Two Competing Models of How People Learn in Games

Two Competing Models of How People Learn in Games

Ed Hopkins*

Department of Economics	Department of Economics
University of Edinburgh	University of Pittsburgh
Edinburgh EH8 9JY, UK	Pittsburgh PA 15260, USA

October, 1999

Abstract

Reinforcement learning and stochastic fictitious play are apparent rivals as models of human learning. They embody quite different assumptions about the processing of information and optimisation. This paper compares their properties and finds that they are far more similar than were thought. In particular, exponential fictitious play and a suitably perturbed reinforcement model have the same expected motion and therefore will have the same asymptotic behaviour. It is also shown that more general models of stochastic fictitious play and perturbed reinforcement learning have identical local stability properties. The main identifiable difference between the two models is speed: stochastic fictitious play gives rise to faster learning.

Journal of Economic Literature classification numbers: C72, D83.

Keywords: Games, Reinforcement Learning, Fictitious Play.

*This paper arose out of extensive discussions with Tilman Börgers. I have also benefitted from comments from Josef Hofbauer, John Duffy and Peyton Young. Errors remain my own. E.Hopkins@ed.ac.uk, http://www.ed.ac.uk/econ/

1 Introduction

What is the best way to model how people play games? In a recent paper, Erev and Roth (1998) analyse experiments where subjects repeatedly play 2×2 games with a unique mixed strategy equilibrium. Their claim is that the data is best described by a model of reinforcement learning, which places very low demands on the rationality of subjects. In contrast, theorists have been more interested in investigating the properties of fictitious play and its variants, which involve agents optimising given their beliefs. Equally some, for example Fudenberg and Levine (1998), have been skeptical about reinforcement learning because it implies that agents ignore potentially useful information. The agents it models are simply too naïve. Camerer and Ho (1999) propose a family of models which contains both approaches as special cases and find that experimental data falls somewhere between the two extremes.[1]

This paper takes a different approach. While remaining agnostic as to which model or models best describe actual human learning behaviour, the theoretical properties of the two different models of learning in games are compared. Existing results are mostly on fictitious play and focus on 2×2 games. This paper looks at reinforcement learning, both the basic model and several extensions, and compares the results obtained with those for stochastic fictitious play. The models differ on two levels, what information agents use and whether agents optimise given that information. Nonetheless, their behaviour both in terms of transition dynamics and asymptotic properties are remarkably similar. It is possible to take the result of Camerer and Ho (1999) a stage further: there is more than just a family resemblance between the two models. As will be seen, the expected motion of both models can be considered as noisy versions of the evolutionary replicator dynamics. This has the result that by the choice of an appropriate noise function, one can construct a pure reinforcement model, with no optimisation and which throws information away, which has identical asymptotic properties to the most commonly used form of stochastic fictitious play.

The second important result of this paper is on local stability of these learning schemes. If any equilibrium is locally stable (unstable) for stochastic fictitious play, it is locally stable (unstable) for perturbed reinforcement learning. In contrast, global results on convergence are only possible for normal form games with a strong structure, of cooperation or competition or dominance solvability. But for these games, convergence results can be obtained for both models. A general principle might be that if a game has sufficient structure that we can be sure of the asymptotic behaviour of fictitious play, then reinforcement learning will exhibit the same behaviour. Though as we will see, small differences in the level of optimisation can lead to significant differences in the speed of convergence. In contrast, the effect of the differing use of information seems negligible.

[1] More recently still, a new wave of research has in various ways disputed the interpretation of the experimental data and indeed the possibility of distinguishing empirically between the two models. See Sarin and Vahid (1998), Feltovich (1999), Salmon (1999).

This paper is also able to resolve some apparently contradictory claims. For example, Erev and Roth (1998) argue that the experimental data on play in 2×2 games does not seem to support traditional equilibrium analysis. Simply put, the experimental subjects often did not play the unique Nash equilibrium even in these simple games. In the short run, the dynamics may seem to move away from equilibrium. Even after a significant length of time, play is not near Nash equilibrium. In seeming complete contrast, Fudenberg and Kreps (1993), Benaïm and Hirsch (1996) have shown that stochastic fictitious play converges to equilibrium for these games.

Theorists like to emphasise convergence and to think of the possibility of agents making mistakes, and the noise this generates, as vanishingly small. However, this is to ignore the data from experiments which suggest that noise is both significant and persistent. Once noise is added to learning, whether reinforcement learning or fictitious play, the steady state of the learning process is a perturbed equilibrium which is typically some distance from Nash equilibrium. Erev and Roth do not calculate the steady states of their model of reinforcement learning. This paper shows that the equilibria of a perturbed reinforcement learning model can be identical to those of stochastic fictitious play, and that, in 2×2 games with a unique equilibrium, both models converge to this equilibrium. Once one realises that the equilibria of these stochastic learning models are not identical to Nash, these convergence results are not in conflict with the experimental data either.

