University of Leicester

DEPARTMENT OF ECONOMICS

Learning without Counterfactuals*

Friederike Mengel, Maastricht University, NL
Javier Rivas, University of Leicester, UK

Working Paper No. 10/15
February 2010

Learning without Counterfactuals[*]

Friederike Mengel
Maastricht University[†]

Javier Rivas
University of Leicester[‡]

February 23, 2010

Abstract

In this paper we study learning procedures when counterfactuals (payoffs of not-chosen actions) are not observed. The decision maker reasons in two steps: First, she updates her propensities for each action after every payoff experience, where propensity is defined as how much she prefers each action. Then, she transforms these propensities into choice probabilities. We introduce natural axioms in the way propensities are updated and the way propensities are translated into choice, and study the decision marker's behavior when such axioms are in place.

JEL Classification: D83.

Keywords: Adaptive Learning, Counterfactuals, Partial Information, Reinforcement Learning.

[*] We thank the audience at the University of Leicester for useful comments and suggestions. Friederike Mengel thanks the European Union for financial support (grant PIEF-2009-235973).

[†] Department of Economics (AE1), Maastricht University, PO Box 616, 6200MD Maastricht, Netherlands. F.Mengel@maastrichtuniversity.nl

[‡] Department of Economics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom. javier.rivas@le.ac.uk., www.le.ac.uk/users/jr168/index.htm.

1 Introduction

In this paper we are concerned with learning by economic decision-makers who have to take decisions in situations where little is known about the performance of each of the alternatives available. For example, consider decisions such as how to conduct business negotiations, which consumption goods to buy, and in which assets to invest our money. In many such situations, we may not know how many different states of nature are possible, the probability distribution over these states or the payoffs associated with each alternative and state. Moreover, even if all this information was available, processing it in order to make optimal choices may prove overly complicated.

Traditional economic theory assumes that even in such situations agents act as if they were maximizing expected utility with a unique subjective probability distribution (Savage, 1954 and Anscombe and Aumann, 1963). In situations like the ones described above it seems unclear why expected utility maximization should be successful in predicting behavior well. An alternative is to study models where agents arrive at their decisions by learning from their own experiences or via communication with others[1]. This is often done by either imposing some optimality properties on the learning rule agents use or by simply positing an ad-hoc model which seems a good description of actual behavior[2].

In this paper we take a somewhat different route. As in other learning models, we propose a setting where the decision maker evaluates the different alternatives according to the payoffs she obtained in the past. Instead of simply positing a model, though, we examine the implications of some natural and simple requirements on how learning occurs. Unlike in the literature on optimal learning, we do not impose any kind of optimality requirement on way learning occurs via our axioms. As opposite to this, we try to find axioms which characterize natural behavior in such situations. The only knowledge the decision maker has in our setting is the set of available actions. She does not necessarily know about the state space, the probability distribution over such state, or the payoff associated with each action at a given state. Over time, the decision maker observes the payoff she obtains in every period and uses that information for choosing next period.

Our approach is similar to Easley and Rustichini (1999). However, unlike them we assume that the decision maker does not observe counterfactuals. That is, at any given period, she does not observe the payoff of the actions she does not choose. This seemingly small difference has a big impact in the way the decision maker tackles the problem of learning: If counterfactuals are observed, the action chosen is completely irrelevant for the learning

[1] Fudenberg and Levine (1998) have surveyed some of the vast literature on learning.

[2] Examples of the first class of models are Börgers et al (2004) or Schlag (1998). Examples for the second class are Roth and Erev (1995), Camerer and Ho (2004) among many others.

process as the decision maker learns the same information independently on her choices. On the other hand, if counterfactual are not observed, learning is crucially affected by the action chosen as the decision maker only learns about the actions she chooses.

Therefore, in an environment where counterfactuals are not observed, the decision maker has a trade-off between exploitation of the currently most preferred action and exploration of the other alternatives[3]. The fact that counterfactuals are not observed, as we explain below, gives rise to a separation between how much the decision maker prefers each action and how these preferences translate into choices. Hence, the learning procedures we study can be characterized through two processes: First, we have the *updating rule*. The updating rule specifies how new information, experiences, etc. affect the decision marker's propensity towards each of the possible actions. These propensities could represent beliefs about the distribution of payoffs associated with each action or a much wider set of feelings such as confirmatory bias or forgetting. Secondly, there is the *choice rule*, which specifies how these propensities are translated into actual choices.

As already mentioned, and quoting Easley and Rustichini (1999), "our interest is not in the existence of a procedure that "works"". Hence, we do not make any requirement regarding optimality from the learning rules *ex ante*, i.e. through our axioms. Instead, the axioms we pose are meant to capture natural features on the agent's behavior rather than desirable properties of the learning rule. However, in a second step, we investigate how these natural requirements relate to optimality.

In our results we also relate how the learning behavior induced by our axioms relates to learning rules that are known in the literature. In particular, we show that under our main axioms the resulting learning procedure is a form of reinforcement learning which approximates the replicator dynamics from evolutionary game theory[4].

There have been other axiomatizations of learning procedures that lead to a behavior that resembles replicator dynamics. To our knowledge, this has been the case mostly in the literature on optimal learning with bounded rational agents (see, for example, Börgers et al (2004) and Schlag (1998) among others). Our approach is different in that we are not interested in characterizing optimal rules. The only axiomatization of the replicator dynamics using natural axioms is due to Easley and Rustichini (1999). However, and as already mentioned, their paper deals with learning when counterfactuals are observed. In section 6.3 we explain the differences between Easley and Rustichini (1999) and our approach in more detail.

[3] Even though we are dealing with a possibly bounded rational decision maker, the rational solution to such a problem is also far from trivial (See e.g. Bergemann and Vaelimaeki, 2006).

[4] See Roth and Erev (1995), Bush and Mosteller (1951) or Sutton and Barto (1998).

The rest of the paper is organized as follows. In section 2 we present the learning environment. Section 3 introduces the axioms on the transition function and give a characterization given such axioms. Section 4 proceeds likewise and a characterization of the choice function is presented. In section 5, we establish efficiency and optimality results for the learning rules resulting from our axioms. A relation between the replicator dynamics and the characterizations resulting from our axioms is presented in section 6. Finally, section 7 concludes.

2 The Model

Consider a decision maker who at each period $t = 0, 1, \ldots$ chooses an action from the finite set $A = \{1, \ldots, n\}$. Every action yields a random payoff $\pi \in \Pi = \mathbb{R}_+$. Denote by π_i^t the realization of π at time t of action i.

Denote by $\bar{H}^t = \{\bar{h}^t\}_{\tau=1}^{t-1}$ the set of all possible histories at $t \geq 1$ with elements $\bar{h}^t \in A \times \Pi^n = \bar{H}^t$ and let $\bar{H} = \bigcup_{t \geq 1} \bar{H}^t$ be the entire set of histories. Denote H^t, h^t, H correspondingly as the part of the history that is observed by the decision maker. The initial histories are assumed to be empty $\bar{H}^0, \bar{h}^0 = \emptyset$ and $H^0, h^0 = \emptyset$. Finally, let $\pi_i \left(h^t\right)$ be the sum of payoffs of action i under history h^t. If action $i \in A$ was not chosen in history h^t then we set $\pi_i \left(h^t\right) = \pi_i^0$. For any history h^t denote by $\#i \left(h^t\right)$ the number of times action i is chosen in h^t.

At each point in time the decision maker is assumed to have propensities $\theta \in \Theta = \mathbb{R}_+^n$ where $\theta_i \in \Theta_i$ is the propensity of action $i \in A$ and $\prod_{i=1}^n \Theta_i = \Theta$. We use θ_i^t to denote the propensity of action i at time t. The initial vector of propensities $\theta^0 \in \Theta$ is given. Propensities can be interpreted as a numerical representation of preferences[5].

Since our agent does not observe counterfactuals, his choices matter crucially for learning. In particular, it is not necessarily optimal for her to always choose the action with the highest propensity (as it is the case under expected utility maximization) as this would preclude learning about other actions. Hence, the agent randomizes his choice every period where the probability of choosing each action is given by $p : \Theta \to \Delta^n$ with $p = (p_1, \ldots, p_n)$ where Δ^n is the n-dimensional unit simplex.

At the end of each period the decision maker observes the payoff of the action played and updates her propensities θ. Propensities over actions change according to the transition function $T : \Theta \times A \times \Pi \to \Theta$. Hence, $T_i \left(\theta, j, \pi_j^t\right)$ is the new propensity of action i if propensities are given by θ and the agent chooses j and obtains a payoff π_j^t. That is,

$$\theta_i^{t+1} = T_i(\theta^t, j, \pi_j^t)$$

[5] Such numerical representation exist under expected utility maximization by the von Neumann-Morgenstern theorem

for all $i \in A$ when action j is chosen and payoff π_j^t is obtained.

