DEPARTMENT OF ECONOMICS

Probability Matching and Reinforcement Learning*

Javier Rivas, University of Leicester, UK

Working Paper No. 11/20
March 2011

Probability Matching and Reinforcement Learning*

Javier Rivas
University of Leicester†

March 2, 2011

Abstract

Probability matching occurs when an action is chosen with a frequency equivalent to the probability of that action being the best choice. This sub-optimal behavior has been reported repeatedly by psychologist and experimental economist. We provide an evolutionary foundation for this phenomenon by showing that learning by reinforcement can lead to probability matching and, if learning occurs sufficiently slowly, probability matching does not only occur in choice frequencies but also in choice probabilities. Our results are completed by proving that there exists no quasi-linear reinforcement learning specification such that behavior is optimal for all environments where counterfactuals are observed.

JEL Classification Number: C73.

Keywords: Probability Matching, Reinforcement Learning.

*I would like to thank Karl Schlag, Larry Samuelson, Mark Le Quement and anonymous referees for useful comments and the audiences at the University of Alicante, Universitat Autónoma de Barcelona, University of Leicester, European University Institute and University of Wisconsin-Madison.

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

1 Introduction

Consider an urn with 60 black balls and 40 white balls. If we were to predict the color of the ball in five draws with replacement, it would be optimal to guess black five times. However, psychologist and experimental economist have reported that in such a situation individuals tend to guess black three times and white two times. Three out of five represents a 60% frequency and two out of five represent a 40% frequency. That is, the frequencies of agents' responses match the frequencies of the balls in the urn. This is what is known as *probability matching*.

Probability matching has been reported repeatedly by psychologist and experimental economist. For example, Rubinstein (2002) conducts several experiments similar to that of the example above and finds that probability matching is present in between 30% and 80% of the population depending on the specifics of the problem. Similar results are obtained by experimental psychologist Siegel and Goldstein (1959) and Gaissmaier and Schooler (2008) among others. Probability matching seems to be an innate characteristic of behavior not only present in the human specie. This phenomenon is also found, for instance, in fish (Behrend and Bitterman (1961)) and pigeons (Bullock and Bitterman (1962)).

In this paper we show that in environments where the payoff of not chosen actions is observed the choices made by an individual who learns by reinforcement, i.e. actions that were successful in the past are more likely to be chosen, converge to probability matching. On top of that, we show that if learning speed is sufficiently slow then this convergency does not only occur in the *frequency* with which each action is chosen but also in the *probability* with which each action is chosen at any given point. This suggests that the probability matching behavior exhibited by some subjects can be explained as the result of reinforcement learning. We also find that this sub-optimality property of reinforcement learning is robust, meaning that it is not possible to design an specification of quasi-linear reinforcement learning such that behavior is optimal for all environments where counterfactuals are observed.

According to reinforcement learning, actions that were more successful in the past are more likely to be adopted in the future. Reinforcement learning has been found to be one of the main driving forces of human behavior in decision problems. For some detailed expositions on reinforcement learning and its relationship with real life behavior the reader is referred to Roth and Erev (1995), Erev and Roth (1998) and Camerer and Ho (1999).

To our knowledge, our paper is the first one to explicitly obtain a direct link between reinforcement learning and probability matching. Nevertheless, there have been previous articles suggesting the possibility of a relationship between these two phenomena. Simon (1959) points out to the fact that under a specific class of reinforcement learning models

the frequency of choices converges to the frequencies with which each of these choices is the best alternative. Apart from the fact that we consider a much general specification of reinforcement learning, what we show in this paper is that not just the frequency converges, but that under some conditions the probability of choosing each action at any given point in time also converges. To understand this difference consider a situation where only two actions exists. Imagine two choice patterns: one such that each action is chosen alternatively and a different one whereby each action is chosen with a 50% probability. In this case, these two different behaviors give rise to the same observed choice frequencies even though the probability of choosing each action at any point in time differs across the two choice patterns.

Other relevant papers indicating the possibility of a relationship between reinforcement learning and probability matchings are those of Börgers and Sarin (2000) and Erev and Barron (2005). Börgers and Sarin find that probability matching can arise in a model of reinforcement learning if agents have aspiration levels. By further exploring the implications of reinforcement learning, we find that probability matching still appears even in the absence of aspiration levels. This suggests that the link between reinforcement learning and probability matching is deeper than initially thought. Erev and Barron (2005) present an experimental exercise where subjects exhibit probability matching behavior and show that reinforcement learning is the behavioral model that better fits the data they observed. Therefore, our theoretical exercise is supported by their findings.

