European University Institute

Department of Economics

EUI Working Papers

ECO 2008/13

Learning within a Markovian Environment

Javier Rivas

EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

Learning within a Markovian Environment

JAVIER RIVAS

EUI Working Paper **ECO** 2008/13

ISSN 1725-6704

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy

http://www.eui.eu/
http://cadmus.eui.eu/

Learning within a Markovian Environment

Javier Rivas*
European University Institute†

February 7, 2008

Abstract

We investigate learning in a setting where each period a population has to choose between two actions and the payoff of each action is unknown by the players. The population learns according to reinforcement and the environment is non-stationary, meaning that there is correlation between the payoff of each action today and the payoff of each action in the past. We show that when players observe realized and foregone payoffs, a suboptimal mixed strategy is selected. On the other hand, when players only observe realized payoffs, a unique action, which is optimal if actions perform different enough, is selected in the long run. When looking for efficient reinforcement learning rules, we find that it is optimal to disregard the information from foregone payoffs and to learn as if only realized payoffs were observed.

JEL Classification Number: C73.

Keywords: Adaptive Learning, Markov Chains, Non-stationarity, Reinforcement Learning.

*Part of this work was written during my visit to the University of Wisconsin-Madison. I'm grateful to the faculty at UW and the participants in the Theory Lunch. I would like to thank Karl Schlag and Larry Samuelson for useful discussions and comments. I would also like to thank Mark Le Quement for useful comments and the seminar audiences at the University of Alicante and at the European University Institute.

†javier.rivas@eui.eu. Department of Economics, European University Institute. Via della Piazzuola 43, 50133 Florence (Italy). www.eui.eu/Personal/Researchers/JavierRivas.

1 Introduction

Imagine the simple decision problem in which every period individuals in a population have to choose between two alternatives. The payoff of these two alternatives is not know by the players. What is more, the payoff of the alternatives could vary over time according to some distribution also unknown for the players.

This decision problem is faced by many of us in our everyday lives: whether to buy a PC or a Mac, whether to have fruit or a cake as a dessert in a restaurant, or whether to watch an action movie or a romantic movie at the theater. Although oblivious of the payoff we will get from making these choices, we might have some information that can help in choosing the better alternative. This information could have been obtained, for instance, from our own experiences in the past or via word-of-mouth communication.

In this paper we study how the choices made by a population evolve in the setting just described. The model we present has two major features about how players learn and about how the payoffs change. First, players learn according to reinforcement, whereby actions that where successful in the past are more likely to be chosen. Second, the underlying distribution determining the payoff of each action is non-stationary. This means that the payoff today of a given action depends on the payoff it yielded in the past. In particular, we consider the case in which payoffs depend deterministically on the state of nature. The state of nature changes following a Markov chain. Hence, the probability of being at a given state tomorrow depends on which state we are in today. Players are ignorant of this fact; they simply observe that the payoff of available actions changes over time.

In the learning literature, as well as in the economic literature in general, randomness determining the outcome of certain events or actions is almost always assumed to follow a stationary i.i.d. process. This assumption is clearly made for the sake of technical simplicity, as real life phenomena, such as financial markets, gambling, population biology, statistical mechanics, etc., quite often follow non-stationary processes. To our knowledge, only Ben-Porath et al. (1993) and Rustichini (1999) deal with the evolutionary properties of models where nature follows a non-stationary process.

Ben-Porath et al. (1993) present an evolutionary model that is framed within a changing environment. They study two types of environments: one in which the change is deterministic and another in which the changes in environment follow a Markov chain. In their model, players' actions are subject to random mutations. They characterize the mutation rate that maximizes population growth in the long run.

Rustichini (1999) presents a paper that focuses on the optimality of two different population dynamics within a Markovian environment. In his model, the environment changes

according to a Markov chain, and for any state in the chain there is a unique action that maximizes payoff. Rustichini (1999) studies the optimality properties of linear and exponential (logit) adjustment process when players have infinite memory. An adjustment process or learning rule is simply a map between information and strategies. Rustichini (1999) considers two different informational settings about payoffs of actions. In one of these settings players observe the performance of all the actions (realized and foregone payoffs are observed), while in the other they only observe the performance of the action chosen (only realized payoffs are observed).

As in Rustichini (1999), we consider two informational settings: one in which both realized and foregone payoffs are observed and another in which only realized payoffs are observed. There are two main differences between Rustichini's work and ours. First, we consider a very general set of learning rules instead of only two specific rules. Second, and most importantly, in our setting players don't use the whole history of past payoff realizations. Instead, as prescribed by reinforcement, players learn using the information they have from their most recent payoff experiences. The reason why we are interested in a setting where players have limited memory is that empirical and theoretical literature in psychology and economics agrees that limited memory is a better assumption for modeling human behavior than infinite memory (see for example, Rubinstein (1998), Hirshleifer and Welch (2002) and Conlisk (1996)).

As already mentioned, the learning rules considered in this paper have the property of being reinforcing. According to reinforcement learning, actions that were more successful today are more likely to be adapted for tomorrow. Reinforcement has been found to be one of the main driving forces of human behavior in repeated 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).

When both realized and foregone payoffs are observed, reinforcement is translated into being more likely to play tomorrow the action that was better today. For this setting, we use a generalization of the best response behavior that we call the Stochastic Better Response. Under the Stochastic Better Response, the probability of playing tomorrow a given action increases if and only if today that action was better than the other one. The magnitude of the change in probabilities of playing either action depends on the specific functional form used. The Stochastic Better Response is a very general learning rule that allows players to respond to the magnitude and not just the ordering of payoffs of each action. Note that the Stochastic Better Response is a different concept from the Stochastic Better Reply Dynamics (Josephson (2007)). The Stochastic Better Reply Dynamics are the dynamics for the evolution of strategies resulting when players use the better response, which is a particular

case of the Stochastic Better Response.

When foregone payoffs are not observed, players can not directly compare the performance of both actions within the same time period. In this case, players reinforce (possibly negatively) the action they played. How much they reinforce this action will depend on the payoff achieved. We use a general case of the Cross (1973) learning rule that also generalizes the rules in Börgers et. al. (2004) (BMS, henceforth). We call this rule the General Reinforcement Rule. Note that players could use the General Reinforcement Rule even if they observe foregone payoffs. While this implies that players are disregarding information, we will show that it may be optimal to do so.

Under the Cross Learning Rule, players increase the probability of playing the action just played by the payoff yielded by that action. An interesting result shown by Börgers and Sarin (1997) is that a population that plays according the Cross Learning Rule exhibits a behavior that converges to replicator dynamics.

The rules in BMS can incorporate aspiration levels (exogenous or endogenous): in other words, if the payoff of the action chosen is higher than the aspiration level, then the probability of playing that action increases for the next period. On the other hand, if the payoff achieved by the action chosen is smaller than the aspiration level, then the probability of playing that action decreases for next period. The rules in BMS are linear on payoffs. We relax this by allowing for any increasing function on realized payoff.

In the case where foregone payoffs are observed, we show that the continuous time limit of the evolution of strategies converges to a situation where every period every action is played with a constant probability bounded away from 1. The specific value of the probability by which each action is played at every period will depend on two things: first, the difference in payoffs between the two actions and the specific form of Stochastic Better Response used, and, second, on the probabilities that the limiting distribution of the Markov chain for states puts on each state. The behavior found in this setting is a generalization to what is know as probability matching. Under probability matching, if an action is best a fraction x of the time, then in any given period it is played with probability x. The best reply matching behavior is clearly suboptimal. While some experimental papers report that this behavior is observed in real life (see, for example, Rubinstein (2002), Siegel and Goldstein (1959)), there does not seem to be consensus as to whether probability matching is in fact present in the behavior of real life agents (see, for instance, Vulkan (2000) and Shanks et. al. (2002)).

The results found in this informational setting are also closely related to the findings by Kosfeld et. al. (2002). They study a setting where a finite set of players repeatedly play a normal-form game. Players adapt their strategies by increasing the probability of playing a

certain action only if this action is a best reply to the actions played by the other agents. Hence, the rule they use is a particular case of the Stochastic Better Response in which the magnitude of payoffs is irrelevant for the updating of strategies. Our setting is also different from theirs in that players do not play against other players but against nature and in that we consider a general class of rules instead of only one. Kosfeld et. al. (2002) find that the continuous time limit of the system converges to a best-reply matching equilibrium. In a best-reply matching equilibrium each player plays an action with a probability that is equal to the probability that this action is a best response to the actions of the other players. The probability matching behavior found in this paper for games against nature is the equivalent to the best-reply matching equilibrium found in Kosfeld et. al. (2002). In Section 5.1 this issue is discussed in more depth.

In our second informational setting, when foregone payoffs are not observed, we show that the population may end up playing a suboptimal action. The population surely selects the action that has higher average payoff only if the difference between the average payoff of the two actions is high enough. Hence, the system may lock-on to a suboptimal action. In this respect, our work extends Ellison and Fudenberg (1995) results to a general set of learning rules and an environment that may not be stationary.

Our results are rounded off by characterizing the efficient rules for both informational settings. A striking result is that when foregone payoffs are observed, it is optimal to ignore the extra information conveyed by the payoff of the action not chosen. That is, players are better off by learning using the General Reinforcement Rule, which only uses the information of the realized payoff. This is due to the fact that observing foregone payoffs leads players to adopt the action that is best today but may be not the best in the long run. That is, players are "distracted" by observing the performance of all the actions. When foregone payoffs are not observed, we show that if players use learning rules that diminish the magnitude of payoffs, that is, that have very cautious and show slow learning, then the population learns the optimal action. These results from are in contrast to those of Rustichini's (1999). In Rustichini (1999), when the population uses the exponential rule (fast learning) the best action is selected only in situations where foregone payoffs are observed, whereas if populations uses the linear rule (slow learning) best action is selected only in situations where foregone payoffs are not observed. Here, instead, we find that under reinforcement learning it is optimal to disregard foregone payoffs and to exhibit slow learning in both informational settings.

This paper's contribution to the literature is twofold. Our first contribution to the literature is the introduction new techniques for dealing with correlated states of nature. As mentioned, very few papers have studied the situation in which the future realization of the state of nature depends on its past realizations. Most papers on learning consider either that

the environment does not change or that it changes independently of past realizations. This is due to the technical difficulties involved in dealing with correlated realizations of states. In this paper we show how these difficulties can, at least partially, be overcome. The proofs for the result for the Stochastic Better Response demonstrate how dependent randomness can be dealt with by showing that for any possible realization of states of nature, the position of the system in the future can be approximated by the differences in speed of convergence towards each action.

