ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

ON THE CONVERGENCE OF REINFORCEMENT LEARNING

A.W. Beggs

Number 96

March 2002

Manor Road Building, Oxford OX1 3UQ

On the Convergence of Reinforcement Learning

A. W. Beggs
Wadham College
Oxford
OX1 3PN
UK

Current Version: March 2002

Abstract

This paper examines the convergence of payoffs and strategies in Erev and Roth's model of reinforcement learning. When all players use this rule it eliminates iteratively dominated strategies and in two-person constant-sum games average payoffs converge to the value of the game. Strategies converge in constant-sum games with unique equilibria if they are pure or in 2×2 games also if they are mixed. The long-run behaviour of the learning rule is governed by equations related to Maynard Smith's version of the replicator dynamic. Properties of the learning rule against general opponents are also studied. In particular it is shown that it guarantees that the lim sup of a player's average payoffs is at least his minmax payoff.

JEL Classification Numbers: C72, D83

Keywords: reinforcement learning, games

I am grateful to Josef Hofbauer, Ed Hopkins and seminar participants at Oxford and Warwick for helpful comments.

1. Introduction

This paper studies the convergence properties of a class of naïve reinforcement learning models in games. These were originally proposed by Roth and Erev (1995) and Erev and Roth (1998) as a means of modelling the observed behaviour of subjects in experiments on games. They argue that their behaviour can be well approximated by a simple model in which players tend to put more weight on strategies that have enjoyed past success, as measured by the cumulated payoffs they have achieved. Harley (1981) proposed a similar model in a biological context.

This model has considerable attraction as a simple model of boundedly-rational players. The amount of information players are assumed to gather is small. Players need only observe their realised payoffs and may not be aware they are even playing a game, let alone the payoff matrix of their opponent or even their actions. It also builds in a certain amount of inertia, in that players are slow to switch from actions that have performed well in the past, which seems a plausible feature of learning. Despite this, little is known about the analytical properties of the model.

This paper aims to reduce this gap. It studies the behaviour of players' payoffs and strategies when other players use the same rule and when they do not.

It shows that when all players use this rule dominated strategies are iteratively deleted. In addition in two-person constant-sum games, players' average payoffs converge to the value of the game. It also shows that their strategies converge to equilibrium in constant-sum games if there is a unique pure strategy equilibrium or in 2×2 games if there is a unique-mixed strategy equilibrium.

In the course of the analysis it is shown that the long-run behaviour of the players' strategies is governed by equations related to Maynard Smith (1982)'s version of the replicator dynamic. This may be of independent interest since in 2×2 constant-sum games mixed equilibria are stable under it, while the ordinary replicator dynamic cycles around it. There have, however, been few derivations of the Maynard Smith dynamic from primitive assumptions. Bjornerstedt and Weibull (1996) and Hofbauer and Schlag (2000), which justify versions from models of imitation, seem the only examples. The current dynamic is similar to the Maynard Smith dynamic and shares its convergence properties.

When opponents do not use the Erev and Roth rule, it is shown that a player using it learns not to play dominated strategies. It also shown that if some strategy of the stage game, pure or mixed, would guarantee a certain payoff against the opponent's play on average, in a sense made precise, then a player using this learning scheme will not do worse than this, in the sense that the lim sup of his average payoffs will be at least this. In particular, his average payoffs cannot be kept away from his minmax payoffs. It is shown, however, that a clever player can exploit the inertia in the Erev and Roth scheme and force the lim inf of the player's average payoffs below his minmax value. In constant-sum games, however, it is not clear that this is of much benefit to the other player and conditions are given on their strategies which rule this out.

There are of course many other models of learning. Much attention recently has

focussed on fictitious play and variations of it. For example Benaïm and Hirsch (1999) study convergence of strategies in games with randomly perturbed games. Although fictitious play itself has poor optimality properties, smoothed versions have quite good properties — stronger than those mentioned above for the procedure studied here. Fudenberg and Levine (1998) provide a good summary of this work. Hart and Mas-Collel (2000) (2001a,2001b) study procedures based on 'regrets', which in some cases share these properties. On the other hand fictitious play and regret-based strategies require greater knowledge of the game and sophistication. Auer et al (1998) and Hart and Mas-Collel (2001b) study versions which do not require knowledge of the game, but still require some sophistication. The feature of inertia which these procedures lack, but is shared by Hart and Mas-Collel (2000) where it also can be exploited by clever players, also seems an appealing feature of a model of learning.

In any case it is not argued that this is the only plausible model of learning, only that it is of enough interest to be make it worth further study. Camerer and Ho (1999) suggest that both it and fictitious play have feature which match the data and present a synthesis.

Börgers and Sarin (1997) analyse a reinforcement learning model and discuss its relationship to the so-called replicator dynamic used in biology (see for example Hofbauer and Sigmund (1988)). Their work is discussed further in Section 6. The principal difference is that in the model of Erev and Roth, less weight is placed on current payoffs as experience accumulates. This implies that the long-run behaviour of the two models is rather different. Rustichini (1999) analyses some properties of reinforcement learning in a single-player context. He assumes that the player faces a stationary environment, which makes it inappropriate when other players may be changing their play. His work is discussed in more detail in Sections 3 and 4. His results on reinforcement learning can be obtained from the results in Section 4. He also compares reinforcement learning with other models of learning not studied here.

Laslier, Topol and Walliser (2001) study the Erev and Roth model under a different name. In the single-player context they show that if the environment is unchanging, the player will learn to play the action with highest expected payoff. This is a special case of the results obtained here. They also show that if a two-person game has a strict Nash equilibrium then there is a positive probability that play will converge to it, if both players learn according to their scheme. Their results leave open the possibility that even when there is a unique strict Nash equilibrium, play does not always converge to it. Their results therefore do not imply the convergence results here, even when the domains of the papers overlap, as conditions are given for convergence with probability 1 here. They do not consider the convergence of payoffs.

Posch (1997) studies the convergence of a related learning model of Arthur (1993) in 2×2 games. In this model equilibrium play cycles round a mixed-equilibrium with positive probability rather than converging. His work is discussd in more detail in Section 5. He does not discuss convergence of payoffs. Ianni (2001) provides some results on convergence to strict equilibria with positive probability for this model,

similar to those of Laslier et al. (2001) for the Erev and Roth model.

Hopkins (2000) argues that some perturbed forms of reinforcement learning are similar to some forms of perturbed fictitious play, in that they have similar local stability properties about rest points. The perturbations imply that in general these models cannot converge to equilibrium and any rest points do not correspond to equilibria. This contrasts with the unperturbed versions studied here, where convergence to equilibrium is possible.

Section 2 outlines the basic model of reinforcement learning. Section 3 draws some connections with the theory of urn models and presents results on its behaviour in a single-player context. In particular, it shows that a player will always learn to play a dominant strategy and will learn not play dominated strategies. If all players use this learning procedure, they will in in the long-run not play strategies that are ruled out by iterative deletion of strictly dominated strategies.

Section 4 analyses the performance of reinforcement learning when other players do not necessarily use the same rule and obtains the results on limiting average payoffs discussed above. Section 5 considers behaviour in a two-person constant-sum game when both players learn according to the reinforcement procedure and demonstrates convergence of average payoffs to the values of the game and of strategies to their equilibrium values when there is a unique equilibrium in the case of pure equilibria in general and mixed equilibria in 2×2 games.

Section 6 discusses some variations on the basic model. It discusses adding the idea of 'forgetting', or putting more weight on more recent observations, to the model. Although payoffs converge at a reasonable rate in simulations of the model of Section 5, and this is presumably what players care about, convergence of strategies to mixed equilibria can be very slow. Adding forgetting, although it leads to different asymptotic behaviour, improves the finite-horizon predictions. The section also discusses briefly the idea of reference points and alternative functional forms.

2. The Learning Model

The basic structure of the Erev and Roth (ER for short) model can be described simply. Suppose that there are m possible actions to be taken. At each stage $n = 1, 2, \ldots$, the decision-maker must decide which to choose. To each action is associated a reinforcement level, $A_i(n)$ to action i at stage n. There are some initial reinforcement levels, $A_i(0)$, $i = 1, \ldots, m$. At each stage the chosen action's reinforcement is increased by the payoff it obtains. Denoting by $\pi_i(n)$ the payoff of action i at time n, one can therefore write

$$A_i(n+1) = \begin{cases} A_i(n) + \pi_i(n+1), & \text{if action } i \text{ is chosen;} \\ A_i(n), & \text{otherwise} \end{cases} \tag{1}$$

That is $A_i(n)$ represents the cumulated payoffs obtained by action i when it is chosen. The probability that action i is chosen at stage $n+1$ is simply

$$P_i(n+1) = \frac{A_i(n)}{\sum_j A_j(n)} \tag{2}$$

It will be assumed that all payoffs, $\pi_j(n)$, and initial reinforcements, $A_j(0)$, are strictly positive so the denominator is non-zero. Note that this implies that there is initially a positive probability of playing every action.

This model embodies some simple features of reinforcement learning. In the first place, the higher the payoff to a chosen action the more it is reinforced and the more likely it is to be chosen in future. In the second place, as more experience accumulates learning slows down, which appears to be true in experiments by psychologists (the so-called 'power law of practice' see Erev and Roth (1998)). For large t the impact of any particular stage's choice on the total cumulated payoffs for any action will be small and so choice probabilities will change little. Some of the drawbacks and possible modifications of the model will be discussed in Section 6.

This model can be linked to so-called urn models, which have been much studied in the probability literature. In these an urn contains balls of various different colours. At each stage a ball is drawn from it at random and then a number of balls are added, the number and colour of which depend on the colour of the ball drawn. One can then study how the long-run composition of the urn depends on the rules for adding new balls.

The model set out above can clearly be regarded as an urn scheme. Each possible action can be thought of as a colour. The total reinforcement for each action can be interpreted as the number of balls of that colour in the urn. At each stage, (2) specifies that the probability that a ball of a certain colour is drawn is simply the fraction of balls of that colour in the urn. (1) specifies that the ball drawn be returned and a certain number of balls of the same colour be added to the urn. The initial reinforcements are simply the initial numbers of each colour of ball in the urn. The reinforcements need not be integers, so the interpretation is not exact, but this is unimportant mathematically. This link will be used in subsequent sections.

For future reference, two simple results will be stated here. More structure will be put on the payoffs each period in future sections but a standing assumption will be that $\pi_i(n)$ is bounded away from zero and infinity: $0 < k_1 \leq \pi_i(n) \leq k_2$ for all i and n, for some k_1 and k_2. The role of this assumption is simply to bound the growth of the reinforcements above and below. It is innocuous in the applications studied.

Now from (2), regardless of history, the probability that action i is chosen at time $n+1$ is at least

$$\frac{A_i(0)}{\sum_j A_j(0) + nk_2} \tag{3}$$

— which corresponds to i never having been chosen previously. Summed over t this diverges, so by the (conditional) Borel-Cantelli lemma

Lemma 1 *Each action is chosen infinitely often with probability 1.*

Since payoffs are bounded below, an immediate consequence, which will be of use later, is

Lemma 2 *A_i tends to infinity for each i with probability 1.*

According to Lemma 1, no matter how bad an action is the decision-maker is will experiment with it infinitely often. Nevertheless, experimentation may become increasingly rare and the next sections investigate the convergence of choice probabilities and frequencies of actions.

3. Single Player Learning

This section uses the theory of urn processes to study the properties of reinforcement learning. This theory is most complete in the case of two colours, so attention will initially be limited to the case of two actions. The results obtained will then be used to derive more general results.

3.1 The Case of Two Actions

This sub-section considers the case of two actions in some detail to develop intuition that will be needed for future discussion. All that is needed directly for Section 3.2 is Theorem 3, so the reader can skip to that if desired.

Suppose therefore that $m = 2$ and suppose further for the moment that action 1 has constant payoff α_1 and action 2 constant payoff α_2, where $\alpha_1 \geq \alpha_2$. The assumption of constant payoffs will be relaxed later.

If $\tilde{N}_i(n)$ is the number of times action i has been played up to and including stage n, then $A_i(n) = \alpha_i \tilde{N}_i(n) + A_i(0)$. Let $N_i(0)$ be defined by $A_i(0) = \alpha_i N_i(0)$ and $N_i(0) = \tilde{N}_i(n) + N_i(0)$. $N_i(0)$ can be regarded as the initial number of times action i has been played (possibly non-integral). If $x_n = N_1(n)/N_1(n) + N_2(n)$ is the fraction of times when action 1 has been played, then from (2) the probability that action 1 is chosen is

$$f(x_n) = \frac{\alpha_1 x_n}{\alpha_1 x_n + \alpha_2(1 - x_n)} \tag{4}$$

Now if the frequency with which action 1 is played converges to x, the probability that action 1 is played at any given stage will converge to $f(x)$, so it is natural to suppose that x must be a fixed point of f.

Hill, Lane and Sudderth (1980) prove the following very general result (interpreted in the current framework):

Result A(Hill et al.) *Suppose action 1 is played at stage $n+1$ with probability $f(x_n)$, where f is continuous*

(a) The frequency with which action 1 is played, x_n, converges with probability 1 to a fixed point of f.

