Learning and Risk Aversion*

Carlos Oyarzun
University of Queensland

Rajiv Sarin
University of Birmingham

May 21, 2012

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 processes 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. Lakshmivarahan, 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, Max Planck Institute (Berlin and Jena), Monash, Rice, UT Dallas, U of Queensland, Texas A&M, Victoria (Wellington), and Warwick 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 payoff 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. In particular, these learning models do not require that the agent has preferences over payoff distributions or a utility function describing such preferences. The learning models we study include many that have been used to describe human behavior in diverse experimental settings in both psychology and economics and which have been studied in the machine learning literature to find optimal actions in complex environments.

We consider an agent who knows the finite set of available 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). The state is transformed to stochastic behavior according to a *choice rule*. 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. Her state of learning is then updated according to her *state transition rule*. 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

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

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.

The state transition rule and the choice rule are taken as the primitive of the model and they define what we call the *learning process*. The learning process determines a (behavioral) *learning rule*, which maps the state of the agent today, the action she chooses, and the outcome she obtains into her behavior tomorrow. Hence, a learning rule describes the short run change in behavior and determines how behavior evolves over time. The learning process is, thus, assumed to be a primitive characteristic of the 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 models 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. Hence, we focus on the properties of the learning rule that the learning process defines. We study learning rules that, in every environment, are expected to increase the probability of choosing those actions whose payoff distribution second order stochastically dominate the distributions of all other actions. Therefore, decision makers whose learning rules satisfy this property are expected to increase the probability of choosing those actions 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 payoff 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 a 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]

Short run properties of learning 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. 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 the learning processes whose learning rules 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 learning process with a 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 at all states. We call this property the *non shrinking condition* (NSC). It requires that the expected change in the probability of choosing any safest action (in proportion to its current probability) is bounded from below away from zero while both safest and riskier actions are chosen with positive probability. And second, we require learning to be *slow* in response to experience, i.e., that the probability of choosing each action does not change much in each period.[3] We provide a number of

[2] EUT 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).

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

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 learning processes with monotonically risk seeking and monotonically risk neutral learning rules.[4]

Our last result (Proposition 5) provides conditions under which the choices of an expected utility maximizer with a given Bernoulli utility function would be obtained in the long run by a learning process. 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 processes would exhibit in the long run. We illustrate this by showing that some well-known learning models converge to choose according to the preferences represented by some widely used utility functions, such as the constant absolute risk aversion utility functions.

Before turning our attention to the related literature, we shall emphasize that, since we do not use any concept of "utility" in our definition of the learning model, our only substantive assumption is that obtained monetary payoffs and the learning process are enough to specify how learning shapes behavior. Any model of human behavior consistent with this general structure can be analyzed using our results. Therefore, the features of the functional form of monotonically risk averse learning rules may have different meanings and interpretations in different learning models. In all of them, however, satisfying these properties has the implications studied here. In particular, our results tell us what these models predict in terms of risk behavior. Thus, to the extent that these models differ in terms of such predictions, our results and the appropriate data can be useful for telling these models apart.

The first paper we know of that investigates the relation between learning

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

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 models with monotonically risk averse learning rules, 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 processes converge to choose the safest action in *any* decision problem with an arbitrary finite number of actions.

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 strict subsets 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

[5] Earlier 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).

[6] For 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.

particular, the result of Lakshmivarahan and Thathachar (1976) is related to our first long run result. They show that a specific class of learning processes, which move slowly, converge with a high probability to expected payoff maximization. The learning processes 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 μ_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. The set of states, denoted by 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 sets of states may be considered, though most of the learning models that we are familiar with are easily accommodated with this set of states. The state of the agent determines her *behavior*, which is described by the probability with which she chooses each action. We, thus, define a function, $\alpha : S \to \Delta(A)$, which maps the state of the agent to her behavior. The component of $\alpha(s)$ corresponding to action $a \in A$, denoted by $\alpha_a(s)$, is the probability of choosing this action at the state $s \in S$. We call α the *choice rule* and it may be thought of as a behavior rule.[7]

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. Formally, we define a second function, $\pi : S \times A \times X \to S$, which maps the state of the agent, the action she chooses and the payoff she obtains into her state in the next period. Therefore, π defines the *state transition rule* and may be interpreted as the primitive learning model. Since the transition of the states is determined by the chosen action and obtained payoff, states evolve randomly. The initial state, however, is assumed to be exogenously given.

We define a *learning process* as a pair of a choice rule and a state transition rule (α, π). The learning process (α, π) and the environment F define the transition probability for the states. Formally, we define $T_{(\alpha,\pi),F} : S \times \mathcal{B}(S) \to [0, 1]$, where $\mathcal{B}(S)$ is the set of Borel subsets of S, such that

$$T_{(\alpha,\pi),F}(s, S') = \sum_{a \in A} \alpha_a(s) \int 1_{\{x' \in X : \pi(s,a,x') \in S'\}}(x) dF_a(x)$$

for all $s \in S$ and $S' \in \mathcal{B}(S)$, where $1_{\{.\}}$ is the indicator function of the set $\{\cdot\}$. Thus, for any learning process (α, π) and environment F, $T_{(\alpha,\pi),F}(s, S')$ gives the probability of reaching a state in the set S' in the next period, given that the current state is s, for all $s \in S$ and $S' \in \mathcal{B}(S)$.

For each learning process (α, π), a *learning rule* L is defined as a composition of π and α, i.e., $L := \alpha \circ \pi$. 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 in the next period. That is, $L : S \times A \times X \to \Delta(A)$. In this paper we focus on the agent's behavior and, consequently, the learning rule plays a critical role in our analysis. Formally, we have the following definition.

Definition 1 *A* learning process (α, π) *is a pair of functions* $\alpha : S \to \Delta(A)$ *and* $\pi : S \times A \times X \to S$. *The* learning rule *associated to the learning process* (α, π) *is a function* $L : S \times A \times X \to \Delta(A)$ *defined as* $L = \alpha \circ \pi$.

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

For any state $s \in S$, 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 $L_a(s, a', x)$ denote the probability with which a is chosen in the next period if the state of the agent is s, action a' is chosen and a payoff of x is received today. Thus, by specifying the learning process (α, π), and hence, the learning rule L, we define a (finite, square) matrix of functions

$$\begin{pmatrix} L_a\left(s, a, \cdot\right) & L_{a'}\left(s, a, \cdot\right) & L_{a''}\left(s, a, \cdot\right) & \cdots \\ L_a\left(s, a', \cdot\right) & L_{a'}\left(s, a', \cdot\right) & \cdots & \cdots \\ L_a\left(s, a'', \cdot\right) & \vdots & \ddots & \\ \vdots & \vdots & & \ddots \end{pmatrix} \tag{1}$$

for all state $s \in S$.

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 $L_a\left(s, a, \cdot\right)$ is concave (increasing) for all $a \in A$ with $\alpha_a(s) > 0$. A learning rule L is* cross-convex *(*cross-decreasing*)* at *s if $L_a\left(s, a', \cdot\right)$ is convex (decreasing) for all $a' \in A$ and $a \in A \backslash \left\{a'\right\}$ with $\alpha_{a'}(s) > 0$.*

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.*

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. The expected change in behavior plays 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 environment F and learning process (α, π), with corresponding learning rule L, the expected movement of probability on action a in state s is denoted by

$$f_a(s) = \sum_{a' \in A} \alpha_{a'}(s) \int L_a(s, a', x)\, dF_{a'}(x) - \alpha_a(s).$$

Therefore, the study of the short run properties only requires knowing the learning rule of a learning process. It will also be useful to extend this definition to any subset of A, hence, for any learning rule L and environment F, we also define $f_{\hat{A}}(s) = \Sigma_{a \in \hat{A}} f_a(s)$ for all $\hat{A} \subset A$ and $s \in S$.

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 $a, a' \in A$.

Definition 3 *A learning rule L is* monotonically risk averse at $s \in S$ *if $f_{A^*}(s) \geq 0$ in all environments.*

A learning rule is monotonically risk averse if the expected change in the probability of choosing the set of the safest actions is non-negative in every environment. Correspondingly, we say that a learning rule is *monotonically risk seeking* if the expected change in the probability of choosing the set of the riskiest actions (those that are second order stochastically dominated by all other actions) is non-negative in every environment. Finally, we say that a learning rule is *monotonically risk neutral* if it is monotonically risk averse and monotonically risk seeking. In EUT, if A^* is empty, just knowing that an individual is risk averse (without knowing her utility function) does not allow us to pin down the action she will choose. 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 at $s \in S$ *if $f_a(s) = 0$ for all a whenever $F_a = F_{a'}$ for all $a, a' \in A$.*

Proposition 1 *A learning rule L is monotonically risk averse at $s \in S$ if and only if (i) L is impartial at s and (ii) L is cross-convex at s.*

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 at $s \in S$ then it is impartial at $s \in S$.*

Proof. The proof is by contradiction. Suppose L is monotonically risk averse at $s \in S$ but there exists an environment F with $A = A^*$ and $f_a(s) > 0$ 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 $(\widehat{\mu}_a - x_{min})\varepsilon/(x_{max} - x_{min})$ on the probability of x_{max} and $(x_{max} - \widehat{\mu}_a)\varepsilon/(x_{max} - x_{min})$ on the probability of x_{min}. By construction, $\widetilde{F}_{a'}$ SOSD $\widetilde{F}_a$ for all $a' \neq a$. It follows that $\widetilde{A}^* = A \backslash \{a\}$. Since $\widetilde{f}_a(s)$ is a continuous function of ε, there exists a small enough ε such that $\widetilde{f}_a(s) > 0$. This contradicts that L is monotonically risk averse at s. ■

Lemma 2 *A learning rule L is impartial at $s \in S$ if and only if $\alpha_a(s) = \sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x)$ 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 $a \in A$ and $s \in S$,

$$f_a(s) = \sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) - \alpha_a(s).$$

Therefore, in order to be impartial, L must satisfy

$$\alpha_a(s) = \sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x)$$

for all $a \in A$ and $x \in X$.