The proof of the result for the case where foregone payoffs are not observed extends Ellison and Fudenberg's (1995) result to the case where the distribution of payoffs is not stationary. We show that the behavior of a system that evolves according to a Markov Chain can be approximated by the behavior of a system in which the probability of each state occurring is independent and equal to the limiting distribution of the Markov Chain.

Our second contribution is the extension of the knowledge about stimulus response learning models and evolutionary models. The differences in the behavior of the population under the two informational settings are very intriguing and of interesting application for real life situations. For instance, why can inferior technologies come to dominate the market? A well known example is that when the video format VHS took over from the superior format Betamax. The model can explain that if the two technologies are not too different in terms of performance, the stochastic evolution of nature can lead the population to lock on the suboptimal choice forever. In the example with video formats, during the first months after the release of both technologies, Betamax tapes could not hold an entire movie. This caused the population to slowly adopt the VHS format. Once the true potential of Betamax was revealed, it was too late, consumers had already locked on the inferior technology.

The rest of the paper is organized as follows. Section 2 presents the model. The two informational settings are introduced in Section 3. Results are developed in Section 4. Section 5 presents a discussion and a deeper comparison of this work with the existing literature. Finally, Section 6 concludes.

2 The Model

Consider a continuum of identical players of measure 1. Every period $t = 0, 1, \ldots$ players in the population have to choose between action 1 or action 2. The payoff of each player at time t depends on her action and on the current state of nature $s^t \in \{1, \ldots, m\}$. If a player chooses action i and the state equals j then she gets a payoff $\pi_{ij} \in [0, 1]$ with $i \in \{1, 2\}$ and $j \in \{1, \ldots, m\}$. Note that the payoff of each player does not depend on the actions played by others but only on her own action and the state of nature. We assume there is

no weakly dominant action. That is, there exists no $i \in \{1,2\}$ such that $\pi_{ij} \geq \pi_{-ij}$ for all $j \in \{1,\ldots,m\}$. Without loss of generality we assume that for some $h < m$, $\pi_{1j} \geq \pi_{2j}$ for $j \leq h$ and $\pi_{2j} > \pi_{1j}$ for $j > h$. That is, in the first h states action 1 yields at least the same payoff as action 2. In the remaining states, action 2 yields more payoff than action 1. Finally, we define π_j as the vector of payoffs of action 1 and action 2 in state j, $\pi_j = (\pi_{1j}, \pi_{2j})$.

The sequence of states of nature $\{s^t\}_{t=0}^{\infty}$ follows a discrete Markov process P with $m \geq 2$ states. The probability of transiting from state i to state j is given by $\theta_{ij} \in [0,1]$. We assume the Markov chain to be irreducible and aperiodic. Hence, if $\theta_{ij} = 0$ for some $i, j \in \{1,\ldots,m\}$ then there exists a sequences of states $k_1, k_2, \ldots, k_n \in \{1,\ldots,m\}$ with $n \leq m$ such that $\theta_{ik_1}, \theta_{k_1,k_2}, \ldots, \theta_{k_n,j} \neq 0$. We define $\lambda \in [0,1]^m$ as the limiting distribution of the Markov chain P where λ_i is the weight the limit distribution puts in state i. An environment is defined then by the payoff vectors together with a transition matrix, $\{(\pi_1,\ldots,\pi_m), P\}$.

A strategy is the probability of playing each action at a given period. We denote by $\sigma_i^t \in [0,1]$ with $i \in \{1,2\}$ and $t \in \{0,1,\ldots\}$ the probability of playing action i at time t. Define $\sigma^* = (\sigma_1^*, \sigma_2^*) \in [0,1]^2$ as the strategy that maximizes payoff in the long run. Formally, for any $(\bar{\sigma}_1, \bar{\sigma}_2) \in [0,1]^2$ we have that

$$\sum_{j=1}^{m} \lambda_j \left(\sigma_1^* \pi_{1j} + \sigma_2^* \pi_{2j}\right) \geq \sum_{j=1}^{m} \lambda_j \left(\bar{\sigma}_1 \pi_{1j} + \bar{\sigma}_2 \pi_{2j}\right).$$

Since we are dealing with a continuum of population, Law of Large Numbers applies and we have that σ_i^t is also the fraction of players playing action i at time t. In an abuse of notation, throughout the paper we will refer to σ_i^t as both the probability for a single player of playing action i at time t and the fraction of the population playing action i at time t.

Note that given our setting, the sequence $\sigma_i = \{\sigma_i^t\}_{t=0}^{\infty}$ is an irreducible and aperiodic Markov process on $[0,1]$ for $i \in \{1,2\}$. The aim of the paper is to characterize, if it exists, the invariant distribution of such process.

The timing within each time period works as follows. First, players choose actions according to their strategies. Then, nature decides the state. Third, payoffs are realized and players observe their payoff and possibly forgone payoffs. The possibility of observing foregone payoffs depends on the informational setting being considered. Finally, players update their strategies.

When updating their strategies, players use the following information: their strategy at the beginning of the period, the action they played and the payoff they got and possibly the payoff the other action would have yielded (foregone payoffs). Formally, a learning rule is a function $b : [0,1]^2 \times \{1,2\}^2 \times [0,1]^2 \to [0,1]^2$. That is, a function that maps three arguments, strategies for the present period, action played and payoff gotten and action not played and

foregone payoff, into the strategies for the following period. The functional form of b will depend on the specific learning rule under consideration.

3 Informational Settings

3.1 Forgone Payoffs are Observed

When both realized and foregone payoffs are observed, players best respond to the environment by increasing the probability of playing at the next period the action that was most successful at the present period. We use a generalization of the best response behavior that we call the Stochastic Better Response.

We write $\sigma_i^{t+1}|_j$ to denote the value of σ_i^{t+1} given that at period t the state of nature, s^t, was j. The Stochastic Better Response is defined by

$$\sigma_1^{t+1}|_j = \begin{cases} \sigma_1^t + \sigma_2^t \mu f(\pi_j) & \text{if } \pi_{1j} \geq \pi_{2j} \\ \sigma_1^t - \sigma_1^t \mu f(\pi_j) & \text{otherwise,} \end{cases}$$

where $\mu > 0$ is a learning speed parameter. The function $f : [0,1]^2 \to [0,1]$ maps the payoff of the action that yielded higher payoff and the payoff of the other action into a number between 0 and 1. This function is interpreted as the probability of adopting or learning the action that was best given today's state of nature. The only requirement on f is that it must be weakly increasing in the payoff of the action that yielded higher payoff and weakly decreasing in the payoff of the other action. That is, f is weakly increasing (decreasing) in π_{ij} only if $\pi_{ij} > (<)\pi_{-ij}$. We set $f(\pi_j) = 0$ if and only if $\pi_{1j} = \pi_{2j}$. In other words, we assume that the population does not change strategies if and only if both actions yielded the same payoff. The function f could also be a constant. In the case where the function f is constant and equals 1, the learning rule is equivalent to the standard best response in which players show inertia with probability $1 - \mu$ (as in Samuelson (1994) and Kosfeld et. al. (2002)).

The intuition behind the Stochastic Better Response is the following. In each period, all players observe the payoff of the action chosen and the payoff of the other action. Then every player updates her strategy in the following way. The probability of playing action i in the next period is increased if and only if action i yielded higher payoff than the other action in the current period. The increase in the probability of playing action i will depend on the difference in payoffs between the two actions.

A different interpretation of this same rule uses the fact that σ_i can be considered as the fraction of population playing action i deterministically. Under this interpretation, at every period, players that did not play the best action will change their actions (best response to

the environment) with some probability. The probability of changing action depends on the difference in payoff between the two actions. The Stochastic Better Response is an individual learning rule because actions played by other players have no effect on the updating of the one's own strategy.

As an example, we can look at two possible ways of writing the Stochastic Better Response. In the first one below, payoffs enter exponentially in the function f.

$$\sigma_1^{t+1}|_j = \begin{cases} \sigma_1^t + \sigma_2^t \mu \frac{e^{\pi_{1j}} - e^{\pi_{2j}}}{e^{\pi_{1j}} + e^{\pi_{2j}}} & \text{if } \pi_{1j} \geq \pi_{2j} \\ \sigma_1^t - \sigma_1^t \mu \frac{e^{\pi_{2j}} - e^{\pi_{1j}}}{e^{\pi_{1j}} + e^{\pi_{2j}}} & \text{otherwise} \end{cases} \tag{1}$$

A second example could be the following, where only the payoff of the best action at the current period enters in f and f is linear.

$$\sigma_1^{t+1}|_j = \begin{cases} \sigma_1^t + \sigma_2^t \mu \pi_{1j} & \text{if } \pi_{1j} \geq \pi_{2j} \\ \sigma_1^t - \sigma_1^t \mu \pi_{2j} & \text{otherwise} \end{cases}$$

3.2 Foregone Payoffs are not Observed

When foregone payoffs are not observed, players have no means of directly comparing the performance of both actions within the same time period. In this case, players reinforce (possibly negatively) the action they played. How much they reinforce this action will depend on the payoff achieved. We use a general case of the Cross (1973) learning rule that also generalizes the rules in BMS. We call this rule the General Reinforcement Rule.

Let $\sigma_i^{t+1}|_{kj}$ be the probability by which a player plays action i at time $t+1$ given that action k was played at time t and state at time t, s^t, was j. The General Reinforcement Rule is defined by

$$\begin{aligned} \sigma_1^{t+1}|_{1j} &= \sigma_1^t + \sigma_2^t g(\pi_{1j}), \\ \sigma_1^{t+1}|_{2j} &= \sigma_1^t - \sigma_1^t g(\pi_{2j}), \end{aligned}$$

and similarly for $\sigma_2^{t+1}|_{1j}$ and $\sigma_2^{t+1}|_{2j}$. The only assumption we make in $g : [0,1] \to [-1,1]$ is that it must be weakly increasing in its argument. If $g(\pi_{ij}) = \pi_{ij}$ then we have the Cross Learning Rule. For the rules in BMS we have that $g(\pi_{ij}) = A_{ij} + B_{ij}\pi_{ij}$ for given $A_{ij} \in \mathbb{R}$ and $B_{ij} \in \mathbb{R}$ for $i \in \{1,2\}$ and $j \in \{1,\ldots,m\}$. BMS show that setting $A_{ij} = -\min\{1-\sigma_1^0, \sigma_1^0\} / \max\{1-\sigma_1^0, \sigma_1^0\}$ and $B_{ij} = 1/\max\{1-\sigma_1^0, \sigma_1^0\}$ for all i,j results in the best monotone rule. A rule is defined to be monotone if the expected probability of playing the action that is best given today's state increases. A rule is said to be the best monotone rule if the expected increase in playing the best action from one period to another is highest among all monotone rules. Since BMS study a setting in which the evolution of

nature follows a stationary distribution, the action that is best today is the action that is best at every period. In our setting the action that is best today may not be the best action tomorrow due to the Markovian evolution of the states of nature. This particular difference will have important consequences in the optimality properties of the rules in BMS.

