Department of Economics
University of Southampton
Southampton SO17 1BJ
UK

Discussion Papers in Economics and Econometrics

REINFORCEMENT LEARNING AND THE POWER OF PRACTICE: SOME ANALYTICAL RESULTS

By

Antonella Ianni

No. 0203

This paper is available on our website
http://www/soton.ac.uk/~econweb/dp/dp02.html

Reinforcement Learning and the Power Law of Practice: Some Analytical Results[1].

Antonella Ianni[2]

University of Southampton (U.K.),

February, 12, 2002

[1]The author thanks D. Balkenborg, T. Börgers, J. Hofbauer, G. Mailath, R. Sarin, J. Välimäki, seminar participants at the 2001 North American Winter Meetings of the Econometric Society, at the University of Exeter and at the University of Southampton for useful comments, and the ESRC for financial support under Research Grant R00022370.

[2]Address for Correspondence: Department of Economics, University of Southampton, Southampton SO17 1BJ, U.K.. e-mail: a.ianni@soton.ac.uk

Abstract

Erev and Roth (1998) among others provide a comprehensive analysis of experimental evidence on learning in games, based on a stochastic model of learning that 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). This note complements this literature by providing an analytical study of the properties of such learning models. Specifically, the paper shows that:

a) up to an error term, the stochastic process is driven by a system of discrete time difference equations of the replicator type. This carries an analogy with Börgers and Sarin (1997), where reinforcement learning accounts only for the Law of Effect.

b) if the trajectories of the system of replicator equations converge sufficiently fast, then the probability that all realizations of the learning process over a possibly infinite spell of time lie within a given small distance of the solution path of the replicator dynamics becomes, from some time on, arbitrarily close to one. Fast convergence, in the form of exponential convergence, is shown to hold for any strict Nash equilibrium of the underlying game.

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. Unlike field data, experimental data allow to focus on the dynamics of the learning process of a well specified and controlled interactive setting, where subjects are typically required to play exactly the same game repeatedly over time, often with varying opponents. Hence the main source of the observed dynamics is the learning process. A question that has received increasing attention is that of how people learn to play games. Various learning models, such as reinforcement learning (Roth and Erev (1995) and Erev and Roth (1998) among others), belief based learning, such as various versions of fictitious play (Fudenberg and Levine (1995) and (1998) among others) and experience weighted attraction learning (Camerer and Ho (1999) among others) have been fitted to the data generated by experiments. These different approaches share the common feature that they aim at providing a learning based foundation to equilibrium theory, by relying heavily on empirical investigation of available data.

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. Specifically, the basic specification of a reinforcement learning model accounts for two main elements: the Law of Effect (positive reinforcement learning) and the Power Law of Practice (learning curves tend to be steeper initially). Within a growing area of research in experimental economics, Roth and Erev (1995), Erev and Roth (1998), Sarin and Vahid (1998), Felthovich (2000), Mookherjhee and Sopher (1997) among others provide a comprehensive analysis of experimental results that are shown to be well explained (ex-ante and ex-post) by these models.

Despite their wide applications, very little is known on the analytical properties of the class of reinforcement learning models. Results to date in this area include Borgers and Sarin (1997) who study the properties of a reinforcement model that accounts

only for the Law of Effect; Posh (1997) who analyzes a reinforcement learning model applied to an underlying Matching Penny game; Rustichini (1999) who characterizes the asymptotic properties of a reinforcement model defined in a non-interactive setting and Hopkins (2000) who investigates the asymptotic properties of a perturbed version of a reinforcement learning model in a two-player setting.

This note contributes to this literature in that it studies the asymptotic behaviour of reinforcement learning models, as well as the dynamics of their sample paths over time. Specifically this paper shows that: a) 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 1) and b) if the trajectories of the system of replicator equations converge sufficiently fast, then the probability that all realization of the learning process over a given spell of time lie within a given small distance of the solution path of the replicator dynamics becomes arbitrarily close to one, from some time on (Theorem 1). In particular, the paper shows that the conditions of which in b) above are always satisfied in proximity of a strict Nash equilibrium of the underlying game (Remark 1). Hence, if the initial condition with which the learning process is started lies within the basin of attraction of a strict Nash equilibrium, then the learning process converges with probability one to that Nash equilibrium.

The objective of the paper is achieved by modeling the learning process in terms of a non-linear urn scheme that formalizes the stochastic process of individual learning. As an example consider a situation where a finite number of players are to play repeatedly over time a normal form game with strictly positive payoffs. Suppose that, at each round of play, players choose actions probabilistically in the following way: player i's behaviour is described by an urn of infinite capacity containing balls of as many colours, as actions available; each action is chosen with probability equal to the proportion of balls of the corresponding coulour in the urn. Furthermore, suppose that the proportions of balls in the urns are updated through time to reflect the payoff obtained by players in the interaction. If player i has sampled a colour j-ball at time t, played action j at time t and received a positive payoff, then she would add a number of colour j-balls, exactly equal to the payoff that she got, to her urn. As a result, at time $t+1$, the proportions of balls in her urn will be different from

what they were at time t. In particular, the proportion of colour j (vs. colour $k \neq j$) balls will be higher (vs. lower) than it was, since action j has been played and has produced a positive payoff (vs. action k has not been played and has not produced any payoff). This formalizes a positive reinforcement effect of actions that are played, commonly referred to as the Law of Effect. Moreover, the increase in the proportion of colour j-balls will be decreasing in the total number of balls in the urn, meaning that an action taken at an early stage of the learning process will have a stronger effect on the proportion than the same action taken at a later stage. This formalizes the fact that learning curves are steeper initially, a property usually labelled as the Law of Practice.

