Learning Strict Nash Equilibria through Reinforcement

Antonella Ianni[*][†]

Economics Division, University of Southampton

February 2012

Abstract

This paper studies the analytical properties of the reinforcement learning model proposed in Erev and Roth (1998), also termed cumulative reinforcement learning in Laslier et al (2001). This stochastic model of learning in games accounts for two main elements: the *law of effect* (positive reinforcement of actions that perform well) and the *law of practice* (the magnitude of the reinforcement effect decreases with players' experience).

The main results of the paper show that, if the solution trajectories of the underlying replicator equation converge exponentially fast, then, with probability arbitrarily close to one, all the pathwise realizations of the reinforcement learning process will, from some time on, lie within an ε band of that solution. The paper improves upon results currently available in the literature by showing that a reinforcement learning process that has been running for some time and is found sufficiently close to a strict Nash equilibrium, will reach it with probability one.

[*] Address for Correspondence: Economics Division, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK, e-mail: A.Ianni@soton.ac.uk, tel: ++44 (0) 2380 592536.

[†] A previous version of the paper appeared as EUI Working Paper 2007/21. The author acknowledges the ESRC for financial support under Research Grant R00022370 and the support of the MIUR under the program 'Rientro dei Cervelli'.

Keywords: Learning, Law of Effect, Power Law of Practice, Strict Nash Equilibrium, Replicator Dynamics.

JEL Classification Numbers: C72, C92, D83.

1 Introduction

Over the last two decades there has been a growing body of research within the field of experimental economics aimed at analyzing learning in games. Various learning models have been fitted to the data generated by experiments with the aim of providing a learning based foundation to classical notions of equilibrium. The family of stochastic learning theories known as positive reinforcement seem to perform particularly well in explaining observed behaviour in a variety of interactive settings. Although specific models differ, the underlying idea of these theories is that actions that performed well in the recent past will tend to be adopted with higher probability by individuals who repeatedly face the same interactive environment. Despite their wide applications, however, little is known on the analytical properties of this class of learning models.

Consider for example a normal form game that admits a strict Nash equilibrium. Suppose players have almost learned to play that equilibrium, meaning that they have been playing for some time and their behaviour is close to that equilibrium prescription. Since the equilibrium is *strict*, any unilateral deviation will necessarily lead to lower payoffs. One would hence expect players to consistently reinforce their choice of their equilibrium action and, by this doing, to eventually learn to play that Nash equilibrium. This seems to be a basic requirement for a learning theory. Yet, it is not satisfied by some reinforcement learning models (e.g. the Cross model as studied in Börgers and Sarin (1997)), and most results available to date can only guarantee that in some reinforcement learning models, it may (e.g. the Erev and Roth model analyzed in Hopkins (2002), Beggs (2005) and Laslier et al. (2001)). This paper complements this literature by providing sufficient conditions under which a strict Nash equilibrium is reached with probability one.

We study the stochastic reinforcement learning model introduced by Roth and Erev (1995) and Erev and Roth (1998), also termed *cumulative proportional reinforcement* in Laslier et al. (2001). In this model, there is a finite number of players who are to repeatedly play a normal form game with strictly positive payoffs. At each round of play, players choose actions probabilistically, in a way that accounts for two main features. The first effect (labelled the *Law of Effect*) is the positive reinforcement of the probability of choosing actions that have been played in the previous round of play, as a function of the payoff they led to. The second effect (labelled the *Law of Practice*) is that the magnitude of this reinforcement is endogenously decreasing over time.

The main results of this paper show that, if players have been learning for sufficiently long, and if play is found close to a strict Nash equilibrium of the underlying game, then players will learn to play it with probability one. While doing so, they will in fact choose actions in a way that is close to a deterministic multi-population replicator dynamics. The latter dynamics have been studied extensively in biology, as well as in economics. The reinforcement learning process we model offers a micro-foundation for replicator dynamics, by showing that they provide a good approximation of the stochastic process of learning that players use to update their action choices. Specifically, our results exploit the fact that in proximity of a strict Nash equilibrium, convergence of the deterministic replicator dynamics occurs at an exponentially fast rate. As in our learning process the step size decreases endogenously, over time (due to the *Law of Practice*), we are able to define a timescale over which the stochastic component of the reinforcement learning process, which in principle could move the process away from any equilibrium, is in fact overcome by this deterministic effect.[1]

The results we obtain rely on stochastic approximation techniques (Ljung (1978), Arthur et al. (1987), (1988), Benaim (1999)) to establish the close connection between the reinforcement learning process and the underlying deterministic replicator equation. We show that, up to an error term, the behaviour of the stochastic process is well described by a system of discrete time difference equation of the replicator type (Lemma 4). The main result (Theorem 1) shows that if the trajectories of the underlying system of replicator equations converge sufficiently fast and if the learning process has been going on for sufficiently long, then the probability that all the pathwise realizations of the learning process over a given spell of time, possibly infinite, lie within a given small distance of the solution path of the

[1] A more detailed account of this logic is offered in Section 4.

replicator dynamics, becomes arbitrarily close to one. The property of fast convergence, as required in the main result, is always satisfied in proximity[2] of a strict Nash equilibrium of the underlying game (Remark 2) and is sufficient to guarantee that the approximation error converges uniformly over any spell of time.

A number of recent studies emphasize the fact that the deterministic replicator dynamics act as a driving force for several stochastic reinforcement learning process (Börgers and Sarin (1997)), Laslier et al. (2001), Hopkins (2002), Beggs (2005)). These results are very compelling, in that they can be used to approximate the dynamics of these learning processes over any finite time interval. For example, Laslier et al (2001), Lemma 1, applies results from Benaïm (1999) to show that the replicator dynamics act as an asymptotic-pseudo-trajectories of the learning process. Since multi-population replicator dynamics are pulled towards asymptotically stable Nash Equilibria, these findings allow to show that the probability that the stochastic reinforcement process gets absorbed in any such state is strictly positive. This is surprising, as Nash behaviour yields even in an environment that imposes very minimal informational and computational requirements on players.

The limit of this approach is, however, that it provides only a partial characterization: as convergence does not necessarily obtain with probability one, it might very well be that the long run behaviour of the learning process is dramatically different from its finite time approximation. This is stressed, for example, in the analytical study of the Cross learning model of Börgers and Sarin (1997), and is validated by the simulations presented in Izquierdo et al. (2007). The results we obtain in this paper improve upon these findings by showing that, for the reinforcement learning model we study, the approximation in terms of replicator dynamics is suitable to describe its transient behaviour over finite time spells, as well as asymptotically. A direct implication is that we are able to identify sufficient conditions under which Nash behaviour obtains with probability one.

