European University Institute

Department of Economics

EUI Working Papers

ECO 2007/21

Learning Strict Nash Equilibria through Reinforcement

Antonella Ianni

EUROPEAN UNIVERSITY INSTITUTE

DEPARTMENT OF ECONOMICS

Learning Strict Nash Equilibria through Reinforcement

ANTONELLA IANNI

EUI Working Paper **ECO** 2007/21

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 Strict Nash Equilibria through Reinforcement[1].

Antonella Ianni[2]

Universita' Ca' Foscari di Venezia, University of Southampton (U.K.), E.U.I.

This version: July 2007

[1] A previous version of the paper was titled Reinforcement Learning and the Power Law of Practice: Some Analytical Results. The author aknowledges the ESRC for financial support under Research Grant R00022370 and the support of the MIUR under the program 'Rientro dei Cervelli'.

[2] Address for Correspondence: Department of Economics, European University Institute, Via della Piazzuola 43, 50133 Florence, Italy.

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). The stochastic model of learning accounts for two main elements: the Law of Effect (positive reinforcement of actions that perform well) and the Power Law of Practice (learning curves tend to be steeper initially).

The paper establishes a relation between the learning process and the underlying deterministic replicator equation. The main results show that if the solution trajectories of the latter converge sufficiently fast, then the probability that all the realizations of the learning process over a given spell of time, possibly infinite, becomes arbitrarily close to one, from some time on. In particular, the paper shows that the property of fast convergence is always satisfied in proximity of a strict Nash equilibrium.

The results also provide an explicit estimate of the approximation error that could prove to be useful in empirical analysis.

JEL: **C72, C92, D83.**

1 Introduction

Over the last decade 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 the stochastic learning process is started in a neighbourhood of it. Since, for each player, any action different from the equilibrium action will necessarily lead to lower payoffs, one would expect players to consistently reinforce their choice of the 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 B\"orgers 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 studies 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 the 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 Power Law of Practice) is that the magnitude of this reinforcement is endogenously decreasing over time. The main results of this paper imply that, if players start close to a strict Nash equilibrium of the underlying game, from some time onwards and with probability one, they will learn to play it. While doing so, players will in fact choose actions in a way that is close to a deterministic replicator dynamics. The latter dynamics have been studied extensively in biology, as well as in economics, and it is known that all, and only those, strict Nash equilibria are their stable rest points. Our results exploit the fact that in proximity of a strict Nash equilibrium, convergence occurs at an exponentially fast rate. If learning has been going on for some time, the stochastic component of the reinforcement learning process, which in principle could move the process away from the equilibrium, is in fact overcome by this deterministic effect.

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. Specifically the paper shows 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 2). The main result of the paper (Theorem 1) shows that if the trajectories of the underlying system of replicator equations converge sufficiently fast, then the probability that all the realization of the learning process over a given spell of time, possibly infinite, lie within a given small distance of the solution path of the replicator dynamics, becomes arbitrarily close to one, from some time on. In particular, the paper shows that the property of fast convergence, as required in the main result, is always satisfied in proximity of a strict Nash equilibrium of the underlying game (Remark 1) and is sufficient to guarantee that the approximation error converges uniformly over any spell of time.

Based on the primitives of the model, the result also offers an explicit estimate of the approximation error, say α, i.e. the probability that all the realization of the learning process are more than ε away from the solution trajectory of a replicator dynamics started with the same initial conditions. Hence, for a given (α, ε), one can compute an estimate of the number of repetitions, say n_0, that are needed in order to guarantee that with probability at least $(1-\alpha)$ all the realizations at any step $n > n_0$ lie within ε distance from the replicator dynamics. This estimate could prove to be useful in empirical analysis, as an alternative to the simulation of learning behaviour, which is typically done on the basis of a law of large numbers and involves averaging over the realizations of play of thousands of simulated players.

The paper is organized as follows. Section 2 describes the reinforcement learning model we study. Section 3 introduces the main result of this paper, which is then stated in Section 4. 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 (Detailed proofs are instead contained in the Appendix). Finally, Section 5 contains some concluding remarks.

2 The model

This paper studies the reinforcement learning process introduced by Roth and Erev (1995), and Erev and Roth (1998), also referred to as cumulative proportional reinforcement learning in Laslier et al. (2001).

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 $\pi^i : \times_i A^i \equiv A \rightarrow \Re$ is player i's payoff function[1]. Given a strategy profile $a \equiv (a_1, ..., a_i, ..., a_N) \in A$, we denote by $\pi^i(a)$ the payoff to player i when a is played. For a given player i, we conventionally denote a generic profile of action a as (a_i, a_{-i}) where the subscript $-i$ refers to all players other than i. Hence $\pi^i(j, a_{-i})$ is the payoff to player i when (s)he chooses action j and all other players play according to a_{-i}. Throughout the paper we assume that payoffs are non negative and bounded.