Despite its simple formulation, this model generates quite a complex dynamics, the study of which is the object of this paper. Understanding the analytical properties of this widely used learning model is also essential as it allows to address some of the issues noted below.

A first question that arises is whether in this model players' behaviour becomes stationary over time: in other words will the proportions of balls in the urn that defines choice behaviour on the part of players converge to a limit? Posh (1997) shows that the answer to this question is 'not necessarily so'. He studies a two player reinforcement learning model where players repeatedly play a Matching Pennies game. His model is analogous to the one sketched above, except that, at each time t, the total number of balls in each urn is renormalized to t. The paper shows that, with positive probability, this learning process cycles around the orbits of the corresponding deterministic replicator equation. It is interesting to notice that the Law of Practice plays a key role here. To see this, consider the same model but where, at each time t, the total number of balls in each urn is renormalized to a constant K (other things being left equal). Hence, the relative effect of the choice of an action on choice probabilities is constant over time (or in other word there is no Law of Practice at work). This different renormalization makes the model analogous to the one studied in Börgers and Sarin (1997). Although analogous in motivation, this model produces different results: by taking a continuous time limit, the authors show that for any underlying game being played players will, eventually and with probability one, play

a pure strategy. Hence, for an underlying Matching Penny game, limit behaviour converges[1].

A second question is whether the behavioural specification of reinforcement learning model leads to optimal choices. For example, suppose there is a 'best' action that, when chosen, delivers the highest possible payoff. Will players who repeatedly face this environment eventually learn to choose it? Rustichini (1998) studies the optimal properties of a single agent reinforcement learning model that accounts for the Law of Effect, as well as for the Law of Practice. He shows that in a linear adjustment model like the one we consider, the process governing the proportions of balls in the urns converges almost surely to the action that maximizes the expected payoff (where the expectation is taken over the states of the world), whenever such an action is unique[2]. This optimality property does not necessarily carry over to an interactive setting. In fact, a key assumption in Rustichini's paper is that 'nature', when generating random states of the world, does so according to an ergodic (invariant) process; this is in general not the case in an interactive setting, where players' behaviour may display phenomena of lock in and path-dependence in choices. Consistently with this intuition, the already mentioned Börgers and Sarin (1997) reinforcement learning model can, with positive probability, be locked into any suboptimal state at the boundary of the simplex. However, by explicitly modeling the Power Law of Practice, results can be substantially strengthened to recover optimality properties of the learning algorithm. Namely, a direct implication of the results we obtain in this paper is that, if the underlying game admits a Nash equilibrium in strictly dominant strategies, then starting from any (interior) initial condition and from some time on, any realization of the learning process will remain arbitrarily close to that Nash equilibrium.

A third point that is worth noticing relates to underlying games that admit multiple Nash equilibria. Issues of multiplicity are endemic in many interactive settings, a leading example being the class of coordination games. Within the literature on learning and evolution, traditional approaches to equilibrium selection often rely on modeled ergodicity properties of the underlying dynamic process (as for example in Kandori, Mailath and Rob (1993) and Young (1993)). A reinforcement learning model is, by its mere definition, non-ergodic: players' behaviour is path-dependent and as

such can be absorbed in different steady configurations of play, depending on the initial condition with which the learning process is started. A first step in handling issues of multiplicity in a path-dependent setting relies on the full characterization of the basins of attraction of different absorbing states (or, more generally, ergodic sets). For a reinforcement learning model that incorporates only the Law of Effect this connection cannot be easily established: Börgers and Sarin (1997) Remark 3 observes that, starting from any interior initial condition, the process can reach any of its absorbing states with positive probability. On the contrary, this paper shows that, once the Power Law of Practice is modeled, the learning process will converge to a strict Nash equilibrium of the underlying game whenever its initial condition lies within its basin of attraction, in a suitably defined neighbourhood.

A point raised by the above questions is that the qualitative features of the stochastic process generated by learning models based on the Law of Effect are very sensitive to whether the model also incorporates the Law of Practice. One way to introduce a taxonomy is to notice that under the Law of Practice, the process shows *decreasing gains*, in the sense that the magnitude of state transitions is decreasing over time, while in the absence of such an effect, the process exhibits *constant gains*, meaning that the effect on the state of each action choice is constant at any point in time. Models that formalize decreasing gains typically arise endogeneously in the framework of fictitious play (see Fudenberg and Kreps (1993), Fudenberg and Levine (1998), Kaniovski and Young (1995), Benaim (1999) and Benaim and Hirsch (1999b)). Constant gains are instead prominent in evolutionary models (see Binmore (1992), Boylan (1995), Binmore, Samuelson and Vaughan (1995), Binmore and Samuelson (1997), Benaim and Hirsch (1999a), Corradi and Sarin (2000), Benaim and Weibull (2000)).