(b) If x is an unstable fixed point of f in the sense that in a neighbourhood of x $f(p) > p$ if $x > p$ and $f(p) < p$ if $x < p$, then if $0 < x_0 < 1$ there is probability zero that x_n converges to x.

The condition in (b) simply says that the slope of f at x is greater than 1 when f is differentiable.

When $\alpha_1 = \alpha_2$, that is payoffs to the two actions are equal, $f(x) = x$ and so every fixed point of $[0, 1]$ is a fixed point. This corresponds to Polya's urn, in which c balls of the colour drawn are added to the urn, for some number c (the same for both colours). In fact it can be shown that x_n is a martingale and that x_n does not converge to a constant (unless $x_0 = 1$ or 0) but rather

Theorem 1 *If $\alpha_1 = \alpha_2$ and $0 < x_0 < 1$, then the frequency with which action 1 is played converges a random variable with mean equal to x_0, with a beta distribution on $[0, 1]$.*

Note that a beta distribution is absolutely continuous on $[0, 1]$, so there is probability zero that x_n converges to any particular value, even though its mean is a constant. This is an example where averaging sample paths to study the long-run behaviour of the system would be misleading.

The Polya urn is somewhat degenerate, in that small perturbations of f will radically change the set of fixed points. On the other hand it is of interest as the case with stochastic but equal reinforcements would arise if one's opponent is playing a mixed strategy. Further discussion will be given in Section 6.

When $\alpha_1 > \alpha_2$, then f has fixed points at 1 and 0, but unless $x_0 = 0$ so the player never tries action 1, by (b) of the Result above there is probability 1 that x_n will converge to 1. Hence

Theorem 2 *If $\alpha_1 > \alpha_2$ and $x_0 > 0$, then the probability that the decision-maker chooses the superior action converges to 1.*

In other words, even though when action 1 has been rarely tried there is very little experimentation with it (since an action is only reinforced when it is tried), ultimately the decision-maker will learn to play it.

This result can be generalised to allow action 1 merely to dominate action 2 on average and indeed allow the process to be non-stationary, which is of importance in learning in games since opponents' strategies may be changing. Let $\mathcal{F}_n$ denote the sigma-field generated by choices up to time n and note that the payoff to an action is only relevant if it is played.

Theorem 3 *Suppose that $E(\pi_1(n+1)|\mathcal{F}_n, \text{action 1 is chosen at time } n+1) > \gamma E(\pi_2(n+1)|\mathcal{F}_n, \text{action 2 is chosen at time } n+1)$, where $\gamma > 1$ is a constant, for all n. Then with probability 1, the probability that the decision-maker plays action 1 converges to 1.*

The proof, and the result, are an easy generalisation of a result in Pemantle and Volkov (1999). The details can be found in the Appendix but the idea is that since the payoff to action 1 is on average greater than that to action to 2, the reinforcement to action 1 will grow at a faster rate. More precisely, for $A_1(n)$ and $A_2(n)$ large enough

(which will always happen eventually by Lemma 2), $A_2(n)^\epsilon/A_1(n)$, for $0 < \epsilon < \gamma$. is a positive supermartingale and hence convergent. It follows from this (choose $1 < \epsilon < \gamma$) that $A_2(n)/A_1(n)$ tends to zero, since $A_2(n)$ tends to infinity (by Lemma 2). Hence the relative probability that action 1 is played tends to 1.

By a Strong Law of Large Numbers for dependent random variables (for example Hall and Heyde (1980) p. 36–37, — note that here the random variables are bounded), it follows that

Corollary *With probability 1 the empirical frequency with which the decision-maker plays action 1 converges to 1.*

The speed of learning under this model is, however, slow. For example suppose action 1 has constant payoff 1.5 and action 2 constant payoff 1. Suppose further that the initial reinforcements for the 2 actions are 1 and 1.5 respectively and that the rule is allowed to run for 10,000 periods. In 100 simulations, the average probability of playing action 2 in period 10,000 was 0.15 but there was considerable variation, the coefficient of variation being 1.4. If the payoffs to the two actions were 2 and 1 (other parameters unchanged) the average probability of playing action 2 was 0.05 with coefficient of variation 2.75.

Two considerations seem to drive these results. In the first place, random chance can lead to the wrong action being reinforced initially. After the process has been running for a while chance effects even out, but since payoffs are small compared to accumulated reinforcements, it moves very slowly. Secondly, one can be quite far away from the optimal action in probability terms yet be close in payoffs. For example if action 1 has payoff 3 and action 2 payoff 2, if one plays action 2 with probability zero one obtains a payoff of 3. If one plays it with probability 0.2, the payoff is 2.8 which is close in proportionate terms. Note that (as discussed in Section 6) only proportional differences between payoffs affect the algorithm.

A similar argument to that in the proof of Theorem 3 shows that if $E(\pi_1(n+1)|\mathcal{F}_n, \text{action 1 is chosen at time } n+1) < \gamma' E(\pi_2(n+1)|\mathcal{F}_n)$ (again also conditional on the actions being played), if $\epsilon > \gamma'$, $A_1(n)/A_2(n)^\epsilon$ is a positive supermartingale eventually and thus convergent. Putting these two observations suggests in the case of two actions with constant ratio of payoffs γ, the rate of convergence of inferior action's probability of being played to zero is $1/n^{1/\gamma}$. So to obtain convergence at rate at least $1/\sqrt{n}$ one requires $\gamma \geq 2$, which is not inconsistent with the simulations mentioned.

3.2 Many Actions

The reinforcement rule has the property of Independence of Irrelevant Alternatives: given that one of actions 1 or 2 is chosen by the decision-maker the conditional probability that action 1 is chosen is from (2) simply

$$\frac{A_1(n)}{A_1(n) + A_2(n)}$$

That is the reinforcements to the other actions are irrelevant. It follows that in order to study the relative choice probabilities of any pair of actions one can consider this pair in isolation and apply the results of the previous sub-section. Note that by Lemma 1, each action is played infinitely often, and simply define stages in the new process to be when one of the two actions is played in the old process.

It follows that

Theorem 4 *Suppose that, for some i, $E(\pi_i(n+1)|\mathcal{F}_n, \text{action } i \text{ is chosen at time } n+1)$ $> \gamma E(\pi_j(n+1)|\mathcal{F}_n, \text{action } j \text{ is chosen at time } n+1)$, where $\gamma > 1$ is a constant, for all n. Then with probability 1, the probability that the decision-maker plays action j converges to zero. The same is true of the empirical frequency of play.*

For by Theorem 3 (and its Corollary) applied to the pair i and j, the relative probability that action j is played when one of i and j is played converges to zero. The same is clearly true when one does not condition on the events that one of i and j is played.

Now in a game with finitely many actions and players, if action i strictly dominates action j the assumptions on payoffs are clearly met, so one obtains

Theorem 5 *In a game with finitely many actions, and players if a player learns according the ER scheme then,*
(a) With probability 1, the probability and empirical frequency that he plays any action that is strictly dominated by another pure strategy converges to zero.
(b) Hence if he has a strictly dominant strategy, with probability 1, the probability and empirical frequency with which he plays that action converges to 1.

That a player will learn to play a dominant strategy seems a minimal requirement of a reasonable learning rule. If all players in a finite game learn according to the ER scheme, then it is easy to extend the Theorem to show that eventually, players will not play strategies that are deleted by iterated domination of dominated strategies. For eventually, the probability that any player plays a dominated strategy will be small and so players will learn not to play those that are dominated once these are deleted. A proof can be found in the Appendix:

Corollary *If all players play according to the ER rule, then with probability 1 the probability and empirical frequency with which any strategy which is eliminated by iterative deletion of strategies strictly dominated by other pure strategies is played tends to zero.*

Laslier et al. (2001) show that a single player playing against nature in an unchanging environment will converge to the action with highest expected payoff, which is a special case of the result above. Rustichini (1999) shows that a single player learning in a stationary Markov environment will converge to the action with highest average

payoff. His result does not imply the result above as it assumes a stationary environment, which the result above does not and is inappropriate when other players' play may be changing. On the other hand, his result does not assume that one action is dominant all the time, only so on average. His result can be obtained from the results of the next section, where it is discussed further. His proof does not in fact seem quite complete. He appeals to results on stochastic approximation. Ruling out convergence to inferior actions requires ruling out convergence to unstable points of a certain differential equation. As discussed in Section 6 before Theorem 12, however, standard results guaranteeing this are not applicable as these points lie on the boundary of the state space, where the variance of the process goes to zero. Standard results require it to bounded away from zero.

One can also show that if a strategy is dominated by a mixed strategy it will be eliminated in the limit. The argument above does not prove this since it is not enough to consider pairs of pure strategies. One can however show that if i is dominated by a mixed strategy then $A^{\epsilon}(n)/A_1^{\sigma_1}(n)\dots A_m^{\sigma_m}(n)$ is a positive supermartingale, where $\sigma_1,\dots,\sigma_n$ is the dominating mixture of strategies $1,\dots,m$ and $\epsilon<\gamma$, where $\gamma>1$ is such that the mixed strategy always earns γ times more than i. It follows that

Theorem 6 *In a game with a finitely players and actions*

(a) With probability 1, the probability and empirical frequency that a player who learns according to the ER scheme plays an action dominated by a mixed strategy converges to zero.

(b) If all players learn according to the ER scheme, then with probability 1, the probability and empirical frequency that any of them plays a strategy eliminated by iterated deletion of dominated strategies tends to zero.

4. Reinforcement Learning in Games: General Results

This section evaluates the performance of reinforcement learning in games. The main application to bear in mind is a finite two-person constant-sum game. The analysis proceeds somewhat more generally, however, and it is simply assumed that the player under consideration learns according to the ER rule. Other players may play differently.

If a player is rational and knows the payoffs of the game, he can ensure that his long-run average payoff is at least his minmax value simply by playing his minmax strategy. In the current framework it is not assumed that the player is rational or that he knows the payoffs of the game, but it will be shown that the lim sup of his long-run average payoffs will be at least his minmax value. In fact a more general result will be proved: it will be shown that if there is some payoff that a strategy will guarantee 'on average' against the play of the opponents, then the ER scheme will achieve that. Conditions will also be given when lim sup can be replaced by lim.

In contrast to Section 3, this Section does not consider the convergence of strategies only of payoffs. The convergence of strategies is considered in Section 5 in the case where both players use the ER rule in a constant-sum game. In addition, as will be

discussed after the statement of Theorem 8, while the framework considered here is some ways more general than that studied in Section 3, the results of Section 3 are sharper for the case of dominant strategies.

Since the focus is on the long-run average payoffs, it is useful to consider the time averages of the reinforcements, $a_i(n) = A_i(n)/n$ for $i = 1, \ldots, m$, instead of the reinforcements themselves. Let $A(n) = A_1(n) + \ldots + A_m(n)$ be the total reinforcement received by the player up to time n. Its time average $a(n) = A(n)/n$ is not necessarily equal to the total average payoff received up to time n, on account of the initial reinforcements, but asymptotically they are the same, so they will not be distinguished carefully in the discussion below. Note that the assumptions in Section 2 imply that $a(n)$ is bounded above and away from zero, a fact which will be used without comment below.

Equation (1) can be re-written in terms of average reinforcements as

$$a_i(n+1) = a_i(n) + \frac{1}{n+1}\left(\tilde{\pi}_i(n+1) - a_i(n)\right) \tag{5}$$

where $\tilde{\pi}_i(n+1)$ equals $\pi_i(n+1)$ if action i is played at stage $n+1$, 0 otherwise. Equivalently

$$a_i(n+1) = a_i(n) + \frac{1}{n+1}\left(p_i\pi_i(n+1) + u_{n+1}\right) \tag{5}'$$

where $p_i = a_i(n)/a(n)$ is the probability that action i is played at stage $n+1$ and $E(u_{n+1}|\mathcal{F}_n) = 0$, $\mathcal{F}_n$ being the σ-field generated by events up to stage n. For large n the change in a_i is small and it is plausible that the noise term washes out, so that long-run behaviour of the reinforcements is governed by

$$\frac{da_i}{dt} = -a_i(t) + p_i\pi_i(t) \tag{6}$$

where t denotes time. Note that all choice probabilities are functions of the reinforcements, so the right-hand simply depends on these.

This is a familiar idea in stochastic approximation. In this section little formal use will be made of this, rather it will used as a guide to intuition and to construction of appropriate martingales.

Consider some mixed strategy $\sigma = \sigma_1, \ldots, \sigma_m$ for player 1. The geometric average of the reinforcements corresponding to its components might be thought of as a measure its fitness. The logarithm of this is slightly more convenient to use, so consider the function

$$V = \sum_i \sigma_i \ln a_i \tag{7}$$

The convention $0 \ln 0 = 0$ is adopted. V is only well-defined on the set where $a_i > 0$ for all i with $\sigma_i > 0$. The assumptions of Section 2 imply all average reinforcements are positive for all finite n and although some care needs to be taken about behaviour near the boundary in the formal discussion, it is enough for discussion purposes to assume that all a_i are strictly positive.

If the system evolves according to (6) then the rate of change of V is

$$\dot{V} = \frac{\pi_\sigma - a}{a} \tag{8}$$

where $\pi_\sigma(t)$ is the expected payoff to σ at time t. If σ guarantees the player at least v, then V will increase if his average payoff is less than v. Since V is bounded above, his average payoff must eventually reach v.

