Attainability of Boundary Points under Reinforcement Learning

Ed Hopkins*
Department of Economics
University of Edinburgh
Edinburgh EH8 9JY, UK

Martin Posch
Department of Medical Statistics
University of Vienna
Vienna 1090, Austria

July, 2003

Abstract

This paper investigates the properties of the most common form of reinforcement learning (the "basic model" of Erev and Roth, *American Economic Review*, **88**, 848-881, 1998). Stochastic approximation theory has been used to analyse the local stability of fixed points under this learning process. However, as we show, when such points are on the boundary of the state space, for example, pure strategy equilibria, standard results from the theory of stochastic approximation do not apply. We offer what we believe to be the correct treatment of boundary points, and provide a new and more general result: this model of learning converges with zero probability to fixed points which are unstable under the Maynard Smith or adjusted version of the evolutionary replicator dynamics. For two player games these are the fixed points that are linearly unstable under the standard replicator dynamics.

Journal of Economic Literature classification numbers: C72, C73, D83
Keywords: Learning in games, reinforcement learning, stochastic approximation, replicator dynamics.

*We thank Josef Hofbauer for his advice and encouragement. We are also grateful to Alan Beggs and two anonymous referees for helpful comments. Errors remain our own. Both authors contributed equally to this work. E.Hopkins@ed.ac.uk, http://homepages.ed.ac.uk/ehk/; Martin.Posch@univie.ac.at, http://www.univie.ac.at/medstat/posch/posch.html

1 Introduction

Whilst equilibrium analysis has been the mainstay of economic theory for many years, economists have more recently turned to non-equilibrium explanations of human behaviour based on learning models. This approach has found considerable success in explaining how people behave in economic experiments (Roth and Erev, 1995; Erev and Roth, 1998; Camerer and Ho, 1999). This in turn leads to the intriguing prospect of using adaptive learning models in economic applications in the wider world outside the laboratory, for example, to explain consumer behaviour (Erev and Haruvy, 2001; Hopkins, 2003) or to design economic mechanisms robust to bounded rationality (Sandholm, 2002). However, such applications are hampered by the fact that learning models are more difficult to work with than equilibrium analysis: one has first to calculate the equilibria and then consider such issues as stability and convergence. Such difficulties are compounded when one works with stochastic rather than deterministic systems. Therefore, recent results that indicate that stochastic learning models can behave in the same way, at least asymptotically, as deterministic models such as the evolutionary replicator dynamics are particular valuable.[1]

In this paper, we highlight some previously hidden technical difficulties in the application of this methodology, and offer some solutions. In particular, we show that two important results in the theory of stochastic approximation cannot be easily applied to the reinforcement learning model popularised by Erev and Roth (1998). This implies that existing results cannot rule out the possibility that this learning process converges to a state which is not a Nash equilibrium, something that is known to be impossible for the deterministic replicator dynamics. Beggs (2002) makes significant progress on this issue. He shows that for single person decision problems the process cannot converge to a suboptimal action.[2] We give a more general result applicable to all normal form games and show that starting from a position where all strategies are played with positive probability this learning process will indeed converge with probability zero to any point which is linearly unstable under the Maynard Smith or adjusted version of the evolutionary replicator dynamics. This rules out survival of suboptimal actions in decision problems and, in games, rules out convergence to points which are not Nash equilibria. We go on to show how these results can be used in a simple practical application.

In this work we are able to clarify some earlier misunderstandings concerning the convergence of reinforcement learning to boundary points. Stochastic approximation examines the behaviour of a learning process by investigation of an ordinary differential equation or ODE derived from the expected motion of the learning process. One

[1] Analysis of single person decision making is found in Arthur (1993), Rustichini (1999) and Sarin and Vahid (1999), for games in Börgers and Sarin (1997), Posch (1997), Ianni (2000) and Hopkins (2002) and for both in Laslier et al. (2001) and Beggs (2002).

[2] He also proves that, in games, strategies which are dominated by a mixed strategy or removed by iterated deletion of dominated strategies are eliminated by reinforcement learning. Additionally he shows, for constant sum games, that in the limit the players play Nash.

classic result is that if the ODE has a global attractor, the learning process converges with probability one to that point. An example of such a result is Corollary 6 of Theorem 4 of Benveniste *et al.* (1990, p45-6; see also Theorem 17, p239). However, the evolutionary replicator dynamics, the system of ODE's associated with the ER model, strictly speaking do not have any global attractors. Proposition 3.2 in Rustichini (1999) claims that in decision problems reinforcement learning converges to the optimal choice. However, as already noted by Beggs (2002), the proof is incomplete, since it appeals to Theorem 4 of Benveniste *et al.* (1990). Second, there exists a very general result in the theory of stochastic approximation due to Pemantle (1990) that proves that there is a zero probability of a stochastic process converging to an equilibrium point unstable under the associated differential equation. However, it is difficult to use this result in learning in games or decision problems as it does not apply at rest points where each agent plays only one action. Therefore, Laslier et al. (2001, Proposition 9) are incorrect to claim that in 2×2 games Pemantle's result ensures that reinforcement learning does not converge to a point which is not a Nash equilibrium. In a similar way, Arthur (1993, Lemma 1 and Theorem 2), in attempting to prove a result in single agent learning, uses the result of Pemantle where it does not apply. Importantly, we do not claim that the results we mention above are wrong. Indeed, our new result shows that this earlier work was essentially correct. However, the methods of proof were not.

The case of single person decision problems has also been studied in two other very different contexts. The first is adaptive sequential randomization schemes for clinical trials. Durham et al. (1998) proposed the so called "randomized Pólya urn" rule for sequential randomization of patients to different treatments. The randomization device is an urn with different types of balls, corresponding to the treatments to be compared. Patients sequentially enter the trial and for each patient a ball is drawn from the urn and replaced. The patient gets the corresponding treatment. If it is a success another ball of the type that was drawn is added to the urn. Otherwise, no balls are added to the urn. This corresponds exactly to the reinforcement learning model of Erev and Roth (1998). Durham *et al.* (1998) show that in the limit with probability one only the treatment with the highest success probability is chosen. They take a different approach to prove their results: embedding the urn process into a continuous time Markov branching process (see Athreya and Ney, 1972) they use known results on these processes to determine the urn dynamics. The other area where similar questions have been considered is in the analysis of learning automata. Indeed, in this context, convergence to boundary points in single agent decision problems has also recently been analysed by Lamberton, Pagès and Tarrès (2002) and in Tarrès (2001).