4 Results

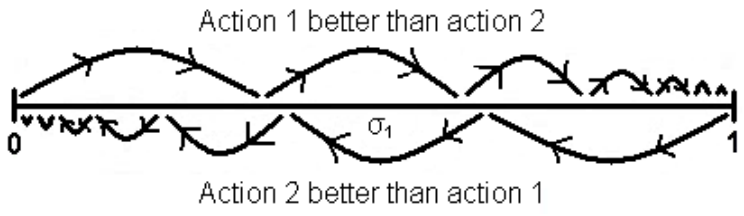
4.1 Results - Foregone payoffs are Observed

Before going to the formal results, we present a small discussion on the behavior of the system under the Stochastic Better Response. First, note that the biggest difference in the behavior of the two rules that we consider lies in the way they behave when σ_i is close to the corners (0 and 1). In particular, under the Stochastic Better Response the corners are not absorbing while the opposite occurs under the General Reinforcement Rule.

Assume for this short discussion that there are only two states of nature. Under the Stochastic Better Response, the speed at which a player adopts an action slows down as the probability of playing that action increases. That is, consider that action 1 is played with a high probability and that today action 1 yielded a higher payoff than action 2. Then the increase in the probability of playing action 1 will be small. On the other hand, consider that action 1 is played with a small probability and today action 1 yielded higher payoff than action 2. In this case the probability of playing action 1 next period increases sharply.

Figure 1 shows the movements of the probability of playing action i (σ_i) as a response to an action being better than the other in the current period. As above, assume that an action is played with a high probability. Then the increase in playing that action in case it yielded a higher payoff than the other action at the present period is low.

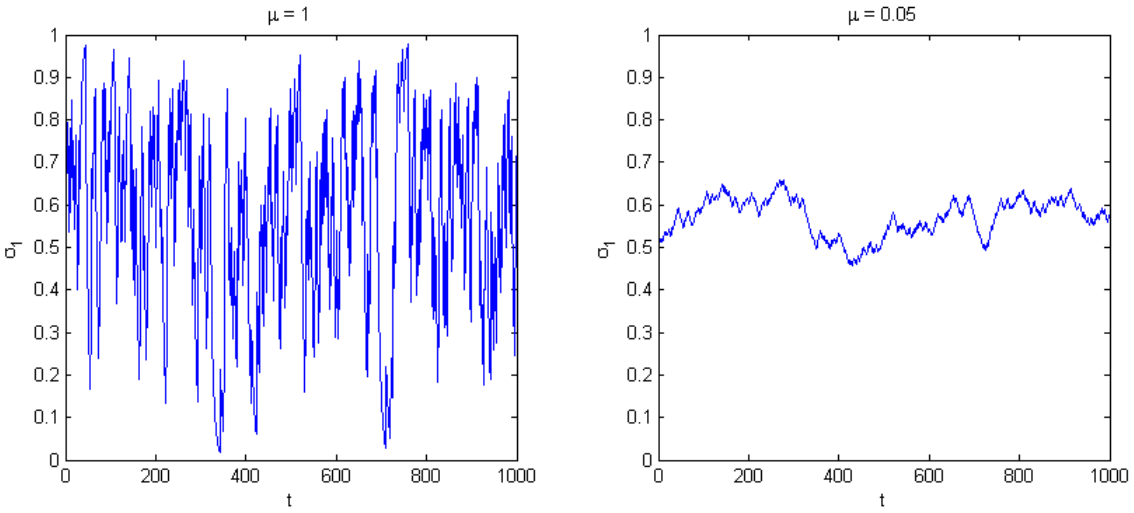
Figure 1: Stochastic Better Response



As one could possibly guess already, the Stochastic Better Response will not converge to any of the corners. To study convergency, we consider the limit case when μ, which can be viewed as the size of the changes in σ_i, gets arbitrarily small. Once such a limit is

taken, the Stochastic Better Response converges to a single point. This issue can be seen much more clear by looking at Figure 2, where a simulation is conducted. The specific learning rule used is given by equation 1. The value of the parameters is set to $m = 2$, $\pi_{11} = 0.5, \pi_{12} = 0.3, \pi_{21} = 0.1, \pi_{22} = 0.6$ and $\theta_{12} = \theta_{21} = 0.3$. The initial value σ_1 was set to $\sigma_1^0 = 0.5$. The figure depicts the same simulation, the same random seed, for two situations: one in which $\mu = 1$ and another in which $\mu = 0.05$.

Figure 2: Simulation - Stochastic Better Response



By studying the behavior of the system when μ is made arbitrarily small we are characterizing the continuous time limit of σ_i. When μ is taken to zero the adjustment in the strategies is made arbitrarily small while keeping constant the speed at which the environment changes. For other papers that use this continuous time limit approximation in settings somewhat different from ours see, for example, Börgers and Sarin (1997) and Benaïm and Weibull (2003).

The following proposition characterizes the convergence of (σ_1, σ_2) under the Stochastic Better Response when μ is arbitrarily small. Later in this section we present a sketch of the proof. The formal proof is contained in the Appendix.

Proposition 1. *Define*

$$\tilde{\sigma} = \frac{\sum_{j:\pi_{1j} \geq \pi_{2j}} \lambda_j f(\pi_j)}{\sum_{j=1}^{m} \lambda_j f(\pi_j)}.$$

For any $\varepsilon > 0$ there exists a $\bar{\mu} > 0$ such that if $\mu < \bar{\mu}$ then

$$P\left(\lim_{t \to \infty} \left|\sigma_1^t - \tilde{\sigma}\right| > \varepsilon\right) = 0.$$

The interpretation of the result is the following. For simplicity of the exposition let us focus on the evolution of the variable σ_1 and assume again that there are only two states of nature. The point $\tilde{\sigma}$ corresponds to the situation where an increase in σ_1^t due to action 1 yielding higher payoff at time t than action 2 would be equivalent to the decrease in σ_1^t from action 2 yielding more payoff than action 1. That is, with $m = 2$, $\tilde{\sigma}$ is the σ_1^t is such that $\left|\sigma_1^{t+1}|_1 - \sigma_1^t\right| = \left|\sigma_1^{t+1}|_2 - \sigma_1^t\right|$. In Figure 1, the point $\tilde{\sigma}$ would be such that the size of the arrows (or jumps) towards the left from a given point σ_1^t is the same as the size of the arrows towards the right from this same point σ_1^t. Hence, $\tilde{\sigma}$ is the point where the marginal movements towards action 1 and towards action 2 are equalized.

One can easily check that $\tilde{\sigma} < 1$, so it will never be the case that the best action in the long run is played with probability 1. For the general case where the Markov chain has m states, action 1 is strictly better than action 2 if and only if $\sum_{j=1}^m \lambda_j \pi_{1j} > \sum_{j=1}^m \lambda_j \pi_{2j}$; this inequality holds in the simulation in Figure 2. However, for that simulation we have that $\tilde{\sigma} = 0.57$. That is, in the long run at any given period action 1 is played with probability of 0.57. This behavior is clearly suboptimal as if $\sum_{j=1}^m \lambda_j \pi_{1j} > \sum_{j=1}^m \lambda_j \pi_{2j}$ then the σ_1^t that maximizes payoff in the long run is $\sigma^* = 1$.

Let us now look at a sketch of the proof. To studying the convergence of the sequence σ_1 we first show that it suffices to study the convergence of a sequence $y = \{y^t\}_{t=\hat{t}}^{\infty}$, for $\hat{t}$ large enough, which evolves in a world with just 2 states of nature and symmetric transition matrix.

First, define the sequence $\hat{\sigma}_1 = \{\hat{\sigma}_1^t\}_{t=\hat{t}}^{\infty}$ as $\hat{\sigma}_1^{\hat{t}} = \sigma_1^{\hat{t}}$ and recursively for $t > \hat{t}$

$$\hat{\sigma}_1^{t+1} = \begin{cases} \hat{\sigma}_1^t + \hat{\sigma}_2^t \mu f(\pi_1) & \text{with probability } \lambda_1 \\ \qquad\qquad\vdots & \\ \hat{\sigma}_1^t + \hat{\sigma}_2^t \mu f(\pi_h) & \text{with probability } \lambda_h \\ \hat{\sigma}_1^t - \hat{\sigma}_1^t \mu f(\pi_{h+1}) & \text{with probability } \lambda_{h+1} \\ \qquad\qquad\vdots & \\ \hat{\sigma}_1^t - \hat{\sigma}_1^t \mu f(\pi_m) & \text{with probability } \lambda_m \end{cases}.$$

Then we have that for any given $t > \hat{t}$,

$$P\left(|E_0(\sigma_1^t) - E_0(\hat{\sigma}_1^t)| > \varepsilon\right) = 0. \tag{2}$$

Hence, the expected value of both σ_1 and $\hat{\sigma}_1$ converge in probability to the same value. This is because the transition matrix P is irreducible and aperiodic. Now define the sequence $y = \{y^t\}_{t=\hat{t}}^{\infty}$ as $y^{\hat{t}} = \hat{\sigma}_1^{\hat{t}}$ and define recursively

$$y^{t+1} = \begin{cases} y^t + 2(1 - y^t)\mu \sum_{j:\pi_{1j} \geq \pi_{2j}} \lambda_j f(\Pi_j) & \text{with probability } 1/2 \\ y^t - 2y^t \mu \sum_{j:\pi_{1j} < \pi_{2j}} \lambda_j f(\Pi_j) & \text{with probability } 1/2 \end{cases}.$$

Note that the variable y evolves according to the expected movement in the long run of the variable $\hat{\sigma}_1$. It can be easily seen that $y^t = \hat{\sigma}_1^t$ implies $E_0(y^{t+1}) = E_0(\hat{\sigma}_1^{t+1})$. Hence, since $y^{\hat{t}} = \hat{\sigma}_1^{\hat{t}}$, the distribution of both y^t and $\hat{\sigma}_1^t$ is aperiodic and both $E_0(y^{\hat{t}+1})$ and $E_0(\hat{\sigma}_1^{\hat{t}+1})$ are linear in their arguments, we can state that $E_0(y^{\hat{t}+k}) = E_0(\hat{\sigma}_1^{\hat{t}+k})$ for any $k \in \mathbb{N}$. Moreover, we have that for any $t > \hat{t}$, equation 2 must hold. Hence, we have that for any $\varepsilon > 0$ and $t > \hat{t}$,

$$P\left(|E_0(\sigma_1^t) - E_0(y^t)| > \varepsilon\right) = 0.$$

Furthermore, by making μ arbitrarily small we make the variance of both random variables y^t and σ_1^t to shrink to zero. Thus, their limiting distribution puts weight on a single point. In other words, y and σ_1 must converge in probability to a fixed value $\bar{y}$ and $\bar{\sigma}$ respectively. Since $E_0(y^{\hat{t}+k})$ converges to $E_0(\sigma_1^{\hat{t}+k})$ for all $k \in \mathbb{N}$, we must have that $\bar{y} = \bar{\sigma}$. Hence, instead of studying the convergence of the variable σ_1 we focus on the convergence of the variable y. This is more formally stated in Lemma 2 in the Appendix.

