s)$.

[10] The posibility of null cross-effects has some "nasty" effects when dealing with abstract environments. If s is the unique optimal action then monotonicity implies that $f(\sigma, s)$ is strictly positive. But if there are multiple optimal actions and s belongs to the set of optimal action then monotonicity implies that $f(\sigma, S^*)$ is strictly positive, but it is possible that $f(\sigma, s) = 0$. All we know for sure is that $f(\sigma, s)$ will be non-negative. The same argument applies if s belongs to the set of the worst actions for given environment. In general, all we will know for sure is that monotonicity implies that $f(\sigma, s)$ will be non-positive.

[11] Look for the relevant formula at remark 1, and focus on the second formula, that is why it is there for.

its meaning in light of the expected movement of the rule. Consider a given environment E, a given state σ and focus your attention on some particular action $s' \in S$. We are to perform a comparative statics analysis by changing the expected payoff of action s' and computing how $f(\sigma, s)$ changes for all $s \neq s'$. Cross-negativity means that $\partial f(\sigma, s)/\partial \pi_{s'} = -\sigma(s')\mathcal{B}(\sigma, s', s) \leq 0$, i.e. whenever the expected payoff of action s' is decreased, the expected change of all other actions' probabilites is (weakly) increased. Assume this were not true, but that to the contrary, the expected change of the probability of some action s was decreased. Now look for an environment $\widetilde{E}$ such that s is the *unique* optimal action. Lower now the expected payoff of action s'. Note that action s will remain the unique optimal action, but with a smaller expected change in its probability. In the appendix we show that by using this procedure, there will always exist an initial environment $\widetilde{E}$ such that $f(\sigma, s)$ will eventually become negative, contradicting monotonicity.

Remark 2 *It is worth noticing that a monotone rule is also own-positive. Note that unbiasedness implies that* $(1 - \sigma(s))\,\mathcal{B}(\sigma, s, s) = \sum_{s' \neq s} \sigma(s')\mathcal{B}(\sigma, s', s)$. *Cross-negativity implies that the summands are non-negative and that there exists at least some* $s' \neq s$ *with* $\mathcal{B}(\sigma, s', s) > 0$. *Therefore* $\mathcal{B}(\sigma, s, s) > 0$. *Note also that own-positivity does no imply that the probability of the played action is always increased, as Cross' rule does. See example 2 below for a monotone learning rule with an "aspiration level".*

Summarising, cross-negativity makes an unbiased rule go into the "right" direction for monotonicity. When an action gets a higher expected payoff, this action "moves" into the direction of becoming the expected payoff maximising action in the new environment. Accordingly, the learning rule increases the expected change in its probability and decreases all others expected changes probabilities. When an action gets a smaller expected payoff, this action does not move into the direction of becoming the expected payoff maximising action, and therefore the learning rule decreases the expected change of its probability but increases all others' expected changes, re-assuring this way that the optimal action gets its expected change in probability increased.

We next show a surprising result: monotonicity –a concept that involves only the expected movement of the optimal action– implies absolute expediency –a concept that involves the expected movements of all actions–.

Proposition 3 *Every* monotone *learning rule is* absolutely expedient.

The above result is straightforward when the decision problem has only two different expected payoffs. Our proof uses this fact and then takes advantage of the cross–negativity property. The intuition would run as follows. The idea is to perform a comparative statics analysis by focusing on how $g(\sigma)$ changes when lowering the expected payoff of one of the actions with the lowest expected payoff, let s' be such action. With a slight abuse of notation, this can be described as

$$\frac{\partial g(\sigma)}{\partial \pi_{s'}} = f(\sigma, s') + \sum_{s \in S} \pi_s \frac{\partial f(\sigma, s)}{\partial \pi_{s'}}$$

Recall that the learning rule is monotone. Then $f(\sigma, s')$ is non-positive, because s' belongs to the set of the worst actions in the original environment.[12][13] Furthermore, Cross-negativity implies that the summatory is strictly positive. Summarising, when lowering the expected payoff of some of the worst actions, the expected change in the expected payoffs increases. In the appendix you will find this idea formalised in an inductive argument.

Note also that when the decision problem has only two different expected payoffs, absolute expendiency trivially implies monotonicity. Hence, for this case, both properties are equivalent. While we have been unable to show that this equivalency also holds for more general decision problems, we have found a number of necessary conditions for absolute expediency. Lemma 2 showed that unbiasedness was one of such conditions. The following proposition shows that own-positivity is another one.

Proposition 4 *Every* absolutely expedient *learning rule is* own-positive.

We include the proof here because it highlights how the three properties we have analysed are interrelated. Let L be an absolutely expedient rule. We know by lema 2 that L is unbiased. This implies that for all trivial environments $f(\sigma, s) = 0$ for all $s \in S$. Add $\varepsilon > 0$ to the expected payoff associated to some action s'. In the new environment, there are only two different expected payoffs. We now that in this sort of decision problems, absolute expediency implies monotonicity. By monotonicity, $f(\sigma, s') = \varepsilon\sigma(s')\mathcal{B}(\sigma, s', s') > 0$, which implies that L is own-positive.

Note that the above lemma clarifies the relationship between absolute expendiency and monotonicity in the general case. If there exists an absolutely

[12]Look at the second formula in Remark 1.

[13]Rea footnote 10 again if needed.

expedient rule which is not monotone -and such a rule has not been shown to exist in in the literature-, it has to display negative cross-effects.

We end this section with the presentation of further examples of monotone learning rules. Although they might look a bit awkward, they are meant to show that, against the impression the reader might have had after reading Proposition 1, there are monotone rules qualitatively different from the Cross' one. Recall that Cross' rule has two main features: (i) It always increases the probability attached to the played action, and (ii) It always decreases the probability attached to all unplayed actions.

As regards point (i), we present in example 2 a rule which incorporates an aspiration level. If the received payoff falls below that aspiration level, the probability of the played action is decreased. As regards point (ii), we present in example (3) a rule which increases the probabilities of some unplayed actions and decreases the probabilities attached of some others actions..

Recall that Proposition 3 implies that these examples are also absolutely expedient rules.

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 $\sigma \in \Delta(S), s, s' \in S$ with $s \neq s'$:*

$$\begin{aligned}
\mathcal{A}(\sigma, s, s) &= -\sigma(s) \sum_{s' \neq s} \left[(\sigma(s'))^2 (1 - \sigma(s')) \right] \alpha \\
\mathcal{B}(\sigma, s, s) &= +\sigma(s) \sum_{s' \neq s} \left[(\sigma(s'))^2 (1 - \sigma(s')) \right] \\
\mathcal{A}(\sigma, s', s) &= -(1 - \sigma(s'))(1 - \sigma(s))\sigma(s')\sigma(s)\alpha \\
\mathcal{B}(\sigma, s', s) &= +(1 - \sigma(s'))(1 - \sigma(s))\sigma(s')\sigma(s)
\end{aligned}$$

Note that $\mathcal{A}(\sigma, s', s) = -\mathcal{B}(\sigma, s', s)\alpha$ for all $s, s' \in S$, which implies that

$$\begin{aligned}
L(\sigma, s, x)(s) &= \sigma(s) + (1 - \sigma(s))\mathcal{B}(\sigma, s, s)\,(x - \alpha) \\
L(\sigma, s', x)(s) &= \sigma(s) - \sigma(s)\mathcal{B}(\sigma, s', s)\,(x - \alpha) \text{ for all } s' \neq s
\end{aligned}$$

Therefore, according to this learning rule, if strategy s was played in iteration n, the decision maker increases (resp. decreases) in period $n+1$ the

probability assigned to s 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 is increased (resp. decreased), the probability of all other strategies is decreased (resp. increased).

We next show that this rule is monotone. We start out by showing first that it is unbiased. Conditions (1) and (2) of Proposition 1 are trivially satisfied. For conditions (3) and (4), note that as $L(\sigma, s, x)$ is a probability distribution, we have the following:

$$\mathcal{A}(\sigma, s, s) = \sum_{s' \in S} \sigma(s') \mathcal{A}(\sigma, s, s')$$

$$\mathcal{B}(\sigma, s, s) = \sum_{s' \in S} \sigma(s') \mathcal{B}(\sigma, s, s')$$

Noting that this rule is symmetric, i.e. $\mathcal{A}(\sigma, s, s') = \mathcal{A}(\sigma, s', s)$ and $\mathcal{B}(\sigma, s, s') = \mathcal{B}(\sigma, s', s)$ for all $s, s' \in S$, we can re-write the above expressions as

$$\mathcal{A}(\sigma, s, s) = \sum_{s' \in S} \sigma(s') \mathcal{A}(\sigma, s', s)$$

$$\mathcal{B}(\sigma, s, s) = \sum_{s' \in S} \sigma(s') \mathcal{B}(\sigma, s', s)$$

which, in fact, are the very conditions (3) and (4) of Proposition 1. Finally, we need to show that this learning rule is cross-negative and irreducible. But this is trivial as $\mathcal{B}(\sigma, s', s) > 0$ for all $s, s' \in S$ and all $\sigma \in \dot{\Delta}(S)$.