Sufficiency.

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

$$\begin{aligned} f_a(s) &= \sum_{a' \in A} \alpha_{a'}(s) \int L_a(s, a', x) dF_{a'}(x) - \alpha_a(s) \\ &= \int \sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) dF_a(x) - \alpha_a(s) \\ &= 0. \end{aligned}$$

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 L is impartial at s and $\alpha_a(s) = 0$, then $L_a(s, a', x) = 0$ for all $a' \neq a$, for all $a \in A$ and $x \in X$. We list this observation for future reference as Remark 1.

Remark 1 *Suppose L is monotonically risk averse at $s \in S$ and for some $a \in A$, $\alpha_a(s) = 0$. Then, for all $a' \in A \diagdown \{a\}$, $\alpha_{a'}(s) > 0$ implies $L_a(s, a', x) = 0$ for all $x \in X$.*

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

Sufficiency.

For all $a \in A^*$,

$$\begin{aligned}
f_a(s) &= \alpha_a(s)\int L_a(s,a,x)dF_a(x) + \sum_{a'\neq a}\alpha_{a'}(s)\int L_a(s,a',x)dF_{a'}(x) - \alpha_a(s) \\
&= \int\left[\alpha_a(s) - \sum_{a'\neq a}\alpha_{a'}(s)L_a(s,a',x)\right]dF_a(x) + \sum_{a'\neq a}\alpha_{a'}(s)\int L_a(s,a',x)dF_{a'}(x) - \alpha_a(s) \\
&= \sum_{a'\neq a}\alpha_{a'}\left[\int L_a(s,a',x)dF_{a'}(x) - \int L_a(s,a',x)dF_a(x)\right] \\
&\geq 0.
\end{aligned}$$

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 $L_a(s,a',\cdot)$ for all $a' \neq a$ such that $\alpha_{a'}(s) > 0$.

Necessity.

We argue by contradiction. Suppose that for some $a \in A$ and some $a' \neq a$ such that $\alpha_{a'}(s) > 0$, $L_a(s,a',\cdot)$ is not convex. Therefore there exists x', x'', $\lambda \in (0,1)$ and $x := \lambda x' + (1-\lambda)x''$ such that $\lambda L_a(s,a',x') + (1-\lambda)L_a(s,a',x'') < L_a(s,a',x)$. 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

$$\begin{aligned}
f_a(s) &= \sum_{a''\neq a}\alpha_{a''}(s)\left[\int L_a(s,a'',x)dF_{a''}(x) - \int L_a(s,a'',x)dF_a(x)\right] \\
&= \alpha_{a'}(s)[\lambda L_a(s,a',x') + (1-\lambda)L_a(s,a',x'') - L_a(s,a',x)] \\
&\quad +\varepsilon\sum_{a''\neq a,a'}\alpha_{a''}(s)[\lambda L_a(s,a'',x') + (1-\lambda)L_a(s,a'',x'') - L_a(s,a'',x)].
\end{aligned}$$

Therefore, for small enough ε, $f_a(s) < 0$. ■

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 $\sum_{a\in A} L_a\left(s, a', x\right) = 1$ for all $s \in S$, $a' \in A$, and $x \in X$, 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* at $s \in S$ if no probability mass is expected to be moved among actions when all have the same expected payoff (i.e., $f_a(s) = 0$ for all a whenever $\mu_a = \mu_{a'}$ 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 at $s \in S$* if and only if *it is monotonically risk neutral at $s \in S$.*

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

$$\begin{aligned} L_a(s,a,x) &= \alpha_a(s) + (1-\alpha_a(s))(A_{aa}(s) + B_{aa}(s)x) \qquad (2) \\ L_a(s,a',x) &= \alpha_a(s) - \alpha_a(s)(A_{a'a}(s) + B_{a'a}(s)x) \qquad \forall a' \neq a, \end{aligned}$$

where the coefficients $A_{a'a}(s)$ and $B_{a'a}(s)$ satisfy $A_{aa}(s) = \Sigma_{a'\in A}\alpha_{a'}(s)A_{a'a}(s)$

and $B_{aa}(s) = \Sigma_{a' \in A} \alpha_{a'}(s) B_{a'a}(s)$.[8] To see that unbiased rules are impartial, we verify that

$$\begin{aligned}
\sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) &= \sum_{a' \neq a} \left[\alpha_{a'}(s) \left(\alpha_a(s) - \alpha_a(s)(A_{a'a}(s) + B_{a'a}(s)x) \right) \right] \\
&\quad + \alpha_a(s) \left(\alpha_a(s) + (1 - \alpha_a(s))(A_{aa}(s) + B_{aa}(s)x) \right) \\
&= \alpha_a(s)
\end{aligned}$$

for all $a \in A$ and $x \in [0, 1]$. 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 $A_{a'a}(s)$ and $B_{a'a}(s)$ such that every monotonically risk neutral rule can be written as (2). To prove that $A_{aa}(s) = \Sigma_{a' \in A} \alpha_{a'}(s) A_{a'a}(s)$ and $B_{aa}(s) = \Sigma_{a' \in A} \alpha_{a'}(s) B_{a'a}(s)$, note that monotonically risk neutral rules are impartial and, hence, they satisfy

$$\sum_{a' \neq a} \alpha_{a'}(s) L_a(s, a', x) + \alpha_a(s) L_a(s, a, x) = \alpha_a(s)$$

for all $a \in A$. Therefore,

$$\begin{aligned}
L_a(s, a, x) &= 1 - \sum_{a' \neq a} \alpha_{a'}(s) L_a(s, a', x) / \alpha_a(s) \\
&= 1 - \sum_{a' \neq a} \alpha_{a'}(s) \left(1 - (A_{a'a}(s) + B_{a'a}(s)x) \right) \\
&= \alpha_a(s) + \sum_{a' \neq a} \alpha_{a'}(s) \left(A_{a'a}(s) + B_{a'a}(s)x \right).
\end{aligned}$$

Equating the RHS of the last equality with the RHS of the first equality in (2), we obtain that $\alpha_a(s) + (1 - \alpha_a(s))(A_{aa}(s) + B_{aa}(s)x) = \alpha_a(s) + \Sigma_{a' \neq a} \alpha_{a'}(s)(A_{a'a}(s) + B_{a'a}(s)x)$. Since this equality has to hold for all $x \in$

[8] Börgers et al. (2004) define unbiasedness for states s such that $\alpha_a(s) > 0$ for all $a \in A$. However, their definition and its equivalence to monotone risk neutrality can be extended to the whole simplex. Also notice that the coefficients $A_{a'a}(s)$ and $B_{a'a}(s)$ are defined for each state $s \in S$, for all $a, a' \in A$ (i.e., they are allowed to be state dependent).

$[0, 1]$ we obtain that $A_{aa}(s) = \Sigma_{a' \in A} \alpha_{a'}(s) A_{a'a}(s)$ and $B_{aa}(s) = \Sigma_{a' \in A} \alpha_{a'}(s) B_{a'a}(s)$, as desired. ■

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 at $s \in S$ *if $f_{A^{**}}(s) \geq 0$ 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 at $s \in S$ if and only if (i) L is impartial at s and (ii) L is cross-decreasing at s.*

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 models that have appeared in the literature. The next section studies the long run properties of some of these models 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). The learning rule of this model was shown to satisfy all the properties studied in Börgers et al. (2004). Example 2 studies Roth and Erev's (1995) learning model 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 learning rule of Roth and Erev (1995) 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 models (see, e.g., Erev and Roth (1998)) that generalizes one of the learning models in March (1996) is also considered.

Example 1 *[Cross, 1973] The set of states is $S = \Delta(A)$, where each component of $s \in \Delta(A)$ corresponds to each action in A. The state transition rule, $\pi : S \times A \times X \to S$, is given by*

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

for all $s \in S$, $a \in A$, and $x \in X = [0,1]$. The choice rule $\alpha : S \to \Delta(A)$ is the identity function, i.e., $\alpha(s) = s$ for all $s \in S$. Thus, $L_a(s, a', x) = \pi_a(s, a', x)$ for all $s \in S$, $a, a' \in A$, and $x \in [0,1]$.