Note that is not asserted that a itself is monotonically increasing when a is below v. $\exp(V)$ can be written as $ap_1^{\sigma_1} \dots p_m^{\sigma_m}$, where $p_i = a_i/a$ is the probability that action i is played. The assertion that V is increasing implies that either a is increasing or the player becomes closer to playing σ (as measured by the geometric average of probabilities) and so eventually his payoff will rise.

The above suggests that if σ guarantees at least v at every stage, then under the ER scheme the player's average payoff must reach v. More generally, this will hold if σ guarantees v on average, in the sense that it satisfies the following condition:

Condition M *$E(\pi_\sigma|\mathcal{F}_{n-1}) \geq v + v_n$ for all n, where $\sum_n v_n/n$ converges.*

Theorem 7 *If condition M holds then* $\limsup_{n\to\infty} a(n) \geq v$ *almost surely.*

The proof is in the Appendix. The idea is that for large n, a Taylor expansion shows that V is a submartingale plus some disturbance terms and one can make an analogous argument to the one based on the differential equation. This is a standard idea though the logarithmic form of V means that one needs to take a little extra care to show that the second-order terms in the Taylor expansion are unimportant.

Condition M is used to show that the disturbances can be neglected. If $\sum_n v_n/n$ converges then[1] $\sum_{i=1}^n v_i/n$ converges to zero almost surely, so by a strong law of large numbers[2], $\liminf_{n\to\infty} \sum_{i=1}^n \pi(n;\sigma)/n \geq v$ almost surely, where $\pi(n;\sigma)$ denotes the payoff a player would earn at stage n if he were to play σ. In other words, the player's long-run average payoff if he were to play σ forever would exceed v almost surely, which is clearly necessary for the Theorem. The condition is a standard one in the stochastic approximation literature. It can probably be weakened but it seems adequate for most examples.

Examples

1. *Minmax strategy.* If σ is the player's minmax strategy, then condition M holds with $v_n = 0$, hence the lim sup of his payoffs is at least v.
2. *Stationary Markov Environment* Suppose that the player is playing in a stationary environment with payoff π_{ij} to action i if the state is j, where j is the state of a finite irreducible, aperiodic Markov chain. This is the framework considered by Rustichini (1999). Condition M is satisfied here (see for example Duflo (1997)

[1] Use Kronecker's Lemma — see Hall and Heyde (1980) p. 31.
[2] For example Hall and Heyde p. 36.

Proposition 9.2.9) and applying the Theorem to the (pure) strategy with the highest long-run average payoff, shows that the lim sup of the player's payoffs is at least this.

3. *General Markov and mixing conditions* Condition M also holds for much more general Markov chains, even one where the transition probabilities depend on the current strategy (provided they are not too sensitive — see Benveniste et al (1990)). This latter case would be natural when the opponent is changing strategy, but verification depends of the required conditions depends on the nature of the opponent's learning process. It also holds under general mixing conditions (see Kuan and White (1994)).
4. *Periodic Play.* The condition also holds in some periodic cases. Consider for example the two-person constant-sum games shown in Figure 1 (Matching Pennies). Suppose the player is the row player. If player 2 plays strategy 1 for two periods then switches to 2 for one period, then back to strategy 1 for two periods and so on for ever, then player 1 can guarantee himself a long-run average payoff of 5/3 by playing 1 forever (and this is the best payoff from a fixed action). Condition M applied here and so lim sup of his payoffs is at least 5/3. Clearly if he knew the pattern of player 2's play he could do better, but that would require a more sophisticated (and quicker adjusting) action rule.
5. The Condition also makes clear that some degree of time variation can be allowed for. Some for example that with probability 1, there is some N such that after it other players always play some specified action (e. g. sooner or later switch to defecting in the prisoner's dilemma). Then the above allows one to conclude that the lim sup of the player's payoffs must be at least that of a best response to this. Occasional play of other strategies after this time will not perturb this result, provided they are rare enough for Condition M still to hold.

If long-run average payoffs converge, then of course one can replace lim sup by lim in all the statements above. If long-run average payoffs do not converge, it is not clear how to define players' payoffs. lim sup is certainly a good candidate for this, but a cautious player might prefer to look at the lim inf.

It is therefore of interest to ask whether one can replace lim sup by lim inf in the examples above. Without some further conditions this cannot in fact be done.

Counterexample

Consider the game in Figure 1 again. Player 1's minmax strategy is $(0.5, 0.5)$ and his minmax payoff is 1.5. Suppose that player 2 a strategy of following form: play action 1 until the probability that player 1 plays action 1 is close to 1, then switch to action 2 until the probability that player 1 plays action 1 is close to zero, then switch to action 1 and so on for ever. A strategy of this form be found that so that player 1's average payoff is below $1.5 - \epsilon$, for some $\epsilon > 0$ infinitely often.

To see this note that one can re-write equation (6) in terms of $p_i = a_i/a$ and a

$$\dot{p}_i = \frac{p_i(\pi_i - \Pi)}{a} \tag{9a}$$

$$\dot{a} = -a + \Pi \tag{9b}$$

p_i can be interpreted as the probability that the player plays action i and Π is his expected payoff if he follows this strategy.

Now if p_2 is close to zero, then its rate of change is very slow. Suppose that player 2 suddenly switches to playing strategy 2. Player 1 will only react very slowly to this, while payoffs adjust relatively quickly, and so player 1's payoffs will fall below 1.5 before he has a chance to adjust back. Player 2 then waits until he is puts almost no weight on strategy 1 and switches to it and so on. A proof using these ideas can be found in the Appendix.

This example shows that one cannot replace lim inf by lim sup in the above. Now some simple procedures, such as for example smoothed fictitious play, but not fictitious play itself, (Auer et al. (1998), Fudenberg and Levine (1998) Ch. 4) or the procedures considered by Hart and MasCollel (2001a,b) do have this property (or strictly almost in the case of smoothed fictitious play). Indeed they have the stronger one of universal (or almost in the case of smoothed fictitious play) consistency - that is one can do as well as one would if one knew the empirical distribution of play and took a best response to it. In that sense, the ER rule has less good properties than these rules. The slow-rate of adjustment to changes in strategy by the opponent, noted in Section 2, can be exploited.

On the other hand, note that the strategy played above requires player 2 to allow large fluctuations in her own payoff below 1.5, while she waits for player 1 to concentrate on one pure strategy before switching back. So if she cares about the lim inf this is undesirable.

Now if one looks at (8), the only reason V is not a global Lyapounov function is that sometimes a may lie above v. If a always lay below v, then one could conclude that a must converge to v. As before one does not require that it always be below v but only that it converge to it fairly rapidly:

Condition A *$a(n) \leq v + \eta_n$ for all n, with $\eta_n \geq 0$, and $\lim_{n\to\infty} \eta_n = 0$, where $\sum \eta_n/n$ converges.*

The condition that $\lim_{n\to\infty} \eta_n = 0$ implies that $\limsup_{n\to\infty} a(n) \leq v$.

A variation on Condition M is used in the proof:

Condition M′ *$E(\pi_\sigma|\mathcal{F}_{n-1}) \geq v + v_n$ for all n, where $\sum_n v_n/na(n-1)$ converges.*

In this condition, the disturbances are measured relative to the current level of reinforcement. If either series converges absolutely, then each condition implies the other. Condition M' holds in the examples given above for Condition M, again see the cited references, and one could have proved Theorem 7 under it, but is perhaps less

intuitive and for that reason was not used there. In the present proof, it is not clear that Condition M is enough to deal with the disturbances so M′ is used.

Theorem 8 *If condition A is satisfied and a player has some strategy that satisfies condition M′ for the same v, then* $\lim_{n\to\infty} a_n = v$.

The following are immediate corollaries:

Corollary 1 *Consider a finite two-person constant-sum game with strictly positive payoffs. If player 1 plays according to the ER rule and player 2 uses her minmax strategy or takes a myopic best response to player 1's current strategy, then player 1's long-run average payoff converges to his minmax value.*

One needs only to verify that condition A is true see Appendix — M′ is immediate. Rustichini (1999)'s result also follows

Corollary 2 *Suppose that player 1 faces an environment where if he plays strategy i he obtains payoff* π_{ij} *and the state j follows a finite irreducible, aperiodic Markov chain, then player 1's long-run average payoff converges to the payoff of the action which yields highest long-run average payoff.*

The proof is in the Appendix. One simply needs to verify condition M′ and A for the player's best strategy. Intuitively, although the state changes and so payoffs change, in a Markov chain convergence to the ergodic distribution is fast, so average payoffs converge rapidly to their long-run values.

It also straightforward to check that the result applied in the periodic case considered above.

Corollary 2 is in one sense stronger than Theorem 4, in other ways weaker. In one way it is stronger, as it applies when one action only dominates an action on average, whereas Theorem 4 requires domination at each stage. On the other hand it may be a weak result when payoffs vary. For example suppose there are two players and each has two actions: action 1 yields payoff 1 or 3 to player 1, according as player 2 chooses 1 or 2, the choice of the other player, while action 2 yields payoff 2 or 4, according as 2 chooses 1 or 2. If player learns according to the ER rule, Theorem 5 implies that the the probability that he plays action 2 converges to 1. If the game is constant-sum and player 2 plays minmax or some other strategy obeying assumption A, then one can also conclude this from Theorem 8. In the absence of any further information about the game or 2's play, however, all that follows from Theorems 6 and 7, is that the lim sup of player 1's average payoffs is at least 2.

These are of course asymptotic results. Nevertheless they seem to work well in practice. For example in simulations of the game in Figure 1 against player 2 where (i) she uses fictitious play, (ii) takes a (myopic) best response to player 1's current strategy, (iii) plays the minmax strategy, player 1's average payoff converges rapidly to 1.5. Against each opponent the ER rule was run 100 times in a run of length 10,000, with initial reinforcements $A_1(0) = 1$, $A_2(0) = 1.5$. Against (i) the mean average payoff

was 1.48, with coefficient of variation 0.04, against (ii) 1.49 with coefficient of variation 0.01, and against (iii) 1.5 with coefficient of variation 0.003.

It is not immediately apparent whether the conditions of Theorem 8 hold if both players use the ER rule. The next section investigates this case in more detail.

5. Learning in Games: the ER rule on both sides

This section considers learning in two-person constant-sum games when both players learn according to the ER rule. It will be shown that in the long-run each earns their minmax payoff. In addition if there is a unique pure-strategy equilibrium play will converge to this. The same is true in 2×2 games with a unique mixed-strategy equilibrium. The analysis uses similar techniques to the previous sections. It is shown that the behaviour of the system is related to a system of equation similar to the adjusted replicator dynamic introduced by Maynard Smith (1982). This may be of some independent interest as in 2×2 games it it well known that a mixed-strategy equilibrium is asymptotically stable in the Maynard-Smith dynamic but not in the ordinary replicator dynamic, where play cycles around it.[3] Several studies have derived the replicator dynamic from a learning model (for example Börgers and Sarin (1997), Gale et al. (1995)) but justification for the Maynard Smith dynamic is rather more elusive. Bjornerstedt and Weibull (1996) and Hofbauer and Schlag (2000) show that dynamics similar to the Maynard Smith ones may arise from imitation dynamics in large population model but there seem few other results in this direction.

5.1 Generalities and Values

To fix notation assume that player 1 has m actions 1 to m, player 2 l actions 1 to l. If 1 plays action i and player 2 action j then player i receives payoff α_{ij} and player 2 payoff β_{ij}, where $\alpha_{ij} > 0$ and $\beta_{ij} > 0$ for all i, j. It is assumed that the game is constant sum, so $\alpha_{ij} + \beta_{ij} = K$ for some constant K, for all i, j. Let $\mathcal{A} = (\alpha_{ij})$ be the $m \times l$ payoff matrix for player 1 and $\mathcal{B} = (\beta_{ij})$ the $m \times l$ matrix for player 2. For convenience a strategy vector $p = p_1, \ldots, p_m$ for player 1 will be regarded a row vector and a strategy vector for player 2 $q = q_1, \ldots, q_l$ will be regarded a column vector, so the expected payoffs to the players from these strategies are $p\mathcal{A}q$ and $p\mathcal{B}q$ respectively.

The two players play the game repeatedly and each learns according to the learning model of Section 2. Let $A_i(n)$ denote the reinforcement to strategy i for player 1 at stage n and $B_j(n)$ that for strategy j for player 2. As in section 4 the time-averaged reinforcements will be considered and are denoted by $a_i(n)$ and $b_j(n)$ respectively. The total average reinforcements for the first n stages are denoted by $a(n)$ and $b(n)$ respectively.

There is no reason to suppose that players' initial reinforcements are consistent with the structure of the game. In particular, there is no reason to suppose that they add up to K. Let $r = a + b$ denote the sum of the total average reinforcements.

[3] See for example Maynard Smith (1982) Appendix J or Hofbauer and Sigmund (1988) ch. 27.

r converges to K asymptotically, since reinforcements after time 0 simply cumulated payoffs, but r will be allowed to differ from K in the analysis.

