Parametric Adaptive Learning
(Draft)

Dana Heller
Department of Economics
University of Chicago

Rajiv Sarin
Department of Economics
Texas A&M University

January 31, 2000

Abstract

We investigate a general parametric model of adaptive learning. The model includes most of the adaptive learning procedures studied in the literature where agents optimize given their ranking over actions, perhaps allowing for experimentation. It provides a convenient parametric framework to analyze experimental data and to compare the performance of previously proposed learning hypotheses. We show that several "parameter clusters" result in qualitatively similarly behavior, hence making precise the important realtions between the di¤erent parameters. We also identify and analyze some previously uninvestigated parameter clusters which lead to empirically plausible behavior, such as "loss aversion."

1 Introduction

There is an extensive and growing theoretical literature on adaptive learning in games. All of the models posit some manner by which players rank their strategies at any point in time. Choice behavior depends on this ranking either deterministically or stochastically. Each model posits the way in which players update their ranking upon receiving new information. The analysis of the models focuses on the strategies the players converge to play over time. A vast complementary literature uses these models to organize experimental data on learning in games, and seeks to evaluate their relative performance.

The theoretical literature has greatly increased our understanding of different adaptive procedures according to which players learn in games. The experimental literature has informed us about which models are more useful in describing subjects behavior in di¤erent contexts. So far, however, there are no theoretical studies which provide a general framework in which the speci…c models of adaptive learning are "naturally" nested. Such a general model would be particularly useful for experimentally distinguishing between di¤erent learning hypotheses, as pointed out most recently by Camerer and Ho (1999).

In this paper, we study a general parametric adaptive learning model according to which players rank their strategies and update their ranking in light of the information they observe after each period of play. The three parameter model we propose nests almost all the ranking and updating procedures that have been proposed in the literature on adaptive learning, and provides a useful framework in which the relative importance of di¤erent learning hypotheses can be distinguished. Our framework also suggests new models, or "parameter clusters" of the general model, that may represent plausible and interesting learning behavior and which have hitherto not been discussed in the literature.

We analyze asymptotic properties of the three-parameter model when a player repeatedly plays a decision problem in which there are a …nite number of possible states of the world, and a …xed probability distribution (over time) on this set of states. The player does not have a model of the environment and need not know whether he is playing against nature or strategic opponent(s), or whether the environment is …xed or changes over time.

The agent associates with each action a scalar, which we refer to as the score, according to which he ranks his actions. The score could, for example, be the average payo¤ the action has historically received. After an action is taken and a state of the world is realized, the agent observes the payo¤ from the action he has taken, and maybe also observes the payo¤s from actions he did not choose. He could, for example, obtain information about unplayed actions from reading newspapers or from talking with other agents. This updating goes through three "cognitive operators" which describe how the agent incorporates this new information in the scores.

The …rst cognitive operator evaluates this period's perceived payo¤ from each action. The perceived payo¤ is used by the agent to update the score. For the action chosen this period, the perceived payo¤ is equal to the objective payo¤. For unplayed actions, the perceived payo¤ is proportional to the objective payo¤ but is not necessarily equal to it. We believe that people may distinguish between payo¤s from chosen actions, which are actually receive, and payo¤s from unchosen actions, regarding which they have only indirect experience. Although the information about the value of these untaken actions might be precise, it is the psychological attitude of the agent towards this source of information that will determine how much "weight" she gives to it. This cognitive operator is summarized in a parameter which measures the discrepancy between the objective payo¤ and the perceived payo¤.

The second cognitive operator measures the amount of subjective experience the agent has had with each action. This includes two features. First, the previously accumulated experience may decay over time. This decay may arise because the agent has a limited memory or because he believes that older information may not be as relevant as new information. Second, the experience he accumulates in the current period with each action. Here again, we believe that the decision maker may view the amount of experience obtained with played and unplayed actions di¤erently. We normalize the amount of experience accumulated with the action chosen in the current period to one, and require that it is assessed to be at least as great as the amount of experience accumulated with unplayed actions. Two parameters of our model capture these two features of the amount of experience the agent has had with her actions.

The third operator computes next period's score of an action given the score this period, its perceived payo¤, and the subjective amount of experience the agent accumulated with the action. This is done by transforming the subjective measure of experience into a pair of weights, one for the perceived payo¤ and the other for the current score. The new score is calculated by combining the current score and the perceived payo¤ using these weights.

The manner in which the agent ranks strategies and how he transforms these rankings given new information may seem very stylized. However, our three-parameter model includes all of the better known adaptive learning procedures. For instance, if initial scores represent the agents ranking of the actions prior to the learning process, the perceived payo¤ for each actions is equal to its objective payo¤, and each period counts for a unit of experience for all actions then we obtain the well known adaptive learning procedure of "…ctitious play." In the next section, after formally stating the model, we show how this model specializes to other well-known learning procedures.

Given the manner in which an individual ranks his strategies, two di¤erent classes of choice rules may be distinguished: deterministic and stochastic. According to the former, the agent chooses, at each time, the action which he ranks the highest. Such a myopic choice rule is utilized in many adaptive learning procedures including Cournot learning and …ctitious play. The stochastic decision rule we consider allows the agent to randomly experiment with actions not currently ranked the highest. However, he must eventually choose the action he ranks the best, i.e. he is asymptotically myopic. Intuitively, random experimentation and asymptotic myopia (…rst discussed by Fudenberg and Kreps (1993)) allow us to consider an agent who acquires su¢cient amount of information about all strategies and eventually becomes con…dent in his ranking.