In the Cross rule, for any $a \in A$, $s \in S$, and $x \in X$,

$$\begin{aligned}\Sigma_{a' \in A}\alpha_{a'}(s)L_a(s, a', x) &= \Sigma_{a' \neq a}\alpha_{a'}(s)(s_a - s_a x) + \alpha_a(s)(s_a + (1 - s_a)x) \\ &= \alpha_a(s).\end{aligned}$$

Hence, by Lemma 2, the Cross learning rule is impartial. Furthermore, for all $a' \neq a$ and $s \in \Delta(A)$, $L_a(s, a', \cdot)$ is an affine function. Therefore, this learning rule is monotonically risk averse and monotonically risk seeking and, hence, it is monotonically risk neutral. As the cross terms are decreasing functions, this 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 $s \in S = R_{++}^{|A|}$. The vector $s = (s_a)_{a \in A}$ describes the decision maker's "attraction" to choose any of her actions. The choice function $\alpha : S \to \Delta(A)$ is given by $\alpha_a(s) = s_a / \Sigma_{a'} s_{a'}$ for all $a \in A$ and $s \in S$. The state transition rule π is defined as follows:*

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

for all $s \in S$, $a \in A$, and $x \in X = [0, x_{\max}]$. Combining α and π we obtain that the Roth and Erev learning rule is given by

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

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

For arbitrary $a \in A$, $s \in S$, and $x \in X$,

$$\begin{aligned}\sum_{a' \in A}\alpha_{a'}(s)L_a(s, a', x) &= \sum_{a' \neq a}\left(\frac{s_{a'}}{\Sigma_{a'' \in A} s_{a''}}\right)\frac{s_a}{\Sigma_{a'' \in A} s_{a''} + x} + \left(\frac{s_a}{\sum_{a'' \in A} s_{a''}}\right)\frac{(s_a + x)}{\sum_{a'' \in A} s_{a''} + x} \\ &= \frac{s_a}{\sum_{a' \in A} s_{a'}} = \alpha_a(s).\end{aligned}$$

Therefore, this learning rule is impartial. Furthermore, for all $a' \neq a$ and $s \in S$, $L_a(s, a', \cdot)$ is decreasing and convex, 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 $X = [x_{\min}, x_{\max}]$, $x_{\min} > 0$, and $x_{\max} < \infty$. If action a is chosen and yields a payoff x, the state transition rule adds $\beta(x - s_a)$ to s_a, where $\beta \in (0,1)$ is a parameter, leaving all other attractions unchanged. Thus,*

$$\begin{aligned} \pi_a(s, a, x) &= s_a + \beta(x - s_a) \\ \pi_a(s, a', x) &= s_a \quad \forall a' \neq a. \end{aligned}$$

for all $s \in S$, $a \in A$, and $x \in X$. The choice rule is given by $\alpha_a(s) = s_a / \Sigma_{a'} s_{a'}$ for all $a \in A$ and $s \in S$.Therefore, the learning rule of this learning process may be written as

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

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

The learning rule of this example 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 $x \in (x_{\min}, x_{\max})$ with probability one and the state s satisfies $s_a < x$ and $s_{a'} > x$ for all $a' \neq a$. With probability one, either the attraction of an action $a' \neq a$ will decrease or the attraction of a will increase. In either case the probability of choosing action a will increase and thus, $f_a(s) > 0$. It follows that this learning 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. It can be obtained as a particular case of a class of *belief-based learning* models considered, for instance, in Erev and Roth (1998), p. 866. For this class of learning models, the set of payoffs is $X := [x_{\min}, x_{\max}]$, with $x_{\min} > 0$ and $x_{\max} < \infty$, and the set of states is $S = (X \times \mathbb{N})^{|A|}$. The state, denoted by $s = (\nu, \kappa)$, is given by the vector of (strictly positive) attractions of each action, denoted by ν, and a vector containing the number of times each action has been chosen in the past, denoted by κ. It turns out to be useful for the discussion below to allow explicitly for a transformation of payoffs. Thus, let $u : X \to X$ be an increasing transformation. Roughly speaking, the attraction of each action corresponds to the average value of $u(x)$ that it has yielded in the past. Thus, the state transition rule is given by

$$\begin{aligned}
\pi_a(s, a, x) &= (\nu_a + (u(x) - \nu_a) / (\kappa_a + 1), \kappa_a + 1) \\
\pi_a(s, a', x) &= (\nu_a, \kappa_a) \qquad \forall a' \neq a.
\end{aligned}$$

for all $s \in S$, $a \in A$, and $x \in X$. The choice rule is given by $\alpha_a(s) = \nu_a^\rho / \Sigma_{a'} \nu_{a'}^\rho$ for all $a \in A$ and $s \in S$, where $\rho \geq 0$ is a parameter. Thus, the learning rule of this learning process may be written as

$$\begin{aligned}
L_a(s, a, x) &= \frac{(\nu_a + (u(x) - \nu_a)/(\kappa_a + 1))^\rho}{\sum_{a'' \neq a} \nu_{a''}^\rho + (\nu_a + (u(x) - \nu_a)/(\kappa_a + 1))^\rho} \\
L_a(s, a', x) &= \frac{\nu_a^\rho}{\sum_{a'' \neq a'} \nu_{a''}^\rho + (\nu_{a'} + (u(x) - \nu_{a'})/(\kappa_{a'} + 1))^\rho} \qquad \forall a' \neq a
\end{aligned}$$

for all $a \in A$, $s \in S$, and $x \in X$. March's (1996) 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 $\rho = 1$, the probabilities of choosing each action are proportional to the past averages of $u(x)$ of each action. In game theory, since payoffs are considered to be von Neumann-Morgestern utility (see, e.g., Fudenberg and Tirole (1991)), u has been 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 March's (1996) Weighted

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

Return over Gains model reveals that this conjecture is false for any $\rho > 0$, because the belief-based learning rule fails to be globally impartial. Hence, belied-based learning models fail to satisfy our short run properties.

Example 4 *[Fractional Adjustment over Gains, March, 1996] This learning model is defined for environments in which the decision maker has only two actions. As in the Cross learning model, the set of states is $S = \Delta(A)$, where each component of $s \in \Delta(A)$ corresponds to the probability of choosing each action in A. The choice rule $\alpha : S \rightarrow \Delta(A)$ is, therefore, the identity function, $\alpha(s) = s$ for all $s \in S$. Over gains, i.e., for $x \geq 0$, the state transition rule of this model is given by*

$$\begin{aligned} \pi_a(s, a, x) &= 1 - (1-\sigma)^x (1 - s_a) \\ \pi_a(s, a', x) &= (1-\sigma)^x s_a \qquad a' \neq a, \end{aligned}$$

for all $a \in A$, and $s \in S$, where $\sigma \in (0,1)$ is a parameter. Thus, $L_a(s, a', x) = \pi_a(s, a', x)$ for all $s \in S$, $a, a' \in A$, and $x \geq 0$.

For all $a \in A$, $s \in S$, and $x \in X$,

$$\begin{aligned} \Sigma_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) &= \alpha_{a'}(s)\left((1-\sigma)^x s_a\right) + \alpha_a(s)\left(1 - (1-\sigma)^x (1 - s_a)\right) \\ &= \alpha_a(s). \end{aligned}$$

Therefore, this learning rule is impartial. For all $a' \neq a$ and $s \in S$, $L_a(s, a', \cdot)$ is decreasing and convex. Hence, this learning rule is globally first order monotone and monotonically risk averse.

Example 5 *[Fractional Adjustment over Losses, March, 1996] The learning process is defined in the same manner as in the Fractional Adjustment over Gains model. The only difference is that the state transition rule is given by*

$$\begin{aligned} \pi_a(s, a, x) &= (1-\sigma)^{-x} s_a \\ \pi_a(s, a', x) &= 1 - (1-\sigma)^{-x} (1 - s_a) \qquad a' \neq a \end{aligned}$$

for all $a \in A$, $s \in S$, and $x < 0$. The learning rule is, thus, given by $L_a(s, a', x) = \pi_a(s, a', x)$ for all $s \in S$, $a, a' \in A$, and $x < 0$.

This learning rule is cross-concave. However, simple algebra reveals that $\Sigma_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) = \alpha_a(s)$ if and only if $x = 0$, for all $a \in A$ and $s \in S$. Therefore, this learning 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 learning processes with 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 processes that, with high probability, converge to choose the corresponding expected utility maximizing action.

5.1 Preliminaries

As we consider the learning process at different periods of time, we will often subscript the state and the probability of choosing the safest actions by the time period $t = 0, 1, ...$ Let $\alpha^* : S \to [0, 1]$, defined by $\alpha^*(s) := \sum_{a \in A^*} \alpha_a(s)$, be the probability of choosing an action in A^*, for all $s \in S$. Also, let α_t^* be the probability of choosing an action in A^* at time t, for all $t \in \mathbb{N}_0$. Hence, $\alpha_t^* = \alpha^*(s_t)$, where s_t is the state at time $t \in \mathbb{N}_0$.

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.[10] For any learning rule L, two actions a and $a' \neq a$, and payoff distributions $F'_{a'}$ and $F_{a'}$, let

$$n_{a'a}(F'_{a'}, F_{a'}) := \inf_{\{s \in S : \alpha_a(s), \alpha_{a'}(s) > 0\}} \left\{ \frac{\int L_a(s, a', x)\, dF_{a'}(x) - \int L_a(s, a', x)\, dF'_{a'}(x)}{\alpha_a(s)} \right\}.$$

[10] In particular, impartiality implies that for all states such that $\alpha_a(s)$ is arbitrarily close to zero, while $\alpha_{a'}(s)$ is bounded away from zero for all $a' \neq a$, $L_a(s, a', x)$ 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.