3 Axioms on Updating Rule T

As mentioned in the introduction, the decision maker separates her reasoning into two steps. First, there are her propensities (Θ), which tells us how much she prefers each action. Second, we have her choices (p), which are simply an application from propensities into probabilities. In this section we introduce a set of axioms on the way the propensities are updated. These axioms are not motivated by optimality considerations but rather by behavioral rules of thumb.

Our first axiom on the way propensities are updated deals with the fact that there is, a priori, no reason why the decision maker would treat propensities differently based only on the label of the actions. This axiom embodies the requirement that the decision-maker should not have preferences which do not stem from her own experience. It avoids all unreasonable bias towards any action.

Axiom 1 (Anonymity). *The transition function T does not depend on actions per se. That is, $T_i\left(\theta, j, \pi_j^t\right) = T_k\left(\theta, j, \pi_j^t\right)$ if and only if $\theta_i = \theta_k$ and either $j \neq i, k$ or $j = i, k$.*

Requiring some kind of Anonymity is standard and can also be found in, for instance, Börgers et al (2004) or Easley and Rustichini (1999). The second axiom is a monotonicity requirement. It implies that, all things equal, starting with a higher propensity results in a higher propensity after the updating takes place. It also implies that, all things equal, higher payoff translates into higher new propensity of the action chosen. Note that the axiom makes claim between different actions. That is, it relates only the updating of the propensity of a given action i with this same action's propensity and payoff θ_i, π_i.

Axiom 2 (Monotonicity). *The transition function T_i is strictly increasing in Θ_i and Π_i.*

Our third axiom deals with how big the changes in propensities can be. We assume the decision maker updates propensities in a sufficiently smooth way. In particular, we assume that T is differentiable with respect to initial propensities θ and payoff π. As an extension to our main results, in section 3.1.1 we characterize the function T when differentiability (nor continuity) is not assumed. We choose to focus on differentiability as the resulting function T is more reasonable as it does not present any sudden changes in slope or jumps.

Axiom 3 (Differentiability). *The transition function T is differentiable in Θ and Π.*

Our final axiom specifies how payoffs are treated when updating propensities. The following axiom states that, for any two possible histories that start with the same propensity for a given action and where the number of times a given action has been played coincides across these two histories, then the sum of payoffs determines under which history the propensity of the given axiom is higher. Note that the axiom makes no claim on: how to compare propensities of different actions and, on how to compare the propensities under two different histories where the given action has been played a different number of times or had different initial propensity. Hence the term Weak.

Axiom 4 (Weak Sum). *For any histories h^t and $\bar{h}^k$ and any action i, T is such that if $\theta_i^0 = \bar{\theta}_i^0$ and $\#i\left(h^t\right) = \#i\left(\bar{h}^k\right)$, then $\theta_i^t \geq \bar{\theta}_i^k$ if and only if $\pi_i\left(h^t\right) \geq \pi_i\left(\bar{h}^k\right)$.*

The aim of this paper is to understand learning without counterfactuals when natural axioms are in place. We are aware that "natural" is an, at most, vague term. Therefore, considering alternatives to our main axioms is an important part of our article. In section 3.1 we consider other alternatives to the axioms presented above.

We move now to present a characterization on the way propensities are updated given the four axioms above. Our first result is that if an action is not chosen, then its propensity does not change. This is a consequence of the Weak Sum axiom.

Lemma 1. *Given any transition T that satisfies Weak Sum, we have that $\theta_i^{t+1} = T_i(\theta^t, j, \pi_j)$ for any θ^t, $j \neq i$ and π_j.*

Proof. Fix the initial propensities θ^t. Assume action i is not played at time t. Consider the alternative event (history) where action i is played at time t. Weak Sum implies that $\theta_i^{t+1} = \hat{\theta}_i^t$ where $\hat{\theta}$ indicates the value of θ under the alternative history. Since, by construction, $\theta_i^t = \hat{\theta}_i^t$, we can conclude that $\theta_i^{t+1} = \theta_i^t$. Thus, we have that the propensity of an action only changes if that action is played. $\square$

The next result states that the propensity of a given action is not affected by the propensities of the other actions. That is, propensities are updated independently across action. This results is again a consequence of the Weak Sum axiom.

Lemma 2. *Assume that T satisfies Weak Sum. For any pair θ_i, π we have that $T_i\left(\theta, i, \pi\right)$, where θ_i is an element of θ, is independent of θ_{-i}.*

Proof. By Weak Sum T, is such that starting from any θ^t and for any two histories h^{t+m} and $\hat{h}^{t+m}$ with $m \in \ltimes^+$ that are different only in the periods where action i is not chosen, the value of θ_i^{t+m} equals the value of $\hat{\theta}_i^{t+m}$. Thus, the propensities of all actions $-i$ do not influence the propensity of action i and, hence, the function T_i does not depend on θ_{-i}. $\square$

We are finally able to present a characterization of the transition function T. Proposition 1 below states that the way propensities are updated is linear in payoffs. The results is a consequence of lemmas 1 and 2 above and axioms 1-4.

Proposition 1. *A transition T satisfies axioms 1- 4 if and only if for all $i \in A$ there exists a $\lambda > 0$ and a $\rho \geq 0$ such that*

$$\theta_i^{t+1} = \begin{cases} \theta_i^t + \lambda \pi + \rho & \textit{if action } i \textit{ is played and payoff } \pi \textit{ is obtained,} \\ \theta_i^t & \textit{otherwise.} \end{cases}$$

Note that the expression above has two free parameters: λ and ρ. The parameter λ represents how payoffs increase propensities while the parameter ρ acts as a counter on how many times an action has been played.

Proof. First note that by Lemma 2 we can find a function $g_i : \Theta_i \times \Pi$ that represents T_i for all $i \in A$. That is, for all $\theta \in \Theta$, $i \in A$, $\pi \in \Pi$ we have that

$$g_i(\theta_i, \pi) \quad = \quad T_i(\theta, i, \pi)$$

where θ_i is an element of θ.

Take any two payoffs $\hat{\pi}, \bar{\pi} \in \Pi$ and consider the following two histories:

$$\begin{aligned} h^2 &= ((i, \hat{\pi}), (i, \bar{\pi})), \\ \hat{h}^2 &= ((i, \bar{\pi}), (i, \hat{\pi})). \end{aligned}$$

By Weak Sum we have that $\theta^2 = \hat{\theta}^2$. Thus, using the definition of g we have that for any $\theta_i \in \Theta_i$ and any $\hat{\pi}, \bar{\pi} \in \Pi$

$$g_i(g_i(\theta_i, \hat{\pi}), \bar{\pi}) \quad = \quad g_i(g_i(\theta_i, \bar{\pi}), \hat{\pi}). \tag{1}$$

Furthermore, Weak Sum implies that for all $\delta, \varepsilon > 0$

$$\begin{aligned} g_i(g_i(\theta_i, \hat{\pi}) + \delta, \bar{\pi}) &= g_i(g_i(\theta_i, \bar{\pi}) + \delta, \hat{\pi}), \\ g_i(g_i(\theta_i, \hat{\pi}), \bar{\pi} + \varepsilon) &= g_i(g_i(\theta_i, \bar{\pi}), \hat{\pi} + \varepsilon). \end{aligned}$$

Thus, since g is differentiable, we have that

$$\left.\frac{\partial g_i}{\partial \theta}\right|_{(g_i(\theta_i,\hat{\pi}),\bar{\pi})} = \left.\frac{\partial g_i}{\partial \theta}\right|_{(g_i(\theta_i,\bar{\pi}),\hat{\pi})}, \tag{2}$$

$$\left.\frac{\partial g_i}{\partial \pi}\right|_{(g_i(\theta_i,\hat{\pi}),\bar{\pi})} = \left.\frac{\partial g_i}{\partial \pi}\right|_{(g_i(\theta_i,\bar{\pi}),\hat{\pi})} \tag{3}$$

for all $\theta_i \in \Theta_i$ and $\tilde{\pi}, \bar{\pi} \in \Pi$.

If we differentiate now (1) with respect to $\hat{\pi}$ we obtain

$$\left.\frac{\partial g_i}{\partial \theta}\right|_{(g_i(\theta_i,\hat{\pi}),\bar{\pi})} \left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\hat{\pi})} = \left.\frac{\partial g_i}{\partial \pi}\right|_{(g_i(\theta_i,\bar{\pi}),\hat{\pi})}. \tag{4}$$

On the other hand, differentiating (1) with respect to $\bar{\pi}$ we obtain

$$\left.\frac{\partial g_i}{\partial \pi}\right|_{(g_i(\theta_i,\hat{\pi}),\bar{\pi})} = \left.\frac{\partial g_i}{\partial \theta}\right|_{(g_i(\theta_i,\bar{\pi}),\hat{\pi})} \left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\bar{\pi})}. \tag{5}$$

Thus, combining (2), (3), (4) and (5) we get that

$$\left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\hat{\pi})} = \left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\bar{\pi})}$$

for all $\theta_i, \hat{\pi}, \bar{\pi}$. Therefore, using (3) and the fact that g is differentiable by Differentiability, there must exist a constant λ such that

$$\left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\pi)} = \lambda \tag{6}$$

for all θ_i, π.