This example arises the question whether within the class of monotone rules, there are more aspiration level learning rules, in particular the existence of endogenous aspiration level learning rules. The following lemma shows that there are no such rules.

Lemma 3 *There are no endogenous aspiration level monotone rules.*

The intuition behind this result relies on the fact that for the aspiration level to be endogenous, it should respond to payoffs, causing the rule be no longer linear in payoffs.

Remark 3 *An alternative model of reinforcement learning with an aspiration level was investigated in Börgers and Sarin [4]. The model of that paper*

postulates that a payoff which is an amount of y below the aspiration level leads to the probability of the action just played to be multiplied by y and all other probabilities to be increased proportionally. If the payoff is above the aspiration level, then the Cross rule is applied.

This model appears plausible, however, in general it fails to be unbiased. To see this, we translate this model into our notation and given that the rule considered in Börgers and Sarin [4] was defined only for the case when the agent had two actions, we will do so.

Let α with $0 \leq \alpha \leq 1$ denote the aspiration level and x be the payoff received. For all $s \in S$,

$$L(\sigma, s, x)(s) = \sigma(s) + (1 - \sigma(s)) \left[\mathcal{A}(\sigma, s, s) + \mathcal{B}(\sigma, s, s)x\right]$$

with

$$\begin{aligned} \mathcal{A}(\sigma, s, s) &= \begin{cases} \frac{-\sigma(s)(1-\alpha)}{1-\sigma(s)} \text{ if } x < \alpha \\ 0 \text{ if } x \geq \alpha \end{cases} \\ \mathcal{B}(\sigma, s, s) &= \begin{cases} \frac{-\sigma(s)}{1-\sigma(s)} \text{ if } x < \alpha \\ 1 \text{ if } x \geq \alpha \end{cases} \end{aligned}$$

Note that this rule with $0 < \alpha < 1$ does not satisfy conditions (1) and (2) of Proposition 1 because, while being linear in payoffs -as required-, the functional form of the functions $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ depends on payoff. This problem is overcome whenever the aspiration level is either 0 or 1. In the case that $\alpha = 0$, the rule reduces to the Cross rule and it is therefore both monotone and absolutely expedient. When $\alpha = 1$, conditions (3) and (4) are not satisfied. These conditions when specialized to the two actions case reduce to $\mathcal{A}(\sigma, s, s) = \mathcal{A}(\sigma, s', s')$ and $\mathcal{B}(\sigma, s, s) = \mathcal{B}(\sigma, s', s')$. It is easily checked that the above rule does not satisfy them.

Note that if some particular action has probability close to one, and has a positive probability of receiving negative payoff, the expected change in this actions probability is negative, independent of the expected payoff of all other actions.[14]

We conclude this section with the following example of a monotone rule which implicitly assumes that actions are related.

[14] Consider the following environment: s gives .7 with probability .5 and .5 with probability .5, s' gives .5 with probability 1 and $\alpha = .6$. We get

$$f(\sigma, s) = \sigma(s) \left[(..5)(.7 - .6)(1 - \sigma(s)) - (..5)(.6 - .5)\sigma(s)\right]$$

Example 3 *Suppose that* $S = \{1, 2, ..., \#S\}$. *For any given strategy* $s \in S$ *we define two strategies* $s \oplus 1$ *and* $s \ominus 1$ *both of which are also contained in* S. *In general,* $s \oplus 1 \equiv s + 1$ *and* $s \ominus 1 = s - 1$.... *But there are two cases in which this is not well-defined, and in these cases we set:* $\#S \oplus 1 \equiv 1$, *and* $1 \ominus 1 \equiv \#S$. *Using the notation of Proposition 1 we can then define a monotone learning rule by setting for all* $\sigma \in \Delta(S)$ *and* $s \in S$*:*

$$\begin{aligned}
\mathcal{A}(\sigma, s, s \oplus 1) &= \prod_{s' \neq s, s \oplus 1} \sigma(s') \\
\mathcal{B}(\sigma, s, s \oplus 1) &= 1 - \mathcal{A}(\sigma, s, s \oplus 1) \\
\mathcal{A}(\sigma, s, s \ominus 1) &= -\prod_{s' \neq s, s \ominus 1} \sigma(s') \\
\mathcal{B}(\sigma, s, s \ominus 1) &= 1 - \mathcal{A}(\sigma, s, s \ominus 1)
\end{aligned}$$

and, for $s' \neq s \ominus 1, s \oplus 1$*:*

$$\begin{aligned}
\mathcal{A}(\sigma, s, s') &= 0 \\
\mathcal{B}(\sigma, s, s') &= 1
\end{aligned}$$

Note that this rule is a modification of Cross' learning rule, the differences being on the way the probabilities of the actions $s \oplus 1$ and $s \ominus 1$ are modified. We write them down, starting out with action $s \oplus 1$:

$$L(\sigma, s, x)(s \oplus 1) = \sigma(s \oplus 1) - \sigma(s \oplus 1)\left[\mathcal{A}(\sigma, s, s \oplus 1) + \mathcal{B}(\sigma, s, s \oplus 1)x\right]$$

This can be written as

$$L(\sigma, s, x)(s \oplus 1) = \sigma(s \oplus 1) - \sigma(s \oplus 1)\left[x + (1 - x)\mathcal{A}(\sigma, s, s \oplus 1)\right]$$

Noting that $\mathcal{A}(\sigma, s, s \oplus 1)$ is positive, the probability of strategy $s \oplus 1$ is reduced by more than it would be the case in Cross' rule.

$$\begin{aligned}
&+ (1 - \sigma(s))\,[.1\,(1 - \sigma(s))] \\
= \;& \sigma(s)\,(1 - \sigma(s))\,(..05) + (1 - \sigma(s))^2\,(..1) - \sigma(s)\,\sigma(s)\,(.05)
\end{aligned}$$

which is clearly negative for large enough p ($p > .67$). Hence, the rule is not monotone.

We now turn our attention to action $s \ominus 1$:

$$L(\sigma, s, x)(s \ominus 1) = \sigma(s \ominus 1) - \sigma(s \ominus 1)\left[\mathcal{A}(\sigma, s, s \ominus 1) + \mathcal{B}(\sigma, s, s \ominus 1)x\right]$$

This can be written as

$$L(\sigma, s, x)(s \ominus 1) = \sigma(s \ominus 1) - \sigma(s \ominus 1)\left[x + (1-x)\mathcal{A}(\sigma, s, s \ominus 1)\right]$$

Noting that $\mathcal{A}(\sigma, s, s \ominus 1)$ is negative, the probability of strategy $s \oplus 1$ is reduced by less than in Cross' rule. In fact, if the payoff x was very low, the probability of $s \oplus 1$ may be increased rather than decreased.

A learning rule of this type might capture the intuitive idea that strategy $s \oplus 1$ is "the opposite" of s, whereas strategy $s \ominus 1$ is "similar" to s. Notice, however, that these relations are not transitive in the example.

We finally check that this rule is monotone. As in the previous example, we first show that this rule is unbiased. Conditions (1) and (2) of Proposition 1 are trivially satisfied. For condition (3) note that

$$\begin{aligned}\sum_{s' \in S} \sigma(s')\mathcal{A}(\sigma, s', s) &= \sigma(s \ominus 1)\mathcal{A}(\sigma, s \ominus 1, s) + \sigma(s \oplus 1)\mathcal{A}(\sigma, s \oplus 1, s) \\ &\quad + \sum_{s' \neq s\ominus, s\oplus 1} \sigma(s')\mathcal{A}(\sigma, s', s)\end{aligned}$$

Taking into account that $\mathcal{A}(\sigma, s', s) = 0$ for $s' \neq s \ominus 1, s \oplus 1$, the above equation reduces to

$$\sum_{s' \in S} \sigma(s')\mathcal{A}(\sigma, s', s) = \sigma(s \ominus 1)\mathcal{A}(\sigma, s \ominus 1, s) + \sigma(s \oplus 1)\mathcal{A}(\sigma, s \oplus 1, s)$$

This can be rewritten as

$$\begin{aligned}\sum_{s' \in S} \sigma(s')\mathcal{A}(\sigma, s', s) &= \sigma(s \ominus 1) \prod_{s' \neq s, s\ominus 1} \sigma(s') - \sigma(s \oplus 1) \prod_{s' \neq s, s\oplus 1} \sigma(s') \\ &= \prod_{s' \neq s} \sigma(s') - \prod_{s' \neq s} \sigma(s') \\ &= 0\end{aligned}$$

Recalling that $\mathcal{A}(\sigma, s, s) = 0$, condition (3) is proven.