Definition 6 *The learning rule L associated to a learning process (α, π) 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 learning rule of the Roth-Erev (1995) model, may not satisfy the NSC. Such rules may asymptotically become insensitive to risk in the environment.

Lemma 3 *Suppose that the learning rule L associated to the learning process (α, π) is globally monotonically risk averse and satisfies the NSC. Then, for arbitrary initial state $s_0 \in S$, α^*_t converges almost surely to a random variable α^*_∞ with support in $\{0, 1\}$.*

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

From the proof of Proposition 1, we know that for any monotonically risk averse learning rule,

$$\begin{aligned} f_{A^*}(s_t) &= \sum_{a \in A^*} \sum_{a' \notin A^*} \alpha_{a'}(s_t) \left(\int L_a(s_t, a', x)\, dF_{a'}(x) - \int L_a(s_t, a', x)\, dF_a(x) \right) \\ &= \sum_{\{a \in A^* | \alpha_a(s_t) > 0\}} \sum_{\{a' \notin A^* | \alpha_{a'}(s_t) > 0\}} \alpha_a(s_t) \alpha_{a'}(s_t) \frac{\int L_a(s_t, a', x)\, dF_{a'} - \int L_a(s_t, a', x)\, dF_a}{\alpha_a(s_t)} \end{aligned}$$

for all $t \in \mathbb{N}_0$, where the second equality follows from Remark 1 and deleting all the $\alpha_{a'}(s_t)$ terms that are equal to zero.

Let

$$\delta := \min_{\{a \in A^*, a' \notin A^*\}} \left\{ \inf_{\{s \in S : \alpha_a(s), \alpha_{a'}(s) > 0\}} \left\{ \frac{\int L_a(s, a', x)\, dF_{a'} - \int L_a(s, a', x)\, dF_a}{\alpha_a(s)} \right\} \right\}.$$

The NSC guarantees that $\delta > 0$. Consequently, for all $t \in \mathbb{N}_0$,

$$f_{A^*}(s_t) \geq \delta \sum_{\{a \in A^* | \alpha_a(s_t)>0\}} \sum_{\{a' \notin A^* | \alpha_{a'}(s_t)>0\}} \alpha_a(s_t)\alpha_{a'}(s_t) = \delta \sum_{a \in A^*} \sum_{a' \notin A^*} \alpha_a(s_t)\alpha_{a'}(s_t) \geq 0. \tag{3}$$

For all $t \in \mathbb{N}$,

$$\begin{aligned} E[\alpha_t^*] &= \alpha_0^* + E[\alpha_1^* - \alpha_0^*] + E[\alpha_2^* - \alpha_1^*] + ... + E[\alpha_t^* - \alpha_{t-1}^*] \\ &= \alpha_0^* + \sum_{\tau=0}^{t-1} E[f_{A^*}(s_\tau)]. \end{aligned}$$

Since $f_{A^*}(s_t) \geq 0$, we have $E[f_{A^*}(s_t)] \geq 0$ for all $t \in \mathbb{N}_0$. It follows that $\lim_{t\to\infty} E[f_{A^*}(s_t)] = 0$, because $E[\alpha_t^*] \leq 1$ for all $t \in \mathbb{N}_0$.

From (3), $E[f_{A^*}(s_t)] \geq E\left[\delta \sum_{a\in A^*} \sum_{a' \notin A^*} \alpha_a(s_t)\alpha_{a'}(s_t)\right] \geq 0$ for all $t \in \mathbb{N}_0$. Since $\lim_{t\to\infty} E[f_{A^*}(s_t)] = 0$, $\lim_{t\to\infty} E\left[\delta \sum_{a\in A^*} \sum_{a' \notin A^*} \alpha_a(s_t)\alpha_{a'}(s_t)\right] = 0$. Since $\delta > 0$, $\lim_{t\to\infty} E\left[\sum_{a\in A^*} \sum_{a' \notin A^*} \alpha_a(s_t)\alpha_{a'}(s_t)\right] = 0$. Equivalently, $\lim_{t\to\infty} E[\alpha_t^*(1-\alpha_t^*)] = 0$. This may be written as $\lim_{t\to\infty} E[\alpha_t^*] = \lim_{t\to\infty} E\left[(\alpha_t^*)^2\right]$.

By the Continuous Mapping Theorem (e.g., Resnick 1999, p. 261), $\lim_{t\to\infty} E[\alpha_t^*] = E[\alpha_\infty^*]$ and $\lim_{t\to\infty} E\left[(\alpha_t^*)^2\right] = E[(\alpha_\infty^*)^2]$.[11] Therefore, $E[\alpha_\infty^*] = E[(\alpha_\infty^*)^2]$. Since $z > z^2$ for all $z \in (0,1)$, $\Pr\{\alpha_\infty^* \in (0,1)\} = 0$. ■

Let $\Gamma(s_0) := \Pr\{\alpha_\infty^* = 1\}$ for arbitrary initial state $s_0 \in S$. Existence and uniqueness of $\Gamma(s_0)$ follow 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).

Lemma 3 guarantees that learning processes with 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 α_t^* converges to one. To obtain more precise bounds on $\Gamma(s_0)$ we consider slow versions of the original learning process.

[11] The 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 $g : \mathbb{R} \to \mathbb{R}$ is discontinuous is 0, then $\lim_{t\to\infty} Eg(X_t) = Eg(X)$. Since α_t^* converges to α_∞^* with probability 1, α_t^* converges to α_∞^* in distribution, hence the Continuous Mapping Theorem applies.

Definition 7 *Let (α, π) be a learning process with learning rule $L = \alpha \circ \pi$ and state space S, and let $\theta \in (0, 1]$. A learning process (α, π^θ) with state space S, is a* slow version *of (α, π) if $L^\theta := \alpha \circ \pi^\theta$ satisfies*

$$L_a^\theta (s, a', x) - \alpha_a(s) = \theta (L_a (s, a', x) - \alpha_a(s))$$

for all $s \in S$, $a', a \in A$ and $x \in X$. Furthermore, in this case we also say that L^θ is a slow version of L.

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

$$L_a^\theta (s, a', x) = (1 - \theta)\alpha_a(s) + \theta L_a (s, a', x) . \tag{4}$$

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 initial state $s_0 \in S$, $\alpha_t^{\theta,*}$ converges almost surely to a random variable, denoted by $\alpha_\infty^{\theta,*}$, whose support is contained in $\{0, 1\}$. Let $\Gamma^\theta(s_0) := \Pr \left\{ \alpha_\infty^{\theta,*} = 1 \right\}$ for a slow learning process (α, π^θ) and initial state $s_0 \in S$.

5.2 Convergence to second order stochastically dominant actions

The main result in this subsection is Proposition 3. It shows that slow versions of learning processes with 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 $\Gamma(s_0)$. Specifically, it shows that we can construct a function $\phi : [0, 1] \to [0, 1]$ such that $\Gamma(s_0) \geq \phi (\alpha^*(s_0))$ for all initial state $s_0 \in S$. The function ϕ satisfies $\phi(0) = 0$ and $\phi(1) = 1$, and is strictly increasing and strictly concave. This function satisfies $\phi(z) \geq z$ for all $z \in [0, 1]$. The proofs of this lemma and the next are provided in the Appendix.

Lemma 4 *Suppose that the learning rule L associated to the learning process* (α, π) *is globally monotonically risk averse and satisfies the NSC. Then, there exists* $\gamma^* > 0$ *such that the function* $\phi : [0,1] \to [0,1]$, *defined as*

$$\phi(z) := \frac{1 - e^{-\gamma z}}{1 - e^{-\gamma}}$$

for all $z \in [0,1]$, *satisfies the following conditions when* $\gamma = \gamma^*$ *:*
(I) $E\left[\phi(\alpha^*_{t+1})\right] - E\left[\phi(\alpha^*_t)\right] \geq 0$ *for all* $t \in \mathbb{N}_0$, *and*
(II) $\Gamma(s_0) \geq \phi(\alpha^*(s_0)) \ \forall \ s_0 \in S$.

The next lemma shows that γ can be taken to be arbitrarily large for suitably slow learning processes while preserving the properties of ϕ obtained in Lemma 4.

Lemma 5 *Suppose that the learning rule L associated to the learning process* (α, π) *is globally monotonically risk averse and satisfies the NSC. Then, for all* $\widehat{\gamma} > \gamma^*$, *with* γ^* *defined as in Lemma 4, there exists a slow version of* (α, π), *denoted by* $\left(\alpha, \pi^{\hat{\theta}}\right)$, *such that* $\Gamma^{\widehat{\theta}}(s_0) \geq \phi(\alpha^*(s_0)) \ \forall \ s_0 \in S$, *when* $\gamma = \widehat{\gamma}$.

