

# Adaptive Learning with Indirect Payoff Information<sup>1</sup>

Dana Heller  
Department of Economics  
University of Chicago  
Chicago, IL 60637

Rajiv Sarin  
Department of Economics  
Texas A&M University  
College Station, TX 77843

October 2001

<sup>1</sup>Dana Heller thanks the NSF (SES-0111781) and Rajiv Sarin thanks the Bush program in the economics of public policy and the Program to enhance scholarly and creative activities at Texas A&M University for financial assistance.

## **Abstract**

We investigate a model of adaptive learning in which agents optimize given their ranking over actions. The updating procedure for the ranking nests a number of previously investigated learning rules. A key feature of the updating procedure is that it allows the agent to treat payoffs experienced directly differently than indirect information about payoffs from other actions.

We show that when the agent does not discount past experience, the learning rule is well behaved and frequencies of choice converge. However, depending on the shape of the distortion of indirect payoff information, behavior can either “lock into” an action or converge to look as if the agent is playing a mixed strategy. The latter form of behavior emerges when the agent views “the grass as being greener on the other side.” When past experience is discounted and indirect payoff information is not used, we point to a sharp discontinuity in behavior as the discount factor approaches one.

# 1 Introduction

People learn what to choose from their own experiences. They also learn from the experiences of others. For example, a farmer will learn from her own past experience about the profitability of differing amounts of fertilizer input. She may also learn from her neighbor's experience. If a neighbor informs her that he produced more output (per unit of land) when he used a smaller amount of fertilizer (per unit of land) then the farmer may change the amount of fertilizer she uses in the direction of her neighbor. Of course, the farmer need not view the information provided by her neighbor in the same manner as the information she obtains from her own choices.

In this paper we propose a model of adaptive learning in which an agent evaluates an action according to the payoffs it obtained in the past, and, how much experience she has had with the action. Payoff information is obtained when an action is chosen and may also be obtained, from "others," when it is not chosen. The latter, which we call "indirect payoff information," need not be treated in the same way as the payoff obtained from the chosen action. That is, indirect payoff information may be distorted by the agent. The particular manner in which it is distorted may depend, for example, on how the agent views her source of information.

The model "counts" the experience the agent has had with each action. Choosing an action gives the agent some experience with it. Even when not chosen, the agent may feel she has obtained experience with an action because of the indirect payoff information she obtains about the action. For example, even though the farmer has never used a particular amount of fertilizer input she may feel she has had some experience with it because her neighbor has informed her about his experience. The model allows for the experience obtained from unchosen actions to differ from that obtained from the chosen action. We also allow for the previously accumulated experience to decay over time. Such decay might occur if the agent views her recent information as more relevant and hence more valuable.

The payoff information and experience weights are combined to give a ranking of the alternative actions, called the *score* of an action. The score summarizes information about the quality of an action, with those having higher scores representing more attractive actions. At each time, the agent chooses the action with the highest score.

We study the model in the context of an agent repeatedly facing a decision problem. The payoff to an action in a period depends on the state of the

world selected in that period, where we assume the decision environment is stationary over time. We do not assume, however, that the agent has any knowledge of it. The agent need not even know whether she is facing a decision problem or a game. She only knows the set of feasible actions. In each period the agent obtains information about the payoff to the action she chooses and she may also obtain indirect payoff information. When she obtains the latter, we assume it to represent the correct information. One objective of this paper is to analyze the consequences of specific manners of distorting indirect payoff information. The effect of such distortions is most clearly illustrated when the agent obtains the correct information. These distortions are assumed to persist over time. We believe this may be a reasonable assumption in situations where the agent may not be able to verify the accuracy of indirect information or it is too costly to do so. We will discuss some experimental evidence supporting this assumption.

After presenting the general model in the next section, we study in Section 3 the implications of treating indirect payoff information differently from direct payoff information. For simplicity, our analysis in this section focuses on the case in which previously accumulated experience does not decay.<sup>1</sup> Specifically, in Section 3, we characterize the asymptotic behavior of learning rules which are allowed to treat indirect payoff information differently from direct payoff information. We show that asymptotic behavior either converges to choose a single action, or converges to choose more than one action. The latter occurs when the agent distorts indirect payoffs upwards. Such an agent, of course, lives under the illusion that “the grass is greener on the other side”. In such a situation choices may not converge to a single action and cyclical choice behavior may emerge. We derive the conditions under which this happens and show that a certain mix of actions gets played, and that the frequencies of choice converge as does the score of each such action. The result provides a bound on the degree of distortion consistent with convergence to the optimal choice.

Cyclical choice behavior has been studied in various contexts. Marketing research often attempts to determine whether consumers’ behavior is characterized by inertia or “variety seeking.” For example, Seetharaman and Chintagunta (1998) add to the updating procedure a parameter, which de-

---

<sup>1</sup>An example of the behavior consistent with this specialization of our model is described by the fictitious play algorithm. Our result concerns a much larger class of adaptive learning rules.

pending on its sign, prefers variety or inertia. Consequently, cyclical behavior can be attributed to preference for variety. Nyarko (1991) obtained cyclical behavior in a Bayesian framework. He shows that a monopolist, who maximizes the sum of discounted profits and faces a linear demand curve, may exhibit cyclical behavior if she has a mis-specified model. Our model provides another rationale for cyclical behavior: The mis-perception of actions not chosen and payoffs not directly obtained.

The updating rule in the event that the agent obtains no information regarding unplayed actions specializes to some of the previously studied models which allow only direct information. In Section 4 we analyze this case and point out a sharp discontinuity in behavior as the amount by which the agent discounts previously accumulated experience vanishes. Specifically, we show that if the agent discounts past experience then choice converges to the maxmin action. However, if she does not discount her past experience, then choice converges to some action depending on the initial ranking of the actions and the stochastic realization of the states.

Models that study learning when indirect payoff information is not available (or is not used, even if available) have been studied by Borgers et. al. (2001), Borgers and Sarin (1997, 2000), Gilboa and Schmeidler (1996), Rustichini (1999), Sarin and Vahid (1999), among others. These models have been found to be useful in describing behavior in a variety of situations by Chen and Khoroshilov (2001), Erev and Roth (1998), Feltovich (2000), Mookherjee and Sopher (1994), Roth and Erev (1995), Sarin and Vahid (2001), among others. Models which take indirect payoff information into account have been studied by Easley and Rustichini (1999), Fudenberg and Kreps (1993), Fudenberg and Levine (1995, 1998) and have been found to be useful in describing behavior by Camerer and Ho (1999), Camerer et. al. (1999), Grosskopf et. al. (2001), Van Huyck et. al. (1996), among others.

The model closest to ours is the “experience-weighted attraction” learning model developed by Camerer and Ho (1999) and used by them to describe data collected in various experiments. While the two models share the same building blocks in terms of the updating procedure, they differ in their specifications. We provide a detailed comparison of the models in the next section.