Since payoffs are positive and bounded above, the time-averaged reinforcements, $a_i(n)$ and $b_i(n)$, are bounded below by zero and bounded above, say by M. In addition, since payoffs are strictly positive and bounded away from zero, the time-averaged total reinforcements, a and b, are bounded away from zero, say by $\kappa > 0$. The state of the system can be therefore assumed to lie in the set

$$\Delta = \{(a_1, \ldots, a_m, b_1, \ldots, b_m) | 0 \leq a_i \leq M, 0 \leq b_j \leq M \quad \text{for all} \quad i, j, \quad a \geq \kappa, b \geq \kappa\}$$

z will be used for a generic element of this set where convenient.

As discussed in Section 4 in the derivation of equation (6), the long-run behaviour of the system is related, at least heuristically, to that of the system

$$\frac{da_i}{dt} = -a_i(t) + p_i \pi_i^I(t) \qquad i = 1, \ldots, m \tag{10a}$$

$$\frac{db_j}{dt} = -b_j(t) + q_j \pi_j^{II}(t) \qquad j = 1, \ldots, l \tag{10b}$$

where $p_i = a_i/a$ and $q_i = b_i/b$ and $\pi_i^I(t) = (\mathcal{A}q)_i$ and $\pi_j^{II}(t) = (p\mathcal{B})_j$ are the expected payoffs to actions i and j.

Since $p_i = a_i/a$ and $q_j = b_j/b$ an equivalent set of equations is

$$\frac{dp_i}{dt} = \frac{p_i(t)\left(\pi_i^I(t) - \Pi^I(t)\right)}{a(t)} \qquad i = 1, \ldots, m \tag{11a}$$

$$\frac{da}{dt} = -a(t) + \Pi^I(t) \tag{11b}$$

$$\frac{dq_j}{dt} = \frac{q_j(t)\left(\pi_j^{II}(t) - \Pi^{II}(t)\right)}{b(t)} \qquad j = 1, \ldots, l \tag{11c}$$

$$\frac{db}{dt} = -b(t) + \Pi^{II}(t) \tag{11d}$$

where $\Pi^I(t) = p\mathcal{A}q$ and $\Pi^{II}(t) = p\mathcal{B}q$ are the expected payoffs of the two players at time t.

In this form, these equations can be related to more familiar dynamics. The ordinary replicator dynamic has the same form as above, except that (11b) and (11d) are replaced by the condition that a and b are constants. In the Maynard Smith dynamic, (11b) and (11d) are replaced by the condition that $a = \Pi^I$ and $b = \Pi^{II}$, that is reinforcements always equal current payoffs. The dynamic considered here might be considered a version of this where instead reinforcements adjust slowly towards current payoffs. As will be seen subsequently, it shares some of the convergence properties of the Maynard Smith dynamic.

For the moment, attention will be focussed on payoffs. Recall (see for example Karlin (1959)) that in a zero-sum game players earn the same expected payoff in any

equilibrium. Denote the values of the game to the players by $v_I = s_I K$ and $v_{II} = s_{II} K$ (that is s_I and s_{II} are the shares of the total payoff of the game the players earn).

Recall also (see Karlin (1959) Ch. 3) that in a zero-sum game one can find an equilibrium pair (p^*, q^*) such that if any strategy appears in the support of any equilibrium strategy it appears with strictly positive weight in this equilibrium pair. Any equilibrium strategy pair would in fact do, but for convenience fix on this one.

Consider the function

$$V = s_I \left(\sum_i p_i^* \ln a_i(t) \right) + s_{II} \left(\sum_j q_j^* \ln b_j(t) \right) - \ln r(t) \tag{12}$$

where the convention $0 \ln 0 = 0$ is adopted. This is a variant on the standard Lyapounov function for the replicator dynamic and indeed of that considered in Section 4. The term $\ln r(t)$ simply caters for the fact that r may not always equal K.

V is well-defined on the set

$$\Delta^* = \left\{ (a,b) | (a,b) \in \Delta, \quad p_i^* > 0 \Rightarrow a_i > 0, \quad q_j^* > 0 \Rightarrow b_j > 0, \quad \text{for all} \quad i,j \right\}$$

Since initial reinforcements are all assumed positive, the initial conditions of the system can be assumed to lie in $\text{int}\,(\Delta) \subseteq \Delta^*$, so V is certainly well-defined initially. In fact (again see Karlin (1959) Ch. 3) for this choice of (p^*, q^*) any pure strategy which does not lie in the support of the equilibrium pair can be deleted without affecting the equilibrium set, so one can allow any initial conditions in Δ^*.

Let $\partial\Delta^*$ be the boundary of Δ^* considered as a subset of Δ:

$$\partial\Delta = \left\{ (a,b) | (a,b) \in \Delta, \exists i,\ a_i = 0 \text{ and } p_i^* > 0 \quad \text{or} \quad \exists j,\ b_j = 0 \text{ and } q_j^* > 0 \right\}$$

It is straightforward to check that (see Appendix)

Lemma 3 *V is bounded above on Δ^* and $V \to -\infty$ as $z \to \partial\Delta^*$. Moreover $\dot{V} \geq 0$ in Δ^*.*

In other words, V is a weak Lyapounov function. By a standard result on differential equations (see for example Hale (1980) p. 317) it follows that:

Lemma 4 *For any initial condition in Δ^* the solution is bounded away from $\partial\Delta^*$ and converges to the largest invariant set in $\{z | \dot{V}(z) = 0\}$.*

Note that $\dot{V}$ can also be written as $\langle \nabla V, h \rangle$, where ∇V is the gradient of V, h the vector-field corresponding to the right-hand side of (10a) and (10b) and $\langle , \rangle$ denotes the inner product.

Furthermore,

Lemma 5 *(a) $\{z | \dot{V}(z) = 0\}$ is contained in the set $a = s_I r$, $p^* \mathcal{A} q = v_I$ and $p \mathcal{B} q^* = v_{II}$. (b) Any invariant set in $\{z | \dot{V}(z) = 0\}$ has $r = K$ and $p \mathcal{A} q = v_I$.*

The proof can be found in the Appendix. It follows that

Theorem 9 *Any solution of the system governed by (10a) and (10b) with initial conditions in Δ^* remains bounded away from $\partial\Delta^*$ and converges to the set of reinforcements where $a = v_I$ and $b = v_{II}$.*

In other words, in the long-run both players earn the value of the game.

These properties can be transferred to the stochastic model. As Section 4, it is shown by a Taylor expansion that V is a sub-martingale plus some perturbations. As there, the logarithmic form of V requires some care with the second-order terms. It is shown in the Appendix that:

Lemma 6 *In the ER model, V converges to a finite random variable almost surely. z converges to the largest set in $\langle \nabla V, h \rangle$ invariant under the flow generated by h.*

It follows that

Theorem 10 *If both players learn according to the ER model, then $\lim_{n\to\infty} a(n) = v_I$ and $\lim_{n\to\infty} b(n) = v_{II}$.*

In other words, the long-run average payoffs of the players converge to their values.

Of course this says nothing about converges of strategies. This will be investigated in the next sub-section. One observation before proceeding, is that if there are multiple equilibria then if play converges to an equilibrium at all, it must converge to one with the same support as p^*, q^*. For the fact that V converges to a finite random variable implies that the system cannot converge to $\partial\Delta^*$.

5.2 Strategies

One can obtain results on the convergence of strategies in games with unique equilibria by characterising the invariant sets in $\langle \nabla V, h \rangle = 0$ more fully. From Lemma 5, it follows that any point in this set corresponds to strategies which earn exactly the value of the game to each player against the opponent's equilibrium strategy. Now from Karlin (1959) Ch. 3 Theorem 3.1.1, if there is a unique pure-strategy equilibrium, these strategies are the only ones that have this property. Hence

Theorem 11 *If the game has a unique pure-strategy equilibrium, then if both players learn according to the ER scheme, the probability that they play these strategies converges to one. The same is true of the empirical frequencies.*

The last statement follows from a strong law of large numbers.

In the case of 2×2 games, one can also prove convergence if there is a unique mixed strategy equilibrium. The following is shown in the Appendix:

Lemma 7 *If the game is 2×2 and has a unique mixed strategy equilibrium, then the unique invariant set in in $\langle \nabla V, h \rangle$ corresponds to both players playing their equilibrium strategy.*

It follows that in 2×2 games a mixed strategy equilibrium is asymptotically stable under the dynamic (10) or equivalently (11), just as it is under the Maynard Smith dynamic, though not the ordinary replicator. The behaviour of both this dynamic and

the Maynard Smith dynamic is unknown in higher-dimensional games, though it is plausible that they are both convergent.[4]

It is perhaps worth noting here a point that has been surpressed in some of the discussion above. The fact that in 2×2 games all trajectories beginning in the interior of Δ converge to the equilibrium point under (11), does not of itself imply that the same is true of the ER algorithm. The boundary of Δ is an absorbing set and although there are results ruling out convergence of stochastic approximation to unstable sets of the corresponding differential equation[5], these are not applicable here as they assume that the variance of the noise of the system (at least in unstable directions) is bounded away from zero in the neighbourhood of the set. Here as the reinforcement of a strategy goes to zero, the probability that it is played and so the variance of its component go to zero. One therefore needs a further argument to rule convergence to the boundary. This is provided by Lemma 6, which shows that V converges to a finite random variable.

In any case, here one has

Theorem 12 *In a 2×2 game with a unique mixed-strategy equilibrium, the probability that each player plays each strategy converges to their equilibrium probabilities almost surely, if players learn according to the ER scheme. The empirical frequency with which each strategy pair (i, j) is played converges to $p_i^* q_j^*$ almost surely.*

The last statement is proved in the Appendix.

This result should, however, be interpreted with caution. If one examines equations (11) in the case of Figure 1 (Matching Pennies), it is easy to check that the eigenvalues of h at the equilibrium point $p_1 = 0.5$, $q_1 = 0.5$, $a = 1.5$, $b = 1.5$ are -1,-1,$+i$ and $-i$. More generally, in 2×2 games with a unique fully-mixed equilibrium the eigenvalues are -1, -1 and a pair of conjugate purely imaginary ones. The real eigenvalues intuitively come from the evolution of a and b in (11b) and (11d). The imaginary eigenvalues come from the evolution of p_1 and q_1 in (11a) and (11c) and correspond to the fact that in the ordinary replicator dynamic, where a and b are fixed constants, the system cycles about the equilibrium. The equilibrium is in fact a centre so small perturbations such as the Maynard Smith dynamic, or the one considered here, make the equilibrium asymptotically stable. Nevertheless, the rate of convergence is slow.

This is reflected in the stochastic algorithm. Results which yield $\sqrt{n}$ convergence for a stochastic algorithm of the current form (see for example Duflo (1997) ch. 2) require the largest real part of the eigenvalues, τ, of the vector field of the corresponding differential equation to be less than $-1/2$. If $-1/2 < \tau < 0$, then one can show that the distance between the solution and the equilibrium goes to zero at rate $n^{-\tau}$. For the case $\tau = 0$, there seem few results (Chen (1998) has some, but these do not cover the current case), but it seems plausible that the rate is extremely slow.

[4] I am grateful to Josef Hofbauer for some helpful correspondence on the Maynard Smith dynamic.

[5] For example Pemantle (1990), Brandiere and Duflo (1996) in the case of unstable equilibria, Benaïm (1999) in the case of more general sets.

Another way of understanding the difficulty is that as the tendency of the deterministic system to converge is weak, it is easy for random shocks to disturb the system away from equilibrium in the initial stages of the algorithm. In order to eliminate these effects one needs a very small step size, $1/n$, for the system (or equivalently start with very large initial reinforcements $A(0)$ and $B(0)$ so that the random perturbations have little influence). For large n, however the system moves very slowly (see (6)). Since $\sum_n 1/n$ diverges the system cannot actually get stuck but since $\sum_n 1/n$ only goes to infinity at rate $\ln n$ movement can become very slow. Unless, therefore, the system converges strongly initially, the algorithm may converge very slowly indeed, which is a well-known difficulty with stochastic approximation algorithms.

From the players' point of view it is not clear that this is a draw-back. They care about payoffs. It is possible to be a long way from equilibrium in probability terms yet be close in terms of average payoffs and also best attainable current payoffs. For example $p_1 = 0.6$, $q_1 = 0.4$ offers yields an expected payoff of 1.48 to player 1 and 1.52 to player 2. If player 1 were to take a best response of $p_1 = 0$, his expected payoff would only rise to 1.6. Given the slow rates of converence noted in Section 3, it is to be expected that convergence of strategies, if not of payoffs, will be slow here.

This behaviour is confirmed in simulations. Average payoffs converge to the value of the a game at a very respectable rate but convergence of strategies is slow. For example the game in Figure 1 was simulated 100 times for 10,000 periods with initial reinforcements $A_1(0) = 3$, $A_2(0) = 7$, $B_1(0) = 6$ and $B_2(0) = 4$. The mean value of player's 1 payoff was 1.52 with coefficient of variation 0.05. The mean value of the probability with which player 1 plays strategy 1 after 10,000 periods was 0.65 with coefficient of variation 0.23. For player 2 the corresponding figure for strategy 1 are 0.62 and 0.29. In other words payoffs converge reasonably quickly, strategies do not.