As Lemma 4 reveals, for every learning process with a globally monotonically risk averse learning rule, we can provide a lower bound for the probability of convergence to choose the safest actions with probability one.[12] This lower bound corresponds to $\phi(\alpha^*_0)$ with $\gamma = \gamma^*$, for some $\gamma^* > 0$. This function is a lower bound because $E\left[\phi(\alpha^*_t)\right]$ is increasing in time and $E\left[\phi(\alpha^*_\infty)\right] = E\left[\alpha^*_\infty\right]$ (because $\phi(z) = z$ for $z \in \{0,1\}$, which is the support of α^*_∞, as Lemma 3 reveals). Then, Lemma 5 shows that, if we consider the slow learning process (α, π^θ), $\phi(\alpha^*_0)$ is a lower bound for the probability of convergence if we set $\gamma = \widehat{\gamma} := \gamma^*/\theta$ instead of $\gamma = \gamma^*$. It follows that by considering slow learning, i.e., θ small enough, $\hat{\gamma}$ can be chosen arbitrarily large. Finally, since $\phi(\alpha^*_0) \to 1$ when $\gamma \to \infty$ for all $\alpha^*_0 \in (0,1]$, we have that $\Gamma^\theta(s_0) \to 1$ as $\theta \to 0$, for all initial state s_0 such that the initial probability of choosing a second order stochastically dominant action is strictly positive.

[12] The 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, $\phi(\alpha_0^*)$, as close to 1 as desired.[13]

Proposition 3 *Suppose that the learning rule associated to the learning process* (α, π) *is globally monotonically risk averse and satisfies the NSC. Then, for any* $\varepsilon > 0$ *and initial state* $s_0 \in S$ *such that* $\alpha^*(s_0) > 0$*, there exists* $\theta_\varepsilon \in (0, 1]$ *such that* $\Gamma^\theta(s_0) > 1 - \varepsilon$ *for any slow version of* (α, π) *with* $\theta \in (0, \theta_\varepsilon)$*.*

Proof. Consider any $\varepsilon > 0$ and s_0 such that $\alpha^*(s_0) > 0$. Lemma 4 and Lemma 5 imply that there exists $\gamma^* > 0$ such that, for all $\widehat{\gamma} > \gamma^*$, there exists $\widehat{\theta} := \gamma^*/\widehat{\gamma}$ which satisfies $\Gamma^{\widehat{\theta}}(s_0) \geq \phi(\alpha^*(s_0))$, when $\gamma = \widehat{\gamma}$. As $\gamma \to \infty$, $\phi(z) \to 1$ for all $z \in (0, 1]$. Thus, we can choose $\gamma_\varepsilon > \gamma^*$ such that, if $\gamma \geq \gamma_\varepsilon$, $\phi(\alpha^*(s_0)) > 1 - \varepsilon$. From the argument above, there exists $\widehat{\theta}_\varepsilon \in (0, 1)$, with $\widehat{\theta}_\varepsilon := \gamma^*/\gamma_\varepsilon$, such that $\Gamma^{\widehat{\theta}_\varepsilon}(s_0) > 1 - \varepsilon$. Furthermore, for each $\theta \in (0, \theta_\varepsilon)$ we can find a $\gamma > \gamma_\varepsilon$ such that $\theta = \gamma^*/\gamma$ and $\Gamma^\theta(s_0) \geq \phi(\alpha^*(s_0)) > 1 - \varepsilon$. ■

Lemmas 3-5 and Proposition 3 reveal the different natures of the NSC and slow learning, and what each of them accomplishes. The NSC plays a critical role in the proof of Lemma 3, which shows that α_∞^* cannot take values in $(0, 1)$ with strictly positive probability. Since the expected changes in α_t^* are positive and α_t^* is bounded above by 1, these expected changes converge to zero as time goes to infinity. The proof of Lemma 3 reveals that, given the formulations of the expected change in α_t^* for monotonically risk averse learning rules (i.e., $f_{A^*}(s_t)$), the NSC provides a sufficient condition for these expected changes to go to zero only when α_t^* converges to 0 or 1. This yields the result of the lemma. In contrast, 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.[14] From (3), it follows that if the sequence $\{\alpha_t^*\}_{t=1}^\infty$ was deterministic and followed the path of its expected value, then the probability of choosing the safest action would converge to one. This allows the lower bound for the probability of convergence to the

[13] Notice that our result provides sufficient conditions for convergence with high probability for slow learning processes. 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.

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

safest action to go to one for slow learning processes with 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 monotonically risk averse and satisfies the NSC. To see this, notice that for $a \neq a'$, $L_a(s, a', x) = (1-\sigma)^x s_a$. This is twice differentiable in x with strictly positive second derivative. Thus, if $F'_{a'}$ strictly second-order stochastically dominates $F_{a'}$, then

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

Slow versions are obtained by specifying the corresponding slow state transition rule. In particular, for any $s \in \Delta(A)$, $a \in A$, and $x \geq 0$, let

$$\begin{aligned} \pi_a^\theta(s, a, x) &= s_a + \theta\left(1 - (1-\sigma)^x\right)(1 - s_a) \\ \pi_a^\theta(s, a'x) &= s_a + \theta\left((1-\sigma)^x - 1\right)s_a \qquad a' \neq a. \end{aligned}$$

The corresponding slow learning rule is given by $L_a^\theta(s, a', x) = \pi_a^\theta(s, a', x)$ for all $s \in S$, $a, a' \in A$, and $x \geq 0$. Proposition 3 implies that sufficiently slow versions of this learning process 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 $s_0 \in S$ such that $\alpha^*(s_0) > 0$, there exists $\theta_\varepsilon \in (0, 1]$ such that $\Pr\left\{\alpha_\infty^{*,\theta} = 1\right\} > 1 - \varepsilon$ for any slow version of this learning process, (α, π^θ), such that $\theta \in (0, \theta_\varepsilon)$.

An analogous version of Proposition 3 holds for learning processes with monotonically risk seeking learning 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 learning process with a monotonically risk seeking learning rule would converge with high probability to this set.

Another analogy can be obtained for first order monotone learning rules. Formally, let $\alpha^{**}(s) := \Sigma_{a \in A^{**}} \alpha_a(s)$ for all $s \in S$ and $\alpha_t^{**} := \alpha^{**}(s_t)$ for all $t \in \mathbb{N}_0$, 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 that the learning rule associated to the learning process (α, π) is globally first-order monotone and $n_{a'a}(F'_{a'}, F_{a'}) > 0$ 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 initial state $s_0 \in S$, α_t^{**} converges almost surely to a random variable α_∞^{**}, with support in $\{0, 1\}$, and (ii) for any $\varepsilon > 0$ and $s_0 \in S$ such that $\alpha^{**}(s_0) > 0$, there exists $\theta_\varepsilon \in (0, 1]$ such that $\Pr\{\alpha_\infty^{\theta,**} = 1\} > 1 - \varepsilon$ for any slow version of (α, π) 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^* be the set of most preferred actions for a decision maker with Bernoulli utility function u, i.e.,

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

Let $\alpha^*(u, s) := \Sigma_{a \in A_u^*} \alpha_a(s)$ for any Bernoulli utility function $u : [x_{\min}, x_{\max}] \to \mathbb{R}$ and $s \in S$; and also let $\alpha_t^*(u) := \alpha^*(u, s_t)$. The following Lemma provides conditions on a learning process such that the limit of $\alpha_t^*(u)$ is 1 with high probability. We say that $c + du$ is a *negative affine transformation* of u if $d < 0$ and $c \in \mathbb{R}$.

Lemma 6 *Suppose that the learning rule L associated to the learning process (α, π) is globally impartial. Furthermore, suppose that (i) $L_a(s, a', \cdot)$ is a negative affine transformation of u for all $s \in S$, $a \in A$ and $a' \in A \backslash \{a\}$, 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* *initial state $s_0 \in S$ with $\alpha^*(u, s_0) > 0$, there exists a $\theta_\varepsilon \in (0,1]$ such that for any slow version $\left(\alpha, \pi^\theta\right)$ of the learning process, with $\theta \in (0, \theta_\varepsilon)$, there exists a random variable, denoted by $\alpha_\infty^{*,\theta}(u)$, such that $\alpha_t^{*,\theta}(u)$ converges almost surely to $\alpha_\infty^{*,\theta}(u)$, and $\Pr\{\alpha_\infty^{*,\theta}(u) = 1\} > 1 - \varepsilon$.*

Proof. Impartiality implies

$$f_a(s) = \sum_{a' \neq a} \alpha_{a'}(s) \left[\int L_a(s, a', x) dF_{a'}(x) - \int L_a(s, a', x) dF_a(x) \right]$$

for all $a \in A$ and $s \in S$. For any action $a \in A_u^*$, we have $f_a(s) \geq 0$ because $L_a(s, a', \cdot)$ is a negative affine transformation of u for all $a \in A$, $a' \neq a$, and $s \in S$. It follows that $f_{A_u^*}(s_t) \geq 0$ for all $t \in \mathbb{N}_0$. The rest of the argument is analogous to the arguments in the proofs of Lemmas 3-5 and Proposition 3, and is omitted.[15] ■

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

Proposition 5 *Consider any environment F and continuous Bernoulli utility function $u : [x_{\min}, x_{\max}] \to \mathbb{R}$. There exists a learning process (α, π), with associated learning rule L, such that $f_a(s) \geq 0$ for all $a \in A_u^*$ and $s \in S$; and for all $\varepsilon > 0$, if s_0 satisfies $\alpha^*(u, s_0) > 0$, then there exists $\theta_\varepsilon \in (0, 1]$ such that for any slow version of (α, π), denoted by (α, π^θ), with $\theta \in (0, \theta_\varepsilon)$, we have that $\alpha_t^{*,\theta}(u)$ converges almost surely to a random variable, denoted by $\alpha_\infty^{*,\theta}(u)$, such that $\Pr\{\alpha_\infty^{*,\theta}(u) = 1\} > 1 - \varepsilon$.*