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 that $a(n) \equiv [j, a_{-i}(n)]$ 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 $l = 1, 2, ...m$. In the terminology of Erev and Roth (1995) the $b^i_k(\cdot)$ are called propensities. Since $\gamma^i(n+1) = \gamma^i(0) + \sum_{r=1,...,n} \sum_l \sigma^i_l(r)$, this learning process is termed cumulative reinforcement learning in Laslier et al. (2001).

If payoffs are positive and bounded (as will be assumed throughout) 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. We label the first effect as *reinforcement* and we refer to the second as the *law of practice.*

To better understand the microfoundation of this learning model, it is instructive to rewrite (1) for j being the action chosen at step n and by recalling that $b^i_j(n) \equiv x^i_j(n)\gamma^i(n)$, as:

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

This shows that conditional upon $a(n)$ being played at step n, player i updates her state by taking a weighted average of her old state and a unit vector that puts mass one on action j, where step n weights depend positively on step n realized payoff and negatively on step n total number of balls contained in urn i^2.

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:

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

Clearly $\gamma \equiv [\gamma^i] \in \Re_+^N$ and $x^i \equiv [x_k^i] \in \Delta_i \equiv \{x^i \in \Re_+^m : \sum_j x_j^i = 1\}$, $x \in \Delta \equiv \times_i \Delta_i$, 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 $[x(n), \gamma(n)]$ (and time inhomogeneous, since $\gamma(n)$ depends on n).

Let $\Im\{n\}$ denote the sigma algebra generated by$\{x(l); \gamma(l) \ l = 1, ..., n\}$. Consider the term in square brackets in equation (3) and compute its expected value conditional on $\Im\{n\}$. It is not difficult to see that $E[\sigma_k^i(n) \mid \Im\{n\}] = x_k^i(n) \sum_{a_{-i}} \pi^i(k, a_{-i}) x_{a_{-i}}(n)$ i.e. it is the expected payoff to player i from playing action k at step n, when all other players are choosing each profile a_{-i} with probability $x_{a_{-i}}(n)$. Since analogous reasoning applies to $\sum_l \sigma_l^i(n)$, one gets:

$$\begin{aligned} E[[\sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n)] \quad | \quad \Im\{n\}] &= \\ &= x_k^i(n)[\sum_{a_{-i}} \pi^i(k, a_{-i}) x_{a_{-i}}(n) - \sum_a \pi^i(a_i, a_{-i}) x_a^i(n)] \end{aligned} \tag{4}$$

where the term of the RHS of this equation defines a (discrete time) system of deterministic replicator dynamics. Its continuous time version $f(x^D) : \Delta \rightarrow \Delta$ defined by $\frac{d}{dt} x^D(t) = f(x^D(t))$ where, for $i = 1, ..., N$ and $k = 1, ..., m$

$$f_k^i(x^D) \equiv x_k^i[\sum_{a_{-i}} \pi^i(k, a_{-i}) x_{a_{-i}}^i - \sum_a \pi^i(a_i, a_{-i}) x_a^i] \tag{5}$$

is a direct generalization of the Taylor (1979) multipopulation replicator dynamics. It has been extensively studied in the literature on evolution, usually in the contest of large population and random matching models (see for ex. Fudenberg and Levine (1998), Ch. 3, Weibull (1996), Ch. 3 and therein references) and has been applied to the study of learning models by Börgers and Sarin (1997)), Posch (1997), Ianni (2002), Hopkins (2002), Vega-Redondo (2003), among others.

This relation is what motivates the study of the dynamic properties of the reinforcement learning process (3) by using known properties of the replicator dynamics (5). In fact, if the step sizes of the reinforcement process, $\gamma^i(n+1)^{-1}$ in (3), were deterministic and equal for all players, for example by a renormalization of the total number of balls in each urn to n that leaves proportions $x(n)$ unaffected, then by arguments that are now standard in the literature on stochastic approximation of Robbins-Monro algorithms, and since replicator dynamics are bounded and Lipschitz-continuous, the $N(m+1)$ system (3) could be approximated by the following Nm system:

$$x^i_k(n+1) = x^i_k(n) + \frac{1}{n} f^i_k(x(n)) + \frac{1}{n}\xi^i(n) + O(\frac{1}{n^2}) \text{ for } i = 1, ..., N \text{ and } k = 1, ..., m$$

where $E[\xi^i(n) \mid \Im\{n\}] = 0$ and $O(n^{-2})$ is a term of the same infinitesimal order of (n^{-2}). This renormalization has been used for example in Arthur (1993), Posh (1997).