Reinforcement or stimulus-response learning is a very simple application of the principle that actions that have led to good outcomes in the past are more likely to be repeated in the future. Agents have a probability distribution over possible actions. When an action is chosen, the probability of that action being taken again rises in proportion to the realised payoff. The action has been "reinforced". Note the very low level of information or processing ability necessary to implement such an algorithm. In the context of game-playing, an agent does not need to know the structure of the game, to calculate best responses or even to know that a game is being played.

This is in contrast to other learning models considered in the literature. Fictitious play and its variants have been perhaps the most popular. Though they involve boundedly-rational behaviour, in particular, stochastic or smooth fictitious play allows for mistakes, they also do involve a degree of optimisation. This emphasises the boldness of Erev and Roth's claims. The experimental data they examine was generated under a wide variety of conditions. In one experiment, subjects were unpaid and not even informed they were playing a game. In others, there were both monetary incentives and much more complete information. In the former experiment, one might think that reinforcement learning would explain the data well, after all, the subjects must be more or less groping in the dark. However, Erev and Roth claim that the same model does well when far more information is available and therefore much more sophisticated behaviour is possible.

There are in fact two fundamental issues which divide the differing models considered

here. First, information: are agents sophisticated enough to work out what payoffs they might have received if they had played some strategy other than they actually did? This is what Camerer and Ho (1999) call "hypothetical reinforcement" and it is present in classical fictitious play. In its absence there will be what Erev and Roth (1998) call "force of habit". That is, agents will be biased to repeat actions they have taken in the past. Second, optimisation: do agents maximise given their beliefs, like in fictitious play or use a more probabilistic rule like reinforcement learning?

Do these differing assumptions mean that the two models give different predictions? If attention is confined to the games that Erev and Roth consider, 2×2 games with a unique mixed strategy equilibrium, this appears to be the case. Both Fudenberg and Kreps (1993) and Benaïm and Hirsch (1996) show convergence of stochastic fictitious play to equilibrium. In contrast, it has been known for some time that the expected motion of reinforcement learning is given by the evolutionary replicator dynamics, see for example, Börgers and Sarin (1997), Posch (1997), Hopkins (1999a). The replicator dynamics do not converge to equilibrium in this class of games, a result that Posch (1997) uses to show that the basic model of reinforcement learning typically will not do so either.

As noted, data from experiments seems to suggest that human learning is noisy, in that adding a noise parameter significantly improves the fit of both fictitious play and reinforcement learning. There are several possible functional forms for perturbations. Here, I show that with a suitable choice of noise function, the asymptotic behaviour of stochastic fictitious play and reinforcement learning is identical. That is, a process without optimisation and without hypothetical reinforcement will have exactly the same outcomes as stochastic fictitious play which has both. It is possible to show global convergence for both stochastic fictitious play and perturbed reinforcement learning for games which are dominance solvable, for rescaled zero sum games and for rescaled partnership games (the idea of rescaling is due to Hofbauer and Sigmund, 1998). In doing so, this paper builds on recent work on the application of the theory of stochastic approximation to reinforcement learning by Posch (1997), and the analysis of stochastic fictitious play by Benaïm and Hirsch (1996) and Hofbauer and Hopkins (1999). The sum total of games included in these three classes could be seen as small relative to the set of all games. However, included are many games of economic interest and, importantly, most games that have been subject to experimental investigation. For example, all the games studied by Erev and Roth (1998) and most of those studied by Camerer and Ho (1999) fall in these categories.

This paper also looks at two possible extensions to stochastic learning models. Both Erev and Roth and Camerer and Ho find that experimental data suggests that subjects discount experience, giving more weight to recent events, rather than the steady accumulation found in fictitious play. In this case, there is no possibility of learning converging to a single point. It is nonetheless possible to demonstrate the existence of an invariant distribution which, in some cases, can be shown to be clustered around a perturbed equilibrium. Last, it has been suggested that the introduction of an aspiration level or reference point may improve

the performance of reinforcement learning. Here, our analysis shows that it has rather less effect than might be thought.

This paper is structured in the following way. Section 2 introduces and compares reinforcement learning and fictitious play. Section 3 outlines the basics of stochastic approximation theory. Section 4 shows how the expected motion of stochastic fictitious play can be expressed as a form of replicator dynamic. Section 5 analyses the impact of noise on reinforcement learning and proves some convergence results for 2×2 games. Section 6 gives the main results on equivalence between the two models. Section 7 gives a closer look at the speed of learning and the role of hypothetical reinforcement. Section 8 demonstrates how a weaker form of stochastic stability can still apply in the case where agents discount their experience. Section 9 examines endogenous aspirations. Section 10 looks to see how far a simplifying assumption, which I call rebalancing, can be relaxed. Section 11 concludes.