When the individual gives each observation the same weight, we show that the asymptotic behavior of the agent depends crucially on the relationship between the "perceived per-period expected payo¤" and the objective expected payo¤. The former can be di¤erent from the latter due to the discrepancy between received payo¤s for played actions versus indirect payo¤s for unplayed action, and also because the subjective experience with the two types of actions di¤er. If the perceived per-period expected payo¤ for an unplayed action is greater than its objective expected payo¤, and there is some action for which this (in‡ated) perceived per-period expected payo¤ is greater than the objective maximal expected payo¤, it must be that the agent converges to play more than one action. Although the scores of this subset of actions converge to the same number and choice is deterministic, play looks like a mixed strategy as these actions are played a …xed proportion of the time. If the perceived per-period expected payo¤ for an unplayed action is smaller than its objective expected payo¤, then play converges to a single action, and it can be any action whose objective expected payo¤ is greater than the (de‡ated) perceived per-period expected payo¤ of the expected-payo¤ maximizing action.

The second goal of the paper is to explore the e¤ect of experimentation on the asymptotic behavior. For some simple updating procedures, such as when the score of an action is its time-average payo¤, we show that behavior in the absence of experimentation can be very suboptimal. The introduction of random experimentation and asymptotic myopia drastically changes this, and results in convergence to the expected payo¤ maximizing action. However, when the procedure is altered slightly, so that the most recent experience gets positive weight, behavior deviates dramatically from expected-payo¤ maximization in the absence of experimentation and is not altered by the introduction of experimentation.

We also identify clusters of parameters that have not received attention in the past which lead to plausible patterns of behavior. These clusters involve a discrepancy between the perceived per-period expected payo¤ for an unplayed action from its expected payo¤. For example, such a procedure can lead to behavior that is "loss averse." In particular, the …rst time the agent plays an action with a positive minimal payo¤, i.e., an action that only ensures him gains, he never switches to a di¤erent action. We are also able to get from this family of learning rules some rules that are sensitive to the size of the expected payo¤, some to its sign, while others are sensitive to the size or sign of the minimal payo¤. Also, some rules, though deterministic, involve endogenous experimentation; after a certain action has not been played for a while it looks more lucrative than it did the last time it was played, and therefore it is revisited. Given such a behavior the parameters of this learning rule can be interpreted as re‡ecting the agent's attitude regarding the nature of the environment. In particular, this can be justi…es as a sensible procedure in a changing environment, though it is suboptimal in a …xed environment.

The theoretical literature is admirably summarized in a recent text by Fudenberg and Levine (1998). Two large classes of learning models are distinguished: belief-based models and reinforcement learning models. In belief based models the agent has a well-formed, though perhaps mis-speci…ed, model of the environment. These models include Cournot learning and …ctitious play. The agent myopically optimizes given his belief regarding how others will play. These beliefs are updated upon observing how the choice of the others players evolves. In reinforcement learning models the beliefs of the agents are left unspeci…ed and only their behavior is studied. In most such models agents choose stochastically. They only use information on the payo¤ obtained from their chosen strategy in updating their behavior.

Papers by Sarin and Vahid (1999), Hopkins (1999) and Rustichini (1999) have further studied these models. Sarin and Vahid present a model in which agents only use information on the payoff from chosen actions and optimize given this information. Their model, hence, combines features of both belief and reinforcement learning models. Hopkins shows that the asymptotic properties of noisy versions of belief based models and reinforcement learning models have similar asymptotic properties. Rustichini considers full information and partial information models of reinforcement learning and discusses their optimality properties. Easley and Rustichini (1999) provide an axiomatic framework in which to model reinforcement learning models where agents observe information of all actions in every period.

Experimental studies have tested the plausibility of these two classes of learning models. These include papers by Camerer and Ho (1999), Erev and Roth (1998), Feltovich (1999) and Sarin and Vahid (1999). The paper by Camerer and Ho evaluates belief based and reinforcement learning models by considering a more general parametric form that nests these two learning hypotheses. Erev and Roth and Feltovich contrast the performance of reinforcement and belief based models in a large class of experiments and both show that learning models explain behavior better than equilibrium predictions. Sarin and Vahid show that their model performs at least as well as reinforcement and belief-based models and is much simpler to analyze.

This paper is organized as follows. The next section presents the model. Section 3 analyzes the model. Section 4 concludes and discusses some extensions of the model.

2 The Model

We suppose that the individual has a finite set of actions, $A = \{a^1, ..., a^I\}$. At each time, the individual takes an action and a state of the world is realized. We suppose that there is a finite set of possible states of the world $\Omega = \{\omega^1, ..., \omega^J\}$. Nature chooses the state of the world according to a fixed probability distribution μ which does not change over time, where μ^j gives the probability that state ω^j is selected in any period. We denote the state of the world in period t by $\omega_t \in \Omega$. The agent in the model is assumed to hold no model of the environment in which he operates. In particular, he does not postulate as to whether he is playing against nature or against a strategic opponent(s). Neither does he deliberate whether the environment is static or changing over time. He only follows an adaptive learning procedure, by which he (almost always) optimizes given the score, where the parameters of this procedure may be interpreted as representing his attitude\belief about the qualitative nature of his environment. After the individual chooses an action, nature selects a state, and the objective payoffs are realized. Denote the state of the world realized in period t by $\omega_t \in \Omega$, then the objective payoff from action a^i is denoted $\pi^i(\omega_t)$. Let $a_t \in A$ represent the action chosen in period t.