In the reinforcement learning model we study in this paper gains decrease endogenously, since the relative effect of payoffs from the interaction on action choices becomes smaller as players gain more experience in the learning routine. 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. However, as we show below, in proximity of an equilibrium towards which the solution trajectories of the replicator dynamics converge sufficiently fast, any renormalization to a common scale would do. In fact, since the property of fast convergence is shown to hold in a neighbourhood of any *strict* Nash equilibrium, we are able to show that, if the learning process is started in its proximity, the probability that any of its realization lie within a small distance from the solution path of the replicator dynamics, over a possibly infinite spell of time, becomes arbitrarily close to one, from some time on. Hence the paper sheds some light on the asymptotics of the reinforcement learning process, as well as on its evolution over time.

3 An overview

Before proceeding to state the main result of this paper, we find it useful to place it in the contest of results already available in the literature on stochastic approximation that have found application to the study of learning dynamics.

First, the results of Arthur et al. (1988) (Theorem 2) applied to our setting guarantee that the learning process converges almost surely to a random vector with support given by the set of rest points of the replicator dynamics[3], i.e. the set:

$$D_R \equiv \{x \in \Delta \mid f(x) = 0\}$$

whenever this consists of isolated points. As it is well known, this set typically includes all the Nash equilibria of the underlying game, as well as all the vertices of the simplex Δ and all the points that are Nash equilibria only with respect to the strategies in their support. Results on 'unattainability' (i.e. convergence with zero probability to a given rest point) within this literature (Arthur et al. (1988), Pemantle (1990)) apply only to interior solutions, and are not of straightforward extension to the boundaries of the simplex Δ (see Hopkins and Posch (2005) for a clarification).

Sufficient conditions that guarantee that the process does not oscillate between different isolated rest points in D_R typically require the existence of a Ljapunov function for the system (5). Theorem 1 of Ljung (1978) or Corollary 6.6 of Benaim (1999) provide convergence conditions, that do straightforwardly apply to our setting whenever a Ljapunov function can be identified. Hence, convergence of the reinforcement learning process obtains for wide classes of underlying games (see Hofbauer and Sigmund (1988) and Weibull (1995) among others on the study of Ljapunov convergence for some classes of games).

Theorem 2 of Ljung (1978) details conditions under which the process converges with probability one to the subset of stable rest points, i.e. the set:

$$\begin{aligned} D_S \equiv \; & \{x \in \Delta \mid f(x) = 0 \text{ and the Jacobian } Df(x) \text{ has} \\ & \text{only eigenvalues with non-positive real part}\} \end{aligned}$$

This set is particularly important in the study of the properties of the reinforcement learning model when applied to an interactive setting, since it consists of all, and only those, strict Nash equilibria of the underlying game. Unfortunately, the result of Ljung (1978) is not easily applicable to our reinforcement learning model (in particular condition D1 cannot be easily checked).

Benveniste et al (1990) also provide a number of interesting results in the theory of stochastic approximation of adaptive algorithms. Although the results we obtain are in line with those of Theorem 14 , Ch. 1, Part 2 of the quoted book, the underlying assumptions are different (in their setup, there is a unique globally stable rest point and the gain sequence, $\gamma(.)^{-1}$, is deterministic and, by assumption, has locally bounded moments).

The types of results available in the literature emphasize the fact that the deterministic replicator dynamics act as a driving force for the stochastic reinforcement learning process, in that it describes its expected motion. How far this relation can be used to understand the asymptotics of the learning process is however, not obvious. The best we can do is in fact to approximate the dynamics of the learning process by replicator dynamics over compact time intervals. This is exactly what is done in Laslier et al (2001), Lemma 1, where, using terminology and results from Benaim (1999), it is shown that the replicator dynamics is an asymptotic-pseudo-trajectory of the learning process. Results of this type are useful in understanding the asymptotics of the reinforcement learning process, since they allow to show that Theorem 7.3 in Benaim (1999) applies and that the probability that the reinforcement learning process gets absorbed in an asymptotically stable Nash Equilibrium is strictly positive. However, to address general properties of convergence to Nash equilibria of reinforcement learning models, one needs to rule out convergence to all the other rest points of the replicator dynamics. While convergence to (linearly) unstable rest points can be ruled out, for example, on the basis of Theorem 5.1. of Benaim and Hirsh (1999), convergence to boundary rest points that are Nash equilibria only with respect to strategies in their support, need also to be ruled out (see Hopkins and Posch (2005) for more on this issue). As it will become clear below, the main result of this paper exploits an additional stability property of strict Nash equilibria under replicator dynamics. By this doing, it improves upon the type of results available in the literature by showing that if the process is started in proximity of a *strict* Nash equilibrium, convergence will obtain with probability arbitrarily close to one.