A fact worth mentioning is that no relationship between reinforcement learning and probability matching occurs in environments where the decision maker has no information about counterfactuals (payoffs of not chosen actions). To our knowledge, this fact was first observed by Rustichini (1998), who showed that in environments where there is no information about counterfactuals linear reinforcement learning results in the decision maker choosing with probability one the action that is best in the long run.

2 The Model

2.1 Environment

Consider a decision maker who every period $t = 0, 1, \ldots$ has to choose an action from the finite set $A = \{1, \ldots, n\}$. The payoff of the decision maker at time t depends on her action and on the state of nature $s^t \in S = \{1, \ldots, m\}$ at time t unknown to the decision maker. If the decision maker chooses action i and the state of nature equals s then her payoff equals π_{is}. To simplify the exposition, we assume that $\pi_{is} \in [0, 1]$ and for any state s the payoff

maximizing action is unique. Define π_s as the vector of payoffs of each action in state s, $\pi_s = (\pi_{1s}, \ldots, \pi_{ns})$.

The sequence of states of nature $\{s^t\}_t$ follows an independent and identically distributed process where $p_s \in [0,1]$ is the probability of each state s occurring at any given t with $\sum_s p_s = 1$[1]. An environment is defined by the payoff vectors together with the probabilities of each state occurring: $\{(\pi_1, \ldots, \pi_m), (p_1, \ldots, p_s)\}$.

Let $\sigma_i^t \in [0,1]$ denote the probability with which the decision maker chooses action i at time t with $\sum_i \sigma_i^t = 1$ for all t. We assume σ_i^0 is given for all i and lies between $(0,1)$ so that all actions have positive initial probability of being chosen. Finally, define $\sigma_i = \{\sigma_i^t\}_t$.

The timing within each time period t goes as follows: First, the decision maker chooses an action according to σ_i^t for all action i. Second, nature decides the state. Third, payoffs are realized and the decision maker observes the payoff of all actions[2]. Finally, the decision maker updates the probability of choosing each action σ_i^{t+1} for all i.

2.2 The Learning Rule

The type of reinforcement learning we consider is such that the next period's likelihood of choosing a given action is quasi-linear in the current likelihood of choosing that action and the payoff each action yielded. This implementation of reinforcement learning is a generalization of the linear reinforcement learning models pioneered by psychologists Bush and Mosteller (1951) and applied to economics first by Simon (1959) and Cross (1973).

By ways of reinforcement, the decision maker increases the probability of choosing the action that yielded a higher payoff in the previous period. As argued above, we focus our attention on a generalization of the most widely used implementation of reinforcement learning, whereby the increase in the probability of choosing a given action next period is quasi-linear in the probability of choosing that action in the current period and the payoff of each action. Let $\sigma_i^{t+1}(s)$ be the value of σ_i^{t+1} if at period t the state of nature is s. We have the following definition:

Definition 1. *The quasi-linear reinforcement learning rule is given for any $s \in S$ by*

$$\sigma_i^{t+1}(s) = \begin{cases} \sigma_i^t + \left(1 - \sigma_i^t\right) \mu f(\pi_s) & \text{if } \pi_{is} > \pi_{js} \forall j \in A, \\ \sigma_i^t - \sigma_i^t \mu f(\pi_s) & \text{otherwise} \end{cases} \tag{1}$$

with $\mu \in (0,1]$ and $f : [0,1]^n \to (0,1]$.

[1] In Rivas (2008) we show that when states of nature follow a Markov process results presented here are still valid.

[2] For the case where foregone payoffs are not observed see Rivas (2008).

The function f above can be seen as the intensity or strength of the reinforcement whilst the parameter μ is interpreted as the learning speed. The reason why μ is not included in f will be clear later on. The function f, as imposed by reinforcement, is weakly increasing in the payoff of the action that yielded the highest payoff and weakly decreasing in the payoff of all the other actions. That is, f is weakly increasing (decreasing) in π_{is} if and only if $\pi_{is} > (<)\pi_{-is}$.