In the next section, we give a heuristic treatment of the technical problems that boundary rest points pose. In Section 3, we give the main result of this paper that even in the case of boundary rest points, reinforcement learning does not converge to points that are unstable under the replicator dynamics. In the final section, we give an application of our results, showing that reinforcement learning converges to pure Nash equilibria in rescaled partnership games.

2 The Problem with Reinforcement Learning

In this section we introduce the model of reinforcement learning now often referred to as the Erev-Roth (ER) model (after Roth and Erev, 1995; Erev and Roth, 1998), and a variant on it due to Arthur (1993). We also give an introduction to the technical problems that we address in this paper.

Consider $N \geq 1$ agents who repeatedly play the same game over a number of time periods indexed by n. We restrict our attention to finite games in the strategic form. Denote agent i's strategy set $s^i = (s^i_1, ..., s^i_{m^i})$, where m^i is the number of strategies for player i. For $s \in \times_{i=1}^N s^i$ the payoff of player i at round n is a random variable $U^i_n(s)$, with $E(U^i_n(s)) = u^i(s)$ for some function $u^i(s)$. We assume that the random variables $U^i_n(s)$ are independent and that there are constants M_0, M_1 such that a.s. $0 < M_0 < U^i(\cdot) < M_1$. It is usually assumed that payoffs are a deterministic function of strategy choices. However, this is a special case of the current setup which has the added advantage of including single person decision problems by assuming that $N = 1$.

Under reinforcement learning, each agent i is assumed in each period to have a propensity q^i_{jn} for each action j out of the set of her possible actions. Let x^i_{jn} be the probability placed by agent i on action j in period n. In the models of reinforcement learning we consider these probabilities are determined by the following choice rule,

$$x^i_{jn} = \frac{q^i_{jn}}{\sum_k q^i_{kn}} = \frac{q^i_{jn}}{Q^i_n} \text{ for } j = 1, ..., m^i, \tag{1}$$

where $Q^i_n = \sum_{k=1}^N q^i_{kn}$. What is needed to complete the learning model is a means by which to update propensities. In this simple model it takes the form that if agent i takes action j in period n, then his jth propensity is increased by an increment equal to his realised payoff. All other propensities are unchanged. Let U^i_n denote the payoff obtained by player i in period n. And let σ^i_{jn} denote the increment to player i's jth propensity, that is, U^i_n if action j is chosen in period n and zero otherwise. Thus, we can write the updating rule for the Erev-Roth model as

$$q^i_{jn+1} = q^i_{jn} + \sigma^i_{jn} \text{ for } j = 1, ..., m^i. \tag{2}$$

This updating rule reveals why in this model of reinforcement learning payoffs must be positive, or there would be the possibility of propensities becoming negative and the choice probabilities would be undefined.

An alternative formulation due to Arthur (1993) assumes that each agent's step size is renormalised every period. That is, as well as propensities being updated according to the rule (2), all propensities are multiplied by the factor $C(n+1)/Q^i_{n+1}$ for some $C > 0$. That is, for each player i,

$$q^i_{jn+1} = \frac{(q^i_{jn} + \sigma^i_{jn})\, C(n+1)}{n\, C + U^i_n}, \text{ for } j = 1, ..., m^i. \tag{3}$$

The effect of the renormalisation is that at each point $Q_n^i = nC$ for all i. We refer to $1/Q_n^i$ as agent's i's step size. Note that in the case of the Erev-Roth model, it is stochastic and of order $1/n$. However, the step size of the Arthur model is deterministic and exactly proportional to $1/n$.

To illustrate the problems we analyse in this paper, let us now consider an extremely simple case. Two players play the following trivial game where both have a dominant strategy.

	U	D
U	2, 2	2, 1
D	1, 2	1, 1

. (4)

It is tempting to conclude that under both reinforcement learning processes that, in the limit, both players would learn to play their dominant strategies and play would converge to the Nash equilibrium (U, U).

The seemingly obvious way to prove such a result would be to use techniques drawn from stochastic approximation theory. The theory largely works by predicting the behaviour of stochastic processes by using an ordinary differential equation (ODE) formed by taking the expected motion of the stochastic process. For example, consider a stochastic process of the form

$$x_{n+1} - x_n = \gamma_n f(x_n) + \eta_n(x_n) + O(\gamma_n^2) \tag{5}$$

for $x_n \in I\!R^n$. We can think of η as the random component of the process with $E[\eta_n|x_n] = 0$. γ_n is the step size of the process. Stochastic approximation theory obtains its strongest results when this step size is of the order of $1/n$. This paper concentrates on this case.[3]

The mean or averaged ordinary differential equations (ODE's) derived from the stochastic process above would be

$$\dot{x} = f(x). \tag{6}$$

The ODE's associated with the two processes, in this sense, are (variants on) the evolutionary replicator dynamics (Lemma 2 below). But despite having ODE's with the same local stability properties, the two processes potentially have different asymptotic behaviour. This crucially shows the limits to predicting the behaviour of a stochastic process using the ODE alone.