A fruitful line of research, alternative to ours, to address general properties of convergence to

[2] More precisely, within an open subset of the basin of attraction of a strict Nash equilibrium, under deterministic multi-population replicator dynamics.

Nash equilibria of reinforcement learning models is to rule out convergence to all the other rest points of the replicator dynamics. As Hopkins and Posch (2005) note, this heuristic approach raises significant issues and can only be done on an ad hoc basis, typically for very simple games (see therein references for further clarifications on this issue). Relative to the above logic, our results provide a more direct and more general way to achieve the aim.

The paper is organized as follows. Section 2 describes the reinforcement learning model we study. Section 3 states the main result of this paper. Since the logic followed in the proof is more general and could fruitfully be applied to the study of other learning models, an explicit outline is provided in Section 4. Detailed proofs are instead contained in the Appendix. Finally, Section 5 contains some concluding remarks.

2 The model

Consider an N-player, M-action normal form game $G \equiv (\{i = 1, ..., N\}; A^i; \pi^i)$, where $A^i = \{j = 1, ..., M\}$ is player i's action space and $\pi^i : \prod_{l=1}^{N} A^l \equiv A \rightarrow \Re$ is player i's payoff function[3]. Given a strategy profile $a \in A$, we denote by $\pi^i(j, a_{-i})$ the payoff to player i when (s)he chooses action j and all other players play according to a_{-i}, where the subscript $-i$ refers to all players other than i. Throughout the paper we assume that payoffs are strictly positive.

We shall think of player i's behaviour as being characterized by urn i, an urn of infinite capacity containing γ^i balls, $b^i_j > 0$ of which are of colour $j \in \{1, 2, ..., M\}$. Clearly $\gamma^i \equiv \sum_j b^i_j > 0$. We denote by $x^i_j \equiv b^i_j / \gamma^i$ the proportion of colour j balls in urn i. Player i behaves probabilistically in the sense that we take the composition of urn i to determine i's action choices and postulate that x^i_j is the probability with which player i chooses action j.

Behaviour evolves over time in response to payoff consideration in the following way. Let $x^i_j(n)$ be the probability with which player i chooses action j at step $n = 0, 1, 2....$ Suppose

[3] We hereby assume that each player's action space has exactly the same cardinality (i.e. M). This is purely for notational convenience.

that $(j, a_{-i}(n))$ is the profile of actions played at step n and $\pi^i(j, a_{-i}(n))$, shortened to $\pi^i_j(n)$, is the corresponding payoff gained by player i who chose action j at step n. Then, exactly $\pi^i_j(n)$ balls of colour j are added to urn i at step n. At step $n+1$ the resulting composition of urn i, will be:

$$x^i_k(n+1) \equiv \frac{b^i_k(n+1)}{\gamma^i(n+1)} = \frac{b^i_k(n) + \sigma^i_k(n)}{\gamma^i(n) + \sum_l \sigma^i_l(n)} \tag{1}$$

where $\sigma^i_k(n) = \pi^i_j(n)$ for $k = j$ (i.e. if action j is chosen at step n) and zero otherwise, and $l = 1, 2, ...M$. Although the interpretation in terms of urns is novel, the model is not: in the terminology of Roth and Erev (1995) the $b^i_k(\cdot)$ are called propensities, and, since $\gamma^i(n+1) = \gamma^i(0) + \sum_{r=1,\ldots,n} \sum_l \sigma^i_l(r)$, this learning process is termed cumulative reinforcement learning in Laslier et al. (2001).

The above new urn composition reflects two facts: first the proportion of balls of colour j (vs. $k \neq j$) increases (vs. decreases) from step n to step $n+1$, formalizing a positive (vs. negative) reinforcement for action j (vs. action k), and second, since γ^i appears at the denominator, the strength of the aforementioned reinforcement is decreasing in the total number of balls in urn i. It is usual to label the first effect as the *Law of Effect (reinforcement)* and the second as the *Law of Practice.*

To better understand the micro-foundation of this learning model, it is instructive to rewrite (1), by recalling that $b^i_j(n) \equiv x^i_j(n)\gamma^i(n)$, as:

$$\begin{aligned} x^i_j(n+1) &= x^i_j(n)\left[1 - \frac{\pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)}\right] + \frac{\pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)} \\ x^i_k(n+1) &= x^i_k(n)\left[1 - \frac{\pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)}\right] \quad \text{for } k \neq j \end{aligned} \tag{2}$$

where j denotes the action chosen at step n.

These equations show that conditional upon the strategy profile $a(n)$ being played at step n, player i updates her choice probabilities to $x^i(n+1)$ by taking a weighted average of $x^i(n)$ and a unit vector that puts mass one on action j (the chosen one) and mass zero on any other action k. Step n weights depend on step n realized payoff and on step n total number of balls contained in urn i [4].

Since the relative effect of payoffs on action choices becomes smaller as players gain more experience in the learning routine, gains decrease endogenously. Since payoffs are random, so are the updated weights given to payoffs experienced at any given point in time. Furthermore, since different players may get different streams of payoffs over time, each player's learning process may display a different sequence of decreasing gains.

Given an initial condition, $[\gamma(0), x(0)]$, for any $n > 0$, the above choice probabilities define a stochastic process over the state space $[x(n), \gamma(n)]$, described by the following system of $N(M+1)$ stochastic difference equations:

$$\begin{cases} x^i_k(n+1) = x^i_k(n) + \frac{[\sigma^i_k(n) - x^i_k(n)\sum_l \sigma^i_l(n)]}{\gamma^i(n+1)} \\ \gamma^i(n+1) = \gamma^i(n) + \sum_l \sigma^i_l(n) \end{cases} \quad i = 1, ..., N \quad k = 1, ..., M \tag{3}$$

Clearly $\gamma \equiv [\gamma^i] \in \Re^N_+$ and $x^i \equiv [x^i_k] \in \Delta_i \equiv \{x^i \in \Re_+ : \sum_j x^i_j = 1\}$, $x \in \Delta \equiv \times_i \Delta_i$, i.e. x lies in the Cartesian product of the N unit simplexes Δ_i. It can be easily checked that, conditional upon a realization of $a(n)$ the system of equations (3) reproduces exactly the system of equations (2). Note that, by construction, the process is Markovian in the state variables $[x(n), \gamma(n)]$.

