

Learning and Risk Aversion*

Carlos Oyarzun

Rajiv Sarin

University of Queensland

University of Birmingham

October 10, 2011

Abstract

We study the manner in which learning shapes behavior towards risk when individuals are not assumed to know, or to have beliefs about, probability distributions. In any period, the behavior change induced by learning is assumed to depend on the action chosen and the payoff obtained. We characterize learning rules that, in expected value, increase the probability of choosing the safest (or riskiest) actions and provide sufficient conditions for them to converge, in the long run, to the choices of risk averse (or risk seeking) expected utility maximizers. We provide a learning theoretic motivation for long run risk choices, such as those in expected utility theory with known payoff distributions.

*We thank two anonymous referees and an associate editor for instructive comments. We also thank Eddie Dekel, John Geanakoplos, Ernan Haruvy, Ed Hopkins, S. Lakshminarayanan, Eric Maskin, Jack Meyer, Robert Reed, Johannes Ruf, Karl Schlag, Joel Sobel, M.A.L. Thathachar, and audiences at the following conferences: ASSET (Alicante), Econometric Society meetings (Minneapolis, Pittsburgh, Chennai), Game theory Society meetings (Stony Brook), Midwest Economic Theory Conference (East Lansing), First Singapore Economic Theory meeting (Singapore), Southern Economic Society meetings (Washington DC) and seminars at Alabama, Alicante, ANU, Arizona, Boston College, Rice, Victoria (Wellington), Max Planck Institute (Berlin and Jena), Monash, UT Dallas, U of Queensland, and Texas A&M for helpful comments. Financial support from the Ministerio de Ciencia e Innovación and FEDER funds under project SEJ-2007-62656 (Spain) are gratefully acknowledged. All errors remain our responsibility.

1 Introduction

In the economics tradition, risk aversion is typically understood in the context of expected utility theory (EUT). This theory is silent on how individuals make choices in the absence of knowledge or beliefs about such distributions, however, and therefore, it cannot explain risk averse choices in complex problems where beliefs are hard to formulate.¹ We provide a learning approach to study choice in such contexts and show that individuals can systematically make more often the safest (or riskiest) choices over time, even in absence of beliefs, depending on how their choices change in response to experience. As the learning models we study in this paper place far fewer cognitive demands on individuals than EUT, they allow us to understand decision making in a wide variety of complex problems. These learning models include many that have been used to describe human behavior in diverse experimental settings in both psychology and economics.

We consider an agent who knows the available, finite set of actions. She does not know, and need not have beliefs about, the distribution of outcomes that would result from the choice of any action. We refer to this set of distributions, one for each action, as the *environment*. The individual chooses among the actions according to her state (of learning). This results in behavior which is assumed to be stochastic. The assumption of stochastic choice is common in learning models. It can, for instance, describe an individual who may not know for sure which is the best action, if any. In every period, the individual chooses an action and receives an outcome from the chosen action. This is the only information the agent uses in updating her *behavior*, i.e., the probabilities of choosing each action. As in EUT, outcomes are assumed to be money (not utility) and we shall often refer to them as “payoffs.” This assumption is crucial for the interpretation of all our results. Our analysis is concerned with how learning responds to the degree of risk of payoff distributions in terms of the amount of money an individual would obtain by choosing an action. This parallels the manner in which risk preferences are defined in EUT, with respect to monetary payoff distributions of lotteries.

A (behavioral) *learning rule* maps the state of the agent today, the action she chooses and the outcome she obtains into her behavior tomorrow. Hence,

¹For a related discussion, see Gilboa and Schmeidler (1995).

a learning rule describes the short run change in behavior and determines how behavior evolves over time. The learning rule is assumed to be a primitive characteristic of individuals. In particular, we do not need to assume that individuals have Bernoulli utility functions or that they play any role in the learning process. Examples of learning rules that satisfy our assumptions include those in Bush and Mosteller (1951), Cross (1973), Roth and Erev (1995), March (1996), and Börgers and Sarin (2000).

In the first part of our analysis we study the short run properties of learning. We focus on learning rules that, in every environment, are expected to increase the probability of choosing those actions whose payoff distribution second order stochastically dominates the distributions of all other actions. Therefore, decision makers whose learning rules satisfy this property are expected to increase the probability of choosing actions among those that any risk averse expected utility maximizer would choose. We refer to learning rules that satisfy this property as *monotonically risk averse learning rules*. We define *monotonically risk seeking* and *monotonically risk neutral* learning rules in an analogous way. These properties refer to the expected change in behavior (from one period to the next), and they impose restrictions on how the learning rule responds to the outcome distributions' risk properties. These restrictions are required to hold in every environment. This is a desirable restriction, given that we do not require the decision maker to know any aspect of the environment, other than the set of actions.

We characterize monotonically risk averse learning rules (Proposition 1). This characterization reveals that how the learning rule updates the probability of *unchosen actions*, in response to the payoff obtained from the chosen action, must be a convex function of the obtained payoff. The convex response of the probability of choosing the actions not chosen in this period, together with the fact that the updated probabilities must sum up to one, imply that every monotonically risk averse learning rule has to update the probability of choosing the chosen action using a concave function of the payoff obtained. The characterization of monotonically risk seeking and monotonically risk neutral learning rules is analogous.²

²EUT also provides a characterization of utility functions that prefer more to less and have a preference for first order stochastically dominant distributions. We define and characterize the analogous property for learning rules (Proposition 2).

Short run properties of learning rules are worth investigating, as behavioral scientists are often interested in how behavior is likely to *change* in response to some experience. In some problems, the number of times an individual faces a decision is very low, such as choosing a realtor for selling a house, or choosing a school for each of her children. In these kinds of problems, short run properties are particularly relevant. The short run properties of learning also have implications for the second part of our analysis to which we now turn our attention.

In the second part of the paper, we focus on the long run properties of learning rules. These properties are relevant in problems that individuals face repeatedly over time, such as choosing the assets in which to invest their wealth. We study whether learning rules that satisfy the above properties *globally*, i.e., at every state, converge to make choices that involve less risk. Our first long run result (Proposition 3) shows that any globally monotonically risk averse learning rule converges with high probability to the set of actions that second order stochastically dominate all others, provided that it satisfies two additional conditions. First, we require the learning rule to be sensitive in a non-trivial manner to second order stochastic dominance. We call this property the *non shrinking condition* (NSC). It imposes the restriction that the expected change in the probability of choosing any safest action is bounded from below away from zero (at least in proportion to the current probability), while both safest and riskier actions are chosen with positive probability. And second, we require the learning rule to be *slow* in response to experience, i.e., that it does not update the probability of choosing each action very much.³ We provide a number of examples illustrating that these conditions are not very restrictive. This long run result reveals that agents who have very limited information about the environment can converge to choose the same actions in the long run as those that would be chosen by risk averse expected utility maximizers who know the true environment. An analogous result holds for monotonically risk seeking and monotonically risk neutral learning rules.⁴

Our last result (Proposition 5) provides conditions under which the choices

³The role of slow learning to achieve convergence has been recognized in the literature and we discuss this further below.

⁴We also provide an analogous result regarding convergence to actions that first order stochastically dominate all others (Proposition 4).

of an expected utility maximizer with a given Bernoulli utility function would be obtained in the long run by a learning rule. The conditions on the learning rule required for this result are similar to those in our first two long run results. Here, however, we require the learning rule to transform payoffs into the updated probability of choosing the unchosen actions using affine transformations of the additive inverse of the Bernoulli utility function. Hence, the updated probability of choosing the chosen action is an affine transformation of the Bernoulli utility function. This result reveals the preferences specific learning rules would exhibit in the long run. We illustrate this by showing that some well-known learning rules converge to choose according to the preferences represented by some widely used utility functions, such as the constant absolute risk aversion utility functions.

The first paper we know of that investigates the relation between learning and risk aversion in expected utility theory is March (1996).⁵ March simulates three classes of learning models popular in the psychology literature. Burgos (2002) provides a similar exercise with two additional models. Most of these simulations find that the models choose the risk free action more frequently than the risky in the domain of gains. Their simulation results, for rules that are monotonically risk averse, are consistent with our findings.⁶

Denrell (2007) shows that a class of learning models that initially choose each action equally often, asymptotically, choose the risk-free action more frequently than the risky action in two action-decision problems with some restrictions on the outcome distributions. In comparison with his paper, ours characterizes a large class of learning rules that satisfy monotone risk aversion. We also provide conditions under which these learning rules converge to choose the safest action in *any* decision problem with an arbitrary finite number of actions.

⁵Earlier attempts to relate decision theory and learning theory analyze specific learning rules to see whether their long run choices coincide with those of specific static decision theories (see, e.g., Thrall, Coombs and Davis (1954) and the related paper by Simon (1956)). This literature finds that different learning rules could be used to motivate the choices predicted by some decision theories, often different from EUT (such as maximin or minimax regret).

⁶For the other rules they consider, which are not monotonically risk averse (or risk seeking), our results imply the more subtle fact that, in spite of the reported simulations, one can find other safe and risky distributions such that these rules would not choose the safest (risky) action more often.

Monotonically risk averse learning rules are closely related to the monotone learning rules studied by Börgers, Morales, and Sarin (2004), which, in every environment, are expected to strictly increase the probability of choosing the expected payoff maximizing actions. They show that monotone learning rules update the probability of the unchosen actions by a negative affine transformation of the obtained payoff. We show that all classes of learning rules studied in their paper are a strict subset of the class of monotonically risk neutral learning rules that we characterize. In contrast to the analysis in their paper, we also provide implications for choices in the long run.

Our long run analysis is related to the convergence results provided in the machine learning literature (e.g., Narendra and Thathachar (1989)). In particular, the result of Lakshmivaran and Thathachar (1976) is related to our first long run result. They show that a specific class of monotonic learning rules, which move slowly, converge with a high probability to expected payoff maximization. The learning rules they consider have a state space equal to the probability simplex, which is much smaller than the state space we allow. They only consider environments in which each action gives two possible outcomes. Hence, our results generalize theirs in several dimensions. To the best of our knowledge, the machine learning literature has no analogue of our other results.

Lastly, our work is related to the literature on the evolution of preferences. Robson (1996a,b) studies how evolution may shape attitudes towards risk and when it may lead to expected utility (and non-expected utility) preferences. Dekel and Scotchmer (1999) consider the impact of evolution on risk attitudes in winner-take-all games. This work complements ours in showing how risk preferences emerge through a dynamic process involving cognitively less sophisticated agents than those assumed in EUT.

2 Framework

Let A be a finite set of actions. If chosen, action $a \in A$ gives an outcome, which we interpret to be a number of units of money, according to the distribution function F_a , whose expected value is denoted by π_a . We assume that the support of F_a is contained in the compact interval $X = [x_{\min}, x_{\max}]$ for all a . We shall refer to $F = (F_a)_{a \in A}$ as the *environment* and we suppose

it does not change from one period to the next. The agent knows the set of actions A but not the distributions F .

All the aspects of each of the actions that are relevant for the individual's behavior are summarized by her state $s \in S$. The state space S is taken to be some subset of $\mathbb{R}^{|A| \cdot K}$, where $|A|$ denotes the number of actions in A and $K \in \mathbb{N}$ denotes the finite number of attributes of each action that are relevant for choice and learning. Larger state spaces may be considered though most of the learning rules that we are familiar with are easily accommodated with this state space. The state s of the agent determines her behavior σ , which is described by the probability with which she chooses each action. Hence, $\sigma \in \Delta(A)$ where $\Delta(A)$ denotes the set of all probability distributions over A .

In every period the agent chooses an action and receives its payoff. This is the only information the agent uses to update her state in any period. A learning rule L takes the state s of the agent, the action a she chooses and the payoff x she obtains in any period and maps them to her behavior σ in the next period. That is, $L : S \times A \times X \rightarrow \Delta(A)$. A learning rule may be seen as the composition of two functions. A first function, $L_1 : S \rightarrow \Delta(A)$ maps the state of the agent to her behavior. L_1 may be thought of as a choice or behavior rule.⁷ The second function, $L_2 : S \times A \times X \rightarrow S$ maps the state of the agent, the action she chooses and the payoff she obtains into her state in the next period. L_2 defines the state updating process and may be considered to be the "true" learning rule. Our focus in this paper is on the agent's behavior and how it changes, and consequently our emphasis is on L , which may be thought of as a behavioral learning rule.

Definition 1 *A learning rule $L : S \times A \times X \rightarrow \Delta(A)$ is a composition of two functions, L_1 and L_2 , with $L = L_1 \circ L_2$, where $L_1 : S \rightarrow \Delta(A)$ and $L_2 : S \times A \times X \rightarrow S$.*

For any state s , a learning rule tells us how the probability of *each* action $a \in A$ is updated upon choosing any action $a' \in A$ and receiving a payoff $x \in X$. Let $L_{(s,a',x)}(a)$ denote the probability with which a is chosen in the

⁷See, e.g., Hopkins (2007) and Jehiel and Samet (2005).

next period if the state of the agent is s , action a' is chosen and a payoff of x is received today. Specifying a learning rule at any state s requires specifying a (finite, square) matrix of functions,

$$\begin{pmatrix} L_{(s,a,x)}(a) & L_{(s,a,x)}(a') & L_{(s,a,x)}(a'') & \cdots \\ L_{(s,a',x)}(a) & L_{(s,a',x)}(a') & \cdots & \cdots \\ L_{(s,a'',x)}(a) & \vdots & \ddots & \\ \vdots & \vdots & & \ddots \end{pmatrix}. \quad (1)$$

Defining a learning rule involves specifying the above matrix of functions for all states $s \in S$.