The first possible method for a proof of such a result would be to use a classic result in stochastic approximation. This states, informally, that if the ODE possesses a global attractor, the stochastic process, if its step size is of order $1/n$, will also converge to that point with probability one. Let us examine the associated ODE, the replicator dynamics, for the simple game (4). If x denotes the probability player 1 places on U

[3]Indeed, without this assumption, more or less anything is possible. Arthur (1993) showed that if the step size is constant in single person decision problems then there is a positive probability of convergence to any suboptimal action. In games, as Börgers and Sarin (1997) show, a learning process with a constant step size may converge to a pure state that is not a Nash equilibrium.

and y denotes the probability player 2 places on U, the standard replicator dynamics in this case would be

$$\dot{x} = x(1-x),\ \dot{y} = y(1-y). \tag{7}$$

Thus, the growth rates of x and y are clearly positive for any $(x, y) \in (0, 1) \times (0, 1)$. However, $(x, y) = (1, 1)$ is not strictly speaking a global attractor in that it does not attract orbits with initial condition with either $x = 0$ or $y = 0$. It might be thought that since (1,1) attracts all the interior, i.e. $(0, 1) \times (0, 1)$, this would be sufficient. But, by constructing a counterexample (see Proposition 1 below), we show that this is not the case.

The second possible strategy involves two stages. The first is to establish the existence of a globally applicable Liapunov function for the ODE. Then, by Corollary 6.6 (Benaïm, 1999) the stochastic process must converge to a fixed point of the ODE.[4] One might then hope to apply the theorem of Pemantle (1990) to show that the probability of convergence to a fixed point unstable under the ODE is zero, implying that the process must converge to a stable fixed point. Unfortunately, this theorem cannot be applied as the three unstable fixed points of the replicator dynamics in this case, (0,0), (0,1) and (1,0), are on the boundary of the state space, which invalidates one crucial technical condition of Pemantle's theorem.

It might be hoped that a new result could be derived of similar generality to Pemantle's but applicable to boundary points. This hope is unlikely to be realised as the following result establishes that there is a stochastic learning process that converges with positive probability to a fixed point that is unstable with respect to its associated ODE.

Proposition 1 *In the game (4), if $C < 1$ the Arthur model of reinforcement learning defined by choice rule (1) and updating rule (3) converges with positive probability to one of the unstable fixed points of the replicator dynamics (7).*

Proof: This is a special case of Proposition 5 in Posch (1997). ∎

Posch (1997) also establishes that for the Arthur model, if the constant C is greater than any of the possible payoffs in any 2×2 game then such convergence to an unstable fixed point is not possible. Our main result here shows that the Erev-Roth model never converges to equilibria which are linearly unstable under the replicator dynamics. So, for example, in the game (4) reinforcement learning does not converge to the corner points (0,0), (0,1) or (1,0). More generally, reinforcement learning will not converge to any point that is not a Nash equilibrium, or in single person decision problems, any point that is not optimal.

[4] If the number of fixed points is finite and certain other technical assumptions are met. Note that in the absence of a global Liapunov function convergence of the solutions of the ODE to an equilibrium does not necessarily imply convergence of the stochastic process. See, e.g., example 5.1 in Benaïm (1999).

3 The Main Result

In this section, we state and prove Theorem 1, that the Erev- Roth (ER) model of reinforcement learning converges with probability zero to a point that is linearly unstable under the adjusted or Maynard Smith form of the evolutionary replicator dynamics (defined below). This is proved through a series of intermediate results. In particular, Proposition 2 shows the ER model cannot converge to a Nash equilibrium unstable under the replicator dynamics. Then, Proposition 3 establishes that the ER model cannot converge to a rest point that is not a Nash equilibrium. In single person decision problems this result implies that the learning process will not converge to an action that is suboptimal.

We now develop our analysis of reinforcement learning on a more formal level. Both forms of the reinforcement model define Markov processes on $I\!R^m$ where $m = \sum_{i=1}^{N} m^i$. The state of the system can be summarised by a vector $q_n = (q_n^1, ..., q_n^N)$ with $q_n^i = (q_{1n}^i, ..., q_{m^i n}^i)$. However, the real interest is in the evolution of each player's mixed strategy $x_n^i \in S_{m^i}$, where S_{m^i} is the simplex $\{y = (y_1, ..., y_{m^i}) \in I\!R^{m^i} : \Sigma y_k = 1, y_k \geq 0\}$. Let $x = (x^1, ..., x^N) \in \Delta$ where $\Delta = \times_{i=1}^{N} S_{m^i}$. Let $x^{-i} = (x^1, ..., x^{i-1}, x^{i+1}, ..., x^N)$. Then, denote the vector of expected payoffs for player i conditional on the other players' mixed strategies as $u^i(x^{-i})$.

We can write the standard evolutionary replicator dynamics as

$$\dot{x}_j^i = f_j^i(x) = x_j^i(u^i(s_j, x^{-i}) - x^i \cdot u^i(x^{-i})). \tag{8}$$

They can also be written in vector form as

$$\dot{x}^i = f^i(x) = R(x^i)u^i(x^{-i}), \tag{9}$$

where $R(x^i)$ is a symmetric matrix for which $R_{jj} = x_j^i(1 - x_j^i)$ and $R_{jk} = -x_j^i x_k^i$. One can verify that if each element of x^i is strictly positive, then R is positive semi definite and that $z \cdot R(x^i)z > 0$ for all $z \in I\!R_0^{m_i} = \{z \in I\!R^{m_i} : \sum z_i = 0\}$ (see Hofbauer and Sigmund (1998, Chapter 9.6)). We also consider a variant known as the adjusted or Maynard Smith version of the replicator dynamics (see, for example, Hofbauer and Sigmund (1998, Chapter 11) or Weibull (1995, Chapter 5.2)) which have the form

$$\dot{x}_j^i = \frac{f_j^i(x)}{x^i \cdot u^i(x^{-i})} = \frac{x_j^i(u^i(s_j, x^{-i}) - x^i \cdot u^i(x^{-i}))}{x^i \cdot u^i(x^{-i})}. \tag{10}$$

It is easily verified that the set of rest points under the standard and adjusted replicator dynamics are identical. However, as we will see, the stability properties of these equilibria may be different under the two dynamics. We will need the following definition.[5]

[5]The requirement for a fixed point to be linearly unstable is slightly stronger than requiring it to be unstable (in the sense of Liapunov). The latter, informally put, requires that there exists a neighbourhood of the fixed point such that one can find orbits starting arbitrarily close to the fixed point which leave this neighbourhood. It is possible for a fixed point, if it has one or more eigenvalues with zero real part and hence is non-hyperbolic, to be unstable without being linearly unstable. However, linear instability implies instability.