As an example, consider the case where payoffs enter exponentially in f in the following form:

$$f(\pi_s) = \frac{\max_i\{e^{\pi_{is}}\}}{\sum_j e^{\pi_{js}}}.$$

Hence, if we define $k(s) \in A$ as $k(s) = \arg\max_i \pi_{is}$, equation (1) becomes

$$\sigma_i^{t+1}(s) = \begin{cases} \sigma_i^t + \left(1 - \sigma_i^t\right) \mu \frac{e^{\pi_{k(s)s}}}{\sum_j e^{\pi_{js}}} & \text{if } i = k(s), \\ \sigma_i^t - \sigma_i^t \mu \frac{e^{\pi_{k(s)s}}}{\sum_j e^{\pi_{js}}} & \text{otherwise.} \end{cases}$$

Another possible specification of f includes the case where $f(\pi_s) = 1$ for all π_s. In this situation, the resulting learning rule is equivalent to what is known in the population games literature as the best response with inertia (see Samuelson (1994) or Kosfeld et al. (2002)).

Previous literature relating reinforcement learning and probability matching assumed that $f(\pi_s) = 1$ and $\mu = 1$. In this case, the frequency with which each action is chosen trivially converges to the frequency with which that action is the best choice. Allowing for a much general specification permits us to better understand the relationship between the two phenomena. In particular, as we shall show, under some circumstances convergence not only occurs in the frequency by which each action is chosen but also in the probability of choosing each action.

3 Results

3.1 Convergence in Frequencies

Our first result is that under reinforcement learning the frequency with which each action is chosen is closely related to the probability of that action being the best choice. This relationship is given by the specific functional form of f used and is independent on the parameter μ. Proposition 1 below states this finding formally.

Proposition 1. *Define E^0 as the expected value operator evaluated at time 0. Furthermore, for all action $i \in A$ define*

$$\bar{\sigma}_i = \frac{\sum_{s:\pi_{is} > \pi_{-is}} p_s f(\pi_s)}{\sum_s p_s f(\pi_s)}.$$

We have that

$$\lim_{t\to\infty} E^0\left(\sigma_i^t\right) = \bar{\sigma}_i.$$

Proof. The result follows directly from applying Breiman's strong law for Markov processes (Breiman (1960)) to the sequence σ_i for all action i. However, in order to improve exposition we show a self-contained proof that uses the well known law of iterated expectations (see, for instance, Ljungqvist and Sarget (2000)).

Applying law of iterated expectations to our model yields

$$E^0\left(\sigma_i^t\right) = E^0 \circ \ldots \circ E^{t-1}\left(\sigma_i^t\right).$$

Simple algebra shows that at any time q

$$\begin{aligned} E^{q-1}\left(\sigma_i^q\right) &= \sigma_i^{q-1} + \left(1-\sigma_i^{q-1}\right)\sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s) - \sigma_i^{q-1}\sum_{s:\pi_{is}<\pi_{-is}} p_s\mu f(\pi_s) \\ &= \sigma_i^{q-1}\left(1-\sum_s p_s\mu f(\pi_s)\right) + \sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s) \\ &= a\sigma_i^{q-1} + b, \end{aligned}$$

with $a = 1-\sum_s p_s\mu f(\pi_s)$ and $b = \sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s)$. Therefore, using the law of iterated expectations we have that

$$E^0\left(\sigma_i^t\right) = (a)^t\sigma_i^0 + b\sum_{r=1}^{t}(a)^{t-r}.$$

Hence, as t grows large and taking into account that $a<1$

$$\begin{aligned} \lim_{t\to\infty} E^0\left(\sigma_i^t\right) &= \lim_{t\to\infty} b\sum_{r=1}^{t}(a)^{t-r} \\ &= \frac{b}{1-a} \\ &= \frac{\sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s)}{1-(1-\sum_s p_s\mu f(\pi_s))} \\ &= \frac{\sum_{s:\pi_{is}>\pi_{-is}} p_s f(\pi_s)}{\sum_s p_s f(\pi_s)} \\ &= \bar{\sigma}_i. \end{aligned}$$

□

That is, the expected probability with which an action is chosen converges to the probability of that action being the best choice, corrected by the specific function f used and

the payoff vectors $(\pi_1, \ldots, \pi_m)$. Previous literature relating reinforcement learning and probability matching assumed $f(\pi_s) = 1$ and $\mu = 1$, which by proposition 1 implies that $\lim_{t\to\infty} E_0\left(\sigma_i^t\right) = \sum_{s:\pi_{is}>\pi_{-is}} p_s$. That is, the frequency with which an action is chosen converges to the probability of that action being the best choice. This is what is known as probability matching.

Proposition 1 generalizes on previous literature by allowing for a more general specification of reinforcement learning. In section 3.3 we study whether or not this general specification is able to select the best action in the long run.