In analogy with Hopkins and Posch (2005), we re-label the system in terms of new variables $\mu^i(n) \equiv n^{-1}\gamma^i(n)$ and re-write the dynamics as a process with a constant step size, equal to n^{-1}. This leads to an $N(M+1)$ system, entirely analog to (3), in the new state variables $[x(n), \mu(n)]$:

$$\begin{cases} x^i_k(n+1) = x^i_k(n) + \frac{1}{n}\mu^i(n)[\sigma^i_k(n) - x^i_k(n)\sum_l \sigma^i_l(n)] \\ \mu^i(n+1) = \mu^i(n) + \frac{1}{n}\mu^i(n)[1 - \mu^i(n)\sum_l \sigma^i_l(n)] \end{cases} \tag{4}$$

[4] The system of equations (2) carries a direct analogy with Börgers and Sarin (1997)'s reinforcement model, where payoffs are assumed to be positive and strictly less than one and the payoff player i gets by playing action j is taken to represent exactly the weights given to the unit vector in the above formulation. In their model the weights do not depend on the step number n and, as a result, the formulation of their model only accounts for the *Law of Effect*.

Let $\Im\{n\}$ denote the sigma algebra generated by$\{x(l); \mu(l)\ l = 1, ..., n\}$. Consider the term in square brackets in the first of the equations in (4) and compute its expected value conditional on $\Im\{n\}$. It is not difficult to see that $E[\sigma_k^i(n) \mid \Im\{n\}] = x_k^i(n)\pi_k^i(n)$, i.e., it is the expected payoff to player i from playing action k at step n (given other players' choices). By the same token, $E[\sum_l \sigma_l^i(n) \mid \Im\{n\}] = \sum_l x_l^i(n)\pi_l^i(n)$, i.e., it is the expected payoff to player i at step n. As a result:

$$\begin{aligned} E[\sigma_k^i(n) - x_k^i(n)\sum_l \sigma_l^i(n) \quad \mid \quad \Im\{n\}] &= \\ &= x_k^i(n)(\pi_k^i(n) - \sum_l x_l^i(n)\pi_l^i(n)) \equiv f_k^i(x(n)) \end{aligned} \tag{5}$$

where the term of the right hand side of this equation defines a (discrete time) system of deterministic replicator dynamics. Its continuous time version $f(x) : \Delta \to \Delta$ is defined by:

$$\frac{d}{dt}x^i(t) = f_k^i(x) \equiv x_k^i(\pi_k^i - \sum_l x_l^i\pi_l^i) \tag{6}$$

and is a direct generalization of the Taylor (1979) multi-population replicator dynamics. It has been extensively studied in the literature on evolution, usually in the context of large population and random matching models (see for ex. Fudenberg and Levine (1998), Weibull (1995), Benaim and Weibull (2003) and references therein) and has been applied to the study of learning models by Börgers and Sarin (1997)), Posch (1997), Ianni (2007), Hopkins (2002), Vega-Redondo (2003), among others.

Lemma 2 in Hopkins and Posch (2005) shows that, under the assumption that payoffs are strictly positive, the ODEs associated with these dynamics are $f^M(x) : \Delta \cup \Re^+ \to \Delta \cup \Re^+$ defined by:

$$\begin{cases} \frac{d}{dt}x^i(t) = \mu^i f_k^i(x) \\ \frac{d}{dt}\mu^i(t) = \mu^i(1 - \mu^i \sum_l x_l^i\pi_l^i) \end{cases} \tag{7}$$

and it can be easily seen that for $\mu^i = (\sum_l x_l^i\pi_l^i)^{-1}$ (i.e., for stationary values of μ^is):

$$\frac{d}{dt}x^i(t) = \frac{f_k^i(x)}{\sum_l x_l^i\pi_l^i} \tag{8}$$

which corresponds exactly to the Maynard Smith replicator dynamics (see, for example, Weibull (1995), and Beggs (2005)), where the excess payoff of action k over the average payoff to player i is renormalized by the latter quantity.

It is clear that the two types of replicator dynamics share the same rest points and it is known (see Hopkins (2002) and Hopkins and Posh (2005)) that the two dynamics share the same set of asymptotically stable rest points. It is also known (see Ritzberger and Weibull (1995)) that a stationary point is asymptotically stable in the Taylor's replicator dynamics if and only if it is a strict Nash equilibrium of the underlying game. What has not been exploited in the literature to date is an additional property of strict Nash equilibria under replicator dynamics: namely, the fact that different trajectories started within the basin of attraction of a strict Nash equilibrium, not only converge, but do so exponentially fast. This property will prove to be key to our results.

In the next Sections we shall state the main result we obtain and outline the logic of its construction.

3 The main result

Let $I = \{n_l \mid l \geq 0\}$ be a collection of indices such that $0 < n_0 < n_1 < .... < n_l <$ ..... Let $x(n_0), x(n_1), ...x(n_l), ...$ denote a realization of the stochastic process (3) at steps $n_0, n_1, ..., n_l, ...$ Introduce the following fictitious time scale: let $t_l = \sum_{k=n_0,n_{l-1}} k^{-1}$ and $\Delta t_l = t_{l+1} - t_l$. Consider the collection of points $\{(x(n_l), t_l) \mid n_l \in I\}$. Suppose also that the solution of the system of differential equations (6), started at time t_0 with initial condition equal to $x(n_0)$ is plotted against the same time scale.

The main result of this paper estimates the probability that all points $x(n_l)$ for $n_l \in I$ simultaneously are within a given distance ε from the trajectory of the solution of the system of differential equations. In words, Theorem 1 shows that, if and whenever, the solutions of the system of differential equations (6) converge sufficiently fast, there exists constants $\overline{\varepsilon}, \overline{n}$

that depend on the payoffs of the game, such that, for $\varepsilon < \overline{\varepsilon}$ and $n_0 > \overline{n}$, the probability that all pathwise realizations of the process in I simultaneously lie in an ε-band of the trajectory of the ODE, becomes arbitrarily close to one.

Theorem 1 *Consider the stochastic learning process x defined by system (3). Suppose payoffs of the underlying game are strictly positive. Let the system of ODE (6) denote a system of deterministic replicator dynamics and $x(t, t_0, x)$ denote any time $t \geq 0$ solution, when the initial condition is taken to be x at time t_0. Suppose that the following property holds over a compact set $D \subseteq \Delta$:*

$$|x(t + \Delta t, t, x + \Delta x) - x(t + \Delta t, t, x)| \leq (1 - \lambda \Delta t) |\Delta x| \tag{9}$$

with $0 < \lambda < 1$ and $|.|$ denoting the Euclidean norm.

Then there exists constants $C, \overline{\varepsilon}, \overline{n}$ that depend on the game, such that, for $\varepsilon < \overline{\varepsilon}$, $n_0 > \overline{n}$ and $x(n_0) \in D$:

$$\Pr\left[\sup_{n_l \in I} |x(n_l) - x(t_{n_l}, t_{n_0}, x(n_0))| > \varepsilon\right] \leq \frac{C}{\varepsilon^2} \sum_{j=n_0}^{\overline{N}} \frac{1}{j^2} \tag{10}$$

for $n_l \in I = \{n_0, n_1, ... \overline{N}\}$, where $\overline{N} = \sup_I n_l$.

