On the convergence of reinforcement learning

A.W. Beggs

Wadham College, Oxford, OX1 3PN, UK

Received 8 April 2002; final version received 30 March 2004
Available online 15 September 2004

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 if they are mixed and the game is $2 \times 2$. 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.

© 2004 Elsevier Inc. All rights reserved.

JEL classification: C72; D83

Keywords: Reinforcement learning; Games

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 [37] and Erev and Roth [14] 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 [20] proposed a similar model in a biological context.

E-mail address: alan.beggs@economics.ox.ac.uk.

0022-0531/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2004.03.008
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 a constant-sum game if it has a unique pure strategy equilibrium or it has a unique-mixed strategy equilibrium and is $2 \times 2$.

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’s [33] version of the replicator dynamic. This may be of independent interest since in $2 \times 2$ 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 [6,25], 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 is also shown that a player’s long-run average payoff cannot be forced permanently below his minmax payoff. That is if a player uses the ER rule, the lim sup of his average payoffs will be at least this. More generally it is shown that his long-run average payoff cannot be forced below any payoff he can guarantee himself on average by playing a fixed action. A clever player can, however, exploit the inertia in the Erev and Roth scheme and force the lim inf of the player’s average payoffs below his minmax value, so that his average payoff is below this infinitely often.

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 [4] study convergence of strategies in games with randomly perturbed payoffs. 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 [16] provide a good summary of this work. Hart and Mas-Collel [21–23] 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. [2] and Hart and Mas-Collel [22] 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 [22] 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 [9] suggest that both it and fictitious play have features which match the data and present a synthesis.
Börgers and Sarin [7] analyse a reinforcement learning model and discuss its relationship to the so-called replicator dynamic used in biology (see for example [26]). 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 [38] 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 et al. [32] 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 [36] studies the convergence of a related learning model of [1] in $2 \times 2$ games. In this model equilibrium play cycles round a mixed-equilibrium with positive probability rather than converging. His work is discussed in more detail in Section 5. He does not discuss convergence of payoffs. Ianni [28] provides some results on convergence to strict equilibria with positive probability for this model, similar to those of Laslier et al. [32] for the Erev and Roth model.  

Hopkins [27] 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 and draws some connections with the theory of urn models. Section 3 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 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

---

1 Skyrms and Pemantle [39] analyse models of reinforcement in the context of network formation.
a unique equilibrium in the case of pure equilibria in general and mixed equilibria in $2 \times 2$ 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

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

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

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 [14]). For large $n$ 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. Eq. (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.

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

\[
\frac{A_i(0)}{\sum_j A_j(0) + nk_2},
\]

which corresponds to \( i \) never having been chosen previously. Summed over \( n \) 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 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

The section focusses on the implications of the ER rule for the learning of a single player. It shows that if he has an action that dominates others, he will learn to play it under very weak assumptions. These results are then applied to learning in games. In particular it shows that if all players use the ER rule, they will not in the long-run play strategies ruled out by iterated elimination of dominated strategies. The relevant single player results are obtained in Section 3.1. They are applied to games in Section 3.2.

3.1. General results

To begin with consider the case \( m = 2 \), that is there are only two actions available to the decision-maker. The key result is that if action 1 dominates action 2, in the sense that its expected payoff is always higher by a constant factor, then the decision-maker will converge
to playing action 1. This is in spite of the inertia of the learning rule that favours actions that have been played in the past. The fact that action 1’s payoff is greater than action 2’s turns out to be enough to ensure that its reinforcement grows faster.

The result makes no requirement that the expected payoffs to the two actions are constant or are even are generated by a stationary process. This is important for learning in games, since opponents’ strategies may be changing.

To state the result formally, 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 1.** Suppose that $E(\pi_1(n + 1) | \mathcal{F}_n$, action 1 is chosen at time $n + 1) > \gamma E(\pi_2(n + 1) | \mathcal{F}_n$, 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 [35]. 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 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), $A_2(n)^\varepsilon/A_1(n)$, for $0 < \varepsilon < \gamma$, is a positive supermartingale and hence convergent. It follows from this (choose $1 < \varepsilon < \gamma$) 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 [19, 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.

Theorem 1 immediately generalises to the case of many actions. For 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 above. 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 2.** 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 1 (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.

An immediate corollary of Theorem 2 is

**Corollary.** Suppose the expected payoff to each action is constant and independent of past history, then, with probability 1, the probability with which the decision-maker plays any action which does not have the maximum expected payoff tends to zero.

This corollary is obtained by Laslier et al. [32]. It is much less general than Theorem 2, which does not require that the expected payoffs be constant or independent of past history.

Rustichini [38] 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 5 before Theorem 9, 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 be bounded away from zero.

The speed of learning under the ER model is, unfortunately, slow. Suppose for example that the decision-maker has only two actions available: 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.
3.2. Application to games

Suppose now that the payoffs to the decision-maker are generated by playing a fixed finite normal-form game, \( \Gamma \), repeatedly at each stage against other players. These other players may use arbitrary, possibly history-dependent, strategies, the only restriction being that at any stage they can only base actions on events prior to that stage (that is they cannot observe the decision-maker’s action before he takes it). In other words they are restricted to behavior strategies (possibly correlated): that is a sequence of measurable mappings, \( n = 0, 1, 2, \ldots \) from \( \mathcal{F}_n \) to the set of mixed actions available at stage \( n + 1 \).

Suppose that at stage \( n \) if the decision-maker takes action \( k, k = 1, \ldots, m \), he obtains payoff \( U(k, \sigma) \), where \( \sigma \) represents the action taken by the other players. Action \( i \) strictly dominates action \( j \) if \( U(i, \sigma) > U(j, \sigma) \) for all \( \sigma \). Now if the other players have only finitely many actions, \( \sigma \) can only take finitely many values, so if \( i \) dominates \( j \) there exists \( \gamma > 1 \) such that \( U(i, \sigma) > \gamma U(j, \sigma) \) for all \( \sigma \). It follows that the expected payoff to action \( i \), for any past history of play, dominates that of \( j \) by a constant factor, \( \gamma \), at each stage, and hence the assumptions of Theorem 2 apply.

One therefore obtains:

**Theorem 3.** 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.

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 \( A^i(n)/A_j^\sigma_1(n) \ldots A^m_m(n) \) is a positive supermartingale, where \( \sigma_1, \ldots, \sigma_n \) is the dominating mixture of strategies 1, \ldots, \( m \) and \( \epsilon < \gamma \), where \( \gamma > 1 \) is such that the mixed strategy always earns \( \gamma \) times more than \( i \). It follows that
Theorem 4. 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

The last section showed that if a player has a dominant strategy then he will learn to play it if he uses the ER rule. This seems a minimal property that should be satisfied by a reasonable learning rule. In many games, however, there is no dominant strategy. It is then less clear what one might expect of a plausible learning rule.