3.2 Convergence in Probabilities

Apart from convergence in frequencies, we find that converge in probabilities is also possible. This means that not only the frequency with which each action is chosen converges but that the probability with which each action is chosen also converges to the probability of that action being a best choice.

The following proposition characterizes the convergence of σ_i for all action i when learning speed μ is arbitrarily small.

Proposition 2. *For any $\varepsilon > 0$ there exists a $\bar{\mu} > 0$ such that for all $\mu < \bar{\mu}$*

$$Pr\left(\lim_{t\to\infty} \left|\sigma_i^t - \bar{\sigma}_i\right| > \varepsilon\right) = 0.$$

Proof. We proceed by showing that for any $\varepsilon > 0$, if $\mu < \varepsilon$ then $\lim_{t\to\infty} E^0\left(\left(\sigma_i^t - E^0\left(\sigma_i^t\right)\right)^2\right) < \varepsilon^2$. That is, σ_i^t converges in 2-nd order mean when learning speed μ converges to zero. In other words, we proceed by showing that for any $\varepsilon > 0$, if $\mu < \varepsilon$ then the variance of σ_i is bounded above by ε^2.

Using the result in proposition 1 it is true that

$$\begin{aligned}
\lim_{t\to\infty} E^0\left(\left(\sigma_i^t - E^0\left(\sigma_i^t\right)\right)^2\right) &= \lim_{t\to\infty} E^0\left(\left(\sigma_i^t\right)^2\right) - \left(\bar{\sigma}_i\right)^2 \\
&= \lim_{t\to\infty} E^0\left(\left(\sigma_i^{t-1}\right)^2 + 2\sigma_i^{t-1}\sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s) - 2\left(\sigma_i^{t-1}\right)^2\sum_s p_s\mu f(\pi_s)\right) \\
&\quad - \left(\bar{\sigma}_i\right)^2 + p^{(2)}(\mu) \\
&= \lim_{t\to\infty} E^0\left(a'\left(\sigma_i^{t-1}\right)^2 + b'\sigma_i^{t-1}\right) - \left(\bar{\sigma}_i\right)^2 + p^{(2)}(\mu) \\
&= \lim_{t\to\infty} a' E^0\left(\left(\sigma_i^{t-1}\right)^2\right) + b'\bar{\sigma}_i - \left(\bar{\sigma}_i\right)^2 + p^{(2)}(\mu)
\end{aligned}$$

where $a' = 1 - 2\sum_s p_s\mu f(\pi_s)$, $b' = 2\sum_{s:\pi_{is}>\pi_{-is}} p_s\mu f(\pi_s)$ and $p^{(2)} : (0,1] \to (0,1]$ is a polynomial whose lowest power is 2. Iterating on the term $E^0\left(\left(\sigma_i^{t-1}\right)^2\right)$ in the equation

above leads to

$$\begin{aligned}
\lim_{t\to\infty} E^0\left(\left(\sigma_i^t - E^0\left(\sigma_i^t\right)\right)^2\right) &= \frac{b'}{1-a'}\bar{\sigma}_i - \left(\bar{\sigma}_i\right)^2 + p^{(2)}\left(\mu\right) \\
&= \frac{\sum_{s:\pi_{is}>\pi_{-is}} p_s \mu f(\pi_s)}{\sum_s p_s \mu f(\pi_s)}\bar{\sigma}_i - \left(\bar{\sigma}_i\right)^2 + p^{(2)}\left(\mu\right) \\
&= \left(\bar{\sigma}_i\right)^2 - \left(\bar{\sigma}_i\right)^2 + p^{(2)}\left(\mu\right) \\
&= p^{(2)}\left(\mu\right) \\
&\leq \mu^2.
\end{aligned}$$

Thus, for any $\varepsilon > 0$ if we choose a learning speed $\mu < \varepsilon$ then

$$\lim_{t\to\infty} E^0\left(\left(\sigma_i^t - E^0\left(\sigma_i^t\right)\right)^2\right) < \varepsilon^2.$$

Therefore, σ_i^t converges in 2-nd order mean when learning speed μ converges to zero. As convergence in r-th order mean with $r > 1$ implies converges in probability, we have that σ_i^t converges in probability to $E^0\left(\sigma_i^t\right)$ when learning speed μ is arbitrarily small. Since by proposition 1 $E^0\left(\sigma_i^t\right) = \bar{\sigma}_i$, the result follows. □

The intuition behind proposition 2 is that as learning speed parameter μ becomes small, the change in the probabilities of choosing each action also becomes small. In the limit this means that the the variance of the stochastic process on σ_i for all i collapses to zero. This fact together with proposition 1 implies than on top of converges in frequencies converge in probabilities also occurs.