Before stating the main result of this paper, we remark on two ingredients that are key to its proof. The first refers to the definition of a suitable time scale upon which we construct our numerical estimate of the probability that the learning process is well approximated (in a sense to be made precise later) by a solution trajectory of the replicator dynamics. The second involves the notion of stability of the solution trajectories of the replicator dynamics.

One way to address this issue of different random step sizes of the reinforcement learning process (3) is the one followed by Hopkins (2002), where the author introduces N new variables $\mu^i(n) = n^{-1}\gamma^i(n)$ to re-write the dynamics as a process with a constant (hence common) step size, equal to n^{-1}. This leads to an $N(m+1)$ system analog to (3) in the new state variables $[x(n), \mu(n)]$. It turns out that having a common deterministic time scale is not a necessary requirement for the decomposition of the expected motion of the process (3) in a deterministic part, $f(x(.))$, which denotes system (5), weighted by random sequences $\gamma^i(.)^{-1}$, plus an error term, that is uniformly bounded. Lemma 2 in the Appendix shows that, under the assumption that payoffs are positive and bounded, the stochastic learning process (3) can still be written as:

$$x_k^i(n+1) = x_k^i(n) + \frac{1}{\gamma^i(n)} f_k^i(x(n)) + \varepsilon_k^i(n) \text{ for } i = 1, ..., N \text{ and } k = 1, ..., m$$

where $\varepsilon(n)$ is proved[4] to be a.s. $O(n^{-2})$.

As a result, the fact that in a reinforcement learning model the step sizes are random, does not prevent the application of the techniques of stochastic approximation, in that they do not alter the order of magnitude of the error term. If, however, numerical bounds are to be provided, we need to fix a time scale that is common to all players[5]. For the reasons hinted at before, and summarized in Remark 3 after the proof of Lemma 2 in the Appendix, any sequence $\{g(n), n > 0\}$ is such that $g(n) > 0$, $\sum_n g(n)^{-1} = \infty$ and $\sum_n g(n)^{-2} < \infty$ would satisfy the necessary conditions. We shall take $g(n) = \underline{\gamma(n)} + n\underline{\pi}$ with $\underline{\gamma(n)} \equiv \inf_i \gamma^i(0)$ where, we recall, $\gamma^i(0)$ is the initial number of balls in player i's urn and $\underline{\pi}$ is the minimum payoff achievable in the game. We obtain this by simply re-adjusting each urn composition, $b^i_k(n)$ to $b^{i\prime}_k(n)$ in such a way as to ensure that $x^i_k(n) \equiv b^i_k(n)\gamma^i(n)^{-1} = b^{i\prime}_k(n)g(n)^{-1}$for all $i = 1, ...N$ and for all $k = 1, ..., m$. The dynamics is hence defined by:

$$\begin{cases} x^i_k(n+1) = x^i_k(n) + \frac{1}{g(n)} f^i_k(x(n)) + \varepsilon^i_k(n) \\ \{g(n)\} \end{cases} \quad i = 1, ..., N \quad k = 1, ..., m \tag{6}$$

Second, we shall use the notion of *exponential stability* as applied to our non linear time varying system. We say that an equilibrium $x = 0$ is exponentially stable for $\frac{d}{dt}x^D(t) = f(x^D(t))$ 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 that 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 on the infinite interval. To see this, consider any two solution trajectories of the replicator dynamics, labelled as $y(t)$ and $z(t)$, with initial conditions $y(t_0)$ and $z(t_0)$ respectively. Since the replicator dynamics are Lipschitz-continuous, the application of the Gronwall-Bellman inequality provides the following upper bound to the time t distance between the two trajectories:

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

where $\delta > 0$ and L is the Lipschitz constant. This bound is valid only on compact time intervals, since the exponential term grows unbounded for $t \to \infty$. If, however, the solutions are exponentially stable, then (see for example Khalil (1996), Chapter 5) such bound can be expressed as:

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

where k, γ and β are positive constants. This bound is clearly valid also on infinite time intervals, and this is key to our proof.

In the next Section we shall state the main result of this paper and outline the logic of its construction.

4 The main result

As described below, our result relates the trajectories of the system of replicator dynamics (5) to the asymptotic paths of the reinforcement learning model defined by (3). By doing this, we are able to show that, provided the process is started within the basin of attraction of an asymptotically stable rest point the probability with which such a rest point is reached can be made arbitrarily close to one.