If a player plays a fixed game repeatedly, then one might argue that a reasonable learning rule should guarantee that in the long run he should do at least as well as he could by playing his maximin strategy. Fudenberg and Levine [16, Chapter 4] refer to this property as ‘safety’.

The criterion explored in this section is a generalisation of this. It is asked whether a player will do at least as well as the payoff he can guarantee himself on average by playing a fixed action. It is shown that employing the ER rule guarantees that a player cannot be forced permanently below this payoff. A clever player may, however, be able to force him repeatedly below it. In particular cases the result is strengthened to show that he will, in fact, do at least as well as this.

Section 4.1 sets out the main results. Section 4.2 gives some intuition; some of this material is useful in Section 5. Section 4.3 analyses a counterexample.

4.1. Main results

The general framework is as in Section 2. Attention is focussed a single decision-maker. The actions of other players, if any, are not restricted, except where mentioned, save that, as in Section 3.2, they can only be based on events prior to the current stage.

Attention will focus on long-run average payoffs. Consider therefore the time averages of the reinforcements, \( a_i(n) = A_i(n)/n \) for \( i = 1, \ldots, m \). Let \( A(n) = A_1(n) + \cdots + A_m(n) \) be the total reinforcement received by the player up to time \( n \). Its time average \( a(n) = A(n)/n \) is asymptotically equal to the total average payoff received up to time \( n \). They are not necessarily the same for finite \( n \), on account of the initial reinforcements, but as the focus is on long-run average payoffs 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.

Consider a fixed mixed strategy \( \sigma = \sigma_1, \ldots, \sigma_m \) for the agent and let \( \pi_\sigma(n) = \sum_i \sigma_i \pi_i(n) \) denote its ‘expected’ payoff at stage \( n \), where the expectation is taken over the player’s own actions not those of his opponents. As in Section 3 let \( F_n \) denote the sigma-field generated
by events, including the agent’s own actions, up to and including stage $n$. $\sigma$ will be said to guarantee a payoff of $v$ on average if the following holds.

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

The interpretation is that playing $\sigma$ in any period guarantees an expected payoff of $v$, regardless of the history of previous play, apart from an asymptotically negligible disturbance. With $v_n = 0$ for all $n$, this corresponds to the usual idea of a maximin payoff. Condition M weakens this so that it need hold only in a long-run average sense: if $\sum_n v_n/n$ converges then $\sum_{i=1}^n v_i/n$ converges to zero almost surely, so by a strong law of large numbers, $\liminf_{n \to \infty} \sum_{i=1}^n \pi(n; \sigma)/n \geq v$ almost surely, where $\pi(n; \sigma)$ denotes the payoff a player would earn at stage $n$ if he were to play $\sigma$. In other words, the player’s long-run average payoff if he were to play $\sigma$ forever would be at least $v$. The condition is a standard one in the stochastic approximation literature. It can probably be weakened but it seems adequate for most examples.

The extra generality gained by allowing $v_n$ to be non-zero is useful as it allows one to treat environments which may be changing in some predictable way: the optimal action may vary from period to period, but one can still ask how good a fixed action is in a long-run average sense. For example assumption $M$ holds in the environment considered by Rustichini [38] where the state evolves according to a finite stationary Markov chain: in each period the payoff to action $i$ is $\pi_{ij}$ if the state is $j$, where $j$ is the state of a finite irreducible, aperiodic Markov chain. It also holds for even more general Markov chains, even with some limited feedback of actions into the environment, see for example [5] and under general mixing conditions (see for example [30]).

The main result is (proof in appendix):

**Theorem 5.** If condition $M$ holds then $\limsup_{n \to \infty} a(n) \geq v$ almost surely.

In particular, if a fixed game is played repeatedly, the average payoff of a player using the ER rule is greater than his minmax payoff infinitely often.

In other words, the player cannot be forced infinitely often below an average payoff which he can guarantee himself by playing a fixed action. Intuition for the result is given in Section 4.2.

The result cannot, however, in general be strengthened to read ‘$\lim inf$’. A counterexample, analysed in Section 4.3, shows that a clever opponent may be able to exploit the ER rule and drive the player’s average payoff below his minmax payoff infinitely often.

The strategies used in the counterexample involve luring the player using in the ER rule into playing a certain action by allowing him a payoff above what he can guarantee himself. The opponent then changes action suddenly, exploiting the slow-adjusting nature of the

---

2 Use Kronecker’s Lemma—see [19, p. 31].

3 For example [19, p. 36].

4 See for example [12, Proposition 9.2.9].

5 It also holds for even more general Markov chains, even with some limited feedback of actions into the environment, see for example [5] and under general mixing conditions (see for example [30]).
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.

Corollary 2. Suppose that player 1 faces an environment in which if he plays strategy \( i \) he obtains payoff \( \pi_{ij} \), where the state \( j \) follows a finite irreducible, aperiodic Markov chain. If he uses the ER rule then player 1’s long-run average payoff converges to the payoff of the action which yields highest long-run average payoff.

Corollary 2 is Rustichini’s [38] result. 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.

The results of this section are in a sense more general than in the last section as they do not require the decision maker to have a dominant strategy each period. Corollary 2 clearly generalises the corollary to Theorem 2 and is not implied by Theorem 2. In other ways though, they are less general. Theorem 2 requires no assumption about stationarity, as does Corollary 2, nor does it require assumption 4. The results here also concern payoffs and do not necessarily imply convergence of actions.

These are of course asymptotic results. Nevertheless, they seem to work well in practice. For example in simulations of the game in Fig. 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.

The results above leave open the case when all players use the ER rule. This is treated separately in Section 5.

4.2. Intuition

This sub-section provides some intuition for the results above. The ideas introduced will be useful in Section 5.
Eq. (1) can be re-written in terms of average reinforcements as

\[ a_i(n + 1) = a_i(n) + \frac{1}{n + 1} (\bar{a}_i(n + 1) - a_i(n)), \tag{4} \]

where \( \bar{a}_i(n + 1) \) equals \( a_i(n + 1) \) if action \( i \) is played at stage \( n + 1 \), 0 otherwise. Equivalently,

\[ a_i(n + 1) = a_i(n) + \frac{1}{n + 1} (\bar{a}_i(n + 1) - a_i(n) + u_{n+1}), \tag{4}^{'} \]

where \( p_i = a_i(n)/a(n) \) is the probability that action \( i \) is played at stage \( n + 1 \), \( E(u_{n+1}|\mathcal{F}_n) = 0 \), \( \mathcal{F}_n \) being the \( \sigma \)-field generated by events up to stage \( n \), and \( \bar{a}_i(n + 1) = E(\pi_i(n + 1)|\mathcal{F}_n) \).