The following definition introduces some terminology for describing the manner in which a learning rule responds to payoffs.

Definition 2 *A learning rule L is own-concave (own-increasing) at s if $L_{(s,a,x)}(a)$ is a concave (increasing) function of x for all $a \in A$ with $\sigma_a > 0$. A learning rule L is cross-convex (cross-decreasing) at s if $L_{(s,a',x)}(a)$ is convex (decreasing) in x for all $a \in A$ and $a' \in A \setminus \{a\}$ with $\sigma_{a'} > 0$.*

Own-increasing and own-concavity place restrictions only on the diagonal terms of the learning matrix (1). The restrictions on the off-diagonal elements of the learning rule are referred to as cross restrictions. Since each row in (1) has to sum to one, a cross-convex (cross-decreasing) learning rule is own-concave (own-increasing). Notice that these properties are *local* in that they apply to a specific state s . A specific learning rule may, for example, be own-concave at some states and not at others. If a property is satisfied at all states $s \in S$ then we will say that the property holds *globally*.

The expected change in behavior will play a central role in the subsequent analysis. This magnitude is useful because it allows us to abstract from the random aspects of both behavior and payoffs. As we shall see in section 5, it also provides pertinent information for the long run analysis of learning. For a given learning rule L and environment F , let the expected movement of probability on action a in state s be denoted by $f_s(a) = \sum_{a' \in A} \sigma_{a'} \int L_{(s,a',x)}(a) dF_{a'}(x) - \sigma_a$.

3 The Short Run

In this section we take the state of the agent as given and study how behavior changes from one period to the next. That is, in this section, we focus on local and *short run* properties of learning rules. Taking the state of the agent today as given, a learning rule specifies her behavior tomorrow as a function only of the action $a \in A$ she chooses and the payoff $x \in X$ she obtains. Abusing notation, we will denote such a local learning rule by $L : A \times X \rightarrow \Delta(A)$ and denote by $f(a)$ the expected change in the probability of playing action $a \in A$. Accordingly, all the statements of this section are local in the sense that they refer to some, arbitrary, given state $s \in S$.

In EUT, a distribution F_a is said to be more risky than another F'_a if and only if both have the same mean and every risk averse agent prefers F'_a to F_a (e.g., Mas-Colell et al., 1995). In this case it is usually said that F'_a second order stochastically dominates (SOSDs) F_a . Let A^* denote the set of actions whose distributions SOSD those of all other actions. That is, $A^* = \{a : F_a \text{ SOSDs } F_{a'} \text{ for all } a' \in A\}$. If $A^* = A$ then $F_a = F_{a'}$ for all $a, a' \in A$. For any subset $\hat{A} \subset A$, let $f(\hat{A}) = \sum_{a \in \hat{A}} f(a)$.

Definition 3 A learning rule L is monotonically risk averse if $f(A^*) \geq 0$ in all environments.

A learning rule is monotonically risk averse if the expected change in the probability of choosing one of the safest actions is positive in every environment. Correspondingly, we say that a learning rule is *monotonically risk seeking* if the expected change in the probability of choosing one of the riskiest actions (those that are second order stochastically dominated by all other actions) is positive in every environment. Finally, we say that a learning rule is *monotonically risk neutral* if it is monotonically risk averse and monotonically risk seeking. When A^* is empty, expected utility maximization by a risk averse agent places no specific restrictions on behavior (without knowledge of her utility function). This has the analogue, in this paper, that no restrictions are placed on the movement of probability when A^* is empty (without knowledge of the learning rule).

We shall see, in the proof of the next result, that all monotonically risk averse learning rules have the feature that if all actions have the same distribution of payoffs then there is no expected movement of probability. We call such learning rules *impartial*.

Definition 4 *A learning rule L is impartial if $f(a) = 0$ for all a whenever $F_a = F_{a'}$ for all $a, a' \in A$.*

Proposition 1 *A learning rule L is monotonically risk averse if and only if (i) $\sigma_a = \sum_{a' \in A} \sigma_{a'} L_{(a', x)}(a)$ for all $a \in A$ and $x \in X$, and (ii) L is cross-convex.*

Our proof begins with two lemmas. The first shows that all monotonically risk averse learning rules are impartial and the second characterizes impartial learning rules.

Lemma 1 *If a learning rule L is monotonically risk averse then it is impartial.*

Proof. The proof is by contradiction. Suppose L is monotonically risk averse but there exists an environment F with $A = A^*$ and $f(a) > 0$ for some $a \in A$. If F_a does not place strictly positive probability on (x_{min}, x_{max}) , then consider the environment \widehat{F} such that, for all action $a' \in A$, the probabilities of x_{min} and x_{max} are $(1-\varepsilon)$ times their corresponding probabilities in the environment F , and the probability of some $x \in (x_{min}, x_{max})$ is ε . If F_a places strictly positive probability on (x_{min}, x_{max}) , then let $\widehat{F} = F$. We now construct the environment \widetilde{F} in which \widetilde{F}_a is a mean preserving spread of \widehat{F}_a and $\widetilde{F}_{a'} = \widehat{F}_{a'}$ for all $a' \neq a$. Specifically, suppose that \widetilde{F}_a is obtained by assigning to every interval $I \subset [x_{min}, x_{max}]$ only $(1-\varepsilon)$ times the probability it had under \widehat{F}_a and then adding $(\widehat{\pi}_a - x_{min})\varepsilon/(x_{max} - x_{min})$ on the probability of x_{max} and $(x_{max} - \widehat{\pi}_a)\varepsilon/(x_{max} - x_{min})$ on the probability of x_{min} . By construction, $\widetilde{F}_{a'}$ SOSD \widetilde{F}_a for all $a' \neq a$. It follows that $\widetilde{A}^* = A \setminus \{a\}$. Since $\widetilde{f}(a)$ is a continuous function of ε , there exists a small enough ε such that $\widetilde{f}(a) > 0$. This contradicts that L is monotonically risk averse. ■

Lemma 2 *A learning rule L is impartial if and only if $\sigma_a = \sum_{a' \in A} \sigma_{a'} L_{(a',x)}(a)$ for all $a \in A$ and $x \in X$.*

Proof. *Necessity.*

Consider an environment where all actions pay x with probability one. Then, for all $a \in A$,

$$f(a) = \sum_{a' \in A} \sigma_{a'} L_{(a',x)}(a) - \sigma_a.$$

Therefore, in order to be impartial, L must satisfy

$$\sigma_a = \sum_{a' \in A} \sigma_{a'} L_{(a',x)}(a).$$

Sufficiency.

Consider the environment F such that $F_a = F_{a'}$ for all $a, a' \in A$.

$$\begin{aligned} f(a) &= \sum_{a' \in A} \sigma_{a'} \int L_{(a',x)}(a) dF_{a'}(x) - \sigma_a \\ &= \int \sum_{a' \in A} \sigma_{a'} L_{(a',x)}(a) dF_a(x) - \sigma_a \\ &= 0. \end{aligned}$$

The second statement follows from the fact that all the distributions are the same, and the third statement follows from the hypothesis. ■

Notice that Lemma 2 implies that impartial rules do not “experiment,” i.e., if $\sigma_a = 0$ then $L_{(a',x)}(a) = 0$ for all $a' \neq a$ and for all x . We list this observation for future reference as Remark 1.

Remark 1 *If L is monotonically risk averse and $\sigma_a = 0$ then $L_{(a',x)}(a) = 0$ for all $a' \in A \setminus \{a\}$ with $\sigma_{a'} > 0$ and for all $x \in X$.*

Proof. We now proceed to complete the proof of Proposition 1.

Sufficiency.

Consider $a \in A^*$

$$\begin{aligned}
f(a) &= \sigma_a \int L_{(a,x)}(a) dF_a(x) + \sum_{a' \neq a} \sigma_{a'} \int L_{(a',x)}(a) dF_{a'}(x) - \sigma_a \\
&= \int [\sigma_a - \sum_{a' \neq a} \sigma_{a'} L_{(a',x)}(a)] dF_a(x) + \sum_{a' \neq a} \sigma_{a'} \int L_{(a',x)}(a) dF_{a'}(x) - \sigma_a \\
&= \sum_{a' \neq a} \sigma_{a'} [\int L_{(a',x)}(a) dF_{a'}(x) - \int L_{(a',x)}(a) dF_a(x)] \\
&\geq 0.
\end{aligned}$$

The second statement follows from (i) in the hypothesis of the proposition and the last inequality follows from the fact that $a \in A^*$ and the convexity of the functions $L_{(a',x)}(a)$ for all $a' \neq a$ such that $\sigma_{a'} > 0$.

Necessity.

We argue by contradiction. Suppose that for some $a \in A$ and some $a' \neq a$ such that $\sigma_{a'} > 0$, $L_{(a',x)}(a)$ is not convex. Therefore there exists x' , x'' , $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda)x''$ such that $\lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a) < L_{(a',x)}(a)$. Consider an environment where $a' \neq a$ pays x' with probability λ , and x'' with probability $(1 - \lambda)$. Action a pays x with probability one, and all the other actions in the set, if any, pay x with probability $1 - \varepsilon$, x' with probability $\varepsilon\lambda$, and x'' with probability $\varepsilon(1 - \lambda)$. Clearly, $A^* = \{a\}$. From the sufficiency part we know

$$\begin{aligned}
f(a) &= \sum_{a'' \neq a} \sigma_{a''} [\int L_{(a'',x)}(a) dF_{a''}(x) - \int L_{(a'',x)}(a) dF_a(x)] \\
&= \sigma_{a'} [\lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a) - L_{(a',x)}(a)] \\
&\quad + \varepsilon \sum_{a'' \neq a, a'} \sigma_{a''} [\lambda L_{(a'',x')}(a) + (1 - \lambda)L_{(a'',x'')}(a) - L_{(a'',x)}(a)].
\end{aligned}$$

Therefore, for small enough ε , $f(a) < 0$. ■

An analogous argument shows that a learning rule is monotonically risk seeking (neutral) *if and only if* it is impartial and cross-concave (cross-convex and cross-concave).

Monotonic risk aversion places restrictions on how a learning rule updates the probability of each unchosen action as a function of the payoff obtained

from the action chosen. In particular, it requires this function to be convex in the payoff received for each unchosen action. Since $\sum_{a \in A} L_{(a',x)}(a) = 1$, every cross-convex learning rule is own-concave. Hence, every monotonically risk averse learning rule is own-concave.

Proposition 1 also shows that all monotonically risk averse learning rules are impartial. The set of impartial learning rules is related to the unbiased learning rules studied in Börgers et al. (2004). A learning rule is *unbiased* if no probability mass is expected to be moved among actions when all have the same expected payoff (i.e., $f(a) = 0$ for all a whenever $\pi_a = \pi_{a'}$ for all $a, a' \in A$). Unbiased learning rules have the restrictive feature that they cannot respond to aspects of the payoff distribution other than the mean. The set of impartial learning rules is considerably larger than the set of unbiased rules.

Börgers et al. (2004) characterize the class of monotone learning rules, defined as those which are expected to strictly increase probability on the set of expected payoff maximizing actions, in every environment. Such rules seem to respect risk neutrality. It may, hence, be conjectured that the set of monotone learning rules is equal to the set of monotonically risk neutral learning rules. This conjecture turns out to be false. In fact, a learning rule is monotonically risk neutral *if and only if* it is unbiased. As unbiased learning rules (strictly) contain the set of monotone learning rules, monotone learning rules are (strictly) contained in the class of monotonically risk neutral rules studied in this paper.

Remark 2 *A learning rule is unbiased if and only if it is monotonically risk neutral.*

Proof. Recall from Proposition 1 in Börgers et al. (2004) that if $\sigma \in \text{int}(\Delta(A))$, a learning rule L is unbiased if it can be written, for all $a \in A$ and $x \in [0, 1]$, as

$$\begin{aligned} L_{(a,x)}(a) &= \sigma_a + (1 - \sigma_a)(A_{aa} + B_{aa}x) \\ L_{(a',x)}(a) &= \sigma_a - \sigma_a(A_{a'a} + B_{a'a}x) \quad \forall a' \neq a, \end{aligned} \tag{2}$$

where the coefficients $A_{a'a}$ and $B_{a'a}$ satisfy $A_{aa} = \sum_{a' \in A} \sigma_{a'} A_{a'a}$ and $B_{aa} = \sum_{a' \in A} \sigma_{a'} B_{a'a}$.⁸ To see that unbiased rules are impartial, we verify that

$$\begin{aligned} \sum_{a' \neq a} \sigma_{a'} L_{(a',x)}(a) + \sigma_a L_{(a,x)}(a) &= \sum_{a' \neq a} [\sigma_{a'} (\sigma_a - \sigma_a (A_{a'a} + B_{a'a}x))] \\ &\quad + \sigma_a (\sigma_a + (1 - \sigma_a)(A_{aa} + B_{aa}x)) \\ &= \sigma_a. \end{aligned}$$

Equations (2) show that unbiased rules are cross-affine, and hence they are cross-concave and cross-convex. Consequently, unbiased rules are monotonically risk averse and monotonically risk seeking and, hence, monotonically risk neutral.