Definition 1 *A rest point $\bar{x}$ is linearly unstable with respect to a dynamic process $\dot{x} = g(x)$ if its linearisation $Dg(x)$ evaluated at $\bar{x}$ has at least one eigenvalue with positive real part.*

We can now state our main result.

Theorem 1 *Let $\bar{x} \in \Delta$ be a rest point for the replicator dynamics (8), or equivalently the adjusted replicator dynamics (10). If, either A) $\bar{x}$ is not a Nash equilibrium; or B) $\bar{x}$ is a Nash equilibrium linearly unstable under the adjusted replicator dynamics (10); or C) $N = 2$ and $\bar{x}$ is a Nash equilibrium linearly unstable under the replicator dynamics (8), then, for the reinforcement learning process defined by the choice rule (1) and updating rule (2), from any initial condition with all propensities positive, that is $q^i_{j0} > 0$ for each player i and all strategies j,* $\Pr(\lim_{n\to\infty} x_n = \bar{x}) = 0$.

The theorem thus links the behaviour of the ER model of reinforcement learning with that of the adjusted or Maynard Smith replicator dynamics.[6] It is possible to show that in two player games (Lemma 4 below) a fixed point is linearly unstable under the standard replicator dynamics (8) if and only if it linearly unstable under the adjusted version (10). This result allows, at least in two player games, the analysis of reinforcement learning using the more common standard replicator dynamics, which are also somewhat simpler than the adjusted version.

Proof of the main theorem will take the rest of this section. As a first step we relate the expected motion of reinforcement learning to the replicator dynamics. Let $\mathcal{F}_n$ denote the σ-algebra generated by $\{q_1, q_2, \ldots, q_n\}$.

Lemma 1 *The expected change in x^i_n under the basic reinforcement model is*

$$E[x^i_{j\,(n+1)}|\mathcal{F}_n] - x^i_{jn} = \frac{x^i_{jn}(u^i(s_j, x^{-i}_n) - x^i_n \cdot u^i(x^{-i}_n))}{Q^i_n} + O(\frac{1}{(Q^i_n)^2}). \tag{11}$$

Under the Arthur version it is

$$E[x^i_{j\,(n+1)}|\mathcal{F}_n] - x^i_{jn} = \frac{x^i_{jn}\,(u^i(s_j, x^{-i}_n) - x^i_n \cdot u^i(x^{-i}_n))}{nC} + O(\frac{1}{(n\,C)^2}) \tag{12}$$

Proof: See Lemma 1 in Hopkins (2002) and Posch (1997). ∎

One implication of this Lemma is that the expected motion of both forms of reinforcement learning is related to the evolutionary replicator dynamics. The difference

[6] Beggs (2002) was the first to make this connection. His approach is slightly different from ours but equally effective.

between the two different forms of learning lies in their different step sizes. While in the Arthur model, each player has the same step size of $1/(n\,C)$, in the Erev-Roth model, each player has a different step size which is determined by her payoff experience. This means that in order to predict the behaviour of the Erev-Roth learning process, one has to correct the replicator dynamics for the differing learning speeds of the different players. This can be done by first taking the step size of the learning process to be $\gamma_n = 1/n$. Second, we introduce N new variables $\mu_n^i = n/Q_n^i$ that account for the varying step sizes. Then the dynamics for the basic reinforcement model can be written as dynamics with constant step size

$$\begin{aligned} x^i_{j(n+1)} &= x^i_{jn} + \frac{1}{n}\,\mu^i_n\{x^i_{jn}[u^i(s_j, x^{-i}_n) - x^i_n \cdot u^i(x^{-i}_n)]\} + \frac{1}{n}\xi^i_n(x_n) + O(\frac{1}{n^2}) \\ \mu^i_{n+1} &= \mu^i_n + \frac{1}{n}\mu^i_n(1 - \mu^i_n\, x^i_n \cdot u^i(x^{-i}_n)) + \frac{1}{n}\,\zeta^i_n(x_n) + O(\frac{1}{n^2}), \end{aligned}$$

where $\xi^i_n(x_n), \zeta^i_n(x_n)$ are sequences of uniformly bounded random variables with $E(\xi^i_n(x_n)|\mathcal{F}_n) = E(\zeta^i_n(x_n)|\mathcal{F}_n) = 0$. Note that by definition the μ^i_n are uniformly bounded away from 0.

Lemma 2 *The ODE associated with the Arthur model of reinforcement learning is the evolutionary replicator dynamics (9). The ODE associated with the Erev-Roth model is the following modification of the replicator dynamics:*

$$\dot{x}^i = \mu^i R(x^i) u^i(x^{-i}), \quad \dot{\mu}^i = \mu^i(1 - \mu^i x^i \cdot u^i(x^{-i})) \tag{13}$$

If $\bar{x}$ is a rest point for (9) then $(\bar{x}, \bar{\mu})$, with $\bar{\mu}^i = 1/(\bar{x}^i \cdot u^i(\bar{x}^{-i}))$, is a rest point for (13). If $\bar{x}$ is linearly unstable under the adjusted replicator dynamics (10), then $(\bar{x}, \bar{\mu})$ is linearly unstable under (13).

Proof: The first claim, that $(\bar{x}, \bar{\mu})$ is a rest point for (13) is easily verified. The second claim follows from the linearization of (13) at any fixed point being of the form

$$\begin{pmatrix} J & 0 \\ d\dot{\mu}/dx & d\dot{\mu}/d\mu \end{pmatrix}, \tag{14}$$

where J is the Jacobian of the equations $\dot{x}^i = \mu^i R(x^i) u^i(x^{-i})$. Because of the block of zeros to the upper right, it can be shown that every eigenvalue of a matrix of the above form is an eigenvalue for either J or $d\dot{\mu}/d\mu$ (see, for example, Hopkins (2002, Proposition 5)). Hence, if J has one or more positive eigenvalues, the equilibrium point is unstable for the joint dynamics. But because the equilibrium value of each μ^i is exactly $1/(\bar{x}^i \cdot u^i(\bar{x}^{-i}))$, J is identical to the Jacobian for the adjusted replicator dynamics (10) and the result follows. ∎

The behaviour of the replicator dynamics has been investigated for many years now. For example, for generic games the number of fixed points of the replicator dynamics

(standard or adjusted) is finite and consists of the Nash equilibria of the game but also points which are Nash equilibria with respect to the strategies in their support. This includes therefore all points representing pure strategy profiles. See, for example, Weibull (1995, Section 3.3.1). Rest points that are not Nash equilibria are always unstable, whereas rest points that are Nash equilibria can be stable or unstable.