This paper is structured as follows. Section 2 presents the model in detail and highlights its relation to some of the previously proposed models. Section 3 considers the affects of distortions of indirect payoff information under the assumption that previously accumulated experience does not decay.

Section 4 considers the effect of allowing previously accumulated experience to decay. Section 5 concludes.

## 2 The Model

The individual has a finite set of actions,  $A = \{a^1, \dots, a^I\}$ . The individual knows  $A$ . In each period  $t$ , after the individual chooses an action, nature chooses a state of the world out of  $\Omega = \{\omega^1, \dots, \omega^J\}$  according to a probability distribution  $\mu = (\mu^1, \dots, \mu^J)$ , where  $\mu^j$  gives the probability that state  $\omega^j$  is realized. The individual does not know  $\mu$  and may not even know  $\Omega$ .

Payoffs are realized according to the state and the chosen action. Let  $a_t \in A$  denote the action chosen in period  $t$ , and  $\omega_t \in \Omega$  denote the state of the world realized in period  $t$ . The objective payoff from action  $a^i$  in period  $t$  is denoted  $\pi^i(\omega_t)$ , which, when no confusion may arise, we shall denote by  $\pi_t^i$ .

The individual ranks each action according to its *score*. The score of an action provides a subjective measure of its historical performance. Let  $S_t^i$  denote the score of action  $a^i$  at time  $t$ , and  $S_t$  the score vector at that time. The score vector is adaptively updated in each period. In updating the score of an action the agent uses subjective information about the payoff of the action and a subjective measure of the experience she has had with the action. We now describe how this is done.

*Perceived payoffs* are used in updating scores in every period. If an action is chosen, the agent directly experiences its objective payoff. For this action, the perceived payoff is assumed to equal its objective payoff. For actions not chosen, the agent may have only indirect information about what the payoffs could have been. For these actions, we allow for the possibility that the agent distorts the information she obtains. The possible information distortion is represented by a *distortion* function  $D : \mathfrak{R} \rightarrow \mathfrak{R}$ . Formally, perceived payoffs  $\tilde{\pi}_t^i$  at time  $t$  are,

$$\tilde{\pi}_t^i = \begin{cases} \pi^i(\omega_t) & \text{for } a^i = a_t \\ D(\pi^i(\omega_t)) & \text{for } a^i \neq a_t \end{cases}$$

The function  $D$  represents the agent's attitude towards indirect payoff information obtained regarding unplayed actions. Such information may be obtained from "others". For example, indirect payoff information concerning movies the agent has not herself watched may have been obtained from her friends. In this case, the agent may distort the indirect information

upward or downward depending on whether she believes her friends typically understate or overstate such information. The agent may continue to systematically distort the information she obtains from her friends because she may never get the opportunity to verify the relevant payoffs for herself.<sup>2</sup>

Indirect payoff information may also be obtained from public sources. In using such information, the agent may discount it if she believes it to overrepresent positive outcomes.<sup>3</sup> In the event that the agent has, perhaps partial, knowledge of the underlying payoff structure, indirect payoff information may be (partly) obtained through introspection.

Experimental evidence suggests that indirect payoff information affects choices. Duffy and Feltovich (1999) show that having indirect payoff information has a systematic effect in the best-shot game and the ultimatum game. Grosskopf et. al. (2000) construct a series of experiments in which indirect payoff information has a significant effect on choice. This effect sometimes improves choice and sometimes worsen it. Experimental evidence also suggests that payoffs obtained from chosen actions and experienced directly may be treated differently from indirect payoff information obtained regarding actions which are not chosen. Camerer and Ho (1999) find that indirect payoff information (or “foregone payoffs”) are typically not given as much weight as actually experienced payoffs in updating scores. Odean (1999) suggests different accounting for unrealized versus realized losses as a possible explanation of investment patterns of individual investors. Such asymmetric effects are also found by Conley and Udry (2001).

We assume that  $D$  is monotonically increasing. Hence, distorted payoffs preserve the order of actual payoffs. Note also that we have assumed that  $D$  is not action dependent, hence it represents the agent’s general attitude towards the source of information (regarding all unplayed actions) rather than her attitude towards the action itself.

Some special cases of  $D$  are of particular interest. In the event that the agent does not have any information about the payoffs from unplayed actions, it would be natural to set  $D(x) = K$  for all  $x$ .<sup>4</sup> In this case,  $K$  serves as

---

<sup>2</sup>Note that verifying the information is not an easy task as the agent does not observe the state of the world, neither does she know the payoff function. Therefore, differences in payoffs due to differences in the realized states at the times the action was chosen may serve as evidence supporting the existence of a distortion.

<sup>3</sup>For example, it is often said that we only observe the payoff obtained by successful companies and executives and not of those who have performed poorly.

<sup>4</sup>Alternatively, the agent may choose not to use any such indirect information, and only

an (exogenous) aspiration level of the individual. An aspiration level of zero is often used in the literature, and especially in models of reinforcement learning (e.g. Börgers and Sarin (1997) and Erev and Roth (1998)).

The case of  $D$  being the identity function, i.e.,  $D(x) = x$  for all  $x$ , represents no distortion. This is the assumption used, for example, in the fictitious play algorithm since the agent is assumed to know the payoff function and observe the action of his opponent, so that she can deduce the payoffs of unplayed actions. If  $D(x) = \delta x$ , then we obtain the distortion function estimated by Camerer and Ho (1998). Inflating ( $\delta > 1$ ) and deflating ( $0 < \delta < 1$ ) positive and negative payoffs has distinct effects. For example, when inflated, positive payoffs become more attractive whereas negative payoffs become less attractive. A more general form of the distortion function arises when it is assumed to be a positive affine transformation of the true payoffs, i.e.,  $D(x) = \beta + \delta x$ . This form yields sharp conclusions in our main result as it preserves the order of expected payoffs.

It is important to observe that we assume the agent obtains the correct indirect information.<sup>5</sup> In this paper we study the consequences on choice of the manner in which the individual distorts indirect payoff information. As discussed above, there is evidence that such distortions are present and that alternative methods of modelling it have so far been proposed in the literature. Our approach is to specify a general form of the distortion function which still allows us to obtain sharp, and interesting, results regarding its effect on choice.

We now discuss the manner in which the agent “counts” the experience she has had with an action. We begin by stating this formally. Let  $N_t^i$  denote the *subjective amount of experience* the agent has had with action  $a^i$  upto time  $t$ . Then,

$$N_{t+1}^i = \begin{cases} \rho N_t^i + 1 & \text{for } a^i = a_t \\ \rho N_t^i + \gamma & \text{for } a^i \neq a_t \end{cases}$$

Hence, the subjective amount of experience the agent has with an action in any period depends on whether it was chosen or not, with all unchosen actions being treated symmetrically. The parameter  $\rho \in [0, 1]$  measures the rate at which the agent discounts her previously accumulated experience.

---

update the score for the chosen action.

<sup>5</sup>While we do not study the effect of mis-specified information regarding the payoffs of unplayed actions, the distortion can be interpreted as representing such information.