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 the realizations of the stochastic process (3) at steps $n_0, n_1, ..., n_l, ...$. Consider the renormalized process defined by (6) and introduce the following fictitious time scale: let $t_l = \sum_{k=n_0}^{n_{l-1}} g(k)^{-1}$ 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 (5), 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 (5) 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 realizations of the process in I simultaneously lie in an ε-band of the trajectory of the ODE, becomes arbitrarily large, after time $\overline{n}$.

Theorem 1 *Consider the stochastic learning process defined by system (6)). Suppose payoffs of the underlying game are bounded and strictly positive. Let the system of*

ODE (5) denote a system of deterministic replicator dynamics and $x^D(t, t_0, x)$ *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* $D \subseteq \Delta$*:*

$$\left|x^D(t + \Delta t, t, x + \Delta x) - x^D(t + \Delta t, t, x)\right| \leq (1 - \lambda \Delta t)\left|\Delta x\right| \tag{7}$$

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

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

$$\Pr\left[\sup_{n_l \in I}\left|x(n_l) - x^D(t_{n_l}, t_{n_0}, x(n_0)\right| > \varepsilon\right] \leq \frac{C}{\varepsilon^2}\sum_{j=n_0}^{\overline{N}}\frac{1}{g(j)^2} \tag{8}$$

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

The above 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 intuition behind the result is that the common gain sequence $g(.)^{-1}$ of the process can be rescaled in such a way as to guarantee that the process $x(.)$ stays close to $x^D(.)$ with an arbitrary high degree of precision.

An important thing to notice is that, since the RHS of inequality (8) 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, by a replicator dynamics over any interval of the form $[t, +\infty[$[7]. The fact that the approximation holds uniformly over a possibly infinite spell of time has significant implication for the characterization of the asymptotic behaviour of the reinforcement learning process: while being an a.s. asymptotic-pseudo-trajectory of the reinforcement learning process guarantees that the probability of the reinforcement learning process gets absorbed in a linearly stable Nash equilibrium is strictly positive, being a limit trajectory, guarantees that such probability converges uniformly to one.

Next, condition (7) is shown to hold for any strict Nash equilibrium of the underlying game:

Remark 1 *Let x^* be a strict Nash equilibrium of G and denote its basin of attraction by:*

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

Then there exist an open set $B_r \equiv \{x \in \Delta \mid |x - x^*| < r\} \subseteq B(x^*)$ such that condition (7) stated in Theorem 1 holds in B_r.

A straightforward implication of the above Remark is that if the stochastic process is started 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 2 *Let x^* be a strict Nash equilibrium of $\mathcal{G}$ and suppose $x(n_0) \in B_r$, defined in Remark 1. Then*

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

Proof For $\overline{N} = \infty$ in inequality (8) reads:

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

Hence:

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

∎

Since convergence occurs uniformly, the upper bound of which in Theorem 8 could prove to be useful to practically control the approximation error. To this aim, let $\alpha \equiv \Pr\left[\sup_{n_l\in I}\left|x(n_l) - x^D(t_{n_l}, t_{n_0}, x(n_0)\right| > \varepsilon\right]$, i.e. the probability that the learning process is more than ε distant from the solution trajectory of the replicator dynamics, both processes started with the same initial condition. The bound in (8) can then be read to provide a lower bound for the number of steps the learning process needs to go through in order to guarantee that, with probability at least α, the approximation error is less than ε:

$$\frac{\varepsilon^2}{C} \geq \sum_{j=n_0}^{\infty}\frac{1}{g(j)^2} \geq \frac{\alpha\varepsilon^2}{C} \tag{9}$$

The first inequality in (9) guarantees that the bound is operative, meaning that the value on the RHS of inequality (8) is less than one; the second is a direct application of Theorem 8. Notice that:

$$\sum_{j=n_0}^{\infty} \frac{1}{(\underline{\gamma(0)}+j\underline{\pi})^2} = \frac{1}{\underline{\pi}^2}\,\mathrm{Psi}\left(n_0 + \frac{\underline{\gamma(0)}}{\underline{\pi}}, 1\right)$$

where Psi(.,.) is a Polygamma function, which takes strictly positive values, it is continuous and it is strictly decreasing in n_0+ $\underline{\gamma(0)\pi}^{-1}$. Hence the error can be controlled by n_0 and / or by $\underline{\gamma(0)}$. As detailed in the proof of Theorem 8 in the Appendix, the constant C is a function of λ (a measure of the speed at which learning takes place, which is to be computed from the payoffs of the underlying game), of L (the Lipschitz constant, which can be taken to be one, without loss of generality), of N (the number of players in the game), of M (the number of actions available) and of $\overline{\pi}$ (the maximum payoff achievable)[8].

Since the logic followed in the proof of Theorem 1 is quite general, we conclude this Section by sketching its outline.

The main result relies on a series of Lemmas.