For the individual's learning behavior, however, it is not objective payoffs that are important but the perceived payoffs. The perceived payoff of action a^i at time t, $\tilde{\pi}_t^i$, is given by,

$$\tilde{\pi}_t^i = \left(\delta + (1-\delta)\, I\left(a^i; a_t\right)\right) \pi^i(\omega_t)$$

where I denotes the indicator function which takes a value of 1 when $a^i = a_t$ and is zero otherwise. Hence, perceived payoffs are equal to the objective payoff for the chosen action, but are equal to a proportion $\delta \geq 0$ of the objective payoffs for the unplayed actions.[1] If the agent does not obtain any information about the payoffs from unplayed actions then it would be natural to have $\delta = 0$. In other cases, when the agent obtains information on these payoffs he may not treat them the same as the payoff he actually receives due to some psychological factors that lead him to discount or inflate the inferred payoffs. Hence the parameter δ represents the agent's attitude towards the information regarding the possible payoffs of unplayed actions to him or his attitude towards the fact that this experience is indirect. This can be interpreted in many ways, such as uncertainty about the validity of the source of information or about the relevance of idiosyncratic components in the utility from a certain outcome (state of the world). Note that this parameter δ is not action dependent, hence it represents the agent's general attitude rather than his attitude towards the action itself. In the special case where $\delta = 1$ he treats the payoffs inferred about unplayed actions in the same way as the payoffs obtained from the played action. Hence, when $\delta = 1$ the agent makes the "correct" use of all the information he obtains.

Let the amount of experience the agent has had with an action a^i upto time t be denoted by N_t^i. The following equation describes the manner in which the agent updates the amount of experience he believes to have had with any action at the beginning of period t + 1:

[1] We suspect that all our results hold if we extend the relation between objective payoffs and perceived payoffs to be $\tilde{\pi}^i = \delta\pi^i$, for $\delta \geq 0$ (or some other order preserving transformation) but have not checked all the details.

$$N^i_{t+1} = \rho N^i_t + \gamma + (1-\gamma)\, I\left(a^i; a_t\right):$$

The parameter ρ measures the rate at which past experience decays, where $\rho \in [0;1]$. When ρ is equal to one, all observations get the same weight in the agent's score. While when $\rho < 1$ the weight put on the current perceived payoff remains uniformly bounded away from zero, though it might evolve over time. A parameter $\rho = 1$ may be interpreted as representing the agent's belief that the environment is fixed over time, therefore all observations carry the same weight, while $\rho < 1$ may represent the belief that the environment is changing hence the last observation weighs more than previous observations. The agent augments his experience counter for the action chosen in period t by one, though he is allowed to augment his experience counter by a fraction γ for an unplayed action. Intuitively, as the agent does not have direct experience with unplayed actions, although he might obtain (perfect) information regarding their performance, he may treat the passing period as providing some experience with unplayed actions. This parameter γ, like ρ, can also be interpreted as capturing the agent's attitude regarding the dynamic nature of the environment; if the environment is believed to be changing then the mere fact that a period has elapsed carries information regarding the possible value of an action. If $\gamma = 0$, the agent behaves as if he had no experience with an unplayed action this past period, whereas if $0 < \gamma < 1$, he only considers that he has had some partial experience with unplayed actions. If $\gamma = 1$ he feels he has had full experience with unplayed actions in the current period.

Suppose that agent partially discounts the payoff information he obtains from unplayed actions. Then, it seems intuitive that he may also partially discount the experience he obtains in the current period from an unplayed action. As we shall see in the next section, as long as $0 < \delta = \gamma < 1$ the agent utilizes the information he obtains in the current period "optimally," i.e., his "perceived per-period payoff" for an unplayed action is equal to its objective payoff.

Scores for any action a^i are updated using information from the previous score, the perceived payoff of the action and the subjective experience that agent has had with that action. Specifically, the agent uses his experience counter with an action to give weights on the previous score and the currently perceived payoff from the action. In particular, the score of action i in period $(t+1)$ is given as:

$$s^i_{t+1} = \frac{\rho N^i_t}{N^i_{t+1}} s^i_t + \frac{1}{N^i_{t+1}} \pi^i(\omega_t):$$

That is, the score in period $(t+1)$ is a convex combination of the previous score and the perceived payoff in the current period for the played action, where the weights on the previous score and the perceived payoff are in accordance to the experience the agent has had so far with this action. The same weights apply for unplayed actions, however, it is a convex combination only in the case that $\gamma = 1$.

So far we have discussed the manner in which the agent ranks his strategies at any time, and how new information causes the ranking to be updated. We now turn to the discuss the behavior rule the agent uses to select among the actions. Denote the behavior rule by $\varphi = (\varphi_1; \varphi_2; \ldots)$ where $\varphi_t(s_t)$ is the behavior rule at time t and s_t is the vector of scores at time t. That is, $\varphi_t(s_t) \in \Delta(A)$; where $\Delta(A)$ is the set of probability distributions over the actions. We first suppose that at each period the agent chooses (deterministically) the action with the highest assessment, i.e., the agent is myopic. Formally,

$$\begin{aligned} \varphi_t(s_t)(a^j) &= 1 \text{ for } j = \arg\max_{i=1;\ldots;I} s^i_t \\ \varphi_t(s_t)(a^k) &= 0 \text{ for } k \neq j: \end{aligned}$$

We also consider a different behavior rule which allows the agent to experiment with each action infinitely often before behaving myopically. Such a stochastic choice rule involves experimentation by the agent with possibly suboptimal actions. However, as experience accumulates, the agent's confidence in his assessments is required to grow, restricting the agent's use of inferior actions. Formally, the behavior rule is assumed to possess the following two properties. Let $\zeta_t = (\omega_1; \omega_2; \ldots)$ be the realization of states of the world up to time t.