We shall suppose that the agent discounts the previously accumulated of all actions at the same rate  $\rho$ . When  $\rho = 1$ , the agent does not discount the previously accumulated experience. Such a value of  $\rho$  seems appropriate if the agent believes her environment is stationary and so past experience with an action is just as useful or informative as current experience.<sup>6</sup>

If  $\rho < 1$  the agent discounts her previously accumulated experience with all actions. Hence, she views the current experience as more valuable than experience accumulated earlier. A  $\rho < 1$  seems appropriate if the agent believes that her environment is changing, and so previously accumulated experience becomes less valuable with time. When  $\rho = 0$  the agent completely discounts previous experience. It may be appropriate in the extreme case that the environment changes very rapidly. Alternatively, this value may represent an individual with severe memory limitations.

We assume that the agent augments her subjective measure of experience for the action chosen in a period by one, and subjective measure of experience of unplayed actions in any period by  $\gamma \in [0, 1]$ . Intuitively, as the agent does not have direct experience with unplayed actions, she may treat the passing period as providing some, but not necessarily complete, experience with unplayed actions. A smaller value of  $\gamma$  indicates that the agent views the current period as providing lesser experience. If the action receives no indirect payoff information a value of  $\gamma = 0$  may be thought suitable, whereas if she believes she has received completely accurate indirect payoff information a value of  $\gamma = 1$  may be appropriate.

We may now state how the agent updates her scores. The score for any action  $a^i$  is updated using information from the previous score, the perceived payoff of the action and the subjective amount of experience that agent has had with that action. The measure of experience with an action is used to give weights on the previous score and the currently perceived payoff from the action. Formally,

$$S_{t+1}^i = \begin{cases} \frac{\rho N_t^i}{N_{t+1}^i} S_t^i + \frac{1}{N_{t+1}^i} \tilde{\pi}^i(\omega_t) & \text{for } a^i = a_t \\ \frac{\rho N_t^i}{N_{t+1}^i} S_t^i + \frac{\gamma}{N_{t+1}^i} \tilde{\pi}^i(\omega_t) & \text{for } a^i \neq a_t \end{cases}.$$

That is, the score in a period is a convex combination of the previous score and the perceived payoff in the current period for played and unplayed actions, where the weights on the previous score and the perceived payoff are

---

<sup>6</sup>Alternatively, estimates of such value for  $\rho$  may be indicative of such a belief.

in accordance to the experience the agent has had so far with this action.<sup>7</sup>

The score gives a ranking of each action in terms of its performance in the past. Performance in the past being evaluated both from directly experienced payoffs and from indirect payoff information. The current scores of the actions are used to select among them, implying that past performance of an action is deemed relevant to evaluate its future performance.

We now turn to describe the behavior rule the agent uses to select among the actions on the basis of their scores. Denote the behavior rule by  $\varphi = (\varphi_1, \varphi_2, \dots)$ , where  $\varphi_t(S_t)$  is the behavior rule at time  $t$  and  $S_t$  is the vector of scores at time  $t$ . We first assume that at each period the agent chooses the action with the highest score, i.e., the agent is myopic. Formally, for every  $t$ ,  $\varphi_t(S_t) = a^i$  for  $i \in \arg \max_{j=1, \dots, I} S_t^j$ .

The assumption that choice is myopic is very common in adaptive learning models. It is, for example, adopted in Cournot learning (see, Fudenberg and Levine (1998)), fictitious play (Brown (1950)), payoff assessment learning (Sarin and Vahid (1999)), case based decision making (Gilboa and Schmeidler (1996)), etc. Apart from the intuitive plausibility of the assumption of myopic behavior when the agent has little information about her environment, formal justifications of it have been provided in different contexts by Ellison (1997) and Sonsino (1999). Sonsino provides a justification in non-strategic contexts, in which the agent has a “large amount” of subjective uncertainty. In contrast, Ellison explains why myopic behavior may be an optimal response in strategic contexts. This assumption is also frequently used in applied work (see, for example, Conley and Udry (2001)).

We now briefly mention the different updating rules “nested” in the adaptive learning model developed in this paper. Consider the case where  $D(x) = I$  and  $\gamma = 1$ . In this situation, the agent correctly evaluates payoffs from unplayed actions and treats the current period as providing full experience also from such actions. Such assumptions are standard in “belief based” models. If we additionally assume that  $\rho = 1$  and initial scores correspond to the expected payoffs given prior beliefs, the model specializes to fictitious play in decision problems. If  $\rho = 0$ , then Cournot learning is obtained.

Now consider the case where  $\gamma = 0$ . In this case the agent does not utilize indirect payoff information. This is often an assumption made in “reinforcement learning” models. If we additionally suppose that  $\rho = 1$ , so

---

<sup>7</sup>Note that the agent only uses information regarding the performance of the action itself when updating its score. That is, information doesn’t “spillover” across actions.

that the agent does not discount previously accumulated experience with the actions, then scores measure the time-average performance of each action. If  $\rho < 1$  is assumed, then previously accumulated information with actions is discounted, and such a model is asymptotically very much like the “payoff assessment” model.

The experience-weighted attraction learning model of Camerer and Ho (1998) also nests several models and is similar to our model in several respects. In both models, scores (called “attractions” by Camerer and Ho) are updated using the experience weights of the actions and indirect payoff information is allowed. There are, however, significant differences between the two models. Firstly, in our model experience weights are action specific whereas in the basic EWA model they are not. As a result, the basic EWA model leaves out the updating rule where the agent tracks the actions’ average payoffs along their history of play. While Camerer et. al. (1999) extends the basic EWA model to allow, among other things, for action-specific experience counters, the time-average updating rule is still left out since the equivalent to our  $\gamma$  parameter, which facilitates the action specific counters, is coupled with the discounting parameter of past experience.<sup>8</sup> Secondly, the specification of perceived payoffs in Camerer and Ho which multiplies the objective payoff by  $\delta$ , sets the exogenous reference point automatically to zero, making the rule possibly more sensitive than needed to payoff transformations. Using a (positive) affine transformation for distortions, as in our model, allows for possibly different reference points for different scales of payoffs. Lastly, while Camerer and Ho couple this updating procedure of the scores with probabilistic choice, we focus on deterministic (or almost deterministic) choice of the action which is ranked highest according to the scores.

### 3 Indirect Payoff Distortions

Our main result in this section focuses on the effect of incorporating distortions of indirect payoff information on choice. For analytical tractability, we suppose that previously accumulated experience is not discounted. That is, we restrict the analysis to  $\rho = 1$ . Consequently, the experience counter

---

<sup>8</sup>Specifically, if indirect experience counts for less than a period it mandates that past experience is discounted. As a result, averaging over the stream of perceived payoffs with equal weights is ruled out.

of  $a^i$  upto time  $t$ ,  $N_t^i$ , is equal to the number of periods action  $a^i$  has been played up to time  $t$  and  $\gamma$  times the number of periods it has not been played. Moreover, the score at time  $t$  is an average of the stream of perceived payoffs, where all observations receive equal weights, equal to  $1/N_t^i$ . For (distorted) indirect payoffs to be incorporated into the ranking of alternative strategies, and hence for them to have an impact on choice, the subjective per-period experience accumulated with unplayed actions,  $\gamma$ , must be positive. Our result concerns all  $\gamma \in (0, 1]$ . Fictitious play, when it is restricted to decision problems, is a special case where  $\rho = \gamma = 1$ ,  $D = I$ , and initial scores are interpreted as expected payoffs given prior beliefs.