In the reinforcement learning model we study in this paper gains decrease endogeneously, 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, once such sequences are renormalized to a common scale (which can for example be the realized sequence of gains of a given player), the results of this paper show that any realisation of the learning process can be suitably approximated by a replicator dynamics, whenever the solutions of the latter converge sufficiently fast. In fact, this condition is shown to hold in a neighbourhood of any *strict* Nash equilibrium. Hence, if the learning process is started in proximity of a strict Nash equilibrium, 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.

The results we obtain rely on stochastic approximation techniques (Ljung (1978), Arthur et al. (1987), (1988)) to establish the close connection between the reinforcement learning process and the underlying deterministic replicator equation. By explicitly modeling the Power Law of Practice we are able to track the magnitude of the jumps of the stochastic process, to obtain the desired result. Since replicator dynamics have been studied extensively in biology, as well as in economics (see for example Hofbauer and Sigmund (1998) and Weibull (1995)), results known in that area are then used to establish that the property we require holds for any strict Nash equilibrium of the underlying game.

The paper is organized as follows. Section 2 describes the reinforcement learning model we study; Section 3 states the main result of the paper, the proof of which is contained in the Appendix; and Section 4 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 $\pi^i : \times_i A^i \equiv A \to \Re$ is player i's payoff function[3]. 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}.

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:

$$\begin{aligned} x^i_j(n+1) &\equiv \frac{b^i_j(n+1)}{\gamma^i(n+1)} = \frac{b^i_j(n) + \pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)} \\ x^i_k(n+1) &\equiv \frac{b^i_k(n+1)}{\gamma^i(n+1)} = \frac{b^i_k(n)}{\gamma^i(n) + \pi^i_j(n)} \quad \text{for } k \neq j \end{aligned} \tag{1}$$

If payoffs are positive (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.*

It is instructive to rewrite (1) by recalling that $b^i_j(n) \equiv x^i_j(n)\gamma^i(n)$, as:

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

This shows that conditional upon $a(n) \equiv [a_i(n) = j, a_{-i}(n)]$ 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. 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[4].

The above reasoning is made conditional on action profile $a(n)$ being played at step n, but this is clearly a random variable. Since actions are chosen at random and independently by players, with each player i adopting, at step n, a mixed strategy defined by the vector of proportions of balls in urn i at step n, $x^i(n)$, we postulate that each $a(n)$ takes values $a \in A$ with probability:

$$\Pr[a(n) = a] = x^1_{a_1}(n)x^2_{a_2}(n)...x^N_{a_N}(n) \equiv x_a(n)$$

As a result the dynamic element of the model is captured by a sequence of random matrix functions $\Pi(a(n)) \equiv [\Pi^i_j(a(n))] : A \to \Re^{m \times N}$, where $\Pi^i_j(a(n))$ is the random number of balls of colour j added to urn i at step n. In particular, at each step n, $\Pi(a(n))$ takes values $\pi(a) \equiv [\pi^i_j(a)]$ (where $\pi^i_j(a)$ is the payoff that player i gets by choosing action j when the realized action profile is a) with probability $x_a(n)$:

$$\Pr[\Pi(a(n)) = \pi(a)] = \Pr[a(n) = a] = x_a(n)$$

Clearly, at any given n, $\sum_a x_a(n) = 1$, as well as $\sum_{a_{-i}} x_{a_{-i}}(n) = 1$.

To lighten the notation let $\gamma \equiv [\gamma^i]$ for $i = 1, ..., N$ and $x \equiv [x^i_j]$ for $i = 1, ..., N$ and $j = 1, ..., m$. Clearly $\gamma \in \Re^N_+$ and, since $x^i \in \Delta_i \equiv \{x^i \in \Re^m_+ : \sum_j x^i_j = 1\}$, $x \in \Delta \equiv \times_i \Delta_i$, i.e. x lies in the cartesian product of the N unit simplexes Δ_i. Given an initial condition, $[\gamma(0), x(0)]$, for any $n > 0$, the above formalization defines a stochastic process over the state space $[x(n), \gamma(n)]$, where the dynamics of the process is determined by the sequence of random matrix functions $\Pi(a(n))$, constructed in relation to the normal form game G, in the way described above.

We call such a process a stochastic urn dynamics $\mathcal{U}$:

$$\mathcal{U} \equiv (\{i = 1, .., N\}; [x(n), \gamma(n)]; \Pi(a(n)); n > 0)$$

for the underlying game $\mathcal{G}$:

$$\mathcal{G} \equiv \left(\{i = 1, .., N\}; A^i = \{j = 1, .., m\}; \Pi^i : A \to \Re\right)$$

and we denote it by $\{\mathcal{U}, \mathcal{G}\}$. Note that, given an initial condition $[x(0), \gamma(0)]$, the dynamics of the process $\{\mathcal{U}, \mathcal{G}\}$ at any $n > 0$ can be described in terms of a system of stochastic difference equations:

$$\begin{cases} x^i_j(n+1) = x^i_j(n) + \frac{1}{\gamma^i(n)} \Phi^i_j(x(n), \gamma(n)) \\ \gamma^i(n+1) = \gamma^i(n) + \sum_j \Pi^i_j(a(n)) \end{cases} \quad i = 1, ..., N \quad j = 1, ..., m \tag{3}$$

where:

$$\Phi^i_j(x(n), \gamma(n)) = \frac{\Pi^i_j(a(n)) - x^i_j(n) \sum_j \Pi^i_j(a(n))}{1 + \frac{1}{\gamma^i(n)} \sum_j \Pi^i_j(a(n))}$$

and the probability distribution of $a(n)$ is a function of $x(n)$. 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).