Note now that the point $y^t = \tilde{\sigma}$, with $\tilde{\sigma}$ as defined in Proposition 1, solves the equation

$$y^t + 2(1 - y^t)\mu \sum_{j:\pi_{1j} \geq \pi_{2j}} \lambda_j f(\pi_j) = y^t - 2y^t\mu \sum_{j:\pi_{1j} < \pi_{2j}} \lambda_j f(\pi_j).$$

Define the sequence $y_1 = \{y_1^t\}_{t=\hat{t}}^{\infty}$ as follows

$$y_1^t = \begin{cases} y^t & \text{if } y^t \geq \tilde{\sigma} \\ \tilde{\sigma} & \text{otherwise} \end{cases}.$$

Hence, we have that $E_0(y^t) \leq E_0(y_1^t)$ for all $t > \hat{t}$. Note that $E_0(y_1^{t+1}) \leq E_0(y_1^t)$. Therefore, y_1 is a super-martingale with lower bound $\tilde{\sigma}$. Thus, by the martingale convergence theorem, y_1 converges in probability to $\tilde{\sigma}$. This implies that for t large enough, $E_0(y^t) \leq \tilde{\sigma}$.

Define now the sequence $y_2 = \{y_2^t\}_{t=\hat{t}}^{\infty}$ as follows

$$y_2^t = \begin{cases} y^t & \text{if } y^t \leq \tilde{\sigma} \\ \tilde{\sigma} & \text{otherwise} \end{cases}.$$

Therefore, we have that $E_0(y^t) \geq E_0(y_2^t)$ for all $t > \hat{t}$. Note that $E_0(y_1^{t+1}) \geq E_0(y_2^t)$. Hence, y_2 is a sub-martingale with upper bound $\tilde{\sigma}$. Thus, by the martingale convergence theorem, y_2 converges in probability to $\tilde{\sigma}$. This implies that for t large enough, $E_0(y^t) \geq \tilde{\sigma}$.

Hence, we know that for t large enough, $E_0(y^t) \leq \tilde{\sigma}$ and $E_0(y^t) \geq \tilde{\sigma}$. This implies that for all $t > \hat{t}$, $E_0(y^t) = \tilde{\sigma}$. Since the variance of y shrinks to zero as μ is made arbitrarily small, we have that y converges in probability to $\tilde{\sigma}$ as μ is made arbitrarily small. Combined with the fact that y converges in probability to σ_1, this implies that σ_1 converges in probability to $\tilde{\sigma}$.

4.2 Results - Foregone Payoffs are not Observed

We recall that the probability by which a player plays action i at time $t+1$ given that action k was played at time t and state at time t was j is denoted by $\sigma_i^{t+1}|_{kj}$ and given by

$$\begin{aligned}\sigma_1^{t+1}|_{1j} &= \sigma_1^t + \sigma_2^t g(\pi_{1j}),\\ \sigma_1^{t+1}|_{2j} &= \sigma_1^t - \sigma_1^t g(\pi_{2j}).\end{aligned}$$

Hence, $\sigma_1^{t+1}|_j$, which is the probability of playing action 1 at time $t+1$ given that state was j, equals $\sigma_1^t + \sigma_2^t g(\pi_{1j})$ if action 1 was played at time t and $\sigma_1^t - \sigma_1^t g(\pi_{2j})$ if action 2 was played at time t. Action i with $i \in \{1,2\}$ is played at time t with probability σ_i^t. Hence, since we are dealing with a continuum of players, we can use Law of Large Numbers to state that

$$\sigma_1^{t+1}|_j = \sigma_1^t \sigma_1^{t+1}|_{1j} + \sigma_2^t \sigma_1^{t+1}|_{2j}.$$

This can be rewritten as

$$\sigma_1^{t+1}|_j = \sigma_1^t \left(\sigma_1^t + (1-\sigma_1^t) g(\pi_{1j})\right) + (1-\sigma_1^t)\left(\sigma_1^t - \sigma_1^t g(\pi_{2j})\right).$$

Thus, it follows that

$$\sigma_i^{t+1}|_j = \sigma_i^t \left(1 + (1-\sigma_i^t)\left[g(\pi_{ij}) - g(\pi_{-ij})\right]\right). \tag{3}$$

Note that if we set $g(\pi_{ij}) = \pi_{ij}$, as in the Cross Learning Rule, the resulting law of motion for σ_i is the discreet time version of the Replicator Dynamics. That is, if $g(\pi_{ij}) = \pi_{ij}$ then we have that

$$\sigma_i^{t+1}|_j = \sigma_i^t + \sigma_i^t \left(\pi_{ij} - \left[\sigma_i^t \pi_{ij} + \sigma_{-i}^t \pi_{-ij}\right]\right).$$

The General Reinforcement Rule behaves completely differently to the Stochastic Better Response. Under the General Reinforcement Rule, the changes in the variable σ_i^t become smaller as σ_i^t gets closer to either bound. For example, consider that action 1 is played with a high probability. Then the change in σ_i will be small independently of whether action 1 yielded higher payoff than action 2 or the other way around. Figure 3 shows the movements of σ_1 under the General Reinforcement Rule as a response to the environment.

As we see, the process will spend almost no time in intermediate values of σ_i. This will allow us to draw our conclusions from analyzing only the behavior of σ_i in the neighborhoods of its bounds. In this respect, our analysis will partially rely on the approach by Ellison and Fudenberg (1995).

Figure 3: General Reinforcement Rule

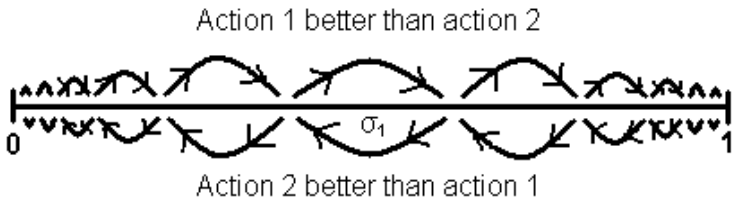


Figure 4: Simulation - General Reinforcement Rule

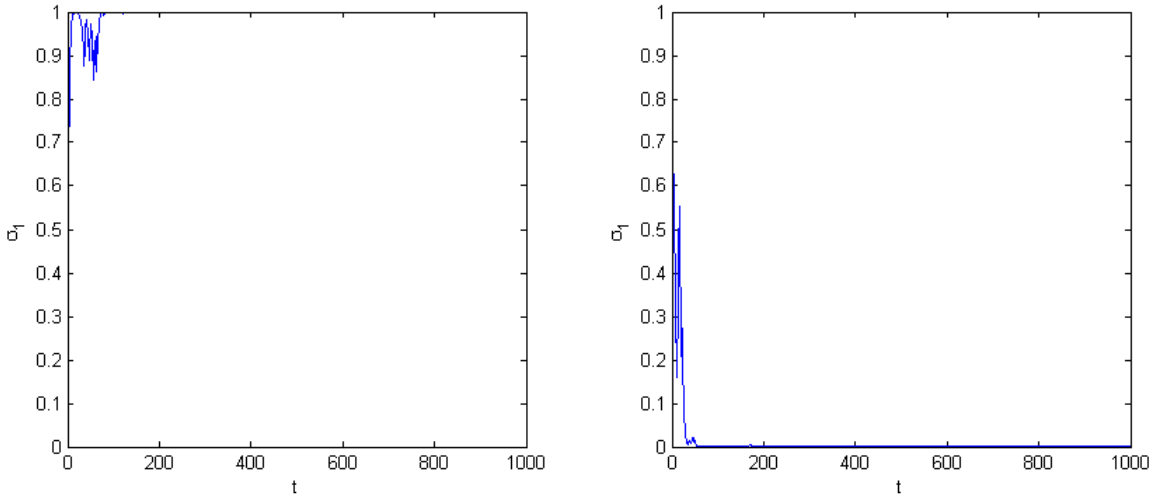


Figure 4 shows a simulation for the General Reinforcement Rule for the case where $g(\pi_{ij}) = \pi_{ij}$ and with the same parameters as the ones used in Figure 2. The figure plots the result of the same simulation performed with two different random seeds.

It can be seen that the General Reinforcement Rule quickly converges to a situation in which all the population plays the same action a fraction 1 of the time. An interesting thing to note is that the action selected by the General Reinforcement Rule does not coincide necessarily with the action that is best in the long run. The simulation on the right-hand side shows a situation in which the General Reinforcement Rule converges to a situation where all players in the population are playing the suboptimal action. As we will see, this is the result of the two actions performing not too differently in terms of payoffs in the long run.

The following proposition, whose proof is presented in the Appendix, characterizes the convergence of the sequence σ_1.

Proposition 2. *Define* $\gamma_j = 1 + g(\pi_{1j}) - g(\pi_{2j})$ *and* $\hat{\gamma}_j = 1 + g(\pi_{2j}) - g(\pi_{1j})$ *and consider*

the two inequalities:

$$\sum_{j=1}^{m} \lambda_j \log \gamma_j \quad > \quad 0, \tag{4}$$

$$\sum_{j=1}^{m} \lambda_j \log \hat{\gamma}_j \quad > \quad 0. \tag{5}$$

1. *If both (4) and (5) hold then* $\lim_{t\to\infty} \sigma_1^t$ *does not exist.*

2. *If (4) holds but (5) does not then* $\lim_{t\to\infty} \sigma_1^t = 1$.

3. *If (5) holds but (4) does not then* $\lim_{t\to\infty} \sigma_1^t = 0$.

4. *If neither (4) nor (5) hold then* $\lim_{t\to\infty} \sigma_1^t$ *has full support over* $\{0, 1\}$.