Figure 3 shows the behaviour of p_1 and q_1 in an extremely long run of the game of Figure 1 — 10^7 periods. Initial reinforcements were $A_1(0) = 200$, $A_2(0) = 800$ and $B_1(0) = 800$, $B_2(0) = 200$ (far away from equilibrium so some reason to change strategy, but large to eliminate random effects). Points are plotted every 10,000 perods and interpolated. The algorithm moves so slowly that it has not had time even to complete a full cycle (the first point plotted is after 10,000 stages so short-run behaviour is surpressed).

On the other hand, consider the game in Figure 2. In this game there are rather greater penalties for being away from equilibrium, so one might expect convergence to be quicker. This appears to be the case. In Figure 4, the same long run, 10^7 iterations, is done as in Figure 3 (initial reinforcements $A_1(0) = 7$, $A_2(0) = 3$, $B_1(0) = 3$ and $B_2(0) = 7$ — here large values are not needed to eliminate random effects). Now strategies appear to spiralling inwards, albeit at a slow rate.

In practice, therefore, it appears that convergence of strategies is likely to be very slow. On the other hand, players may not care much about this. Convergence to the value of the game is much quicker.[6]

[6]The simulations presented are for symmetric games. A similar picture emerges for asymmetric

Against other learning rules, for example against an opponent employing fictitious play or myopic best response, the play of the player using the ER rule converges rapidly to the mixed-strategy equilbrium in simulations. These dynamics react strongly to even small deviations from the mixed strategy equilibrium and so force play back there quickly.

Posch (1997) studies a related model of Arthur (1993) in the context of 2×2 games. This is similar to Erev and Roth's model, except that reinforcements are re-scaled in every period so that each player's total reinforcement grows at rate C, for some C.[7] As a result he obtains the ordinary replicator dynamic as governing the stochastic evolution of the system rather than that above — in effect a and b are both always equal to C. In his model, strategies cycle around the equilibrium levels. Here convergence obtains, though this is distinctly asymptotic result and in practice the difference in conclusion may be small. He does not consider the convergence of payoffs.

6. Discussion

This section comments on the previous results and considers variations of the basic model.

(a) Forgetting

In the ER model as time goes on the rate of change of the state becomes less and less, since the step-size $(1/n)$ tends to zero. This is necessary to ensure that random effects are eliminated, but as seen in the previous section can lead to slow convergence. In practice it might be preferable to have step sizes which are small but do not tend to zero. Another consideration is that the ER scheme puts equal weight on all observations, while if the environment may be changing it may be preferable to put more weight on recent ones. Both these issues can be dealt with by introducing 'forgetting' into the model.

Suppose that (1) is replaced by

$$A_i(n+1) = \begin{cases} \phi A_i(n) + \pi_i(n+1), & \text{if action } i \text{ is chosen;} \\ \phi A_i(n), & \text{otherwise} \end{cases} \qquad (1)'$$

where ϕ is a constant less than 1 in absolute value. $1 - \phi$ measures the rate at which past experience is discounted or forgotten.

Instead of the average reinforcements per unit time, it is natural to consider the normalised reinforcement $a_i(n) = (1 - \phi)A_i(n)$. $A_i(n)$ essentially cumulates past payoffs discounted at the rate ϕ and the factor $(1 - \phi)$ normalises them so they are are comparable for different values ϕ.

Exactly the same equations are obtained as in Section 4, with $\psi = 1 - \phi$ replacing $1/(n+1)$. For example (5)' becomes

$$a_i(n+1) = a_i(n) + \psi \left(p_i \pi_i(n+1) + u_{n+1} \right) \qquad (5)''$$

games.

[7] Ianni (2000) and, for some results, Hopkins (2000) also use a similar normalisation.

For small ψ, that is ϕ close to 1, one would again expect (6) to describe the evolution of a_i well.

Assume therefore that both players use the ER rule with the same forgetting rate ψ.[8] Let $t_n = n\psi$ for $n = 0, 1, 2, \ldots$ and fix a finite time $T > 0$. Let z_n be the vector of normalised reinforcements $a_1, \ldots, a_m, b_1, \ldots, b_l$ at stage n generated by the ER scheme with forgetting from some initial position. Let $\Theta(t, z_0)$ denote the position of the system governed by (10) at time t with the same initial position. $\| \ldots \|$ denotes the Euclidean norm.

The following is an immediate consequence of standard results, for example Benveniste et al. (1990), Theorem 1 p. 43,

Theorem 13 *For any $\epsilon > 0$ and $T > 0$, there is constant $C(\psi)$ with $C \to 0$ as $\psi \to 0$ such that for any z_0, $Pr\left(\sup_{\{n:t_n \leq T\}} \|z_n - \Theta(t_n, z_0)\| > \epsilon\right) < C(\psi)$.*

In other words, by making the step-size small enough one can the paths of the stochastic and deterministic systems can be made arbitrarily (uniformly) close with arbitrarily high probability over a finite horizon.

Note the restriction that T be finite. Even if the deterministic system converges to a unique equilibrium, for any fixed positive ψ there is always a small probability that shocks will knock the stochastic system away from the equilibrium, at least temporarily. If one wishes to eliminate fluctuations, one must let the step size go to zero.

In fact here, the situation is even starker here: the deterministic and stochastic systems may have completely different asymptotic behaviour. For example in Section 5.2 it was seen that an interior fully-mixed equilibrium was globally asymptotically stable in 2×2 constant-sum games. For any fixed ψ, however, the stochastic process is a Markov chain with absorbing states corresponding to pure strategies and it easy to check that the model is distance-diminishing in the sense of Norman (1972), and applying Theorems 4.3 and 6.1 of Chapter 3 of that book, one obtains

Theorem 14 *For any $\psi > 0$, in a 2×2 constant-sum game, play converges to a pure-strategy combination.*

This result is very similar to that obtained by Börgers and Sarin (1997) in the context of the Bush and Mosteller learning model. They show that finite-horizon behaviour be described by the replicator dynamic, but asymptotically their stochastic model converges to the boundary. The contrast is perhaps even starker here as the equilibrium is actually globally stable here, while in the replicator dynamic play cycles around it. The contrast with the decreasing step-size model is at first sight puzzling, but is perhaps best understood by noting that for any fixed ψ, no matter how much time has passed there is a constant probability that there will a large chance fluctuation

[8] Allowing different rates would only introduce different scale factors in the corresponding differential equations.

away from the deterministic path which results in the process becoming stuck near the boundary, so eventually this happens with probability 1. In the decreasing-step case, the deterministic model becomes a better and better approximation as time goes on.

On the other hand, although the constant-step model converges to the boundary it will take a long while to do so for small ψ if it starts near the equilibrium point. More precisely one can show the following. Let $B_{\nu(\mu)}$ denote a neighbourhood of the equilibrium point such that any trajectory of the differential equation which starts in it remains within in a distance μ of the equilibrium point permanently. Let G be any open set containing the equilibrium point. Let $\tau^{\psi} = \inf\{t_n : z_n \notin G\}$, where t_n is defined before Theorem 13, be the first time the stochastic process escapes from G. Let $E_x\tau^{\psi}$ be the expected escape time with starting point x. One can then apply a Theorem of Kushner and Yin (1997) (see Appendix) and show

Theorem 15 *In a 2×2 constant-sum game with unique equilibrum point, for small enough μ there is constant $c > 0$ such that $E_x\tau^{\psi} \geq \exp(c/\psi)$ for all $x \in B_{\nu(\mu)}$.*

In any case, Theorem 14 should not be taken too literally. Simulations for the game in Figure 1 ('Matching Pennies') suggest that for moderate values of ψ, the process indeed tends to get stuck near the boundaries. On the other hand for small ψ the differential equation provides a good approximation. Figure 5 shows the path of the algorithm with $\psi = 0.0005$ and 100,000 periods and initial values $A_1(0) = 0.1$, $A_2(0) = 1.8$, $B_1(0) = 1.1$ and $B_2(0) = 0.5$. The points are plotted at intervals of 1,000. The path is spiralling inwards as the differential equation suggests it should. The fact that ψ needs to be chosen small reflects the difficulties noted in Section 5: for the deterministic process to be a good approximation the step-size must be small, but in the decreasing-step model, this implies system moves very slowly.

This suggests that if one wishes for a model which predicts play of mixed equilibria in reasonable, albeit large, time horizons the model with forgetting is a more promising candidate in practice, despite the theoretical properties of the two models. The model of Section 5 is of course satisfactory if one is interested in payoffs.

(b) Invariance Properties and Reference Points

In the case of rational agents, multiplying all their payoffs in a game by a constant or adding a constant to the payoffs when opponent takes action j, regardless of their own action, will not change their behaviour. In particular, neither change affects the set of equilibrium points. In the ER scheme agents are not fully rational and their behaviour is not affected by the first change but may be by the second.

From (2), multiplying all a player's payoffs by a constant leaves choice probabilities unchanged if the initial reinforcements are also re-scaled, so learning behaviour is the same. If the latter are not re-scaled, then this is equivalent to a new starting set of reinforcements. It follows that the results in Section 5 for constant-sum games also hold for all games for which one can find (positive) λ and μ for which $\lambda\alpha_{ij} + \mu\beta_{ij} = K$

for all i,j. For players' learning is the same as in the constant-sum game where they have payoffs $\lambda\alpha_{ij}$ and $\mu\beta_{ij}$.

Adding a constant to all payoffs, so for example player i's payoffs become $\alpha_{ij}+M$, will change choice probabilities. The game remains constant-sum, however, and so the results of Section 5 remain valid. This can be given an interpretation in terms of reference points. Erev and Roth (1998), for example, suggest that the reinforcement should be the extent to which payoff exceed some reference point. So if y_{ref} is the reference payoff, the reinforcement to i if i and j are played is $\alpha_{ij}-y_{ref}$. The rationale for this is that if all payoffs are high the impact of a big payoff on choices may be less than if all the rest are low. Now this is exactly equivalent to subtracting y from all payoffs, so (provided these all remain positive) the results of Section 5, which correspond to reference payoff $y_{ref}=0$, remain valid with non-zero reference point. Erev and Roth also suggest that the reference point may shift in response to the history of payoffs. This is not covered by the results but could be analysed by similar techniques.

Adding a constant to player 1's payoffs if 2 plays a certain action, so that they become $\alpha_{ij}+\delta_k$ if $j=k$, α_{ij} otherwise is also strategically irrelevant, but the learning algorithm is affected, so it is not clear whether the results cover games which are equivalent, in this sense, to constant-sum ones.

(c) Modifications of the Choice Function

The model in the previous sections assumes that the choice probabilities are determined by the reinforcements via (2). This sections considers some possible modifications. It is shown, that though special, its form is well-suited for learning to play mixed strategies. For ease of exposition, attention is restricted to the case of two actions.

One natural variation would be to consider a more general power form, that is the probability of choosing action 1 is

$$\frac{A_1^\epsilon(n)}{A_1^\epsilon(n)+A_2^\epsilon(n)} \qquad (2)'$$

Erev and Roth (1998), for example, consider this. This could easily be analysed by the techniques of Section 3, since this can simply be written as a function of the average reinforcements per unit time:

$$\frac{a_1^\epsilon(n)}{a_1^\epsilon(n)+a_2^\epsilon(n)}$$

Here only a heuristic treatment will be given. Now by the argument in Section 3.1 one would expect the system to converge to a fixed point of this function, if it converges. Suppose that the payoffs are of the two actions are equal to a constant α and that the player converges to playing action 1 with probability and empirical frequency x. Then we must have $a_1=\alpha x$ and $a_2=\alpha(1-x)$, so

$$x=\frac{x^\epsilon}{x^\epsilon+(1-x)^\epsilon}$$

Now for $\epsilon\neq 1$, the only fixed points of this system $x=0$, $x=1$ and $x=1/2$. In a mixed-strategy equilibrium, the expected payoff to playing each action is constant, so

this suggests that the system could only converge to a mixed-strategy equilibrium where each strategy is played with probability 1/2 or a pure strategy profile. By contrast when $\epsilon = 1$, every point in $[0, 1]$ is a fixed point. In a single-player context this means it is indeterminate which point the system converges to (see the discussion of the Polya urn in Section 3.1), but in a two-player context this freedom makes it possible to converge to an arbitrary mixed-strategy point.

Of course in practice if the equilibrium is not too asymmetric and ϵ is close to 1, this may not be too noticeable in practice in small samples. Indeed if the equilibrium is symmetric convergence may be aided as when payoffs are equal the system is pushed towards the equilibrium point, unlike the linear case.

One could also imagine other functional forms, for example using a Logit formulation where each reinforcement is replaced by an exponential. Camerer and Ho (1999), for example, considers this. The techniques of Section 5 would not then be applicable directly as the choices are not functions of the average reinforcements. If one assumed players used these, rather than the totals, then the methods could be applied.

Another variation would be to reinforce actions which are not played as well as those that are. Roth and Erev (1995) and Erev and Roth (1998) consider this. This might be thought likely to speed up convergence since, although the results of Section 5 show that this will take place eventually, this may be slow as the linear form means the system can remain stuck close to the boundary for a long while if some actions have been rarely played.