Definition 1 Given a vector of score vectors $s = \{s_t\}^\infty_{t=0}$, we say that the behavior rule $\varphi = \{\varphi_t\}^\infty_{t=1}$ is asymptotically myopic relative to s if for some sequence of strictly positive numbers $\{\varepsilon_t\}$ with limit zero, for every t, $\varphi_t(s_t)$ comes within ε_t of maximizing the agent's payoff given the score vector s_t. That is

$$\sum_{a^j \in A} \varphi_t(s_t)(a^i)s_t(a^i) + \varepsilon_t \geq \max_{a^j \in A} s_t(a^i):$$

In the definition of asymptotic myopia we are following Fudenberg and Kreps (1993). Asymptotic myopia allows the agent to play slightly inferior actions with large probability, or he can use grossly inferior actions, relative to the scores, with very small probability, as long as the average suboptimality is getting arbitrarily small. When the score vector is derived from an updating rule, i.e., $s_t(\zeta_t)$, asymptotic myopia requires that ε-optimality holds for each ζ_t.

Definition 2 A behavior rule φ follows random experimentation if for some strictly positive number α, each action $a^i \in A$ is played with a probability not smaller than α/t at time t, i.e., $\varphi_t(s_t)(a^i) \geq \alpha/t$.

The simplest rule that satisfies asymptotic myopia and follows randome experimentation is the following:

$$\begin{aligned} \varphi_t(s_t)(a^i) &= 1 - \alpha/(|S|t) \text{ for } k = \arg\max_{a^j \in A} s_t(a^i) \\ &\text{and} \\ \varphi_t(s_t)(a^i) &= \alpha/(|S|t) \text{ for } i \neq k: \end{aligned}$$

A direct application of Borel-Cantelli lemma implies that rate of experimentation required from a behavior rule following random experimentation is sufficient to ensure that each action is played infinitely often. The main goal of introducing experimentation is to understand which patterns of behavior result from an agent that is myopic before enough experience with different actions is accumulated from patterns of behavior that are robust to this amount of experience with all actions. As the analysis will show, some learning rules, time-averaging to name one, perform poorly in the absence of experimentation but asymptotically pick the optimal outcome with this amount of experimentation. For other rules, adding experimentation is not enough to induce convergence to the action with the highest expected payoff.

We briefly mention the different learning rules nested in the parametric adaptive form presented above. In particular, when $\delta = 1$ and $\gamma = 1$ different models of belief learning are spanned by different values of ρ. The rule specializes to fictitious play when $\rho = 1$, and initial scores correspond with prior beliefs about the value of the different actions. Cournot learning is achieved for $\rho = 0$.

Reinforcement-learning type models are realized when $\delta = 0$. For example, scores that measure the time-average performance of each action are a special case where $\delta = 0; \gamma = 0; \rho = 1$: Also, averaging where the weight on current perceived payoff does not vanish, while assessments of unplayed actions remain unchanged, such as in Sarin and Vahid (1999),[2] correspond to $\delta = 0; \gamma = 0; \rho < 1$. Other reinforcement learning models which use the cumulative reinforcement learning rule, where current scores are discounted by ϕ; and the payoff of the action currently taken is added to the scores of that action, are easily included by the addition of one parameter.

Also, our stochastic decision choice rule with random experimentation and asymptotic myopia allows choice to be stochastic as is often assumed in traditional reinforcement learning models and is introduced in the stochastic version of belief-learning rules like fictitious play. We postpone the discussion of how several stochastic choice rules (e.g. stochastic fictitious play) are nested in this choice rule.

3 Analysis

Some additional notation and definitions will prove to be useful in the analysis of the model. When the individual chooses deterministically, some of the results will depend upon the initial scores of the agent. We say that the initial scores of the agent are realistic if, for each action, they are not below the lowest payoff an action can give. Initial scores are said to be pessimist if they are below the minimum payoff from an action, for all actions.

Definition 3 Initial scores are realistic if $s_0^j \geq \pi_{min}^j$ for all j. Initial scores are pessimistic if $s_0^j \leq \pi_{min}^j$ for all j.

Let π_{min}^i denote the minimum payoff that action a^i gives. Then the maxmin action $a^{\max\min}$ and the maxmin payoff $\pi^{\max\min}$ are defined as follows.

[2] Sarin and Vahid (1999) assumes fixed weights on current assessments and current payoffs, therefore, our case is asymptotically equivalent to their model. This is all that is needed to get qualitatively similar behavior.

Definition 4 $a^{\max\min}$ is the maxmin action if it gives the highest minimum payoff, i.e. it solves $\arg\max_{a_i} \pi^i_{\min}$. The maxmin payoff is the minimum payoff that $a^{\max\min}$ gives.

We shall denote the objective expected payoff of a^i by $\bar{\pi}^i$. For convenience, we assume that all $\bar{\pi}^i$ are distinct and finite. We also suppose that the minimum payoff the agent may obtain from the choice of any action a^i, $\pi^i_{\min}$ is unique. As we had assumed that all minimum payoffs are distinct, the $a^{\max\min}$ is unique. We now begin our analysis of the model. We first consider the following parameter cluster.