Since $\sigma_2 = 1 - \sigma_1$ the convergence of the sequence σ_2 follows for the proposition above. An important fact revealed by proposition above is that the process may fail to converge to the best action. Consider for simplicity the Cross Learning Rule, where $g(\pi_{ij}) = \pi_{ij}$. Action 1 is weakly better than action 2 in the long run if and only if $\sum_{j=1}^{m} \lambda_j \pi_{1j} \geq \sum_{j=1}^{m} \lambda_j \pi_{2j}$. This condition can be rewritten as $\sum_{j=1}^{m} \lambda_j \gamma_j \geq 1$. However, even if $\sum_{j=1}^{m} \lambda_j \gamma_j \geq 1$ holds, it may still happen that $\sum_{j=1}^{m} \lambda_j \log \gamma_j < 0$ holds and hence σ_1 may not converge to 1. To make this point more clear consider the case in which $m = 2$ and $\lambda_1 = \lambda_2 = 0.5$. That is, there are only two states of nature and both states are equally likely in the long run. The following corollary characterizes the convergence of σ_1 in this case when action 1 is better in the long run than action 2.

Corollary 1. *Assume* $g(\pi_{ij}) = \pi_{ij}$, $m = 2$, $\lambda_1 = \lambda_2 = 0.5$ *and* $\pi_{11} + \pi_{12} > \pi_{21} + \pi_{22}$.

- *If* $\pi_{11} + \pi_{12} - \pi_{21} - \pi_{22} - (\pi_{11} - \pi_{21})(\pi_{22} - \pi_{12}) > 0$ *then* $\lim_{t\to\infty} \sigma_1 = 1$.

- *Otherwise,* $\lim_{t\to\infty} \sigma_1$ *has full support over* $\{0, 1\}$.

Proof. We can rewrite inequalities (4) and (5) from Proposition 2 for the case with $m = 2$ and $\lambda_1 = \lambda_2 = 0.5$ as follows:

$$\log \gamma_1 + \log \gamma_2 \quad > \quad 0 \tag{6}$$

$$\log \hat{\gamma}_1 + \log \hat{\gamma}_2 \quad > \quad 0. \tag{7}$$

The conditions (6) and (7) can be rewritten as $\gamma_1\gamma_2 > 1$ and $\hat{\gamma}_1\hat{\gamma}_2 > 1$. These in turn can be rewritten as

$$\pi_{11} + \pi_{12} - \pi_{21} - \pi_{22} - (\pi_{11} - \pi_{21})(\pi_{22} - \pi_{12}) \quad > \quad 0, \tag{8}$$

$$\pi_{21} + \pi_{22} - \pi_{11} - \pi_{12} - (\pi_{11} - \pi_{21})(\pi_{22} - \pi_{12}) \quad > \quad 0. \tag{9}$$

It can be easily seen that equation (9) is never holding. Hence, by Proposition 2, if the inequality (8) holds then we have that $\lim_{t\to\infty}\sigma_1 = 1$, whereas if (8) does not hold we have that $\lim_{t\to\infty}\sigma_1$ has full support over $\{1,2\}$. □

For the process to select the best action, the two actions need to perform significantly differently. That is, having action 1 better than action 2, $\pi_{11}+\pi_{12}-\pi_{21}-\pi_{22}>0$, is not enough for the process to select the best action.

Now we present the intuition for the proof of Proposition 2 for the case where $g(\pi_{ij})=\pi_{ij}$. The proof of Proposition 2 relies partially on the analysis by Ellison and Fudenberg (1995).

In Ellison and Fudenberg (1995), the realization of states of nature is independent of past values of states. In order to be able to apply Ellison and Fudenberg's analysis to our setting, we proceed as follows. Given that the transition matrix P is irreducible and aperiodic, the state of nature many periods ahead is independent of the state of nature today. This means that by the law of large numbers, we can take the probability of each state being realized many periods ahead as the limiting probability placed on it by the Markov chain. Therefore, for the rest of the exposition we consider that the realization of states is independent of past values. For a formal proof the reader is referred to Lemma 4 in the Appendix.

Assume, for the simplicity of the exposition, that there are only two states of nature. Let $1-p$ be the probability by which state 1 occurs. Since the process spends almost no time at its intermediate values, it suffices to examine the convergence of the variable σ_i when it is close to its boundary values (0 and 1). To make the exposition clearer, we focus on the sequence $\sigma_2 = 1-\sigma_1$. Imagine that σ_2 is arbitrarily close to 0. Then we can rewrite (3) as follows:

$$\sigma_2^{t+1}|_j = \gamma_j\sigma_2^t + o(\sigma_2^t) \tag{10}$$

where $\gamma_j = 1+g(\pi_{2j})-g(\pi_{1j})$ for $j\in\{1,2\}$ and $o(\sigma_2^t)$ is a term of order higher than σ_2 and hence is negligible when σ_2 is arbitrarily small. Without loss of generality we can assume that $\pi_{11}>\pi_{21}$, which implies $\pi_{12}<\pi_{22}$. Then we can rewrite (10) as

$$\sigma_2^{t+1}|_j = \begin{cases} \gamma_1\sigma_2^t + o(\sigma_2^t) & \text{if } \pi_{1j}\geq\pi_{2j} \\ \gamma_2\sigma_2^t + o(\sigma_2^t) & \text{otherwise} \end{cases}.$$

Since $\pi_{11}>\pi_{21}$ and $\pi_{12}<\pi_{22}$ we have that $\gamma_2>1>\gamma_1>0$. Finally, note that $\pi_{1j}\geq\pi_{2j}$ with probability $1-p$ and $\pi_{1j}<\pi_{2j}$ with probability p.

The sequence σ_2 converges to 0, or σ_1 converges to 1, if and only if the sequence $x=\{x^t\}_{t=0}^{\infty}$ with $x^t=\log\sigma_2^t$ converges to $-\infty$. The process for x when σ_2^t is close to 0 can be

approximated by

$$x^{t+1} = \begin{cases} \log \gamma_1 + x^t & \text{with probability } 1-p \\ \log \gamma_2 + x^t & \text{with probability } p \end{cases}.$$

Therefore, $E_t(x^{t+1}) = (1-p)\log\gamma_1 + p\log\gamma_2 + x^t$. Hence, if $(1-p)\log\gamma_1 + p\log\gamma_2 > 0$ then $E_t(x^{t+1}) > x^t$, which implies that x is a sub-martingale. Thus, by the Martingale Convergence Theorem, if $(1-p)\log\gamma_1 + p\log\gamma_2 > 0$ then x cannot converge to $-\infty$ and hence σ_2 cannot converge to 0. Which implies that σ_1 does not converge to 1.

Ellison and Fudenberg's (1995) result is presented here for the readers' convenience.

Lemma 1 (Ellison and Fudenberg (1995))**.** *Let z^t be a Markov Process on (0,1) with*

$$z^{t+1} = \begin{cases} \gamma_1 z^t + o(z_t) & \textit{with probability } 1-p \\ \gamma_2 z^t + o(z_t) & \textit{with probability } p \end{cases}.$$

Suppose that $\gamma_1 < 1 < \gamma_2$.

(a) If

$$\frac{p}{1-p} > -\frac{\log(\gamma_1)}{\log(\gamma_2)},$$

then z^t cannot converge to 0 with positive probability.

(b) If

$$\frac{p}{1-p} < -\frac{\log(\gamma_1)}{\log(\gamma_2)},$$

then there are $\delta > 0$ and $\varepsilon > 0$ such that if $z^0 < \delta$ then $P\left(\lim_{t\to\infty} z^t = 0\right) \geq \varepsilon$.

(c) If

$$\frac{p}{1-p} > -\frac{\log(\gamma_1)}{\log(\gamma_1)},$$

there is a $\bar{z} > 0$ such that for all $z^0 > 0$ and all $t \in \{0, 1, \dots\}$, $P\left(z^t < \bar{z}\right) = 0$.

4.3 Efficient Learning Rules

We say that a learning rule is efficient if it is able to select to optimal action in the long run. An interesting result is that if foregone payoffs are observed, then it is optimal to disregard this information and to act as if only realized payoffs were observed.

When players observe the performance of both actions they can be "distracted" towards the suboptimal action by the Markov chain. This is because even if the population plays the optimal action with a high probability they can still observe the performance of the

suboptimal action. Hence, since the suboptimal action is the best action for some states of nature, randomness can constantly lead some players in the population to adopt the suboptimal action for many periods in time. Thus, the continuous time limit of the process converges to a situation in which the suboptimal action is played with a positive probability. This is formally proven in the next proposition.

Proposition 3. *Under the Stochastic Better Response, for some $\varepsilon > 0$ there exists no $f : [0,1]^2 \to [0,1]$ such that for all the environments $(\{\pi_1, \ldots, \pi_m\}, P)$ we have that $|\tilde{\sigma}_1 - \sigma_1^*| < \varepsilon$.*

Proof. Assume, without loss of generality, that $\sum_{j=1}^{m} \lambda_j \pi_{1j} > \sum_{j=1}^{m} \lambda_j \pi_{2j}$. Hence, we have that $\sigma_1^* = 1$.

The proof goes by contradiction. Assume that for all $\varepsilon > 0$ there exists a function $f : [0,1]^2 \to [0,1]$ such that for all the environments $(\{\pi_1, \ldots, \pi_m\}, P)$, $|\tilde{\sigma}_1 - \sigma_1^*| < \varepsilon$. This can be rewritten as follows: there exists a sequence of functions $f = \{f_n\}_{n=0}^{\infty}$ with $f_n : [0,1]^2 \to [0,1]$ for all $n \geq 0$ such that for all the environments we have that

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \tilde{\sigma}_1(f_n) = \sigma_1^* = 1,$$

where $\tilde{\sigma}_1(f_n)$ is the value of $\tilde{\sigma}_1$ associated with the function f_n.

The limit above holds if and only if

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{\sum_{j:\pi_{1j} \geq \pi_{2j}} \lambda_j f_n(\pi_j)}{\sum_{j:\pi_{1j} < \pi_{2j}} \lambda_j f_n(\pi_j)} = \infty \tag{11}$$

holds.

Take now an environment $E = (\{\pi_1, \pi_2\}, P)$ where $0 < \pi_{11} < \pi_{22}$ and $\pi_{ij} = 0$ for all $i \neq j$. We could consider more general environments but that will only complicate the exposition leaving the logic of the proof unchanged. P is such that action 1 is the optimal one in the long run. That is, given $\pi_{11} < \pi_{22}$ and $\pi_{ij} = 0$ for all $i \neq j$, P is such that $\sum_{j=1}^{2} \lambda_j \pi_{1j} > \sum_{j=1}^{2} \lambda_j \pi_{2j}$. In this situation, equation (11) implies that

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{\lambda_1 f_n(\pi_1)}{1 - \lambda_1 f_n(\pi_2)} = \infty. \tag{12}$$

Given that the transition matrix P is irreducible we have that $\lambda_1 \in (0,1)$. Thus, we must have that (12) holds if and only if the following limit holds.