\( \bar{a}_i(n + 1) \) is the expected payoff to playing \( i \) at time \( n + 1 \) conditional on events at time \( n \). Note that all players must use strategies which are measurable with respect to \( \mathcal{F}_n \).

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

\[ \frac{da_i}{dt} = -a_i(t) + p_i \bar{a}_i(t), \tag{5} \]

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 is 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

\[ V = \sum_i \sigma_i \ln a_i. \tag{6} \]

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 (5) then the rate of change of \( V \) is

\[ \dot{V} = \frac{\bar{a}_\sigma - a}{a} \tag{7} \]

where \( \bar{a}_\sigma(t) \) is the expected payoff to \( \sigma \) at time \( t \). If \( \sigma \) 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 \( a p_1^{\sigma_1} \ldots p_m^{\sigma_m} \), 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 \( \sigma \) (as measured by the geometric average of probabilities) and so eventually his payoff will rise.
The above suggests that if $\sigma$ 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 $\sigma$ guarantees $v$ on average, in the sense that condition M holds.

The formal proof of Theorem 5 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 disturbance terms can be neglected.

Now if one looks at (7), 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 $v$. More generally, one does not require that it always be below $v$ but only that the fluctuations below it tend to zero sufficiently rapidly. This idea is used in the proof of Corollaries 1 and 2.6

4.3. Counter example and discussion

As noted above, one cannot in general replace lim sup by lim inf in Theorem 5:

Counterexample: Consider the game in Fig. 1. 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 $1.5 - \varepsilon$, for some $\varepsilon > 0$ infinitely often.

To see this note that one can re-write Eq. (5) in terms of $p_i = a_i/a$ and $a$:

\[
\dot{p}_i = \frac{p_i(\bar{p}_i - \Pi)}{a},
\]

\[
\dot{a} = -a + \Pi,
\]

$p_i$ can be interpreted as the probability that the player plays action $i$ and $\Pi$ 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 Theorem 5. Now some simple procedures, such as for example smoothed fictitious play, but not fictitious play itself, [2,16, Chapter 4], or the procedures considered by Hart and Mas-Collel [22,23] do have this property (or strictly almost in the case of smoothed fictitious play). Indeed they have the

---

6 A slightly more general result on these lines is given in the appendix.
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 in the example 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. It is not clear therefore how serious a deficiency this reveals in the ER rule.

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 if there is a unique mixed-strategy equilibrium when the game is $2 \times 2$. The analysis uses similar techniques to the previous sections. It is shown that the behaviour of the system is related to a system of equations similar to the adjusted replicator dynamic introduced by Maynard Smith [33]. This may be of some independent interest as in $2 \times 2$ games 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.7 Several studies have derived the replicator dynamic from a learning model (for example [7,17] but justification for the Maynard Smith dynamic is rather more elusive. Björnerstedt and Weibull [6] and Hofbauer and Schlag [25] 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 1 receives payoff $\alpha_{ij}$ and player 2 payoff $\beta_{ij}$, 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 $A = (\alpha_{ij})$ be the $m \times l$ payoff matrix for player 1 and $B = (\beta_{ij})$ 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 $p^t A q$ and $p^t B q$, 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

---

7 See for example [33, Appendix J] or [26, Chapter 27].
Let \( r = a + b \) denote the sum of the total average reinforcements. \( r \) converges to \( K \) asymptotically, since reinforcements after time 0 simply cumulate 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 \( b_j(n) \), 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

\[
\mathcal{A} = \{(a_1, \ldots, a_m, b_1, \ldots, b_l) | 0 \leq a_i \leq M, 0 \leq b_j \leq M \text{ for all } i, j, \ a \geq \kappa, b \geq \kappa\}
\]

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

As discussed in Section 4 in the derivation of Eq. (5), the long-run behaviour of the system is related, at least heuristically, to that of the system \(^8\)

\[
\begin{align*}
\frac{da_i}{dt} &= -a_i(t) + p_i \pi^I_i(t), \quad i = 1, \ldots, m, \quad (9a) \\
\frac{db_j}{dt} &= -b_j(t) + q_j \pi^{II}_j(t), \quad j = 1, \ldots, l, \quad (9b)
\end{align*}
\]

where \( p_i = a_i/a \) and \( q_j = b_j/b \) and \( \pi^I_i(t) = (Aq)_i \) and \( \pi^{II}_j(t) = (pB)_j \) are the expected payoffs of actions \( i \) and \( j \).

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

\[
\begin{align*}
\frac{dp_i}{dt} &= \frac{p_i(t)(\pi^I_i(t) - \Pi^I(t))}{a(t)}, \quad i = 1, \ldots, m, \quad (10a) \\
\frac{da}{dt} &= -a(t) + \Pi^I(t) \quad (10b) \\
\frac{dq_j}{dt} &= \frac{q_j(t)(\pi^{II}_j(t) - \Pi^{II}(t))}{b(t)}, \quad j = 1, \ldots, l, \quad (10c) \\
\frac{db}{dt} &= -b(t) + \Pi^{II}(t), \quad (10d)
\end{align*}
\]

where \( \Pi^I(t) = pAq \) and \( \Pi^{II}(t) = pBq \) 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 (10b) and (10d) are replaced by the condition that \( a \) and \( b \) are constants. In the Maynard Smith dynamic, (10b) and (10d) 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.

---

\(^8\) This can be made rigorous, see for example [3], but in this paper the equations are simply used to guide the construction of appropriate martingales, so this is not considered in detail.
For the moment, attention will be focussed on payoffs. Recall (see for example [29]) 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 [29, Chapter 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
\[ V = s_I \left( \sum_i p^*_i \ln a_i(t) \right) + s_{II} \left( \sum_j q^*_j \ln b_j(t) \right) - \ln r(t) , \]  

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 $\mathcal{A}^* = \{(a, b) | (a, b) \in \mathcal{A}, \ p^*_i > 0 \Rightarrow a_i > 0, \ q^*_j > 0 \Rightarrow b_j > 0, \ \text{for all } i, j\}$

Since initial reinforcements are all assumed positive, the initial conditions of the system can be assumed to lie in $\text{int}(\mathcal{A}) \subseteq \mathcal{A}^*$, so $V$ is certainly well defined initially. In fact (again see [29, Chapter 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 $\mathcal{A}^*$.