The following notation and definitions prove useful in this section. The objective expected payoff of  $a^i$  is denoted by  $\bar{\pi}^i = \sum_{\omega} \mu(\omega) \pi^i(\omega)$  and  $\bar{\pi} = (\bar{\pi}^1, \dots, \bar{\pi}^I)$ . We shall assume that all  $\bar{\pi}^i$  are distinct and finite. Without loss in generality, order the actions so that  $\bar{\pi}^1 > \dots > \bar{\pi}^I$ . Let  $\bar{D}^i = \sum_{\omega} \mu(\omega) D(\pi^i(\omega))$  and  $\bar{D} = (\bar{D}^1, \dots, \bar{D}^I)$ . Denote by  $\bar{R}$  the ordered vector that merges  $\bar{\pi}$  and  $\bar{D}$ .

Lemma 1 identifies the limit choice frequencies. Roughly speaking, it shows that if the agent were to converge to play more than one action with positive frequency in the limit then the scores of all such actions must be the same and above the scores of all actions played with zero frequency, in the limit. It also identifies situations in which a unique action is selected in the limit and when this action is the expected payoff maximizing one. The proof of the lemma is constructive. It provides an algorithm for simultaneously finding the limit scores and the frequencies of play. Hence, it also establishes the existence of a limit point.

Let the function  $S^i : [0, 1] \rightarrow R$  be defined as follows:

$$S^i(\alpha) = \frac{\alpha \bar{\pi}^i + (1 - \alpha) \gamma \bar{D}^i}{\alpha + (1 - \alpha) \gamma}.$$

Consider the score of action  $a^i$  that had been played with frequency  $\alpha$  up to some finite but large time  $t$ . The score  $S_t^i$  would approximately equal  $S^i(\alpha)$  because a fraction  $\alpha$  of the time the objective payoffs were accumulated, averaging roughly  $\bar{\pi}^i$ , and a fraction  $(1 - \alpha)$  of the time the  $\gamma$  weighted expected distorted payoffs,  $\gamma \bar{D}^i$ , are accumulated. That is, the accumulated perceived payoffs for action  $a^i$  is approximately  $S_0^i + (\alpha \bar{\pi}^i + (1 - \alpha) \gamma \bar{D}^i) t$ .<sup>9</sup>

<sup>9</sup>Clearly, if payoffs from  $a^i$  are deterministic and equal to  $\bar{\pi}^i$ , and the frequency upto any time  $t$  were given by  $\alpha_t$  then this holds exactly for all  $t$ .

The measure of perceived experience with  $a^i$  at such a time is given by  $N_t^i = N_0^i + (\alpha + (1 - \alpha)\gamma)t$ . In the limit as  $t \rightarrow \infty$ , the effect of the initial score  $S_0^i$  and initial experience  $N_0^i$  vanish and the score is approximated by  $S^i(\alpha)$ .

**Lemma 1** *A vector of limit frequencies of play,  $\alpha^i$   $i = 1, \dots, I$ , is a solution to the following system of equations*

$$S^i(\alpha^i) = S^j(\alpha^j) \text{ for all } a^i, a^j \in H^* \subset \{a^1, \dots, a^I\} \quad (1)$$

$$\alpha^i = 0 \text{ and } \bar{D}^i < S^j(\alpha^j), \text{ for } a^i \notin H^*, a^j \in H^* \quad (2)$$

$$\alpha^i \geq 0, \sum_{i=1}^I \alpha^i = 1. \quad (3)$$

For  $D(x) = c + dx$ ,  $d > 0$  the solution has the following characteristics: (i) When  $\bar{\pi}^1 \geq \bar{D}^1$ ,  $H^*$  consists of a single action  $a^1$  such that  $\bar{\pi}^1 > \bar{D}^1$ . In particular,  $H^*$  is not necessarily unique; (ii) When  $\bar{\pi}^1 < \bar{D}^1$ , there exists a unique solution to the system where  $H^* = \{a^1, \dots, a^k\}$  for some  $1 \leq k \leq I$ .

The algorithm used in the proof operates in the following manner. If the expected payoff maximizing action is deflated (in expectation) when not played, then all other unplayed actions have an even lower expected perceived payoff (in expectation). Consequently, all actions with an expected payoff higher than the deflated expected perceived payoff of  $a^1$  are candidates for a limit point depending on initial conditions and the realizations of the stochastic environment. Once such an action is played its score gets closer to the objective expected payoff (in expectation) while all the other scores go towards the expected perceived payoffs which are below the deflated perceived payoff of the expected payoff maximizing action. The process is reinforcing. Consequently, the frequency of choosing this action converges to one.

When the opposite is true, i.e., when the expected perceived payoff of the expected payoff maximizing action and at least one more action is inflated relative to the true objective expected payoff, play must involve a mix of actions. Once an action is played frequently, its score gets close to its objective expected payoff while the other scores get close to the perceived expected payoffs which are inflated and hence possibly above the chosen action's score. Consequently, choice must eventually turn away from each single action. As scores of unplayed actions rise relative to their expected

payoff we can show that frequencies of choice settle down to values that give equal scores to a subset of actions, where each score is a weighted average of objective and distorted expected payoffs.

The following two corollaries of Lemma 1 are straightforward and collect some of its important implications.

**Corollary 2** *When the first two elements of  $\bar{R}$  are  $\bar{\pi}^1$  and  $\bar{D}^1$ , in whatever order, the unique limit point is the expected-payoff maximizing action.*

Suppose that the distortion function is the identity and  $\gamma = 1$ . Then, the learning rule is equivalent to fictitious play with initial scores interpreted as the agent's expected payoff for each action given her prior belief about the environment. Corollary 2 points out that the unique limit point in this case is the expected payoff maximizing action. This can easily be established directly. Corollary 2 also shows that the expected payoff maximizing action is the unique limit whenever the objective and distorted expected payoffs of the optimal action are ranked at the top (in whatever order), hence providing an upper bound on the amount of distortion allowed with out affecting the optimality of the behavior.<sup>10</sup>

**Corollary 3** *When  $D(x) = K$ , any action  $a^i$  such that  $\bar{\pi}^i > K$  is a limit point. If  $K > \bar{\pi}^1$ , then  $H^*$  includes all actions.*

Consider a constant distortion function, and suppose the constant,  $K$ , is relatively low. All unplayed actions seem (equally) disappointing. Depending on initial conditions and the stochastic realization of states, Corollary 3 states that play can converge to any action whose objective expected payoff above this (disappointing) constant level. If, on the other hand,  $K > \bar{\pi}^1$ , play will asymptotically involve choosing all actions with positive frequency. Once any action has not been played for a while, so that its frequency of play to date becomes low, its score gets closer to  $K$  and hence this action will be revisited. Suppose that  $\bar{\pi}^1 > 0$  and  $K = 0$ , then play converges only to actions with a positive expected payoff. If, in addition,  $\gamma = 1$ , behavior exhibits the following form of aversion to losses: Any action with  $\pi_{\min}^i > 0$  is absorbing, that is, if an action that ensures only gains is ever chosen it will never be abandoned.<sup>11</sup> This happens since unplayed actions

---

<sup>10</sup>Note that this is true for any manner in which the experience with unplayed is counted, i.e. for all  $\gamma \in (0, 1]$ .