As already mentioned, Lemma 2 shows that, whenever payoffs are positive and bounded, the motion of the stochastic system $x^i(n)$ is driven by the deterministic system of $f^i(x(n))$, rescaled by a random sequence $\gamma^i(n)^{-1}$, up to a convergent error term. The key to the proof of convergence is the coupling of the error term with the sum of a supermartingale and a quadratically integrable martingale. Lemma 2 allows us to re-write the process as:

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

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 3 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 3 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 derive global 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 noise 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. Hence 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 3, 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 noise effect. To this aim, Lemma 4 shows that if the realization of our process $x(.)$ lies within ε distance from the corresponding trajectory of $x^D(.)$ 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 $x^D(.)$ 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 3 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 of this paper is 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 started in a neighbourhood of a strict Nash equilibrium, then convergence to that equilibrium takes 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. This also provides a practical way to control the approximation error.

The convergence result we obtain relies on two main facts: first by explicitly modelling the Law of Practice, we are able to construct a fictituous 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 acts as a driving force. 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 peculiar to the reinforcement learning model analyzed in this paper, it could be fruitfully applied to the study of different learning models (for example 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 $f^D x(n)$). Although conceptually this would amount to considering different learning models, the results of Benaim (1999) on *shadowing* do support this conjecture.

Appendix

Lemma 2 *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:

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

where:

$$f_k^i(x(n)) = x_k^i(n)[\sum_{a_{-i}} \pi^i(k, a_{-i}) x_{a_{-i}}(n) - \sum_a \pi^i(a_i, a_{-i}) x_a(n)]$$

and:

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

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

Proof. The dynamics (3) is defined by:

$$\begin{cases} x_k^i(n+1) = x_k^i(n) + \frac{1}{\gamma^i(n)} \Phi_k^i(n) & n \geq 1 \\ 0 < x_k^i(0) < 1 & n = 0 \end{cases} \tag{10}$$

for all $i = 1, ..., N$ and $k = 1, ..., m$, where:

$$\Phi_k^i(n) = [\sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n)] + \delta_k^i(n) \tag{11}$$

with:

$$\delta_k^i(n) \equiv -\frac{1}{\gamma^i(n)} [\sigma_k^i(n) - x_k^i(n) \sum_l \sigma_l^i(n)] \left[\frac{\sum_l \sigma_l^i(n)}{1 + \frac{\sum_l \sigma_l^i(n)}{\gamma^i(n)}} \right]$$

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

$$\begin{aligned} E[[\sigma_k^i(n) - x_k^i(n) \sum_k \sigma_k^i(n)] \quad | \quad \Im\{n\}] &= \\ &= x_k^i(n)[\sum_{a_{-i}} \pi^i(k, a_{-i}) x_{a_{-i}}^i(n) - \sum_a \pi^i(a_i, a_{-i}) x_a^i(n)] \\ &\equiv f_k^i(x(n)) \end{aligned}$$

Also, since:

$$\begin{aligned}
\frac{\sum_l \sigma_l^i(n)}{1+\frac{\sum_l \sigma_l^i(n)}{\gamma^i(n)}} &\leq \sum_l \sigma_l^i(n) \leq \overline{\pi} \\
\sigma_k^i(n) - x_k^i(n)\sum_l \sigma_l^i(n) &\leq \sigma_k^i(n) \leq \overline{\pi}
\end{aligned}$$

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

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

As a result, we can now write:

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

where:

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

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

$$\begin{aligned}
\Xi(n) &\equiv \sum_{l=1}^{n-1} \varepsilon_l^i(l) \\
&\equiv \sum_{l=1}^{n-1} \frac{1}{\gamma^i(l)} \delta_l^i(l) + \sum_{l=1}^{n-1} \frac{1}{\gamma^i(l)} \eta_l^i(l) \\
&\equiv \Xi_\delta(n) + \Xi_\eta(n)
\end{aligned}$$

Note that:

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

and since by construction, δ_k^i is bounded as in eq. (5), it follows that:

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

where $g^i(n) \equiv \gamma^i(0) + n\underline{\pi}$ is deterministic.