The study of this learning process is the object of this paper. Specifically, the main result of the paper will relate the stochastic dynamics of this process to those of the system of deterministic replicator dynamics $f(x^D) : \Delta \to \Delta$ defined by:

$$\frac{d}{dt} x^D(t) = f(x^D(t)) \tag{4}$$

where for $i = 1, ..., N$ and $j = 1, ..., m$

$$f^i_j(x^D) \equiv x^i_j [\sum_{a_{-\mathrm{i}}} \pi^i(j, a_{-i}) x^i_{a_{-\mathrm{i}}} - \sum_a \pi^i(a_i, a_{-i}) x^i_a]$$

System (4), a direct generalization of the Taylor (1979) multipopulation replicator dynamics, 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 recently applied to the study of learning models (Börgers and Sarin (1997)).

Lemma 1 in the Appendix characterizes the process $\{\mathcal{U}, \mathcal{G}\}$ under the assumption that payoffs are positive and bounded. It shows that, conditionally on the past, the expected motion of this process can be decomposed in a deterministic part, $f(x(.))$, which denotes system (4) for the underlying game $\mathcal{G}$, weighted by random sequences $\gamma^i(.)^{-1}$, plus an error term, that is uniformly bounded. The actual process can then by written as:

$$x^i(n+1) = x^i(n) + \frac{1}{\gamma^i(n)} f^i(x(n)) + \varepsilon^i(x(n), \gamma(n)) \tag{5}$$

for $i = 1, ..., N$, where $\varepsilon(x(.)), \gamma(.))$ is shown to converge in the limit, for n becoming large.

Lemma 1 introduces an important relation between the learning process we study and the replicator dynamics, since it states that, up to an error term, the deterministic replicator equation fully characterizes its expected motion. We may then conjecture that the dynamics of our learning model may be suitably described by studying the dynamics of its deterministic counterpart. To pursue this intuition, we first notice that the process defined by (5) is characterized by N different random step sizes: $[\gamma^i(n)]^{-1}$ for $i = 1, ..., N$, where we recall $\gamma^i(n)$ denotes the total number of balls in player i's urn. This makes standard techniques of stochastic approximation theory (see for example Fudenberg and Levine (1998), Chapter 4, or Benveniste at al (1990) and Ljung (1978) among others) not directly applicable. In order to overcome this problem, we consider a renormalized process, that we denote by $\{\mathcal{U}_g, \mathcal{G}\}$, and we construct as follows. Suppose that, at each step n, the total number of balls in each urn is renormalized to some positive value, $g(n)$, leaving proportions, i.e. $x(n)$, unaffected. This is obtained by simply re-adjusting each urn composition, $b^i_j(n)$ to $b^{ir}_j(n)$ in such a way as to ensure that:

$$x^i_j(n) \equiv \frac{b^i_j(n)}{\gamma^i(n)} = \frac{b^{ir}_j(n)}{g(n)}$$

for all $i = 1, ...N$ and for all $j = 1, ..., m$ and leads to the renormalized urn process:

$$\mathcal{U}_g \equiv (\{i = 1, .., N\}; [x(n), g(n)]; \Pi(a(n)); n > 0) \tag{6}$$

which dynamics is defined by:

$$\begin{cases} x_j^i(n+1) = x_j^i(n) + \frac{1}{g(n)}\Phi_j^i(x(n), g(n)) \\ \{g(n)\} \end{cases} \quad i = 1, ..., N \quad j = 1, ..., m$$

where $\Phi_j^i(.,.)$ is as defined in (3).

By this doing, the sequence of step sizes of the renormalized process becomes equal to $\{g(n)^{-1}, n > 0\}$ and, as such, it is exactly the same for all players. A standard choice used for example in Arthur (1993), Posh (1997) and Hopkins (2000) is to take the deterministic sequence $g(n) \equiv n$. It turns out that the results we obtain in this paper hold for any (possibily random) sequence $\{g(n), n > 0\}$, that is square summable, but not summable satisfying the following Assumption:

Assumption $\{g(n), n > 0\}$ is such that $g(n) > 0$, $\sum_n g(n)^{-1} = \infty$ and $\sum_n g(n)^{-2} < \infty$ (with probability one).

This assumption is consistent with standard techniques used in stochastic approximation theory. We note that, since any $\gamma^i(n)$, regarded as a sequence over n, $\{\gamma^i(n), n > 0\}$, satisfies this assumption whenever payoffs in the underlying game are strictly positive and bounded[5], we may allow for the choice of any realized sequence of $\gamma^i(n)$, as long as, at each step n, the urn composition in all urns $j \neq i$ is renormalized to this chosen $\gamma^i(n)$ (See Remark 2 after the proof of Lemma 1 in the Appendix).