<sup>11</sup>This result holds as well when  $\rho < 1$ .

are in fact averaged with zero using the same weights as those used by the played actions.<sup>12</sup>

The classification of the limit points, depending on the relationship between  $\bar{\pi}^1$  and  $\bar{D}^1$ , relies on the assumption that the distortion function  $D$  preserves the order of the expectations, i.e., that  $\bar{\pi}^i > \bar{\pi}^j$  implies  $\bar{D}^i > \bar{D}^j$ . Imposing that this order is preserved over all possible payoff matrices and all possible probability distributions over states of the world requires to restrict  $D$  to be a positive affine transformation. In the absence of this assumption, while the system of equations remains the same, the nature of the solution might be different as demonstrated in the following example.<sup>13</sup> Suppose there are three actions and three equal probability states of the world. The payoffs to the actions are as listed in the table with the distorted payoffs in brackets:<sup>14</sup>

	$\omega_1$	$\omega_2$	$\omega_3$
$a^1$	18 (9)	0 (0)	3 (3)
$a^2$	0 (0)	9 (18)	6 (18)
$a^3$	0 (0)	5 (8)	9 (18)

Example 1

The solutions to the system characterized in the lemma is  $H^* = \{a^1\}$  or  $H^* = \{a^2, a^3\}$ . That is, play can converge to either the expected payoff maximizing action ( $a^1$ ) or to a mix of the other two actions, where  $a^2$  is played (asymptotically) 2/3 of the times and  $a^3$  is played the rest of the time.

Next we show that the frequencies of play converge to a limit point identified in Lemma 1, and, in particular, we rule out the possibility of limit cycles.

**Proposition 4** *The limit frequencies  $\alpha^i$ ,  $i = 1, \dots, I$ , converge a.s. to a limit point characterized in Lemma 1.*

<sup>12</sup>In the case of  $\rho < 1$ , if there is an action with  $\pi_{\min}^i > 0$ , then any action with  $\pi_{\min}^k < 0$  will be eventually abandoned. Consequently, if the minimal payoff of the expected-payoff maximizing action is even slightly negative, it will eventually be abandoned regardless of initial conditions. If  $\pi_{\min}^i < 0$  for all actions  $a^i$ , every action is chosen infinitely often.

<sup>13</sup>The arguments used in the algorithm to partition the actions remain true for consecutive subsets of actions.

<sup>14</sup>Note that this corresponds to a (weakly) monotonic distortion.

As it is the case for many learning rules, establishing that play indeed converges is not always an easy task. Convergence in the first case covered by Lemma 1 is straightforward to show. Consider now the second case, where behavior “cycles” among a subset of actions. For the actions in this limit subset, the scores are monotonically decreasing in the frequencies. Intuitively, if an action is played too frequently at some point, its score must go below the equating limit score, while there must be another action whose score goes above. The latter is then chosen more frequently, pulling the scores of this subset of actions closer together. It is left to be shown that actions which are supposed to be played with zero frequency are indeed abandoned since from some point onwards the scores of all other actions are above theirs and since the weight of each observation goes to zero there could not be a sequence of good draws that will sway that. Moreover, in order to rule out the possibility of limit cycles one must show that the maximal score at any point of time, belonging to the chosen action, cannot be too far away from the scores of the other actions that are supposed to be played. Therefore, there cannot be increasingly long cycles of play, and so the frequencies of play converge and hence they must converge, as well as the scores, to the unique limit point identified in Lemma 1. Ruling out the possibility of increasingly long cycles of play leads to many of the complications in the proof. We relegate the details to the appendix.

## 4 Continuity in $\rho$

In this section we point to a sharp discontinuity in behavior, depending on whether  $\rho = 1$  or  $\rho < 1$ , in the event that the agent obtains no experience from unplayed actions ( $\gamma = 0$ ).

Some additional notation is useful for the analysis of this section. Let  $\pi_{\min}^i$  denote the minimum payoff of action  $a^i$ . Then the maxmin action,  $a^{\max\min}$ , is the action with the highest minimal payoff, i.e., it solves  $\arg \max_i \pi_{\min}^i$ . We assume that all minimum payoffs are distinct, so  $a^{\max\min}$  is unique. Initial scores of the agent are called realistic if, for each action, they are (weakly) above the action’s minimal payoff, i.e.,  $S_0^j \geq \pi_{\min}^j$  for all  $j$ .

Suppose  $\gamma = 0$  and  $\rho = 1$ . Then, previously accumulated experience is not discounted and, hence, each period of experience with an action chosen earlier is given the same weight as given to the experience from the currently chosen action. Consequently, the score of an action is equal to the time-

average of the payoffs it has yielded when taken. Since the score of an action only updates when it is chosen, payoffs experienced in earlier periods may influence (irreversibly) the choice of actions. This also forces choice to converge to a single action. The score of this action converges to its objective expected payoff, while the scores of other actions must remain lower and unchanged since the last time each such action was taken. Note that, choice can settle to any action depending on initial scores and the realization of payoffs.

When  $\gamma = 0$  and  $0 \leq \rho < 1$ , the agent places positive weight bounded away from zero on the current payoff while the scores of unchosen actions remain unchanged.<sup>15</sup> In this case, choice cannot converge to an action that does not give at least the maxmin payoff because if it did the agent would eventually experience a long enough sequence of a worse payoff from any other action she might choose. As she gives the current payoff a weight bounded away from zero, while the scores of unplayed actions remain unchanged, such a sequence of bad payoffs ensures that she will eventually abandon such an action. Hence play cannot converge to an action that does not give the maxmin payoff. While it is clear that once the maxmin action is chosen it is chosen asymptotically, it remains to show that play indeed selects it at least once. The following result proves this and gives the precise effect of the initial scores.