Using (4) and (6) yields

$$\left.\frac{\partial g_i}{\partial \theta}\right|_{(g_i(\theta_i,\hat{\pi}),\bar{\pi})} = 1.$$

Since this is true for any $\theta_i, \hat{\pi}, \bar{\pi}$, we have that

$$g_i(\theta_i, \pi) = \theta_i + f(\pi) \tag{7}$$

for some strictly increasing (Axiom 2) and everywhere differentiable (Axiom 3) function $f : \Pi \to \Theta$. But by equation (6) we must have that

$$f(\pi) = \lambda\pi + \rho. \tag{8}$$

for some λ and ρ. By Monotonicity we have that $\lambda > 0$, and by the fact that propensities are defined in $\mathbb{R}_+$, $\rho \geq 0$. Combining (7), (8) and Lemma 1 gives the desired result. □

3.1 Discussion on the Axioms of T

3.1.1 Differentiability

The differentiability axiom has been assumed as it seems natural that the decision maker would want to update propensities in a sufficiently smooth way. However, one might argue that this may not necessarily be the case. We now deal with how the decision maker updates propensities when there is no assumption regarding the smoothness of T. The following result has an answer:

Proposition 2. *A transition T satisfies axioms 1, 2 and 4 if and only if for all $i \in A$ and almost all θ_i^t there exists a $\lambda > 0$, $\rho \in \Pi$ and $\varepsilon > 0$ such that*

$$\theta_i^{t+1} = \begin{cases} \theta_i^t + \lambda\pi + \rho & \text{when } i \text{ is played and } \pi \text{ is obtained,} \\ \theta_i^t & \text{otherwise} \end{cases}$$

for $\pi \in B_\varepsilon(0)$.

Proof. A known result is that any monotone function is almost everywhere differentiable (see, for example, Theorem 14a in Chabrillac and Crouzeix (1987)). Thus, since T is monotonous, it is almost everywhere differentiable. Once this fact is established, the rest of the proof is a straightforward extension to the proof of Proposition 1. □

If differentiability is not assumed, then the functional form in Proposition 1 is still valid for small neighborhoods of given propensities. However, the functional form of T for a given neighborhood of θ, π may not be valid for the entire $\Theta \times \Pi$ space. It may happen that T is neither differentiable nor continuous for some finite set of points θ as the following example shows:

Example 1.

$$\theta_i^{t+1} = \begin{cases} \theta_i^t + \lambda_1\pi & \text{if } \theta_i^t + \lambda_1\pi < \alpha, \\ \alpha + \left(\theta_i^t + \lambda_1\pi - \alpha\right)\frac{\lambda_2}{\lambda_1} + \rho & \text{if } \theta_i^t < \alpha \text{ and } \theta_i^t + \lambda_1\pi \geq \alpha, \\ \theta_i^t + \lambda_2\pi & \text{otherwise} \end{cases}$$

for some $\lambda_1, \lambda_2, \rho > 0$ and $\alpha > 0$.

It is easy to check that the transition T represented above satisfies axioms 1, 2 and 4 but it doesn't satisfy Differentiability (T is not even continuous at $\theta = \alpha$). In this case, each payoff π increases the propensity of its respective action by $\lambda\pi$ where the value of λ is not constant throughout the whole $\Theta \times \Pi$ space. The transition T is has also a discontinuity at $\theta = \alpha$ where propensities are increased by ρ. We choose to avoid such non-smooth behavior by assuming the transition T to be differentiable.

3.1.2 Weak Sum

The Weak Sum axiom has the feature that all payoffs are treated equally independently on when they happened. Thus, a t-period old payoff affects propensities in the same way as a payoff obtained in the present period. In this subsection we consider an alternative to the Weak Sum axiom where recent payoffs affect the current propensities more than older ones.

Define $\pi\left(\delta, h^t\right)$ as the discounted sum of payoffs of action i in history h^t where each payoff is discounted at the rate of $\delta \in (0,1)$ per period. Thus, a t-period old payoff of π equals to a payoff of $\delta^t \pi$ today. An alternative to Weak Sum is the following:

Axiom' 1 (Weak Weighted Average). *For any histories h^t and $\bar{h}^t$ and any action i, T is such that if $\theta_i^0 = \bar{\theta}_i^0$ and $\#i\left(h^t\right) = \#i\left(\bar{h}^t\right)$, then $\theta_i^t \geq \bar{\theta}_i^t$ if and only if $\pi_i\left(\delta, h^t\right) \geq \pi_i\left(\delta, \bar{h}^t\right)$ for some $\delta \in (0,1)$.*

Note that two histories can be compared using the Weak Weighted Average only if they are the same length as otherwise it can be shown that no transition T exists (see Appendix 8.1.2). If we replace the Weak Sum axiom with the Weak Weighted Average axiom we obtain the following result (proof in Appendix 8.1.1):

Proposition 3. *If transition T satisfies axioms 1-3 and 1' then for all $i \in A$ there exists a $\lambda > 0$ and $\rho > 0$ such that*

$$\theta_i^{t+1} = \begin{cases} \delta\theta_i^t + \lambda\pi + \rho & \textit{if action } i \textit{ is played and payoff } \pi \textit{ is obtained,} \\ \delta\theta_i^t + \rho & \textit{otherwise.} \end{cases}$$

Note that the functional form of the transition T_i in this case has three free parameters: λ, ρ and δ.

The Weak Weighted Average is not the only reasonable alternative to the Weak Sum axiom. So far we have assumed that the decision maker has no concern about risk, that is, we have taken her to be risk neutral. A natural alternative to this is to consider a risk averse decision maker. In particular, we could assume that the decision maker attitude towards risk is lexicographic in the sum of payoff and variance of payoffs. That is,

Axiom' 2 (Risk Averse Sum). *For any histories h^t and $\bar{h}^k$ and any action i, T is such that if $\theta_i^0 = \bar{\theta}_i^0$ and $\#i\left(h^t\right) = \#i\left(\bar{h}^k\right)$ then $\pi_i\left(h^t\right) > \pi_i\left(\bar{h}^k\right)$ implies $\theta_i^t > \bar{\theta}_i^k$, and $\pi_i\left(h^t\right) = \pi_i\left(\bar{h}^k\right)$ implies $\theta_i^t \geq \bar{\theta}_i^k$ if and only if $var\left(\pi_i\left(h^t\right)\right) \leq var\left(\pi_i\left(\bar{h}^k\right)\right)$.*

One can show, however, that there exists no transition T satisfying axioms 1-3 and 2' (see Appendix 8.1.3). An alternative way of treating risk could be assuming the decision maker treats sum of payoffs and variance as substitutes. This would imply that she is willing to give up higher payoffs in favor of less variance. However, an axiom targeting a representation of such behavior is troublesome. In particular, it is not clear what is the natural way of exchanging risk into payoff and vice versa.

4 Axioms on Choice Function p

We proceed now to study how propensities are translated into choices. How the decision maker chooses between each of the available actions is crucial for the learning process as the only way the decision maker can gain information about an action is by choosing it.

Our first two axioms are equivalent to the Anonymity and Monotonicity axioms for T. As for the transition functions, Anonymity implies that all actions are given equal treatment in the sense that their label does not matter. Monotonicity means that, other things equal, the decision maker tends to assign more probability to the action she prefers more. That is, higher propensities implies higher probability.

Axiom 5 (Anonymity). *The choice rule p does not depend on actions per se.*

Axiom 6 (Monotonicity). *Each choice function p_i is strictly increasing in θ_i for all $i \in A$.*

Our third axiom, Continuity, relates to how smooth Behaviour is. In particular, Continuity means that there are no discrete jumps in how propensities are translated into choices. Later in section 4.1.1 we discuss what are the implications of dropping the Continuity axiom.

Axiom 7 (Continuity). *Each choice function p_i is continuous in Θ.*

The final axiom relates to how the relative differences in propensities are translated into relative differences in choice probabilities. Boundedness means that the decision maker should not exaggerate the relative differences in propensities when translating these into choices. This has the natural interpretation that the decision maker is always inclined to "try" the different actions, albeit with small probability, even if the propensities of these actions is low. We believe that in an environment where counterfactuals are not observed such cautiousness or willingness to investigate is reasonable. In section 4.1.2 we prove that Boundedness can be thought as a consequence of another two axioms that we introduce later: Independence and Lipschitz Continuity.

Axiom 8 (Boundedness). *For all $i, j \in A$, p is such that there exists a $\kappa_{ij} > 0$ such that if $\frac{\theta_i^t}{\theta_j^t} < \delta$ for some $\delta > 0$ then $\frac{p_i}{p_j} < \kappa_{ij}\delta$.*

In other words, Boundedness states that if the relative propensities of two actions are bound by some number $\delta \in R$, then relative choice probabilities should be bound by a finite (but possibly arbitrarily large) multiple of δ. Hence, the requirement is that the decision maker does not ever fully discard an action given current propensities. However, she may assign an arbitrarily small probability to some actions. Hence, loosely speaking, the requirement is to be at least "a bit cautious".