Consider therefore adding a constant ϵ to the reinforcement of the action that is not played. Again consider the case when the payoffs of each action are constant at α. If in the long-run action 1 is played a fraction x of the time, its average reinforcement will be $\alpha x + \epsilon(1 - x)$. (2) then becomes

$$\frac{\alpha x + \epsilon(1 - x)}{\alpha + \epsilon}$$

Now if the system converges it should be to a fixed point of this. Yet it is easy to check that the only fixed point of this system is $x = 1/2$ for any value of $\epsilon > 0$, no matter how small. In the literature on urns, this is referred to as the so-called Friedman urn model. It corresponds to always adding some balls of the opposite colour to that drawn and it is well-known that it has radically different properties to the Polya urn (see Freedman (1965) for example for an extended discussion).

Again, this suggests it will be hard to learn asymmetric mixed strategy equilibria. In small samples this may of course not be too much of an issue if the equilibrium is not too asymmetric and the advantage of quicker escape from the boundary may outweigh the long-run effects. Hopkins (2000) contains some results on local stability of the rest points of this kind of model.

Another variation would be to simply change the probability function near the boundary — for example by preventing the probability of any action falling to zero. If the equilibrium is not on the boundary, this would leave the fixed points of the

system unchanged. Provided that this change is done in such a way that the system cannot become stuck near the boundary (for example by inducing a cycle), then the global stability shown in Section 5 would continue to hold and so there would still be convergence to equilibrium.

7. Conclusion

This paper has studied the properties of a simple model of a reinforcement learning. It has shown that it has quite reasonable properties. It is certainly not the only plausible model of learning but study in more general games seems worthwhile.

References

Arthur, W. B. (1993) 'On Designing Economic Agents that Behave Like Human Agents', *Journal of Evolutionary Economics*, **3**, 1–22

Auer, P., Cesa-Bianchi, N., Freund Y. and R. Schapire (1998) 'Gambling in a Rigged Casion: The Adversarial Multi-Armed Bandit Problem', mimeo, AT&T laboratories

Benaïm, M. (1999) 'Dynamics of Stochastic Approximation Algorithms', in *Séminaire de Probabilités, XXXIII*, 1–68, *Lecture Notes in Mathematics* **1709**, Springer Verlag, Berlin

Benaïm, M. and M. Hirsch (1999)'Mixed Equilibria and Dynamical Systems Arising from Fictitious Play in Perturbed Games', *Games and Economic Behavior*', **29**, 36–72

Benveniste, A., Metivier, M. and P. Prioret (1990) *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin and New York

Bjornerstedt, J. and J. Weibull (1996) 'Nash Equilibrium and Evolution by Imitation', in (ed.) Arrow, K. et al. *The Rational Foundations of Economic Behaviour*, Macmillan, London, 155–71

Börgers, T. and R. Sarin (1997) 'Learning through Reinforcement and Replicator Dynamics', *Journal of Economic Theory*, **77**, 1–14

Brandiere, O. (1998) 'Some Pathological Traps for Stochastic Approximation', *SIAM Journal on Control and Optimization*, **36**, 1293–1314

Brandiere, O. and M. Duflo (1996) 'Les Algorithmes Stochastiques Contournent-ils les Pieges', *Ann. Inst. Henri Poincaré*', **32**, 395–427

Camerer, C. and T.-H. Ho (1999) 'Experience-weighted Attraction Learning in Normal Form Games', *Econometrica*, **67**, 827–874

Chen, H.-F. (1998) 'Convergence Rate of Stochastic Algortithms in Degenerate Cases', *SIAM Journal on Control and Optimization*, **36**, 100–114

Duflo, M. (1996) *Algorithmes Stochastiques*, Springer Verlag, Berlin

Duflo, M. (1997) *Random Iterative Models*, Springer Verlag, Berlin

Dupuis, P. and H. Kushner (1989) 'Stochastic Approximations and Large Deviations: Upper Bounds and w. p. 1 Convergence', *SIAM Journal on Control and Optimization*, **27**, 1108–35

Erev, I. and A. Roth (1998) 'Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria', *American Economic Review*, **88**, 848–881

Freedman, D. (1965) 'Bernard Friedman's Urn', *Annals of Mathematical Statistics*, **36**, 956–70

Fudenberg, D. and D. Levine (1998) *Learning and the Theory of Games*, MIT Press, Cambridge, MA

Gale, J., Binmore, K. and L. Samuelson (1995) 'Learning to be Imperfect: The Ulitmatum Game', *Games and Economic Behavior*, **8**, 56–90

Hale, J. (1980) *Ordinary Differential Equations*, Robert E. Krieger Publishing Co., Huntingdon, New York

Hall, P. and C. Heyde (1980) *Martingale Limit Theory and its Applications*, Academic Press, New York

Harley, C. (1981) 'Learning the Evolutionarily Stable Strategy', *Journal of Theoretical Biology*, **89**, 611–33

Hart, S. and A. Mas-Collel (2000) 'A Simple Adaptive Procedure Leading to Correlated Equilibrium', *Econometrica* **68**, 1127–1150

Hart, S. and A. Mas-Collel (2001a) 'A General Class of Adaptive Strategies', *Journal of Economic Theory*, **98**, 26–54

Hart, S. and A. Mas-Collel (2001b) 'A Reinforcement Procedure Leading to Correlated Equilibrium', mimeo, Hebrew University

Hill, B., Lane, D. and W. Sudderth (1980) 'A Strong Law for Some Generalized Urn Processes', *Annals of Probability*, **8**, 214–26

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

Hofbauer, J. and K. Schlag (2000) 'Sophisticated Imitation in Cyclic Games', *Journal of Evolutionary Economics*, **10**, 523–43

Hopkins, E. (2000) 'Two Competing Models of How People Learn in Games', mimeo, University of Edinburgh

Ianni, A. (2001) 'Reinforcement Learning and the Power Law of Practice', mimeo, University of Southampton

Karlin, S. (1959) *Mathematical Methods and Theory in Games, Programming and Economics*, Volume I, Addison-Wesley, Reading, MA

Kuan, C.-M. and H. White (1994) 'Adaptive Learning with Nonlinear Dynamics Driven by Dependent Processes', *Econometrica*, **62**, 1087–1114

Kushner, H. and G. Yin (1997) *Stochastic Approximation Algorithms and Applications*, Springer Verlag, New York

Laslier, J.-F., Topol, R. and B. Walliser (2001) 'A Behaviorial Learning Process in Games', *Games and Economic Behavior*, **37**, 340–66

Maynard Smith, J. (1982) *Evolution and the Theory of Games*, Cambridge University Press, Cambridge

Pemantle, R. (1990) 'Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations', *The Annals of Probability*, **18**, 698–712

Pemantle, R. and S. Volkov (1999) 'Vertex-reinforced Random Walk on **Z** has Finite Range', *Annals of Probability*, **27**, 1368–88

Posch, M. (1997) 'Cycling in a Stochastic Learning Algorithm for Normal Form Games', *Journal of Evolutionary Economics*, **7**, 193–207

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**, 164–212

Rustichini, A. (1999) 'Optimal Properties of Stimulus-Response Learning Models', *Games and Economic Behavior*, **29**, 244–73

Appendix

Proof of Theorem 3

By a Taylor expansion, there is a constant c_1 such that for any value of $\pi_i(n+1)$ and sufficiently large $A_1(n)$

$$\frac{1}{A_1(n)+\pi_1(n+1)} \leq \frac{1}{A_1(n)} - \frac{\pi_1(n+1)}{A_1^2(n)} + c_1\frac{\pi_1(n+1)^2}{A_1^3(n)} \tag{A1}$$

By another Taylor expansion, since by assumption $\pi_2(n+1)$ is bounded, there is a constant c_2 such that

$$(A_2(n)+\pi_2(n+1))^\epsilon \leq A_2(n)^\epsilon + \epsilon A_2(n)^{\epsilon-1}\pi_2(n+1) + c_2\epsilon A_2(n)^{\epsilon-2}\pi_2(n+1)^2 \tag{A2}$$

The expected increment $A_2(n+1)^\epsilon/A_1(n+1) - A_2(n)^\epsilon/A_1(n)$, conditional on $\mathcal{F}_n$ is

$$\begin{aligned}&\frac{A_1(n)}{A_1(n)+A_2(n)}E\left(\frac{A_2(n)^\epsilon}{A_1(n)+\pi_1(n+1)} - \frac{A_2(n)^\epsilon}{A_1(n)}\right)\\ &\quad + \frac{A_2(n)}{A_1(n)+A_2(n)}E\left(\frac{(A_2(n)+\pi_2(n+1))^\epsilon}{A_1(n)} - \frac{A_2(n)^\epsilon}{A_1(n)}\right)\end{aligned} \tag{A3}$$

Using $(A1)$ and $(A2)$, this is at most

$$\begin{aligned}&\frac{A_1(n)}{A_1(n)+A_2(n)}\frac{A_2(n)^\epsilon}{A_1(n)^2}\left(-E\pi_1(n+1) + c_1\frac{E\pi_1(n+1)^2}{A_1(n)}\right)\\ &\quad + \frac{A_2(n)}{A_1(n)+A_2(n)}\frac{\epsilon A_2^{\epsilon-1}}{A_1(n)}\left(E\pi_2(n+1) + \frac{c_2E\pi_2(n+1)^2}{A_2(n)}\right)\end{aligned} \tag{A4}$$

Using the assumptions in the theorem, the boundedness above of $\pi_1(n+1)$ and $\pi_2(n+1)$, and the boundedness away from zero of $\pi_2(n+2)$ this is at most

$$\frac{A_2(n)^\epsilon}{(A_1(n)+A_2(n))\,A_1(n)}\left(K_1\,(\epsilon-\gamma) + \frac{K_2}{A_1(n)} + \frac{K_3}{A_2(n)}\right) \tag{A5}$$

for some constants $K_1 > 0$, K_2 and K_3. For sufficiently large $A_1(n)$ and $A_2(n)$ this is non-positive.

Choosing M large enough, therefore, $A_2(n)^\epsilon/A_1(n)$ is a positive super-martingale when $A_1(n) > M$ and $A_2(n) > M$. By Lemma 2, this is true for all but finitely many n. It follows that $\frac{A_2(n)^\epsilon}{A_1(n)}$ converges to a finite limit almost surely. Picking $1 < \epsilon < \gamma$ shows that $A_2(n)/A_1(n)$ converges to zero almost surely.

Proof of Corollary to Theorem 5

The proof is by induction. Suppose that strategy 1 (say) of player 1 is dominates strategy 2 if a certain dominated strategy of another player is deleted. By Theorem 4, the probability that the latter strategy is played almost surely converges to zero. Hence there is almost surely some N, such that for $n \geq N$,

$E(\pi_1(n)|\text{action 1 is chosen at stage } n) > \gamma E(\pi_2(n)|\text{action 2 is chosen at stage } n)$, for some constant γ. Hence, applying the argument of the proof of Theorem 3, $A_2(n)^\epsilon/A_1(n)$ is almost surely a postive supermartingale for all but finitely many n, hence convergent. It follows that the probability that action 1 is played converges to zero. Induction proves the result.

Proof of Theorem 6

Suppose that strategy i is dominated by the mixed strategy $\sigma_1, \ldots, \sigma_m$. A Taylor expansion, as in the proof of Theorem 3, shows that $A_i(n)^\epsilon/A_1(n)^{\sigma_1} \ldots A_m(n)^{\sigma_m}$ is a positive supermartingale for some $\epsilon > 1$, hence convergent. It follows that $A_i(n)/A_1(n)^{\sigma_1} \ldots A_m(n)^{\sigma_m}$ converges to zero almost surely. Now, since payoffs are bounded, $A_1(n)^{\sigma_1} \ldots A_m(n)^{\sigma_m}/n$ is bounded. It follows that $A_i(n)/n$, and therefore the probability that strategy i is played, converges to zero. Arguing as in Theorem 5 proves the theorem for iteratively dominated strategies.

Proof of Theorem 7

A well-known result of Robbins and Siegmund states (see for example Duflo (1997) Theorem 1.3.12)

Result RS *Let $\{\Omega, \mathcal{F}, \mathcal{P}\}$ be a probability space and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ a sequence of sub-σ-algebras of $\mathcal{F}$. For each $n = 1, 2, \ldots$ let z_n, β_n, ξ_n and ζ_n be non-negative $\mathcal{F}_n$-measurable random variables such that*

$$E(z_{n+1}|\mathcal{F}_n) \leq z_n(1+\beta_n) + \xi_n - \zeta_n$$

then $\lim_{n\to\infty} z_n$ exists and is finite and $\sum_n \zeta_n < \infty$ on the event $\{\sum_n \beta_n < \infty, \sum_n \xi_n < \infty\}$.

It is easy to check from the proof that one can relax the requirement on ξ_n to $\sum_n \xi_n$ converges, that is one need not require ξ_n to be non-negative. A re-arrangement clearly implies:

Result RS′ *Let z'_n, ζ'_n and ξ'_n be $\mathcal{F}_n$-measurable random variables, with z'_n bounded above by a non-random constant C and ζ'_n non-negative, such that*

$$E\left(z'_{n+1}|\mathcal{F}_n\right) \geq z'_n + \zeta'_n - \xi'_n$$

then $\lim_{n\to\infty} z'_n$ exists and is finite and $\sum_n \zeta'_n < \infty$ on the event $\{\sum_n \xi'_n$ converges$\}$.