Let $\partial \mathcal{A}^*$ be the boundary of $\mathcal{A}^*$ considered as a subset of $\mathcal{A}$:
\[ \partial \mathcal{A}^* = \{(a, b) | (a, b) \in \mathcal{A}, \ \exists i, \ a_i = 0 \text{ and } p^*_i > 0 \text{ or } \exists j, \ b_j = 0 \text{ and } q^*_j > 0\} \]

It is shown in the appendix that $V$ is a weak Lyapounov function: that is $\dot{V} \geq 0$. It follows that the system converges to the largest invariant set in $\dot{V} = 0$ and it is shown in the appendix that at any point in this set each player earns the value of the game to them. One therefore obtains (details in appendix):

**Theorem 6.** Any solution of the system governed by (9a) and (9b) in a constant-sum game with initial conditions in $\mathcal{A}^*$ remains bounded away from $\partial \mathcal{A}^*$ 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 in Section 4, it is shown by a Taylor expansion that $V$ is a sub-martingale plus some perturbations. As there, the logarithmic form of $V$ means some care is required with the second-order terms. It is shown in the appendix that:

**Theorem 7.** If both players learn according to the ER model in a constant-sum game, then $\lim_{n \to \infty} a(n) = v_I$ and $\lim_{n \to \infty} b(n) = v_{II}$ with probability 1.
In other words, the long-run average payoffs of the players converge to their values. Of course this says nothing about convergence of strategies. This will be investigated in the next sub-section.

5.2. Strategies

One can obtain results on the convergence of strategies in games with unique equilibria by characterising the invariant sets in $\dot{V} = 0$ more fully. 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 [29, Chapter 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 8.** *If a constant-sum 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.

Laslier et al. [32] show that if a game has a strict equilibrium, then play converges to it with positive probability. If a constant-sum game has a unique pure strategy equilibrium, then it must be strict. The result above is not, however, implied by Laslier et al.’s [32] result, as they do not show that if a game has a unique equilibrium which is strict then play converges to it with probability 1. Theirs is essentially a local analysis: they show that if the process starts close enough to the equilibrium point it converges to it with positive probability and then argue that there is positive probability that the process reaches this neighbourhood if it starts outside it. A sharper result is possible here as the global behaviour of the process can be characterised.

In the case of (constant-sum) $2 \times 2$ games, one can also prove convergence if there is a unique mixed strategy equilibrium. The following is shown in the appendix:

**Lemma 3.** *If a game is $2 \times 2$, constant-sum and has a unique mixed strategy equilibrium, then the unique invariant set in $\dot{V} = 0$ corresponds to both players playing their equilibrium strategy. The mixed strategy equilibrium is therefore asymptotically stable under dynamic (9).*

It follows that in $2 \times 2$ games a mixed strategy equilibrium is asymptotically stable under dynamic (9) or equivalently (10), 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.9

It is perhaps worth noting here a point that has been suppressed in some of the discussion above. The fact that in $2 \times 2$ games all trajectories beginning in the interior of $A$ converge to the equilibrium point under (10), does not of itself imply that the same is true of the ER algorithm. The boundary of $A$ is an absorbing set and although there are results ruling out

---

9 I am grateful to Josef Hofbauer for some helpful correspondence on the Maynard Smith dynamic.
convergence of stochastic approximation to unstable sets of the corresponding differential equation. 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, given in the appendix, to rule convergence to the boundary.

In any case, here one has (proof in appendix)

**Theorem 9.** In a $2 \times 2$ constant-sum game with a unique mixed-strategy equilibrium, the probabilities with which each player plays each strategy converge to their equilibrium values 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.

5.3. Discussion

This sub-section discusses the results above, particularly Theorem 9, and relates them to the literature.

Posch [36] studies a related model of Arthur [1] in the context of $2 \times 2$ 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 $C$. As a result he obtains the ordinary replicator dynamic as governing the stochastic evolution of the system rather than the system above—in effect $a$ and $b$ are both always equal to $C$. In his model, therefore, strategies cycle around the equilibrium levels, just as in the deterministic replicator. The model here yields dynamics closer to the Maynard Smith replicator and so it converges to the mixed strategy equilibrium. Posch does not consider the convergence of payoffs.

Börgers and Sarin [7] consider a somewhat different model of learning, in which they find that even if there is a unique mixed equilibrium, it always converges to a pure strategy combination. Their model is discussed in more detail in Section 6.1, but the essential difference with the current model, and Posch’s, is that the impact of a given observation on learning does not decline with experience as it does here (see for example Eq. (4) in Section 4.2). As shown in Section 6.1, the current model yields a similar prediction if it is modified to have this feature.

The differences between these models may, however, not be as great in practice as their asymptotic behaviour might suggest. In finite samples, the convergence of strategies in the ER model appears to be extremely slow. Similarly in Section 6.1, where the model is modified in a way which makes it closer to [7], convergence to the boundary does not appear relevant in finite samples. The speed of convergence in the current model is considered in more detail in the rest of this sub-section.

If one examines Eqs. (10) in the case of Fig. 1 (Matching Pennies), it is easy to check that the eigenvalues of the system 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 $2 \times 2$ games with a unique fully-mixed equilibrium the eigenvalues are $−1, −1$ and a pair of conjugate purely imaginary

---

10 For example [34,8] in the case of unstable equilibria, [3] in the case of more general sets.

11 Ianni [28] and, for some results, Hopkins [27] also use a similar normalisation.
ones. The real eigenvalues intuitively come from the evolution of \( a \) and \( b \) in (10b) and (10d). The imaginary eigenvalues come from the evolution of \( p_1 \) and \( q_1 \) in (10a) and (10c) 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 [12, Chapter 2]) require the largest real part of the eigenvalues, \( \tau \), 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 ([10] has some, but these do not cover the current case), but it seems plausible that the rate is extremely slow.

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 (4)). 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 convergence 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 a game at a very respectable rate but convergence of strategies is slow. For example the game in Fig. 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.

Fig. 3 shows the behaviour of \( p_1 \) and \( q_1 \) in an extremely long run of the game of Fig. 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 periods 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 suppressed).
On the other hand, consider the game in Fig. 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 Fig. 3, the same long run, $10^7$ iterations, is done as in Fig. 4 (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.\textsuperscript{12}

\textsuperscript{12}The simulations presented are for symmetric games. A similar picture emerges for asymmetric games.
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 equilibrium in simulations. These dynamics react strongly to even small deviations from the mixed strategy equilibrium and so force play back there quickly.

6. Discussion

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

6.1. 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

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

where \( \phi \) 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

\[ a_i(n) = \frac{1}{n+1} \sum_{i=1}^{m} k_i = \frac{1}{n+1} \sum_{i=1}^{m} \frac{p_i \pi_i(n+1) - a_i(n) + u_{n+1}}{a_i(n)} \]

where \( a_i(n) \) essentially cumulates past payoffs discounted at the rate \( \phi \) and the factor \( 1 - \phi \) normalises them so they are comparable for different values of \( \phi \).