Our first result states that the probability of choosing each action negatively depends on the propensities of the other actions. Given Monotonicity and Anonymity, the result is due to the fact that the choice probabilities must lie in the n-dimensional unit simplex.

Lemma 3. *If p satisfies Anonymity and Strict Monotonicity then for all $i \in A$ we have that p_i is strictly decreasing in θ_k, $\forall k \neq i$.*

Proof. Since $\sum p_i = 1$ and since p_k is strictly increasing in θ_k, at least some $p_j \neq p_k$ has to be strictly decreasing in θ_k. But then by Anonymity this has to be true for all p_j. $\square$

Next we show that if an action has a greater propensity than another action, then the probability of choosing it must also be greater. This result is again a consequence of Monotonicity and Anonymity.

Lemma 4. *If p satisfies Monotonicity and Anonymity then $\theta_i^t > \theta_j^t$ implies $p_i^t > p_j^t$.*

Proof. The proof is by contradiction. Assume that $\theta_i^t > \theta_j^t$ but $p_i(\theta_i^t, \theta_j^t, \theta^t) < p_j^t(\theta_i^t, \theta_j^t, \theta^t)$, where the first argument is the value of θ for action i, the second argument for action j and the third argument the value of θ for all other actions. Now by Strict Monotonicity and Lemma 3 we have that $p_j^t(\theta_i^t, \theta_j^t, \theta^t) < p_j^t(\theta_j^t, \theta_j^t, \theta^t)$. Furthermore (by Strict Monotonicity) we have $p_i^t(\theta_i^t, \theta_j^t, \theta^t) > p_i^t(\theta_j^t, \theta_j^t, \theta^t)$. This implies

$$p_i(\theta_j^t, \theta_j^t, \theta^t) < p_i(\theta_i^t, \theta_j^t, \theta^t) < p_j(\theta_i^t, \theta_j^t, \theta^t) < p_j(\theta_j^t, \theta_j^t, \theta^t) \tag{9}$$

violating Anonymity. $\square$

We are now able to provide a characterization of the choice probabilities p. The class of rules satisfying Anonymity, Monotonicity and Continuity is quite large. The Boundedness axiom, however, turns out to induce a unique characterization. Proposition 4 below states that the probability of choosing each action is linearly proportional to its propensity.

Proposition 4. *A choice rule p satisfies axioms 5-8 if and only if*

$$p_i = \frac{\theta_i}{\sum_j \theta_j}.$$

Proof. First we show necessity. It is obvious that $p_i = \frac{\theta_i}{\sum_j \theta_j}$ satisfies Strict Monotonicity and Anonymity. For Boundedness note that $\frac{p_i^t}{p_j^t} = \frac{\theta_i^t}{\theta_j^t} = \kappa\delta$, $\forall \kappa > 1$. Next we show sufficiency. Note that Cautious Choice implies that $\forall i, j : \frac{p_i^t}{p_j^t} \leq \kappa \frac{\theta_i^t}{\theta_j^t}$ for some $\kappa' \in \mathbb{R}$. (If $\exists i, j$ s.t.

$\frac{p_i^t}{p_j^t} \geq \kappa' \frac{\theta_i^t}{\theta_j^t}, \forall \kappa' > 0$, then assume that $\frac{\theta_i^t}{\theta_j^t} = \frac{\delta\kappa}{\kappa'} < \delta$. This implies $\frac{p_i^t}{p_j^t} \geq \kappa' \frac{\theta_i^t}{\theta_j^t} = \delta\kappa$, a contradiction). Next note the self consistency condition

$$p_i^t \leq \kappa' \left(\frac{\theta_i^t}{\theta_j^t} \right) p_j^t \leq \kappa' \left(\frac{\theta_i^t}{\theta_j^t} \right) \kappa' \left(\frac{\theta_j^t}{\theta_i^t} \right) p_i^t = \left(\kappa'\right)^2 p_i^t \tag{10}$$

implies that $\kappa' \geq 1$. Now consider the smallest $\kappa' \geq 1$ for which $\frac{p_i^t}{p_j^t} \leq \kappa' \frac{\theta_i^t}{\theta_j^t}$ holds true $\forall i, j \in A$ and denote it by $\kappa^{\min}$. Next note that this is equivalent to

$$\forall i, j : p_i^t \theta_j^t \leq \kappa^{\min} \theta_i^t p_j^t. \tag{11}$$

Since this is true $\forall i, j$, summing both sides over j delivers

$$p_i^t \sum_j \theta_j^t \quad \leq \quad \kappa^{\min} \theta_i^t \sum_j p_j^t \Longleftrightarrow \tag{12}$$

$$p_i^t \quad \leq \quad \frac{\kappa^{\min} \theta_i^t}{\sum_j \theta_j^t}, \tag{13}$$

since $\sum_i p_i^t = 1$. But now if $\kappa^{\min} > 1$, then $\exists \kappa'' \in [1, \kappa^{\min}]$ and $i \in A$ s.t. $p_i^t > \kappa'' \frac{\theta_i^t}{\sum_j \theta_j^t} > \frac{\theta_i^t}{\sum_j \theta_j^t}$. This also implies (since $\sum_i p_i = 1$) that there exists $j \neq i$ s.t. $p_j^t < \frac{\theta_j^t}{\sum_h \theta_h^t}$. Fix all $\theta_j^t, \forall j \neq i$. Now since $p_i(\theta)$ is continuous and mapping Θ into $[0, 1]$ then (by Anonymity and monotonicity) it follows that $\exists \hat{\theta}$ s.t. $\forall \theta_i > \hat{\theta} : p_i > \frac{\kappa'' \theta_i}{\sum_j \theta_j}$. [6] But then we have that

$$\forall \theta_i^t > \frac{1}{\kappa'' - 1} \sum_{j \neq i} \theta_j^t : p_i^t > 1. \tag{14}$$

what contradicts p being a probability. Hence we need $\kappa^{\min} = 1$. But then $\sum_i p_i = 1$ implies $p_i^t = \frac{\theta_i^t}{\sum_i \theta_j^t}$. □

4.1 Discussion on the Axioms of p

4.1.1 Continuity

What happens if we do not require the continuity axiom, i.e. if we allow the agent to respond to small changes in propensities with "big" changes in choice probabilities? Without Continuity, another class of rules is possible, such as, for example, the following: Denote by $\theta_{\max}$ the maximum of $\{\theta_1, \ldots, \theta_n\}$ and define p_i as follows:

$$p_i = \begin{cases} 0.9 & \text{if } \theta_{\max} \text{is unique and } \theta_i = \theta_{\max} \\ \frac{(0.1)\theta_i}{\sum_{\theta_j \neq \theta_{max}} \theta_j} & \text{if } \theta_{\max} \text{is unique but } \theta_i \neq \theta_{\max} \\ \frac{\theta_i}{\sum_j \theta_j} & \text{if } \theta_{\max} \text{is not unique} \end{cases}$$

[6] Assume there existed an interval $[\widetilde{\theta} - x, \widetilde{\theta} + x]$ s.t. $\forall \widetilde{\theta} + x > \theta_j > \widetilde{\theta} : p_j < \frac{\theta_j}{\sum \theta_h}$ and $\forall \widetilde{\theta} - x < \theta_i < \widetilde{\theta} : p_i > \frac{\theta_i}{\sum \theta_h}$, then $\lim_{\theta_i, \theta_j \to \widetilde{\theta}} \frac{p_j}{p_i} < \lim_{\theta_i, \theta_j \to \widetilde{\theta}} \frac{\theta_j \left(\sum_{h \neq i} \theta_h + \theta_i\right)}{\theta_i (\sum_{h \neq j} \theta_h + \theta_j)} = 1$ which contradicts Strict Monotonicity (using Lemma 4). Note also that the limit exists by continuity

The rule above satisfies Strict Monotonicity and Anonymity and also the Boundedness axiom since if $\theta_i = \theta_{\max}$, then

$$\frac{p_i}{p_j} = \frac{0.9 \sum_{\theta_h \neq \theta_{max}} \theta_h}{(0.1)\, \theta_j} < \frac{(0.9)\,(|A|-1)\,\theta_i}{(0.1)\,\theta_j} < 9\,(|A|-1)\,\delta$$

whenever $\frac{\theta_i}{\theta_j} < \delta$ and hence the axiom holds for $\kappa = 9\,(|A|-1)$. (If $\theta_i \neq \theta_{\max}$, then the axiom holds for $\kappa = 1$).

4.1.2 Boundedness

Essentially, the Boundedness axiom rules out choice rules where "too little" exploration is performed. The simple choice rule where the action with the highest propensity is chosen with probability one or the exponential choice rule are two examples of rules where the decision maker converges to a single action quickly without exploring the environment. Obviously, without the Boundedness axiom the class of admissible choice rules is massive.