To prove that monotonically risk neutral learning rules are unbiased, note that every monotonically risk neutral rule is cross-concave and cross-convex. Therefore, there exist coefficients $A_{a'a}$ and $B_{a'a}$ such that every monotonically risk neutral rule can be written as (2). To prove that $A_{aa} = \sum_{a' \in A} \sigma_{a'} A_{a'a}$ and $B_{aa} = \sum_{a' \in A} \sigma_{a'} B_{a'a}$, note that monotonically risk neutral rules are impartial and, hence, they satisfy

$$\sum_{a' \neq a} \sigma_{a'} L_{(a',x)}(a) + \sigma_a L_{(a,x)}(a) = \sigma_a.$$

Therefore,

$$\begin{aligned} L_{(a,x)}(a) &= 1 - \sum_{a' \neq a} \sigma_{a'} L_{(a',x)}(a) / \sigma_a \\ &= 1 - \sum_{a' \neq a} \sigma_{a'} (1 - (A_{a'a} + B_{a'a}x)) \\ &= \sigma_a + \sum_{a' \neq a} \sigma_{a'} (A_{a'a} + B_{a'a}x). \end{aligned}$$

Equating the RHS of the last equality with the RHS of the first equality in (2), we obtain that $\sigma_a + (1 - \sigma_a)(A_{aa} + B_{aa}x) = \sigma_a + \sum_{a' \neq a} \sigma_{a'} (A_{a'a} + B_{a'a}x)$. Since this equality has to hold for all x we obtain that $A_{aa} = \sum_{a' \in A} \sigma_{a'} A_{a'a}$ and $B_{aa} = \sum_{a' \in A} \sigma_{a'} B_{a'a}$, as desired. ■

⁸Börgers et al. (2004) define unbiasedness for behaviors in the interior of the simplex. However, their definition and its equivalence to monotone risk neutrality can be extended to the whole simplex.

In EUT, a distribution F'_a is said to first order stochastically dominate (FOSD) another F_a *if and only if* every individual with an increasing Bernoulli utility function prefers the former. We conclude this section with the analogue of Proposition 1 as it pertains to first-order stochastic dominance. Specifically, we would like to identify the learning rules that are expected to add probability mass on the set of actions whose distributions FOSD the distributions of all the other actions, in every environment. We call such learning rules *first order monotone*. Let $A^{**} := \{a \in A : F_a \text{ FOSDs } F_{a'} \text{ for all } a' \in A\}$.

Definition 5 *A learning rule L is first order monotone if $f(A^{**}) \geq 0$ in every environment.*

First-order monotone learning rules can be characterized in the same manner as monotonically risk averse learning rules. In particular, these rules need to be impartial but instead of being cross-convex they require the response of the probabilities of playing the unchosen actions to be decreasing in the obtained payoff.

Proposition 2 *A learning rule L is first order monotone if and only if (i) $\sigma_a = \sum_{a' \in A} \sigma_{a'} L_{(a',x)}(a)$ for all $a \in A$ and $x \in X$, and (ii) L is cross-decreasing.*

The proof of this result is based on an argument that is analogous to the one used in the proof of Proposition 1 and is provided in the Appendix.

The class of monotone learning rules studied in Börgers et al. (2004) is contained in the set of first order monotone rules which we characterize above. Whereas monotone learning rules are unbiased and affine in payoffs, the first-order monotone rules we characterize are impartial and do not require affine-ness in payoffs.

In EUT requiring that an individual prefers more to less and is risk averse restricts her Bernoulli utility function to be increasing and concave. Our results show that requiring a learning rule to be first order monotone and monotonically risk averse restricts how each element of the learning matrix

(1) responds to payoffs. Specifically, each of the off-diagonal terms have to be decreasing and convex. Hence, the diagonal terms have to be increasing and concave. These restrictions on learning rules should prove useful in further theoretical and empirical work on learning.

4 Examples

In this section we study the short run properties of some learning rules that have appeared in the literature. The next section studies the long run properties of some of these rules or their straightforward modifications and extensions.

Example (1) considers the Cross (1973) learning model. Previous work has studied its long run properties and shown its relation to the replicator dynamics of evolutionary game theory (e.g., Börgers and Sarin, 1997). This rule was shown to satisfy all the properties studied in Börgers et al. (2004). Example (2) studies the Roth and Erev (1995) learning rule which has been used extensively in the experimental analysis of learning in games (e.g., Erev and Roth, 1998). Its long run properties in decision problems, normal form and extensive form games have been studied by Rustichini (1999), Laslier, Topol and Walliser (2001), Beggs (2005), Hopkins and Posch (2005) and Laslier and Walliser (2005). The Roth and Erev (1995) learning rule does not satisfy any of the properties studied by Börgers et al. (2004) who show this rule is not unbiased. Both of our first two examples are discussed in Fudenberg and Levine (1998). Examples (3) - (5) study three learning models analyzed by March (1996) and which have been used widely in the psychological literature on learning. A class of belief-based learning rules (see, e.g., Erev and Roth (1998)) that generalizes one of the learning rules in March (1996) is also considered.

Example 1 [*Cross, 1973*] For any action $a \in A$,

$$\begin{aligned} L_{(s,a,x)}(a) &= \sigma_a + (1 - \sigma_a)x \\ L_{(s,a',x)}(a) &= \sigma_a - \sigma_a x \quad \forall a' \neq a, \end{aligned}$$

where $s := (\sigma_a)_{a \in A} \in \Delta(A)$ and $x \in [0, 1]$.

In the Cross rule, the state of the player is given by her behavior σ . For an arbitrary s and x , $\sum_{a' \in A} \sigma_{a'} L_{(s,a',x)}(a) = \sum_{a' \neq a} \sigma_{a'} (\sigma_a - \sigma_a x) + \sigma_a (\sigma_a + (1 - \sigma_a) x) = \sigma_a$. Hence, by Lemma 2, the Cross learning rule is impartial. Furthermore, for all $a' \neq a$, $L_{(s,a',x)}(a)$ is an affine transformation of x . Therefore, the rule is monotonically risk averse and monotonically risk seeking and, hence, it is monotonically risk neutral. As the cross terms are decreasing in x , the rule is first order monotone. As all these properties hold for all states $s \in \Delta(A)$, each of them also holds globally.

Example 2 [Roth and Erev, 1995] *The state of an agent is given by a vector $s \in R_{++}^{|A|}$. The vector $s = (s_a)_{a \in A}$ describes the decision makers “attraction” to choose any of her $|A|$ actions. The function L_1 which maps the state of the agent to her behavior is given by $\sigma_a = s_a / \sum_{a'} s_{a'}$ for all a . The function L_2 which updates the state s is defined as follows: if the agent plays a and receives a payoff of x then she adds x to her attraction to play a , leaving all other attractions unchanged. Combining L_1 and L_2 we see that the Roth and Erev learning rule, for any action a , is given by*

$$\begin{aligned} L_{(s,a,x)}(a) &= \frac{s_a + x}{\sum_{a''} s_{a''} + x} \\ L_{(s,a',x)}(a) &= \frac{s_a}{\sum_{a''} s_{a''} + x} \quad \forall a' \neq a, \end{aligned}$$

where $x \in [0, x_{\max}]$.

For arbitrary s and x , $\sum_{a' \in A} \sigma_{a'} L_{(s,a',x)}(a) = \sum_{a' \neq a} (s_{a'} / \sum_{a''} s_{a''}) (s_a / \sum_{a''} s_{a''} + x) + (s_a / \sum_{a''} s_{a''}) (s_a + x) / (\sum_{a''} s_{a''} + x) = s_a / \sum_{a'} s_{a'} = \sigma_a$. Therefore, this rule is impartial. For all $a' \neq a$, $L_{(s,a',x)}(a)$ is a decreasing and convex function of x , which implies that this learning rule is globally first order monotone and monotonically risk-averse.

Example 3 [Weighted Return over Gains, March, 1996] *The state of learning is described by a vector of attractions $s \in X^{|A|}$, where $X = [x_{\min}, x_{\max}]$ and $x_{\min} > 0$. Given s , the agents behavior is given by $\sigma_a = s_a / \sum_{a'} s_{a'}$ for all a . If action a is chosen and receives a payoff of x , the updating process of the state adds $\beta(x - s_a)$ to s_a , where $\beta \in (0, 1)$ is a parameter, leaving all*

other attractions unchanged. Thus, for any $a \in A$, this learning rule may be written as

$$\begin{aligned} L_{(s,a,x)}(a) &= \frac{s_a + \beta(x - s_a)}{\sum_{a'' \in A} s_{a''} + \beta(x - s_a)} \\ L_{(s,a',x)}(a) &= \frac{s_a}{\sum_{a'' \in A} s_{a''} + \beta(x - s_{a'})} \quad \forall a' \neq a, \end{aligned}$$

This learning rule is cross-convex. Similar computations to those provided above for the Cross learning rule and the Roth and Erev learning rule show that this learning rule is monotonically risk averse and first order monotone if $s_a = s_{a'}$ for all $a, a' \in A$. However this learning rule is not globally impartial and thus is not globally monotonically risk averse or globally first-order monotone. To see this, consider an environment in which all actions pay x with probability one and the state s has $s_a \in (0, x)$ and $s_{a'} > x$ for all $a' \neq a$. With probability one, either the attraction of an action a' will decrease or the attraction of a will increase. In either case σ_a will increase and thus $f_s(a) > 0$. It follows that this rule cannot be globally impartial. March (1996) also proposes a weighted return model over losses that can be analyzed in the same way.

The analysis of the *average return model* studied by March (1996) is similar. Here β is replaced by $\beta_a = 1/(\kappa_a + 1)$, where κ_a is the number of times action a has been chosen in the past, and hence the attraction of each action is the average of its historical payoffs. This model is a particular case of a class of *belief-based learning rules* considered, for instance, in Erev and Roth (1998), p. 866. For this class of learning rules, the state, denoted by (s, κ) , is given by the vector of (strictly positive) attractions, denoted by s , and a vector containing the number of times each action has been chosen in the past, denoted by κ . In particular, for any $a \in A$,

$$\begin{aligned} L_{((s,\kappa),a,x)}(a) &= \frac{(s_a + (u(x) - s_a)/(\kappa_a + 1))^\rho}{\sum_{a'' \neq a} s_{a''}^\rho + (s_a + (u(x) - s_a)/(\kappa_a + 1))^\rho} \\ L_{((s,\kappa),a',x)}(a) &= \frac{s_a^\rho}{\sum_{a'' \neq a'} s_{a''}^\rho + (s_{a'} + (u(x) - s_{a'})/(\kappa_{a'} + 1))^\rho} \quad \forall a' \neq a, \end{aligned}$$

where $u : X \rightarrow X$ is an increasing transformation defined on $X := [x_{min}, x_{max}]$, with $x_{min} > 0$, κ_a is the number of times action a has been chosen in the past, and $\rho \geq 0$ is a parameter. The average return model is obtained when

$\rho = 1$ and $u(x) = x$ for all $x \in X$. As ρ goes to infinity, the model approximates arbitrarily close fictitious play (the action with the highest attraction is chosen with probability arbitrarily close to one). If $\rho = 1$, each action is chosen proportionally to the past average of $u(x)$ for each action. In game theory, since payoffs are considered to be von Neumann-Morgenstern utility (see, e.g., Fudenberg and Tirole (1991)), u may be interpreted as a Bernoulli utility function.⁹ One may conjecture that if u is concave, then the resulting learning rule is monotonically risk averse. An argument analogous to the one we used for the March's (1996) weighted return over gains model reveals that this conjecture is false for any $\rho \geq 0$, because the belief-based learning rule fails to be globally impartial. Hence, belief-based learning rules fail to satisfy our short run properties.¹⁰

Example 4 [*Fractional Adjustment over Gains, March, 1996*] *This learning rule is defined for environments in which the decision maker has only two actions. As in the Cross learning rule, $s := (\sigma_a)_{a \in A} \in \Delta(A)$. Over gains, i.e., for $x \geq 0$, this model is specified for any $a \in A$ by*

$$\begin{aligned} L_{(s,a,x)}(a) &= 1 - (1 - \alpha)^x (1 - \sigma_a) \\ L_{(s,a',x)}(a) &= (1 - \alpha)^x \sigma_a \quad a' \neq a, \end{aligned}$$

where $\alpha \in (0, 1)$ is a parameter.