3.1 $\delta = 0; \gamma = 0; \rho \in [0; 1]$

This corresponds to the case where the agent does not update his scores of unplayed actions (as $\delta = 0$), and where the current period does not count as experience for unplayed actions (as $\gamma = 0$). The agent may not update his scores of unplayed actions simply because he may not know what these payoffs would have been. Hence, this case is relevant for situations where the agent does not know the payoff matrix and he does not observe the state of the world. It may also be relevant in situations where the agent knows the payoff matrix but does not update his information regarding it because of the deliberations costs involved.[3] Given that the agent does not use this period's information in updating his assessments of unplayed actions, it seems natural that he does not take into account this period's experience to update his (experience) counters for the unplayed actions.[4]

Two distinct cases arise for this parameter cluster. When $\rho = 1$, each unit of experience with an action chosen previously is given the same weight as the current experience from the chosen action. Hence, past experience is not discounted relative to the current experience. As this rule gives decreasing weight to current payoffs relative to the entire past, the payoffs the agent experiences early may influence the choice of actions and therefore future scores and consequently the action the agent converges to choose. In particular, the

[3] See Conlisk (1996) for a discussion of the importance of deliberational costs in economic decision making.

[4] Camerer and Ho (1999) refer to the case where $\delta = 0$ as the reinforcement learning case because of the minimal information the agent uses in updating her assessments. Most authors, however, define reinforcement learning in a different way, even though updating assessments in such models uses only minimal information.

score of each action a^i is the time average of the payoffs the agent has received from a^i so far. Hence, if an agent converges to choose an action a^j, its score converges to its expected payoff π^j. Proposition 1 states that choice does indeed converge in this case. However, we cannot state which action the agent converges to choose because of the importance of initial periods of play.

The other case we consider involves supposing that $0 \leq \rho < 1$. In this case, the agent always places positive weight bounded away from zero on the current payoff.[5] In this case, Proposition 1 shows that choice does converge even while assessments do not. If the agent is realistic then he converges to the maxmin action. We can also show that he converges to his maxmin action among all the actions he has ever chosen, which in return depends on the particular history of realizations. It is easy to see that a pessimist will choose only one action forever: The action that has chosen initially since it had the highest initial score.

Proposition 1 If $\rho = 1$, the agent converges to choose some action, the assessment of which converges to its expected payoff. 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, even though his assessment for this action does not converge. If the agent is an optimist he converges to $a_{\max\min}$.

Proof. Suppose $\rho = 1$ and that the individual does not converge to any action. Then he will choose more than one action infinitely often given that A is finite. The assessment of each of these actions will converge to their expected values which we have assumed to be distinct. But, this is a contradiction as the individual will choose only the strategy with the highest assessment.

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 $a^k \in \bar{A}$, and that $\pi^i_{\min} > \pi^k_{\min}$ for all $a^k \in \bar{A}$. Suppose that the individual converges to a strategy $a^k \in \bar{A}$, $a^i \neq a^k$. Then, at some time the agent will experience a long enough run on the worst payoff a^k can give and this will ensure that $s^k_T < \pi^i_{\min} \leq s^i_T$ for some finite time T. The latter inequality can be deduced from the fact that action i has been played in the past and has been abandoned.

[5] Note that when $\rho = 0$, the score and an action is equal to the most recent payoff it recieved.

Consider that last occurrence of this; at that point, it must be that the payo¤ the agent got from the action was below its assessment at the time. At time T, the agent will switch to a^i. Hence, the agent cannot converge to any action other than a^i. The above argument also applies for any action a^k 6= a^i that the agent plays in…nitely often. Hence, the individual cannot cycle among actions. To see that the agent can converge to a^i, it su¢ces to consider the situation in which the individual assesses the payo¤ from all a^k 2 $\bar{A}$ to be lower than $¼^i_{min}$, and that of action a^i as being higher. At such states, which clearly have a positive probability of being reached from all other states, the individual will choose only a^i.

Suppose 0 · ½ < 1, and that the individual is an optimist. This ensures that the individual will converge to play his maxmin strategy at some time, because in…nite play of any other strategy would result in a long enough run of the worst possible payo¤ from that strategy. Now, the argument in the above paragraph su¢ces to conclude the proof. ¥

The result reveals the sharp contrast in behavior induced by ½. In particular, it reveals that behavior is not continuous in ½. An interesting case arises where 0 · ½ < 1 and ½ converges to zero. For any positive value of ½, as long as $N^j_0 > 0$, we get that in the limit, the agent's next period assessment for an action he chose in the current period is equal to the payo¤ he obtained from that action, and the assessments of unplayed actions remain unchanged. It is readily seen that even in this case, a pessimist will stick with the action chosen initially and an optimist will converge to playing the maxmin action.

Proposition 2 With asymptotic myopia and random experimentation, (a) if ½ = 1, play converges to the payo¤-maximizing action. (b) if 0 · ½ < 1, the proportion of time in which the agent chooses $a^{\max\min}$ converges to one.

Proof. For part (a), note that the assessment of each action is a time average of its history of payo¤s, and since each action is played in…nitely often, all assessments converge to the objective expected payo¤. Since expected payo¤s are distinct, asymptotic myopia implies that play converges to the payo¤-maximizing action. As for part (b) the following steps are a sketch of the proof. First, note that random experimentation ensures that the score of each action is in the range of the support of the payo¤s of that action eventually. Also, for each action, the weight placed on the current payo¤ in the updated score is bounded away from zero. Then, there exists a time T_1 < 1 (not necessarily bounded) such that s^j < $¼^{\max\min}_{\min}$ with probability one for all

$a^j \neq a^{\max\min}$. This follows from considering a sequence of play in which the agent does not experiment and the realization of the state of the world is such that the score of the intended action goes below $\pi_{\min}^{\max\min}$ for all actions. This sequence has a positive probability bounded away from zero, and therefore, it will eventually happen with probability one. Finally, at this point the agent intends to play $a^{\max\min}$ but could experiment with some action a^j often enough so that the score $s^j > s^{\max\min}$. Note that the probability of this event is declining to zero over time. If this happens, the agent will intentionally play a^j. Therefore, the probability that $s^j < \pi_{\min}^{a^{\max\min}}$ is positive and bounded away from zero and increasing. Hence the probability of leaving $a^{\max\min}$ declines to zero over time while the probability of leaving any other a^j goes to one. This ensures that the proportion of time goes to one as argued. ■

Combined with Proposition 1, this result reveals that random experimentation and asymptotic myopia results in better choices when $\rho = 1$, whereas it has no significant effect when $0 \leq \rho < 1$. The latter reveals the robustness of the Sarin and Vahid (1999) maxmin result.