If instead of a general function f we consider the case where $f(\pi_s) = 1$ for all π_s, we have the following corollary:

Corollary. *If $f(\pi_s) = 1$ then for any $\varepsilon > 0$ there exists a $\bar{\mu} > 0$ such that for $\mu < \bar{\mu}$*

$$Pr\left(\lim_{t\to\infty}\left|\sigma_i^t - \sum_{s:\pi_{is}>\pi_{-is}} p_s\right| > \varepsilon\right) = 0.$$

That is, the probability of choosing action i converges to the probability with which that action is a best response to the environment. This is a stronger result than that of probability matching: not only there is convergency in frequencies but also in probabilities if learning speed is sufficiently slow.

3.3 Optimality

We continue our analysis by formulating the following question: Is it possible to design an specification of f such that the resulting probability of choosing the action that has the

highest expected payoff converges to 1? For understanding this issue we use the concept of optimality:

Definition 2. *We say that the quasi-linear reinforcement learning rule is optimal if there exists a function f such that for all environment $\{(\pi_1, \ldots, \pi_m), (p_1, \ldots, p_s)\}$ and all $\varepsilon > 0$,*

$$Pr\left(\lim_{t\to\infty} |\sigma_k^t - 1| > \varepsilon\right) = 0$$

with $k = \arg\max_i \sum_{s:\pi_{is}>\pi_{-is}} p_s \pi_{is}$.

A feature of reinforcement learning is that the decision maker can be "distracted" towards non-optimal actions by the random process on the states of nature. This is because non-optimal actions can be the best action for some states of nature. Therefore, randomness can lead the decision maker to increase the probability of choosing a non-optimal action even if she is currently choosing the optimal action with probability one. As a consequence, there are environments such that for any rule the limit of the learning process converges to a situation where non-optimal actions are chosen with some positive probability. This is formally proven in our next proposition.

Proposition 3. *The quasi-linear reinforcement learning rule is not optimal.*

Proof. Assume, without loss of generality, that $k = \arg\max_i \sum_{s:\pi_{is}>\pi_{-is}} p_s \pi_{is}$, so that action k has the highest expected payoff.

The proof goes by contradiction. Assume that for all $\varepsilon > 0$ there exists a function f such that for all the environments $\{(\pi_1, \ldots, \pi_m), (p_1, \ldots, p_s)\}$, $|\bar{\sigma}_k - 1| < \varepsilon$. This can be rewritten as follows: for any sequence $\{\varepsilon_r\}_{r=0}^{\infty}$ converging to 0 with $\varepsilon_r > \varepsilon_{r+1} > 0$ for $r \in \{0, \ldots, \infty\}$ and ε_0 given, we have that there exists an associated sequence of functions $\{f_{\varepsilon_r}\}_{r=0}^{\infty}$ with $f_{\varepsilon_r} : [0,1]^n \to [0,1]$ such that $\bar{\sigma}_k(f_{\varepsilon_r}) < \bar{\sigma}_k(f_{\varepsilon_{r+1}})$ with $r \in \{0, \ldots, \infty\}$ and

$$\lim_{n\to\infty} \bar{\sigma}_k(f_{\varepsilon_r}) = 1,$$

where $\bar{\sigma}_k(f_{\varepsilon_r})$ is the value of $\bar{\sigma}_k$ associated with the function f_{ε_r}.

The limit above holds if and only if

$$\lim_{n\to\infty} \frac{\sum_{s:\pi_{ks}>\pi_{-ks}} p_s f_{\varepsilon_r}(\pi_s)}{\sum_{s:\pi_{ks}<\pi_{-ks}} p_s f_{\varepsilon_r}(\pi_s)} = \infty \tag{2}$$

holds.

Take now an environment $\{(\pi_1, \pi_2), (p_1, p_2)\}$ where $0 < \pi_{11} < \pi_{22}$ and $\pi_{12} = \pi_{21} = 0$. We could consider more general environments but that will only complicate the exposition

leaving the logic of the proof unchanged. The probabilities of each state occurring are such that $\sum_{s=1}^{2} p_s \pi_{1s} > \sum_{s=1}^{2} p_s \pi_{2s}$ with $p_1, p_2 > 0$. In this situation, equation (2) implies that

$$\lim_{n\to\infty} \frac{p_1 f_{\varepsilon_r}(\pi_1)}{(1-p_1) f_{\varepsilon_r}(\pi_2)} = \infty.$$