For all s , $\sum_{a' \in A} \sigma_{a'} L_{(s,a',x)}(a) = \sigma_{a'} ((1 - \alpha)^x \sigma_a) + \sigma_a (1 - (1 - \alpha)^x (1 - \sigma_a)) = \sigma_a$. Therefore, this rule is impartial. For all $a' \neq a$, $L_{(s,a',x)}(a)$ is a decreasing and convex function of x . Hence, this learning rule is globally first order monotone and monotonically risk averse.

⁹As we do not assume risk preferences, however, this is not the interpretation we consider along the paper.

¹⁰In contrast, for every learning rule L that is both monotonically risk averse and first-order monotone, and increasing and concave function $u : X \rightarrow X$, the learning rule \hat{L} , such that $\hat{L}_{(s,a',x)}(a) = L_{(s,a',u(x))}(a)$ for all $a, a' \in A$, $s \in S$, and $x \in X$, is both monotonically risk averse and first-order monotone. This follows from the fact that impartiality, cross-convexity and cross-decreasing are preserved when payoffs are transformed using such a function u .

Example 5 [*Fractional Adjustment over Losses, March, 1996*] The learning rule, for any $a \in A$, is specified as

$$\begin{aligned} L_{(s,a,x)}(a) &= (1 - \alpha)^{-x} \sigma_a \\ L_{(s,a',x)}(a) &= 1 - (1 - \alpha)^{-x} (1 - \sigma_a) \quad a' \neq a. \end{aligned}$$

where $x < 0$.

This rule is cross-concave. However, $\sum_{a' \in A} \sigma_{a'} L_{(s,a',x)}(a) = (1 - \alpha)^{-x} \sigma_a^2 + (1 - \sigma_a) (1 - (1 - \alpha)^{-x} (1 - \sigma_a)) = (1 - \sigma_a) - (1 - 2\sigma_a) (1 - \alpha)^{-x}$. This expression is equal to σ_a only for $x = 0$, therefore this rule is not impartial at any state.

5 The Long Run

In subsection 5.1 we introduce some concepts that play an important role in the subsequent analysis. Subsection 5.2 proves that under some conditions globally monotonically risk averse (respectively, globally first-order monotone) learning rules converge, with high probability, to the set of actions that SOSD (respectively, FOSD) all others. Subsection 5.3 shows that for every continuous Bernoulli utility function, there exist learning rules that, with high probability, converge to choose the corresponding expected utility maximizing action.

5.1 Preliminaries

As we consider the learning rule at different periods of time, we will often subscript the state and behavior by the time period $t = 0, 1, \dots$. Let σ_t^* denote the probability with which an action in A^* is chosen at time t . Hence, $\sigma_t^* = \sum_{a \in A^*} \sigma_{at}$, where σ_{at} is the probability of choosing action a at time t .

Recall the standard concept of strict second-order stochastic dominance. A distribution $F'_{a'}$ strictly-SOSDs $F_{a'}$ if $F'_{a'}$ SOSDs $F_{a'}$ and there is some $x \in X$ such that $F'_{a'}(x) \neq F_{a'}(x)$. The following definition formalizes the idea that if we replace the distribution of an action a' with one which strictly-SOSDs it, then the learning rule should respond to this change in a non-trivial manner. Specifically, it requires that, at all states in which a and a' are chosen with positive probability, if the distribution of a' is replaced by

a more “attractive” distribution then less probability should be added to any action $a \neq a'$, when a' is chosen. The condition requires that this difference in the response has to be bounded away from zero. It utilizes a normalization which ensures that it is consistent with impartiality.¹¹ For any learning rule L , two actions a and $a' \neq a$, and payoff distributions $F'_{a'}$ and F_a , let

$$n_{a'a}(F'_{a'}, F_a) := \inf_{\{s: \sigma_a, \sigma_{a'} > 0\}} \left\{ \left(\int L_{(s, a', x)}(a) dF'_{a'}(x) - \int L_{(s, a', x)}(a) dF_a(x) \right) / \sigma_a \right\}.$$

Definition 6 *A learning rule L satisfies the non-shrinking condition (NSC) if for any pair of actions $a \neq a'$ and payoff distributions $F'_{a'}$ and F_a such that $F'_{a'}$ strictly-SOSDs F_a , $n_{a'a}(F'_{a'}, F_a) > 0$.*

Notice that learning rules in which the change in probabilities becomes arbitrarily small, such as the Roth-Erev (1995) learning rule, may not satisfy the NSC. Such rules may asymptotically become insensitive to risk in the environment.

Lemma 3 *Suppose L is a globally monotonically risk averse learning rule and satisfies the NSC. Then, for arbitrary $s_0 \in S$, σ_t^* converges almost surely to a random variable σ_∞^* with support in $\{0, 1\}$.*

Proof. For any monotonically risk averse learning rule L , σ_t^* is a submartingale bounded above by 1. Consequently, σ_t^* converges with probability 1 to a random variable σ_∞^* (e.g., Billingsley, 1995, p. 468). We shall prove that the support of σ_∞^* is contained in $\{0, 1\}$ by showing that $\Pr\{\sigma_\infty^* \in (0, 1)\} = 0$.

Let, $f_{s_t}(A^*) := E[\sigma_{t+1}^* - \sigma_t^* | s_t]$ denote the expected change in the probability of choosing an action in A^* given that the state at time t is s_t . From the proof of Proposition 1, we know that for any monotonically risk averse learning rule,

$$\begin{aligned} f_{s_t}(A^*) &= \sum_{a \in A^*} \sum_{a' \notin A^*} \sigma_{a't} \left(\int L_{(s_t, a', x)}(a) dF'_{a'}(x) - \int L_{(s_t, a', x)}(a) dF_a(x) \right) \\ &= \sum_{\{a \in A^* | \sigma_{at} > 0\}} \sum_{\{a' \notin A^* | \sigma_{a't} > 0\}} \sigma_{at} \sigma_{a't} \left(\int L_{(s_t, a', x)}(a) dF'_{a'} - \int L_{(s_t, a', x)}(a) dF_a \right) / \sigma_{at}. \end{aligned}$$

¹¹In particular, impartiality implies that for all states such that σ_a is arbitrarily close to zero, while $\sigma_{a'}$ is bounded away from zero for all $a' \neq a$, $L_{(s, a', x)}(a)$ must be arbitrarily close to zero for all x (see Lemma 2). Hence, without this normalization, the learning rule would not be able to respond to the “more attractive” distribution.

The second equality follows from Remark 1 and deleting all the $\sigma_{a't}$ terms that are equal to zero.

Let

$$\delta := \min_{\{a \in A^*, a' \notin A^*\}} \left\{ \inf_{\{s: \sigma_a, \sigma_{a'} > 0\}} \left\{ \left(\int L_{(s, a', x)}(a) dF_{a'} - \int L_{(s, a', x)}(a) dF_a \right) / \sigma_a \right\} \right\}.$$

The NSC guarantees that $\delta > 0$. Consequently, for all $s_t \in S$,

$$f_{s_t}(A^*) \geq \delta \sum_{\{a \in A^* | \sigma_{at} > 0\}} \sum_{\{a' \notin A^* | \sigma_{a't} > 0\}} \sigma_{at} \sigma_{a't} = \delta \sum_{a \in A^*} \sum_{a' \notin A^*} \sigma_{at} \sigma_{a't} \geq 0. \quad (3)$$

For all t ,

$$\begin{aligned} E[\sigma_t^*] &= \sigma_0^* + E[\sigma_1^* - \sigma_0^*] + E[\sigma_2^* - \sigma_1^*] + \dots + E[\sigma_t^* - \sigma_{t-1}^*] \\ &= \sigma_0^* + \sum_{\tau=0}^{t-1} E[f_{s_\tau}(A^*)]. \end{aligned}$$

Since $f_{s_t}(A^*) \geq 0$ for all s_t , $E[f_{s_t}(A^*)] \geq 0$ for all t . It follows that $\lim_{t \rightarrow \infty} E[f_{s_t}(A^*)] = 0$, because $E[\sigma_t^*] \leq 1$ for all t .

From (3), $E[f_{s_t}(A^*)] \geq E[\delta \sum_{a \in A^*} \sum_{a' \notin A^*} \sigma_{at} \sigma_{a't}] \geq 0$ for all t . Since $\lim_{t \rightarrow \infty} E[f_{s_t}(A^*)] = 0$, $\lim_{t \rightarrow \infty} E[\delta \sum_{a \in A^*} \sum_{a' \notin A^*} \sigma_{at} \sigma_{a't}] = 0$. Since $\delta > 0$, $\lim_{t \rightarrow \infty} E[\sum_{a \in A^*} \sum_{a' \notin A^*} \sigma_{at} \sigma_{a't}] = 0$. Equivalently, $\lim_{t \rightarrow \infty} E[\sigma_t^* (1 - \sigma_t^*)] = 0$. This may be written as $\lim_{t \rightarrow \infty} E[\sigma_t^*] = \lim_{t \rightarrow \infty} E[(\sigma_t^*)^2]$.

By the Continuous Mapping Theorem (e.g., Resnick 1999, p. 261), $\lim_{t \rightarrow \infty} E[\sigma_t^*] = E[\sigma_\infty^*]$ and $\lim_{t \rightarrow \infty} E[(\sigma_t^*)^2] = E[(\sigma_\infty^*)^2]$.¹² Therefore, $E[\sigma_\infty^*] = E[(\sigma_\infty^*)^2]$. Since $\sigma^* > (\sigma^*)^2$ for all $\sigma^* \in (0, 1)$, $\Pr\{\sigma_\infty^* \in (0, 1)\} = 0$. ■

Let $\Gamma(s_0) := \Pr\{\sigma_\infty^* = 1\}$ for arbitrary $s_0 \in S$. Existence and uniqueness of $\Gamma(s_0)$ follows from Lemma 3, the fact that convergence with probability 1 implies weak convergence and the uniqueness of weak limits (e.g., Resnick 1999, p. 252).

¹²The Continuous Mapping Theorem (e.g., Resnick, 1999) establishes that if the sequence of random variables $\{X_t\}_{t=1}^\infty$ converge in distribution to the random variable X and the probability that X is in the set of points where the bounded function $g: \mathbb{R} \rightarrow \mathbb{R}$ is discontinuous is 0, then $\lim_{t \rightarrow \infty} E g(X_t) = E g(X)$. Since σ_t^* converges to σ_∞^* with probability 1, σ_t^* converges to σ_∞^* in distribution, hence the Continuous Mapping Theorem applies.

Lemma 3 guarantees that monotonically risk averse learning rules will converge to choose the safest actions with probability zero or one. However, it is uninformative about the probability with which σ_t^* converges to one. To obtain more precise bounds on $\Gamma(s_0)$ we consider slow versions of the original learning rule L .

Definition 7 For any learning rule $L = L_1 \circ L_2$, with state space S , and $\theta \in (0, 1]$, a learning rule $L^\theta = L_1 \circ L_2^\theta$, with state space $S^\theta \subseteq S$, is a slow version of L if

$$L_{(s,a',x)}^\theta(a) - \sigma_a = \theta (L_{(s,a',x)}(a) - \sigma_a)$$

for all $s \in S^\theta$, $a', a \in A$ and $x \in X$.

Equivalently, a slow version L^θ of L satisfies,

$$L_{(s,a',x)}^\theta(a) = (1 - \theta)\sigma_a + \theta L_{(s,a',x)}(a). \quad (4)$$

for all $x \in X$, $a', a \in A$, and $s \in S^\theta$. Notice that, $L_{(s,a',x)}^1(a) = L_{(s,a',x)}(a)$ for all $s \in S$, $a, a' \in A$ and $x \in X$.

From now on we use the superscript θ to refer to the slow version of the relevant variable. If L is monotonically risk averse, then so is L^θ . Furthermore, if L satisfies the NSC then so does L^θ for all $\theta \in (0, 1]$. Hence, we can apply Lemma 3 to show that for arbitrary $s_0^\theta \in S^\theta$, $\sigma_t^{\theta,*}$ converges almost surely to a random variable, denoted by $\sigma_\infty^{\theta,*}$, whose support is contained in $\{0, 1\}$. Let $\Gamma^\theta(s_0) := \Pr\{\sigma_\infty^{\theta,*} = 1\}$ for a slow learning rule L^θ and $s_0 \in S$.

5.2 Convergence to second order stochastically dominant actions

The main result in this subsection is Proposition 3. It shows that slow versions of globally monotonically risk averse learning rules which satisfy the NSC converge, with high probability, to the set of actions that second order stochastically dominate all others.

The following lemma plays a critical role in the proof of Proposition 3. It provides a lower bound for $\Gamma^\theta(s_0)$ when $\theta = 1$. Specifically, it shows that we can construct a function $\phi : [0, 1] \rightarrow [0, 1]$ such that $\Gamma^\theta(s_0) \geq \phi(\sigma_0^*)$ for all s_0 . $\phi(\cdot)$ satisfies $\phi(0) = 0$ and $\phi(1) = 1$, and is strictly increasing and

strictly concave within an interval whose lower and upper bounds can be chosen to be arbitrarily close to zero and one, respectively. This function satisfies $\phi(\sigma_0^*) \geq \sigma_0^*$ for all $\sigma_0^* \in [0, 1]$. The proofs of this lemma and the next are provided in the Appendix.