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{f_n(\pi_1)}{f_n(\pi_2)} = \infty \tag{13}$$

However, given that $\pi_{11} < \pi_{22}$ and $\pi_{ij} = 0$ for all $i \neq j$, we have that $f_n(\pi_1) < f_n(\pi_2)$ for all $n > 0$. Hence, the sequence f is such that equation (13) cannot hold for the environment E, a contradiction. □

The logic behind the proof is that if a learning rule makes the population to select the optimal action in a given environment E', then the rule must magnify the payoffs of each action. This can be seen in equation (11), where, according to the learning rule, payoffs are magnify to infinity. However, if this is the case, an environment E can be found such that there is a very rare state for which the payoff of the suboptimal action is much bigger than the payoff of the optimal action for that state. In this situation, the learning rule that make the population to select the best action for environment E' will fail to do so in environment E.

When only realized payoffs are observed, a different force operates. Once the population is almost always playing the optimal action, it is very difficult for players to take notice of the periods in which the suboptimal action is giving more payoff than the optimal action. A drawback for the population under this informational setting is that if both actions perform not too differently in terms of payoffs, the population may lock on the suboptimal action forever. However, a learning rule can be designed such that this inefficiency is avoided.

The next result states two important features about efficiency rules under the General Reinforcement Rule. The first one is that if learning is sufficiently cautious in that the magnitude of payoffs is diminished then the population will select the optimal action. The second important feature is that how cautious the learning has to be depends on how big the difference in the long run average payoff of both actions is. The more both actions differ in terms of long run performance, the more cautious the learning has to be. This implies that while a learning rule that is very cautious may not be able to make the population to select the best action, this will only happen in environments where the two actions perform very similarly in the long run. Hence, when cautious learning is exhibited, the possible loss in payoff from not selecting the best action is small.

Proposition 4. *Under the General Reinforcement Rule, assume* $g : [0, 1] \to [-1, 1]$ *is given by*

$$g(\pi_{ij}) = x\pi_{ij}$$

where

$$x = \frac{1 + 4\varepsilon - \sqrt{1 + 8\varepsilon}}{4\varepsilon}$$

for some $\varepsilon > 0$. *If* $|\sum_{j=1}^{m} \lambda_j \pi_{1j} - \sum_{j=1}^{m} \lambda_j \pi_{2j}| > \varepsilon$, *then we have that* $\lim_{t\to\infty} \sigma_1^t = \sigma_1^*$.

Proof. Assume, without loss of generality, that $\sum_{j=1}^{m} \lambda_j \pi_{1j} > \sum_{j=1}^{m} \lambda_j \pi_{2j}$. Hence, we have that $\sigma_1^* = 1$. Moreover, given the inequality $|\sum_{j=1}^{m} \lambda_j \pi_{1j} - \sum_{j=1}^{m} \lambda_j \pi_{2j}| > \varepsilon$, we must have that $\sum_{j=1}^{m} \lambda_j (x\pi_{1j} - x\pi_{2j}) > x\varepsilon$ for all $x > 0$.

Using the first order Taylor series for the logarithmic function around 1 we get that

$$\log(1 + x\pi_{1j} - x\pi_{2j}) = x\pi_{1j} - x\pi_{2j} + R_1(1 + x\pi_{1j} - x\pi_{2j}),$$

where $R_1(1 + x\pi_{1j} - x\pi_{2j})$ is the remainder term and $x > 0$. Using the Lagrange form we can rewrite the remainder term as

$$R_1(1 + x\pi_{1j} - x\pi_{2j}) = \frac{-1/y^2}{2}(1 + x\pi_{1j} - x\pi_{2j} - 1)^2,$$

where y lies between 1 and $1+x\pi_{1j}-x\pi_{2j}$. We can bound the absolute value of the remainder term in the following way:

$$\begin{aligned} |R_1(1 + x\pi_{1j} - x\pi_{2j})| &\leq \frac{1/(1-x)^2}{2}(x\pi_{1j} - x\pi_{2j})^2 \\ &\leq \frac{x^2}{2(1-x)^2}. \end{aligned}$$

Moreover, we have that

$$\begin{aligned} \log(1 + x\pi_{1j} - x\pi_{2j}) &= x\pi_{1j} - x\pi_{2j} + R_1(1 + x\pi_{1j} - x\pi_{2j}) \\ &\geq x\pi_{1j} - x\pi_{2j} - |R_1(1 + x\pi_{1j} - x\pi_{2j})|. \end{aligned}$$

This can be rewritten as

$$\begin{aligned} \sum_{j=1}^{m} \lambda_j \log(1 + x\pi_{1j} - x\pi_{2j}) &\geq \sum_{j=1}^{m} \lambda_j \left(2x\pi_{1j} - x\pi_{2j} - |R_1(1 + x\pi_{1j} - x\pi_{2j})|\right) \\ &> x\varepsilon - \frac{x^2}{2(1-x)^2}. \end{aligned}$$

If we take $x > 0$ to be the minimum solution to the equation

$$x\varepsilon - \frac{x^2}{2(1-x)^2} = 0,$$

we get that

$$x = \frac{1 + 4\varepsilon - \sqrt{1 + 8\varepsilon}}{4\varepsilon}. \tag{14}$$

Thus, setting $x > 0$ as in equation (14) yields

$$\sum_{j=1}^{m} \lambda_j \log \gamma_j > 0. \tag{15}$$

Similar arguments show that

$$\begin{aligned} \sum_{j=1}^{m} \lambda_j \log(1 - x\pi_{1j} + x\pi_{2j}) &\leq -\sum_{j=1}^{m} \lambda_j x\pi_{1j} - x\pi_{2j} + |R_1(1 - x\pi_{1j} + x\pi_{2j})| \\ &< -x\varepsilon + \frac{x^2}{2(1-x)^2}. \end{aligned}$$

Hence, setting again $x > 0$ as in equation (14) yields

$$\sum_{j=1}^{m} \lambda_j \log \hat{\gamma}_j \quad < \quad 0. \tag{16}$$

Finally, combining inequalities (15) and (16) with Proposition 2 we get that if $g(\pi_{ij}) = x\pi_{ij}$, where we set $x > 0$ as in equation (14), and if $|\sum_{j=1}^{m} \lambda_j \pi_{1j} - \sum_{j=1}^{m} \lambda_j \pi_{2j}| > \varepsilon$, then we have that $\lim_{t\to\infty} \sigma_1^t = \sigma_1^*$. □

Note that if we set $g(\pi_{ij})$ as in Proposition 4, then $\lim_{\varepsilon\to 0} g(\pi_{ij}) = 0$. That is, a rule that makes the population able to select the best action in all the environments must exhibit arbitrarily slow learning.

5 Discussion

A way of enriching the model could be by adding idiosyncratic perturbations to payoffs. This could be done by adding ε_{ht} to each payoff π_{ij}. ε_{ht} are normally distributed zero mean random variables that are independent across players h and time t. Since the rules we consider under both scenarios can treat payoffs in a non-linear way, it is not true that the process will converge to the same values as compared to the case without noise. The reason is the same as why, for instance, $E(x^2) \neq E\left((x+\varepsilon)^2\right)$ with $E(\varepsilon) = 0$. However, it can easily be verified that adding noise makes no difference to our results for all the learning rules that treat payoffs linearly. Rules that treat payoffs linearly include the standard best response and the bernoulli best response, for the case where foregone payoffs are observed, and the Cross Learning Rule and the rules in BMS, for the case where foregone payoffs are not observed.

One might argue that if players had means of comparing the payoff of the same action across different time periods, they could recall different payoff realizations over time and have significantly more information about the world they are living in. However, as showed by Rustichini (1999) in a setting very similar to ours, even if players had infinite memory and could make this comparison, it is not true that they will learn the best action for sure.

5.1 Relating our results for the Stochastic Better Response with Kosfeld et. al. (2002)

Kosfeld et. al. (2002) present a setting where a finite set of players play a normal-form game. Each period players update their strategies myopically in the following way. They increase the probability of playing an action if and only if that action is a best response to the action

played by the other players. If there are many actions that are a best response, the increase in probability is shared equally among the actions that are a best response. Formally, let $\sigma_i^t(j)$ be the probability by which player j plays action i at time t. Define s_{-j} as the actions played by all the players but j. Finally, let $B_j(s_{-j})$ be the set of actions that are a best response for player j to s_{-j} and let $|B_j(s_{-j})|$ be the cardinality of $B_j(s_{-j})$. The evolution in the strategies of every player j is governed by

$$\sigma_i^{t+1}(j) = \begin{cases} (1-\mu)\sigma_i^t(j) + \mu/|B_j(s_{-j})| & \text{if } s_j \in B_j(s_{-j}) \\ (1-\mu)\sigma_i^t(j) & \text{otherwise,} \end{cases} \tag{17}$$

where $\mu \in (0,1)$ is exogenously given.

Comparing this rule with the Stochastic Better Response there are two points worth noting. First, the rule in Kosfeld et. al. (2002) is a particular case of the Stochastic Better Response. Second, and most importantly, in our model players play against nature and not against themselves. Hence, in Kosfeld et. al.'s (2002) setting, players best respond to the actions of other players while in our setting players best respond to the actions of nature.

Kosfeld et. al. (2002) show that the continuous time limit of their process, when μ is made arbitrarily small, converges to a so-called Best-Reply Matching Equilibrium. In a Best-Reply Matching Equilibrium, for every player, the probability of playing a given action is equal to the probability by which that action is a best response given the strategies of the other players.

Their result and our result for the Stochastic Better Response have the same intuition behind them and in some situations are equivalent. Given that in our setting there are only two action we can rewrite (17) as follows.

$$\sigma_1^t|_j = \begin{cases} \sigma_1^t + \sigma_2^t\mu & \text{if } \pi_{1j} \geq \pi_{2j} \\ \sigma_1^t - \sigma_1^t\mu & \text{otherwise} \end{cases}$$

In Proposition 1 we proved that the sequence σ_1 defined above converges in probability to

$$\begin{aligned} \hat{\sigma} &= \frac{\sum_{j:\pi_{1j}\geq\pi_{2j}} \lambda_j}{\sum_{j=1}^m \lambda_j} \\ &= \sum_{j:\pi_{1j}\geq\pi_{2j}} \lambda_j. \end{aligned}$$

That is, σ_i^t, which is the probability of playing action i, converges to the limiting probability that action i is a best response to the environment. Hence, the population strategies match the nature's strategies, exactly as predicted by the Best-Reply Matching Equilibrium.

In our results for the Stochastic Better Response we consider a much bigger set of rules than do Kosfeld et. al. (2002). In particular, Kosfeld et. al. (2002) only consider one rule. However, for the specific rule used by Kosfeld et. al. (2002), their results and ours come from two different settings, as in their setting players play against each other while in our setting players play against nature.