**Proposition 5** *Suppose  $\gamma = 0$ . (i) If  $0 \leq \rho < 1$ , then, along any path of play, the agent converges to the action with the highest minimal payoff among all actions taken along the path. If initial scores are realistic then choice converges to  $a^{\max \min}$ . (ii) If  $\rho = 1$ , the agent converges to playing a single action, the identity of which depends on the initial condition and the stochastic realization of the states.*

The result reveals the discontinuity in behavior as  $\rho$  approaches one. For any  $\rho < 1$ , choice converges to the maxmin action whereas when  $\rho = 1$  this is not true. It may be shown that this discontinuity is robust to the introduction of random experimentation that decays at a rate of  $1/t$  and behavior that is (asymptotically) myopic.<sup>16</sup> This is in contrast to the case of  $\rho = 1$ , where in the presence of random experimentation and asymptotic myopia the agent is assured to converge to the expected payoff maximizing

---

<sup>15</sup>When  $\rho = 0$ , the score and an action is equal to the most recent payoff it received.

<sup>16</sup>For a precise definition of this term see Fudenberg and Kreps (1993).

action, since this amount of experimentation is just enough to ensure that each action is taken infinitely often so that its score must converge to its expected payoff.

## 5 Summary and Conclusions

In this paper we studied a model of adaptive learning in which an agent may obtain indirect payoff information regarding some actions. This indirect information is allowed to be treated differently from directly experienced information. Payoff information from an action is combined with a measure of the experience the agent has had with the action to provide a score concerning the attractiveness of the action. The measure of experience is itself allowed to depend on whether the information was obtained directly or not, and is allowed to decay over time. The action with the highest score is assumed to be chosen.

Our analysis of the model provides insight into how distortions of indirect payoffs affect behavior. In particular, we were able to give precise conditions under which behavior cycles among actions or “locks” into a particular action. We also point out a sharp discontinuity in behavior with respect to the degree of discounting of previously accumulated experience.

One particular extension of our work involves consideration of a nondeterministic choice rule. There are two ways in which this may be incorporated. First, we may suppose that the agent experiments at random with actions other than the one with the highest score. Specifically, we may assume the rate of experimentation dies down over time in such a way that the choice procedure to be asymptotically myopic (see, for example, Fudenberg and Kreps (1993)). We believe both our main results are robust to this extension. Proposition 4 only depends on the expected score, which in turn depends on the frequencies of choice. Hence, as long as experimentation dies out, it should not have an impact. As for the discontinuity as  $\rho$  approaches one; events which are not driven by intended choice, i.e., experimentation, happen with a probability that declines to zero while the probability of events which depend on the environment remains constant, and in particular bounded away from zero. Therefore, the latter events must affect the results much more than the former.

An alternative nondeterministic choice rule is probabilistic choice, for example, with scores entering a logit function (see, for example, Camerer

and Ho (1999)). The analysis of this case would require the application of stochastic approximation techniques. We postpone the investigation of this case for future work.

## 6 Appendix

**Proof.** Lemma 1.

Examination of the equations reveals that Part (a) is immediate given that the limit scores are a convex combination of objective and perceived expected payoffs. For part (b), the following algorithm will establish existence and uniqueness of a solution for the system (1)-(3). Recall the definition of  $S^i$ :

$$S^i(x) = \frac{x\bar{\pi}^i + (1-x)\gamma\bar{D}^i}{x + (1-x)\gamma}.$$

The function is monotonically decreasing (increasing) in  $x$  if  $\bar{D}^i > (<)\bar{\pi}^i$ . In Step (1), if  $S^1(1) = \bar{\pi}^1 > \bar{D}^2$  then we are done and the solution is  $\alpha_*^1 = 1, \alpha_*^j = 0$  for  $j > 1$ . Otherwise, move to Step (2); since  $S^1$  is monotonically decreasing there exists  $\beta_2 \in [0, 1)$  such that  $S^1(\beta_2) = \bar{D}^2$ . Monotonicity of  $S^1, S^2$  implies that for any  $\beta \in [\beta_2, 1]$  there is a split  $(\alpha_\beta^1, \alpha_\beta^2)$  such that  $\alpha_\beta^1 + \alpha_\beta^2 = \beta$  and  $S^1(\alpha_\beta^1) = S^2(\alpha_\beta^2)$ . Denote the common score as a function of the probability to be divided between the first two actions by  $S(\beta)$ . If  $S(1) > \bar{D}^3$  then we are done and the solution is  $\alpha_*^1 = \alpha_{\beta=1}^1, \alpha_*^2 = \alpha_{\beta=1}^2, \alpha_*^j = 0$  for  $j > 2$ . Otherwise, continue the process in the same manner where, in step  $k$ ,  $S(\beta)$  is the common score when there is a probability  $\beta$  to be split between the first  $k$  actions. The algorithm will stop since there is a finite number of actions. It must yield one (and only one) solution that satisfies the requirements, i.e., that each action played a positive frequency of the time must have the same limit score and all unplayed actions have a limit score which is lower than the common score of played actions. ■

**Proof.** Proposition 4.

We first prove the proposition for the deterministic case, i.e., suppose that the payoff to action  $a^i$  in each state of the world is just  $\bar{\pi}^i$ . Also, suppose that  $S_0^i = 0, N_0^i = 0$ . Consider first part (a) where  $\bar{D}^1 < \bar{\pi}^1$ . An action  $a^i$  with  $\bar{\pi}^i < \bar{D}^1$  is eventually abandoned since its score must be below the score of  $a^1$  from some point onwards. Each action with  $\bar{\pi}^i > \bar{D}^1$  is a potential limit point. Suppose that such an action  $a^i$  has the maximal score at some large  $T$ . This action is chosen and by monotonicity its score increases while the

scores of other candidate actions decrease. This repeats and consequently, the frequency in which this action is played converges to one.

Next we prove part (b). Define the function

$$L^i(\alpha^i) = \int_{\alpha_*^i}^{\alpha^i} (S^i(x) - S_*^i) dx ,$$

where  $S_*^i$  and  $\alpha_*^i$  are the limit score and frequency (correspondingly) of action  $a^i$  characterized by the Lemma. For the actions with  $\bar{D}^i > \bar{\pi}^i$  the function  $S^i$  is monotonically decreasing, all other actions will be abandoned at a bounded time. The function  $S^i$  is also Lipschitz continuous;  $|S^i(x) - S^i(x')| \leq K|x - x'|$ . Note that  $S^i(\alpha_t^i)$  is the score of action  $a^i$  at time  $t$  when payoffs are deterministic and initial conditions are normalized. Let us examine the difference

$$L^i(\alpha_{t+1}^i) - L^i(\alpha_t^i) = \int_{\alpha_t^i}^{\alpha_{t+1}^i} (S^i(x) - S_*^i) dx .$$

If action  $a^i$  is played in period  $t$  then  $\alpha_{t+1}^i - \alpha_t^i = (1 - \alpha_t^i)/(t + 1)$  and so

$$L^i(\alpha_{t+1}^i) - L^i(\alpha_t^i) \geq \frac{(1 - \alpha_t^i)}{t + 1} (S^i(\alpha_t^i) - S_*^i) - \left( \frac{1 - \alpha_t^i}{t + 1} \right)^2 K .$$