For some equilibrium point $\bar{x} \in \Delta$, its support K is the set of strategies given positive support at $\bar{x}$, or $K = \times_{i=1}^{N} K^i$, with $\bar{x}_j^i > 0$ if and only if $j \in K^i$. Let I be the complement of K and we write $\Delta(K)$ for the face of Δ where only strategies in K have positive representation. The next lemma summarises some properties of the replicator dynamics.

Lemma 3 *Let $\bar{x} \in \Delta$ be a fixed point of either the replicator dynamics (8) or the adjusted replicator dynamics (10) with support K. Then if $\bar{x}$ is a Nash equilibrium and it is linearly unstable, the eigenvectors corresponding to the positive eigenvalues of $Df(\bar{x})$ are contained in $\Delta(K)$. If $\bar{x}$ is not a Nash equilibrium, then it is linearly unstable under the dynamics (8) and (10).*

Proof: Every face of Δ is invariant under both forms of the replicator dynamics (8) and (10). Therefore the Jacobian evaluated at boundary rest points will have two distinct sets of eigenvectors, the first set spanning $\Delta(K)$, the second the rest of Δ. More specifically, as Hofbauer and Sigmund (1998, Chapter 13) note, the Jacobian for the replicator dynamics (8) will possess eigenvalues of the form $u^i(s_j, \bar{x}^{-i}) - \bar{x}^i \cdot u^i(\bar{x}^{-i})$ for each $i \in I$. These eigenvalues are called "transversal" by Hofbauer and Sigmund as each has a corresponding (left) eigenvector e_i (that is, the vector with 1 at position i and zero elsewhere) that points away from $\Delta(K)$. Now, $\bar{x}^i \cdot u^i(\bar{x}^{-i})$ is the payoff for player i at $\bar{x}$, and if $\bar{x}$ is a Nash equilibrium, then $u^i(s_j, \bar{x}^{-i})$ the payoff to strategy j that is not in the support of the equilibrium cannot be higher and so these eigenvalues are non-positive. Hence, if this Nash equilibrium is unstable, then the eigenvectors corresponding to positive eigenvalues must be contained in $\Delta(K)$. To establish the second claim, consider that if $\bar{x}$ is not a Nash equilibrium, it must be true that for some player i and some $j \in I^i$ that $u^i(s_j, \bar{x}^{-i}) > u^i(s_k, \bar{x}^{-i})$ for $k \in K^i$, which by the above argument would give rise to positive eigenvalues. Both results also hold for the adjusted replicator dynamics (10) as the transversal eigenvalues are in that case equal to $(u^i(s_j, \bar{x}^{-i}) - \bar{x}^i \cdot u^i(\bar{x}^{-i}))/\bar{x}^i \cdot u^i(\bar{x}^{-i})$, and so are non-positive for a Nash equilibrium and positive otherwise. ∎

	U	D	B
U	4, 4	2, 2	2, a
D	2, 2	4, 4	2, a
B	a, 2	a, 2	a, a

. (15)

This lemma enables us to classify boundary rest points unstable under the replicator

dynamics into those that are Nash equilibria and those that are not. We can illustrate the distinction using another simple game (15). The first type are Nash equilibria of the game of the whole. For example, in (15) if $a = 1$ then B is dominated for both players, and the replicator dynamics will approach the face where only U and D are represented. However, we want to exclude the possibility of convergence to points such as the mixed Nash equilibrium which places probability 1/2 on both U and D as this point is unstable under the replicator dynamics for the game where U and D are the only strategies. This is relatively easy to do.

We employ a theorem due to Brandière (1998), that generalises the earlier results by Pemantle (1990) and Arthur et al. (1988). The result of Pemantle, stated informally, is that the stochastic process should diverge from an unstable fixed point of the ODE if the stochastic process has positive variance in every direction around that fixed point. This is what makes it inappropriate for use with boundary rest points of the Erev-Roth model. Because if play is at a boundary fixed point, it cannot enter the interior as strategies that start with a zero propensity are never played. Importantly for our purposes, in contrast Brandière (1998) has the weaker condition that there should be positive variance of the stochastic process in unstable directions.

Proposition 2 *Let $\bar{x} \in \Delta$ be a Nash equilibrium that is linearly unstable under the adjusted replicator dynamics (10). Then, for the reinforcement learning process defined by the choice rule (1) and updating rule (2), from any initial condition with all propensities positive, that is $q^i_{j0} > 0$ for each player i and all strategies j,* $\Pr(\lim_{n\to\infty} x_n = \bar{x}) = 0$.

Proof: Any rest point linearly unstable under the replicator dynamics (10) is linearly unstable under the modified dynamics (13) by Lemma 2. By Lemma 3, the eigenvectors corresponding to the positive eigenvectors of $\bar{x}$ are contained in $\Delta(K)$ where K is the support of $\bar{x}$. The result then follows from Theorem 5 of Brandière (1998). ∎

In the case of two player games, we can extend Proposition 2. From Lemma 3 we know that pure Nash equilibria are never linearly unstable. However, mixed Nash equilibria may be stable or unstable and, in $N-$player games the exact stability properties are potentially different under the standard and adjusted replicator dynamics. We can, nonetheless, establish the following equivalence when $N = 2$. This result, together with Proposition 2, of course implies that in two player games reinforcement learning will never converge to a Nash equilibrium linearly unstable under the standard replicator dynamics.