We now turn to condition (4).

$$\begin{aligned}\sum_{s'\in S}\sigma(s')\mathcal{B}(\sigma,s',s) &= \sigma(s\ominus 1)\mathcal{B}(\sigma,s\ominus 1,s)+\sigma(s\oplus 1)\mathcal{B}(\sigma,s\oplus 1,s)\\ &\quad+\sum_{s'\neq s\ominus,s\oplus 1}\sigma(s')\mathcal{B}(\sigma,s',s)\end{aligned}$$

This can be rewritten as

$$\begin{aligned}\sum_{s'\in S}\sigma(s')\mathcal{B}(\sigma,s',s) &= \sigma(s\ominus 1)\mathcal{A}(\sigma,s\ominus 1,s)+\sigma(s\oplus 1)\mathcal{A}(\sigma,s\oplus 1,s)\\ &\quad+\sigma(s\ominus 1)+\sigma(s\oplus 1)+\sum_{s'\neq s\ominus,s\oplus 1}\sigma(s')\mathcal{B}(\sigma,s',s)\end{aligned}$$

The first line of the rhs of this equation add up to zero by condition (3). The remaining terms add up to one because $\mathcal{B}(\sigma,s',s)=1$ for $s'\neq s\ominus 1, s\oplus 1$. Therefore we arrive at

$$\sum_{s'\in S}\sigma(s')\mathcal{B}(\sigma,s',s)=1$$

Recalling that $\mathcal{B}(\sigma,s,s)=1$ we have proven condition (4). We finally need to show that this learning rule is cross-negative and irreducible. But this is trivial as $\mathcal{B}(\sigma,s',s)>0$ for all $s,s'\in S$ and all $\sigma\in\dot{\Delta}(S)$.

5 Stationary Environments

Absolute expediency and monotonicity are both intermediate run properties. However, in stationary environments, we show in this section that they also have desirable long run implications. In this section we study these implications. Our main result is Proposition 8 which shows that in stationary environments a monotone rule converges to the expected payoff maximizing actions with high probability provided it moves in small steps. We begin our analysis with the following Proposition.

Proposition 5 *Suppose that L is a monotone learning rule, let $\sigma_0\in\dot{\Delta}(S)$ be the initial state, and consider some environment E. Let $\{\sigma_n\}_{n\in I\!N_0}$ be the learning process which corresponds to L, σ_o and constant environment E. Then the event*

$$\text{“}\sigma_n(S^*)\to 0 \text{ or } \sigma_n(S^*)\to 1\text{”}$$

has probability 1.

For given learning rule L and given environment E we define a function ϕ which assigns to every state $\sigma_0 \in \dot{\Delta}(S)$ the probability $\phi(\sigma_0)$ of the event $\sigma_n(S^*) \to 1$ if the initial state is σ_0, and if the environment is equal to E for all n.

Proposition 6 *Suppose that L is a monotone learning rule, and that E is an environment such that $S \neq S^*$. Then, for any $\sigma_0 \in \dot{\Delta}(S)$, $\phi(\sigma_0) > \sigma_0(S^*)$.*

This proposition describes a very weak, but certainly desirable property of monotone learning rules: in all non-trivial, stationary environments the probability with which the decision maker ends up playing an expected payoff maximising strategy is larger than the probability with which the decision maker played some such strategy initially.

Stronger results can be proved for monotone learning rules which move in small steps. Such learning rules can be derived from a given learning rule L in the following way:

Definition 10 *For given learning rule L we define for every $\varepsilon \in (0,1)$ a new learning rule L^ε by setting*

$$L^\varepsilon(\sigma, s, x) - \sigma = \varepsilon(L(\sigma, s, x) - \sigma).$$

Intuitively, L^ε describes a learning process which moves into the same direction as L, but at speed ε. We are interested in limit properties of the learning process corresponding to L^ε for fixed environment, and fixed initial state, where the limit which we wish to take is: $\varepsilon \to 0$.

For our next result we introduce a continuous time variable $t \geq 0$, and we adapt the learning process introduced in Definition 4 so that it is a continuous time learning process. If the learning rule is L^ε, we assume that the amount of "real time" which passes between two iterations of the decision problem equals ε. In the time interval which passes between two iterations of the decision problem the state of the decision maker remains constant. This motivates the following definition:

Definition 11 *The* continuous time learning process *corresponding to a constant environment E, a learning rule L^ε, and an initial state $\sigma_0 \in \dot{\Delta}(S)$ is the stochastic process $\{\tilde{\sigma}_t^\varepsilon\}_{t \geq 0}$ whose initial distribution assigns probability 1 to σ_0, and which satisfies for any $t \geq 0$:* [15]

$$\tilde{\sigma}_t^\varepsilon = \sigma_{[\frac{t}{\varepsilon}]}^\varepsilon$$

[15] For $x \in \mathbb{R}$ we denote by $[x]$ the largest integer smaller or equal to x.

where $\{\sigma_n^\varepsilon\}_{n\in\mathbb{N}_0}$ *is the (discrete time) learning process corresponding to* L^ε, σ_0 *and constant environment* E.

This definition of the continuous time learning process has the following desirable feature. If we investigate the process in the case that ε is close to zero, but fix some time interval $[0,t]$, then, as ε is reduced, the number of iterations over which we keep track of the decision maker's behaviour is correspondingly increased. If we didn't increase the number of iterations, but instead kept it fixed, then, if ε were close to zero, almost no change in the decision maker's behaviour would be observed.

To characterize the limit as $\varepsilon \to 0$ we introduce a deterministic dynamic process which starts in σ_0, and which moves into the direction of the *expected movement* of L^ε.

Definition 12 *The* deterministic continuous time learning process *corresponding to a constant environment* E, *a learning rule* L^ε, *and an initial state* $\sigma_0 \in \dot{\Delta}(S)$ *is the continuous time process* $\{\hat{\sigma}_t^\varepsilon\}_{t\geq 0}$ *which satisfies* $\hat{\sigma}_0 = \sigma_0$, *and, for every* $t > 0$, $\hat{\sigma}_t^\varepsilon = \hat{\sigma}_{[\frac{t}{\varepsilon}]}^\varepsilon$ *if* $\frac{t}{\varepsilon}$ *is not an integer, and* $\hat{\sigma}_t^\varepsilon = \hat{\sigma}_{[\frac{t}{\varepsilon}]-1}^\varepsilon + f^\varepsilon(\sigma_{[\frac{t}{\varepsilon}]-1}^\varepsilon)$ *otherwise. Here,* f^ε *is the function that describes the expected movement of the learning rule* L^ε *in the environment* E.

Proposition 7 *Suppose that* L *is a monotone learning rule, that* E *is an environment, and that* $\sigma_0 \in \dot{\Delta}(S)$. *For any* $\varepsilon > 0$ *let* $\{\tilde{\sigma}_t^\varepsilon\}_{t\geq 0}$ *and* $\{\hat{\sigma}_t^\varepsilon\}_{t\geq 0}$ *be the corresponding continuous time learning process and deterministic continuous time learning process. Then for any* $t > 0$, $\delta > 0$ *and* $p \in [0,1)$ *there is an* $\bar{\varepsilon} > 0$ *such that* $\varepsilon \leq \bar{\varepsilon}$ *implies that the probability of the event*

$$\text{“}\max_{0\leq\tau\leq t} \|\tilde{\sigma}_\tau^\varepsilon - \hat{\sigma}_\tau^\varepsilon\| \leq \delta\text{”}$$

is at least p.

This proposition shows that over any finite time horizon the stochastic learning process $\{\tilde{\sigma}_t^\varepsilon\}_{t\geq 0}$ stays with high probability close to the deterministic process $\{\hat{\sigma}_t^\varepsilon\}_{t\geq 0}$, provided that ε is close to zero. Standard results in numerical mathematics show moreover that over any finite time horizon the deterministic process $\{\hat{\sigma}_t^\varepsilon\}_{t\geq 0}$ stays close to the solution of the differential equation

$$\frac{d\bar{\sigma}_t}{dt} = f(\bar{\sigma}_t),$$

provided that ε is close to zero.[16] Thus we can conclude that over finite time intervals and for small ε the solution of the differential equation constitutes a good approximation to the learning process.

We omit the proof of Proposition 7. The proof follows directly from Proposition 1 of Corradi and Sarin [10]. The result is also closely related to Proposition 1 in Börgers and Sarin [3], which was proved using Theorem 1.1 in Chapter 8 of Norman [22]. Whereas that result applies only to particular points in time, t, whereas in Proposition 7 we refer to a time interval $[0, t]$.