The Theorem shows that the learning process stays close to the corresponding trajectory of the replicator dynamics with higher probability as n_0 increases, for a given ε.

The first important thing to notice is that, since the right hand side of inequality (10) is square summable, the statement holds for any $\overline{N}$, possibly infinite. This amount to saying that, under the assumptions of the Theorem, the reinforcement learning process is stochastically approximated, to an arbitrarily high degree of precision[5], by a replicator dynamics over any interval of the form $[t, +\infty[$. This property is known not to hold for a learning process à la Cross as, there, the step size does not decrease endogenously.

[5] In the terminology of Benaim (1999) and applied in Laslier et al. (2001), the replicator dynamics constitutes a.s. a *limit trajectory* (and not only an *asymptotic-pseudo-trajectory)* for the process.

The second important thing to notice is that the result holds, *after* some n, and it would not, in general, be correct to infer any global convergence property. Global convergence for models à la Erev and Roth could be established for games with a unique strict Nash equilibrium along the lines of Laslier et al. (2001) and Hopkins and Posch (2005).

Next, condition (9) is shown to hold for any strict Nash equilibrium of the underlying game:

Remark 2 *Let x^* be a strict Nash equilibrium of $\mathcal{G}$ and denote its basin of attraction by:*

$$B(x^*) \equiv \{x \in \Delta \mid \lim_{t\to\infty} x(t, t_0, x) = x^*\}$$

Then there exists an open set $B_r(x^*) \equiv \{x \in \Delta \mid |x - x^*| < r\} \subseteq B(x^*)$ such that condition (9) stated in Theorem 1 holds in $B_r(x^*)$.

A straightforward implication of the above Remark is that if the stochastic process is found, at some step of its dynamics, in a suitably defined neighbourhood of a strict Nash equilibrium, then the probability with which the process converges to that Nash equilibrium can be made arbitrarily close to one.

Remark 3 *Let x^* be a strict Nash equilibrium of $\mathcal{G}$ and suppose $x(n_0) \in B_r(x^*)$, defined in Remark 2. Then, for $n > n_0$:*

$$\lim_{n\to\infty} \Pr[x(n) = x^* \mid x(n_0) \in B_r(x^*)] = 1$$

Proof For $\overline{N} = \infty$ in inequality (10) reads:

$$\Pr\left[\sup_{n_l \in I} |x(n_l) - x(t_{n_l}, t_{n_0}, x(n_0))| \leq \varepsilon\right] \geq 1 - \frac{C}{\varepsilon^2} \sum_{j=n_0}^{\infty} \frac{1}{j^2}$$

Hence:

$$\lim_{n_0\to\infty} \Pr\left[\sup_{n_l \in I} |x(n_l) - x(t_{n_l}, t_{n_0}, x(n_0))| \leq \varepsilon\right] \geq 1 - \lim_{n_0\to\infty} \frac{C}{\varepsilon^2} \sum_{j=n_0}^{\infty} \frac{1}{j^2} = 1$$

■

4 An outline of the Proof

The technical proofs of the main results are contained in the Appendix. In this Section, we remark on the key ingredient that is novel, i.e. the use of the notion of *exponential stability* as applied to our non linear time varying system, and we outline the logic we follow.

We say that an equilibrium $x = 0$ is exponentially stable for $\frac{d}{dt}x(t) = f(x(t))$ if there exists positive constants, c, k and γ, independent of the initial condition t_0, such that $\mid x(t) \mid \leq k \mid x(t_0) \mid \exp[-\gamma(t-t_0)]$ for all $t \geq t_0 \geq 0$ and for all $\mid x(t_0) \mid < c$. It can be shown (see Khalil (1996)) that, for linear systems, this requirement is equivalent to asymptotic stability of the solution, holding uniformly with respect to the initial condition[6]. Exponential stability is what will allow us to extend the approximation result to the infinite interval. To see this, consider any two solution trajectories, labelled as $y(t)$ and $z(t)$, with initial conditions $y(t_0)$ and $z(t_0)$ respectively. If the solutions are exponentially stable, then the following is an upper bound to the time t distance between the two trajectories:

$$\mid y(t) - z(t) \mid \leq \mid y(t_0) - z(t_0) \mid k \exp[-\gamma(t-t_0)] + \beta$$

where k, γ and β are positive constants. This bound is valid also on infinite time intervals (for $t \to \infty$)[7].

The main result relies on a series of Lemmas. Lemma 4 allows us to re-write the process x

[6] We say that a solution $x = 0$ is stable if for each $\varepsilon > 0$, there is $\delta = \delta(\varepsilon, t_0)$ such that $\mid x(t_0) \mid < \delta \Rightarrow \mid x(t) \mid < \varepsilon$ for all $t \geq t_0 \geq 0$. We say that a solution $x = 0$ is asymptotically stable if it is stable and there is $c = c(t_0) > 0$ such that $x(t) \to 0$ as $t \to \infty$ for all $\mid x(t_0) \mid < c$. We say that a solution is uniformly asymptotically stable if c does not depend on t_0, i.e. if for each $\varepsilon > 0$ there is $T = T(\varepsilon) > 0$ such that $\mid x(t) \mid < \varepsilon$ for all $t \geq t_0 + T(\varepsilon)$ and for all $\mid x(t_0) \mid < c$. These definitions are standard and can be found for example in Khalil (1996) p. 134.

[7] Note that this bound is tighter than the one obtained by applying the Gronwall-Bellman inequality:

$$\mid y(t) - z(t) \mid \leq \mid y(t_0) - z(t_0) \mid \exp[L(t-t_0)] + \frac{\delta}{L}\{\exp[L(t-t_0)] - 1\}$$

where $\delta > 0$ and L is the Lipschitz constant. This latter bound is valid only on compact time intervals, since the exponential term grows unbounded for $t \to \infty$.

as:

$$x^i(j(n)) = x^i(n) + \sum_{s=n}^{j(n)-1} \frac{1}{s} f^i(x(s)) + \sum_{s=n}^{j(n)-1} \varepsilon^i(x(s), s)$$

for $j(n) \geq n+1$, where the last term can be made arbitrarily small by an appropriate choice of n, since it is the difference between two converging martingales.

Lemma 5 then proceeds to show that if the process is, at step n of its dynamics, within a small ρ-neighbourhood of some value x, then it will remain within a ρ-neighbourhood of x for some time after n. As such, Lemma 5 provides information about the local behaviour of the stochastic process $x(.)$ around x, by characterizing an upper bound to the spell of re-scaled time within which the process stays in a neighbourhood of x.