Proof. Consider a continuous function $u : [x_{\min}, x_{\max}] \to \mathbb{R}$. By Lemma 6, we just need to provide a learning process with a globally impartial learning rule L such that conditions (i) and (ii) of the lemma are satisfied. In order to construct such a learning process, for each $a \in A$ and $a' \in A \setminus \{a\}$, consider an affine transformation of u, denoted by $u_{a'a} : [x_{\min}, x_{\max}] \to [0, 1]$ such that $u_{aa'} = u_{a'a}$. Let the set of states be given by $S = \Delta(A)$, with state

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

transition rule $\pi_a(s, a', x) = s_a(1 - u_{a'a}(x))$ for all $s \in S,\ a \in A,\ a' \in A \setminus \{a\}$ and $x \in [x_{\min}, x_{\max}]$. Furthermore, define the choice rule as $\alpha(s) = s$ for all $s \in S$. These conditions imply that the associated learning rule L satisfies (i) and (ii) in Lemma 6. Finally, impartiality follows from

$$\begin{aligned}\sum_{a' \in A} \alpha_{a'}(s) L_a(s, a', x) &= \sum_{a' \neq a} \alpha_{a'}(s) s_a(1 - u_{a'a}(x)) + \alpha_a(s) \left(1 - \sum_{a' \neq a} L_{a'}(s, a, x)\right) \\ &= \sum_{a' \neq a} \alpha_{a'}(s) s_a(1 - u_{a'a}(x)) + \alpha_a(s) - \sum_{a' \neq a} \alpha_a(s) s_{a'}(s)(1 - u_{a'a}(x)) \\ &= \alpha_a(s)\end{aligned}$$

for all $s \in S$, $a \in A$, and $x \in X$. ■

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

$$\begin{aligned}L_a(s, a', x) &= (1 - \sigma)^x s_a \\ &= s_a e^{-\left(\ln \frac{1}{1-\sigma}\right)x}\end{aligned}$$

for all $s \in S$ and $x \in X$. 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 $b = \ln \frac{1}{1-\sigma}$. For any pair of distributions $F_{a'}$ and $F'_{a'}$ such that $\int u_{CARA}(x) dF'_{a'}(x) > \int u_{CARA}(x) dF_{a'}(x)$,

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

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, $b = \ln \frac{1}{1-\sigma}$, can be made as close to 1 as desired for a suitably slow version of March's fractional adjustment learning process.

Next, we consider a simple modification of the Roth and Erev learning model (Section 4, Example 2) in which we keep the sum of attractions constant. Additionally, suppose all payoffs are strictly positive. For this modified learning process, the set of states is $S = \left\{(s_a)_a \in \mathbb{R}^{|A|}_{++} : \sum_a s_a = \overline{S}\right\}$, where $\overline{S} > 0$ is an exogenously given constant. If the state is s, action a is chosen and it gets a payoff of x, then its attraction becomes $\pi_a(s,a,x) = (s_a + x)\left(\Sigma s_{a'}/\left(\Sigma s_{a'} + x\right)\right) = (s_a + x)\left(\overline{S}/\left(\overline{S} + x\right)\right)$. If instead $a' \neq a$ is chosen, then the attraction of action a becomes $\pi_a(s,a',x) = s_a\left(\overline{S}/\left(\overline{S} + x\right)\right)$. The choice rule α is as in the original model, i.e., $\alpha_a(s) = s_a/\sum_{a' \in A} s_{a'}$ for all $a \in A$ and $s \in S$. The corresponding learning rule is, thus, given by

$$\begin{aligned} L_a(s,a,x) &= \frac{s_a + x}{\overline{S} + x} \\ L_a(s,a',x) &= \frac{s_a}{\overline{S} + x} \qquad \forall\ a' \neq a, \end{aligned}$$

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

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

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

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 $u(x) = \frac{-1}{\overline{S}+x}$, can be made as close to one as desired by choosing a small enough θ to construct an slow version of this learning process.

As $\overline{S} \to 0$, since we have assumed that $0 < x_{\min} < x_{\max} < \infty$, $u(x) = \frac{-1}{\overline{S}+x}$ converges uniformly to the Constant Relative Risk Aversion (CRRA) utility function $\widetilde{u}(x) = \frac{x^{1-\sigma}}{1-\sigma}$ with coefficient of relative risk aversion $\sigma = 2$. Hence, for every environment F, there exists a small enough $\overline{S}$, call it $\overline{S}_F$, such that at least one of the maximizers of $\widetilde{u}$ is a maximizer of u. Therefore, for a slow enough version of the learning process constructed with $\overline{S}_F$, the probability of converging to one of the maximizers of $\widetilde{u}$ in F can be made arbitrarily close to one.[16]

[16] As $\overline{S} \to 0$, slow versions of this learning rule converge to the Bush-Mosteller (1955) learning rule, i.e., $\lim_{\overline{S} \to 0} L_a^\theta(s,a,x) = (1-\theta)s_a + \theta$ and $\lim_{\overline{S} \to 0} L_a^\theta(s,a',x) \to (1-\theta)s_a$ for $a' \neq a$.

Finally, consider a "generalized" version of Cross' learning model (Section 4, Example 1). This learning process 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 learning process constructed in the proof of Proposition 5. The associated learning rule specifies

$$\begin{aligned} L_a(s, a, x) &= s_a + (1 - s_a) u(x) \\ L_a(s, a', x) &= s_a - s_a u(x) \qquad \forall a' \neq a, \end{aligned}$$

for all $s \in S$, $a \in A$, and $x \in X$, where $s := (s_a)_{a \in A} \in \Delta(A)$ and $u : X \rightarrow [0, 1]$ is a continuous function. This rule 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 this learning process.

The proof of Proposition 5 reveals that for every utility function one can consider several different learning processes that converge to choose the most preferred actions for such a decision maker. There are other learning processes in the literature that converge to choose, with high probability, expected utility maximizing actions, including, for instance, the belief-based learning model that we discussed in Section 4.[17] Nevertheless, in contrast to the learning rules that satisfy the hypothesis of Lemma 6, the learning rules of belief-based models are not impartial and hence, cannot satisfy the short run property $f_a(s) \geq 0$ for all $a \in A_u^*$ 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

[17] For 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.

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 (or beliefs about) payoff distributions could address this issue. This is left for further research.

An approach that we do not pursue in this paper is defining the learning process as a function of utility functions rather than directly over monetary payoffs. Some of our results hold under both manners of defining the learning process. For instance, consider a learning process whose learning rule $\hat{L}$ is both monotonically risk averse and first-order monotone, but instead of being defined over monetary payoffs, it is defined over an increasing and concave Bernoulli utility function $u : X \to X$. The corresponding learning rule, as defined in our framework, is given by $L_a(s, a', x) = \hat{L}_a(s, a', u(x))$ for all $a, a' \in A$, $s \in S$, and $x \in X$. Impartiality, cross-convexity, and cross-decreasing are preserved when payoffs are transformed using such a function u. Therefore, L is both monotonically risk averse and first-order monotone. In general, however, whether the learning process is defined over the utility function or directly over monetary payoffs can make a substantive difference. For example, let $\hat{L}$ be the learning rule of the Cross model, as defined in Example 1, but now defined over a strictly increasing and strictly concave Bernoulli utility function. Then, the corresponding learning rule, defined over payoffs, is monotonically risk averse and first-order monotone, yet, not monotonically risk neutral as the Cross learning rule. Defining learning processes over utility functions, however, requires establishing conditions that define whether and how individuals use their utility functions (provided that they exist) to update behavior. This analysis is different to what we do here and, hence, we leave it for future 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. Learning models with risk averse learning rules that are not monotonically risk averse, however, do not, in general, 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 given the current state s, namely, $\left(\alpha_a(s) + f_a\left(s\right)\right)_{a \in A}$. 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, $\sum_a \left(\alpha_a(s) + f_a\left(s\right)\right) F_a$. 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 $\sum_a \left(\alpha_a(s) + f_a\left(s\right)\right) F_a$ SOSDs $\sum_a \alpha_a(s) F_a$, 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).[18] 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.[19] 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.[20] 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

[18] See Claim 1 in the Appendix for the proof.

[19] It 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 $\sum_a \left(\alpha_a(s) + f_a\left(s\right)\right) F_a$ sosd $\sum_a \alpha_a(s) F_a$ in every environment. For details, see Claim 2 in the Appendix.

[20] For example, we could consider learning rules that satisfy $\sum_a \left(\alpha_a(s) + f_a\left(s\right)\right) F_a$ sosd $\sum_a \alpha_a(s) F_a$ 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 at that state.

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 processes 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 and Rustichini (1999) make important contributions toward this objective.[21] 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 processes within the framework of this paper.

Finally, our results on convergence of learning processes with 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 *If a learning rule L is first order monotone at $s \in S$ then it is impartial at s.*

[21] Easley 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.