Notice that the deterministic process to which Proposition 7 refers has the property:

$$\lim_{t\to\infty} \hat{\sigma}_t^{\varepsilon}(S^*) = 1$$

provided that the initial state is interior. Thus, Proposition 7 comes very close to asserting that for small ε the learning process converges to the expected payoff maximizing actions if the process moves in slowly. However, Proposition 7 considers only *finite* time intervals $[0, t]$. Therefore, we provide a further result which concerns the *asymptotics* for $t \to \infty$ of the decision maker's behaviour.

Proposition 8 *Suppose that L is a monotone learning rule, and that E is an environment. Then for all $\sigma_0 \in \dot{\Delta}(S)$:*

$$\lim_{\varepsilon\to 0} \phi^{\varepsilon}(\sigma_0) = 1.$$

Here, the function ϕ^{ε} assigns to every initial state σ_0 the probability of the event "$\sigma_n^{\varepsilon}(S^*) \to 1$" if the learning rule is L^{ε} and the environment is constant and equal to E.

6 Larger State Spaces

In this section we will study learning rules which are defined over state spaces larger than the strategy simplex. The decision maker will be therefore described by a vector $v_n \in V$, where V is a subset of a finite dimensional Euclidean space. We refer to v_n as the *state of the decision maker at iteration n*. As before, the decision maker updates her state v_n in response to the strategy which she chose at iteration n, and the payoff which she received. Formally speaking,

[16] See, for example, Theorem 203A in Butcher [8].

Definition 13 *A* learning rule L *is a function:* $L : V \times S \times [0,1] \to V$

The vector $L(v_n, s, x)$ is thus the state of the decision maker at iteration $n+1$ if her state at iteration n was v_n, the pure strategy which she chose at iteration n was s, and the payoff which she received was x.

As we did before, we shall focus on learning rules which satisfy the following assumption:

Assumption 2. *For any $s \in S$ the learning rule L is continuously differentiable in (v_n, x).*

How does the decision maker translate her state v_n into *behaviour*, i.e. how does the decision maker chooses the pure strategy she plays from state v_n?. This is done by means of a function h which maps V into $\Delta(S)$. That is, there exists some $h : V \to \Delta(S)$ where for all $v_n \in V$, $h(v_n)$ is the probability distribution which specifies for each pure strategy s how likely it is that the decision maker chooses s at iteration n. Some more notation, for $s \in S$, let $h(v_n)(s)$ denote the probability which $h(v_n)$ assigns to s.

We shall deal with learning rules such that their behavioral rule satisfy the following assumption:

Assumption 3. *The behavioural function h is continuous in v_n.*

The first thing to note about the *behavioural* function h is that, by definition, every state $v \in S$ is mapped into only one probability distribution $\sigma \in \Delta(S)$. However, it might be the case that several $v \in V$ give rise to the same behaviour, i.e. there exist some $v, v' \in V$ with $v \neq v'$ such that $h(v) = h(v')$.. This non-invertibility of the behavioural function h is something we will have to live with for the rest of this section and will show up as a feature of every learning rule defined on non-trivial larger states spaces. .

We are interested in extending the analysis of the properties we have defined on this paper to cover these more general learning rules. Recalling that both monotonicity and absolute expediency properties refer to expected changes in the decision maker's *behaviour*, it is trivial to see how "easy" the analysis goes by using the function h.

For every learning rule L and environment E, let $f_L(v, s)$ be the function which assigns to every possible state $v \in V$ and every pure strategy $s \in S$, the expected change in the probability attached to strategy s if the current

state is v.

$$f_L(v,s) = \sum_{s' \in S} h(v)(s') \int_0^1 \left[h\left(L\left(v, s', x \right)(s) \right) - h(v)(s) \right] d\mu_{s'}$$

We also define

$$g_L(v) = \sum_{s \in S} f_L(v,s)\pi_s$$

Note that these two functions are always well-defined. We are now in the position to generalise our properties to larger state spaces in the most natural way.

Definition 14 *A learning rule L is* monotone *if for all environments E with $S \neq S^*$ and all states $v \in V$ with $h(v) \in \dot{\Delta}(S)$: $f_L(v, S^*) > 0$.*

Definition 15 *A learning rule L is* absolutely expedient *if for all environments E with $S \neq S^*$ and all states $v \in V$ with $h(v) \in \dot{\Delta}(S)$: $g_L(v) > 0$.*

Again, note that the above properties refer to changes in the decision maker's state mapped into the strategy simplex, i.e. when converted into behaviour between to consecutive rounds. Then for any state v_n at iteration n, how do we obtain the mapped state at iteration $n+1$?. We first apply the learning rule L to obtain the state at iteration $n+1$ and then apply the behaviour mapping h to it.

This raises one interesting question: Can the mapped learning process on the strategy simplex be represented by a learning rule defined entirely on the strategy simplex?. This being the case, the analysis of general learning rules could be done by analysing the associated learning rule defined on the simplex.

Definition 16 *Let $L : V \times S \times [0,1] \to V$ be a learning rule defined on the state space V. Then the rule $L_v^p : \Delta(S) \times S \times [0,1] \to \Delta(S)$ is called the learning rule* induced *by L* at state *v if for the unique $\overline{\sigma} = h(v)$, all $s \in S$ and all $x \in [0,1]$*

$$L_v^p(\overline{\sigma}, s, x) = h\left(L\left(v, s, x \right) \right)$$

Note that we do not talk of induced learning rules in a general sense, but rather we condition its definition to the particular state v the decision maker

is at any particular stage (and that therefore they are referred to the associated probability distribution $\overline{\sigma}$). There will therefore be, associated to the same general learning rule L, as many induced learning rules L_v^p as possible states v, and in each of these induced rules, the probability distribution will be a fixed parameter $\overline{\sigma}$.

An induced learning rule L_v^p will be called *monotone (absolutely expedient)* if for all environments E with $S \neq S^*$, $f_v^p(\overline{\sigma}, S^*) > 0$ ($g_v^p(\overline{\sigma}) > 0$). Note that we do not require the condition to hold for every state $\sigma \in \Delta(S)$ (as we required in Section 3) but only for the probability distribution for which the induced learning rule is defined, $\overline{\sigma} = h(v)$. Note however, that this "restrictive" definition of monotonicity and absolute expediency does not make much difference in the characterisation of these properties when compared to the analogous properties defined in Section 3.[17] They will also be characterised by the Propositions 1, 3 and 4 of Section 4, allowance made of the fact that they are defined only on the associated probability distribution $\overline{\sigma}$.

With this qualification, we can now state our first Theorem.

Theorem 1 *A learning rule L is* monotone (absolutely expedient) *if and only if for every $v \in V$, the induced learning rule L_v^p is* monotone (absolutely expedient).

The idea of the proof is quite simple and strightforward. We will prove that for all $v \in V$, $f_L(v, s) = f_v^p(\overline{\sigma}, s)$ where $\overline{\sigma} = h(v)$. As all the relevant properties are stated in terms of the expected changes in probabilities, this will suffice to get the desired results.

Fix $\widehat{v} \in V$ and compute the expected change in the probability attached to s when the state is $\widehat{v}$

$$f_L(\widehat{v}, s) = \sum_{s' \in S} h(\widehat{v})(s') \int_0^1 \left[h\left[L\left(\widehat{v}, s', x\right)(s) \right] - h(\widehat{v})(s) \right] d\mu_{s'}$$

Let $\overline{\sigma}$ be the *unique* probability distribution associated to $\widehat{v}$, i.e. $\overline{\sigma} = h\left(\widehat{v}\right)$. By plugging this into the right hand side of the above expression and recalling that by definition $L_v^p(\overline{\sigma}, s', x) = h\left(L\left(\widehat{v}, s', x\right)\right)$ we get

$$f_L(\widehat{v}, s) = \sum_{s' \in S} \widehat{\sigma}(s') \int_0^1 \left[L_v^p(\widehat{\sigma}, s', x)(s) - \widehat{\sigma}(s) \right] d\mu_{s'}$$

[17] Note that assumption 2 make the induced learning rule f_v^p be continuous in payoff. (Assumption 3 is not needed right now)

Note that the right hand side is $f_v^p(\overline{\sigma}, s)$, i.e. the expected change in the probability attached to strategy s when the state of the decision maker is $\overline{\sigma}$. ■

6.1 On declining step-size learning rules

Learning rules with randomly, and perhaps endogenously declining step sizes may be considered plausible for many reasons. They seem particularly appropriate if the agent believes that her environment is stationary. In this case, declining step sizes can ensure that more recent experiences are not given too much weight when the agent has already has significant experience with the environment. Such rules require larger state spaces than the probability simplex that we have considered so far.

A learning rule with a variable step size is a function $L^\gamma : \Delta(S) \times S \times [0,1] \times Z \to \gtrneqq (S)$, where $\gamma : Z \to \mathbb{R}$ is the step-size function. The next definition formally states what we mean by a variable step-size rule. It provides a definition of such rules in terms of the learning rules L we have considered so far.