Lemma 4 *Let L be a globally monotonically risk averse learning rule which satisfies the NSC and let $\varepsilon_1 \in (0, 1/2)$. Then, there exists $\gamma^* > 0$ such that the function $\phi : [0, 1] \rightarrow [0, 1]$ defined as*

$$\phi(\sigma_0^*) := \begin{cases} \frac{1-e^{-\gamma\sigma_0^*}}{1-e^{-\gamma}} & \sigma_0^* \in (\varepsilon_1, 1 - \varepsilon_1), \\ \sigma_0^* & \sigma_0^* \in [0, 1] \setminus (\varepsilon_1, 1 - \varepsilon_1), \end{cases}$$

satisfies the following conditions when $\gamma = \gamma^$:*

- (I) $E[\phi(\sigma_{t+1}^*)] - E[\phi(\sigma_t^*)] \geq 0$ for all t , and
- (II) $\Gamma(s_0) \geq \phi(\sigma_0^*) \forall s_0 \in S$.

The next lemma shows that γ can be taken to be arbitrarily large for slow learning rules while preserving the properties of $\phi(\cdot)$ obtained in Lemma 4.

Lemma 5 *Let L be a globally monotonically risk averse learning rule which satisfies the NSC and $\varepsilon_1 \in (0, 1/2)$. Then, for all $\hat{\gamma} > \gamma^*$, there exists a slow version of L , $L^{\hat{\theta}}$, such that $\Gamma^{\hat{\theta}}(s_0) \geq \phi(\sigma_0^*) \forall s_0 \in S$ when $\gamma = \hat{\gamma}$.*

As Lemma 4 reveals, for every globally monotonically risk averse learning rule, we can provide a lower bound for the probability of convergence to choose the safest action with probability one.¹³ This lower bound corresponds to $\phi(\sigma_0^*)$ for some $\gamma^* > 0$. This function is a lower bound because $E\phi(\sigma_t^*)$ is increasing in time and $E\phi(\sigma_\infty^*) = E[\sigma_\infty^*]$ (because $\phi(\sigma^*) = \sigma^*$ for $\sigma^* \in \{0, 1\}$, which is the support of σ_∞^* , as Lemma 3 reveals). Then, Lemma 5 shows that, if we consider the slow learning rule L^θ , $\phi(\sigma_0^*)$ is a lower bound for the probability of convergence if we set $\gamma = \hat{\gamma} := \gamma^*/\theta$ instead of $\gamma = \gamma^*$. It follows that by considering slow learning, i.e., θ small enough, $\hat{\gamma}$ can be chosen arbitrarily high. Finally, notice that, as $\gamma \rightarrow \infty$, $\phi(\sigma_0^*) \rightarrow 1$ for all $\sigma_0^* \in (\varepsilon_1, 1 - \varepsilon_1)$ and thus, $\Gamma(s_0) \rightarrow 1$, for all s_0 such that $\sigma_0^* \in (\varepsilon_1, 1 - \varepsilon_1)$.

¹³The technique for providing lower bounds is well known in computer science (e.g., Lakshmivarahan and Thathachar (1976) and Narendra and Thathachar (1989)).

As explained below, in the proof of Proposition 3, this yields the lower bound for the probability of convergence, $\phi(\sigma_0^*)$, as close to 1 as desired.¹⁴

Proposition 3 *Let L be a globally monotonically risk averse learning rule which satisfies the NSC. Then, for any $\varepsilon > 0$ and $s_0 \in S$ such that $\sigma_0^* > 0$, there exists a $\theta_\varepsilon \in (0, 1]$ such that $\Gamma^\theta(s_0) > 1 - \varepsilon$ for any slow version of L with $\theta \in (0, \theta_\varepsilon)$.*

Proof. Consider any $\varepsilon > 0$ and s_0 such that $\sigma_0^* > 0$. If we chose ε_1 such that $\varepsilon_1 < \sigma_0^* < 1 - \varepsilon_1$, Lemma 4 and Lemma 5 imply that there exists $\gamma^* > 0$ such that, for all $\hat{\gamma} > \gamma^*$, there exists $\hat{\theta} := \gamma^*/\hat{\gamma}$ which satisfies $\Gamma^{\hat{\theta}}(s_0) \geq \phi(\sigma_0^*) \forall s_0 \in S$ when $\gamma = \hat{\gamma}$. As $\gamma \rightarrow \infty$, $\phi(\sigma_0^*) \rightarrow 1$ for all $\sigma_0^* \in (\varepsilon_1, 1 - \varepsilon_1)$. Thus, we can choose $\gamma_\varepsilon > \gamma^*$ such that, if $\gamma = \gamma_\varepsilon$, $\phi(\sigma_0^*) > 1 - \varepsilon$. From the argument above, there exists $\hat{\theta}_\varepsilon \in (0, 1)$ such that $\Gamma^{\hat{\theta}_\varepsilon}(s_0) \geq \phi(\sigma_0^*) > 1 - \varepsilon$. Furthermore, since $\hat{\theta}_\varepsilon := \gamma^*/\gamma_\varepsilon$ and $\phi(\sigma_0^*) \rightarrow 1$ as $\gamma \rightarrow \infty$, for each $\theta \in (0, \hat{\theta}_\varepsilon)$ we can find a $\gamma > \gamma_\varepsilon$ such that $\theta = \gamma^*/\gamma$ and $\Gamma^\theta(s_0) \geq \phi(\sigma_0^*) > 1 - \varepsilon$. ■

Lemmas 3-5 and Proposition 3 reveal the different natures of the NSC and slow learning, and what each of them accomplishes. The NSC, as the proof of Lemma 3 reveals, impedes the expected change in the probability of choosing the safest action to go to zero while the probability of choosing such actions is in the interior of $[0, 1]$. In contrast, slow learning requires learning to be slow along the whole path of the learning process. Slow learning is often used in stochastic learning models to ensure that the actual path of the learning process stays close to the expected path with high probability.¹⁵ From (3), it follows that if the sequence $\{\sigma_t^*\}_{t=1}^\infty$ was deterministic and followed the path of its expected value, then the probability of choosing the safest action would converge to one. This allows our lower bound for the probability of convergence to the safest action to go to one when we consider slow monotonically risk averse learning rules.

To illustrate Proposition 3 consider March's (1996) fractional adjustment model over gains (Example 4 in Section 4). This learning rule is globally

¹⁴Notice that our result provides sufficient conditions for convergence with high probability for slow learning rules. This does not imply that slow learning is a necessary condition for convergence, yet the previous results in the machine learning literature and our analysis suggest that convergence is difficult to obtain without considering slow learning.

¹⁵Early references are Norman (1968, 1974). For more recent related results, see Izquierdo et al. (2007) and the references therein.

monotonically risk averse and satisfies the NSC. To see this, notice that for $a \neq a'$, $L_{(s,a',x)}(a) = (1 - \alpha)^x \sigma_a$. This is twice differentiable in x with strictly positive second derivative. Thus, if $F'_{a'}$ strictly second-order stochastically dominates $F_{a'}$, then

$$\inf_{\{s \in S: \sigma_a, \sigma_{a'} > 0\}} \left\{ \int (1 - \alpha)^x dF_{a'}(x) - \int (1 - \alpha)^x dF'_{a'}(x) \right\} > 0.$$

Slow versions of this rule for any $s \in \Delta(A)$, $a \in A$, and $x \geq 0$ are specified as

$$\begin{aligned} L_{(s,a,x)}^\theta(a) &= \sigma_a + \theta(1 - (1 - \alpha)^x)(1 - \sigma_a) \\ L_{(s,a',x)}^\theta(a) &= \sigma_a + \theta((1 - \alpha)^x - 1)\sigma_a \quad a' \neq a. \end{aligned}$$

Proposition 3 implies that sufficiently slow versions of it converge with high probability to the set of actions that SOSD all others, provided this set is chosen initially with positive probability. That is, Proposition 3 implies that for any $\varepsilon > 0$, and s_0 such that $\sigma_0^* > 0$, there exists a $\theta_\varepsilon \in (0, 1]$ such that $\Pr\{\sigma_\infty^{*,\theta} = 1\} > 1 - \varepsilon$ for any slow version of this rule, L^θ , such that $\theta \in (0, \theta_\varepsilon)$.

An analogous version of Proposition 3 holds for monotonically risk seeking rules. Such a result would pertain to an environment in which there is at least one action whose distribution is a mean preserving spread of all other actions. An action in this set would be the riskiest in terms of second order stochastic dominance. If this set of actions is initially chosen with positive probability and the inequality in the NSC is reversed, then a suitably slow version of any monotonically risk seeking rule would converge with high probability to this set.

Another analogy can be obtained for first order monotone rules. Formally, let $\sigma_t^{**} := \sum_{a \in A^{**}} \sigma_{at}$ and recall that a distribution F'_a is said to strictly first order stochastically dominate a distribution F_a if F'_a FOSDs F_a and there is some $x \in X$ such that $F'_a(x) \neq F_a(x)$.

Proposition 4 *Suppose L is a globally first-order monotone learning rule such that $n_{a'a}(F'_{a'}, F_{a'}) > 0$ for any pair of actions $a \neq a'$ and payoff distributions $F'_{a'}$ and $F_{a'}$ such that $F'_{a'}$ strictly-FOSDs $F_{a'}$. Then, (i) for arbitrary*

$s_0 \in S$, σ_t^{**} converges almost surely to a random variable σ_∞^{**} , with support in $\{0, 1\}$, and (ii) for any $\varepsilon > 0$ and $s_0 \in S$ such that $\sigma_0^{\theta, **} > 0$, there exists a $\theta_\varepsilon \in (0, 1]$ such that $\Pr\{\sigma_\infty^{\theta, **} = 1\} > 1 - \varepsilon$ for any slow version of L with $\theta \in (0, \theta_\varepsilon)$.

The argument of the proof of this proposition is analogous to the argument in the proof of Lemmas 3-5 and Proposition 3 and is omitted.

5.3 Convergence to expected utility maximization

Consider an expected utility maximizer whose preferences are represented by the Bernoulli utility function u and who knows the environment $F = (F_a)_{a \in A}$. Let $A_u^*(F)$ be the set of most preferred actions for a decision maker with Bernoulli utility function u , i.e.,

$$A_u^*(F) := \left\{ a \in A : \int u(x) dF_a(x) \geq \int u(x) dF_{a'}(x) \forall a' \in A \right\}.$$

Let $\sigma_t^*(u, F) := \sum_{a \in A_u^*(F)} \sigma_{at}$. The following Lemma provides conditions on a learning rule L which ensure that $\sigma_t^*(u, F)$ is close to 1 in the long run. We say that $c + du$ is a *negative affine transformation* of u if $d < 0$ and $c \in \mathbb{R}$.

Lemma 6 *Let L be a globally impartial learning rule such that (i) $L_{(s, a', x)}(a)$ is a negative affine transformation of $u(x)$ for all $s \in S$, $a \in A$ and $a' \in A \setminus \{a\}$, and (ii) $n_{a'a}(F'_{a'}, F_{a'}) > 0$ for any pair of actions $a \neq a'$ and payoff distributions $F_{a'}$ and $F'_{a'}$ such that $\int u(x) dF'_{a'}(x) > \int u(x) dF_{a'}(x)$. Then, for all $\varepsilon > 0$ and s_0 with $\sigma_0^*(u, F) > 0$, there exists a $\theta_\varepsilon \in (0, 1]$ such that for any slow version L^θ with $\theta \in (0, \theta_\varepsilon)$, $\Pr\{\sigma_\infty^{*, \theta}(u, F) = 1\} > 1 - \varepsilon$.*

Proof. Impartiality implies

$$f_s(a) = \sum_{a' \neq a} \sigma_{a'} \left[\int L_{(s, a', x)}(a) dF_{a'}(x) - \int L_{(s, a', x)}(a) dF_a(x) \right]$$

for all $a \in A$. For any action $a \in A_u^*(F)$, we have $f_s(a) \geq 0$ because $L_{(s, a', x)}(a)$ is a negative affine transformation of $u(x)$ for all $a \in A$ and $a' \neq a$. Since, this is required to hold for all s , it follows that $f_{s_t}(A_u^*(F)) \geq 0$ for all

t. The rest of the argument is analogous to the arguments in the proofs of Lemmas 3-5, and Proposition 3 and is omitted.¹⁶ ■

Our last result shows that for any continuous Bernoulli utility function u there exists a learning rule L such that $\sigma_t^*(u, F)$ is close to 1 in the long run, with high probability.