Lemma 4 *For $N = 2$, that is for two player games, a rest point $\bar{x}$ of the replicator dynamics (8) is linearly unstable if and only if it is linearly unstable for the adjusted replicator dynamics (10).*

Proof: If $\bar{x}$ is not a Nash equilibrium the statement follows from Lemma 3. It is left to show that if the linearisation of the original replicator dynamics at any Nash

equilibrium has positive eigenvalues so will the adjusted version. If $N = 2$, at any mixed Nash equilibrium, we can write J, the Jacobian of the adjusted replicator dynamics (10) as

$$J = MRU = \begin{pmatrix} \mu^1 & 0 \\ 0 & \mu^2 \end{pmatrix} \begin{pmatrix} R(x^1) & 0 \\ 0 & R(x^2) \end{pmatrix} \begin{pmatrix} 0 & u_2^1 \\ u_1^2 & 0 \end{pmatrix}, \tag{16}$$

where $u_j^i = \partial u^i(x^{-i})/\partial x^j$. The eigenvalues of RU in this case are the square roots of the eigenvalues of the matrix $R(x_1)u_2^1R(x_2)u_1^2$ (see, for example, Hofbauer and Hopkins, 2004). Therefore, the eigenvalues of J are only different by the positive multiple $\sqrt{\mu^1\mu^2}$ (note that since all payoffs are strictly positive it follows that for all solutions of (13) with $x(0) \in \Delta$, $\mu^i(0) > 0, i = 1, \ldots, n$ the terms $\mu^i(t)$, $i = 1, \ldots, n$ are uniformly bounded away from zero). We can extend this result to unstable partially mixed Nash equilibria by noting that by Lemma 3, we can partition the Jacobian into two sections corresponding to K and I. By Lemma 3, the positive eigenvalues will be contained in the section corresponding to K, which will have the same structure as (16) as $\bar{x}$ is a fully mixed Nash equilibrium with respect to its support K. Hence, it is still the case that RU has positive eigenvalues if and only if MRU has them also. ∎

The second type of boundary rest points are only Nash equilibria with respect to strategies in their support. Hence, there is at least one other strategy which gives a higher payoff. Consider the pure profile (U,U) in game (15). Deviations to D make either player worse off, so that on the face of Δ where only U and D are represented, (U,U) will be asymptotically stable. However, if for example $a = 5$ then, from any initial conditions which give positive representation to B, the replicator dynamics will diverge from (U,U).

More generally, if a rest point is not a Nash equilibrium of the game as a whole, then it must be true that for some player i and some $j \in I^i$ that $u^i(s_j, \bar{x}^{-i}) > u^i(s_k, \bar{x}^{-i})$ for $k \in K^i$. Then by the definition of the replicator dynamics (8), there exists a neighbourhood U of $\bar{x}$, and a number $d > 0$ such that

$$f_j^i(x) > d\, x_j^i \quad \forall x \in U. \tag{17}$$

In single person decision problems, this condition will hold in the neighbourhood of any profile that places probability one on an action that does not have the highest expected return. So, while we state the following result in terms of Nash equilibria, it equally implies that in single person decision problems, reinforcement learning cannot converge to a point that is not optimal.[7]

Proposition 3 *Let $\bar{x}$ be a fixed point of the replicator dynamics (8) which is not a Nash equilibrium. Then, for the reinforcement learning process defined by the choice rule (1) and updating rule (2), from any initial condition with all propensities positive, that is $q_{j0}^i > 0$ for each player i and all strategies j,* $\Pr(\lim_{n\to\infty} x_n = \bar{x}) = 0$.

[7] This result for single person decision problems and games with a unique pure Nash equilibrium can be inferred from the results of Laslier et al. (2001) and Beggs (2002). However, our general result on fixed points that are not Nash equilibria is new.

Proof. A proof can be adapted from Theorem 2 of Posch (1997), which establishes a similar result for reinforcement learning with normalisation and is based on the construction of a supermartingale. Similar arguments have also been used in Pemantle and Volkov (1999), Beggs (2002), and Lamberton et al. (2002). Suppose that, first, by relabelling if necessary, $\bar{x}^1_1 = 0$ and that (17) holds for $i = j = 1$ and some $d > 0$ in a neighbourhood U of $\bar{x}$. We have, $E[u^1_n|\mathcal{F}_n] = x^1_n \cdot u^1(x^{-1}_n)$, $E[\sigma^1_{1n}|\mathcal{F}_n] = x^1_{1n} u^1(s^1_1, x^{-1})$ and $f^1_1(x) = E[\sigma^1_{1n} - x^1_1 u^1_n|\mathcal{F}_n]$.

The proof is by contradiction. Suppose that in fact $\Pr(\lim_{n\to\infty} x^1_{1n} = 0) > 0$. Then one can choose an $L > 0$ such that

$$\Pr\left(\{\lim_{n\to\infty} x^1_{1n} = 0\} \cap \{x_l \in U, \forall l > L\}\right) > 0 \tag{18}$$

We have assumed that all payoffs are bounded above by M_1. Note that, due to our assumption that $q^1_{j0} > 0$ for all j, $x^1_{1n} = q^1_{1n}/Q^1_n \geq q^1_{10}/Q^1_n \geq q^1_{10}/(nM_1 + Q^1_0) > 0$. Since, clearly, $\sum_{n=0}^{\infty} q^1_{10}/(nM_1 + Q^1_0) = \infty$, we also have $\sum_{n=0}^{\infty} x^1_{1n} = \infty$. Hence, by the conditional Borel-Cantelli lemma (see e.g. Durrett, 1991, p. 207), player 1 chooses action 1 an infinite number of times and, consequently, $\Pr(\lim_{n\to\infty} q^1_{1n} = \infty) = 1$. Let $E_n = \{q^1_{1n} > (M_1)^2/d\}$. Since by the above argument $\lim_{n\to\infty} P(E_n) = 1$ there is an L' such that

$$\Pr\left(\{\lim_{n\to\infty} x^1_{1n} = 0\} \cap \{x_l \in U, \forall l > L\} \cap E_{L'}\right) > 0 \tag{19}$$