3 The main result

The study of the asymptotics of the process $\{\mathcal{U}_g, \mathcal{G}\}$ naturally involves a deal of technicalities. 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[6], 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 $\mathcal{G}$, as well as all the vertices of the simplex Δ. Results on 'attainability' (i.e. convergence with positive 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.

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 (4). Theorem 1 of Ljung (1978) provides 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:

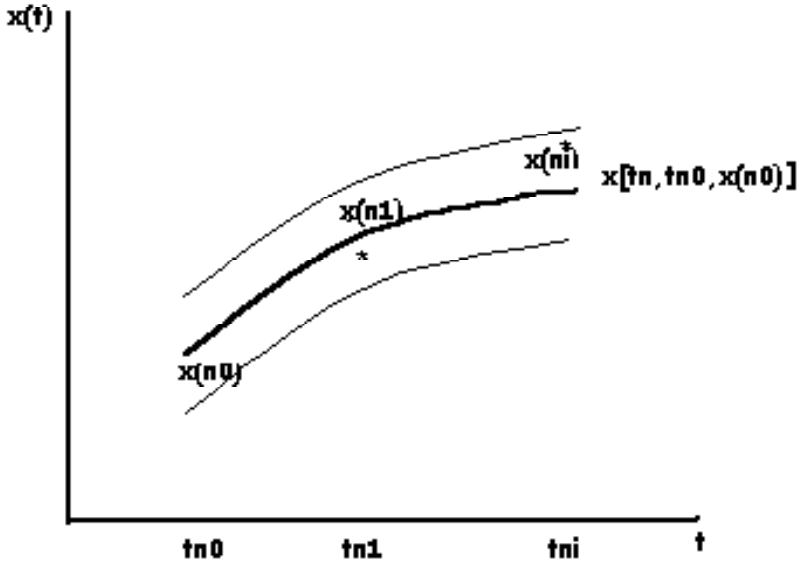
$$\begin{aligned} D_S \equiv \{&x \in \Delta \mid f(x) = 0 \text{ and the Jacobian } Df(x) \text{ has} \\ &\text{only eigenvalues with strictly negative real part}\} \end{aligned}$$

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 of Ljung (1978) cannot be easily checked).

As described below, whenever it applies, our result relates the trajectories of the system of replicator dynamics (4) 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 (and any strict Nash equilibrium is as such), 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 $\{\mathcal{U}_g, \mathcal{G}\}$ 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 (4), started at time t_0 with initial condition equal to $x(n_0)$ is plotted against the same time scale and labelled as $x^D(t, t_0, x(n_0))$ in the picture that follows.



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 (4) 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 small, after time $\overline{n}$.

Theorem 1 *Consider the stochastic process $\{\mathcal{U}_g, \mathcal{G}\}$ (defined by system (6)). Suppose payoffs of $\mathcal{G}$ are bounded and strictly positive. Let the system of ODE (4) 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 $\mathcal{G}$, 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 $\{\mathcal{U}_g, \mathcal{G}\}$ 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. Notice that, since the RHS of inequality (8) is square summable, the statement holds for any $\overline{N}$, possibly infinite.

Next, condition (7) is shown to hold for any strict Nash equilibrium of the underlying game $\mathcal{G}$:

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

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

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 $\{\mathcal{U}_g, \mathcal{G}\}$ 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.

3.1 Outline of the Proof of Theorem 1

The main result relies on a series of Lemmas.

As already mentioned, Lemma 1 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, defined in relation to the sigma-algebra generated by $\{x^i(k), \gamma^i(k); k = 1, ..., n\}$. Lemma 1 applied to $\{\mathcal{U}_g, \mathcal{G}\}$ allows us to re-write the process as:

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

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 2 then proceeds to show that if the process $\{\mathcal{U}_g, \mathcal{G}\}$ 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 2 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 2, 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 3 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 2 holds simultaneously for all n_l.

4 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 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.

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 drives the stochastic dynamics. Both requirements are shown to be essential to establish the result.

Appendix

Lemma 1 *Consider a Stochastic Urn Dynamics $\{\mathcal{U}, \mathcal{G}\}$. 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:

$$\begin{cases} x^i_j(n+1) = x^i_j(n) + \frac{1}{\gamma^i(n)} f^i_j(x(n)) + \varepsilon^i_j(x(n)), \gamma(n)) & n \geq 1 \\ 0 < x^i_j(0) < 1 & n = 0 \end{cases}$$

where:

$$f^i_j(x(n)) = x^i_j(n)[\sum_{a_{-\mathrm{i}}} \pi^i(j, a_{-i}) x_{a_{-\mathrm{i}}}(n) - \sum_a \pi^i(a_i, a_{-i}) x_a(n)]$$

and:

$$\Pr[\lim_{n\to\infty} \sum_{k=n}^{\infty} \varepsilon^i_j(x(k), \gamma^i(k)) = 0] = 1$$

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

Proof. First notice that, since $\Pr[\Pi(a(n)) = \pi(a(n))] = \prod_{i=1,...,N} \Pr[\Pi^i(a(n)) = \pi^i(a(n))]$, we can study the dynamics of $x(n)$ by looking at each single $x^i(n)$ separately.