6 Conclusions

In this paper we investigated learning within an environment that changes according to a Markov chain and where players learn according to reinforcement. The payoff of each possible action depends on the state of nature. Since transition between states follows a Markov Chain, there is correlation between today's state and tomorrow's state of nature. We studied two different scenarios, one in which realized and foregone payoffs are observed and another in which only realized payoffs are observed. Our contribution to the literature relies on the fact that we studied reinforcement learning in a setting where the realization of the state of nature is correlated with the past.

The literature has focused on the study of learning only in a setting where the realization of states (or the shocks to payoffs) is independent of its past values. The reason for this is the technical complexities involved in dealing with the correlated realization of states.

There are several questions left for further research. For the case where foregone payoffs are observed, we only characterized the asymptotic distribution when the learning step goes to zero. For the case where foregone payoffs are not observed we are unable to quantify the probabilities of reaching each endpoint where the process does not converge deterministically to a single point.

The present piece of work explores learning in two very general scenarios but there are other settings that could be of interest. For instance, how does local interaction affect learning when the environment changes according to a Markov chain? What if there are non-stochastic idiosyncratic payoff differences among players? Our paper also tried to shed some light on the techniques that could be used for dealing with such environments. We expect that in the future more papers dealing with non stationary environments will appear.

References

1. Ben-Porath, E., Dekel, E. & Rustichini, A. (1993): "On the Relationship between Mutation Rates and Growth Rates in Changing Environment". *Games and Economic Behavior* 5 (4), 576-603.

2. Benaïm, M. & Weibull, J. (2003): "Deterministic Approximation of Stochastic Evolution in Games". *Econometrica* 71 (3), 873-903.

3. Börgers, T., Morales, A. & Sarin, R. (2004): "Expedient and Monotone Learning Rules". *Econometrica* 71 (2), 383-405.

4. Börgers, & Sarin, R. (1997): "Learning Through Reinforcement and Replicator Dynamics". *Journal of Economic Theory* 77, 1-14.

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

6. Conlisk, J. (1996): "Why Bounded Rationality?". *The Journal of Economic Literature* 34, 669-700.

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

8. Ellison, G. & Fudenberg, D. (1995): "Word-of-Mouth Communication and Social Learning". *The Quarterly Journal of Economics* 110 (1), 93-125.

9. Erev, I. & 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.

10. Fudenberg, D. & Harris, C. (1992): "Evolutionary Dynamics with Aggregate Shocks". *Journal of Economic Theory* 57, 420-441.

11. Hirshleifer, D. & Welch, I. (2002): "An Economic Approach to the Psychology of Change: Amnesia, Inertia, and Impulsiveness". *Journal of Economics & Management Strategy* 11 (3), 379-421.

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

13. Rubinstein, A. (1998): "Modeling Bounded Rationality". The MIT press, Cambridge, Massachusetts.

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

15. Rustichini, A. (1999): "Optimal Properties of Stimulus Response Learning Models". *Games and Economic Behavior* 29, 244-273.

16. Roth and Erev (1995): "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". *Games and Economic Behavior* 8, 164-212.

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

18. Shanks, D., Tunney, R. & McCarthy, J. (2002): "A Re-examination of Probability Matching and Rational Choice". *Journal of Behavioral Decision Making* 15, 233-250.

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

20. Vulkan, N. (2000): "An Economist's Perspective on Probability Matching". *Journal of Economic Surveys* 14 (1), 101118.

Appendix

Proof of Proposition 1

We begin by proving the following lemma.

Lemma 2. *For any $\varepsilon > 0$ there exists a $\hat{\mu} > 0$, $\hat{t}(\varepsilon) > 0$ and a sequence $y = \{y^t\}_{t=\hat{t}}^{\infty}$ given by $y^{\hat{t}} = \sigma_1^{\hat{t}}$ and recursively for $t > \hat{t}$*

$$y^{t+1} = \begin{cases} y^t + 2(1-y^t)\mu \sum_{j:\pi_{1j} \geq \pi_{2j}} \lambda_j f(\pi_j) & \text{with probability } 1/2 \\ y^t + 2y^t \mu \sum_{j:\pi_{1j} < \pi_{2j}} \lambda_j f(\pi_j) & \text{with probability } 1/2 \end{cases},$$

such that for any $\mu < \hat{\mu}$ we have that

$$P\left(\lim_{t\to\infty} |\sigma_1^t - y^t| > \varepsilon\right) = 0.$$

Proof. In the main text we defined $h < m$ as the minimum natural number such that $\pi_{1j} \geq \pi_{2j}$ for $j \leq h$ and $\pi_{2j} > \pi_{1j}$ for $j > h$. For any given $\varepsilon > 0$ define now the sequence $\hat{\sigma}_1 = \{\hat{\sigma}_1^t\}_{t=\hat{t}(\varepsilon)}^{\infty}$ as $\hat{\sigma}_1^{\hat{t}(\varepsilon)} = \sigma_1^{\hat{t}(\varepsilon)}$ and recursively for $t > \hat{t}(\varepsilon)$

$$\hat{\sigma}_1^{t+1} = \begin{cases} \hat{\sigma}_1^t + \hat{\sigma}_2^t \mu f(\pi_1) & \text{with probability } \lambda_1 \\ \vdots & \\ \hat{\sigma}_1^t + \hat{\sigma}_2^t \mu f(\pi_h) & \text{with probability } \lambda_h \\ \hat{\sigma}_1^t - \hat{\sigma}_1^t \mu f(\pi_{h+1}) & \text{with probability } \lambda_{h+1} \\ \vdots & \\ \hat{\sigma}_1^t - \hat{\sigma}_1^t \mu f(\pi_m) & \text{with probability } \lambda_m \end{cases}.$$

Fix $\hat{t}(\varepsilon)$ to be the minimum natural number such that for any given $t > \hat{t}(\varepsilon)$,

$$P\left(|E_0(\sigma_1^t) - E_0(\hat{\sigma}_1^t)| > \varepsilon\right) = 0. \tag{18}$$

The existence of such $\hat{t}(\varepsilon)$ is guaranteed by the fact that the transition matrix P is irreducible and aperiodic and by the Perron-Frobenius theorem applied to P. In an abuse of notation, from now on we will simply write $\hat{t}$ to denote $\hat{t}(\varepsilon)$.

Since E_0 is linear in both $\hat{\sigma}_1^t$ and y^t, we have that for all $t > \hat{t}$, $\hat{\sigma}_1^t = y^t$ if and only if $E_0(\hat{\sigma}_1^{t+1}) = E_0(y^{t+1})$. Thus, given that $y^{\hat{t}} = \hat{\sigma}_1^{\hat{t}}$, that E_0 is linear in both $\hat{\sigma}_1^t$ and y^t and that the distribution of both y and $\hat{\sigma}_1$ is aperiodic, we have that

$$E_0(y^{\hat{t}+k}) = E_0(\hat{\sigma}_1^{\hat{t}+k}) \tag{19}$$

for all $k \in \mathbb{N}$.

Given the definition of y and equations (18) and (19) we must have that for any $\varepsilon > 0$ and any $t > \hat{t}$,

$$P\left(|E_0(\sigma_1^t) - E_0(y^{t+1})| > \varepsilon\right) = 0. \tag{20}$$

Given the specification of σ_1 and the definitions of $\hat{\sigma}_1$ and y, as μ gets arbitrarily small, the variance of σ_1, $\hat{\sigma}_1$ and y gets arbitrarily small as well. Formally, for any $\varepsilon > 0$ there exists a $\hat{\mu} > 0$ and a $t > \hat{t}$ such that for any $\mu < \hat{\mu}$ and $k \in \mathbb{N}$ we have that $Var_t(\sigma_1^{t+k}) < \varepsilon$, $Var_t(\hat{\sigma}_1^{t+k}) < \varepsilon$ and $Var_t(y^{t+k}) < \varepsilon$.

Assume that σ_1 does not converge in probability to y. As μ goes to zero the variance of both σ_1 and y goes to zero. Hence, both variables will converge in probability to a single point. That is, for all $\delta > 0$ there exists $\bar{\sigma}_1$, $\bar{y}$, $\bar{\mu} > 0$ and $\bar{t} \in \mathbb{N}$ such that for all $\mu < \bar{\mu}$ and $t > \bar{t}$, $P\left(|\sigma_1^t - \bar{\sigma}_1| > \delta\right) = 0$ and $P\left(|y_1^t - \bar{y}| > \delta\right) = 0$. This can also be rewritten as $P\left(|E_0(\sigma_1^t) - \bar{\sigma}_1| > \delta\right) = 0$ and $P\left(|E_0(y_1^t) - \bar{y}| > \delta\right) = 0$.

If $\bar{\sigma}_1 \neq \bar{y}$, them we must have that exists a $\gamma > 0$ and a $t \in \mathbb{N}$ such that

$$P\left(|E_0(\sigma_1^{t+k}) - E_0(y^{t+k})| > \gamma\right) > 0$$

for all $k \in \mathbb{N}$, which contradicts equation (20). Hence, given that $P\left(|\sigma_1^t - \bar{\sigma}_1| > \delta\right) = 0$, $P\left(|y_1^t - \bar{y}| > \delta\right) = 0$ and $\bar{\sigma}_1 = \bar{y}$, we must have that for any $\varepsilon > 0$ there exists a $\hat{\mu}$ such that for all $\mu < \hat{\mu}$,

$$P\left(\lim_{t\to\infty} |\sigma_1^t - y^t| > \varepsilon\right) = 0.$$

□

In the next lemma we establish that y converges in probability to $\tilde{\sigma}$.

Lemma 3. *For any $\varepsilon > 0$ there exists a $\hat{\mu} > 0$ such that for any $\mu < \hat{\mu}$ we have that*

$$P\left(\lim_{t\to\infty} |y^t - \tilde{\sigma}| > \varepsilon\right) = 0.$$

Proof. First, note that the point $y^t = \tilde{\sigma}$, with $\tilde{\sigma}$ as defined in Proposition 1, solves the equation

$$y^t + 2(1-y^t)\mu \sum_{j:\pi_{1j}\geq\pi_{2j}} \lambda_j f(\pi_j) = y^t - 2y^t\mu \sum_{j:\pi_{1j}<\pi_{2j}} \lambda_j f(\pi_j).$$

Define now the sequence $y_1 = \{y_1^t\}_{t=\hat{t}}^{\infty}$ as follows

$$y_1^t = \begin{cases} y^t & \text{if } y^t \geq \tilde{\sigma} \\ \tilde{\sigma} & \text{otherwise} \end{cases}.$$

Hence, we have that $E_0(y^t) \leq E_0(y_1^t)$ for all $t > \hat{t}$. Note that if $y^t > \tilde{\sigma}$ then we have that $E_0(y^{t+1}) < E_0(y^t)$. This implies that $E_0(y_1^{t+1}) < E_0(y_1^t)$ for all $y_1^t > \tilde{\sigma}$ and $E_0(y_1^{t+1}) = E_0(y_1^t)$ for $y_1^t = \tilde{\sigma}$. Therefore, y_1 is a super-martingale with lower-bound $\tilde{\sigma}$. Thus, by the Martingale convergence theorem, $\lim_{t\to\infty} y_1^t$ exists. Given that $E_0(y_1^{t+1}) < E_0(y_1^t)$ for all $y_1^t > \tilde{\sigma}$ and $E_0(y_1^{t+1}) = E_0(y_1^t)$ for $y_1^t = \tilde{\sigma}$, we must have that $\lim_{t\to\infty} y_1^t = \tilde{\sigma}$. This implies that y_1 converges in probability to $\tilde{\sigma}$.