We define the stopping time

$$\tau = \begin{cases} \text{argmin}_{l>L} x_l \notin U & \text{if there is an } l > L \text{ s.t. } x_l \notin U \\ \infty & \text{otherwise} \end{cases}$$

which is the first time after L that x_l leaves U. Let $G_n = \{n < \tau\} \cap E_{L'}$ denote the set of all paths in $E_{L'}$ for which the process stayed in U from time L until n. Since $G_n \supset G_{n+1}$ and since for $n > L'$ the sets G_n are $\mathcal{F}_n$ measurable we have for $n > L'$

$$E\left[1_{G_{n+1}} \frac{1}{x^1_{1n+1}} - 1_{G_n} \frac{1}{x^1_{1n}}|\mathcal{F}_n\right] \leq 1_{G_n} E\left[\frac{1}{x^1_{1n+1}} - \frac{1}{x^1_{1n}}|\mathcal{F}_n\right] = 1_{G_n} E\left[\frac{Q^1_n + U^1_n}{q^1_{1n} + \sigma^1_{1n}} - \frac{Q^1_n}{q^1_{1n}}|\mathcal{F}_n\right]$$

$$= 1_{G_n} E\left[\frac{U^1_n q^1_{1n} - \sigma^1_{1n} Q^1_n}{q^1_{1n}(q^1_{1n} + \sigma^1_{1n})}|\mathcal{F}_n\right] \leq 1_{G_n} \frac{1}{q^1_{1n}} \left(E[U^1_n|\mathcal{F}_n] - E[\sigma^1_{1n}|\mathcal{F}_n] \frac{Q^1_n}{q^1_{1n} + M_1}\right) = \%.$$

Now, on U, from (17), we have $E[\sigma^1_{1n}|\mathcal{F}_n] > (E[U^1_n|\mathcal{F}_n] + d)\, x^1_{1n}$ and substituting x^1_{1n} with q^1_{1n}/Q^1_n,

$$\% \leq 1_{G_n} \frac{1}{q^1_{1n}(q^1_{1n} + M_1)} \left(E[U^1_n|\mathcal{F}_n] M_1 - d\, q^1_{1n}\right) \leq 1_{G_n} \frac{1}{q^1_{1n}(q^1_{1n} + M_1)} \left((M_1)^2 - d\, q^1_{1n}\right).$$

On G_n the last term is negative for all $n > L'$ such that $1_{G_n}\, 1/x^1_{1n}$, $n \geq L'$, is a non-negative supermartingale. Hence, $\lim_{n\to\infty} 1_{G_n}\, 1/x^1_{1n} < \infty$ with probability one. But this is a contradiction to (19). ∎

4 An Application

In this section we give an example of how our main result can be used to gain new insights into the behaviour of reinforcement learning. The idea of rescaled partnership games was introduced by Hofbauer and Sigmund (1998). These games can be considered as a subset of the class of potential games identified by Monderer and Shapley (1996). As the work of Sandholm (2002) shows, the convergence of evolutionary processes such as learning in this type of game has a number of interesting economic applications including congestion pricing. Partnership games are games of common interest and/or coordination. Rescaled partnership games are games that can be made into partnership games by simple linear transformations.

Definition 2 *A two player game with payoff matrices* (A, B) *is a partnership game if* $A = B^T$. *It is a rescaled partnership game if there exist constants* c_j, d_i *and* $\alpha > 0, \beta > 0$ *such that*

$$a'_{ij} = \alpha\, a_{ij} + c_j, b'_{ji} = \beta\, b_{ji} + d_i. \tag{20}$$

and the resulting transformed game (A', B') *is a partnership game.*

Hofbauer and Sigmund, (1998, Theorem 11.2.2)) show that rescaled partnership games possess a potential, given by $V(x) = x^1 \cdot A'x^2$, which they use to show that for these games the standard replicator dynamics will converge to a Nash equilibrium. As Monderer and Shapley (1996) show, fictitious play also must converge in potential games. Using the results of the previous section, we are able to show a new and slightly stronger result for rescaled partnership games. Because we can rule out convergence to any mixed strategy equilibria, including those with less than full support found on the boundary, reinforcement learning must converge to a pure strategy equilibrium.[8]

Proposition 4 *In generic rescaled partnership games, reinforcement learning defined by choice rule (1) and updating rule (2), from any fully mixed initial conditions converges with probability one to a pure Nash equilibrium.*

Proof: Generic partnership games have a finite number of Nash equilibria, at least one of which is pure (this follows from Lemma 2.1 of Monderer and Shapley, 1996). So, the rest points of the replicator dynamics will also be finite in number. It is easy to verify from the definition (20) that $\alpha R(x^1)Ax^2 = R(x^1)A'x^2$. That is, the replicator dynamics for the original game are the same as for its rescaling up to a positive multiplicative constant. Taking the function $V(x)$ we have for the modified dynamics (13), given that $R(\cdot)$ is positive semi definite,

$$\dot{V}(x) = \mu^1 A'x^2 \cdot R(x^1)Ax^2 + \mu^2 x^1 A' \cdot R(x^2)Bx^1 = \frac{\mu^1}{\alpha} A'x^2 \cdot R(x^1)A'x^2 + \frac{\mu^2}{\beta} A'^T x^1 \cdot R(x^2)A'^T x^1 \geq 0$$

[8] A similar result for stochastic fictitious play is in Hofbauer and Hopkins (2004). These predictions are tested experimentally in Duffy and Hopkins (2004).

with equality only at the fixed points of the replicator dynamics. Thus, $V(x)$ is a Liapunov function for those dynamics. Then by Corollary 6.6 (Benaïm, 1999), the learning process converges with probability one to one of these equilibrium points. It remains to establish that it does not converge a) to any mixed strategy equilibria or b) any pure strategy profiles that do not represent pure Nash equilibria.

The linearisation of the adjusted replicator dynamics (10) at a mixed equilibrium will have the form (16). Hofbauer and Hopkins (2004) show that at any mixed equilibrium of a partnership game the matrices of the form RU have both positive and negative eigenvalues. This implies that any mixed equilibrium of a rescaled partnership game is a saddlepoint for the standard replicator dynamics. Then by Lemmas 2 and 4, it is also unstable for the augmented dynamics (13).