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

\[ a_i(n+1) = a_i(n) + \psi(p_i \pi_i(n+1) - a_i(n) + u_{n+1}). \tag{4}'' \]

For small \( \psi \), that is \( \phi \) close to 1, one would again expect (5) to describe the evolution of \( a_i \) well.

Assume therefore that both players use the ER rule with the same forgetting rate \( \psi \).\(^\text{13}\) 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 \( a_1, a_2, \ldots, a_m, b_1, \ldots, b_l \) 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 (9) at time \( t \) with the same initial position. \( \| \ldots \| \) denotes the Euclidean norm.

The following is an immediate consequence of standard results, for example [5, Theorem 1, p. 43].

**Theorem 10.** For any \( \varepsilon > 0 \) and \( T > 0 \), there is constant \( C(\psi) \) with \( C \to 0 \) as \( \psi \to 0 \) such that for any \( z_0 \),

\[ \Pr \left( \sup_{n \leq T} \| z_n - \Theta(t_n, z_0) \| > \varepsilon \right) < C(\psi). \]

In other words, by making the step-size small enough 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 \( \psi \) 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, 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 \( 2 \times 2 \) constant-sum games. For any fixed \( \psi \), 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). Applying Theorems 4.3 and 6.1 of Chapter 3 of that book, one obtains.

\(^{13}\) Allowing different rates would only introduce different scale factors in the corresponding differential equations.
Theorem 11. For any $\psi > 0$, in a $2 \times 2$ constant-sum game, play converges to a pure-strategy combination.

This result is very similar to that obtained by Börgers and Sarin [7] 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, while in the replicator dynamic play cycles around it. The contrast with the decreasing step-size model is at first sight puzzling. It is perhaps best understood by noting that for any fixed $\psi$, no matter how much time has passed there is a constant probability that there will a large chance fluctuation away from the deterministic path which results in the process becoming stuck near the boundary. This therefore happens eventually 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 $\psi$ if it starts near the equilibrium point. More precisely one can show that its expected escape time from a small neighbourhood of the equilibrium tends to infinity at an exponential rate.

Let $B_{\sqrt{\psi}}(\mu)$ denote a neighbourhood of the equilibrium point such that any trajectory of the differential equation which starts in it remains within a distance $\mu$ 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 10, 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 [31] (see appendix) and show:

Theorem 12. In a $2 \times 2$ constant-sum game with a unique equilibrium point, for small enough $\mu$ there is constant $c > 0$ such that $E_x \tau^{\psi} \geq \exp(c/\psi)$ for all $x \in B_{\sqrt{\psi}}(\mu)$.

In any case, Theorem 11 should not be taken too literally.

Simulations for the game in Fig. 1 (‘Matching Pennies’) suggest that for moderate values of $\psi$, the process indeed tends to get stuck near the boundaries. On the other hand for small $\psi$ the differential equation provides a good approximation. Fig. 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 $\psi$ 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.

6.2. 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) \( \lambda \) and \( \mu \) for which \( \lambda x_{ij} + \mu \beta_{ij} = K \) for all \( i, j \). For players’ learning is the same as in the constant-sum game where they have payoffs \( \lambda x_{ij} \) and \( \mu \beta_{ij} \).

Adding a constant to all payoffs, so for example player \( i \)’s payoffs become \( x_{ij} + M \), 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 [14], 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 \( x_{ij} - y_{ref} \). 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 \( z_{ij} + \delta_k \) if \( j = k \), \( z_{ij} \) 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.

6.3. Modifications of the choice function

The form of the choice function used is somewhat special. This section illustrates this and discusses briefly possible modifications.

Consider, for illustrative purposes, a special case of the framework in Section 3. The decision maker has available two actions, with constant payoffs \( z_1 \) and \( z_2 \), respectively, with \( z_1 \geq z_2 \).

If \( \tilde{N}_i(n) \) is the number of times action \( i \) has been played up to and including stage \( n \), then \( A_i(n) = z_i \tilde{N}_i(n) + A_i(0) \). Let \( N_i(0) \) be defined by \( A_i(0) = z_i N_i(0) \) and \( N_i(n) = \tilde{N}_i(n) + N_i(0) \). \( N_i(0) \) 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

\[
f(x_n) = \frac{z_1 x_n}{z_1 x_n + z_2 (1 - x_n)}. \tag{12}
\]

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 et al. [24] show this, but this result will not be needed in the current discussion.

Consider the special case when \( z_1 = z_2 \), 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 the so-called Polya urn scheme and one can show that \( x_n \) converges to every point of \([0, 1]\) with positive probability. While this lack of determinancy may be seem annoying, it is useful in convergence to mixed strategy equilibria. Eq. (10a) reveals that if all actions have equal expected payoffs, then any mixed strategy combination is a rest point. If the player had a unique best response, then there is no guarantee that it would be the combination required to induce the other player to play his equilibrium strategy.

This lack of determinancy is, however, not robust to small perturbations of the choice function. Consider instead a more general power form:

\[
\frac{A_i^e(n)}{A_i^e(n) + A_2^e(n)}. \tag{2}'
\]

Erev and Roth [14], for example, consider this. Eq. (12) becomes

\[
f(x) = \frac{z_1^e x_n^e}{z_1^e x_n^e + z_2^e (1 - x_n)^e}. \tag{12}'
\]

When \( z_1 = z_2 \), this has fixed points only at \( x = 0, 1/2 \) and 1.
Similarly suppose that all actions are reinforced, even when not played. Erev and Roth [14] and Roth and Erev [37] consider this and it might be considered desirable to speed up convergence by preventing initial bad luck from leading the process to be stuck for a long while near the boundary. Suppose therefore that each action receives a reinforcement of \( \varepsilon \) when it is not played. Eq. (12) becomes

\[
 f(x_n) = \frac{\alpha_1 x_n + \varepsilon(1 - x)}{\varepsilon + \alpha_1 x_n + \alpha_2(1 - x)}, \tag{12}''
\]

when \( \alpha_1 = \alpha_2 \) this has a unique fixed point at \( x = 1/2 \) for all \( \varepsilon > 0 \). In the literature on urns this is referred to as the Friedman urn scheme (see for example [15] for a detailed discussion).

Small perturbations of the model therefore imply that unless the mixed equilibrium is exactly at \( x = 1/2 \), the ER scheme will not converge to it. It may rather converge to some approximate equilibrium close by. Hopkins [27] contains some results on local stability of perturbed equilibria in schemes similar to (12)''.