Definition 17 *A learning rule L^γ has a variable step size $\gamma(z)$ if*

$$L^\gamma(\sigma, s, x, z) - \sigma = \gamma(z)\,[L(\sigma, s, x) - \sigma]$$

Hence, a learning rule has a *variable step-size* if, conditional on a given state, the change in state induced by the action chosen this period and the payoff received, depends upon the value of the step-size function. Note that so far we have not specified the set Z and nor have we specified the step-size function γ. We next define the properties of interest in this paper for L^γ.

Definition 18 *A learning rule $L^\gamma(\sigma, s, x, z)$ is* monotone *if for all environments E with $S^* \neq S$, for all states $\sigma \in \dot{\Delta}(S)$ and for all z: $f(\sigma, S^*) > 0$.*

In the above definition, $f(\sigma, S^*)$ is defined with respect to the learning rule L^γ in the obvious manner. The same is true for $g(\sigma)$ in the following definition.

Definition 19 *A learning rule $L^\gamma(\sigma, s, x, z)$ is* absolutely expedient *if for all environments E with $S^* \neq S$, for all states $\sigma \in \dot{\Delta}(S)$ and for all z: $g(\sigma) > 0$.*

For any z, we can apply our original analysis, and therefore our results concerning monotone and absolute expedient rules are the same as before. It should be noted that when we consider L^γ the functions $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ depend on time z. and we have

$$\frac{\mathcal{A}(\sigma, s, s')}{\mathcal{A}(\sigma, s, s', z)} = \frac{\mathcal{B}(\sigma, s, s')}{\mathcal{B}(\sigma, s, s', z)} = \gamma(z)$$

In the remaining of this section, we will refer to declining step-size learning rules, that is rules for which $\gamma(z') < \gamma(z)$ if $z' > z$, although some of the results also apply to general time-dependent learning rules.

The following lemma states when L^γ in monotonic and absolutely expedient. The result is straightforward given the relation between L^γ and L and the definiiton of these properties in their respective environments.

Lemma 4 *A learning rule L^γ is monotone (resp. absolutely expedient) if and only if L is monotone (resp. absolutely expedient) and γ does not depend on payoffs.*

If the step size parameter γ were to depend on payoffs then L^γ would no longer be linear in payoffs and so could not be monotone (resp. absolutely expedient) even while L is. An example of a learning rule with a declining step size, which does not depend on payoff, is given in the next example. It is simply the Cross rule with a declining step size. In this rule $Z = \mathbb{N}$, and $\gamma(z) = 1/z$.

Example 4 *For all $\sigma \in \dot{\Delta}(S)$ and all $s \neq s'$ and $x \in [0, 1)$, if s is chosen at time n and receives x,*

$$\begin{aligned} L^\gamma(\sigma, s, x, z)(s) &= \sigma(s) + \frac{1}{z}(1 - \sigma(s))\, x \\ L^\gamma(\sigma, s, x, z)(s') &= \sigma(s') - \frac{1}{z}\sigma(s')\, x \end{aligned}$$

The above rule is clearly absolutely expedient and monotone. One may wonder if the above rule converges to the optimal action in stationary environments. As our next result shows that this is true for a larger class of learning rules L^γ.

Proposition 9 *Suppose L^{γ} is monotone or absolutely expedient. If $\gamma(z) = \frac{1}{Kz}$, $Z = \mathbb{N}$ and $K > 1$, then for all $\sigma \in \dot{\Delta}(S)$, and all E*

$$\Pr\left[\lim_{n\to\infty} \sigma_n(S^*)\right] = 1.$$

Proof. The result follows from a theorem of Arthur (1993) and Posch (1997) noting that $K > 1$ is larger than all possible payoffs. ♠

An important class of step-size rules are those in which the step size is stochastic, and endogously determined. The leading example of such a step-size rule is the Roth and Erev (1995) rule. Their reinforcement learning rule is described in the next example.

Example 5 *(Roth-Erev (1995)) At time n the agent is characterized by a vector of positive numbers $q = (q(s))_{s\in S}$. At time n the agent chooses action s with probability $\sigma(s) = \frac{q(s)}{\Sigma_s q(s)}$. If the agent received a payoff of $x \geq 0$ from the chosen action s then her next period propensities $\hat{q}(s) = q(s) + x$, and $\hat{q}(s') = q(s')$ for all $s' \neq s$.*

The above rule can be described as

$$\begin{aligned} L^{\gamma}(\sigma, s, x, z)(s) &= \sigma(s) + \frac{1}{z}(1 - \sigma(s))x \\ L^{\gamma}(\sigma, s, x, z)(s') &= \sigma(s') - \frac{1}{z}\sigma(s')x \end{aligned}$$

where $Z = \mathbb{R}$ and $\gamma(z) = \Sigma_s q(s) + x$. As the payoff x is random, depending both of the randomly chosen strategy and the randomly chosen state of the world, the step-size of this rule evolves endogenously and stochastically. As payoffs are assumed to be positive this rule has a declining step-size. It is immediate from our earlier results, that the Roth-Erev rule is neither monotonic nor absolutely expedient.[18] In particular, this rule does not satisfy contidions (1) and (2) of proposition 1, i.e. it is not *linear* in payoffs, as z depends on x.[19][20] However, the non-linearity of Roth-Erev rule "gradually

[18] Evenmore, it is not unbiased.

[19] As a matter of fact, conditions (3) and (4) are satisfied.

[20] Following Sarin (1995), suppose $q(s) = 2$, and $q(s') = 3$. Suppose s gives each of 1 and 9 with probability 1/2, and that s' gives 4 for sure. It is easily calculated that $f(\sigma, s)$ and $g(s)$ is negative revealing that the Roth-Erev rule is neither monotonic nor absolutely expedient.

disappears" as n grows to infinity. Hence, it can be thought to asymptotically behaves like a monotone or absolutely expendient rule.

We may wonder whether declining step-size learning rules of the Roth-Erev type, which are not monotone or absolutely expedient, and in which the step size stochasticaly declines, behave well in the long term in stationary environments. This is indeed the case as the following result shows.

7 Appendix

Proof of Proposition 1.

Sufficiency: From the Remark following Proposition 1, for every $\sigma \in \Delta(S)$ and every $s \in S$ the expected movement of the probability of s is given by:

$$f(\sigma, s) = \sigma(s)[\mathcal{B}(\sigma, s, s)\pi_s - \sum_{s' \in S} (\sigma(s')\mathcal{B}(\sigma, s', s)\pi_{s'})]$$

Using condition (4) in Proposition 1 we can re-write this as:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} (\sigma(s')\mathcal{B}(\sigma, s', s)(\pi_s - \pi_{s'}))$$

It is trivial now to show that $f(\sigma, s) = 0$ because $\pi_s = \pi_{s'}$ for all $s, s' \in S$.

Necessity: In the remainder of the proof we consider some given unbiased learning rule L, and we show that L has to have the properties listed in Proposition 1. We proceed in three steps.

Step 1: There are functions $\tilde{\mathcal{A}} : \Delta(S) \times S \times S \to \mathbb{R}$ and $\tilde{\mathcal{B}} : \Delta(S) \times S \times S \to \mathbb{R}$ such that for every $(\sigma, s, x) \in \dot{\Delta}(S) \times S \times (0, 1)$

(1) $L(\sigma, s, x)(s) = \tilde{\mathcal{A}}(\sigma, s, s) + \tilde{\mathcal{B}}(\sigma, s, s)x$

(2) $L(\sigma, s, x)(s') = \tilde{\mathcal{A}}(\sigma, s, s') - \tilde{\mathcal{B}}(\sigma, s, s')x$

Proof: Let $a, b, c \in [0, 1]$, and suppose $a < b < c$. Let $\widehat{s} \in S$, and consider two environments, E and $\widetilde{E}$, both of which have the property that all strategies have the same expected payoff, i.e. in both environments $S^* = S$. Suppose also that the payoff distributions of any strategy $s \in S$ with $s \neq \widehat{s}$ is the same in E and in $\widetilde{E}$. Finally, suppose that in environment E strategy $\widehat{s}$

yields payoff a with probability p and payoff c with probability $1-p$, whereas in environment $\tilde{E}$ strategy $\widehat{s}$ yields payoff b with certainty. Here, p is given by: $p = \frac{c-b}{c-a}$. This ensures that the expected payoff of $\hat{s}$ is the same in the two environments.