Proof. We argue by contradiction. Consider an environment F with $A = A^{**}$ and suppose that $f_a(s) < 0$ 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 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 $a' \in A \backslash \{a\}$ 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 $\widehat{f}_a(s)$ can be written as a continuous function in ε, for small enough ε we have $\widehat{f}_a(s) < 0$. Therefore L is not first order monotone at s. ■

Now we provide the proof of the proposition.

Proof. *Necessity.*

The necessity of (i) follows from Lemma 7. To prove the necessity of (ii) we argue by contradiction. Suppose that for some $a \in A$ and $a' \in A \backslash \{a\}$ such that $\alpha_{a'}(s) > 0$, there are x and x' with $x' < x$ and $L_a(s, a', x) > L_a(s, a', x')$. Consider the environment F where action a pays x with probability one and action a' pays x' with probability one. All the other actions $a'' \in A \backslash \{a, a'\}$, 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

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

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

Sufficiency.

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

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

The last inequality follows from the fact that $a \in A^{**}$ and the fact that $L_a(s, a', \cdot)$ is decreasing for all $a' \in A \backslash \{a\}$ such that $\alpha_{a'}(s) > 0$. ■

Proof of Lemma 4

Proof. (I) From (3), with $\delta > 0$ defined as in the proof of Lemma 3, we have

$$E\left[\alpha_{t+1}^{*}|s_t\right] \geq \alpha_t^{*} + \delta\alpha_t^{*}(1-\alpha_t^{*})$$

and, hence,

$$E\left[\alpha_{t+1}^{*}|s_t\right] - \frac{1-e^{-\gamma\alpha_t^{*}}}{1-e^{-\gamma}} \geq \alpha_t^{*} + \delta\alpha_t^{*}(1-\alpha_t^{*}) - \frac{1-e^{-\gamma\alpha_t^{*}}}{1-e^{-\gamma}} \tag{5}$$

for all $\gamma > 0$ and $t \in \mathbb{N}_0$.

Consider the function $G : [0,1] \times \mathbb{R}_{++} \to \mathbb{R}$, such that

$$G(z,\gamma) := z + \delta z(1-z) - \frac{1-e^{-\gamma z}}{1-e^{-\gamma}}$$

for all $z \in [0,1]$ and $\gamma > 0$. For this function, $G(0,\gamma) = G(1,\gamma) = 0$,

$$\frac{\partial G(z,\gamma)}{\partial z} = 1 + \delta\left(1-2z\right) - \frac{\gamma e^{-\gamma z}}{1-e^{-\gamma}},$$

and

$$\frac{\partial^2 G(z,\gamma)}{\partial z^2} = -2\delta + \frac{\gamma^2 e^{-\gamma z}}{1-e^{-\gamma}} = -2\delta + \frac{\gamma^2}{e^{\gamma z} - e^{-\gamma(1-z)}}$$

for all $z \in [0,1]$ and $\gamma > 0$. It follows that

$$\lim_{\gamma\to 0}\frac{\partial^2 G(z,\gamma)}{\partial z^2} = -2\delta < 0$$

for all $z \in [0,1]$ and uniformly. Hence for small enough $\gamma^{*} > 0$, $G(\cdot,\gamma^{*})$ is strictly concave. Since $G(0,\gamma^{*}) = G(1,\gamma^{*}) = 0$, it follows that $G(z,\gamma^{*}) > 0$ for all $z \in (0,1)$. Therefore, from (5) and the definition of G, we have

$$E\left[\alpha_{t+1}^{*}|s_t\right] - \frac{1-e^{-\gamma^{*}\alpha_t^{*}}}{1-e^{-\gamma^{*}}} \geq G(\alpha_t^{*},\gamma^{*}) \geq 0$$

for all $t \in \mathbb{N}_0$, and thus,

$$E\left[\alpha_{t+1}^{*}|s_t\right] \geq \frac{1-e^{-\gamma^{*}\alpha_t^{*}}}{1-e^{-\gamma^{*}}} = \phi(\alpha_t^{*}).$$

Since $\phi(z) \geq z$ for all $z \in [0,1]$, we obtain $E\left[\phi(\alpha_{t+1}^{*})|s_t\right] \geq E\left[\alpha_{t+1}^{*}|s_t\right] \geq \phi(\alpha_t^{*})$ and $E\left[\phi(\alpha_{t+1}^{*})\right] \geq E\left[\phi(\alpha_t^{*})\right]$ for all $t \in \mathbb{N}_0$.

(II) For all $\gamma > 0$, ϕ is bounded. Hence, from Lemma 3 and the Continuous Mapping Theorem (footnote 11), $\lim_{t\to\infty} E\left[\phi(\alpha_t^*)\right] = E\left[\phi(\alpha_\infty^*)\right]$. Therefore, from (I), $E\left[\phi(\alpha_\infty^*)\right] \geq \phi(\alpha_0^*) = \phi(\alpha^*(s_0))$. From Lemma 3, we also know that the support of α_∞^* is contained in $\{0,1\}$. Thus, since $E\left[\phi(\alpha_\infty^*)\right] = [1-\Gamma(s_0)]\phi(0) + \Gamma(s_0)\phi(1)$, $\phi\left(0\right) = 0$, and $\phi\left(1\right) = 1$, we obtain $E\left[\phi(\alpha_\infty^*)\right] = \Gamma(s_0)$. It follows that $\Gamma(s_0) \geq \phi(\alpha^*(s_0))$ for all $s_0 \in S$. ■

Proof of Lemma 5

Proof. From Lemma 4, when $\gamma = \gamma^*$, $E[\phi(\alpha_{t+1}^*)|s_t] - \phi(\alpha_t^*) \geq 0$ for all $t \in \mathbb{N}_0$. Therefore,

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

and hence, $-e^{-\gamma^*\alpha_t^*}\left(E\left[e^{-\gamma^*\left(\alpha_{t+1}^*-\alpha_t^*\right)}|s_t\right]-1\right) \geq 0$, for all $t \in \mathbb{N}_0$.

For any $\widehat{\gamma} > \gamma^*$, let $\widehat{\theta} := \gamma^*/\widehat{\gamma}$, hence $\gamma^* = \widehat{\gamma}\widehat{\theta}$. Therefore,

$$\begin{aligned}
-e^{-\widehat{\gamma}\widehat{\theta}\alpha_t^*}\left(E\left[e^{-\widehat{\gamma}\widehat{\theta}\left(\alpha_{t+1}^*-\alpha_t^*\right)}|s_t\right]-1\right) &\geq 0 \\
-e^{-\widehat{\gamma}\alpha_t^*}\left(E\left[e^{-\widehat{\gamma}\widehat{\theta}\left(\alpha_{t+1}^*-\alpha_t^*\right)}|s_t\right]-1\right) &\geq 0,
\end{aligned}$$

for all $t \in \mathbb{N}_0$. It follows that for the learning process $(\alpha, \pi^{\widehat{\theta}})$,

$$E\left[\frac{1-e^{-\widehat{\gamma}\alpha_{t+1}^*}}{1-e^{-\widehat{\gamma}}}|s_t\right] - \frac{1-e^{-\widehat{\gamma}\alpha_t^*}}{1-e^{-\widehat{\gamma}}} \geq 0,$$

for all $t \in \mathbb{N}_0$. Then, as in the proof of Lemma 4, we obtain that for $\gamma = \widehat{\gamma}$, $(\alpha, \pi^{\widehat{\theta}})$ satisfies $E\left[\phi(\alpha_{t+1}^*)\right] - E\left[\phi(\alpha_t^*)\right] \geq 0$ for all $t \in \mathbb{N}_0$, and hence, $\Gamma^{\widehat{\theta}}(s_0) \geq \phi(\alpha^*(s_0))\ \forall\ s_0 \in S$. ■

Claim 1 *For problems such that $|A| = 2$, the only learning rules which satisfy $\Sigma_a\left(\alpha_a(s) + f_a(s)\right)F_a$ SOSD $\Sigma_a\alpha_a(s)F_a$ in the state s, in every environment, are the unbiased rules at s.*

Proof. From the proof of Lemma 1, we have that impartiality at s is a necessary condition for a learning rule to satisfy $\Sigma_a\left(\alpha_a(s) + f_a(s)\right)F_a$ SOSD $\Sigma_a\alpha_a(s)F_a$ in every environment, with any finite number of actions.

Now we suppose $|A| = 2$ and prove that $L_a(s, a', \cdot)$ is both concave and convex for all $a', a \in A$ is a necessary condition too. Since L has to be impartial at s,

$$f_a(s) = \alpha_{a'}(s) \left[\int L_a(s, a', x) dF_{a'}(x) - \int L_a(s, a', x) dF_a(x) \right],$$

where $a' \neq a$, for all $a \in A$. Suppose $L_a(s, a', \cdot)$ is not convex, thus there exists x'', $x' \in X$ such that $L_a(s, a', x) > \lambda L_a(s, a', x') + (1 - \lambda) L_a(s, a', x'')$ 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

$$f_a(s) = \alpha_{a'}(s) \left[\lambda L_a(s, a', x') + (1 - \lambda) L_a(s, a', x'') - L_a(s, a', x) \right] < 0.$$

It follows that $\Sigma_a \alpha_a(s) F_a$ strictly-SOSD $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$. Therefore, a necessary condition for a learning rule to satisfy $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ SOSD $\Sigma_a \alpha_a(s) F_a$ in every environment is that $L_a(s, a', \cdot)$ is convex for $a' \neq a$.