The intuition used to extend the results runs as follows. Suppose time t realization of the process, x, belongs to some interval A. Within a time interval Δt two factors determine the subsequent values of the process: a) the deterministic part of the dynamics, i.e. the functions $f(x(t))$ started with $f(x(t))$ in A and b) the stochastic component. If the trajectories of $f(x)$ converge, then after this time interval, $f(x(t+\Delta t))$ will be in some interval $B \subset A$, for all x that started in A. Exponential stability guarantees that the distance between any two such trajectories will decrease over this time interval, the more so, the longer is the time interval. According to Lemma 5, the realization of the stochastic process will differ from the corresponding trajectories by a small quantity, say $\pm C$, the more so, the smaller is the time interval. Hence the stochastic process will not diverge from its deterministic counterpart if $B + 2C \leq A$. In order for this to hold, the time interval Δt needs to be large enough to let the trajectories of the deterministic part converge sufficiently, but small enough to limit the stochastic effect. To this aim, Lemma 6 shows that if the realization of our process $x(.)$ lies within ε distance from the corresponding trajectory of $x(.)$ at time n_l, then this will also be true at time n_{l+1}, provided ε is small enough to guarantee that Δt_l is

a) big enough for any two trajectories of $x(.)$ to converge sufficiently, and

b) small enough to limit second order effects and the effects of the noise.

To conclude the proof of Theorem 1 it is then sufficient to estimate the probability that Lemma 5 holds simultaneously for all n_l.

5 Conclusions

This paper studies the analytical properties of a reinforcement learning model that incorporates the Law of Effect (positive reinforcement of actions that perform well), as well as the Law of Practice (the magnitude of the reinforcement effect decays over repetitions of the game). The learning process models interaction, among a finite set of players faced with a normal form game, that takes place repeatedly over time. The main contribution to the literature relies on the full characterization of the asymptotic paths of the learning process in terms of the trajectories of a system of replicator dynamics applied to the underlying game. Regarding the asymptotics of the process, the paper shows that if the reinforcement learning model is found, after sufficiently many steps, in a neighbourhood of a strict Nash equilibrium, then convergence to that equilibrium will take place with probability arbitrarily close to one. As for the dynamics of the process, the results show that, from some time on, any realization of the learning process will be arbitrarily close to the trajectory of the replicator dynamics started with the same initial condition.

The convergence result we obtain relies on two main facts: first by explicitly modelling the Law of Practice, we are able to construct a fictitious time scale over which any realization of the process can be studied; second, the observation that whenever the solution of the system of replicator dynamics converge exponentially fast, the deterministic part of the process drives the stochastic dynamics. Both requirements are shown to be essential to establish the result.

We conclude with two further remarks. First, since the methodology we used is not specific to the reinforcement learning model analyzed in this paper, it could be fruitfully applied to the study of different learning models. This is an intriguing prospect in relation to the analogies between fictitious play and a perturbed version of reinforcement learning, identified in Hopkins (2002), or to the study of the Experience Weighted Attraction model proposed in Camerer et al. (1999). Second, and more technically, we conjecture that an alternative sufficient condition to achieve the results we obtain in this paper could rely on modelled fast convergence properties of the learning algorithm (for example a sequence of weights given by $[\gamma(n)]^{-p}$ for $p > 1$), rather than on those of the underlying deterministic dynamics (i.e. the properties of the $f(x(n))$. Although conceptually this would amount to considering different learning models for which the micro-foundations used in this paper would not directly apply, the results of Benaïm (1999) on *shadowing* do support this conjecture.

Appendix

Lemma 4 *Consider the reinforcement learning model defined by (3) and suppose that* $x(0) > 0$ *component-wise, and for all i's and for all* $a \in A$, $0 < \underline{\pi} \leq \pi^i(a) \leq \overline{\pi} < \infty$.

Then the following holds:

$$\begin{cases} x_k^i(n+1) = x_k^i(n) + \frac{1}{n} f_k^i(x(n)) + \varepsilon_k^i(n) & n \geq 1 \\ 0 < x_k^i(0) < 1 & n = 0 \end{cases}$$

where $f_k^i(x(n))$ *is defined in (5) and:*

$$\Pr[\lim_{n\to\infty} \sum_{k=n}^{\infty} \varepsilon_k^i(k) = 0] = 1$$

for all $i = 1, ..., N \quad$ *and* $k = 1, ..., M$ *and* $n \geq 1$.

Proof. Simple algebra shows that the dynamics is defined by:

$$\begin{cases} x_k^i(n+1) = x_k^i(n) + \frac{1}{n} \Phi_k^i(n) & n \geq 1 \\ 0 < x_k^i(0) < 1 & n = 0 \end{cases} \tag{11}$$

for all $i = 1, ..., N \quad$ and $k = 1, ..., M$, where:

$$\Phi_k^i(n) \equiv [\sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n)] + \delta_k^i(n) \tag{12}$$

with:

$$\delta_k^i(n) \equiv (\mu^i(n) - 1)[\sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n)]$$

We then study the conditional expectation $E[\Phi_k^i(n) \mid \Im\{n\}]$ by looking at the two additive components separately. As already mentioned, simple algebra shows that:

$$E[\sigma_k^i(n) - x_k^i(n) \sum_k \sigma_k^i(n) \quad | \quad \Im\{n\}] = f_k^i(x(n))$$

Also, since:

$$\begin{aligned} \mu^i(n) &\leq \overline{\pi} \\ \sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n) \quad &\leq \quad \sigma_k^i(n) \leq \overline{\pi} \end{aligned}$$

it follows that, for all i and for all k:

$$\mid \delta_k^i(n) \mid \leq \frac{1}{n}[\overline{\pi}]^2$$

As a result, we can now write:

$$x_k^i(n+1) = x_k^i(n) + \frac{1}{n} f_k^i(x(n)) + \varepsilon_k^i(n)$$

where:

$$\begin{aligned} \varepsilon_k^i(n) \quad &= \quad \frac{1}{n}[\delta_k^i(n) + \eta_k^i(n)] \\ \eta_k^i(n) \quad &\equiv \quad \Phi_k^i(n) - E[\Phi_k^i(n) \mid \Im\{n\}] \end{aligned}$$

For $n \geq 2$, for $\Xi(0) \equiv 0$, and for each given i, k we then construct:

$$\Xi(n) \equiv \sum_{l=1}^{n-1} \varepsilon_l^i(l) \equiv \Xi_\delta(n) + \Xi_\eta(n)$$

Note that:

$$\begin{aligned} \Xi_\delta(n+1) \quad &= \quad \Xi_\delta(n) + \frac{1}{n}\delta_k^i(n) \\ \Xi_\eta(n+1) \quad &= \quad \Xi_\eta(n) + \frac{1}{n}\eta_k^i(n) \end{aligned}$$

and since δ_k^i is bounded by $\frac{1}{n}[\overline{\pi}]^2$, it follows that:

$$\Xi_\delta(n+1) \leq \Xi_\delta(n) + \frac{\overline{\pi}^2}{n^2}.$$

Hence, we can construct an auxiliary stochastic process:

$$Z(n) \equiv \Xi_\delta(n) + \overline{\pi}^2 \sum_{k \geq n} \frac{1}{k^2}$$

where the series of which in the second term converges, and show that this is a supermartingale relative to $\Im\{n\}$. By the convergence theorem for supermartingales, there exists a random variable $Z(\infty)$ and, for $n \to \infty$, $Z(n)$ converges pointwise to $Z(\infty)$ with probability one. Hence, also $\Xi_\delta(n)$ converges to $\Xi_\delta(\infty)$ with probability one.