One may wonder why we do not assume Lipschitz continuity instead of the Boundedness axiom. Next we look at the relationship between the Boundedness axiom and Lipschitz continuity. The first thing to note is that Lipschitz continuity does not imply the Boundedness axiom. To see this consider the following counterexample where $p_i = \frac{\theta_i - \frac{1}{2}\sum_j \theta_j}{\frac{2}{3}\sum_j \theta_j}$. This function is Lipschitz continuous for $\kappa = 3$ (since $(p_i - p_j) = \frac{3}{2\sum_j \theta_j}(\theta_i - \theta_j)$), but it does not satisfy the Boundedness axiom. Note that this rule is also Anonymous and Monotone.

The example presented in subsection 4.1.1 has already shown that it is also not the case that Boundedness implies Lipschitz continuity. Furthermore, Boundedness and Continuity are also not enough to imply Lipschitz continuity. Consider a the following axiom

Axiom' 3 (Independence). *If* $\frac{\theta_i^t}{\theta_j^t} = \frac{\theta_i^{t+1}}{\theta_j^{t+1}}$, *then also* $\frac{p_i^t}{p_j^t} = \frac{p_i^{t+1}}{p_j^{t+1}}$.

Clearly this axiom is weaker than Boundedness and is, for example, satisfied by the exponential choice rule (in conjunction with the other axioms). What about, though, if we require both Independence and Lipschitz continuity? The following proposition shows that taken together these axioms are stronger than the Boundedness axiom.

Proposition 5. *Any choice rule satisfying Independence and Lipschitz continuity also satisfies Boundedness.*

Proof. Note first that independence requires that $\frac{p_i}{p_j}$ is not a function of θ_k for any $k \neq i, j$. Furthermore we know by Rademacher's theorem that every Lipschitz continuous function is almost everywhere differentiable. Now taking the partial derivative of $\frac{p_i}{p_j}$ with respect to θ_k for some $k \neq j, i$ and requiring it to be zero we get the following equation.

$$\frac{\frac{\partial p_i}{\partial \theta_k}}{\frac{\partial p_j}{\partial \theta_k}} = \frac{p_i}{p_j}, \forall k \neq j, i. \tag{15}$$

Thus, all we need to show is that there is a constant bounding the RHS of equation (15). But we also know that $\|\nabla p_i(\theta)\|$ is bound by the Lipschitz constant L. Hence $\sqrt{\sum_j \left(\frac{\partial p_i}{\partial \theta_j}\right)^2} \leq L$ which implies that every partial derivative must be bound by a constant and hence also the RHS of equation (15). Now assume that the derivative $\frac{\partial p_i}{\partial \theta_j}$ fails to exist on some set $C \subset \Theta$. It should be easy to see that the above arguments still apply. □

Hence, Boundedness is implied by Independence and Lipschitz continuity taken together while the reverse is not true. Note that Lipschitz continuity also implies Continuity, but even if we take Continuity and Boundedness together they do not imply Lipschitz continuity. Our example in subsection 4.1.1 above violated Continuity. Next we present an example of a rule which satisfies Boundedness and Continuity but violates Lipschitz continuity. Note that any such example must violate either Anonymity or Strict Monotonicity since the only rule satisfying all four of these is the rule from Proposition 4 which is Lipschitz continuous. Consider the following example, where Anonymity is violated,

$$\begin{aligned} p_1 &= \tfrac{2\theta_1}{\theta_1 + \sum_j \theta_j} \\ p_i &= \tfrac{\theta_i}{\theta_1 + \sum_j \theta_j} \quad \text{if } i \neq 1 \end{aligned}$$

Clearly this rule satisfies Continuity and Boundedness. To see that it violates Lipschitz continuity, note that

$$|p_1 - p_j| = \frac{1}{\theta_1 + \sum_j \theta_j} |2\theta_1 - \theta_j| .$$

Hence, we have seen that requiring Lipschitz continuity and Independence together is a stronger requirement than Continuity and Boundedness. As a matter of fact, Continuity and Boundedness together do not even imply Lipschitz continuity. This should have convinced the reader that our Boundedness axiom is relatively weak.[7]

[7] One could also consider a stronger version of Lipschitz continuity which is sometimes called Bi-Lipschitz. A function is Bi-Lipschitz if there exists κ such that $\forall i, j : \frac{1}{\kappa} |\theta_i - \theta_j| \leq |p_i - p_j| \leq \kappa |\theta_i - \theta_j|$. This could be interpreted as saying that choice probabilities should not be too far nor too close together. The condition again is not implied by Boundedness. The proportional choice rule for example violates it since $|p_i - p_j| = \frac{|\theta_i - \theta_j|}{\sum_j \theta_j}$, but no κ can be found which bounds $\sum_j \theta_j$.

5 Efficiency and Optimality

5.1 Efficiency

As we mentioned in the introduction, our target is not to axiomatize learning procedures that "work". Instead, our aim is to provide a characterization of learning rules that satisfy certain natural axioms. A question that arises once such characterization is carried out is then: do the procedures that satisfy natural axioms work? This is the target to be studied in this subsection.

We say that a transition function T together with a choice function p is efficient if it selects the action with highest average payoff in the long run. Let $E(\pi_i)$ be the expected payoff of action i at any random period. We have then the following definition:

Definition 1. *A pair of transition function together with a choice function, (T, p), is efficient if, for $k = \arg\max_j E(\pi_j)$,*

$$\lim_{t \to \infty} p_k^t = 1.$$

Our next result shows that indeed the natural axioms we placed do make the decision maker to select the efficient action in the long run. Proposition 6 below is a consequence of our characterization and Rustichini's (1998) result on linear procedures without counterfactuals[8].

Proposition 6. *A pair (T, p) that satisfies axioms 1-4 and 5-8 is efficient.*

Proof. Define $k = \arg\max_i E(\pi_i)$. Rustichini (1998, Proposition 3.2) shows that if

$$p_i^t(R) = \frac{\theta_i^0 + \pi_i(h^t)}{\sum_j \left(\theta_j^0 + \pi_j(h^t)\right)} \tag{16}$$

for all $i \in A$ then in the limit when t grows large and independently on θ_i^0 and h^t:

$$\lim_{t \to \infty} p_k = 1.$$

Using propositions 1 and 4 we have that if axioms 1-8 are satisfied then

$$p_i^t = \frac{\theta_i^0 + \lambda \pi_i(h^t) + \#i(h^t)\rho}{\sum_j \left(\theta_j^0 + \pi_j(h^t) + \#j(h^t)\rho\right)}$$

for some $\lambda > 0$ and $\rho \geq 0$.

[8] Rustichini (1998) refers to the situation where counterfactuals are not observed as the partial information case.

Consider now a different environment where we multiply times λ all the payoffs of all actions and add ρ to them. It is clear that in this new environment propensities and payoffs still belong to $\mathbb{R}_+$, so our characterization (propositions 1 and 4) can be applied. It is also easy to see that for the new environment, $\arg\max_i \hat{E}(\pi_i) = \arg\max_i E(\pi_i)$, where $\hat{E}$ denotes the expected payoff in the modified environment.

Thus, Rustichini's result implies that (16) puts probability 1 to action k in the limit in this new environment. That is,

$$\begin{aligned}\lim_{t\to\infty} p_k^t(R) &= \lim_{t\to\infty} \frac{\theta_i^0 + \hat{\pi}_i\left(h^t\right)}{\sum_j \left(\theta_j^0 + \hat{\pi}_j\left(h^t\right)\right)} \\ &= 1\end{aligned}$$

where $\hat{\pi}$ denotes the payoff in the modified environment.

However, we have that

$$\lim_{t\to\infty} \frac{\theta_i^0 + \hat{\pi}_i\left(h^t\right)}{\sum_j \left(\theta_j^0 + \hat{\pi}_j\left(h^t\right)\right)} = \lim_{t\to\infty} \frac{\theta_i^0 + \lambda\pi_i\left(h^t\right) + \#i(h^t)\rho}{\sum_j \left(\theta_j^0 + \pi_j\left(h^t\right) + \#j(h^t)\rho\right)}.$$

Therefore,

$$\begin{aligned}\lim_{t\to\infty} p_k^t &= \lim_{t\to\infty} \frac{\theta_i^0 + \lambda\pi_i\left(h^t\right) + \#i(h^t)\rho}{\sum_j \left(\theta_j^0 + \pi_j\left(h^t\right) + \#j(h^t)\rho\right)} \\ &= 1.\end{aligned}$$

That is, the learning procedure that satisfies axioms 1-8 chooses action k in the limit in the original environment. Since k is the efficient action in the original environment, the result follows. □

5.2 Optimality

There is a sense in which the Boundedness axiom could be interpreted as the agent being cautious in that she small differences in propensities do not translate into large differences in choice. A more standard interpretation of being cautious may be that the choice functions should be Lipschitz continuous as mentioned above. We now take up this definition of being 'cautious' and show that the proportional choice rule is optimal among all 'cautious' (i.e. Lipschitz continuous) rules. This result will give us clearer interpretation of what Boundedness adds over Lipschitz continuity. Consider first the following definition.