Proposition 5 *Consider any environment F and continuous Bernoulli utility function $u : [x_{\min}, x_{\max}] \rightarrow \mathbb{R}$. There exists a learning rule L such that $f_s(a) \geq 0$ for all $a \in A_u^*(F)$ and $s \in S$; and for all $\varepsilon > 0$, if s_0 satisfies $\sigma_0^*(u, F) > 0$, then there exists $\theta_\varepsilon \in (0, 1]$ such that for any slow version L^θ with $\theta \in (0, \theta_\varepsilon)$, $\Pr\{\sigma_\infty^{*,\theta}(u, F) = 1\} > 1 - \varepsilon$.*

Proof. Consider a continuous function $u : [x_{\min}, x_{\max}] \rightarrow \mathbb{R}$. By Lemma 6, we just need to provide a globally impartial learning rule L such that conditions (i) and (ii) of the lemma are satisfied. In order to construct such a learning rule, for each $a \in A$ and $a' \in A \setminus \{a\}$, consider an affine transformation of u , denoted by $u_{a'a} : [x_{\min}, x_{\max}] \rightarrow [0, 1]$ such that $u_{aa'} = u_{a'a}$. Let $L_{(s,a',x)}(a) = \sigma_a(1 - u_{a'a}(x))$ for all $s \in S$, $a \in A$, $a' \in A \setminus \{a\}$ and $x \in [x_{\min}, x_{\max}]$. Notice that these conditions define a learning rule L which satisfies (i) and (ii) in Lemma 6. Finally, impartiality follows from

$$\begin{aligned} \sum_{a' \neq a} \sigma_{a'} L_{(s,a',x)}(a) + \sigma_a L_{(s,a,x)}(a) &= \sum_{a' \neq a} \sigma_{a'} \sigma_a (1 - u_{a'a}(x)) + \sigma_a \left(1 - \sum_{a' \neq a} L_{(s,a,x)}(a') \right) \\ &= \sum_{a' \neq a} \sigma_{a'} \sigma_a (1 - u_{a'a}(x)) + \sigma_a - \sum_{a' \neq a} \sigma_a \sigma_{a'} (1 - u_{a'a}(x)) \\ &= \sigma_a. \end{aligned}$$

■

To illustrate Proposition 5, we provide 3 examples. First, we revisit the March (1996) fractional adjustment learning rule over gains (Section 4, Example 4). We have shown that this learning rule is globally impartial. The off-diagonal terms of this rule can be written as

$$\begin{aligned} L_{(s,a'x)}(a) &= (1 - \alpha)^x \sigma_a \\ &= \sigma_a e^{-\left(\ln \frac{1}{1-\alpha}\right)x}. \end{aligned}$$

¹⁶Note that condition (ii) plays the role that the NSC plays in the proof of Proposition 3.

This expression corresponds to a negative affine transformation of the *Constant Absolute Risk Aversion (CARA)* Bernoulli utility function $u_{CARA}(x) = -e^{-bx}$ with Arrow-Pratt coefficient of absolute risk aversion $b = \ln \frac{1}{1-\alpha}$. For any pair of distributions $F_{a'}$ and $F'_{a'}$ such that $\int u_{CARA}(x)dF'_{a'}(x) > \int u_{CARA}(x)dF_{a'}(x)$,

$$n_{a'a}(F'_{a'}, F_{a'}) = \int e^{-(\ln \frac{1}{1-\alpha})x} dF_{a'}(x) - \int e^{-(\ln \frac{1}{1-\alpha})x} dF'_{a'}(x) > 0.$$

Hence, condition (ii) in Lemma 6 is satisfied. The results above imply that the asymptotic probability of choosing an action which maximizes the expected utility of a CARA decision maker with Arrow-Pratt coefficient of absolute risk aversion, $b = \ln \frac{1}{1-\alpha}$, can be made as close to one as desired for a suitably slow version of March's fractional adjustment learning rule.

Next, we consider a simple modification of the Roth and Erev learning rule (Section 4, Example 2) in which we keep the sum of attractions constant. Additionally, suppose all payoffs are strictly positive. For this rule, the state space is $S = \left\{ (s_a)_a \in \mathbb{R}_{++}^{|A|} : \sum_a s_a = \bar{S} \right\}$. If action a is chosen and it gets a payoff of x then its attraction becomes $(s_a + x) (\sum s_{a'} / (\sum s_{a'} + x)) = (s_a + x) (\bar{S} / (\bar{S} + x))$. If a is not chosen then its attraction becomes $s_a (\bar{S} / (\bar{S} + x))$. L_1 is as in the original model, i.e., $\sigma_a = s_a / \sum_{a'} s_{a'}$ for all a . This learning rule is given by

$$\begin{aligned} L_{(s,a,x)}(a) &= \frac{s_a + x}{\bar{S} + x} \\ L_{(s,a',x)}(a) &= \frac{s_a}{\bar{S} + x} \quad \forall a' \neq a. \end{aligned}$$

The functions $L_{(s,a',x)}(a)$ are negative affine transformations of the utility function $u(x) = -1/(\bar{S} + x)$, for all $a' \neq a$. To see that this rule satisfies condition (ii) in Lemma 6, notice that

$$\begin{aligned} n_{a'a}(F'_{a'}, F_{a'}) &= \int \bar{S} / (\bar{S} + x) dF_{a'}(x) - \int \bar{S} / (\bar{S} + x) dF'_{a'}(x) \\ &= \bar{S} \left(\int 1 / (\bar{S} + x) dF_{a'}(x) - \int 1 / (\bar{S} + x) dF'_{a'}(x) \right). \end{aligned}$$

If $\int u(x) dF'_{a'}(x) > \int u(x) dF_{a'}(x)$, this expression is strictly greater than zero. Our results imply that the asymptotic probability of choosing an

action, which maximizes the expected utility of a decision maker whose preferences are represented by the Bernoulli utility function $u(x) = \frac{-1}{\bar{S}+x}$, can be made as close to one as desired by choosing a small enough θ to construct an slow version of this rule.

As $\bar{S} \rightarrow 0$, since we have assumed that $0 < x_{\min} < x_{\max} < \infty$, $u(x) = \frac{-1}{\bar{S}+x}$ converges uniformly to the Constant Relative Risk Aversion (CRRA) utility function $\tilde{u}(x) = \frac{x^{1-\alpha}}{1-\alpha}$ with coefficient of relative risk aversion $\alpha = 2$. Hence, for every environment F , there exists a small enough \bar{S} , call it \bar{S}_F , such that at least one of the maximizers of \tilde{u} is a maximizer of u . Therefore, for a slow enough version of the learning rule associated to \bar{S}_F , the probability of converging to one of the maximizers of \tilde{u} in F can be made arbitrarily close to one.¹⁷

Finally, consider a “generalized” version of the Cross learning rule (Section 4, Example 1). For any action $a \in A$, define the generalized Cross rule as

$$\begin{aligned} L_{(s,a,x)}(a) &= \sigma_a + (1 - \sigma_a) u(x) \\ L_{(s,a',x)}(a) &= \sigma_a - \sigma_a u(x) \quad \forall a' \neq a, \end{aligned}$$

where $s := (\sigma_a)_{a \in A} \in \Delta(A)$ and $u : X \rightarrow [0, 1]$ is a continuous function. Notice that the generalized Cross rule arises when we set $u_{a'a}(x) = u(x)$ for all $x \in X$, $a \in A$, and $a' \in A \setminus \{a\}$ in the construction of the proof of Proposition 5. Furthermore, it contains the above two examples as special cases. As in our previous examples, this rule can be shown to satisfy the hypotheses of Lemma 6. Hence, the asymptotic probability of choosing an action which maximizes the expected utility of a decision maker, whose Bernoulli utility function is u , can be made as close to one as desired by choosing a small enough θ for a slow version of the generalized Cross learning rule.

The proof of Proposition 5 reveals that for every utility function one can consider several different learning rules that converge to choose the most preferred actions for such a decision maker. There are other learning rules in the literature that converge to choose with high probability expected utility maximizing actions, including, for instance, the belief-based learning rules

¹⁷As $\bar{S} \rightarrow 0$, slow versions of this rule converge to the Bush-Mosteller (1955) learning rule, i.e., $\lim_{\bar{S} \rightarrow 0} L_{(s,a,x)}(a) = (1-\theta)\sigma_a + \theta$, and $\lim_{\bar{S} \rightarrow 0} L_{(s,a',x)}(a) \rightarrow (1-\theta)\sigma_a$ for $a' \neq a$.

that we discussed in Section 4.¹⁸ Nevertheless, in contrast to the rules that satisfy the hypothesis of Lemma 6, belief-based learning rules are not impartial and hence, cannot satisfy globally the short run property $f_s(a) \geq 0$ for all $a \in A_u^*(F)$ and $s \in S$ in every environment.

6 Discussion

Our results do not have implications for the risk attitudes we observe when individuals decide among known payoff distributions. In particular, our study is not aimed to provide an explanation for the origin of risk preferences in EUT. It instead explains how the risk choices of individuals, who do not have knowledge of payoff distributions, change in response to experience and how individuals exhibit risk averse (or risk seeking) behavior that mimics the choices of expected utility maximizers who know the payoff distributions. Neither do our results imply that individuals are more likely to behave as risk averse (or risk seeking) individuals as a consequence of learning than when distributions are known. An approach that accommodates decision making both with and without knowledge (our beliefs about) payoff distributions could address this issue. This is left for further research.

A weaker notion of risk aversion for learning rules requires that, for any payoff distribution, the learning rule is expected to add more probability to an action when it gives the expected value of the distribution with certainty than when it gives a random payoff drawn from that distribution, provided that all the other aspects of the environment are the same. We refer to the learning rules that satisfy this property as *risk averse* learning rules. A learning rule is risk averse if and only if it is own-concave. Therefore monotonically risk averse learning rules are risk averse. The expected value of the probability of choosing an action whose payoff distribution becomes riskier decreases when the learning rule is risk averse. Risk averse learning rules that are not monotonically risk averse, however, do not, in general,

¹⁸For the belief-based learning rules we described in Section 4, this amounts to take arbitrarily large values of ρ . As $\rho \rightarrow \infty$ the probability that the learning rule chooses the action with the highest average utility goes to one. Furthermore, since $u(x_{min}) > 0$ and $u(x_{max}) < \infty$, the probability of choosing each action is bounded from below away from zero, which guarantees that each action is chosen infinitely often and that the average of $u(x)$ converges to $\int u(x)dF_a$ for all $a \in A$. This yields the result.

converge to choose the safest actions with high probability, even for slow versions that satisfy the NSC.

An alternative definition of when a learning rule may be considered to be risk averse can be constructed looking at the expected behavior in the next period, $(\sigma_a + f(a))_{a \in A}$. This probability distribution on actions generates a (reduced) distribution over payoffs, which is a weighted average of the payoff distribution of each of the actions, $\sum_a (\sigma_a + f(a)) F_a$.¹⁹ We could ask whether the learning rule is such that expected behavior tomorrow generates a distribution over payoffs which second-order stochastically dominates that of today, i.e., whether $\sum_a (\sigma_a + f(a)) F_a$ SOSDs $\sum_a \sigma_a F_a$, in every environment. It can be shown that, in environments with only two actions, the only learning rules which satisfy this condition are the unbiased rules studied by Börgers et al. (2004).²⁰ Unbiased rules exhibit zero expected movement in probability when all actions have the same expected payoffs. Such rules satisfy the above condition in a trivial manner because the expected distribution tomorrow is the same as today.²¹ We conclude, therefore, that such a property is too restrictive. Restricting the set of environments on which the improvement is required would lead us to identify a larger class of learning rules.²² We do not pursue such an approach in this paper.

Whereas EUT provides restrictions on the von Neumann and Morgenstern utility function of risk averse individuals, the analysis of monotonically risk averse learning provides restrictions on the matrix of functions that specifies a learning rule. Hence, our analysis provides a similar level of generality and detail in describing learning rules as the analysis of utility functions in EUT. This paper, however, does not provide an axiomatic foundation for the learning rules we study. The properties studied here, while normatively appealing, are not intended to be behaviorally descriptive. This is a subject for further study. The papers by Gilboa and Schmeidler (1995) and Easley

¹⁹As in section 3, we do not introduce here explicitly the state in the notation.

²⁰See Claim 1 in the Appendix for the proof.

²¹It can also be shown that unbiased learning rules are the only learning rules which are continuous in x for all $a, a' \in A$ and satisfy $\sum_a (\sigma_a + f(a)) F_a$ sosd $\sum_a \sigma_a F_a$ in every environment. For details, see Claim 2 in the Appendix.

²²For example, we could consider learning rules that satisfy $\sum_a (\sigma_a + f(a)) F_a$ sosd $\sum_a \sigma_a F_a$ in every environment that is completely ordered by the sosd relation. It can be shown that a learning rule is continuous in payoffs and satisfies this condition if and only if it is monotonically risk-averse.

and Rustichini (1999) make important contributions toward this objective.²³ Both those papers, however, adopt slightly different frameworks from the one we consider. More work needs to be done to provide an axiomatic foundation for learning rules within the framework of this paper.