To complete the proof we need to prove that the concavity of $L_a(s, a', \cdot)$ for $a' \neq a$ is necessary as well. The convexity of $L_a(s, a', \cdot)$ implies that it is continuous on $(x_{\min}, x_{\max})$ for all a and $a' \neq a$. First, we prove that $L_a(s, a', \cdot)$ is concave on $(x_{\min}, x_{\max})$ for all a and $a' \neq a$. Suppose not, i.e., for some $a \in A$ and $a' \neq a$, there exists $x', x'' \in (x_{\min}, x_{\max})$, with $x' < x''$, such that $L_a(s, a', x) < \lambda L_a(s, a', x') + (1 - \lambda) L_a(s, a', x'')$ for some $\lambda \in (0, 1)$ and $x := \lambda x' + (1 - \lambda) x''$. Consider an environment where F_a corresponds to the lottery $(x', \delta\lambda; x, 1 - \delta; x'', (1 - \lambda)\delta)$ for some $\delta \in (0, 1)$ and $F_{a'}$ corresponds to the lottery $(x' + \frac{1-\lambda}{\lambda}\rho, \lambda; x'' - \rho, 1 - \lambda)$ for $\rho > 0$ such that these payoffs are contained in $(x_{\min}, x_{\max})$. Therefore,

$$\begin{aligned}
f_a(s) &= \alpha_{a'}(s)[\lambda L_a(s, a', x' + \frac{1-\lambda}{\lambda}\rho) + (1 - \lambda) L_a(s, a', x'' - \rho) - \\
&\quad (\delta\lambda L_a(s, a', x') + (1 - \delta) L_a(s, a', x) + (1 - \lambda)\delta L_a(s, a', x''))] \\
&= \alpha_{a'}(s)[\lambda L_a(s, a', x' + \frac{1-\lambda}{\lambda}\rho) + (1 - \lambda) L_a(s, a', x'' - \rho) - L_a(s, a', x) - \\
&\quad \delta \left(\lambda L_a(s, a', x') - L_a(s, a', x) + (1 - \lambda) L_a(s, a', x'') \right)].
\end{aligned}$$

Since $L_a(s, a', \cdot)$ is continuous on $(x_{\min}, x_{\max})$, for small enough ρ, in the RHS of the last equality, we have $\lambda L_a(s, a', x' + \frac{1-\lambda}{\lambda}\rho) + (1 - \lambda) L_a(s, a', x'' - \rho) - L_a(s, a', x) > 0$. Thus for small enough ρ and δ, we have $f_a(s) > 0$. 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 $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ does not SOSD $\Sigma_a \alpha_a(s) F_a$. Hence $L_a(s, a' \cdot)$ needs to be both concave and convex on $(x_{\min}, x_{\max})$ for all $a' \neq a$.

Thus, we are left to prove that $\lim_{x \to x_{\min}} L_a(s, a', x) = L_a(s, a', x_{\min})$ and $\lim_{x \to x_{\max}} L_a(s, a', x) = L_a(s, a', x_{max})$. Since $L_a(s, a', \cdot)$ is convex we only need to rule out $L_a(s, a', x_{min}) > \lim_{x \to x_{\min}} L_a(s, a', x)$ and $\lim_{x \to x_{\max}} L_a(s, a', x) < L_a(s, a', x_{max})$. Consider an environment F such that the payoff distribution of a' yields $x_{\min} + \varepsilon$ and $x_{\max} - \varepsilon$ with probability $\varepsilon/2$, and $x_{\min} + \frac{1}{4}(x_{\max} - x_{\min})$ and $x_{\min} + \frac{3}{4}(x_{\max} - x_{\min})$ with probability $(1 - \varepsilon)/2$ for some $\varepsilon \in (0, \min\{1, (x_{\max} - x_{\min})/2\})$; and action a yields $x_{\min}$ and $x_{\max}$ with probability $\varepsilon/2$, and $x_{\min} + \frac{1}{2}(x_{\max} - x_{\min})$ with probability $1 - \varepsilon$. If $L_a(s, a', x_{min}) = \lim_{x \to x_{\min}} L_a(s, a', x)$ and $\lim_{x \to x_{\max}} L_a(s, a', x) = L_a(s, a', x_{max})$, the learning rule would be unbiased and we would have $f_a(s) = 0$. Instead, if $L_a(s, a', x_{min}) > \lim_{x \to x_{\min}} L_a(s, a', x)$, or $\lim_{x \to x_{\max}} L_a(s, a', x) < L_a(s, a', x_{max})$, or both, we have $f_a(s) < 0$, and, for some small enough ε, $\int u(x) dF_a > \int u(x) dF_{a'}$ for the concave function $u : X \to \mathbb{R}$, $u(x) = (x - x_{\min})^{1/2}$. This yields $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ does not SOSD $\Sigma_a \alpha_a(s) F_a$. ■

Claim 2 *The only learning rules such that $L_a(s, a', \cdot)$ is continuous for all $a, a' \in A$ and satisfy $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ SOSD $\Sigma_a \alpha_a(s) F_a$ in every environment, in the state s, are the learning rules that are unbiased at s.*

Proof. From the proof of Claim 1, learning rules which satisfy $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ SOSD $\Sigma_a \alpha_a(s) F_a$ in every environment are impartial. Now we prove that continuous rules which satisfy $\Sigma_a (\alpha_a(s) + f_a(s)) F_a$ SOSD $\Sigma_a \alpha_a(s) F_a$ in every environment also need to be both cross-convex and cross-concave. Since these rules are impartial, we have

$$f_a(s) = \sum_{a' \neq a} \alpha_{a'}(s) \left[\int L_a(s, a', x) dF_{a'}(x) - \int L_a(s, a', x) dF_a(x) \right]$$

for all $a \in A$. Suppose there is $a \in A$ and $a' \neq a$ such that $L_a(s, a', \cdot)$ is not convex. Thus there exists $x'' > x' \in X$ such that $L_a(s, a', x) > \lambda L_a(s, a', x') + (1 - \lambda) L_a(s, a', x'')$ 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

$$\begin{aligned} f_a(s) &= \alpha_{a'}(s)[L_a(s,a',x) - \lambda L_a(s,a',x') - (1-\lambda)L_a(s,a',x'')] + \\ &\sum_{a'' \neq a,a'} \alpha_{a''}(s)\left(\lambda L_a(s,a'',x' + (1-\lambda)\lambda^{-1}\varepsilon) + (1-\lambda)L_a(s,a'',x''-\varepsilon)\right. \\ &\left.-\lambda L_a(s,a',x') - (1-\lambda)L_a(s,a',x'')\right). \end{aligned}$$

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, $f_a(s) > 0$. 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''), $\Sigma_a\left(\alpha_a(s) + f_a(s)\right)F_a$ does not SOSD $\Sigma_a \alpha_a(s) F_a$.

Suppose there is $a \in A$ and $a' \neq a$ such that $L_a(s,a',\cdot)$ is not concave. Thus there exists $x'' > x' \in X$ such that $L_a(s,a',x) < \lambda L_a(s,a',x') + (1-\lambda)L_a(s,a',x'')$ 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

$$\begin{aligned} f_a(s) &= \alpha_{a'}(s)[\lambda L_a(s,a',x' + \frac{1-\lambda}{\lambda}\rho) + (1-\lambda)L_a(s,a',x''-\rho) \\ &-\delta\lambda L_a(s,a',x') - (1-\delta)L_a(s,a',x) - (1-\lambda)\delta L_a(s,a',x'')] + \\ &\sum_{a'' \neq a,a'} \alpha_{a''}(s)(\delta\lambda L_a(a'',x' + \frac{1-\lambda}{\lambda}\varepsilon) + (1-\delta)L_a(s,a'',x) + \delta(1-\lambda)L_a(s,a'',x''-\varepsilon) \\ &-\delta\lambda L_a(s,a'',x') - (1-\delta)L_a(s,a'',x) - (1-\lambda)\delta L_a(s,a'',x'')) \\ &= \alpha_{a'}(s)[\lambda L_a(s,a',x' + \frac{1-\lambda}{\lambda}\rho) + (1-\lambda)L_a(s,a',x''-\rho) - L_a(s,a',x) + \\ &\delta\left(L_a(s,a',x) - \lambda L_a(s,a',x') - (1-\lambda)L_a(s,a',x'')\right)] + \\ &\sum_{a'' \neq a,a'} \alpha_{a''}(s)(\delta\lambda L_a(s,a'',x' + \frac{1-\lambda}{\lambda}\varepsilon) + (1-\delta)L_a(s,a'',x) + \delta(1-\lambda)L_a(s,a'',x''-\varepsilon) \\ &-\delta\lambda L_a(s,a'',x') - (1-\delta)L_a(s,a'',x) - (1-\lambda)\delta L_a(s,a'',x'')). \end{aligned}$$

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, $f_a(s) > 0$.

Thus $\Sigma_a\left(\alpha_a(s)+f_a(s)\right)F_a$ does not SOSD $\Sigma_a\alpha_a(s)F_a$, 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. Lakshmivarahan, 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.