Perturbations to the ER scheme may well be advantageous in small samples if they prevent the process spending becoming stuck for too long near the boundary. As noted in previous samples, convergence of the ER scheme can be slow.

Acknowledgments

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

Appendix

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

\[
\frac{1}{A_1(n) + \pi_1(n + 1)} \leq \frac{1}{A_1(n)} - \frac{\pi_1(n + 1)}{A_1^2(n)} + c_1 \frac{\pi_1(n + 1)^2}{A_1^3(n)}. \tag{A.1}
\]

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

\[
(A_2(n) + \pi_2(n + 1))^\varepsilon \leq A_2(n)^\varepsilon + \varepsilon A_2(n)^{\varepsilon - 1} \pi_2(n + 1)
+ c_2 \varepsilon A_2(n)^{\varepsilon - 2} \pi_2(n + 1)^2. \tag{A.2}
\]

The expected increment \( A_2(n + 1)^\varepsilon / A_1(n + 1) - A_2(n)^\varepsilon / A_1(n) \), conditional on \( F_n \) is

\[
\frac{A_1(n)}{A_1(n) + A_2(n)} E \left( \frac{A_2(n)^\varepsilon}{A_1(n + \pi_1(n + 1))} - \frac{A_2(n)^\varepsilon}{A_1(n)} \right)
+ \frac{A_2(n)}{A_1(n) + A_2(n)} E \left( \frac{(A_2(n) + \pi_2(n + 1))^\varepsilon}{A_1(n)} - \frac{A_2(n)^\varepsilon}{A_1(n)} \right). \tag{A.3}
\]
Using (A.1) and (A.2), this is at most
\[
\frac{A_1(n)}{A_1(n) + A_2(n)} \left( -E \pi_1(n + 1) + c_1 \frac{E \pi_1(n + 1)^2}{A_1(n)} \right) + \frac{A_2(n)}{A_1(n) + A_2(n)} \left( E \pi_2(n + 1) + \frac{c_2 E \pi_2(n + 1)^2}{A_2(n)} \right) .
\]  
(A.4)

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
\[
\frac{A_2(n)^{\varepsilon}}{(A_1(n) + A_2(n))A_1(n)} \left( K_1(\varepsilon - \gamma) + \frac{K_2}{A_1(n)} + \frac{K_3}{A_2(n)} \right) .
\]  
(A.5)

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, \(A_2(n)^{\varepsilon}/A_1(n)\) 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 \(\frac{A_2(n)^{\varepsilon}}{A_1(n)}\) converges to a finite limit almost surely. Picking \(1 < \varepsilon < \gamma\) shows that \(A_2(n)/A_1(n)\) converges to zero almost surely. \(\square\)

**Proof of Corollary to Theorem 3.** The proof is by induction. Suppose that strategy 1 (say) of player 1 dominates strategy 2 if a certain dominated strategy of another player is deleted. By Theorem 2, the probability that the latter strategy is played almost surely converges to zero. Hence there is almost surely some \(N\), such that for \(n \geq N\),
\[
E(\pi_1(n) \mid \text{action 1 is chosen at stage } n) > \gamma E(\pi_2(n) \mid \text{action 2 is chosen at stage } n)
\]
for some constant \(\gamma\). Hence, applying the argument of the proof of Theorem 1, \(A_2(n)^{\varepsilon}/A_1(n)\) is almost surely a positive 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. \(\square\)

**Proof of Theorem 4.** Suppose that strategy \(i\) is dominated by the mixed strategy \(\sigma_1, \ldots, \sigma_m\). A Taylor expansion, as in the proof of Theorem 1, shows that \(A_i(n)^\varepsilon/A_1(n)^{\sigma_1} \ldots A_m(n)^{\sigma_m}\) is a positive supermartingale for some \(\varepsilon > 1\), 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 the Corollary to Theorem 3 proves the theorem for iteratively dominated strategies. \(\square\)

**Proof of Theorem 5.** A well-known result of Robbins and Siegmund states (see for example [12] Theorem 1.3.12):
Result RS. Let \( \{ \Omega, \mathcal{F}, P \} \) be a probability space and \( \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \) a sequence of sub-\( \sigma \)-algebras of \( \mathcal{F} \). For each \( n = 1, 2, \ldots \) let \( z_n, \beta_n, \xi_n \) and \( \zeta_n \) 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 \( \sum_n \xi_n < \infty \) on the event \( \{ \sum_n \beta_n < \infty, \sum_n \zeta_n < \infty \} \).

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

Result RS'. Let \( z'_n, \xi'_n \) and \( \zeta'_n \) be \( \mathcal{F}_n \)-measurable random variables, with \( z'_n \) bounded above by a non-random constant \( C \) and \( \xi'_n \) non-negative, such that

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

then \( \lim_{n \to \infty} z'_n \) exists and is finite and \( \sum_n \xi'_n < \infty \) on the event \( \{ \sum_n \zeta'_n \) 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 (4)

\[
\ln a_i(n + 1) = \ln a_i(n) \left( 1 + \frac{\bar{a}_i(n + 1) - a_i(n)}{(n + 1)a_i(n)} \right) \quad \text{(A.6)}
\]

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

\[
\ln a_i(n + 1) \geq \ln a_i(n) + \frac{\bar{a}_i(n + 1) - a_i(n)}{(n + 1)a_i(n)} - c \frac{(\bar{a}_i(n + 1) - a_i(n))^2}{(n + 1)^2 a_i(n)^2} \quad \text{(A.7)}
\]

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

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

\[
\frac{(\bar{a}_i(n + 1) - a_i(n))^2}{(n + 1)^2 a_i(n)^2} = \frac{1}{(n + 1)^2} \quad \text{(A.8)}
\]

If action \( i \) is played, then \( \pi_i(n + 1) - a_i(n) \) is bounded above, say by \( \bar{K} \), as payoffs and reinforcements are bounded. Also \( a_i(n) \geq (k_1N_i(n) + A_i(0))/n \), 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 $n + 1$,
\[
\frac{(\pi_i(n+1) - a_i(n))^2}{(n+1)^2 a_i(n)^2} \leq \tilde{K}^2 \frac{n^2}{(n+1)^2 (k_1 N_i(n) + A_i(0))^2}.
\] (A.9)

One therefore has
\[
0 \leq \frac{(\pi_i(n+1) - a_i(n))^2}{(n+1)^2 a_i(n)^2} \leq K_n, \tag{A.9}'
\]

where
\[
K_n = \begin{cases} 
\frac{1}{(n+1)^2} & \text{if } i \text{ is not played}, \\
D \frac{n^2}{(k_1 N_i(n) + A_i(0))^2} & \text{if } i \text{ is played}
\end{cases}
\] (A.10)

for some constant $D$. Now
\[
\sum_{\{n: i \text{ is played}\}} \frac{1}{(N_i(n) + A_i(0))^2} \tag{A.11}
\]