Hence, we can construct an auxiliary stochastic process:

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

where the series of which in the second term converges, and show that this is a supermartingale relative to $\Im\{n\}$. In fact:

$$\begin{aligned} E[Z(n+1) \quad | \quad & \Im\{n\}] = \\ &= E[\Xi_\delta(n+1) \mid \Im\{n\}] + \overline{\pi}^2 \sum_{k \geq n+1} \frac{1}{g^i(k)^2} \\ &\leq \Xi_\delta(n) + \overline{\pi}^2 \frac{1}{g^i(n)^2} + \overline{\pi}^2 \sum_{k \geq n+1} \frac{1}{g^i(k)^2} \\ &= \Xi_\delta(n) + \overline{\pi}^2 \sum_{k \geq n} \frac{1}{g^i(k)^2} \equiv Z(n) \end{aligned}$$

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 $E[\eta^i_k(n) \mid \Im\{n\}] = 0$, $\Xi_\eta(n)$ 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 $\Xi(\infty) - \Xi(n) \equiv \sum_{l=n}^{\infty} \varepsilon^i_k(l)$, the assert follows. ∎

Remark 3 *Let Ω^* be a subspace of the sample space of the process $\{x(n), \gamma(n)\}$ such that the assumptions of Lemma 2 hold. For a given initial condition $[x(0), \gamma(0)]$, consider a fixed realization $\omega^* \in \Omega^*$ and the corresponding sequence $\{x(n,\omega^*), \gamma(n,\omega^*)\}$. Any component of the vector $\gamma(n,\omega^*) \equiv [\gamma^i(n,\omega^*),\ i = 1,2,...,N]$, regarded as a sequence over n, satisfies the following:*

$$0 < \frac{1}{\overline{\gamma(0)} + n\overline{\pi}} \leq \frac{1}{\gamma^i(n,\omega^*)} \leq \frac{1}{\underline{\gamma(0)} + n\underline{\pi}} \equiv \frac{1}{g(n)}$$

where $\overline{\gamma(0)} \equiv \sup_i \gamma^i(0)$ *and* $\underline{\gamma(0)} \equiv \inf_i \gamma^i(0)$*. Hence:*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{\gamma^i(n,\omega^*)} &= 0 \\ \sum_{n=0}^{\infty} \frac{1}{\gamma^i(n,\omega^*)} &= \infty \\ \sum_{n=0}^{\infty} \frac{1}{(\gamma^i(n,\omega^*))^2} &< \infty \end{aligned}$$

Lemma 3 *Consider the reinforcement learning model defined by (3) under the assumptions of Lemma 2. Define the number* $m(n,\Delta t)$ *such that*

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

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 2, for $j(n) \geq n+1$, the process can be re-written as:

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

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

Since the function f is Lipschitz in x:

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

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

$$\begin{aligned} |x(j(n))| \leq |x(n)| + \Delta t(n,j(n)) \mid f(x') \mid + \\ +\Delta t(n,j(n)) L \max_{n\leq k\leq j(n)-1} \mid x(k) - x' \mid + \\ + \left| \sum_{s=n}^{j(n)-1} \varepsilon(s) \right| \end{aligned} \qquad (12)$$

As for the last term, from Lemma 2 we know that, for all $\alpha > 0$ there exists an $n = n(\alpha)$ such that for all $n > n(\alpha)$ with probability one:

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

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 $\sum_s g(s)^{-1} = \infty$ and $\sum_s g(s)^{-2} < \infty$ by assumption. 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 $n \leq k \leq m(n, \Delta t) - 1$.

Inequality (12) states that:

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

Hence:

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

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 $n > N_0$:

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

Hence if $x(k) \in \mathcal{B}(x', 2\rho)$ for all $n \leq k \leq m - 1$, 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 4 *Beyond the assumptions of Lemma 3, suppose that the system of ODE (5) satisfies property (7) on a compact set $D \subseteq \Delta$. Suppose $x(n_l) \in D$ with probability one, and $x_0^D(l) \in D$.*

Then,

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

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 3.*

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 $t_l = \sum_{k=n_0}^{n_{l-1}} g(k)^{-1}$. Lemma 2 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 3 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 (5) from t_l to $t_l + \Delta t_l$ Since $f(.)$ is Lipschitz continuous:

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

where $x^D(t + \Delta t, t, \overline{x})$ 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, $x^D(t_{l+1}, t_l, \overline{x})$ shortened to $x_l^D(l+1)$:

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

As a result:

$$\begin{aligned} \left| x_0^D(l+1) - x(n_{l+1}) \right| &\leq \left| x_0^D(l+1) - x_l^D(l+1) \right| + \left| x_l^D(l+1) - x(n_{l+1}) \right| \\ &\leq \left| x_0^D(l+1) - x_l^D(l+1) \right| + L\Delta t_l^2 + \alpha(n_l) \qquad (13) \end{aligned}$$