If action  $a^i$  is not played at time  $t$  then  $\alpha_{t+1}^i - \alpha_t^i = -\alpha_t^i/(t + 1)$  and

$$L^i(\alpha_{t+1}^i) - L^i(\alpha_t^i) \geq \frac{-\alpha_t^i}{t + 1} (S^i(\alpha_t^i) - S_*^i) - \left( \frac{-\alpha_t^i}{t + 1} \right)^2 K .$$

Let  $\alpha$  denote the vector of frequencies  $(\alpha^1, \alpha^2, \dots, \alpha^I)$ ,  $L(\alpha) = \sum_{i=1}^I L^i(\alpha^i)$  and  $a^{i(t)}$  is the action played at time  $t$ . Then, using the bounds derived above,

$$\begin{aligned} L(\alpha_{t+1}) - L(\alpha_t) &\geq \frac{(1 - \alpha_t^{i(t)})}{t + 1} (S_t^{i(t)} - S_*^{i(t)}) - \sum_{j \neq i(t)} \frac{\alpha_t^j}{t + 1} (S_t^j - S_*^j) - O\left(\frac{1}{(t + 1)^2}\right) \\ &\geq \sum_{j=1}^I \frac{\alpha_t^j}{t + 1} (S_t^{i(t)} - S_*^{i(t)} - S_t^j + S_*^j) - O\left(\frac{1}{(t + 1)^2}\right) \end{aligned}$$

where the second inequality is implied by  $\alpha_t^{i(t)} = 1 - \sum_{j \neq i(t)} \alpha_t^j$ . Let

$$v_t = \sum_{j=1}^I \alpha_t^j (S_t^{i(t)} - S_*^{i(t)} - S_t^j + S_*^j)$$

then

$$v_t \leq (t+1)(L(\alpha_{t+1}) - L(\alpha_t)) + O\left(\frac{1}{t+1}\right).$$

We want to show that

$$-\infty < \sum_{t=1}^{\infty} \frac{v_t}{t} < \infty .$$

The above upper bound on  $v_t$  and some algebra manipulations imply that  $\sum_{t=1}^{\infty} \frac{v_t}{t} < \infty$ . We are left to show that the sum is bounded from below. Let

$$v_t = v_t^{H^*} + v_t^{NH^*} := \sum_{j \in H^*} \alpha_t^j (S_t^{i(t)} - S_*^{i(t)} - S_t^j + S_*^j) + \sum_{j \notin H^*} \alpha_t^j (S_t^{i(t)} - S_*^{i(t)} - S_t^j + S_*^j)$$

where each of the elements of  $v_t^{H^*}$  is non negative since  $S_t^{i(t)} - S_t^j \geq 0$  by the choice of  $a^{i(t)}$  and  $S_*^j \geq S_*^{i(t)}$  for  $j \in H^*$ . Also, each element of  $v_t^{NH^*} \geq -\max_{j \notin H^*} |S - \bar{D}^j|$ . There exists a time  $T$  such that only actions such that  $j \in H^*$  are played thereafter; suppose not, then there exists an action  $a^j, j \notin H^*$  which is played infinitely often. Hence its score is the maximal infinitely often and it is lower than  $S$  (by the construction of the solution). Since the scores of actions in  $H^*$  are monotonically increasing it must be that  $\alpha_t^k > \alpha_*^k$  for all  $a^k \in H^*$ , which is a contradiction since frequencies should sum to one. Consequently,

$$\alpha_t^j \leq \frac{T}{t} \alpha_T^j \leq \frac{T}{t} \text{ for all } t \geq T \text{ and } j \notin H^*,$$

and so

$$\sum_{t=1}^{\infty} \frac{v_t}{t} \geq \sum_{t=1}^{\infty} \frac{v_t^{NH^*}}{t} \geq \sum_{j \notin H^*} \left( -\sum_{t=1}^T \frac{\alpha_t^j}{t} \max_{j \notin H^*} |S - \bar{D}^j| - \sum_{t=T+1}^{\infty} \frac{T}{t^2} \max_{j \notin H^*} |S - \bar{D}^j| \right) > -\infty.$$

Next we show that the fact that  $-\infty < \sum_{t=1}^{\infty} \frac{v_t}{t} < \infty$  which we have just established implies

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T v_t = 0 .$$

Let  $w_T = \sum_{t=T}^{\infty} \frac{v_t}{t}$ . The finiteness of  $\sum_{t=1}^{\infty} \frac{v_t}{t}$  implies that  $\lim_{t \rightarrow \infty} w_t = 0$ . Algebraic manipulations show that

$$\frac{1}{T} \sum_{t=1}^T w_t = \frac{1}{T} \sum_{t=1}^T v_t + w_{T+1} .$$

Since  $\lim_{T \rightarrow \infty} w_T = 0$ , and hence also the left hand side converges to zero, it must be that  $\frac{1}{T} \sum_{t=1}^T v_t$ .

The next step is showing that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T v_t = 0 \Rightarrow \lim_{T \rightarrow \infty} \frac{\#\{1 \leq t \leq T : \alpha_t \notin \alpha_*(\varepsilon)\}}{T} = 0 \text{ for any } \varepsilon > 0,$$

where  $\alpha_*(\varepsilon)$  is an  $\varepsilon$ -neighborhood of the limit frequencies defined by the system (1)-(3). Although  $v_t$  could be negative, since  $\lim_{t \rightarrow \infty} v_t^{NH^*} = 0$ ,

$$\lim_{T \rightarrow \infty} \frac{\#\{1 \leq t \leq T : v_t \notin \underline{0}_+(\varepsilon)\}}{T} = 0 .$$

We are left to show that this implies the result for  $j \in H^*$  since for an action  $j \notin H^*$  we have established that  $\lim_{t \rightarrow \infty} \alpha_t^j = 0$ . In particular we need to show that when the product  $\alpha_t^j (S_t^{i(t)} - S_*^{i(t)} - S_t^j + S_*^j)$  is arbitrarily small it is because  $(S_t^{i(t)} - S_t^j)$  is arbitrarily small for every  $j \in H^*$ . Suppose not, i.e., there exists an action  $a^j$  with  $j \in H^*$  for which the product is small since  $\alpha_t^j$  is arbitrarily small. But then  $S_t^j > S$  since it is arbitrarily close to  $\bar{D}^j > S$ . Then, for each  $j \in H^*$  either  $\alpha_t^j$  is arbitrarily small or  $(S_t^{i(t)} - S_t^j)$  is arbitrary small and since  $S_t^{i(t)} \geq S_t^j > S$  its score is bounded away above  $S$  and so by monotonicity  $\alpha_t^j$  is bounded below  $\alpha_*^j$ . Since  $\sum_{j \in H^*} \alpha_*^j = 1$ , it implies that  $\sum_{j=1}^I \alpha_t^j < 1$  which is a contradiction. Hence

$$\lim_{T \rightarrow \infty} \frac{\#\{1 \leq t \leq T : |S_t^{i(t)} - S_t^j| < \varepsilon \text{ for all } j \in H^*\}}{T} = 1$$

and the result follows from the properties of the unique solution to the system (1)-(3) in which all these scores are equal to each other and the sum of the frequencies is one. The proof for the deterministic case is complete once we establish that convergence in Cesaro mean as shown above implies

convergence in frequencies to the same limit set. This is a direct result of the fact that the step size at stage  $t$  is  $1/t$ .<sup>17</sup>