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 (A.11) 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 [19, Corollary 2.3, p. 32], that $\sum_n E(K_n|\mathcal{F}_{n-1})$ converges almost surely as well. Hence it follows from (A.6) and (A.9)' that one has
\[
E(\ln a_i(n+1)|\mathcal{F}_n) \geq \ln a_i(n) + E \left( \frac{\tilde{\pi}_i(n+1) - a_i(n)}{(n+1) a_i(n)} | \mathcal{F}_n \right) - H_n, \tag{A.12}
\]

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 $p_i(n+1) = a_i(n)/a(n)$,
\[
E(V_{n+1}|\mathcal{F}_n) \geq V_n + E \left( \frac{\pi_\sigma(n+1) - a(n)}{(n+1) a(n)} | \mathcal{F}_n \right) - \kappa_n, \tag{A.13}
\]

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

If $a(n) \leq v'$, $v' < v$,
\[
E(V_{n+1}|\mathcal{F}_n) \geq V_n + E \left( \frac{\pi_\sigma(n+1) - v'}{(n+1)v'} | \mathcal{F}_n \right) - \kappa_n. \tag{A.14}
\]

By Assumption M,
\[
E(V_{n+1}|\mathcal{F}_n) \geq V_n + \frac{v - v'}{(n+1)v'} - \tilde{\zeta}_n, \tag{A.15}
\]

where $\sum \tilde{\zeta}_n$ converges.

Hence on the event \{a(n) \leq v' \ \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 \(\lim \sup_{n \to \infty} a(n) \geq v\) almost surely. □

**Proof of Corollaries 1 and 2.** Corollaries 1 and 2 follow from the following more general result. Consider the following conditions:

**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 \(\lim \sup_{n \to \infty} a(n) \leq v\).

**Condition M′.** \(E(p_n|F_{n-1}) \geq v + v_n\) for all \(n\), where \(\sum n v_n / na(n - 1)\) converges.

Condition M′ is a variant on condition M, more suited to this result.

**Theorem A.1.** 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 same argument as in the proof of Theorem 5 yields (A.13) again. Assumption M′ implies that if one defines \(\tilde{a}(n) = \min\{a(n), v\}\), then

\[
E(V_{n+1}|F_n) \geq V_n + \frac{v - \tilde{a}(n)}{(n + 1)\tilde{a}(n)} - \xi_n, \tag{A.16}
\]

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

\[
E(V_{n+1}|F_n) \geq V_n + \frac{v - \tilde{a}(n)}{(n + 1)\tilde{a}(n)} - \xi'_n, \tag{A.17}
\]

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

It follows from R-S that \(\sum \xi_n\) converges. Now if \(\lim \inf_{n \to \infty} a(n) < v\), then \(\tilde{a}(n) < v' < v''\) infinitely often, some \(v', v'' < v\). 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 \(m_{2k+1} = \inf\{n : a(n) < v', \ n > m_{2k}\}\) and \(m_{2k} = \inf\{n : a(n) > v''\}, \ n > m_{2k-1}\), 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′.

To prove Corollaries 1 and 2, it is therefore enough to verify Conditions M′ and A.

Assumption M′ is immediate in the case of Corollary 1 and in the case of Corollary 2 follows from Proposition 9.2.9 of [12] 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:

\[
\pi(n) = \tilde{\pi}(n) + \varepsilon_n, \tag{A.18}
\]
where \( E(\xi_n|\mathcal{F}_{n-1}) = 0 \). In the case of Corollary 1, \( \tilde{\pi}_n \leq v \). In the case of Corollary 2, \( \pi^B(n) \leq v \), where \( \pi^B(n) \) is the expected payoff of the current strategy if the chain were distributed according to its invariant measure. One can therefore write

\[
\pi(n) \leq v + \psi_n + \varepsilon_n ,
\]

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

Now \( \varepsilon_n \) is a bounded martingale with mean zero, so by Theorem 1.3.17 of \cite{12}, \( \sum_{i=1}^{n} \varepsilon_i/n^{3/4} \) tends to zero. Also, by Proposition 9.2.9 of \cite{12}, \( \sum_{i=1}^{n} \psi_i n^{-3/4} \) 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 \( \chi_n = |\sum_{i=1}^{n} \psi_i|/n^{3/4} \) and \( v_i = |\sum_{i=1}^{n} \varepsilon_i|/n^{3/4} \) then \( \chi_n \) and \( v_n \) tend to zero and one has

\[
a(n) \leq v + \frac{A(0)}{n} + \frac{\chi_n}{n^{1/4}} + v_n/n^{1/4}.
\]

This implies condition A. \( \Box \)

### A.1. Analysis of counter example

Let \( \zeta \), 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 \( \tau_{2k} = \inf \{n | n > \tau_{2k-1} \} \), \( p_1(n) \geq 1 - \zeta \), and \( \tau_{2k+1} = \inf \{n | n > \tau_{2k} \} \), \( p_1(n) \leq \zeta \). Player 2 adopts the following strategy: play strategy 1 in periods \( \tau_{2k-1} \) to \( \tau_{2k} - 1 \) and play strategy 2 in periods \( \tau_{2k} \) to \( \tau_{2k+1} - 1 \), for all \( k \). Note that \( \tau_k \) are predictable stopping times and that by Theorem 1 each \( \tau_k \) is finite almost surely. To prove that under this strategy \( a(n) \leq 1.5 - \varepsilon \) infinitely often, for suitable choice of \( \zeta \) and \( \varepsilon \), it is enough to show that for \( k \geq K \), some \( K \), \( P(\inf \{a(n) | \tau_k \leq n < \tau_{k+1} \} \leq 1.5 - \varepsilon | \) events up to time \( \tau_k \) > \( \delta \), for some \( \delta > 0 \). (Apply the conditional Borel–Cantelli lemma.)

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

\[
a(n + 1) = a(n) + \frac{-a(n) + 1 + p_1(n)}{(n + 1)a(n)} + \frac{v_n}{n + 1} , \quad (A.21)
\]

\[
p_1(n + 2) = p_1(n + 1) + \frac{p_1(n + 1)(1 - p_1(n + 1))}{(n + 1)a(n)} + \frac{\eta_n}{n + 1} + \frac{K_n}{(n + 1)^2} , \quad (A.22)
\]

where \( E(v_n|\mathcal{F}_n) = E(\eta_n|\mathcal{F}_n) = 0 \) 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 \( m(l, T) = \inf \{ n : \sum_{i=1}^{n} \gamma_i \geq T \} \). Let \( \Theta(t, z_0) \) be the flow generated by the system of differential equations

\[
\dot{a} = -a + 1 + p_1, \quad (A.23)
\]

\[
\dot{p}_1 = \frac{p_1(1 - p_1)}{a}, \quad (A.24)
\]

with initial condition \( z_0 \). Let \( z_n(l) \) denote the pair of random variables generated by (A.21) and (A.22) with the same initial condition \( z_0 \) at time \( l \).