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 3 that the last two terms on the RHS of (13) 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 RHS of (13) converge, their distance will become increasingly small. An assumption that is sufficient to establish the result that follows requires:

$$\left|x^D(t+\Delta t, t, x+\Delta x) - x^D(t+\Delta t, t, x)\right| \leq (1-\lambda\Delta t)\left|\Delta x\right| \qquad (14)$$

with $0 < \lambda < 1$. If property (7) holds, then:

$$\left|x_0^D(l+1) - x_l^D(l+1)\right| \leq (1-\lambda\Delta t_l)\left|x_0^D(l) - x(n_l)\right|$$

and as a result, inequality (13) can be rewritten as:

$$\left|x_0^D(l+1) - x(n_{l+1})\right| \leq (1-\lambda\Delta t_l)\left|x_0^D(l) - x(n_l)\right| + L\Delta t_l^2 + \alpha(n_l) \qquad (15)$$

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 (15) yields:

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

By Lemma 3 $\alpha(n_l) < r(\varepsilon) \equiv \frac{\lambda^2\varepsilon^2 3}{4L} < \frac{\rho}{2}$, 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 3, 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 Lipschit z 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 3 holds for all $n_l \in I$. To this aim note that:

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

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

From Lemma 3:

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

and from Lemma 2:

$$E[\varepsilon_k^i(l)]\leq\frac{\overline{\pi}^2}{g(l)^2}$$

As a result:

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

By Markov's inequality:

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

Hence:

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

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 3 and $\overline{\varepsilon}=\min\{(3\lambda)^{-1}2L,\sqrt{(6\lambda^2)^{-1}4\rho L}\}$ as from Lemma 4. ▮

Proof of Remark 1

To prove the statement we need to show that every strict Nash equilibrium satisfies condition (7), i.e.:

$$\left|x^D(t+\Delta t, t, x+\Delta x) - x^D(t+\Delta t, t, x)\right| \leq (1-\lambda\Delta t)\left|\Delta x\right| \tag{16}$$

This condition holds if the system of ODE (5) admits the following quadratic Ljapunov function (see, for example, Ljung (1977)):

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

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

Suppose x^* is a strict Nash equilibrium and w.l.g. let $x^* = 0$. Consider the linearization of the system (5) around $x^* = 0$:

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

where $A \equiv \left. Df(x)\right|_{x^*=0}$ 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). The scalar product (5) 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. ∎

Notes

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

[2] The system of equations (2) carries a direct analogy with Börgers and Sarin (1997) 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. Hence in their model these weights do not depend on the step number n, and as a result, the formulation of their model only accounts for the reinforcement effect.

[3] Convergence in the sense that:

$$\inf_{x \in D_R} |x(t) - x| \to 0 \text{ for } t \to \infty$$

[4] In essence, the reason why this approximation result holds for random step sizes is related to that stated in Remark 4.3 of Benaim (1999) and therein references, namely that the approximation results of this type hold also for stochastic step sizes, when these are measurable with respect to $\Im\{x(n)\}$ and square summable. This could offer an alternative way to prove Lemma 1.

[5] In essence, the reason is that we cannot translate the $O(n^{-2})$ order of magnitude statement into a numerical bound on the error term that we make by approximating the learning process by the replicator dynamics. Knowing that the approximation error is $O(n^{-2})$, means that we know that the norm of the error il less than kn^{-2} for some positive k. However, we do not know k, and this may not be independent of n.

[6] We say that a solution $x = 0$ is stable if for each $\varepsilon > 0$, there is $\delta = \delta(\varepsilon, t_0)$ such that $| x(t_0) | < \delta \Rightarrow | x(t) | < \varepsilon$ 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 $| x(t_0) | < c$. 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 $| x(t) | < \varepsilon$ for all $t \geq t_0 + T(\varepsilon)$ and for all $| x(t_0) | < c$.

These definitions are standard and can be found for example in Khalil (1996) p. 134.

[7] In the terminology of Benaim (1999) and applied in Laslier et al. (2001), p. 347, the replicator dynamics constitutes a.s. a *limit trajectory* (and not only an *asymptotic-pseudo-trajectory)* for the process.

[8] Although not pursued in this paper, these considerations could pave the way to a number of

interesting comparative statics exercises.

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.

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.

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.

HOFBAUER J. AND K. SIGMUND (1988), *"The Thoery of Evolution and Dynamical Systems,* . Cambridge University 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. (2002), “Reinforcement Learning and the Power LAw of Practice: some Analytical Result," *Discussion Papers in Economics and Econometrics 0203,* University of Southamtpon.

Karlin, S. and H. Taylor (1981), *“A Second Course in Stochastic Processes, .* Academic Press.

Kahlil, 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. (1977), “Analysis of recursive stochastic algorithms," *IEEE Trans. Automatic Control,* **AC22**, 551-575.

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

Pemantle, R. (1990), “Non-convergence to unstable points in urn models and stochastic approximation," *The Annals of Probability,* **18**, 698-712.

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.

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

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

YOUNG, H. P. (1993), “The Evolution of Conventions," *Econometrica,* **61**, 57-84.

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

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