Since $p_1 \in (0,1)$, the equation above holds if and only if the following limit holds:

$$\lim_{n\to\infty} \frac{f_{\varepsilon_r}(\pi_1)}{f_{\varepsilon_r}(\pi_2)} = \infty. \tag{3}$$

However, given that $\pi_{11} < \pi_{22}$ and $\pi_{12} = \pi_{21} = 0$, we have that $f_{\varepsilon_r}(\pi_1) \leq f_{\varepsilon_r}(\pi_2)$ for all $r > 0$, a contradiction. □

The logic behind the proof is that if the quasi-linear reinforcement learning rule is optimal then there exists a function f that magnifies the payoffs of each action. This can be seen in equation (3), where the finite difference in payoffs is magnified to infinity. However, if this is the case, an environment can be found such that there is a rare state of nature for which the payoff of the suboptimal action is greater than that of the optimal action. In such environment, f cannot result in the decision maker choosing the optimal action with probability one in the long run.

4 Conclusions

We studied the relationship between probability matching and reinforcement learning in an environment where counterfactuals are observed and found that the two phenomena are significantly related. In particular, under a general class of reinforcement learning rules that we called the quasi-linear reinforcement learning rule, the expected probability with which an action is chosen converges to the probability of that action being the best choice, corrected by the specific learning rule used and the payoff vectors. Moreover, if the decision maker's learning speed is sufficiently slow then convergence not only occurs in the frequency with which each action is chosen but also in its probability. We concluded our results by showing that this sub-optimality property of reinforcement learning is robust, meaning that it is not possible to design an specification of quasi-linear reinforcement learning such that behavior is optimal for all environments.

References

1. Börgers, T. and Sarin, R. (2000): "Naive Reinforcement Learning with Endogenous Aspirations". *International Economic Review* 41, 921-950.

2. Breiman, L. (1960): "The Strong Law of Large Numbers for a Class of Markov Chains" *Annals of Mathematical Statistics* 31 (3), 801-803.

3. Bush, R. R. and Mosteller, F. (1951): "A Mathematical Model for Simple Learning". *Psychological Review* 58, 313-323.

4. Camerer, C. and Ho, T. H. (1999): "Experienced-Weighted Attraction Learning in Normal Form Games". *Econometrica* 67 (4), 827-874.

5. Cross, J. G. (1973): "A Stochastic Learning Model of Economic Behavior". *The Quarterly Journal of Economics* 87 (2), 239-266.

6. Easley, D and Rustichini, A. (1999): "Choice Without Beliefs", *Econometrica* 67 (5), 1157-1184.

7. Erev, I. and Barron, G. (2005): "On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies". *Psychological Review* 112 (4), 912931.

8. Erev, I. and Roth, A. (1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria". *The American Economic Review* 88 (4), 848-881.

9. Gaissmaier, W. and Schooler, L. J. (2008):"The Smart Potential Behind Probability Matching". *Cognition* 109, 416422.

10. Josephson, J. (2008): "Stochastic Better-Reply Dynamics in Finite Games". *Economic Theory* 35, 381389.

11. Kimura, M. and Ihara, Y. (2009): "Replicator Dynamics Models of Sexual Conflict". *Journal of Theoretical Biology* 260, 9097.

12. Kosfeld, M., Droste, E. and Voorneveld, M. (2002): "A Myopic Adjustment Leading to Best Reply Matching". *Games and Economic Behavior* 40, 270-298.

13. Ljungqvist, L. and Sarget, T (2000): "Recursive Macroeconomic Theory". *The MIT Press.*

14. Rubinstein, A. (2002): "Irrational Diversification in Multiple Decision Problems". *European Economic Review* 46, 1369-1378.

15. Rivas, J. (2008): "Learning within a Markovian Environment". *EUI Working paper 2008/13.*

16. Rustichini, A. (1998): "Optimal Properties of StimulusResponse Learning Models". *Games and Economic Behavior* 29 (1-2), 244-273.

17. Samuelson, L. (1994): "Stochastic Stability in Games with Alternative Best Replies". *Journal of Economic Theory* 64 (1), 35-65.

18. Schuster, P. and Sigmund, K. (1983): "Replicator dynamics". *Journal of Theoretical Biology* 100 (3), 533-538.

19. Siegel, S. and Goldstein, D. A. (1959): "Decision Making Behavior in a Two-Choice Uncertain Outcome Situation". *Journal of Experimental Psychology* 57 (1), 37-42.