With regard to $\Xi_\eta(n)$, since $E[\eta_k^i(n) \mid \Im\{n\}] = 0$, $\Xi_\eta(n)$ is a quadratically integrable martingale relative to $\Im\{n\}$. Hence (see for ex. Karlin and Taylor (1975), p. 282), there exists a random variable $\Xi_\eta(\infty)$ and $\Xi_\eta(n) \to \Xi_\eta(\infty)$ for $n \to \infty$ a.s..

Since $\Xi(\infty) - \Xi(n) \equiv \sum_{l=n}^{\infty} \varepsilon_k^i(l)$, the assert follows. ■

Lemma 5 *Consider the reinforcement learning model defined by (3) under the assumptions of Lemma 4. Define the number $m(n, \Delta t)$ such that*

$$\lim_{n\to\infty} \sum_{k=n}^{m(n,\Delta t)-1} \frac{1}{k} = \Delta t$$

Assume that, for $\rho = \rho(x') > 0$ and sufficiently small, $x(n) \in \mathcal{B}(x', \rho) = \{x : |x - x'| < \rho\}$. Then there exists a value $\Delta t_0(x', \rho)$ and a number $N_0 = N_0(x', \rho)$ such that, for $\Delta t < \Delta t_0$ and $n > N_0$, $x(k) \in \mathcal{B}(x', \rho)$ for all $n \leq k \leq m(n, \Delta t)$.

Proof. By Lemma 4, for $j(n) \geq n + 1$, the process can be re-written as:

$$x(j(n)) = x(n) + \sum_{s=n}^{j(n)-1} \frac{1}{s} f(x') + \sum_{s=n}^{j(n)-1} \frac{1}{s}[f(x(s)) - f(x')] + \sum_{s=n}^{j(n)-1} \varepsilon(s)$$

and an upper bound for $x(j(n))$ can be constructed as follows.

Since the function f is Lipschitz in x:

$$\sum_{s=n}^{j(n)-1} \frac{1}{s} |f(x(s)) - f(x')| \leq L \max_{n\leq k\leq j(n)-1} \mid x(k) - x' \mid \sum_{s=n}^{j(n)-1} \frac{1}{s}$$

where L is global Lipschitz constant. Hence, by letting $\Delta t(n, j(n)) \equiv \sum\limits_{s=n}^{j(n)-1} s^{-1}$ we obtain:

$$|x(j(n))| \leq |x(n)| + \Delta t(n, j(n)) \mid f(x') \mid +$$

$$+\Delta t(n,j(n))L \max_{n\le k\le j(n)-1} \mid x(k)-x' \mid + \qquad (13)$$

$$+\left|\sum_{s=n}^{j(n)-1} \varepsilon(s)\right|$$

As for the last term, from Lemma 4 we know that, for all $\alpha > 0$ there exists an $n = n(\alpha)$ such that for all $n > n(\alpha)$ with probability one:

$$\left|\sum_{s=n}^{j(n)-1} \varepsilon(s)\right| \le \alpha$$

since these are differences between converging martingales.

Now consider $j(n) = m(n, \Delta t)$, where m is such that $\lim_{n\to\infty} \Delta t(n, m(n, \Delta t)) = \Delta t$. Note that the number m is finite for any n and for any $\Delta t < \infty$, since $\sum_s s^{-1} = \infty$ and $\sum_s s^{-2} < \infty$. Denote $\left|\sum_{s=n}^{j(n)-1} \varepsilon(s)\right|$ by $\alpha(n)$ and suppose $x(k) \in \mathcal{B}(x', 2\rho)$ for all $n \le k \le m(n, \Delta t)-1$. Inequality (13) states that:

$$|x(m)| \le |x(n)| + \Delta t \mid f(x') \mid + \Delta t 2\rho L + \alpha(n)$$

Hence:

$$\begin{aligned} |x(m)-x'| &\le |x(m)-x(n)| + |x(n)-x'| \\ &\le \Delta t\,|f(x')| + \Delta t 2L\rho + \alpha(n) + \rho \end{aligned}$$

and, as a result, we can choose $N_0(\rho) = n(\frac{\rho}{2})$ such that, for all $n > N_0, \alpha(n) < \frac{\rho}{2}$ and $\Delta t_0(x', \rho) = \frac{\rho}{2}(|f(x')| + 2L\rho)^{-1} > 0$ and show that, for all $\Delta t < \Delta t_0$ and $n > N_0$:

$$|x(m)-x'| \le \frac{\rho}{2} + \frac{\rho}{2} + \rho = 2\rho$$

Hence if $x(k) \in \mathcal{B}(x', 2\rho)$ for all $n \le k \le m-1$, this implies that also $x(m) \in \mathcal{B}(x', 2\rho)$. By induction it then follows that $x(k)$ remains in $\mathcal{B}(x', 2\rho)$ also for all k up to $m(n, \Delta t) - 1$. ■

Lemma 6 *Beyond the assumptions of Lemma 5, suppose that the system of ODE (6) satisfies property (9) on a compact set $D \subseteq \Delta$. Suppose $x(n_l) \in D$ with probability one, and $x_0(l) \in D$.*

Then,

$$\text{if } |x_0(l) - x(n_l)| \leq \varepsilon, \text{ also } \left|x_0(l+1) - x(n_{l+1})\right| \leq \varepsilon$$

for $\frac{\lambda\varepsilon}{2L} \leq \Delta t_l \leq \frac{3\lambda\varepsilon}{2L}$, where $0 < \lambda < 1$, L is the Lipschitz constant of $f(.)$ on D, and $0 < \varepsilon < \overline{\varepsilon} = \min\{\sqrt{(6\lambda^2)^{-1}4\rho L}, (3\lambda)^{-1}2L\overline{\Delta t_0}\}$ with $\overline{\Delta t_0} = \inf_{x \in D, \rho=\rho(x)} \Delta t_0(x, \rho) > 0$ defined in Lemma 5.

Proof.