Simple algebra shows that the dynamics is defined by:

$$\begin{cases} x^i_j(n+1) = x^i_j(n) + \frac{1}{\gamma^i(n)} \Phi^i_j(x(n), \gamma(n)) & n \geq 1 \\ 0 < x^i_j(0) < 1 & n = 0 \end{cases} \tag{9}$$

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

$$\Phi^i_j(x(n), \gamma(n)) = [\Pi^i_j(n) - x^i_j(n) \sum_j \Pi^i_j(n)] + \delta^i_j(x(n), \gamma(n)) \tag{10}$$

with:

$$\delta^i_j(x(n), \gamma(n)) \equiv -\frac{1}{\gamma^i(n)} [\Pi^i_j(n) - x^i_j(n) \sum_j \Pi^i_j(n)] \left[\frac{\sum_j \Pi^i_j(n)}{1 + \frac{\sum_{\mathrm{j}} \Pi^{\mathrm{i}}_{\mathrm{j}}(n)}{\gamma^{\mathrm{i}}(n)}} \right]$$

where $\Pi(a(n))$ is shortened to $\Pi(n)$.

We then study the conditional expectation $E[\Phi^i_j(x(n), \gamma(n)) \mid x(n), \gamma(n)]$ by looking at the two additive components separately. Simple algebra shows that:

$$E[[\Pi^i_j(n) - x^i_j(n)\sum_j \Pi^i_j(n)] \quad | \quad x(n), \gamma(n))] =$$

$$= \quad x^i_j(n)[\sum_{a_{-\mathsf{i}}} \pi^i(j, a_{-i}) x^i_{a_{-\mathsf{i}}}(n) - \sum_a \pi^i(a_i, a_{-i}) x^i_a(n)]$$

$$\equiv \quad f^i_j(x(n))$$

Also, since:

$$\frac{\sum_j \Pi^i_j(n)}{1 + \frac{\sum_{\mathsf{j}} \Pi^{\mathsf{i}}_{\mathsf{j}}(n)}{\gamma^{\mathsf{i}}(n)}} \quad \leq \quad \sum_j \Pi^i_j(n) \leq \overline{\pi}$$

$$\Pi^i_j(n) - x^i_j(n)\sum_j \Pi^i_j(n) \quad \leq \quad \Pi^i_j(n) \leq \overline{\pi}$$

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

$$\mid \delta^i_j(x(n), \gamma(n)) \mid \leq \frac{1}{\gamma^i(n)}[\overline{\pi}]^2$$

As a result, we can now write:

$$x^i_j(n+1) = x^i_j(n) + \frac{1}{\gamma^i(n)} f^i_j(x(n)) + \varepsilon^i_j(x(n)), \gamma(n))$$

where:

$$\varepsilon^i_j(x(n)), \gamma(n)) \quad = \quad \frac{1}{\gamma^i(n)}[\delta^i_j(x(n)), \gamma(n)) + \eta^i_j(x(n), \gamma(n))]$$

$$\eta^i_j(x(n), \gamma(n)) \quad \equiv \quad \Phi^i_j(x(n), \gamma(n)) - E[\Phi^i_j(x(n), \gamma(n)) \mid (x(n), \gamma(n)]$$

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

$$\Xi(n) \quad \equiv \quad \sum_{k=1}^{n-1} \varepsilon^i_j(x(k), \gamma(k))$$

$$\equiv \quad \sum_{k=1}^{n-1} \frac{1}{\gamma^i(k)} \delta^i_j(x(k)), \gamma(k)) + \sum_{k=1}^{n-1} \frac{1}{\gamma^i(k)} \eta^i_j(x(k), \gamma(k))$$

$$\equiv \quad \Xi_\delta(n) + \Xi_\eta(n)$$

Let $\mathsf{z}(n)$ be the sigma-algebra generated by$\{x^i(k), \gamma^i(k); k = 1, ..., n\}$.

Note that:

$$\begin{aligned}\Xi_\delta(n+1) &= \Xi_\delta(n) + \frac{1}{\gamma^i(n)}\delta_j^i(x(n),\gamma(n)) \\ \Xi_\eta(n+1) &= \Xi_\eta(n) + \frac{1}{\gamma^i(n)}\eta_j^i(x(n),\gamma(n))\end{aligned}$$

and since by construction, δ_j^i is bounded as in eq. (4), 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 $g^i(n) = \gamma^i(0) + n\overline{\pi}$ 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 $\mathsf{z}(n)$. In fact:

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

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 $E[\eta_j^i(x(n),\gamma(n)) \mid \mathsf{z}(n)] = 0$, $\Xi_\eta(n)$ is a quadratically integrable martingale relative to $\mathsf{z}(n)$. 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 $\Xi(\infty) - \Xi(n) \equiv \sum_{k=n}^{\infty} \varepsilon_j^i(x(k),\gamma^i(k))$, the assert follows. ∎

Remark 2 *Let Ω^* be a subspace of the sample space of the process $\{x(n), \gamma(n)\}$ such that the assumptions of Lemma 1 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:*

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

Hence:

$$\lim_{n\to\infty} \frac{1}{\gamma^i(n, \omega^*)} = 0$$

$$\sum_{n=0}^{\infty} \frac{1}{\gamma^i(n, \omega^*)} = \infty$$

$$\sum_{n=0}^{\infty} \frac{1}{(\gamma^i(n, \omega^*))^2} < \infty$$

As a result, any sequence $\{\gamma^i(n, \omega^*), n > 0\}$ satisfies Assumption 2, for any i and for any realization ω^*.