3.2 $\gamma \in (0; 1]; \delta \geq 0; \rho = 1$

This wide range of the parameters includes the familiar fictitious play as a special case where $\gamma = 1; \delta = 1$. Note first that when $\gamma = 1$ all the action-specific counters are equal and measure time, while when $\delta = 1$, all scores are being updated with the correct payoffs regardless of whether the action is played that period or not. Hence, the rule behaves like fictitious play with initial scores interpreted as the agent's expected payoff for each action given his prior belief about the environment. While it is immediate to see why fictitious play converges in this environment to the expected-payoff maximizing action, we will show that this is the case for any learning rule for which $\gamma = \delta > 0$ (and $\rho = 1$), since the "perceived per-period expected payoff" when the action is not played and its objective expected payoff are identical. The only candidate for a limit of such learning rules is the expected payoff maximizing action.

As $\gamma > 0$ a period counts as a positive fraction of experience for each unplayed action. The degree of payoff updating for unplayed actions varies with δ: it can go from no updating at all ($\delta = 0$) to inflation of the objective payoffs ($\delta > 1$). The main result in this section is that asymptotic behavior is of two qualitative types depending on whether the ratio δ/γ. This ratio determines the relationship between the "perceived per-period expected payoff" when the action is not played and its objective expected payoff. Play converges either to a single action or to a subset of actions that are played with positive frequencies. Loosely speaking, the set of candidates to be played asymptotically is determined according to how their "perceived per-period expected payoff" (when played and when not) relates to that of the expected-payoff maximizing action.

To illustrate this point, assume for simplicity that expected payoffs are positive and consider the case that $\delta/\gamma < 1$. If the expected-payoff maximizing action is not played asymptotically it will be shown that its score converges to $(\delta/\gamma)\pi^{max}$ which is its "perceived per-period expected payoff". Any other action with an objective expected payoff above this threshold is a potential limit of play; since once such an action is played with high frequency, its score gets closer to its expected value, while the score of all other actions becomes lower than $(\delta/\gamma)\pi^{max}$, which implies that the action is likely to be played even more frequently.. Hence the agent converges to play one of these potential actions. Note that when $\delta = \gamma$, the only such action is the expected-payoff maximizing action. When $\delta/\gamma > 1$, it can easily be shown why play cannot converge to a single action if there is at least one action besides the optimal one, say a^i, for which $(\delta/\gamma)\pi^i \geq \pi^{max}$. Suppose play converges to a single action, then its score must be approaching its expected value, however the score of a^i converges to δ/γ times its expected payoff, hence eventually a^i appears better than the action to which play converges, which is a contradiction. Hence, it must be that play switches between at least two actions, i.e., the asymptotic frequency of more than one action is positive although play is deterministic.. To summarize,

Proposition 3 (a) When $\pi^{max} > 0$ and $\delta > \gamma > 0$, or $\pi^{max} < 0$ and $\gamma \geq \delta \geq 0$ a subset of actions for which $(\delta/\gamma)\pi^i \geq \pi^{max}$ are played a positive fraction of the time. The scores of these actions converge to the same number, solving the system of equations

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

(b) When $\pi^{max} > 0$ and $\gamma \geq \delta \geq 0$, or $\pi^{max} < 0$ and $\delta > \gamma > 0$ play converges to some action which satisfies $\pi^i \geq \frac{\delta}{\gamma}\pi^{max}$.

Proof. Given $\rho = 1$, the score of any action a^i at time $(t+1)$ is,

$$S^i_{t+1} = \frac{1}{N^i_{t+1}} \sum_{\tau=1}^{t} \left[I^i_\tau \tilde{\pi}^i_\tau + (1 - I^i_\tau)\delta \tilde{\pi}^i_\tau \right]$$

where I^i_τ is the indicator function for playing action a^i at time τ, and

$$N^i_{t+1} = t\alpha^i_t + t(1-\alpha^i_t)\gamma$$

where α^i_t denotes the frequency in which action a^i has been played up to time t. S^i_{t+1} can be re-written as,

$$S^i_{t+1} = \frac{1}{\alpha^i_t + (1-\alpha^i_t)\gamma} \left(\frac{1}{t} \sum_{\tau=1}^{t} \delta \tilde{\pi}^i_\tau + \frac{1}{t} \sum_{\tau=1}^{t} (1-\delta) I^i_\tau \tilde{\pi}^i_\tau \right)$$

As t becomes large, the law of large numbers implies that the first term in the brackets gets closer to $\delta\bar{\pi}^i$. Since the choice of action a^i at time t is independent of the realization of the state at time t, condition on α^i_t, it must be that as t becomes large, $\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} (1-\delta) I^i_\tau \tilde{\pi}^i_\tau$ gets closer $(1-\delta)\alpha^i_t\bar{\pi}^i$; that is, each term can be viewed as a product of two independent random variables where I^i_τ gets values of 1 and 0 with probabilities α^i_t and $(1-\alpha^i_t)$ respectively. Therefore, the score at time $t+1$ far in the future is approximately given by

$$S^i_{t+1} \simeq \frac{1}{\alpha^i_t + (1-\alpha^i_t)\gamma} \left[\alpha^i_t\bar{\pi}^i + (1-\alpha^i_t)\delta\bar{\pi}^i \right]. \tag{1}$$

Hence, if the frequencies of play converge, the score converges to a number. We are left to show that the frequencies converge and then to observe what are the possible limits for both the frequencies and the scores.