Definition 2. *We say choice rule p is more cautious than choice rule p' if $\forall i, j and \theta_i \neq \theta_j$: $\kappa(\theta_i, \theta_j) \leq \kappa'(\theta_i, \theta_j)$ where $\kappa_{ij} := \kappa(\theta_i, \theta_j) := \frac{p_i - p_j}{\theta_i - \theta_j}$.*

There are several things worth noticing about this definition: First, unlike the Boundedness axiom (and in the definition of Lipschitz Differentiability) κ can depend on θ and is allowed to be infinite. Second, we can set $\kappa_{ii} = 1$ for simplicity, which will not affect any of our results. We also need $\kappa_{ij} = \kappa_{ji}$ (which follows immediately from self-consistency).

Remark *The proportional choice rule $p_i = \frac{\theta_i}{\sum_j \theta_j}$ can be interpreted as the least cautious rule, which is Lipschitz continuous and satisfies Anonymity and Strict Monotonicity.*

To understand the remark above note that (for strictly monotonous choice rules) $\sum_i (p_i^t - p_j^t) = \kappa \sum_i \theta_i^t - \theta_j^t$ is maximal if $\theta_j^t = 0$ and hence $\sum_i (p_i^t - p_j^t) = \kappa \sum_i \theta_i^t \leq 1$ implying that $\kappa \leq \left(\sum_i \theta_i^t\right)^{-1}$ which is exactly the parameter from the proportional choice rule. Hence, this suggests that we could replace the Boundedness axiom above with Lipschitz continuity and then single out from the remaining set of possible rules the ones which are "least cautious" according to the definition above.

Next we look at optimality properties to see how being cautious matters in terms of the probability of choosing the efficient action. Denote by i^* the efficient action (and assume for now that it is unique) and by π_i^e the expected payoff of action i at any given period. For simplicity, consider choice rules that are differentiable everywhere. Then we can state the following result:

Proposition 7. *Consider choice rules where $\frac{\partial \kappa_{ij}(t)}{\partial t} \leq 0$ and assume that at $t = 0$ all actions are chosen with the same probability. Then, there exists $\underline{t}$ such that for all $t > \underline{t} : p_{i^*}^t \geq p_{i^*}'^t$ whenever p is less cautious than p'.*

Proof. First note that

$$\begin{aligned}\kappa_{ij}(\theta_i^t - \theta_j^t) &= \kappa_{ij}\left(\theta_i^{t-1} + p_i^{t-1}\lambda\pi_i^e - \theta_j^{t-1} - p_j^{t-1}\lambda\pi_j^e\right)\\ &= \lambda\kappa_{ij}\left(p_i^{t-1}\pi_i^e - p_j^{t-1}\pi_j^e\right) + \kappa_{ij}\frac{1}{\kappa_{ij}}\left(p_i^{t-1} - p_j^{t-1}\right).\end{aligned}$$

We assume for now that there are only two actions.[9] Then using this expression we can write

$$\begin{aligned}\langle p_i^t - p_i^{t-1}\rangle &= \kappa_{ij}(\theta_i^t - \theta_j^t) + (1 - p_i^t) - p_i^{t-1}\\ &= \lambda\kappa_{ij}\left(p_i^{t-1}\pi_i^e - p_j^{t-1}\pi_j^e\right) + \left(p_i^{t-1} - p_j^{t-1}\right) + (1 - p_i^t) - p_i^{t-1}\\ &= \lambda\kappa_{ij}\left(p_i^{t-1}\pi_i^e - (1 - p_i^{t-1})\pi_j^e\right) + \left(p_i^{t-1} - (1 - p_i^{t-1})\right) + (1 - p_i^t) - p_i^{t-1}\\ &= -\lambda\kappa_{ij}\pi_j^e + p_i^{t-1}\left(\lambda\kappa_{ij}\left(\pi_i^e + \pi_j^e\right)\right) + p_i^{t-1} - p_i^t\end{aligned}$$

[9] Note that we have assumed for simplicity that $\widetilde{\pi} = 0$.

Taking the continuous time limit we can write

$$\begin{aligned} \dot{p}_i &= \lambda(p_i \kappa_{ij} \pi_i^e - (1-p_i)\kappa_{ij}\pi_i^e) && (17) \\ &= -\lambda(\kappa_{ij}\pi_j^e + p_i \kappa_{ij} \left(\pi_i^e + \pi_j^e\right)). && (18) \end{aligned}$$

By noting that

$$\int \kappa_{ij} \left(\pi_i^e + \pi_j^e\right) dt = \left(\pi_i^e + \pi_j^e\right) \int \kappa_{ij} dt$$

And by denoting $\int -\kappa_{ij}\pi_j^e \exp - \left(\left(\pi_i^e + \pi_j^e\right) \int \kappa_{ij} dt \right) dt =: A(t)$ we can write the solution to (17) as

$$p_i(t) = (A(t) + c) \exp \left(\left(\pi_i^e + \pi_j^e\right) \int \kappa_{ij} dt \right)$$

where c stands for a constant that depends on the initial condition $p_i(0)$. This equation is a good approximation of actual behavior whenever (i) $\frac{\partial \kappa_{ij}(t)}{\partial t} = \left(\frac{\partial \kappa_{ij}}{\partial \theta_i} \frac{\partial \theta_i}{\partial t} + \frac{\partial \kappa_{ij}}{\partial \theta_j} \frac{\partial \theta_j}{\partial t} \right) \leq 0$ (decreasing step sizes) and (ii) whenever t is "sufficiently" large. Note that condition (i) is satisfied whenever the associated choice rule is Lipschitz continuous.

Furthermore, note that those functions $\kappa_{ij}(\theta_{i,}\theta_j)$ satisfying $\frac{\partial \kappa_{ij}(t)}{\partial t} \leq 0$ do have a supremum in Θ since $\frac{\partial \theta_i}{\partial t} \geq 0, \forall i$.

We are interested in how $p_i(t)$ varies with κ_{ij}. Since we say a rule is more cautious than another rule whenever $\kappa_{ij} < \kappa'_{ij}, \forall i, j$ and $\forall \theta_i \neq \theta_j$ we can replace each function κ_{ij} by its supremum $\sup \kappa_{ij} := \kappa_s$. If we do this, then

$$p_i(t) = \frac{\lambda \pi_j^e}{\pi_i^e + \pi_j^e} + c \exp\left(\lambda \left(\pi_i^e + \pi_j^e\right) t \kappa_s\right)$$

Setting $t = 0$ we obtain $c = p_i(0) - \frac{\lambda \pi_j^e}{\pi_i^e + \pi_j^e}$ and hence

$$p_i(t) = \begin{pmatrix} p_i(0) \exp\left(\lambda \left(\pi_i^e + \pi_j^e\right) t\kappa_s\right) \\ + \frac{\lambda \pi_j^e}{\pi_i^e + \pi_j^e} \left(1 - \exp\left(\lambda \left(\pi_i^e + \pi_j^e\right) t\kappa_s\right)\right) \end{pmatrix}.$$

The derivative

$$\begin{aligned} \frac{\partial p_i(t)}{\partial \kappa_s} &= t\lambda \left(p_i(0)\pi_i^e - \pi_j^e(1 - p_i(0)\right) \exp\left(\lambda \left(\pi_i^e + \pi_j^e\right) t\kappa_s\right) > 0 \\ \iff & \pi_i^e > \pi_j^e \frac{1 - p_i(0)}{p_i(0)}. \end{aligned}$$

Hence the probability to choose i increases with κ if and only if action i is optimal, i.e. yields higher payoffs in expectation and if initial conditions are not too biased against it. It should also be clear that the proof goes through (with heavier notation) if there are more than two actions, but if i is uniquely optimal. □

If i is optimal and initial conditions are not too biased against it, then being less cautious (higher κ) is always better among Lipschitz continuous rules. This is intuitive, because if initially the optimal action is chosen with very small probability, then it is optimal to be more cautious, i.e. to explore also actions which have initially low propensities (and hence probabilities) more. If initial propensities towards all actions are approximately the same, then it is best to be least cautious within the bounds of Lipschitz continuity.

Note that our optimality result is quite revealing. We already know from section 5.1 that the updating and choice rule derived from our main axioms are efficient in the sense that they yield the optimal action in the long run. What the previous proposition shows is that the the rule from axioms 1-8 yields higher expected payoffs than any other Lipschitz continuous rule after some finite time[10].

Proposition 7 also implies that the behavior of the decision maker under axioms 1-8 also has a justification in terms of optimality among Lipschitz continuous rules: as we have seen, it is the least cautious rule which satisfies Anonymity and Strict Monotonicity.