Lemma 2 *Consider the stochastic process $\{\mathcal{U}_g, \mathcal{G}\}$ under the assumptions of Lemma 1. 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 1, for $j(n) \geq n + 1$, the process $\{\mathcal{U}_g, \mathcal{G}\}$ can be re-written as:

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

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

Since the function f is Lipschitz in x:

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

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:

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

$$+\Delta t(n, j(n)) L \max_{n \leq k \leq j(n)-1} \mid x(k) - x' \mid + \tag{11}$$

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

As for the last term, from Lemma 1 we know that, for all $\alpha > 0$ there exists an $n = n(\alpha)$ such that for all $n > n(\alpha)$ with probability one:

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

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 $\left| \sum_{s=n}^{j(n)-1} \varepsilon(x(s), g(s)) \right|$ 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 (11) states that:

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

Hence:

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

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 3 *Beyond the assumptions of Lemma 2, suppose that the system of ODE (4) 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,

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

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 2.

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 1 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 2 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 (4) 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:

$$\begin{aligned} \left|x_0^D(l+1) - x(n_{l+1})\right| &\leq \left|x_0^D(l+1) - x_l^D(l+1)\right| + \left|x_l^D(l+1) - x(n_{l+1})\right| \\ &\leq \left|x_0^D(l+1) - x_l^D(l+1)\right| + L\Delta t_l^2 + \alpha(n_l) \end{aligned} \tag{12}$$

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 2 that the last two terms on the RHS of (12) 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 (12) converge, their distance will become increasingly small. An assumption that is sufficient to establish the result that follows requires:

$$\left|x^D(t+\Delta t, t, x+\Delta x) - x^D(t+\Delta t, t, x)\right| \leq (1-\lambda\Delta t)\left|\Delta x\right| \tag{13}$$

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 (12) can be rewritten as:

$$\left|x_0^D(l+1) - x(n_{l+1})\right| \leq (1-\lambda\Delta t_l)\left|x_0^D(l) - x(n_l)\right| + L\Delta t_l^2 + \alpha(n_l) \tag{14}$$

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 (14) 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 2 $\alpha(n_l) < r(\varepsilon) \equiv \frac{\lambda^2\varepsilon^2 3}{4L} < \frac{\rho}{2}$, 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 2, 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 2 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 2:

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

and from Lemma 1:

$$E[\varepsilon_j^i(x(k), g(k)] \leq \frac{\overline{\pi}^2}{g(k)^2}$$

As a result:

$$\begin{aligned}\alpha(n_l) &\leq \sqrt{NM}\sup_i\sup_j\varepsilon_j^i(x(k), g(k) \leq \sqrt{NM}\frac{\overline{\pi}^2}{g(n_l)^2} \\ E[\alpha(n_l)] &\leq \sqrt{NM}\frac{\overline{\pi}^2}{g(n_l)^2}\end{aligned}$$

By Chebyshev'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 $r(\varepsilon) \equiv (4L)^{-1}\lambda^2\varepsilon^2$. In the statement of the theorem $\overline{n} = N_0(\rho)$, defined in Lemma 2 and $\overline{\varepsilon} = \min\{(3\lambda)^{-1}2L, \sqrt{(6\lambda)^{-1}4\rho L}\}$ as from Lemma 3. ∎

Proof of Remark 1

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

$$\left|x^D(t+\Delta t, t, x+\Delta x) - x^D(t+\Delta t, t, x)\right| \leq (1-\lambda\Delta t)\left|\Delta x\right| \tag{15}$$

This condition holds if the system of ODE (4) 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 (4) around $x^* = 0$:

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

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 (see for example Walter (1998), p. 321) 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 (4) also satisfies condition (b), since:

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

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

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

which shows that condition (b) holds. ∎

Notes

[1]A result along these lines is also obtained in Beggs (2001), where the author shows that strategies converge in constant-sum games with a unique equilibrium and in 2-by-2 games also if they are mixed.

[2]Rustichini's result is of interest as it also indirectly complements the literature on non-linear urn processes (see for example Arthur et al. (1988a), Pemantle (1990) and the application of the latter in Benaim and Hirsh (1999), Theorem 5.1) by providing sufficient conditions for attainability and unattainability of fixed points at the boundary of the simplex.