Let $I = \{n_l \mid l \geq 0\}$ be a collection of indices such that $0 < n_0 < n_1 < .... < n_l < n_{l+1} < .....$ and let $\Delta t_l = t_{l+1} - t_l$, with $t_l = \sum_{k=n_0}^{n_{l-1}} k^{-1}$. Lemma 4 states that the value of the process at time n_{l+1} is given by:

$$x(n_{l+1}) = x(n_l) + \Delta t_l f(x(n_l)) + \alpha(n_l)$$

and Lemma 5 shows that, for Δt_l small and n_l large, $\alpha(n_l) < \rho/2$, meaning that if the process is started at $x(n_l)$, it stays close to it for some time.

Solve the system of differential equations (6) from t_l to $t_l + \Delta t_l$ Since $f(.)$ is Lipschitz continuous:

$$|x(t + \Delta t, t, \overline{x}) - (\overline{x} + \Delta t f(\overline{x}))| \leq L\Delta t^2$$

where $x(t + \Delta t, t, \overline{x})$ denotes the solution at time $t + \Delta t$, when the initial condition is taken to be $\overline{x}$ at time t and L is a constant.

Now take $x(n_l) = \overline{x}$ and compute the distance between the stochastic process at step n_{l+1}, $x(n_{l+1})$, and the differential equation at time t_{l+1}, with initial condition $\overline{x}$ at time t_l, $x(t_{l+1}, t_l, \overline{x})$ shortened to $x_l(l+1)$:

$$\begin{aligned} |x(n_{l+1}) - x_l(l+1)| &= |x(n_l) + \Delta t_l f(x(n_l)) + \alpha(n_l) - x_l(l+1)| \\ &\leq L\Delta t^2 + \alpha(n_l) \end{aligned}$$

As a result:

$$\begin{aligned}\left|x_0(l+1) - x(n_{l+1})\right| &\leq |x_0(l+1) - x_l(l+1)| + |x_l(l+1) - x(n_{l+1})| \\ &\leq |x_0(l+1) - x_l(l+1)| + L\Delta t_l^2 + \alpha(n_l) \qquad (14)\end{aligned}$$

where the first term is the distance between two trajectories of the ODE, one started at $x(n_0)$ and one at $x(n_l)$ at time t_0 and t_l respectively, and the second term is the distance between the ODE and the stochastic process at time t_{l+1}. We know from Lemma 5 that the last two terms on the right hand side of (14) can be made arbitrarily small by an appropriate choice of Δt_l and n_l. We also know that, if the two trajectories of which in the first term of the right hand side of (14) converge, their distance will become increasingly small. An assumption that is sufficient to establish the result that follows requires:

$$|x(t+\Delta t, t, x+\Delta x) - x(t+\Delta t, t, x)| \leq (1-\lambda\Delta t)\,|\Delta x| \qquad (15)$$

with $0 < \lambda < 1$. If property (9) holds, then:

$$|x_0(l+1) - x_l(l+1)| \leq (1-\lambda\Delta t_l)\,|x_0(l) - x(n_l)|$$

and as a result, inequality (14) can be rewritten as:

$$\left|x_0(l+1) - x(n_{l+1})\right| \leq (1-\lambda\Delta t_l)\,|x_0(l) - x(n_l)| + L\Delta t_l^2 + \alpha(n_l) \qquad (16)$$

We can now show that, if $x(n_l)$ lies in an ε-neighbourhood of the trajectory of the ODE, so will $x(n_{l+1})$, for a suitable choice of ε and Δt.

Under the assumptions of this Lemma, inequality (16) yields:

$$\left|x_0(l+1) - x(n_{l+1})\right| \leq (1-\lambda\Delta t_l)\varepsilon + L\Delta t_l^2 + \alpha(n_l)$$

By Lemma 5 $\alpha(n_l) < r(\varepsilon) \equiv \frac{\lambda^2\varepsilon^2 3}{4L} < \frac{\rho}{2}$, which holds for $0 < \varepsilon < \sqrt{\frac{4\rho L}{6\lambda^2}}$ as assumed. Hence:

$$\begin{aligned}(1-\lambda\Delta t_l)\varepsilon + L\Delta t_l^2 + \alpha(n_l) &\leq \varepsilon - \lambda\Delta t_l\varepsilon + L\Delta t_l^2 + \frac{\lambda^2\varepsilon^2 3}{4L} \\ &= \varepsilon + L\left[\left(\Delta t_l - \frac{\lambda\varepsilon}{2L}\right)\left(\Delta t_l - \frac{3\lambda\varepsilon}{2L}\right)\right] < \varepsilon\end{aligned}$$

as stated.

We also need to show that for $\lambda\varepsilon(2L)^{-1} \leq \Delta t_l \leq 3\lambda\varepsilon(2L)^{-1}$, Δt_l also satisfies Lemma 5, i.e. $\Delta t_l < \Delta t_0(x,\rho)$ for all $x \in D$. The radius ρ depends on x and is a measure of how fast $f(x)$ changes in a neighbourhood of x. Since $f(x)$ is Lipschitz and D is compact, this radius will have a positive lower bound, as x moves in D. Let this be $\overline{\rho} > 0$. Hence:

$$\overline{\Delta t_0} = \inf_{x\in D} \Delta t_0(x) \equiv \inf_{x\in D} \left(\frac{\overline{\rho}}{2[|f(x)| + 2L\overline{\rho}]} \right) > 0$$

and since $\varepsilon < (3\lambda)^{-1} 2L\overline{\Delta t_0}$ by assumption, the assert follows. ■

Proof of Theorem 1

To proof the Theorem we need to estimate the probability that Lemma 5 holds for all $n_l \in I$. To this aim note that:

$$\Pr\left[\sup_{n_l\in I} |x(n_l) - x_0(l)| \leq \varepsilon\right] = \Pr\left[\sup_{n_l\in I} \alpha(n_l) < r(\varepsilon)\right]$$

where, as before $x_0(l) \equiv x(t_l, t_0, x(n_0))$.