Axioms 5-8 seem to yield the optimal level of 'cautiousness' in choice. What if we do not restrict to Lipschitz continuous rules? Is it still better to be less cautious? The answers is negative. Note that if we do not restrict to Lipschitz continuous rules, the least cautious rule is always the rule that chooses the action with the highest propensity with probability one. Clearly this cannot be optimal, since it implies choosing the action that initially has the highest propensity forever.[11]

It should also be made clear that no non-anonymous rules may do better (unless the decision maker has some ex ante knowledge about which action is best, which is something we rule out). By the same token, there can also be no rule which is anonymous but not monotonic, since this (together with our updating rule) would imply that for some range of θ higher payoffs lead to lower probabilities which is not optimal. Overall, hence, axioms 5-8 lead to a rule that yields the optimal level of caution among Lipschitz continuous rules.

6 Relation to Other Procedures and Further Discussion

6.1 Replicator Dynamics

Adapting the results from Hopkins (2002) or Börgers and Sarin (1996) it is easy to show that the behavior resulting from axioms 1-8 approximates the Replicator Dynamics in the long

[10] More precisely, it will attach a higher probability to the action with higher expected value.

[11] Even with an updating rule that allows for decreasing propensities, though, this choice rule does generally not lead to the optimal action as has been shown by Sarin and Vahid (1999).

run. Thus, in this respect we provide an axiomatization to the replicator dynamics.

Proposition 8. *A pair (T, p) that satisfies axioms 1-4 and 5-8 creates a sequence of choices that can be approximated, in the sense of stochastic approximation, by the replicator dynamics.*

Proof. First note that we can write down the expected change in choice frequencies as follows.

$$\langle p_i^{t+1} - p_i^t \rangle = p_i^t \frac{\theta_i^t + \pi_i^t}{\sum_i \theta_j^t + \pi_i^t} + (1 - p_i^t) \frac{\theta_i}{\sum_i \theta_j^t + \pi_j^t} - \frac{\theta_i^t}{\sum_i \theta_j^t} \tag{19}$$

$$= P_i^t \frac{\theta_i^t + \pi_i^t - \theta_i^t \left(1 + \frac{\pi_i^t}{\sum \theta_j^t}\right)}{\sum_j \theta_j^t + \pi_j^t} + (1 - p_i) \frac{\theta_i^t - \theta_i^t \left(1 + \frac{\pi_j^t}{\sum \theta_j^t}\right)}{\sum_j \theta_j^t + \pi_j^t}. \tag{20}$$

$$= p_i^t (1 - p_i^t) \pi_i^t - (1 - p_i^t) p_i^t \pi_j^t + O(\sum \theta_j)^{-2}. \tag{21}$$

Taking the continuous time limit of the last equation and neglecting the term of order $(\sum \theta_j)^{-2}$ we get the evolutionary Replicator Dynamics. We are allowed this since by Proposition 1 all θ_i are strictly increasing and hence the property of decreasing step sizes is satisfied. (See e.g. Hopkins, 2002). □

6.2 Separation between Updating and Choice Rule

In a sense the separation between updating and choice rule is a classical separation in Economics. In standard decision theory, options are evaluated by assigning probabilities to different events. These probabilities then imply some (expected) levels of utility of the different options. The choice rule then simply prescribes to choose the option with the highest evaluation (assigned utility value). However, if there is not enough information about states, outcomes etc. then typically it will not be optimal to choose the option with the highest valuation with probability one. This is the case since one should explore the state space and learn about other options.

All classical learning rules do have this separation in a more or less explicit manner as well. In (stochastic) fictitious play, players update their beliefs about the choices of others, which translates into an update about the expected profitability of actions via an updating rule, and then choose the action they assign the highest expected payoff to. Similarly, a Bayesian learner will update her beliefs about the world using Bayes rule as an updating rule and then choose whichever option seems best given the updated beliefs. This separation seems implicit in any learning rule.

If the decision maker observes information about counterfactuals after each choice, then her decisions become irrelevant for the learning process because her information at the end of a period is independent of the action she chose. Hence, she can simply choose the action for which she has the highest propensity (preference). As mentioned several times, this is the main difference between our approach and the approach of Easley and Rustichini (1999).

If agents do not observe counterfactuals, such a separation is needed. One may suggest a different approach where one *imposes* that the action with the highest propensity is chosen with probability one and tries to include the exploitation/exploration trade-off in the way propensities are formed. Such an approach will fail in the following sense: either propensities will have to be non-monotonous in payoffs, in which case clearly convergence to the optimal action will not occur, or an agent can get stuck with a suboptimal action if all actions tried previously have led to lower payoffs (See e.g. Sarin and Vahid, 1999). I our view, these two failures require a quite unreasonable decision-maker.

6.3 Further Comparison with Literature

Let us compare our axioms with those of Easley and Rustichini (1999). Their most significant axioms are Symmetry, Monotonicity, Independence and Exchangeability. Symmetry and Monotonicity are similar to our axioms of Anonymity and Monotonicity. Weak Sum is comparable to Exchangeability but somewhat weaker. Exchangeability requires that past and current payoffs have the same effects. Weak Sum requires that everything else equal (i.e. given two histories of the same length where the action in question was chosen equally often) the action should have a higher propensity if the sum of payoffs under one history is larger than under the other. Clearly, this implies that past and current payoffs have the same effect. Furthermore, note that Proposition 3 shows that Weak Sum can be weakened without fundamentally altering the results.

6.4 Discussion - Allowing for negative propensities

What if some propensities could be negative? This might lead to conceptual mistakes as it is not necessarily true that agents treat negative propensities the same as positive propensities. In order to circumvent this a possible approach is to normalize propensities according to some function. Note that this would require extra assumptions on the normalization function. Now we suggest a possible normalization.

For any t we order all θ_i^t from smallest to largest and denote by $\theta_{(k)}^t$ the $k-$th smallest θ_i^t. Then we can define the following normalization recursively.

$$\begin{aligned} \theta_{(1)}'^t &= 0 \\ \theta_{(2)}'^t &= \left|\theta_{(1)}^t - \theta_{(2)}^t\right| \\ \theta_{(3)}'^t &= \theta_{(2)}'^t + \left|\theta_{(2)}^t - \theta_{(3)}^t\right| = \theta_{(3)}^t - \theta_{(1)}^t ... \end{aligned} \tag{22}$$

Note that this normalization is the only possible normalization which (i) respects cardinal differences in propensities and (ii) is minimally distortive in the sense that the sum of changes made to all the θ_i is minimal. However, the normalization above is arbitrary, this is why we chose to deal with positive payoffs only.

7 Conclusions

We have presented a model where a decision maker, oblivious of the environment she lives in, learns about the payoff of the alternative options by own experience when counterfactuals are not observed. The reasoning process of the decision maker was separated into two parts: First, she has propensities over action. These represent her preferences for the different alternatives and are based on her past experiences. Second, the decision maker then translates these propensities into choice, the source of her experiences.

We established natural axioms in the way propensities are updated and the way propensities are translated into choice and characterized the behavior of the decision maker. Furthermore, we considered alternatives to our main axioms and studied the efficiency and optimality of the learning procedures resulting from our axioms. Finally, we related our results to known leading rules in the literature providing, for instance, an axiomatization of the replicator dynamics.

This paper targeted covering the gap in the literature whereby learning without counterfactuals had only been studied from the optimality point of view. We posed natural axioms and characterized behavior. The approach we followed had only been used so far in situations where counterfactuals are observed.

References

[1] Anscombe, F.J. and R. Aumann (1963), A definition of subjective probability, *Annals of Mathematical Statistics.*

[2] Bergemann, D. and J. Vaelimaeki (2006), Bandit Problems, Cowles Foundation Discussion paper 1551.

[3] Börgers, T., A. Morales and R. Sarin (2004): "Expedient and Monotone Learning Rules", *Econometrica* 72 (2), 383-405.

[4] Bush, R.R. and F. Mosteller (1951): A mathematical model for simple learning, Psychological Review, 58, 313-323.

[5] Camerer, T. and Ho, T-H. (1999): “Experience Weighted Attraction Learning in Normal Form Games”, *Econometrica* 67, 827-874.

[6] Easley, D. and A. Rustichini (1999): “Choice without Beliefs”, *Econometrica* 67 (5), 1157-1184.

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

[8] Geneakoplos, J. and H. Polemarchakis (1982), ”We Can’t Disagree Forever”, Journal of Economic Theory.

[9] Hopkins, E. (2002): “Two Competing Models on how People Learn in Games”, *Econometrica* 70 (6), 2141-2166.

[10] Karandikar, R., D. Mokerjee, D. Ray and F. Vega Redondo (1998): ”Evolving Aspirations and Cooperation”, *Journal of Economic Theory* 80, 292–331.

[11] Roth, A. 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.

[12] Rustichini, A. (1999), ”Optimal Properties of Stimulus-Response Models”, *Games and Economic Behavior....*

[13] Sarin, R. and F. Vahid (1999): “Payoff Assessment without Probabilities: A Simple Dynamic Model of Choice”, *Games and Economic Behavior* 28, 294-309.