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

[4]One immediate relation between Börgers and Sarin (1997)'s model and the one we study, can be drawn as follows. Let $f(x,\Pi)$ denote a standard discrete time system of replicator equations for an underlying game with strictly positive payoff functions Π. At step n of the repetition of the game, suppose $\gamma^{\mathsf{i}}(n) = n$ for all is and consider the following renormalization of the payoffs of the game:

$$\widetilde{\Pi} \equiv \frac{\Pi}{n+\Pi} : A \to (0,1)$$

These renormalized payoffs clearly satisfy Borgers and Sarin's assumption; hence their result of Section 2 applies and the expected motion of the system (2), conditional upon step n being reached is given by:

$$E[x(n+1) \mid x(n) = x, \gamma(n) = n] = x + f(x, \widetilde{\Pi})$$

i.e. is given by a replicator dynamics in the renormalized variables $\widetilde{\Pi}$. It is worth noticing that, since asymmetric replicator dynamics are not invariant under positive affine transformations of payoffs, in general, the solution orbits of $f(x,\widetilde{\Pi})$ differ from those of $f(x,\Pi)$. However, since:

$$\frac{\Pi}{n+\Pi} = \frac{\Pi}{n} - \frac{\Pi^2}{n^2 + n\Pi}$$

the renormalization is the sum of a $\bigcirc(n^{-1})$ component that affects all payoffs for all players in the same fashion, and hence only alters the speed at which the state moves along the orbits, and a $\bigcirc(n^{-2})$ component which is instead payoff specific. Hence, up to an $\bigcirc(n^{-2})$ error term, the expected increment of the process, x, given $x(n)$, is given by n times $f(x,\Pi)$, i.e.:

$$E[x^{\mathsf{i}}(n+1) \mid x(n) = x, \gamma(n) = n] = x^{\mathsf{i}} + \frac{1}{n} f^{\mathsf{i}}(x,\Pi) + \bigcirc(n^{-2})$$

[5]If, for any profile of actions a, $0 < \underline{\pi} \leq \pi^{\mathsf{i}}(a) \leq \overline{\pi} < \infty$, then the random variable $\gamma^{\mathsf{i}}(n)$ is such that:

$$\gamma^{\mathsf{i}}(0) + n\underline{\pi} \leq \gamma^{\mathsf{i}}(n) \leq \gamma^{\mathsf{i}}(0) + n\overline{\pi}$$

for any n, with probability one.

[6]Convergence in the sense that:

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

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. (2001), "On the Convergence of Reinforcement Learning.," *mimeo,* Oxford University.

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

Benaim, M. and M. Hirsch (1999a), "Stochastic Approximation algorithms with constant step size whose average is cooperative," *Annals of Applied Probability,* 9, 216-241.

Benaim, M. and M. Hirsch (1999b), "Mixed Equilibria and dynamical systems arising from fictitious play in perturbed games," *Games and Economic Behavior,* 29, 36-72.

Benaim, M and J. Weibull (2000), "Deterministic Approximation of Stochastic Evolution in Games," *mimeo,* Stockholm School of Economics.

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

Binmore, K., Samuelson, L. and R. Vaughan (1995), "Musical Chairs: Modelling noisy evolution," *Games and Economic Behavior,* 11, 1-35.

Binmore, K. and L. Samuelson (1997), "Muddling through: Noisy equilibrium selection," *Journal of Economic Theory,* 74, 235-265.

Binmore, K. and L. Samuelson (1999), "Evolutionary Drift and Equilibrium Selection," *Review of Economic Studies,* 66, 363-393.

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

Boylan R. (1995), "Continuous approximation of dynamical systems with randomly mathced individuals," *Journal of Economic Theory,* 66, 615-625.

Bush, R.R. and F. Mosteller (1955), *"Stochastic Models for Learning,* . New York: Wiley.

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

Corradi V. and R. Sarin (1999), "Continuous approximations of stochastic evolutionary game dynamics," *Journal of Economic Theory,,* ?, ?-?.

Cross, J.G. (1973), "A Stochastic Learning Model of Economic Behavior," *Quaterly 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.

Feltovich, N. (2000), "Reinforcement-based vs. belief-based learning models in experimental asymmetric information games," *Econometrica,* 68, 605-641.

Fudenberg D. and D. Kreps (1993), "Learning Mixed Equilibria," *Games and economic Behavior,* 5, 320-367.

Fudenberg D. and D. Levine (1995), "Consistency and Cautious Fictitious Play," *Journal of Economic Dynamic and Control,* 19, 1065-1089.

Fudenberg D. and D. Levine (1998), *"Theory of Learning in Games,* . MIT Press.

Hill, B.M., Lane, D. and W. Sudderth (1980), "A strong law for some generalized urn processes," *The Annals of Probability,* 8, 214-226.

Hofbauer J. and K. Sigmund (1988), *"The Thoery of Evolution and Dynamical Systems,* . Cambridge University Press.

Hopkins, E. (2000), "Two competing models of how people learn in games," *mimeo,* University of Edinburg (fortcoming in Econometrica).

Kandori, M., G. Mailath and R. Rob (1993), "Learning, Mutations, and Long Run Equilibria in Games," *Econometrica,* 61, 29-56.

Kaniovski Y. and P. Young (1995), "Learning dynamics in games with stochastic perturbations," *Games and Economic Behavior,* 11, 330-363.

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

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 Probability,* 6, 680-696.

Mookherjee, D. and B. Sopher (1997), "Learning and decision costs in experimental constant sum games," *Games and Economic Behavior,* 19, 97-132.

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.

Sarin, R. and F. Vahid (1998), "Predicting how people play games: a procedurally rational model of choice.," *mimeo,* Texas AM University.

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.

Walter, W. (1998), *"Ordinary Differential Equations,* . Springer-Verlag.

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