Journal of Mathematical Economics 50 (2014) 148–155

Contents lists available at ScienceDirect

Journal of Mathematical Economics

journal homepage: www.elsevier.com/locate/jmateco

Learning strict Nash equilibria through reinforcement

Antonella Ianni *

Economics Division, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 6 September 2012
Received in revised form
9 April 2013
Accepted 10 April 2013
Available online 19 April 2013

Keywords:
Learning
Law of effect
Power law of practice
Strict Nash equilibrium
Replicator dynamics

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

1. Introduction

Over the last two decades there has been a growing body of research within the field of experimental economics aimed at analyzing learning in games. Various learning models have been fitted to the data generated by experiments with the aim of providing a learning based foundation to classical notions of equilibrium. The family of stochastic learning theories known as positive reinforcement seems 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 application, however, the analytical properties of this class of models have not been fully characterized.

Consider for example a normal form game that admits a strict Nash equilibrium. Suppose players have almost learned to play that equilibrium, meaning that they have been playing for some time and their behaviour is close to that equilibrium prescription. Since the equilibrium is *strict*, any unilateral deviation will necessarily lead to lower payoffs. One would hence expect players to consistently reinforce their choice of their equilibrium action and, by doing this, to eventually learn to play that Nash equilibrium. This seems to be a basic requirement for a learning theory. Yet, it is not satisfied by some reinforcement learning models (e.g. the Cross model as studied in Börgers and Sarin, 1997; Cross, 1973, 1983), and most results available to date can only guarantee that in some reinforcement learning models, it may (e.g. the Erev and Roth model analyzed in Hopkins, 2002, Beggs, 2005 and Laslier et al., 2001). This paper complements this literature by providing sufficient conditions under which a strict Nash equilibrium is reached with probability one.

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

The main results of this paper show that, if players have been learning for sufficiently long, and if play is found close to a strict Nash equilibrium of the underlying game, then players will learn to play it with probability one. While doing so, they will in fact choose actions in a way that is close to a deterministic multi-population replicator dynamics. The latter dynamics have been studied extensively in biology, as well as in economics. The reinforcement learning process we model offers a micro-foundation for replicator dynamics, by showing that they provide a good approximation of the stochastic process of learning that players use to update their action choices. Specifically, our results exploit the fact that in proximity of a strict Nash equilibrium, convergence of the deterministic replicator dynamics occurs at an exponentially fast rate. As

* Tel.: +44 0 2380 592536.
E-mail addresses: ianni@soton.ac.uk, A.Ianni@soton.ac.uk.

0304-4068/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jmateco.2013.04.005

in our learning process the step size decreases endogenously, over time (due to the *Law of Practice*), we are able to define a timescale over which the stochastic component of the reinforcement learning process, which in principle could move the process away from any equilibrium, is in fact overcome by this deterministic effect.[1]

The results we obtain rely on stochastic approximation techniques (Ljung, 1978; Arthur et al., 1987, 1988; Arthur, 1993; Benaim, 1999; Benveniste et al., 1990) to establish the close connection between the reinforcement learning process and the underlying deterministic replicator equation. We show that, up to an error term, the behaviour of the stochastic process is well described by a system of discrete time difference equation of the replicator type (Lemma 4). The main result (Theorem 1) shows that if the trajectories of the underlying system of replicator equations converge sufficiently fast and if the learning process has been going on for sufficiently long, then the probability that all the pathwise realizations of the learning process over a given spell of time, possibly infinite, lie within a given small distance of the solution path of the replicator dynamics, becomes arbitrarily close to one. The property of fast convergence, as required in the main result, is always satisfied in proximity[2] of a strict Nash equilibrium of the underlying game (Remark 2) and is sufficient to guarantee that the approximation error converges uniformly over any spell of time.

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

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

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

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

2. The model

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

We shall think of player i's behaviour as being characterized by urn i, an urn of infinite capacity containing γ^i balls, $b^i_j > 0$ of which are of colour $j \in \{1, 2, \ldots, M\}$. 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 $n = 0, 1, 2 \ldots$. Suppose that $(j, a_{-i}(n))$ is the profile of actions played at step n and $\pi^i(j, a_{-i}(n))$, shortened to $\pi^i_j(n)$, is the corresponding payoff gained by player i who chose action j at step n. Then, exactly $\pi^i_j(n)$ balls of colour j are added to urn i at step n. At step $n + 1$ the resulting composition of urn i, will be:

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

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

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

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

$$x^i_j(n+1) = x^i_j(n)\left[1 - \frac{\pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)}\right] + \frac{\pi^i_j(n)}{\gamma^i(n) + \pi^i_j(n)} \tag{2}$$

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

where j denotes the action chosen at step n.

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

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

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