[14] Savage, L.J. (1954). The Foundations of Statistics. New York, NY: Wiley. Sawyer, A. G.

[15] Sutton, R. and A. Barto (1998): Reinforcement Learning, MIT Press.

[16] Wilson, R. (1968), The Theory of Syndicates, Econometrica, Vol. 36(1), 119-132.

8 Appendix

8.1 Proofs and Extra Results on Propensities

8.1.1 Proof of Proposition 3

First note that Lemma 2 is still valid under Weak Weighted Average. Thus, we can find a function $g_i : \Theta_i \times \Pi$ that represents T_i for all $i \in A$.

Take any two payoffs $\hat{\pi}, \bar{\pi} \in \Pi$ and consider the following two histories:

$$\begin{aligned} h^2 &= \left(\left(i, \hat{\pi}\right), \left(i, \delta\bar{\pi}\right)\right), \\ \hat{h}^2 &= \left(\left(i, \bar{\pi}\right), \left(i, \delta\hat{\pi}\right)\right). \end{aligned}$$

By Weak Weighted Sum we have that $\theta^2 = \hat{\theta}^2$. Thus, using the definition of g we have that for any $\theta_i \in \Theta_i$ and any $\hat{\pi}, \bar{\pi} \in \Pi$

$$g_i\left(g_i\left(\theta_i, \hat{\pi}\right), \delta\bar{\pi}\right) = g_i\left(g_i\left(\theta_i, \bar{\pi}\right), \delta\hat{\pi}\right). \tag{23}$$

Proceeding as in the proof of Proposition 1 we have to

$$\left.\frac{\partial g_i}{\partial \theta}\right|_{\left(g_i\left(\theta_i, \hat{\pi}\right), \bar{\pi}\right)} = \delta.$$

Since this is true for any $\theta_i, \hat{\pi}, \bar{\pi}$, we have that

$$g_i\left(\theta_i, \pi\right) = \delta\theta_i + f(\pi) \tag{24}$$

for some strictly increasing (Axiom 2) and everywhere differentiable (Axiom 3) function $f : \Pi \to \Theta$.

Proceeding again as in the proof of Proposition 1 we can show that

$$f\left(\pi\right) = \lambda\pi + \rho. \tag{25}$$

for some λ and ρ. By Monotonicity we have that $\lambda > 0$, and by the fact that propensities are defined in $\mathbb{R}_{+}$, $\rho \geq 0$. Combining (24), (25) and Lemma 1 gives the desired result.

8.1.2 A Variation of the Weak Weighted Average Axiom

Consider the following variation of the Weak Weighted axiom:

Axiom' 4 (Weak Weighted Average 2). *For any histories h^t and $\bar{h}^k$ and any action i, T is such that if $\theta_i^0 = \bar{\theta}_i^0$ and $\#i\left(h^t\right) = \#i\left(\bar{h}^k\right)$, then $\theta_i^t \geq \bar{\theta}_i^k$ if and only if $\pi_i\left(\delta, h^t\right) \geq \pi_i\left(\delta, \bar{h}^k\right)$.*

The only difference between the Weak Weighted Average axiom and the alternative above lies in the fact that under the original Weak Weighted Average axioms the two histories that are compared must have the same length. A consequence of this is that under Weak Weighted Average' a payoff obtained at the same point in time is treated differently when comparing histories of different lengths. That is, assume any histories with different lengths h^t and $\bar{h}^k$ with $k < t$. If under both histories a payoff π is obtained at time $l < k$ then in history $\pi\left(h^t\right)$ the payoff π is discounted by δ^{t-l} while in $\pi\left(\bar{h}^k\right)$ the payoff is discounted by δ^{k-l}.

Proposition 9. *There exists no transition T that satisfies axioms 1-3 and 4'.*

Proof. We proceed by contradiction. Since axiom 1' follows from axiom 4", we can use the proof of Proposition 3 to prove that there exists a $\lambda > 0$ and $\rho \in \mathbb{R}$ such that

$$\theta_i^{t+1} = \begin{cases} \delta\theta_i^t + \lambda\pi + \rho & \text{if action } i \text{ is played and payoff } \pi \text{ is obtained,} \\ \delta\theta_i^t + \rho & \text{otherwise.} \end{cases}$$

Take any $\theta_i^0 \neq \rho/(1-\delta)$ and any $\pi \in \Pi$ and consider two histories h^2 and $\hat{h}^1$ such that

$$\begin{aligned} h^2 &= \left((i,\pi), \left(-i,\pi'\right)\right), \\ \hat{h}^1 &= \left((i,\delta\pi)\right). \end{aligned}$$

By Weak Weighted Average axiom 2 we have that $\theta_i^2 = \hat{\theta}^1$. Thus, using the result from Proposition 3 we have that

$$\delta^2\theta_i^0 + \delta\pi + 2\rho = \delta\theta_i^0 + \delta\pi + \rho.$$

Therefore,

$$\rho = (1-\delta)\,\theta_i^0.$$

Which contradicts the fact that $\theta_i^0 \neq \rho/(1-\delta)$. The key argument is that ρ and δ are exogenous constants and their value cannot depend on θ_i^0. □

8.1.3 Risk Averse Sum Axiom

Proposition 10. *There exists no transition T that satisfies axioms 1-3 and 2'.*

Proof. Proceeding as in the proof of Proposition 1 we can show that there must exist a constant λ such that

$$\left.\frac{\partial g_i}{\partial \pi}\right|_{(\theta_i,\pi)} = \lambda \tag{26}$$

for all θ_i, π. Furthermore, we can show that

$$g_i(\theta_i,\pi) = \theta_i + f(\pi) \tag{27}$$

for some strictly increasing (Axiom 2) and everywhere differentiable (Axiom 3) function $f : \Pi \to \Theta$.

Take again any $\hat{\pi}, \bar{\pi}$. By Risk Averse Sum we have that

$$f\left(\hat{\pi}\right)+f\left(\bar{\pi}\right) \quad > \quad f\left(0\right)+f\left(\hat{\pi}+\bar{\pi}\right).$$

Differentiating both sides with respect to $\hat{\pi}$ leads to

$$\left.\frac{\partial f}{\partial \pi}\right|_{(\hat{\pi})} \quad > \quad \left.\frac{\partial f}{\partial \pi}\right|_{(\hat{\pi}+\bar{\pi})}.$$

However, this contradicts (26). Thus, no transition exists satisfying axioms 1, 3 and 2'. □

8.2 Proofs and Extra Results on Choice Functions

8.2.1 More on Boundedness and Lipschitz continuity

The Boundedness axiom with $\kappa \leq 1$ does imply Lipschitz continuity with parameter κ. To see this note that

$$\frac{p_i^t}{p_j^t}-\frac{p_j^t}{p_i^t}<\kappa\frac{\theta_i^t}{\theta_j^t}-\frac{p_j^t}{p_i^t}<\kappa\frac{\theta_i^t}{\theta_j^t}-\frac{\theta_j^t}{\kappa\theta_i^t}$$

since by the Boundedness axiom there exists a $\kappa : \frac{p_j^t}{p_i^t} < \kappa\frac{\theta_j}{\theta_i}$ and $\frac{p_j^t}{p_i^t} > \frac{1}{\kappa}\frac{\theta_j}{\theta_i}$ for some κ.. But then

$$\begin{aligned}
\frac{\left(p_i^t\right)^2-\left(p_j^t\right)^2}{p_i^t p_j^t} &< \frac{\kappa^2\left(\theta_i^t\right)^2-\left(\theta_j^t\right)^2}{\kappa\theta_i^t\theta_j^t} \Leftrightarrow \\
\frac{\left(p_i^t-p_j^t\right)\left(p_i^t+p_j^t\right)}{p_i^t p_j^t} &< \frac{\left(\kappa\theta_i^t-\theta_j^t\right)\left(\kappa\theta_i^t+\theta_j^t\right)}{\kappa\theta_i^t\theta_j^t} \Longleftrightarrow \\
\left(p_i^t-p_j^t\right) &< \left(\kappa\theta_i^t-\theta_j^t\right)\underbrace{\frac{\left(\kappa\theta_i^t+\theta_j^t\right)p_i^t p_j^t}{\kappa\theta_i^t\theta_j^t\left(p_i^t+p_j^t\right)}}_{<1 \text{ if } \theta_i^t\geq 1, \forall i.}
\end{aligned}$$

But then

$$\left(p_i^t-p_j^t\right)<\left(\kappa\theta_i^t-\theta_j^t\right)<\kappa\left(\theta_i^t-\theta_j^t\right)\Leftrightarrow\kappa\leq 1.$$

Now together with Anonymity we have that

$$p_i\left(\theta_i,\theta_j,\theta\right)-p_j\left(\theta_j,\theta_i,\theta\right)=p_i\left(\theta_i,\theta_j,\theta\right)-p_i\left(\theta_j,\theta_i,\theta\right).$$

implying Lipschitz continuity.