Note first, that the score is monotonic in the frequencies. It is monotonically increasing (decreasing) when $(\gamma - \delta)\bar{\pi}^i$ is positive (negative). Assuming that the frequencies of play indeed converge, the score of unplayed actions converge to $\delta\bar{\pi}^i$, while the scores of all actions for which the frequencies converge to a positive number must converge to the same number, which should be at least as high as the score of all unplayed actions since choice is myopic. These conditions imply that the limit scores frequencies and scores must solve the system of equations induced by these conditions, i.e.,

$$S^i = \lim_{t\to\infty} S^i_t = \delta\bar{\pi}^i \text{ for } a^i \text{ such that } \alpha^i = \lim_{t\to\infty} \alpha^i_t = 0 \text{ and}$$

$$S^j = \lim_{t\to\infty} S_t^j = \frac{\delta\bar{\pi}^j + (1-\delta)\alpha^j\bar{\pi}^j}{\alpha^j + (1-\alpha^j)\gamma} \text{ for } a^j \text{ such that } \alpha^j = \lim_{t\to\infty}\alpha_t^j > 0$$

subject to $\alpha^k \geq 0$, $\sum \alpha^k = 1$, $S^j = S^{j'}$ for all $a^j; a^{j'}$ such that $\alpha^j; \alpha^{j'} > 0$,

and $S^j \geq S^i$ if $\alpha^j > 0$ and $\alpha^i = 0$.

Consider the case that $\bar{\pi}^{max} > 0$ and $\delta > \gamma > 0$, and suppose that there is at least one action a^i such that $\frac{\delta}{\gamma}\bar{\pi}^i \geq \bar{\pi}^{max}$. Suppose that this action is played at a frequency lower than the frequency implied by the solution to the above system of equations. Since the score is monotonically decreasing in the frequency, as long as this frequency is below this limit, it must be that eventually this score is above the scores of actions that are supposed to be unplayed; this follows since the score of the action must be at least as high as its limit value which is in return higher that the score of the action that are unplayed in the limit. Hence the frequency of these action will decline from that point onwards towards zero. Moreover, this score must then eventually becomes higher than the score of other actions that are supposed to be played with positive frequency. Hence, the frequency of action a^i must increase while the other frequencies decrease towards their limit values. ■

Several special cases of this Proposition are worth highlighting. When $\gamma = 1$ and $\delta = 0$ we get a behavior that exhibits "loss aversion." Specifically, for both $\rho = 1$ and $\rho < 1$, any action with $\pi^i_{min} > 0$ is absorbing. That is, the first time the agent chooses an action that ensures him only gains he never switches away from it. This arises because with this parameter cluster unplayed actions are averages with zero using the same weights as are used for played actions, since $\gamma = 1$. In the case of $\rho < 1$, if there is an action with $\pi^i_{min} > 0$, then any action with $\pi^k_{min} < 0$ will be eventually abandoned. Consequently, if the minimal payoff of the expected-payoff maximizing action is negative, this implies that play will never converge to it regardless of initial conditions.[6]

In the case of $\rho = 1$, if $\bar{\pi}^{max} < 0$, the Proposition implies that each action a^i is played with positive frequency and we get an analytical solution for the frequencies:

$$\alpha^i = \frac{\prod_{i\neq j}\bar{\pi}^j}{\sum_k\left(\prod_{j\neq k}\bar{\pi}^j\right)}$$

Conjecture 1 Proposition 3 holds under random experimentation and asymptotic myopia.

[6] If $\pi^i_{min} < 0$, every action is chosen infinitely often.

This conjecture still needs to be verified. However, it should be true since the limit points of the learning process are identified by a system of equations concerning the limit scores which are (fully) determined by the limit frequencies in which the actions are played, which in turn are not effected by random experimentation.

3.3 $\gamma = 0; \delta > 0; \rho = 1$

This parameter cluster corresponds to the case in which the agent averages the payoff received for the played action with its past score with the weights implied by the subjective experience with the action. However, unplayed actions are treated differentially, for which the perceived payoff is equal to δ times the objective payoff. In particular, it possesses the same sign as the objective payoff. However, the subjective experience with the action does not update. Consequently, the perceived payoff is added to the past score of the unplayed action. The implied behavior of such a rule resembles the old saying "the grass is greener on the other side." An action with a positive expected payoff, say a^i, that has not been played for a while looks lucrative with time, since while the played action(s) are being averaged with the received payoffs, hence are moving towards their expected payoffs, the score of a^i rises in positive amount on average.

A number of results consequently arise. If at least one action has a positive expected payoff then all actions with negative payoffs are eventually abandoned. Also, each action with positive expected payoff is played a positive fraction of the time. Similarly to the behavior investigated in the previous section, the asymptotic behavior looks like a mixed strategy. Formally,

Proposition 4 If $\bar{\pi}^{max} > 0$, each (and only) action a^i with $\bar{\pi}^i > 0$ is played with a positive frequency asymptotically.. The frequency of play increases monotonically in the expected payoff. If $\bar{\pi}^{max} < 0$, then play converges to an action and it can be any action.

The proof of the proposition is in the spirit of the proof of Proposition 3 with the observation that the expected move in the score of an unplayed action is in the direction of its expected payoff and at a rate that is bounded away from zero.