Define now the sequence $y_2 = \{y_2^t\}_{t=\hat{t}}^{\infty}$ as follows:

$$y_1^t = \begin{cases} y^t & \text{if } y^t \leq \tilde{\sigma} \\ \tilde{\sigma} & \text{otherwise} \end{cases}.$$

Hence, we have that $E_0(y^t) \geq E_0(y_1^t)$ for all $t > \hat{t}$. Note that if $y < \tilde{\sigma}$ then we have that $E_0(y^{t+1}) > E_0(y^t)$. This implies that $E_0(y_2^{t+1}) > E_0(y_2^t)$ for all $y_2^t < \tilde{\sigma}$ and $E_0(y_2^{t+1}) = E_0(y_2^t)$ for $y_2^t = \tilde{\sigma}$. Therefore, y_2 is a sub-martingale with upper-bound $\tilde{\sigma}$. Thus, by the Martingale convergence theorem, $\lim_{t\to\infty} y_2^t$ exists. Given that $E_0(y_2^{t+1}) > E_0(y_2^t)$ for all $y_2^t < \tilde{\sigma}$ and $E_0(y_2^{t+1}) = E_0(y_2^t)$ for $y_1^t = \tilde{\sigma}$, we must have that $\lim_{t\to\infty} y_2^t = \tilde{\sigma}$. This implies that y_2 converges in probability to $\tilde{\sigma}$.

Hence, we have that for any $\varepsilon > 0$ exists a $\hat{\mu}$ such that for all $\mu < \hat{\mu}$,

$$\begin{aligned} P\left(\lim_{t\to\infty} |y_1^t - \tilde{\sigma}| > \varepsilon\right) &= 0 \\ P\left(\lim_{t\to\infty} |y_2^t - \tilde{\sigma}| > \varepsilon\right) &= 0. \end{aligned}$$

We know, given the definition of y, that for any $\varepsilon > 0$ there exists a $\hat{\mu} > 0$ and a $t > \bar{t}$ such that for any $\mu < \hat{\mu}$ and $h > t$ we have that $Var_t(y^{t+h}) < \varepsilon$. This, together with the fact that $E_0(y^t) \leq E_0(y_1^t)$ and $E_0(y^t) \geq E_0(y_1^t)$ for all $t > \hat{t}$ implies that for all $t > \max\{\bar{t}, \hat{t}\}$ we must have that $\lim_{t\to\infty} y^t = \tilde{\sigma}$. This implies that y converges in probability to $\tilde{\sigma}$. □

Now we are able to prove the result in Proposition 1.

Proof of Proposition 1. We know from Lemma 2 that σ_1 converges in probability to y. From Lemma 3 we also know that y converges in probability to $\tilde{\sigma}$. Hence, we must have that σ_1 converges in probability to $\tilde{\sigma}$. This is the result of the Proposition. □

Proof of Proposition 2

Whenever σ_1^t is arbitrarily close to 0 we have that

$$\sigma_1^{t+1}|_j = \sigma_1^t(1 + g(\pi_{2j}) - g(\pi_{1j})) + o(\sigma_1^t).$$

Define $\gamma_j = 1 + g(\pi_{2j}) - g(\pi_{1j})$ for all $j \in \{1, \ldots, m\}$. Hence, given that g is increasing, we have that $\gamma_i \leq 1 < \gamma_j$ if and only if $\pi_{1i} \geq \pi_{2i}$ and $\pi_{1j} < \pi_{2j}$. We can approximate the equation for the evolution of the sequence σ_1 when σ_1^t is arbitrarily close to 0 as follows:

$$\sigma_1^{t+1}|_j = \gamma_j \sigma_1^t.$$

Lemma 4. *For any $\bar{\sigma}_1^t \in (0,1)$ and any $\varepsilon > 0$ there exists a $\sigma_1^t < \bar{\sigma}_1^t$ and a $\bar{k} \in \mathbb{N}$ such that for $k > \bar{k}$*

$$P\left(\left|\sigma_1^{t+k} - \hat{\sigma}_1^{t+k}\right| > \varepsilon\right) = 0,$$

where $\hat{\sigma}_1^{t+\bar{k}} = \sigma_1^{t+\bar{k}}$ and

$$\hat{\sigma}_1^{t+k+1} = \begin{cases} \gamma_1 \sigma_1^{t+k} & \text{with probability } \lambda_1 \\ \vdots & \\ \gamma_m \sigma_1^{t+k} & \text{with probability } \lambda_m \end{cases}$$

for $k > \bar{k}$.

Proof. Given that the transition matrix P is irreducible and aperiodic and that the number of states is finite, we have the standard result that the empirical distribution of states converges to the limiting distribution of states. This can be rewritten as: for any $\delta > 0$ there exists a $\bar{k}(\delta) \in \mathbb{N}$ such that for $k > \bar{k}(\delta)$,

$$P\left(\left|\frac{\sum_{t=0}^{k} \mathbb{1}_{\{s^t = j\}}}{k+1} - \lambda_j\right| > \delta\right) = 0 \tag{21}$$

for all $j \in \{1, \ldots, m\}$.

We have seen before that if σ_1^t is arbitrarily close to 0 we can write $\sigma_1^{t+1}|j = \gamma_j \sigma_1^t$. In other words, for any $\kappa > 0$ there exists a $\bar{\sigma}_1(\kappa) \in (0,1)$ such that if $\sigma_1^t < \bar{\sigma}_1(\kappa)$ then

$$P\left(\left|\sigma_1^{t+1}|_j - \gamma_j \sigma_1^t\right| > \kappa\right) = 0$$

for all $j \in \{1, \ldots, m\}$. This result can also be expressed as follows. For any $\kappa > 0$ and any $k \in \mathbb{N}$ there exists a $\bar{\sigma}_1(\kappa) \in (0,1)$ such that if $\sigma_1^t < \bar{\sigma}_1(\kappa)$ then

$$P\left(\left|\sigma_1^{t+k+1}|_j - \gamma_j \sigma_1^{t+k}\right| > \kappa\right) = 0. \tag{22}$$

Hence, we have the following two facts. First, the probability of a state being realized a sufficiently far way number of periods converges to the limiting distribution of the Markov chain. Second, that $\sigma_1^{t+1}|j$ behaves as $\gamma_j \sigma_1^t$ if σ_1^t is sufficiently small. Then, for k sufficiently large and σ_1^t sufficiently close to 0 we have that for all $j \in \{1, \ldots, m\}$, $\sigma_1^{t+k+1} = \gamma_j \sigma_1^{t+k}$ with probability λ_j. In other words, combining the results in equations (21) and (22) we can write that for all $\varepsilon > 0$ there exists a $\bar{k}(\varepsilon) \in \mathbb{N}$ and $\bar{\sigma}_1(\varepsilon) \in (0,1)$, such that for all $k > \bar{k}(\varepsilon)$ and $\sigma_1^t < \bar{\sigma}_1(\varepsilon)$ we have that

$$P\left(\left|\sigma_1^{t+k} - \hat{\sigma}_1^{t+k}\right| > \varepsilon\right) = 0,$$

where $\hat{\sigma}_1^{t+\bar{k}} = \sigma_1^{t+\bar{k}}$ and

$$\hat{\sigma}_1^{t+k+1} = \begin{cases} \gamma_1 \sigma_1^{t+k} & \text{with probability } \lambda_1 \\ \vdots & \\ \gamma_m \sigma_1^{t+k} & \text{with probability } \lambda_m \end{cases}$$

for $k > \bar{k}$. □

Lemma 5. *The sequence σ_1 cannot converge to 0 if*

$$\sum_{j=1}^{m} \lambda_j \log \gamma_j > 0.$$

There is a positive probability that the sequence σ_1^t converges to 0 if

$$\sum_{j=1}^{m} \lambda_j \log \gamma_j < 0.$$

Proof. Reasoning as in the proof of Lemma 1 in Ellison and Fudenberg (1995), the sequence σ_1 can converge to zero if and only if the sequence $y = \log \sigma_1$ can converge to $-\infty$. Using again the proof from Lemma 1 in Ellison and Fudenberg (1995) and Lemma 4 in this appendix, the sequence y can converge to $-\infty$ only if $\sum_{j=1}^{m} \lambda_j \log \gamma_j < 0$. The result follows. □

To study the situation in which the process is arbitrarily close to 1, we proceed as follows. First, we define $w^t = 1 - \sigma_1^t$. Then we apply the analysis above to the variable w^t. Define $\hat{\gamma}_j = 1 + g(\pi_{2j}) - g(\pi_{1j})$. Then we have that for all $\varepsilon > 0$ there exists a $\bar{k} \in \mathbb{N}$ and $\bar{w} \in (0,1)$ such that for all $k > \bar{k}$ and $w^t < \bar{w}$ we have that

$$P\left(\left|w^{t+k} - \hat{w}^{t+k}\right| > \varepsilon\right) = 0,$$

where $\hat{w}^{t+\bar{k}} = w^{t+\bar{k}}$ and

$$\hat{w}^{t+k+1} = \begin{cases} \hat{\gamma}_1 w^{t+k} & \text{with probability } \lambda_1 \\ \vdots & \\ \hat{\gamma}_m w^{t+k} & \text{with probability } \lambda_m \end{cases}$$

for $k > \bar{k}$.

An analogous to Lemma 5 when σ_1^t is close to 1 is the following:

Lemma 6. *The sequence σ_1 cannot converge to 1 if*

$$\sum_{j=1}^{m} \lambda_j \log \hat{\gamma}_j > 0.$$

There is a positive probability that the sequence σ_1 converges to 1 if

$$\sum_{j=1}^{m} \lambda_j \log \hat{\gamma}_j < 0.$$

Summing up the results from lemmas 5 and 6 the result in Proposition 2 follows.