Result RS′ will applied to

$$V_n = \sum_i \sigma_i \ln a_i(n)$$

with V_n in the role of z'_n (note that V_n is bounded above). To verify that the conditions are satisfied, consider the evolution of the components of V_n.

From (5)

$$\ln a_i(n+1) = \ln a_i(n) \left(1 + \frac{\tilde{\pi}_i(n+1) - a_i(n)}{(n+1)a_i(n)}\right) \qquad (A6)$$

Now for $|x| < \epsilon$ (say) $\ln(1+x) \geq x - cx^2$, where c is a constant. $\tilde{\pi}_i(n+1) - a_i(n)$ is bounded and so on the event $(n+1)a_i(n) > M$, some M,

$$\ln a_i(n+1) \geq \ln a_i(n) + \frac{\tilde{\pi}_i(n+1) - a_i(n)}{(n+1)a_i(n)} - c\frac{\left(\tilde{\pi}_i(n+1) - a_i(n)\right)^2}{(n+1)^2 a_i(n)^2} \qquad (A7)$$

Now by Lemma 2 $(n+1)a_i(n)$ tends to infinity, so this event has probability 1 for large enough n.

Now $\tilde{\pi}_i(n+1) = 0$ if action i is not played, so in this case

$$\frac{\left(\tilde{\pi}_i(n+1) - a_i(n)\right)^2}{(n+1)^2 a_i(n)^2} = \frac{1}{(n+1)^2} \qquad (A8)$$

If action i is played, then $\pi_i(n+1) - a_i(n)$ is bounded above, say by $\tilde{K}$, as payoffs and reinforcements are bounded. Also $a_i(n) \geq \left(k_1 N_i(n) + A_i(0)\right)/n$, where $k_1 > 0$ is the minimum payoff to any action (see Section 2) and $N_i(n)$ is the number of times that action i has been played up to stage n. Hence if action i is played at stage $n+1$,

$$\frac{\left(\pi_i(n+1) - a_i(n)\right)^2}{(n+1)^2 a_i(n)^2} \leq \tilde{K}^2 \frac{n^2}{(n+1)^2 \left(k_1 N_i(n) + A_i(0)\right)^2} \qquad (A9)$$

One therefore has

$$0 \leq \frac{\left(\pi_i(n+1) - a_i(n)\right)^2}{(n+1)^2 a_i(n)^2} \leq K_n \qquad (A9)$$

where

$$K_n = \begin{cases} \frac{1}{(n+1)^2} & \text{if } i \text{ is not played} \\ \frac{D}{\left(k_1 N_i(n) + A_i(0)\right)^2} & \text{if } i \text{ is played} \end{cases} \qquad (A10)$$

for some constant D. Now

$$\sum_{\{n:\ i \text{ is played}\}} \frac{1}{\left(N_i(n) + A_i(0)\right)^2} \qquad (A11)$$

converges since each time i is played $N_i(n)$ increases by 1 and so this at most some multiple of $\sum_n \frac{1}{n^2}$ (if one summed over all n this sum in (A11) need not converge since $N_i(n)$ does not change when i is not played but one counts here terms only when it changes).

It follows that $\sum_n K_n < \infty$ almost surely. Since $0 \leq K_n \leq F$, for some constant F, it follows from Hall and Heyde (1980) Corollary 2.3 on p. 32, that $\sum_n E(K_n|\mathcal{F}_{n-1})$ converges almost surely as well. Hence it follows from (A6) and (A9) that one has

$$E(\ln a_i(n+1)|\mathcal{F}_n) \geq \ln a_i(n) + E\left(\frac{\tilde{\pi}_i(n+1) - a_i(n)}{(n+1)a_i(n)}|\mathcal{F}_n\right) - H_n \qquad (A12)$$

where $H_n \geq 0$ and $\sum_n H_n < \infty$.

Applying this argument to each term in V_n and adding up one obtains, noting that $p_i(n+1) = a_i(n)/a(n)$,

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + E\left(\frac{\pi^\sigma(n+1) - a(n)}{(n+1)a(n)}\Big|\mathcal{F}_n\right) - \kappa_n \tag{A13}$$

where $\kappa_n \geq 0$ and $\sum_n \kappa_n < \infty$.

If $a(n) \leq v'$, $v' < v$,

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + E\left(\frac{\pi^\sigma(n+1) - v'}{(n+1)v'}\Big|\mathcal{F}_n\right) - \kappa_n \tag{A14}$$

By Assumption M,

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + \frac{v - v'}{(n+1)v'} - \xi_n \tag{A15}$$

where $\sum_n \xi_n$ converges.

Hence on the event $\{a(n) \leq v' \quad \forall n\}$ by R-S′, V_n and $\sum_n (v-v')/(n+1)v$ converge. The latter is a contradiction, so it follows that $a(n) > v'$ for some n with probability 1. If follows that $a(n) > v'$ infinitely often with probability 1, since the argument can be applied again on exit from this region. Since $v' < v$ is arbitrary it follows that that $\limsup_{n\to\infty} a(n) \geq v$ almost surely.

Analysis of Counter-Example

Let ζ, be a small number, to be determined. Without loss of generality suppose $p_1(1) < 1 - \zeta$ and set $\tau_1 = 1$. For $k \geq 1$, let $\tau_{2k} = \inf\{n | n > \tau_{2k-1}, \quad p_1(n) \geq 1 - \zeta\}$, and $\tau_{2k+1} = \inf\{n | n > \tau_{2k}, \quad p_1(n) \leq \zeta\}$. Player 2 adopts the following strategy: play strategy 1 in periods τ_{2k-1} to $\tau_{2k} - 1$ and play strategy 2 in periods τ_{2k} to $\tau_{2k+1} - 1$, for all k. Note that τ_k are predictable stopping times and that by Theorem 3 each τ_k is finite almost surely. To prove that under this strategy $a(n) \leq 1.5 - \epsilon$ infinitely often, for suitable choice of ζ and ϵ, it is enough to show that for $k \geq K$, some K, $P(\inf\{a(n)|\tau_k \leq n < \tau_{k+1}\} \leq 1.5 - \epsilon | \text{events up to time } \tau_k) > \delta$, for some $\delta > 0$. (Apply the conditonal Borel-Cantelli Lemma.)

Consider, without loss of generality, an odd cycle, τ_{2k-1} to τ_{2k}, so that player 2 plays strategy 1. By a calculation, the evolution of $a(n)$ and $p_1(n)$ is governed by:

$$a(n+1) = a(n) + \frac{-a(n) + 1 + p_1(n)}{(n+1)a(n)} + \frac{\upsilon_n}{n+1} \tag{A16}$$

$$p_1(n+2) = p_1(n+1) + \frac{p_1(n+1)(1 - p_1(n+1))}{(n+1)a(n)} + \frac{\eta_n}{n+1} + \frac{K_n}{(n+1)^2} \tag{A17}$$

where $E(\upsilon_n|\mathcal{F}_n) = E(\eta_n|\mathcal{F}_n) = 0$ and K_n is bounded. Recall that $a(n)$ is bounded away from zero as payoffs are bounded away from zero and that $p(n+1)$ is a function of reinforcements at time n.

Let $\gamma_n = 1/n$ and $t_n = \sum_{i=1}^{n} \gamma_i$ and $m(l,T) = \inf\{n : \sum_{i=l}^{n} \gamma_i \geq T\}$. Let $\Theta(t, z_0)$ be the flow generated by the system of differential equations

$$\dot{a} = -a + 1 + p_1 \tag{A18}$$

$$\dot{p_1} = \frac{p_1(1-p_1)}{a} \tag{A19}$$

with initial condition z_0. Let $z_n(l)$ denoted the pair of random varables generated by (A16) and (A17) with the same initial condition z_0 at time l.

By a standard result (see Benveniste et. al (1990) Theorem 9 p. 232 and the remarks after it), for all $T > 0$, $\delta_1 > 0$ and $\delta_2 > 0$, there exists L large enough such that for all $l \geq L$,

$$P(\sup_{l \leq n \leq m(l,T)} \|z_n(l) - \Theta(t_n, z_0)\| > \delta_1) < \delta_2 \tag{A20}$$

That is, far enough along the sequence, the path of (A16) and (A17) is close to that of (A18) and (A19) in a finite interval with arbitrarily high probability.

Now in (A18) and (A19), given $\epsilon < 0.25$ there is ζ' such that if $p(t) \leq \zeta'$ for all $t \in [0,2]$, $a(2) \leq 1.5 - 2\epsilon$ if $a(0) \geq 1.5 - 2\epsilon$. Also since a is bounded below, there is ζ, such that for all feasible $a(0)$ $p(t) \leq \zeta'/2$ for all $t \in [0,2]$ if $p(0) \leq \zeta$.

The statement at the beginning will be proved with these values of ϵ and ζ. By choosing δ_1 small enough one can ensure that on an event of probability at least $1-\delta_2$, in the stochastic system if $l \geq L$ and $a(l) \geq 1.5-2\epsilon$ and $p_1(l) \leq \zeta$ then $a(m(l,2)) \leq 1.5-\epsilon$ and $p_1(n) < 1 - \zeta$ for all n with $l \leq n \leq m(l,2)$. Now τ_k is finite almost surely and $\tau_k \geq k$, so let $2k - 1 \geq L$. If $l = \tau_{2k-1}$, the result on the path of p_1 implies that $\tau_{2k} > \tau_{2k-1} + m(\tau_{2k-1}, 2)$, so indeed a falls below $1.5 - \epsilon$ on the cycle of between τ_{2k-1} and τ_{2k} (if $a(\tau_{2k-1}) < 1.5 - 2\epsilon$ there is nothing to prove). Hence the statement to be proved is true with $\delta = 1 - \delta_2$.

Proof of Theorem 8

The same argument as in the proof of Theorem 7, yields (A13) again. Assumption M′ implies that

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + \frac{v - a(n)}{(n+1)a(n)} - \xi_n \tag{A21}$$

where $\sum_n \xi_n$ converges. Assumption A implies that if one defines $\tilde{a}(n) = \min\{a(n), v\}$, then

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + \frac{v - \tilde{a}(n)}{(n+1)\tilde{a}(n)} - \xi'_n \tag{A22}$$

where $\sum_n \xi'_n$ converges (note that $\eta_n \geq 0$, so $\sum_n \eta_n$ is absolutely convergent).

It follows from R-S′ that $\sum_n v - \tilde{a}(n)/(n+1)\tilde{a}(n)$ converges. Now if $\liminf_{n\to\infty} a(n) < v$, then $\tilde{a}(n) < v' < v''$ infinitely often, some v', $v'' < v$ Now if $a(n) \leq v''$, $v - \tilde{a}(n)/\tilde{a}(n)$ is bounded below by L_1 say. Since payoffs are bounded above $a(n+1) \leq a(n) + L_2/(n+1)$, some L_2. Therefore if $m_0 = 0$ and $m_{2k+1} = \inf\{n : a(n) < v', \quad n >$

$m_{2k}\}$ and $m_{2k} = \inf\{n : a(n) > v'', \quad n > m_{2k-1}\}$, then $\sum_{n=m_{2k-1}}^{m_{2k}-1} 1/(n+1) \geq L_3$, some $L_3 > 0$. It follows that if $\tilde{a}(n) < v'$ infinitely often, $\sum_n v - \tilde{a}(n)/(n+1)\tilde{a}(n)$ diverges, in contradiction to R-S$'$.

Proof of Corollaries 1 and 2

Assumption M$'$ is immediate in the case of Corollary 1 and in the case of Corolloray 2 follows from Proposition 9.2.9 of Duflo (1997) as $1/a$ is Lipschitz since a is bounded away from zero.

Let $\pi(n)$ be the player's realised payoff at stage n and let $\tilde{\pi}(n)$ denote his expected payoff under his current strategy. One can write by definition:

$$\pi(n) = \tilde{\pi}(n) + \epsilon_n \tag{A23}$$

where $E(\epsilon_n|\mathcal{F}_{n-1}) = 0$. In the case of Corollary 1, $\tilde{\pi}_n \leq v$. In the case of Corollary 2, $\pi^{\mu}(n) \leq v$, where $\pi^{\mu}(n)$ is the expected payoff of the current strategy if the chain were distributed according to its invariant measure. One can therefore write

$$\pi(n) \leq v + \psi_n + \epsilon_n \tag{A24}$$

where $\psi_n = \tilde{\pi}_n - \pi^{\mu}(n)$ in the second case and is zero in the first case.

Now ϵ_n is a bounded martingale with mean zero, so by Theorem 1.3.17 of Duflo (1997), $\sum_{i=1}^{n} \epsilon_i/n^{3/4}$ tends to zero. Also, by Proposition 9.2.9 of Duflo $\sum_n \psi_n n^{-3/4}$ converges (note that the chain is aperiodic irreducible, $\tilde{\pi}_n$ is linear in the reinforcements, and the reinforcements change by at most a constant times $1/n \leq 1/n^{3/4}$, so the conditions of the Proposition are satisfied), hence by Kronecker's Lemma $\sum_{i=1}^{n} \psi_i/n^{3/4}$ tends to zero.