Denote by $f(\sigma, s)$ (resp. $\tilde{f}(\sigma, s)$) the expected change in the probability of any strategy $s \in S$ in the environment E (resp. $\tilde{E}$). As for both environment $S^* = S$, we must have for all $s \in S$:

$$f(\sigma, s) = \tilde{f}(\sigma, s) = 0 \Rightarrow$$

$$f(\sigma, s) - \tilde{f}(\sigma, s) = 0 \Leftrightarrow$$

$$pL(\sigma, \widehat{s}, a)(s) + (1-p)L(\sigma, \widehat{s}, c)(s) = L(\sigma, \widehat{s}, b)(s)$$

Replacing p by $\frac{c-b}{c-a}$ and re-arranging yields:

$$(c-a)(L(\sigma, \widehat{s}, b)(s) - L(\sigma, \widehat{s}, a)(s)) = (c-b)(L(\sigma, \widehat{s}, c)(s) - L(\sigma, \widehat{s}, a)(s))$$

This implies that either

$$L(\sigma, \widehat{s}, a)(s) = L(\sigma, \widehat{s}, b)(s) = L(\sigma, \widehat{s}, c)(s)$$

or

$$\frac{L(\sigma, \widehat{s}, c)(s) - L(\sigma, \widehat{s}, b)(s)}{L(\sigma, \widehat{s}, c)(s) - L(\sigma, \widehat{s}, a)(s)} = \frac{c-b}{c-a}$$

As this must be true for all a, b, c with $a < b < c$, it follows that $L(\sigma, \widehat{s}, x)(s)$ must be linear in x, as asserted.

Step 2: For every $\sigma \in \dot{\Delta}(S)$ and $s \in S$:

(1) $\sum_{s' \in S} \sigma(s') \tilde{\mathcal{A}}(\sigma, s', s) = \sigma(s)$

(2) $\sum_{s' \neq s} \sigma(s') \tilde{\mathcal{B}}(\sigma, s', s) = \sigma(s) \tilde{\mathcal{B}}(\sigma, s, s)$

Proof: Consider an environment E such that all actions give the same, deterministic payoff, i.e. for some x: $\mu_s(x) = 1$ for all $s \in S$. Then for this environment, we must have, for every $s \in S$ and $\sigma \in \dot{\Delta}(S)$, $f(\sigma, s) = 0$. By the previous step we know that:

$$f(\sigma, s) = \sum_{s' \in S} \sigma(s') \tilde{\mathcal{A}}(\sigma, s', s) - \sigma(s) + x[\sigma(s) \tilde{\mathcal{B}}(\sigma, s, s) - \sum_{s' \neq s} \sigma(s') \tilde{\mathcal{B}}(\sigma, s', s)] = 0$$

This can be true for all x only if Step 2 is true.

Step 3: To complete the proof we define the functions $\mathcal{A}$ and $\mathcal{B}$ by setting for all $\sigma \in \dot{\Delta}(S)$ and $s, s' \in S$ with $s \neq s'$:

$$\mathcal{A}(\sigma, s, s) = \frac{\tilde{\mathcal{A}}(\sigma, s, s) - \sigma(s)}{1 - \sigma(s)}$$

$$\mathcal{A}(\sigma, s', s) = \frac{\sigma(s) - \tilde{\mathcal{A}}(\sigma, s', s)}{\sigma(s)}$$

$$\mathcal{B}(\sigma, s, s) = \frac{\tilde{\mathcal{B}}(\sigma, s, s)}{1 - \sigma(s)}$$

$$\mathcal{B}(\sigma, s', s) = \frac{\tilde{\mathcal{B}}(\sigma, s', s)}{\sigma(s)}$$

Step 1 implies that with these definitions $\mathcal{A}$ and $\mathcal{B}$ satisfy conditions (1) and (2) of Proposition 1. And step 2 implies that $\mathcal{A}$ and $\mathcal{B}$ also satisfy conditions (3) and (4) of Proposition 1.

Proof of Lemma 1. The proof is indirect. Suppose there were an environment with $S^* = S$, a $\sigma \in \dot{\Delta}(S)$, and an $s \in S$ such that $f(\sigma, s) \neq 0$. Then there has to be some $s \in S$ such that $f(\sigma, s) < 0$. Now suppose that we change payoffs slightly, so that s becomes the *unique* expected payoff maximising action. Because of the continuity of the learning rule, the expected movement in the probability of s will remain negative, contradicting monotonicity.

Proof of Lemma 2. The proof is indirect. Suppose there is an environment E with $S^* = S$, a $\sigma \in \dot{\Delta}S$, and an $s \in S$ such that $f(\sigma, s) \neq 0$. Let π denote the expected payoff associated with E. Let $S^+(E) = \{s \in S : f(\sigma, s) \geq 0\}$ and $S^-(E) = \{s \in S : f(\sigma, s) < 0\}$. Note that $S^-(E) \neq \emptyset$. Now suppose we change payoffs to strategies in $S^-(E)$ slightly by adding $\varepsilon > 0$ to them. Denote the resulting environment $\tilde{E}$. Now, $S^* = S^-(E)$. Denote by $\tilde{g}(\sigma)$ the expected change in payoffs in $\tilde{E}$. Then, because of the continuity of the learning rule,

$$\tilde{g}(\sigma) = \sum_{s \in S^+(\tilde{E})} f(\sigma, s)\, \pi + \sum_{s \in S^-(\tilde{E})} f(\sigma, s)\, (\pi + \varepsilon)$$

$$
\begin{aligned}
&= \sum_{s\in S^+(E)} f(\sigma,\, s)\,\pi + \sum_{s\in S^-(E)} f(\sigma,\, s)\,\pi + \sum_{s\in S^-(E)} f(\sigma,\, s)\,\varepsilon \\
&= \pi \sum_{s\in S} f(\sigma,\, s) + \varepsilon \sum_{s\in S^-(E)} f(\sigma,\, s)
\end{aligned}
$$

As we are dealing with probability distributions, it is true that $\sum_{s\in S} f(\sigma,\, s) = 0$. Furthermore, $\sum_{s\in S^-(E)} f(\sigma,\, s) < 0$ by definition, which implies that $\tilde{g}(\sigma) < 0$, contradicting absolute expediency.

Proof of Proposition 2.

Sufficiency: From the Remark following Proposition 1, for every $\sigma \in \Delta(S)$ and every $s \in S$ the expected movement of the probability of s is given by:

$$f(\sigma, s) = \sigma(s) \sum_{s' \neq s} (\sigma(s')\mathcal{B}(\sigma, s', s)(\pi_s - \pi_{s'}))$$

We need to prove that for all environments E with $S^* \neq S$ and all states $\sigma \in \dot{\Delta}(S)$: $f(\sigma, S^*) > 0$.

$$
\begin{aligned}
f(\sigma, S^*) &= \sum_{s\in S^*} \sigma(s) \sum_{s' \neq s} (\sigma(s')\mathcal{B}(\sigma, s', s)(\pi_s - \pi_{s'})) \\
&= \sum_{s\in S^*} \sum_{s'\notin S^*} (\sigma(s)\sigma(s')\mathcal{B}(\sigma, s', s)(\pi_s - \pi_{s'}))
\end{aligned}
$$

Since $\pi_s - \pi_{s'} > 0$ for all $s \in S^*$, $s' \notin S^*$, we only need to find some $s \in S^*$, $s' \notin S^*$ such that $\mathcal{B}(\sigma, s', s) > 0$. But the irreducibility of L, assures this because otherwise the subset $S \backslash S^*$ would be closed. This implies that L is monotone.

Necessity: In the remainder of the proof we consider some given monotone learning rule L, and we show that L has to have the properties listed in Proposition 2. Lemma 1, we already have shown that L is unbiased. We now prove that it has to be cross-negative. We start out by condition (i).

Condition (i): Consider any $\sigma \in \dot{\Delta}(S)$. Our proof is indirect. Suppose there were $s', s \in S$ with $s' \neq s$ such that $\tilde{\mathcal{B}}(\sigma, s', s) < 0$. Consider an environment E such that $\mu_s(x) = 1, \mu_{s'}(x - \delta) = 1$ and $\mu_{s''}(x - \varepsilon) = 1$ for all $s'' \neq s, s'$. Suppose $\delta, \varepsilon > 0$. Then:

$$f(\sigma, s) = \sigma(s)(\tilde{\mathcal{A}}(\sigma, s, s) + \tilde{\mathcal{B}}(\sigma, s, s)x)$$

$$
\begin{aligned}
& +\sigma(s')(\tilde{A}(\sigma,s',s)-\tilde{\mathcal{B}}(\sigma,s',s)(x-\delta)) \\
& +\sum_{s''\neq s,s'}\sigma(s'')(\tilde{\mathcal{A}}(\sigma,s'',s)-\tilde{\mathcal{B}}(\sigma,s'',s)(x-\varepsilon)) \\
& -\sigma(s) \\
= & \sum_{\hat{s}\in S}\left[\sigma(\widehat{s})\tilde{\mathcal{A}}(\sigma,\hat{s},s)\right]-\sigma(s) \\
& +x\left[\sigma(s)\tilde{\mathcal{B}}(\sigma,s,s)-\sum_{\hat{s}\neq s}\sigma(\widehat{s})\tilde{\mathcal{B}}(\sigma,\hat{s},s)\right] \\
& +\sigma(s')\tilde{\mathcal{B}}(\sigma,s',s)\delta+\sum_{s''\neq s,s'}\sigma(s'')\tilde{\mathcal{B}}(\sigma,s'',s)\varepsilon
\end{aligned}
$$

As L is unbiased, the first two lines of this sum add up to zero. Therefore:

$$
f(\sigma,s)=\sigma(s')\tilde{\mathcal{B}}(\sigma,s',s)\delta+\sum_{s''\neq s,s'}\sigma(s'')\tilde{\mathcal{B}}(\sigma,s'',s)\varepsilon
$$

If $\tilde{\mathcal{B}}(\sigma,s',s)<0$, then this term becomes negative for sufficiently small ε, contradicting monotonicity.