When $\delta = 1$, we can solve analytically for the asymptotic frequencies of play. Each action with a positive expected payoff is played with a probability that is proportional to its expected payoff, i.e.

$$\alpha^i = \frac{\bar{\pi}^i}{\Sigma_k \bar{\pi}^k}$$

It is interesting to note the similarity between the asymptotic behavior of this rule and that of a variation of "reinforcement learning" where scores are updated for all actions in a cumulative manner and are normalized to a probability vector using linear adjustment. This is the full information case with linear adjustment analyzed in Rustichini (1999).

3.4 The RE model

Much attention has been given to the reinforcement learning model proposed by Roth and Erev (1995, 1998). In this section we show that with an additional our parametric form spans their basic model and many others (including the "experience-weighted attraction learning model of Camerer and Ho (1999)). We briefly discuss the formulation and results for this variation. Suppose we add a parameter ϕ so that the scores are updated in the following manner

$$s^i_{t+1} = \frac{\phi N^i_t}{N^i_{t+1}} s^i_t + \frac{1}{N^i_{t+1}} \pi^i_t.$$

Note that when $\phi = \rho$ we are back to our parametric adaptive model. If we set $\rho = 0$, $\phi = (0,1]$, $N^i_t = 1$, then

$$s^i_{t+1} = \phi s^i_t + \pi\left(a^i, \omega_t\right) \quad (2)$$

$$s^k_t = \phi s^k_t \quad (3)$$

which gives us several variants of the score updating procedure of "basic model" of Roth and Erev. However, in the spirit of this paper, rather than considering a specific function converting scores to choice probabilities, we investigate the behavior of myopic agents and agents who are asymptotically myopic and experiment.

With myopic behavior the following behavior arises: (a) Each action such that $\pi^i_{min} > 0$ is absorbing;[7] (b) If for all actions a^i, $\bar{\pi}^i > 0$ and $\pi^i_{min} < 0$,

[7] However, play can converge to some action a^j with $\pi^j_{min} < 0$ with $\bar{\pi}^j > 0$.

then play can converge to any action; (c) If for all actions a^i, $\bar{¼}^i < 0$ then each action is played in…nitely often.

The intuition for the second result is that from any initial scores, any action gets played for the …rst time with positive probability since $¼^k_{min} < 0$ for all other actions. Conditional on being play, there is a positive probability that it is never abandoned since $\bar{¼}^i > 0$. The third part follows immediately since the scores of actions which are played in…nitely often declines without bound.

For the case of $Á = 0$ we get a di¤erent pattern of behavior — any action a^i with $¼^i_{min} > 0$ is absorbing, while any action a^k with $¼^k_{min} < 0$ is transient.

Roth and Erev transform scores into probabilities of choice in a very speci…c manner — they assume

$$P^i_t = \frac{S^i_t}{\sum_k S^k_t}$$

where P^i_t is the probability of choosing action a^i in time t. In the case where all payo¤s are positive, this rule leads to convergence to the expected-payo¤ maximizing action (Rustichini 1999). In stark contrast to the behavior implied by this speci…c choice rule, agents who randomly experiment and are asymptotically myopic do not behave qualitatively di¤erent from myopic agents. In the particular case of all positive payo¤s, such agents could converge to anything.

4 Extensions

There are several extensions to the model suggested in this paper that may be considered. Firstly, the analysis could be extended to decision environments which are not stationary. It appears to us that our qualitative results should also hold in environments which are Markovian. The analyses should also be extended to games with many players. Many of the recent applications of adaptive learning models have been to games. It would be interesting to see how our results extend to various classes of games.

Thirdly, it would be nice to consider more general perceived payo¤ operators. As we mentioned in footnote 1, we believe that most of our results would tend to the framework in which the perceived payo¤ of an action is a positive linear function of its expected payo¤, and more generally, to a non-linear sign-preserving monotonically-increasing function of the expected payo¤. This allows us to consider a much richer class of rules by which objective payo¤s are transformed to perceived payo¤s.

5 References

1. T. Borgers, A. Morales and R. Sarin (1998): "Simple behavior rules which lead to expected payo¤ maximizing choices," mimeo, University College London and Texas A&M University.

2. T. Borgers and R. Sarin (1999): "Naive reinforcement learning with endogenous aspirations," International Economic Review, forthcoming

3. Brown (1950):

4. R. Bush and F. Mosteller (1955): Stochastic Models of Learning, Wiley.

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 di¤erences in EWA learning with partial payo¤ information," mimeo, Caltech and Wharton.

7. D. Easley and A. Rusticini (1999): "Choice without beliefs," Econometrica, 67, 1157-1184.

8. 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.

9. N. Feltovich (1999): "Reinforcement-based vs. Belief-based learning models in experimental asymmetric games," Econometrica, forthcoming.

10. D. Fudenberg and D. Kreps (1993): "Learning mixed equilibria," Games and Economic Behavior, 5, 320-367.

11. D. Fudenberg and D. Levine (1998): The theory of learning in games, MIT press.

12. E. Hopkins (1999): "Two competing models of how people learn in games," mimeo, University of Edinburgh and Pittsburgh.

13. J. Robinson (1950): "An iterative method of solving a game," Annals of Mathematics, 54, 296-301.

14. A. Rustichini (1999): "Optimal properties of stimulus-response learning models," Games and Economic Behavior, 29, 244-273.

15. R. Sarin and F. Vahid (1999): "Payo¤ assessments without probabilities: A simple dynamic model of choice," Games and Economic Behavior, 28, 294-309.

16. R. Sarin and F. Vahid (1999): "Predicting how people play games: A simple dynamic model of choice," Games and Economic Behavior, forthcoming.