Now one can write

$$a(n) = \frac{A(0)}{n} + \sum_i \frac{\pi(n)}{n}$$

Hence if $\chi_n = |\sum_{i=1} \psi_i|/n^{3/4}$ and $\nu_i = |\sum_{i=1}^{n} \epsilon_i|/n^{3/4}$ then χ_n and ν_n tend to zero and one has

$$a(n) \leq v + \frac{A(0)}{n} + \frac{\chi_n}{n^{1/4}} + \frac{\nu_n}{n^{1/4}} \tag{A25}$$

This implies condition A.

Proof of Lemma 3

All the statements in Lemma 3, except the fact that $\dot{V} \geq 0$ are obvious. From From (10a) and (10b), since $p_i = a_i/a$ and $q_j = b_j/b$ and payoffs sum to K,

$$\dot{V} = \frac{s_I}{a}\left(\sum_i p_i^* \pi_i^I\right) + \frac{s_{II}}{b}\left(\sum_j q_j^* \pi_j^{II}\right) - \frac{K}{r} \tag{A26}$$

Now since the game is constant sum p^* and q^* earn at least the equilibrium payoff for each player against any strategy, in particular against the current strategies, so

$$\dot{V} \geq \frac{s_I^2 K}{a} + \frac{s_{II}^2 K}{b} - \frac{K}{r} \tag{A27}$$

Now $a + b = r$ and in the set defined by this constraint, the right-hand side of $(A27)$ is minimised when $a = s_I r$ and $b = s_{II} r$. Hence

$$\dot{V} \geq \frac{s_I K}{r} + \frac{s_{II} K}{r} - \frac{K}{r} = 0 \tag{A28}$$

Thus $\dot{V} \geq 0$, as was to be shown.

Proof of Lemma 5

$\dot{V} = 0$ if and only if (A27) and (A28) hold with equality. This implies the statements in the text.

From (11b) and (11d) added together, r converges to K, so no set with points with $r \neq K$ can be invariant. Suppose $r = K$. Since $a = s_I r$ and $b = s_I b$, it follows that $\dot{a} = 0$ and $\dot{b} = 0$. From $(11b)$ and $(11d)$, and part (a) of the Lemma, this can only be true if $\Pi^I = s_I K$ and $\Pi^{II} = s_{II} K$, which implies the result in the text.

Proof of Lemma 6

Let $V_n = V(z_n)$. The same argument as in the proof of Theorem 7 applied to the components of V_n and use of the definition of h yields

$$E(V_{n+1}|\mathcal{F}_n) \geq V_n + \frac{1}{n+1}\langle \nabla V_n, h \rangle - \kappa_n \tag{A29}$$

where $\kappa_n \geq 0$, $\sum_n \kappa_n < \infty$ and $\langle \nabla V_n, h \rangle \geq 0$ (as V is a Lyapounov function).

Result RS$'$ implies that with probability 1 (i) $\lim_{n\to\infty} V_n$ exists and is finite (ii) $\sum_n \frac{1}{n+1}\langle \nabla V_n, h \rangle < \infty$. Since $\langle \nabla V, h \rangle$ is continuous on Δ^*, (ii), Lemma 3.III.8 on p. 102 of of Duflo (1996) implies that $\lim_{n\to\infty}\langle \nabla V_n, h \rangle$ is zero.

The fact that z_n converges to an invariant set of h follows from Kushner and Yin (1997) Theorem 2.1 on p. 95.

Proof of Lemma 7

From Lemma 5 parts (a) and (b) and the fact that p^* and q^* yields an equilibrium of the game

$$(p - p^*)\mathcal{A}(q - q^*) = 0 \tag{A30}$$

if (p, q) corresponds to a point in any invariant subset of $\langle \nabla V, h \rangle = 0$.

Since the case of pure equilibria is covered in Theorem 11, one can assume that at least one of p^* and q^* is fully mixed. In that case, since equilibrium is unique and the game is constant-sum and 2×2, both must be fully-mixed. It follows that p^* and q^* are the unique solutions of $\mathcal{A}q = v_I e$ and $q\mathcal{A} = v_{II} e'$, where e is a column vector of ones and $'$ denotes transpose. Hence $\mathcal{A}$ is non-singular.

Now in the 2×2 case, it is easy to check that

$$\xi_0 \mathcal{A} \eta_0 = 0$$

where ξ_0 and η_0 are vectors each of whose components add to zero, cannot have a solution with both ξ_0 and η_0 non-trivial if A is non-singular.

It follows that either $p = p^*$ or $q = q^*$. Say $p = p^*$. Now if $q \neq q^*$, either $(\mathcal{A}q)_1 > v_I$ or $(\mathcal{A}q)_2 > v_I$. Suppose the former. It follows from (11a) and $p^* \mathcal{A} q = v_I$, that $\dot{p}_1 > 0$. Hence the the system will not stay in this invariant set, which contradicts the assumption that it is invariant.

Proof of Theorem 12

The only fact not proved elsewhere is the convergence of emprical frequencies. Let x_{ij} denote the number of times that strategy i is played at the same time as strategy j. One can apply an argument of Benaim and Hirsch (1999) Theorem 4.1 and note that the long-run evolution of x_{ij} is governed by

$$\dot{x}_{ij} = -x_{ij} + p_i q_j$$

and deduce from this that x_{ij} converges to $p_i^* q_j^*$.

Proof of Theorem 15

This follows directly from Kushner and Yin (1997) Theorem 6.10.6. Verification of the conditions of Theorem 6.10.3 follows, using the remarks after it for the constant-step case, from the upper bound in Dupuis and Kushner (1989) Theorem 5.3 (note that the distribution of the random terms in the model under consideration depends solely on the current state and is weakly continuous in it).

Figures

	1	2
1	$2, 1$	$1, 2$
2	$1, 2$	$2, 1$

Figure 1

	1	2
1	$2, 0.1$	$0.1, 2$
2	$0.1, 2$	$2, 0.1$

Figure 2

Figure 3

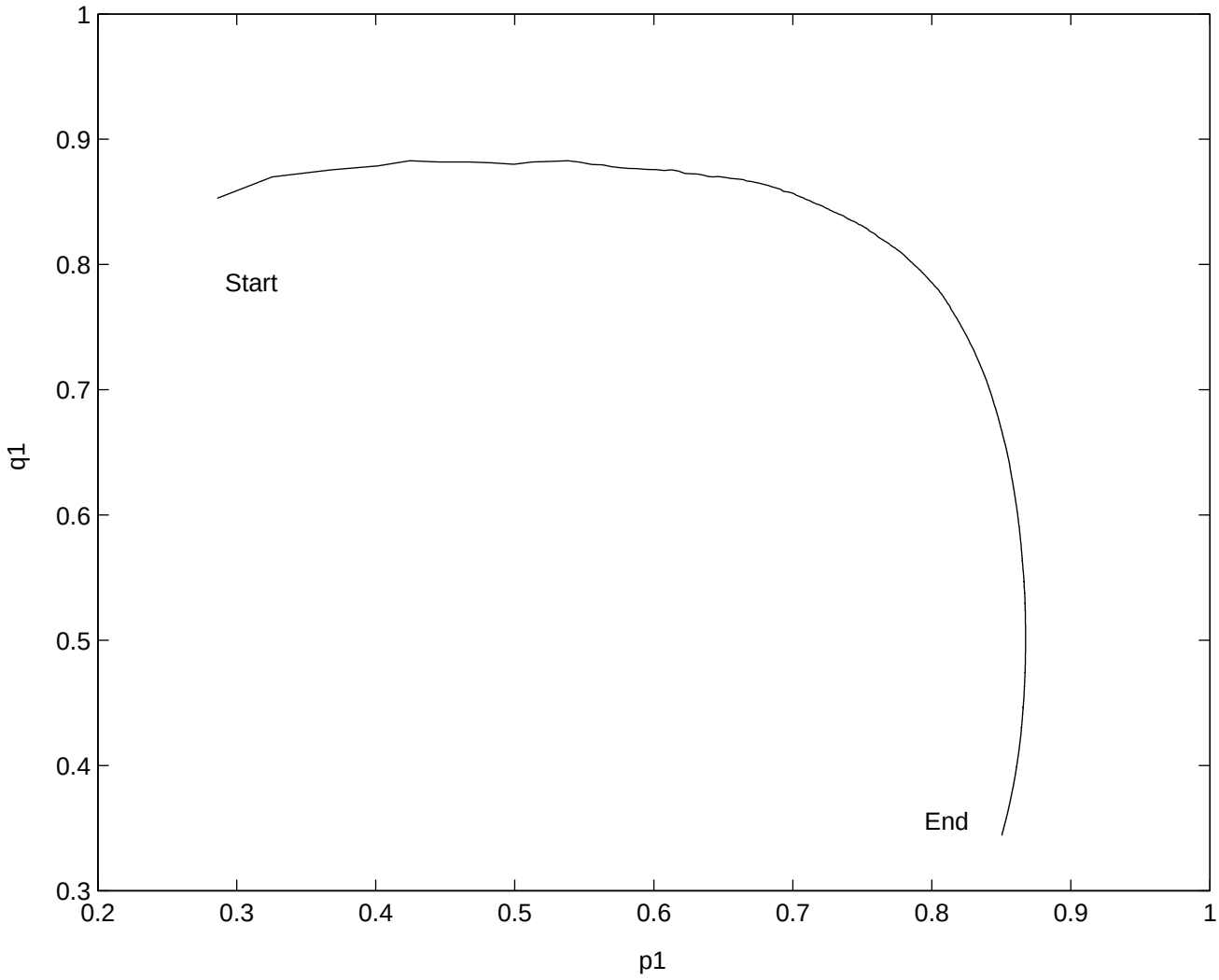

Figure 4

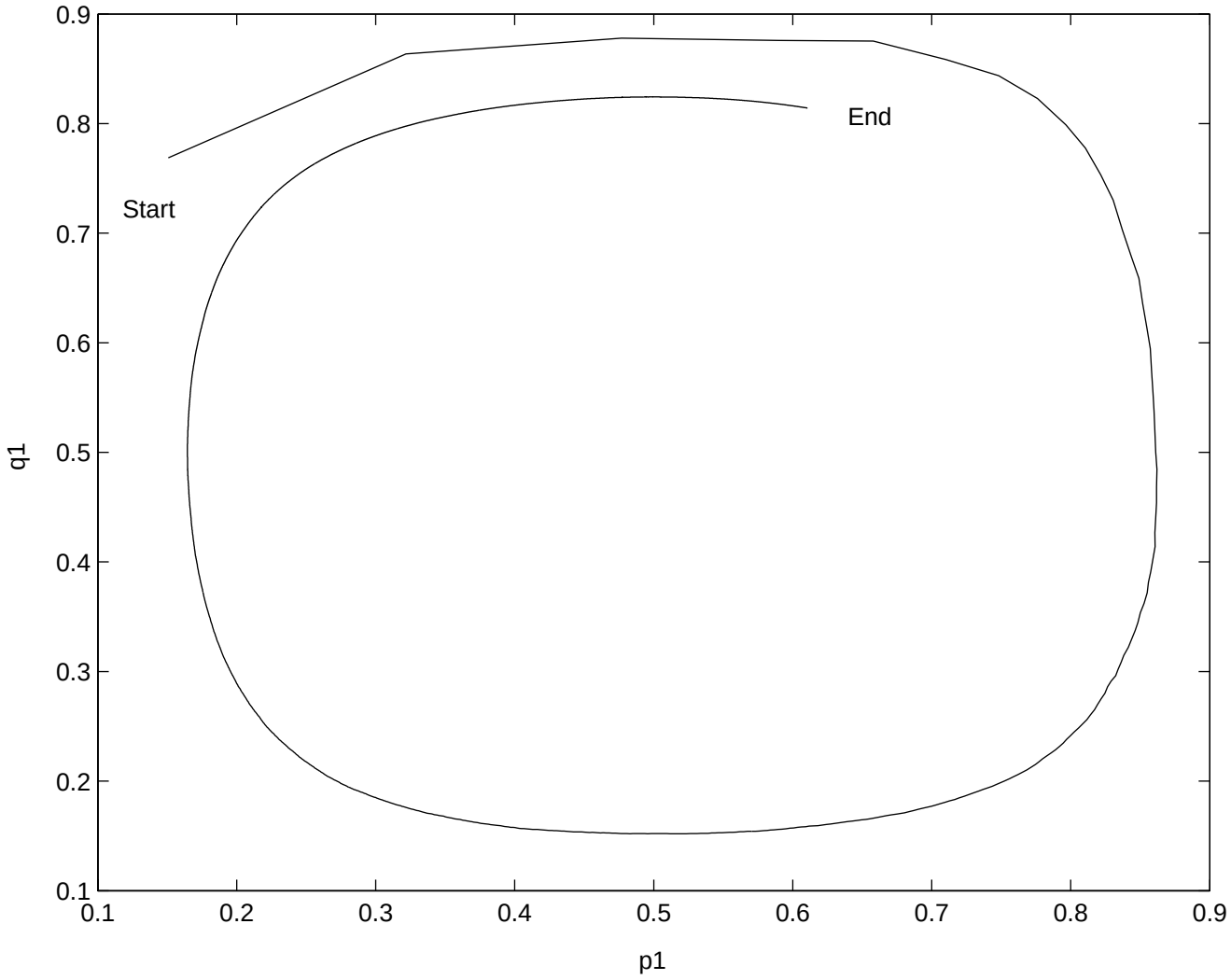

Figure 5