From Lemma 5:

$$\alpha(n_l) \equiv |\varepsilon(n_l)| \equiv \left| \sum_{l=n_0}^{n_l} \varepsilon(l) - \sum_{l=n_0}^{n_{l-1}} \varepsilon(l) \right|$$

and from Lemma 4:

$$E[\varepsilon^i_k(l)] \leq \frac{\overline{\pi}^2}{l^2}$$

As a result:

$$\begin{aligned} \alpha(n_l) &\leq \sqrt{NM} \sup_i \sup_k \varepsilon^i_k(l) \leq \sqrt{NM}\frac{\overline{\pi}^2}{n_l^2} \\ E[\alpha(n_l)] &\leq \sqrt{NM}\frac{\overline{\pi}^2}{n_l^2} \end{aligned}$$

By Chebyshev's inequality:

$$\Pr[\alpha(n_l) > r(\varepsilon)] \leq \frac{\sqrt{NM}}{r(\varepsilon)} \frac{\overline{\pi}^2}{n_l^2}$$

Hence:

$$\Pr[\alpha(n_l) \geq r(\varepsilon); n_l > n_0, n_l \in I] \leq \frac{C}{\varepsilon^2} \sum_{j=n_0}^{\overline{N}} \frac{1}{j^2}$$

where $C = (3\lambda^2)^{-1} 4L\sqrt{NM}\overline{\pi}^2$ since $r(\varepsilon) \equiv 3(4L)^{-1}\lambda^2\varepsilon^2$. In the statement of the theorem $\overline{n} = N_0(\rho)$, defined in Lemma 5 and $\overline{\varepsilon} = \min\{(3\lambda)^{-1}2L, \sqrt{(6\lambda^2)^{-1}4\rho L}\}$ as from Lemma 6. ■

Proof of Remark 2

To prove the statement we need to show that every strict Nash equilibrium satisfies condition (9), i.e.:

$$|x(t + \Delta t, t, x + \Delta x) - x(t + \Delta t, t, x)| \leq (1 - \lambda\Delta t) |\Delta x| \tag{17}$$

This condition holds if the system of ODE (6) admits the following quadratic Ljapunov function (see, for example, Ljung (1977)):

$$V(\Delta x, t) = |\Delta x|^2 \tag{a}$$

$$\frac{d}{dt} V(\Delta x, t) < -C |\Delta x|^2 \quad C > 0 \tag{b}$$

Suppose x^* is a strict Nash equilibrium and w.l.g. let $x^* = 0$. Consider the linearization of the system (6) around $x^* = 0$:

$$\frac{d}{dt} x(t) = Ax + g(x)$$

where $A \equiv Df(x)|_{x^*=0}$ denotes the Jacobian matrix of $f(x)$ at x^* and $\lim_{x \to 0} \frac{g(x)}{|x|} = 0$. From Ritzberger and Weibull (1995), Proposition 2, we know that a Nash equilibrium is asymptotically stable in the replicator dynamics if and only if it is strict. Hence we also know that all the eigenvalues of A at x^* have negative real part and we can consider the following scalar product in $\Re^{NM}$:

$$\langle x, y \rangle = \int_0^\infty (e^{At}x, e^{At}y) dt$$

and choose:

$$V(x, t) = \langle x, x \rangle$$

which satisfies condition (a). This scalar product also satisfies condition (b), since:

$$\frac{d}{dt}V(x,t) \leq -\left|x\right|^2 + 2\left\langle x, g(x)\right\rangle \leq -\left|x\right|^2 + 2\sqrt{\langle x, x\rangle}\sqrt{\langle g(x), g(x)\rangle}$$

By the equivalence of norms in $\Re^N$, there exists a $c > 0$ s.t. $\sqrt{\langle x, x\rangle} \leq c\left|x\right|$. For $r > 0$, consider an open ball $B_r = \{x \in \Delta : |x| < r\}$ such that $B_r \subset D$ and $|g(x)| \leq (1/(4c^2))\left|x\right|$ in B_r. Then:

$$\frac{d}{dt}V(x,t) \leq -\left|x\right|^2 + 2c^2\left|x\right|\left|g(x)\right| \leq -\frac{1}{2}\left|x\right|^2 \leq -\frac{1}{2c^2}V(x,t) \text{ in } B_r$$

which shows that condition (b) holds. ■

REFERENCES

Arthur, W.B. (1993), “On designing economic agents that behave like human agents," *Journal of Evolutionary Economics,* **3**, 1-22.

Arthur, W.B. Yu., M. Ermoliev and Yu. Kaniovski (1987), “Non-linear Urn Processes: Asymptotic Behavior and Applications," *mimeo,* IIASA WP-87-85.

Arthur, W.B. Yu., M. Ermoliev and Yu. Kaniovski (1988), “Non-linear Adaptive Processes of Growth with General Increments: Attainable and Unattainable Components of Terminal Set.," *mimeo,* IIASA WP-88-86.

Beggs, A.W. (2005), “On the Convergence of Reinforcement Learning.," *Journal of Economic Theory,* **122**, 1-36.

Benaim, M. (1999), *“Dynamics of Stochastic Approximation,* Le Seminaire de Probabilite', Springer Lecture Notes in Mathematics.

Benaim, M and J. Weibull (2003), “Deterministic Approximation of Stochastic Evolution in Games," *Econometrica,* **71**, 873-903.

Benveniste, A., Metivier, M. and P. Priouret (1990), *“Adaptive Algorithms and Stochastic Approximation,* . Springer-Verlag.

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

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

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

Cross, J.G. (1983), *“A Theory of Adaptive Economic Behavior,* . Cambridge: Cambridge University Press.

EREV, I. AND A.E. ROTH (1998), "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review,* **88(4)**, 848-881.

FUDENBERG D. AND D. LEVINE (1998), *"Theory of Learning in Games, .* MIT Press.

HOPKINS, E. (2002), "Two competing models of how people learn in games," *Econometrica,* **70**, 2141-2166.

HOPKINS, E. AND M. POSCH (2005), "Attainability of boundary points under reinforcement learning," *Games and Economic Behavior,* **53**, 110-125.

IANNI, A. (2007), "Learning Strict Nash Equilibrium through Reinforcement," *mimeo,* EUI Working Paper 2007/21.

IZQUIERDO, L.R., IZQUIERDO, S.S., GOTTS, N.M. AND J.G. POLHILL (2007), "Transient and asymptotic dynamics of reinforcement learning in games," *Games and Economic Behavior,* **61**, 259-276.

KHALIL, H.K. (1996), *"Nonlinear Systems, .* Prentice Hall.

LASLIER, J.F., TOPOL R. AND B. WALLISER (2001), "A Behavioral Learning Process in Games," *Games and Economic Behavior,* **37**, 340-366.

LJUNG, L. (1978), "Strong Convergence of a Stochastic Approximation Algorithm," *Annals of Statistics,* **6**, 680-696.

POSH, M. (1997), "Cycling in a stochastic learning algorithm for normal form games," *Journal of Evolutionary Dynamics,* **7**, 193-207.

RITZBERGER K. AND J. WEIBULL (1995), "Evolutionary Selection in normal form games," *Econometrica,* **63**, 1371-1399.

Roth, A. and I. Erev (1995), “Learning in Extensive Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior,* **8(1)**, 164-212.

Taylor, P. (1979), “Evolutionary stable strategies with two types of player," *Journal of Applied Probability,* **16**, 76-83.

Vega-Redondo, F. (2003), *“Economics and the Theory of Games,* . Cambridge University Press.

Weibull J. (1995), *“Evolutionary Game Theory,* . MIT Press.