The complete the proof note that

$$E(S_{t+1}^i - S_t^i | S_t) = \begin{cases} \frac{1}{N_t^i + 1}(\bar{\pi}^i - S_t^i) & \text{for } i = \arg \max_j S_t^j \\ \frac{1}{N_t^i + \gamma}(\gamma \bar{D}^i - S_t^i) & \text{for } i \neq \arg \max_j S_t^j \end{cases}$$

Hence the expected law of motion of the stochastic system is the same as the law of motion of the deterministic system. Since the deterministic system has a global attractor and  $\varepsilon_t^i = 1/N_t^i$  is decreasing to zero at the rate of  $1/t$ , the stochastic system must be approaching the same global attractor with probability one. ■

**Proof.** Proposition 5.

(i) Let  $0 \leq \rho < 1$ . Suppose that the individual plays strategy  $a^i$  at some time, and suppose that the individual has only ever chosen strategies  $a^k \in \tilde{A} \subset A$ , and that  $\pi_{\min}^i \geq \pi_{\min}^k$  for all  $a^k \in \tilde{A}$ . Suppose that the individual converges to a strategy  $a^k \in \tilde{A}$ ,  $a^k \neq a^i$ . Then, at some time the agent will experience a long enough run of a low enough (possibly worst) payoff  $a^k$  can give and this will ensure that  $s_T^k < \pi_{\min}^{k'} \leq s_T^{k'}$  for some  $a^{k'} \in \tilde{A}$ , in finite time  $T$ . Note also that  $\pi_{\min}^i \geq \pi_{\min}^{k'}$ . Hence, the individual cannot converge to any action other than  $a^i$ . The above argument also applies for any action  $a^k \neq a^i$  that the agent plays infinitely often. Hence, the individual cannot cycle among actions. To see that the agent can converge to  $a^i$ , it suffices to consider the situation in which the  $s^k < \pi_{\min}^i < s^i$ . At such states, which clearly have a positive probability of being reached from all other states, the individual will choose only  $a^i$ .

If  $0 \leq \rho < 1$  and the agent's initial scores are realistic then argument in the above paragraph ensures that she will converge to play the maxmin strategy, if she chooses  $a^{\max \min}$  at some time. This happens because there is always a positive probability that the scores of all  $a^k \neq a^{\max \min}$  fall below  $\pi^{\max \min}$ . When (or before) this happens, the individual will choose her  $a^{\max \min}$ .

(ii) Suppose  $\rho = 1$ . Since scores track the time-average of the history of payoffs of each action, the score of each action which is taken infinitely often must converge to its expected payoff. We assumed that the latter are distinct and the agent chooses the one with the highest score, so play must converge to a single action. ■

---

<sup>17</sup>This argument is identical to Lemma 1 in Monderer and Shapley (1996).

## 7 References

1. T. Borgers and R. Sarin (1997): “Learning through reinforcement and replicator dynamics,” *Journal of Economic Theory*, 77, 1-14.
2. T. Borgers and R. Sarin (2000): “Naive reinforcement learning with endogenous aspirations,” *International Economic Review*, 41, 921-950.
3. T. Borgers, A. Morales and R. Sarin (2001): “Expedient and monotone learning rules,” mimeo, University College London.
4. G. Cachon and C. F. Camerer (1996): “Loss-avoidance and forward induction in experimental coordination games” *Quarterly Journal of Economics* v111, 165-94.
5. C. Camerer and T. Ho (1999): “Experience weighted attraction learning in normal form games, *Econometrica*, 67, 827-874.
6. C. Camerer, T. Ho and X. Wang (1999): “Individual differences in EWA learning with partial payoff information,” mimeo, Caltech.
7. Y. Chen and Y. Khoroshilov (2001): Asynchrony and learning in serial and average cost pricing mechanisms: An experimental study,” mimeo, Michigan.
8. T. Conley and C. Udry (2000): “Learning about a new technology: Pineapple in Ghana,” mimeo, Northwestern.
9. J. Duffy and N. Feltovich (1999): “Does observation of others affect learning in strategic environments? An experimental study,” *International Journal of Game Theory*, 28, pp. 131-152.
10. D. Easley and A. Rustichini (1999): “Choice without beliefs,” *Econometrica*, 67, 1157-1184.
11. G. Ellison (1997): “Learning from personal experience: One rational guy and the justification of myopia,” *Games and Economic Behavior*, 19, 180-210.
12. I. Erev and A. Roth (1998): “Predicting how people play games: Reinforcement learning in games with a unique mixed strategy equilibrium,” *American Economic Review*, 88, 848-881.

13. N. Feltovich (2000): "Reinforcement-based vs. belief-based learning models in experimental asymmetric games," *Econometrica*, 68, 605-64.
14. D. Fudenberg and D. Kreps (1993): "Learning mixed equilibria," *Games and Economic Behavior*, 5, 320-367.
15. D. Fudenberg and D. Levine (1995): "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, 19, 1065-1090.
16. D. Fudenberg and D. Levine (1998): *The theory of learning in games*, MIT press.
17. I. Gilboa and D. Schmeidler (1996): "Case based optimization," *Games and Economic Behavior*, 15, 1-26.
18. B. Grosskopf, I Erev and E. Yechiam (2001): "Foregone with the wind," mimeo, Harvard.
19. D. Monderer and L. Shapley (1996): "Fictitious play property in games of identical interests," *Journal of Economic Theory*, 68, 258-265.
20. D. Mookherjee and B. Sopher (1994): "Learning behavior in an experimental matching pennies game," *Games and Economic Behavior* v7, 62-91.
21. Y. Nyarko (1991): "Learning in mis-specified models and the possibility of cycles," *Journal of Economic Theory*, 55, 416-427.
22. T. Odean (1998): "Are investors reluctant to realize their losses?," *Journal of Finance*, 53, 1775-1798.
23. A. Rustichini (1999): "Optimal properties of stimulus-response learning models," *Games and Economic Behavior*, 29, 244-273.
24. R. Sarin and F. Vahid (1999): "Payoff assessments without probabilities: A simple dynamic model of choice," *Games and Economic Behavior*, 28, 294-309.
25. R. Sarin and F. Vahid (1999): "Predicting how people play games: A simple dynamic model of choice," *Games and Economic Behavior*, 34, 104-122.

26. P. Seetharaman and P. Chintagunta (1998): "A model of inertia and variety-seeking with marketing variables," *International Journal of Research in Marketing*, 15, 1-17.
27. D. Sonsino (1999): "Uncertainty and the foundations of myopic behavior," mimeo Technion.
28. J. Van Huyck, R. Battalio and F. Rankin (1996): "On the evolution of convention: Evidence from coordination games," mimeo, Texas A&M University.