Condition (ii). The proof is indirect. Assume that there exists some closed subset C of S.. Consider environment E such that $\mu_s(x)=1$ for $s\notin C$ and $\mu_{s'}(x-\varepsilon)=1$ for $s'\in C$ with $\varepsilon>0$. Note that $S^*=S\backslash C$.. Then for all $s\in S$

$$
\begin{aligned}
f(\sigma,s) = & \sigma(s)\left[\tilde{\mathcal{A}}(\sigma,s,s)+\tilde{\mathcal{B}}(\sigma,s,s)x\right] \\
& +\sum_{\substack{s''\neq s\\ s''\notin C}}\sigma(s'')\left[\tilde{\mathcal{A}}(\sigma,s'',s)-\tilde{\mathcal{B}}(\sigma,s'',s)x\right] \\
& +\sum_{s'\in C}\sigma(s')\left[\tilde{\mathcal{A}}(\sigma,s',s)-\tilde{\mathcal{B}}(\sigma,s',s)\,(x-\varepsilon)\right] \\
& -\sigma(s) \\
= & \sum_{s''\in S}\sigma(s')\tilde{\mathcal{A}}(\sigma,s'',s)-\sigma(s) \\
& +x\left[\sigma(s)\tilde{\mathcal{B}}(\sigma,s,s)-\sum_{s''\neq s}\sigma(s'')\tilde{\mathcal{B}}(\sigma,s'',s)\right]
\end{aligned}
$$

$$+\sum_{s'\in C}\sigma(s')\tilde{\mathcal{B}}(\sigma,s',s)\varepsilon$$

As L is unbiased, the first two lines add up to zero. Then

$$f(\sigma,s)=\sum_{s'\in C}\sigma(s')\tilde{\mathcal{B}}(\sigma,s',s)\varepsilon$$

and

$$f(\sigma,S^*)=\sum_{s\notin C}\sum_{s'\in C}\sigma(s')\tilde{\mathcal{B}}(\sigma,s',s)\varepsilon$$

If C is closed, $\mathcal{B}(\sigma,s',s)=0$ for all $s'\in C$, $s\notin C$, and then $f(\sigma,S^*)=0$, contradicting monotonicity.

To complete the proof of the proposition, we define the functions $\mathcal{A}$ and $\mathcal{B}$ by setting for all $\sigma\in\dot{\Delta}(S)$ and $s,s'\in S$ with $s\neq s'$:

$$\mathcal{A}(\sigma,s,s)=\frac{\tilde{\mathcal{A}}(\sigma,s,s)-\sigma(s)}{1-\sigma(s)}$$

$$\mathcal{A}(\sigma,s',s)=\frac{\sigma(s)-\tilde{\mathcal{A}}(\sigma,s',s)}{\sigma(s)}$$

$$\mathcal{B}(\sigma,s,s)=\frac{\tilde{\mathcal{B}}(\sigma,s,s)}{1-\sigma(s)}$$

$$\mathcal{B}(\sigma,s',s)=\frac{\tilde{\mathcal{B}}(\sigma,s',s)}{\sigma(s)}$$

Proof of Proposition 3.

Consider a monotone learning rule. Recall that we denote by $g(\sigma)$ the expected change in payoffs if the state is σ. We need to show that $g(\sigma)>0$ whenever $\pi_s\neq\pi_{s'}$ for some $s,s'\in S$. We shall prove this by induction over the number of expected payoff values which are possible in the given environment, i.e. by induction over $\sharp\{x\in\Re|\pi_s=x \text{ for some } s\in S\}$.

We begin with the case $\sharp\{x\in\Re|\pi_s=x \text{ for some } s\in S\}=2$.... Recall that for monotone learning rules the expected change in a strategy's probability is:

$$f(\sigma,s)=\sigma(s)\sum_{s'\neq s}\sigma(s')\mathcal{B}(\sigma,s',s)(\pi_s-\pi_{s'})$$

where $\mathcal{B}(\sigma, s', s)$ is nonnegative for all $s' \neq s$ and strictly positive for at least some $s' \neq s$. From this formula it is obvious that, if there are only two possible expected payoff levels, the expected change in the probability of a strategy with the higher expected payoff level will be non negative. The expected change in the probability of a strategy with the lower expected payoff level will be no positive. Condition (ii) of Cross-negativity will imply that the expected change in expected payoff will be strictly positive.

Now suppose we had shown the result for the case that $\sharp\{x \in \Re | \pi_s = x \text{ for some } s \in S\} = n-1$. We need to show it for the case that $\sharp\{x \in \Re | \pi_s = x \text{ for some } s \in S\} = n$. 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 $g'(\sigma)$. By the inductive assumption we know that $g'(\sigma) > 0$. We shall now show that $g(\sigma) - g'(\sigma) > 0$. This then obviously implies the claim.

To calculate $g(\sigma) - g'(\sigma)$ we denote for every $s \in S$ by $f'(\sigma, s)$ the expected change in the probability of strategy s in the modified environment if the current state is σ. Then:

$$\begin{aligned} g(\sigma) - g'(\sigma) &= \sum_{s \notin \overline{S}} f(\sigma, s)\pi_s + \sum_{s \in \overline{S}} f(\sigma, s)\overline{\pi} \\ &\quad - \sum_{s \notin \overline{S}} f'(\sigma, s)\pi_s - \sum_{s \in \overline{S}} f'(\sigma, s)(\overline{\pi} + k) \\ &= \sum_{s \notin \overline{S}} (f(\sigma, s) - f'(\sigma, s))\pi_s \\ &\quad + \sum_{s \in \overline{S}} (f(\sigma, s) - f'(\sigma, s))\overline{\pi} - \sum_{s \in \overline{S}} f'(\sigma, s)k \end{aligned}$$

The formula which we gave earlier for the expected change in a strategy's probability makes it obvious that for strategies $s \notin \overline{S}$ we have:

$$f(\sigma, s) - f'(\sigma, s) = \sigma(s) \sum_{\overline{s} \in \overline{S}} \sigma(\overline{s}) \mathcal{B}(\sigma, \overline{s}, s)k$$

Because the sum of the probabilities can't change, we can conclude that

$$\begin{aligned}\sum_{\overline{s}\in\overline{S}}(f(\sigma,\overline{s})-f'(\sigma,\overline{s})) &= -\sum_{s\notin\overline{S}}(f(\sigma,s)-f'(\sigma,s)) \\ &= -\sum_{s\notin\overline{S}}\sum_{\overline{s}\in\overline{S}}\sigma(s)\sigma(\overline{s})\mathcal{B}(\sigma,\overline{s},s)k\end{aligned}$$

Using these formulas, we can rewrite our earlier equation as:

$$\begin{aligned}g(\sigma)-g'(\sigma) &= \sum_{s\notin\overline{S}}\sum_{\overline{s}\in\overline{S}}\sigma(s)\sigma(\overline{s})\mathcal{B}(\sigma,\overline{s},s)k\pi_s \\ &\quad -\sum_{s\notin\overline{S}}\sum_{\overline{s}\in\overline{S}}\sigma(s)\sigma(\overline{s})\mathcal{B}(\sigma,\overline{s},s)k\overline{\pi}-\sum_{s\in\overline{S}}f'(\sigma,s)k \\ &= \sum_{s\notin\overline{S}}\sum_{\overline{s}\in\overline{S}}\sigma(s)\sigma(\overline{s})\mathcal{B}(\sigma,\overline{s},s)k(\pi_s-\overline{\pi}) \\ &\quad -\sum_{s\in\overline{S}}f'(\sigma,s)k\end{aligned}$$

The first term in this difference is evidently strictly positive. The second term, which is subtracted, is negative because for every $s\in\overline{S}$ the expected change $f'(\sigma,s)$ is non-positive and condition (ii) of cross-negativity applies. This is because in the modified environment strategies $s\in\overline{S}$ are among the strategies with the lowest expected payoff, and the same simple argument that we used at the beginning of the proof to deal with the case $\sharp\{x\in\Re|\pi_s=x \text{ for some } s\in S\}=2$ proves in the general case that the expected change in the probability of strategies which have the lowest expected payoff must be negative. We can conclude that $g(\sigma)-g'(\sigma)>0$, as required.