Finally, our results on convergence of slow globally monotonically risk averse learning rules provide only sufficient conditions. We are not aware of convergence results, at least for a suitably general class of payoff distributions, that allow for convergence without imposing some form of slow learning. Furthermore, Narendra and Thathachar (1989) show existence of upper bounds away from one for the probability of convergence to optimal actions for non-slow learning automata in a binary outcome model. This suggests that slow learning also may be a necessary condition in our framework as well. Yet, we have not been able to generalize their analysis of upper bounds to our setting. Whether slow learning is necessary for convergence is a question that deserves further attention in future research.

7 Appendix

Proof of Proposition 2

We begin with the following lemma, the proof of which closely parallels that of Lemma 1.

Lemma 7 *Every first order monotone learning rule L is impartial.*

Proof. We argue by contradiction. Consider an environment F with $A = A^{**}$ and suppose that $f(a) < 0$ for some $a \in A$. Now we construct an environment \widehat{F} where $\widehat{A}^{**} = \{a\}$. We construct \widehat{F}_a by assigning to every

²³Easley and Rustichini (1999) analyze an alternative setting where individuals can observe forgone payoffs. Yet, their analysis, in contrast to ours, focuses on the implications of certain axioms on the learning rules rather than in their properties. Extending our results to a setting with information about forgone payoffs is also an interesting direction for future research. Oyarzun and Ruf (2009) provide some progress in that direction in the context of boundedly rational social learning models.

interval $I \subset X$ only $(1 - \varepsilon)$ times the probability it had under F_a and adding ε to the probability of x_{\max} . We construct $\widehat{F}_{a'}$ for all $a' \in A \setminus \{a\}$ by assigning to every interval $I \subset X$ only $(1 - \varepsilon)$ times the probability it had under $F_{a'}$ and then adding ε to the probability of x_{\min} . Clearly, $\widehat{A}^{**} = \{a\}$. Since $\widehat{f}(a)$ can be written as a continuous function in ε , for small enough ε we have $\widehat{f}(a) < 0$. Therefore L is not first order monotone. ■

Now we provide the proof of the proposition.

Proof. *Necessity.*

The necessity of (i) follows from Lemma 7 and Lemma 2. To prove the necessity of (ii) we argue by contradiction. Suppose that for some $a \in A$ and $a' \in A \setminus \{a\}$ such that $\sigma_{a'} > 0$, there are x and x' with $x' < x$ and $L_{(a',x)}(a) > L_{(a',x')}(a)$. Consider the environment F where action a pays x with probability one and action a' pays x' with probability one. All the other actions $a'' \in A \setminus \{a, a'\}$, if any, pay x with probability $1 - \varepsilon$ and x' with probability ε . Clearly, $A^{**} = \{a\}$. From Lemma 7 and Lemma 2, we have that

$$\begin{aligned} f(a) &= \sum_{a'' \neq a} \sigma_{a''} \left[\int L_{(a'',x)}(a) dF_{a''}(x) - \int L_{(a'',x)}(a) dF_a(x) \right] \\ &= \sigma_{a'} [L_{(a',x')}(a) - L_{(a',x)}(a)] + \varepsilon \sum_{a'' \neq a, a'} \sigma_{a''} [L_{(a'',x')}(a) - L_{(a'',x)}(a)]. \end{aligned}$$

For small enough ε , $f(a) < 0$, which contradicts first order monotonicity.

Sufficiency.

As in the proof of Proposition 1, consider $a \in A^{**}$, then

$$\begin{aligned} f(a) &= \sum_{a' \neq a} \sigma_{a'} \left[\int L_{(a',x)}(a) dF_{a'}(x) - \int L_{(a',x)}(a) dF_a(x) \right] \\ &\geq 0. \end{aligned}$$

The last inequality follows from the fact that $a \in A^{**}$ and the fact that $L_{(a',x)}(a)$ is decreasing for all $a' \in A \setminus \{a\}$ such that $\sigma_{a'} > 0$. ■

Proof of Lemma 4

Proof. (I) For $n = 1, 2, \dots$ let the function $\phi_n : [0, 1] \rightarrow [0, 1]$ be defined as ϕ with $\gamma = \gamma_n$ and $\gamma_n := n^{-1}$, i.e.,

$$\phi_n(\sigma_0^*) := \begin{cases} \frac{1 - e^{-\gamma_n \sigma_0^*}}{1 - e^{-\gamma_n}} & \sigma_0^* \in (\varepsilon_1, 1 - \varepsilon_1), \\ \sigma_0^* & \sigma_0^* \in [0, 1] \setminus (\varepsilon_1, 1 - \varepsilon_1). \end{cases}$$

ϕ_n converges to the identity function on $[0, 1]$. Then, by the Dominated Convergence Theorem (for example, see Resnick 1999, p.133),

$$\lim_{n \rightarrow \infty} E [\phi_n(\sigma_{t+1}^*) | s_t] = E [\sigma_{t+1}^* | s_t]$$

for all s_t . Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} [E [\phi_n(\sigma_{t+1}^*) | s_t] - \phi_n(\sigma_t^*)] &= \lim_{n \rightarrow \infty} E [\phi_n(\sigma_{t+1}^*) | s_t] - \lim_{n \rightarrow \infty} \phi_n(\sigma_t^*) \\ &= E [\sigma_{t+1}^* | s_t] - \sigma_t^* \\ &\geq \delta \sigma_t^* (1 - \sigma_t^*), \end{aligned}$$

where the last inequality follows from (3), and $\delta > 0$ is defined as in the proof of Lemma 3. Hence, there exists $n^* \in \mathbb{N}$ such that $E [\phi_{n^*}(\sigma_{t+1}^*) | s_t] - \phi_{n^*}(\sigma_t^*) \geq 0$ for all $s_t \in S$ such that $\sigma_t^* \in (\varepsilon_1, 1 - \varepsilon_1)$.²⁴ For $\sigma_t^* \in [0, 1] \setminus (\varepsilon_1, 1 - \varepsilon_1)$, $\phi_n(\sigma_t^*) = \sigma_t^*$. Since $\phi_n(\sigma_{t+1}^*) \geq \sigma_{t+1}^*$ for all $\sigma_{t+1}^* \in [0, 1]$, $E [\phi_n(\sigma_{t+1}^*) | s_t] \geq E [\sigma_{t+1}^* | s_t] \geq \sigma_t^* = \phi_n(\sigma_t^*)$ for all n . It follows that

$$E [\phi_{n^*}(\sigma_{t+1}^*) | s_t] - \phi_{n^*}(\sigma_t^*) \geq 0 \tag{5}$$

for all $\sigma_t^* \in [0, 1]$, for all t . Hence, if we set $\gamma = \gamma^* := 1/n^*$, we obtain that $E [\phi(\sigma_{t+1}^*) | s_t] - \phi(\sigma_t^*) \geq 0$, for all t . Taking the expected value of the left hand side, we obtain $E [E [\phi(\sigma_{t+1}^*) | s_t]] - E [\phi(\sigma_t^*)] \geq 0$. Thus, $E [\phi(\sigma_{t+1}^*)] - E [\phi(\sigma_t^*)] \geq 0$ for all t and $s_0 \in S$.

(II) For $\gamma > 0$, ϕ is a bounded and it is discontinuous only at ε_1 and $1 - \varepsilon_1$. Hence, from Lemma 3 and the Continuous Mapping Theorem (footnote 12), $\lim_{t \rightarrow \infty} E [\phi(\sigma_t^*)] = E [\phi(\sigma_\infty^*)]$. Therefore, from (I), $E [\phi(\sigma_\infty^*)] \geq \phi(\sigma_0^*)$. From Lemma 3, we also know that the support of σ_∞^* is contained in $\{0, 1\}$. Thus, since $E [\phi(\sigma_\infty^*)] = [1 - \Gamma(s_0)]\phi(0) + \Gamma(s_0)\phi(1)$, $\phi(0) = 0$, and $\phi(1) = 1$, we obtain $E [\phi(\sigma_\infty^*)] = \Gamma(s_0)$. It follows that $\Gamma(s_0) \geq \phi(\sigma_0^*)$. ■

²⁴This statement uses the fact that ϕ converges uniformly to the identity function.

Proof of Lemma 5

Proof. From Lemma 4, when $\gamma = \gamma^*$, $E[\phi(\sigma_{t+1}^*)|s_t] - \phi(\sigma_t^*) \geq 0$ for all $s_t \in S$, therefore for all $\sigma_t^* \in (\varepsilon_1, 1 - \varepsilon_1)$ we have that L satisfies

$$E \left[\frac{1 - e^{-\gamma^* \sigma_{t+1}^*}}{1 - e^{-\gamma^*}} | s_t \right] - \frac{1 - e^{-\gamma^* \sigma_t^*}}{1 - e^{-\gamma^*}} \geq 0$$

and hence, $-e^{-\gamma^* \sigma_t^*} \left(E \left[e^{-\gamma^* (\sigma_{t+1}^* - \sigma_t^*)} | s_t \right] - 1 \right) \geq 0$.

For any $\hat{\gamma} > \gamma^*$, let $\hat{\theta} := \gamma^* / \hat{\gamma}$, hence $\gamma^* = \hat{\gamma} \hat{\theta}$. Therefore for all $\sigma_t^* \in (\varepsilon_1, 1 - \varepsilon_1)$ we obtain

$$\begin{aligned} -e^{-\hat{\gamma} \hat{\theta} \sigma_t^*} \left(E \left[e^{-\hat{\gamma} \hat{\theta} (\sigma_{t+1}^* - \sigma_t^*)} | s_t \right] - 1 \right) &\geq 0 \\ -e^{-\hat{\gamma} \sigma_t^*} \left(E \left[e^{-\hat{\gamma} \hat{\theta} (\sigma_{t+1}^* - \sigma_t^*)} | s_t \right] - 1 \right) &\geq 0. \end{aligned}$$

It follows that for $L^{\hat{\theta}}$,

$$E \left[\frac{1 - e^{-\hat{\gamma} \sigma_{t+1}^*}}{1 - e^{-\hat{\gamma}}} | s_t \right] - \frac{1 - e^{-\hat{\gamma} \sigma_t^*}}{1 - e^{-\hat{\gamma}}} \geq 0$$

for all $s_t \in S$ such that $\sigma_t^* \in (\varepsilon_1, 1 - \varepsilon_1)$. Then, as in the proof of Lemma 4, we obtain that for $\gamma = \hat{\gamma}$, $L^{\hat{\theta}}$ satisfies $E[\phi(\sigma_{t+1}^*)] - E[\phi(\sigma_t^*)] \geq 0$ for all $\sigma_t^* \in [0, 1]$, $t = 0, 1, \dots$, and hence, $\Gamma^{\hat{\theta}}(s_0) \geq \phi(\sigma_0^*) \forall s_0 \in S$. ■

Claim 1 *For problems such that $|A| = 2$, the only learning rules which satisfy $\Sigma_a(\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment are the unbiased rules.*

Proof. First we prove that impartiality is a necessary condition for a learning rule to satisfy $\Sigma_a(\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment, for any finite number of actions. Suppose L is not impartial, then there exists an environment F , such that $F_a = F_{a'}$ for all $a, a' \in A$; and $f(a) > 0$ for some $a \in A$. Let $\varepsilon \in (0, 1)$ and consider the environment \tilde{F} such that \tilde{F}_a is constructed as a compound distribution of F_a and $\left(x_{\min}, \frac{x_{\max} - \pi_a}{x_{\max} - x_{\min}}; x_{\max}, \frac{\pi_a - x_{\min}}{x_{\max} - x_{\min}} \right)$ with respective weights $1 - \varepsilon$ and ε , while $\tilde{F}_{a'}$ is a compound distribution of $F_{a'}$ and $(\pi_a, 1)$ with respective weights $1 - \varepsilon$ and ε for all $a' \in A \setminus \{a\}$. Notice that $\tilde{F}_{a'}$

strictly-SOSD \tilde{F}_a for all $a' \neq a$. Since $\tilde{f}(a)$ is continuous in ε for all $a \in A$, for small enough ε , $\tilde{f}(a) > 0$. Thus, $\Sigma_a \sigma_a \tilde{F}_a$ strictly-SOSD $\Sigma_a (\sigma_a + f(a)) \tilde{F}_a$.

Now we suppose $|A| = 2$ and prove that $L_{(a',x)}(a)$ is both a concave and a convex function of x for all a' and a . Since L has to be impartial,

$$f(a) = \sigma_{a'} \left[\int L_{(a',x)}(a) dF_{a'}(x) - \int L_{(a',x)}(a) dF_a(x) \right]$$

where $a' \neq a$. Suppose $L_{(a',x)}(a)$ is not convex, thus there exists $x'', x' \in X$ such that $L_{(a',x)}(a) > \lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a)$ for some $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda)x''$. Consider an environment where F_a provides x with probability 1 and $F_{a'}$ pays x' with probability λ and x'' with probability $(1 - \lambda)$. Therefore

$$\begin{aligned} f(a) &= \sigma_{a'} [\lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a) - L_{(a',x)}(a)] \\ &< 0. \end{aligned}$$

It follows that $\Sigma_a \sigma_a F_a$ strictly-SOSD $\Sigma_a (\sigma_a + f(a)) F_a$. It follows that a necessary condition for a learning rule to satisfy $\Sigma_a (\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment is that $L_{(a',x)}(a)$ is convex for $a' \neq a$.