We can extend this argument to any partially mixed equilibrium with less than full support. Note that if each player i uses $k < m^i$ strategies with positive probability, the resulting game is also a partnership game. Hence any mixed strategy equilibrium is a saddlepoint by the above argument. So, by Proposition 2 it is a limit point for reinforcement learning with probability zero. Finally, by Proposition 3, the learning process does not converge to any vertex which is not a Nash equilibrium. ∎

2×2 games, that is, two player games with two strategies per player, have been the class of games subject to the most analytical and experimental investigation in the recent literature. Behaviour in these simple games under other learning models has been well understood for some time. For example, Ellison and Fudenberg (2000) summarise the known results for stochastic fictitious play. They note that generically 2×2 games fall into one of two classes, in our current terminology, rescaled partnership games and rescaled zero sum games.[9] The latter are games that can be made into zero sum games by linear transformations of the form (20). In 2×2 games, they are those games which have a unique Nash equilibrium in mixed strategies. In 2×2 games, stochastic fictitious play always converges to a rest point corresponding to a Nash equilibrium.

Since we have just established a result for rescaled partnership games, it might seem that it would be possible to obtain a similar convergence result for reinforcement learning. This hope is reinforced by Beggs' (2002) recent result that the ER model of reinforcement learning must converge to the Nash equilibrium in 2×2 constant sum games with a unique equilibrium. It would seem a simple matter to extend his result to rescaled zero sum games. Unfortunately, this is not the case. The standard replicator dynamics (8) are not affected by rescaling, but this is not true for the adjusted version (10) or indeed for the dynamic system (13). Thus, though the results in this paper have extended our understanding of reinforcement learning, we still lack complete results on its behaviour in quite simple cases. There is still much to learn.

[9] Ellison and Fudenberg (2000) use the terms "games of conflict" and "games of coordination".

References

Arthreya, K.B. and P. Ney (1972). *Branching Processes*. New York: Springer-Verlag

Arthur, W.B. (1993). "On Designing Economic Agents that Behave like Human Agents," *J. Evol. Econ.*, **3**, 1-22.

Arthur, W. B., Y. M. Ermoliev, and Y. M. Kaniovski (1988). "Nonlinear adaptive processes of growth with general increments. Attainable and unattainable components of terminal set," Technical Report WP-88-86, IIASA, Laxenburg, Austria.

Beggs, A. (2002). "On the Convergence of Reinforcement Learning", working paper, University of Oxford.

Benaïm, M. (1999). "Dynamics of Stochastic Algorithms," in *Séminaire de Probabilités XXXIII*, J. Azéma et al. Eds, Berlin: Springer-Verlag.

Benveniste, A., M. Métivier, and P. Priouret (1990). *Adaptive Algorithms and Stochastic Approximations*. Berlin: Springer-Verlag.

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

Brandière, O. (1998). "The Dynamic System Method and the Traps," *Adv. Appl. Prob.*, **30**, 137-151.

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

Duffy, J., and E. Hopkins (2004). "Learning, Information and Sorting in Market Entry Games: Theory and Evidence," forthcoming *Games Econom. Behav.*

Durham, S.D., Flournoy, N., and Li, W. (1998): "A Sequential Design for Maximizing the Probability of a Favourable Response," *The Canad. J. of Statistics*, **26**, 479-495.

Durret, R. (1991). *Probability: Theory and Examples*. Pacific Grove, California: Wadsworth.

Ellison, G., and Fudenberg, D. (2000). "Learning Purified Mixed Equilibria," *J. Econom. Theory*, **90**, 84-115.

Erev, I., and Haruvy, E. (2001). "Variable Pricing: a Customer Learning Perspective", working paper.

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**, 848-881.

Hofbauer, J., and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge, UK: Cambridge University Press.

Hofbauer J., and Hopkins, E. (2004). "Learning in Perturbed Asymmetric Games," forthcoming *Games Econom. Behav.*

Hopkins, E. (2002). "Two Competing Models of How People Learn in Games," *Econometrica*, **70**, 2141-2166.

Hopkins, E. (2003). "Adaptive Learning Models of Consumer Behavior", working paper.

Ianni, A. (2000). "Reinforcement Learning and the Power Law of Practice: some Analytic Results," working paper.

Lamberton, D., G. Pagès and P. Tarrès (2002). "When Can the Two-Armed Bandit Algorithm be Trusted?" forthcoming *Ann. Applied Probability*

Laslier, J-F., R. Topol, and B. Walliser (2001). "A Behavioral Learning Process in Games," *Games Econom. Behav.*, **37**, 340-366.

Monderer, D., and Shapley, L. (1996). "Potential games," *Games Econom. Behav.*, **14**, 124-143.

Pemantle, R. (1990). "Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations," *Ann. Probability*, **18**, 698-712.

Pemantle, R., and Volkov, S. (1999). "Vertex-Reinforced Random Walk on Z Has Finite Range," *Ann. Probability*, **27**, 1368-1388.

Posch, M. (1997). "Cycling in a Stochastic Learning Algorithm for Normal Form Games," *J. Evol. Econom.*, **7**, 193-207.

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

Rustichini, A. (1999). "Optimal Properties of Stimulus-Response Learning Models," *Games Econom. Behav.*, **29**, 244-73.

Sandholm, W. (2002). "Evolutionary Implementation and Congestion Pricing," *Rev. Econom. Studies*, **69**, 667-689.

Sarin, R., and F. Vahid (1999). "Payoff Assessments Without Probabilities: a Simple Dynamic Model of Choice", *Games Econ. Behav.*, **28**, 294-309.

Tarrès P. (2001). "Pièges des algorithmes stochastiques et marches aléatoires renforcées par sommets" PhD Thesis, Ecole Normale Supérieure de Cachan.

Weibull, J.W. (1995). *Evolutionary Game Theory*, Cambridge, MA: MIT Press.