By a standard result (see [5, 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 \( l \geq L \),

\[
P \left( \sup_{l \leq n \leq m(l, T)} \|z_n(l) - \Theta(t_n, z_0)\| > \delta_1 \right) < \delta_2. \quad (A.25)
\]

That is, far enough along the sequence, the path of (A.21) and (A.22) is close to that of (A.23) and (A.24) in a finite interval with arbitrarily high probability.

Now in (A.23) and (A.24), given \( \varepsilon < 0.25 \) there is \( \zeta' \) such that if \( p(t) \leq \zeta' \) for all \( t \in [0, 2] \), \( a(2) \leq 1.5 - 2\varepsilon \) if \( a(0) \geq 1.5 - 2\varepsilon \). Also since \( a \) is bounded below, there is \( \zeta \), 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 \( \varepsilon \) and \( \zeta \). By choosing \( \delta_1 \) 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 \( a(l) \geq 1.5 - 2\varepsilon \) and \( p_1(l) \leq \zeta \) then \( a(m(l, 2)) \leq 1.5 - \varepsilon \) and \( p_1(n) < 1 - \zeta \) for all \( n \) with \( l \leq n \leq m(l, 2) \). Now \( \tau_k \) is finite almost surely and \( \tau_k \geq k \), so let \( 2k - 1 \geq L \).

Proof of Theorem 6. Theorem 6 follows from the following lemmas

**Lemma A.1.** \( V \) is bounded above on \( \Delta^* \) and \( V \rightarrow -\infty \) as \( z \rightarrow \partial \Delta^* \). Moreover \( \dot{V} \geq 0 \) in \( \Delta^* \).

**Proof.** All the statements in Lemma A.1 except the fact that \( \dot{V} \geq 0 \) are obvious. From (10a) and (10b), since \( p_i = a_i/a \) and \( q_j = b_j/b \) and payoffs sum to \( K \),

\[
\dot{V} = \frac{s_i}{a} \left( \sum_i p_i^* \pi_i^* \right) + \frac{s_{II}}{b} \left( \sum_j q_j^* \pi_j^* \right) - \frac{K}{r}. \quad (A.26)
\]

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

\[
\dot{V} \geq \frac{s_i^2 K}{a} + \frac{s_{II}^2 K}{b} - \frac{K}{r}. \quad (A.27)
\]
Now \( a + b = r \) and in the set defined by this constraint, the right-hand side of (A.27) is minimised when \( a = s_I r \) and \( b = s_{II} r \). Hence

\[
\dot{V} \geq \frac{s_I K}{r} + \frac{s_{II} K}{r} - \frac{K}{r} = 0. \tag{A.28}
\]

Thus \( \dot{V} \geq 0 \), as was to be shown. \( \square \)

Lemma A.1 states that \( V \) is a weak Lyapounov function. By a standard result on differential equations (see for example [18, p. 317]) it follows that:

**Lemma A.2.** For any initial condition in \( \Delta^* \) 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 \( \nabla V \) 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 A.3.** (a) \( \{ z | \dot{V}(z) = 0 \} \) is contained in the set \( a = s_I r, p^* Aq = v_I \) and \( p B q^* = v_{II} \).

(b) Any invariant set in \( \{ z | \dot{V}(z) = 0 \} \) has \( r = K \) and \( p A q = v_I \).

**Proof.** \( \dot{V} = 0 \) if and only if (A.27) and (A.28) hold with equality. This implies (a). From (10b) and (10d) 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 \( b = s_{II} r \), it follows that \( \dot{a} = 0 \) and \( \dot{b} = 0 \). From (10b) and (10d), and part (a) of the lemma, this can only be true if \( I^I = s_I K \) and \( II^I = s_{II} K \), which implies (b).

Theorem 6 follows easily from the above. \( \square \)

**Proof of Theorem 7.** Theorem 7 follows from the following lemma:

**Lemma A.4.** In the ER model, \( V \) converges to a finite random variable almost surely. \( z \) converges to the largest set in \( \langle \nabla V, h \rangle = 0 \) invariant under the flow generated by \( h \).

**Proof.** Let \( V_n = V(z_n) \). The same argument as in the proof of Theorem 5 applied to the components of \( V_n \) and use of the definition of \( h \) yields

\[
E(V_{n+1} | F_n) \geq V_n + \frac{1}{n+1} \langle \nabla V_n, h \rangle - \kappa_n, \tag{A.29}
\]

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) \( \sum_{n=1}^{\infty} \frac{1}{n} \langle \nabla V_n, h \rangle < \infty \). Since \( \langle \nabla V, h \rangle \) is continuous on \( \mathcal{A}^* \), (ii), Lemma 3.3.III.8 on p. 102 of [11] 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 [31, Theorem 2.1, p. 95]. □

Theorem 7 follows immediately. The fact that \( V \) converges to a finite random variable implies that the system cannot converge to the boundary. The result now follows from Lemma A.4. □

**Proof of Lemma 3.** From Lemma A.3 parts (a) and (b) and the fact that \( p^* \) and \( q^* \) yields an equilibrium of the game

\[
(p - p^*), A(q - q^*) = 0, \tag{A.30}
\]

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

Since the case of pure equilibria is covered in Theorem 8, 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 \( 2 \times 2 \), both must be fully-mixed. It follows that \( p^* \) and \( q^* \) are the unique solutions of \( Aq = v_I e \) and \( qA = v_{II} e' \), where \( e \) is a column vector of ones and \( ' \) denotes transpose. Hence \( A \) is non-singular.

Now in the \( 2 \times 2 \) case, it is easy to check that

\[
\zeta_0 A \eta_0 = 0,
\]

where \( \zeta_0 \) and \( \eta_0 \) are vectors each of whose components add to zero, cannot have a solution with both \( \zeta_0 \) and \( \eta_0 \) 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 \( (Aq)_1 > v_I \) or \( (Aq)_2 > v_I \). Suppose the former. It follows from (10a) and \( p^* Aq = v_I \), that \( p_1 > 0 \). Hence the system will not stay in this invariant set, which contradicts the assumption that it is invariant. □

**Proof of Theorem 9.** The only fact not proved elsewhere is the convergence of empirical 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 [4, 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 12.** This follows directly from [31, 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 [13, 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). □
References