Notice that this implies that $L_{(a,x)}(a)$ is concave for $a \in A$ and hence, $L_{(a,x)}(a)$ and $L_{(a',x)}(a)$ are continuous functions of x for all a and $a' \neq a$.

Suppose $L_{(a',x)}(a)$ is not concave, thus there exists $x'' > x' \in X$ such that $L_{(a',x)}(a) < \lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a)$ for some $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda)x''$. Consider an environment where F_a corresponds to the lottery $(x', \delta\lambda; x, 1 - \delta; x'', (1 - \lambda)\delta)$ and $F_{a'}$ corresponds to the lottery $(x' + \frac{1-\lambda}{\lambda}\rho, \lambda; x'' - \rho, 1 - \lambda)$ for $\rho > 0$. Therefore,

$$\begin{aligned} f(a) &= \sigma_{a'} [\lambda L_{(a',x'+\frac{1-\lambda}{\lambda}\rho)}(a) + (1 - \lambda)L_{(a',x''-\rho)}(a) - \\ &\quad (\delta\lambda L_{(a',x')}(a) + (1 - \delta)L_{(a',x)}(a) + (1 - \lambda)\delta L_{(a',x'')}(a))] \\ &= \sigma_{a'} [\lambda L_{(a',x'+\frac{1-\lambda}{\lambda}\rho)}(a) + (1 - \lambda)L_{(a',x''-\rho)}(a) - L_{(a',x)}(a) - \\ &\quad \delta (\lambda L_{(a',x')}(a) - L_{(a',x)}(a) + (1 - \lambda)L_{(a',x'')}(a))]. \end{aligned}$$

Since $L_{(a',x)}(a)$ is continuous, for small enough ρ , in the RHS of the last equality, we have $\lambda L_{(a',x'+\frac{1-\lambda}{\lambda}\rho)}(a) + (1 - \lambda)L_{(a',x''-\rho)}(a) - L_{(a',x)}(a) > 0$. Thus for small enough ρ and δ , we have $f(a) > 0$. But F_a puts weights on the boundaries of the interval $[x', x'']$ while the support of $F_{a'}$ is contained in (x', x'') . It follows that $\Sigma_a (\sigma_a + f(a)) F_a$ does not SOSD $\Sigma_a \sigma_a F_a$. Hence $L_{(a',x)}(a)$ is both a concave and a convex function of x for all a' and a . ■

Claim 2 *The only learning rules which satisfy $\Sigma_a(\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment such that $L_{(a',x)}(a)$ is continuous on x for all $a, a' \in A$ are the unbiased rules.*

Proof. From the proof of Claim 2, learning rules which satisfy $\Sigma_a(\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment are impartial. Now we prove that continuous rules which satisfy $\Sigma_a(\sigma_a + f(a)) F_a$ SOSD $\Sigma_a \sigma_a F_a$ in every environment also need to be both cross-convex and cross concave. Since these rules are impartial, we have

$$f(a) = \sum_{a' \neq a} \sigma_{a'} \left[\int L_{(a',x)}(a) dF_{a'}(x) - \int L_{(a',x)}(a) dF_a(x) \right]$$

for all a . Suppose there is $a \in A$ and $a' \neq a$ such that $L_{(a',x)}(a)$ is not convex. Thus there exists $x'' > x' \in X$ such that $L_{(a',x)}(a) > \lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a)$ for some $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda)x''$. Consider an environment where F_a corresponds to the distribution $(x', \lambda; x'', 1 - \lambda)$ and $F_{a'}$ provides x with probability 1. All the other actions $a'' \neq a, a'$, if any, have payoff distributions $F_{a''}$ corresponding to $(x' + \frac{1-\lambda}{\lambda}\varepsilon, \lambda; x'' - \varepsilon, 1 - \lambda)$ for $\varepsilon > 0$. Therefore

$$f(a) = \sigma_{a'} [L_{(a',x)}(a) - \lambda L_{(a',x')}(a) - (1 - \lambda)L_{(a',x'')}(a) + \sum_{a'' \neq a, a'} \sigma_{a''} \left(\lambda L_{(a'', x' + \frac{1-\lambda}{\lambda}\varepsilon)}(a) + (1 - \lambda)L_{(a'', x'' - \varepsilon)}(a) - \lambda L_{(a',x')}(a) - (1 - \lambda)L_{(a',x'')}(a) \right)].$$

The first term of the sum in the RHS is positive and all the other terms go to zero as ε goes to zero. Therefore, for small enough ε , $f(a) > 0$. Thus, since F_a is the only distribution that puts weights on the boundaries of the interval $[x', x'']$ and the support of all the other distributions is contained in (x', x'') , $\Sigma_a(\sigma_a + f(a)) F_a$ does not SOSD $\Sigma_a \sigma_a F_a$.

Suppose there is $a \in A$ and $a' \neq a$ such that $L_{(a',x)}(a)$ is not concave. Thus there exists $x'' > x' \in X$ such that $L_{(a',x)}(a) < \lambda L_{(a',x')}(a) + (1 - \lambda)L_{(a',x'')}(a)$ for some $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda)x''$. Consider an environment where F_a corresponds to the lottery $(x', \delta\lambda; x, 1 - \delta; x'', (1 - \lambda)\delta)$ and $F_{a'}$ corresponds to the lottery $(x' + \frac{1-\lambda}{\lambda}\rho, \lambda; x'' - \rho, 1 - \lambda)$ for $\rho > 0$. All the other actions $a'' \neq a, a'$, if any, have payoff distributions $F_{a''}$ corresponding

to $(x' + \frac{1-\lambda}{\lambda}\varepsilon, \delta\lambda; x, 1 - \delta; x'' - \varepsilon, \delta(1 - \lambda))$. Therefore

$$\begin{aligned}
f(a) &= \sigma_{a'}[\lambda L_{(a', x' + \frac{1-\lambda}{\lambda}\rho)}(a) + (1 - \lambda)L_{(a', x'' - \rho)}(a) \\
&\quad - \delta\lambda L_{(a', x')}(a) - (1 - \delta)L_{(a', x)}(a) - (1 - \lambda)\delta L_{(a', x'')}(a)] + \\
&\quad \sum_{a'' \neq a, a'} \sigma_{a''}(\delta\lambda L_{(a'', x' + \frac{1-\lambda}{\lambda}\varepsilon)}(a) + (1 - \delta)L_{(a'', x)}(a) + \delta(1 - \lambda)L_{(a'', x'' - \varepsilon)}(a) \\
&\quad - \delta\lambda L_{(a'', x')}(a) - (1 - \delta)L_{(a'', x)}(a) - (1 - \lambda)\delta L_{(a'', x'')}(a)) \\
&= \sigma_{a'}[\lambda L_{(a', x' + \frac{1-\lambda}{\lambda}\rho)}(a) + (1 - \lambda)L_{(a', x'' - \rho)}(a) - L_{(a', x)}(a) + \\
&\quad \delta(L_{(a', x)}(a) - \lambda L_{(a', x')}(a) - (1 - \lambda)L_{(a', x'')}(a))] + \\
&\quad \sum_{a'' \neq a, a'} \sigma_{a''}(\delta\lambda L_{(a'', x' + \frac{1-\lambda}{\lambda}\varepsilon)}(a) + (1 - \delta)L_{(a'', x)}(a) + \delta(1 - \lambda)L_{(a'', x'' - \varepsilon)}(a) \\
&\quad - \delta\lambda L_{(a'', x')}(a) - (1 - \delta)L_{(a'', x)}(a) - (1 - \lambda)\delta L_{(a'', x'')}(a)).
\end{aligned}$$

In the last equality, for small enough ρ and δ , the term inside the square bracket is strictly greater than zero. Furthermore, all the other terms go to zero as ε goes to zero. Therefore, for small enough ε , ρ , and δ , $f(a) > 0$. Thus $\Sigma_a(\sigma_a + f(a))F_a$ does not SOSD $\Sigma_a\sigma_a F_a$, since F_a is the only distribution that puts weights on the boundaries of the interval $[x', x'']$ and the support of all the other payoff distributions is contained in (x', x'') . ■

REFERENCES

1. Beggs, A (2005): "On the convergence of reinforcement learning," *Journal of Economic Theory*, 122, 1-36.
2. Billingsley, P. (1995): *Probability and Measure, Third Edition*. John Wiley & Sons, Inc.
3. Börgers, T., A. J. Morales and R. Sarin (2004): "Expedient and monotone learning rules," *Econometrica*, 72, 383-405.
4. Börgers, T. and R. Sarin (1997): "Learning through reinforcement and replicator dynamics," *Journal of Economic Theory*, 77, 1-14.
5. Börgers, T. and R. Sarin (2000): "Naive reinforcement learning with endogenous aspirations," *International Economic Review*, 41, 921-950.
6. Burgos, A. (2002): "Learning to deal with risk: What does reinforcement learning tell us about risk attitudes," *Economics Bulletin*, 4, 1-13.
7. Bush, R. R. and F. Mosteller (1951): "A mathematical model of simple learning," *Psychological Review*, 58, 313-323.
8. Bush, R. R. and F. Mosteller (1955): *Stochastic Models for Learning*, New York: Wiley.
9. Cross, J. G. (1973): "A stochastic learning model of economic behavior," *Quarterly Journal of Economics*, 87, 239-266.
10. Dekel, E. and S. Scotchmer (1999): "On the evolution of attitudes towards risk in winner-take-all games," *Journal of Economic Theory*, 87, 125-143.
11. Denrell, J. (2007): "Adaptive learning and risk taking," *Psychological Review*, 114, 177-187.
12. Easley, D. and A. Rustichini (1999): "Choice without beliefs," *Econometrica*, 67, 1157-1184.
13. Erev, I. and A. Roth (1998): "Predicting how people play games: Reinforcement learning in experimental games with a unique mixed strategy equilibria," *American Economic Review*, 88, 848-881.

14. Fudenberg, D. and D. Levine (1998): *The Theory of Learning in Games*, Cambridge: MIT Press.
15. Fudenberg, D. and J. Tirole (1991): *Game Theory*, Cambridge: MIT Press.
16. Gilboa, I., and D. Schmeidler (1995): "Case-based decision theory", *Quarterly Journal of Economics*, 110 , 605-639.
17. Hopkins, E. (2007): "Adaptive learning models of consumer behavior," *Journal of Economic Behavior and Organization*, 64, 348-368.
18. Hopkins, E. and M. Posch (2005): "Attainability of boundary points under reinforcement learning," *Games and Economic Behavior*," 53, 110-125.
19. Izquierdo, L., Izquierdo, S., Gotts, N. and Polhill, J. (2007): "Transient and asymptotic dynamics of reinforcement learning in games," *Games and Economic Behavior*, 61, 259-276.
20. Jehiel, P. and D. Samet (2005): "Learning to play games in extensive form by valuation," *Journal of Economic Theory*, 124, 129-148.
21. Lakshminarayanan, S. and M. A. L. Thathachar (1976): "Bounds on the convergence probabilities of learning automata," *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6, 756-763.
22. Laslier, J-F., R. Topol, and B. Walliser (2001): "A behavioral learning process in games," *Games and Economic Behavior*, 37, 340-366.
23. Laslier, J-F. and B. Walliser (2005): "A reinforcement learning process in extensive form games," *International Journal of Game Theory*, 33, 219-227.
24. March, J. G. (1996): "Learning to be risk averse," *Psychological Review*, 103, 309-319.
25. Mas-Colell, A., M. D. Whinston and J. R. Green (1995): *Microeconomic Theory*. New York: Oxford University Press.
26. Narendra, K. S. and M. A. L. Thathachar (1989): *Learning Automata: An Introduction*, Prentice-Hall.

27. Norman, M. (1968): "Slow learning," *British Journal of Mathematical and Statistical Psychology*, 2, pp. 141–159.
28. Norman, M. (1974): "A central limit theorem for Markov processes that move by small steps," *The Annals of Probability*, 2, pp. 1065–1074.
29. Oyarzun, C. and J. Ruf (2009): "Monotone imitation," *Economic Theory*, 41, 411-441.
30. Resnick, S. (1999): *A Probability Path*. Birkhäuser.
31. Robson, A. J. (1996a): "The evolution of attitudes towards risk: Lottery tickets and relative wealth," *Games and Economic Behavior*, 14, 190-207.
32. Robson, A. J. (1996b): "A biological basis for expected and non-expected utility preferences," *Journal of Economic Theory*, 68, 397-424.
33. Roth, A. E. and I. Erev (1995): "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, 8, 164-212.
34. Rustichini, A. (1999): "Optimal properties of stimulus-response learning models," *Games and Economic Behavior*, 29, 244-273.
35. Simon, H. A. (1956): "A comparison of game theory and learning theory," *Psychometrika*, 21, 267-272.
36. Thrall, R. M., C. H. Coombs and R. L. Davis (1954): *Decision Processes*, Wiley.