Proof of Proposition 5

By the definition of monotone learning rules the stochastic process $\{\sigma_n(S^*)\}_{n\in\mathbb{N}_0}$ is a submartingale which is bounded from below by 0 and from above by 1. Therefore, by the *Martingale Convergence Theorem* (Grimmett and Stirzaker [16], p. 454), $\{\sigma_n(S^*)\}_{n\in\mathbb{N}_0}$ converges almost surely to a limit random variable σ_∞.

It remains to show that $\sigma_\infty(S^*)=0$ or 1 with probability 1. This follows if we can show that for every pair $\alpha,\beta\in(0,1)$ with $\alpha<\beta$ the probability of $\sigma_\infty(S^*)\in[\alpha,\beta]$ is zero. Let a,b satisfy: $0<a<\alpha<\beta<b<1$. For every $\bar{n}\in\mathbb{N}$ and $\eta>0$ let $\Phi^\eta_{\bar{n}}$ denote the event "$\sigma_n(S^*)\in[a,b]$ and

$\mid \sigma_n(S^*) - \sigma_{n-1}(S^*) \mid \leq \eta$ for all $n \geq \bar{n}$". Clearly, for every $\eta > 0$ the event $\sigma_\infty \in [\alpha, \beta]$ is contained in the event $\cup_{\bar{n} \in \mathbb{N}} \Phi^\eta_{\bar{n}}$, and therefore it suffices to show that for some $\eta > 0$ the probability of $\cup_{\bar{n} \in \mathbb{N}} \Phi^\eta_{\bar{n}}$ is zero. This follows if we show that for some $\eta > 0$ the probability of $\Phi^\eta_{\overline{n}}$ is zero for every $\overline{n} \in \mathbb{N}$.

Fix $\overline{n}$. For every $n \geq \overline{n}$ we denote by Ψ^η_n the event "$\sigma_n(S^*) \in [a, b]$ and $\mid \sigma_n(S^*) - \sigma_{n-1}(S^*) \mid \leq \eta$". Write $\Pr(\Psi^\eta_n)$ for the probability of Ψ^η_n and write $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ for the probability of Ψ^η_{n+1} conditional on Ψ^η_n. Then the probability of $\Phi^\eta_{\overline{n}}$ can be written as: $\Pr(\Psi^\eta_{\bar{n}}) \cdot \Pr(\Psi^\eta_{\bar{n}+1} \mid \Psi^\eta_{\bar{n}}) \cdot \Pr(\Psi^\eta_{\bar{n}+2} \mid \Psi^\eta_{\bar{n}+1}) \cdot \Pr(\Psi^\eta_{\bar{n}+3} \mid \Psi^\eta_{\bar{n}+2}) \cdot \ldots$. Suppose we can show that there is some ζ with $0 < \zeta < 1$ such that for all $n \in \mathbb{N}$ the conditional probability $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ is bounded from above by ζ. Then the above infinite product converges to zero, and therefore the proof is complete.

Consider the expected value of $\sigma_{n+1}(S^*) - \sigma_n(S^*)$ conditional on Ψ^η_n. This is bounded from below by $\xi \equiv \min_{\sigma \in \Delta(S) \text{ and } \sigma(S^*) \in [a,b]} f(\sigma, S^*)$. Because L is monotone, $\xi > 0$. Now consider the probability that $\sigma_{n+1}(S^*) - \sigma_n(S^*) < \frac{\xi}{2}$, conditional on Ψ^η_n. Intuitively, for the *expected* change to be at least ξ, the probability that the *actual* change is less than $\frac{\xi}{2}$ must not be too large. In fact, a simple calculation shows that it cannot be more than $\frac{1-\xi}{1-\frac{\xi}{2}}$.

Now set $\eta = \frac{\xi}{2}$. Then the preceding paragraph implies that $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ is bounded from above by $\frac{1-\xi}{1-\frac{\xi}{2}}$. Thus we have found a uniform upper bound for $\Pr(\Psi^\eta_{n+1} \mid \Psi^\eta_n)$ which is less than one, and the proof is complete.

Proof of Proposition 6

Because L is monotone, the unconditional expected values satisfy: $E(\sigma_n(S^*)) < E(\sigma_{n+1}(S^*))$ for all $n \in \mathbb{N}_0$. Hence: $\lim_{n\to\infty} E(\sigma_n(S^*)) > \sigma_0(S^*)$. Proposition 2 implies: $\lim_{n\to\infty} E(\sigma_n(S^*)) = \phi(\sigma_0)$. Thus, we can conclude: $\phi(\sigma_0) > \sigma_0(S^*)$.

Proof of Proposition 8

Consider a given and fixed initial state $\sigma_0 \in \dot{\Delta}(S)$. Recall that this implies: $\lim_{t\to\infty} \hat{\sigma}^\varepsilon_t(S^*) = 1$. Therefore for every $\delta > 0$ there will be a $t > 0$ such that $\hat{\sigma}^\varepsilon_t(S^*) \geq 1 - \delta$. By Proposition 6 there will then be for every $\delta > 0$ and $p \in [0, 1)$ a $t > 0$ and an $\bar{\varepsilon} > 0$ such that $\varepsilon \leq \bar{\varepsilon}$ implies that the probability of the event "$\tilde{\sigma}^\varepsilon_t(S^*) \geq 1 - \delta$" is at least p. Now recall from Proposition 5 that, conditional on "$\tilde{\sigma}^\varepsilon_t(S^*) \geq 1 - \delta$" the probability of "$\tilde{\sigma}^\varepsilon_t \to 1$" is at least $1 - \delta$. Thus we can conclude that for every $\delta > 0$ and $p \in [0, 1)$ there will be an $\bar{\varepsilon} > 0$ such that $\varepsilon \leq \bar{\varepsilon}$ implies that the probability

of the event "$\tilde{\sigma}_t^\varepsilon \to 1$" is at least $p(1-\delta)$. This implies the claim.

References

[1] Barron, and I. Erev, On the relationship between decisions in one-shot and repeated tasks: Experimental results and the possibility of general models, mimeo, Technion.

[2] Benveniste, A., M. Metivier and P. Piouret (1990): *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin/Heidelberg.

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

[4] Börgers, T. and R. Sarin, Naive Reinforcement Learning With Endogenous Aspirations, *International Economic Review*, forthcoming.

[5] Brown, G. W., Iterative Solutions of Games by Fictitious Play, in: T. C. Koopmans (editor), *Activity Analysis of Production and Allocation,* New York: Wiley, 1951.

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

[7] Bush, R. and F. Mosteller, *Stochastic Models for Learning*, New York: John Wiley & Sons, 1955.

[8] Butcher, J.C., *The Numerical Analysis of Ordinary Differential Equations*, Chichester, etc.: John Wiley & Sons, 1987.

[9] Camerer C., and Ho, T.-H.: Experience-weighted Attraction Learning in Normal Form Games, *Econometrica* 67 (1999), 837-874.

[10] Corradi, V. and R. Sarin, Continuous Approximations of Stochastic Evolutionary Game Dynamics, *Journal of Economic Theory*, 94 (2000), 163-191.

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

[12] Easley, D. and A. Rustichini, Choice Without Beliefs, *Econometrica* 67 (1999), 1157-1184.

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

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

[15] Feltovich, N., Reinforcement-Based vs. Beliefs-Based Learning Models in Experimental Asymmetric-Information Games, *Econometrica* 68 (2000), 605-641.

[16] Grimmett, G. and D. Stirzaker, *Probability and Random Processes* (second edition), Oxford: Clarendon Press, 1992.

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

[18] Ljung, L. Analysis of recursive stochastic algorithms, *IEEE Transactions of Automatic Control*, 22, 551-575.

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

[20] Mookherjee, D. and B. Sopher, Learning and Decision Costs in Experimental Constant Sum Games, *Games and Economic Behavior* 19 (1997), 97-132.

[21] Narendra, K. and M. Thathachar, *Learning Automata: An Introduction.* Englewood Cliffs: Prentice Hall, 1989.

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

[23] Robinson, J., An Iterative Method of Solving a Game, *Annals of Mathematics* 54 (1951), 296-301.

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

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

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

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

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

[29] Schlag, K., Why Imitate, and if so, How? A Boundedly Rational Approach to Multi-armed Bandits, *Journal of Economic Theory* 78 (1998), 130-156.

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