2 Two Competing Models of Learning

This paper examines learning in the context of 2-person normal form games. This section introduces two rival models, reinforcement learning and fictitious play. There are two agents A and B who play a game repeatedly at discrete time intervals, indexed by n. The first player, A, has N strategies, B has M. In period n, A's mixed strategy is written $x_n \in S_N$, and the strategy of B, $y_n \in S_M$ where S_N is the simplex $\{x = (x_1, ..., x_N) \in I\!R^N : \Sigma x_i = 1, x_i \geq 0$, for $i = 1, ..., N\}$. Let A be the $N \times M$ payoff matrix for the first player, with typical element a_{ij}, and B be the $M \times N$ payoff matrix for the second player with typical element b_{ji}. Expected payoffs for A will be $x \cdot Ay$, and for B, $y \cdot Bx$.

Each player has a propensity for each of his strategies which determine the probability of the choices he makes. If A has N strategies, her propensities in period n are given by a vector $q_n^A = (q_{1n}^A, ..., q_{Nn}^A)$. Then the state of the system can be written $q_n = (q_n^A, q_n^B)$. Fictitious play and reinforcement learning differ both in terms of how these propensities determine choice probabilities and how these propensities are updated as a result of realised play.

First, consider what Roth and Erev (1995), Erev and Roth (1998) call the basic stimulus response or reinforcement learning model. The easiest way of describing how the algorithm works is to imagine that the agent in period n has an urn containing a total of Q_n balls of N different colours with q_{in} balls of the ith colour. Each period the agent draws one ball at random (with replacement) from the urn and takes the action corresponding to the colour drawn. The probability of the player A, respectively B, taking his ith action in period n is therefore,

$$x_{in} = \frac{q_{in}^A}{\sum_{j=1}^N q_{jn}^A} = \frac{q_{in}^A}{Q_n^A},\ y_{in} = \frac{q_{in}^B}{\sum_{j=1}^M q_{jn}^B} = \frac{q_{in}^B}{Q_n^B}. \tag{1}$$

Strategies with higher propensities are played with higher probability. It is always assumed

that all initial propensities are strictly positive, so at all times, there will be a positive probability of a strategy being picked. In contrast the choice rule for fictitious play, for player one for example, is given by

$$x_{in} = 1 \text{ if } q^A_{in} = \max q^A_n, \text{ else } x_{in} = 0. \tag{2}$$

That is, the strategy that appears the best is played with probability one.[2]

The models also differ in the method of updating. For the basic reinforcement model, one can imagine after each period of play the agent adding a number of balls equal to the payoff received of the colour corresponding to the action taken to the urn. That is, for example, if player A takes action i, and player B chooses j in period n, describe this as the event ij. The function f_{ij} is known as the event operator associated with the event ij. The ith propensity is updated thus,

$$q^A_{in+1} = f_{ij}(q^A_{in}) = q^A_{in} + a_{ij} \tag{3}$$

but all other propensities remain unchanged. Since the other actions were not chosen, the payoff they would have earned is not observed. However, if the opponent's choice of action is observed, and one's own payoff matrix is known, a player could calculate what he would have received had he chosen some other action. This is what is normally assumed to happen in fictitious play. For example, if player A were to observe that B is playing her jth strategy, under fictitious play all propensities would be updated thus,

$$q^A_{in+1} = f_{ij}(q^A_{in}) = q^A_{in} + a_{ij} \text{ for } i = 1, ..., N \tag{4}$$

In contrast, standard reinforcement learning assumes that such hypothetical reinforcement does not take place even when the necessary information is available.

Fictitious play is often presented slightly differently with players keeping track of the frequency of opponents' choices. Let $u_n \in S_N$ be the vector of relative frequencies of the actions of the first player up to period n. That is, if after 100 rounds of play A has played the first of two strategies 30 times, then $u_n = (0.3, 0.7)$. Let $v_n \in S_M$ be the vector of the relative frequencies of the choices of the second player. However, given (4), for player A it is true that $q^A_{in} = q^A_{i0} + n(Av_n)_i$ (However, this relationship does not hold if the reinforcement learning updating rule (3) is used). That is, both q^A_{in}/n and $(Av_n)_i$ give the average past return to strategy i. Note that the changes to the propensities of A are independent of the choices made by A. In contrast, the reinforcement learning rule (3) exhibits what Erev and Roth call "force of habit". The actions which are chosen more frequently are reinforced more frequently.

Force of habit is important in what it indicates about agents' processing of information. When information is available about choices of opponents, the standard reinforcement updating rule (3) throws this information away. The evidence from experiments as to what

[2]Implicit here are additional assumptions to ensure that there are no ties for first place. I don't detail them as ties are not an issue in stochastic fictitious play which is the main object of interest and which is described later on in this section.

people actually do is mixed. Erev and Roth's (1998) detection of force of habit in the learning behaviour exhibited in their data is matched by Van Huyck et al. (1997) who find that force of habit is statistically insignificant in data from their experiments. Camerer and Ho (1999) claim the data supports an intermediate case. Furthermore, the differences between the two learning models in their treatment of information is in practice blurred. In fact, both updating rules have been used in the reinforcement learning literature (see Vriend, 1997) and, as Fudenberg and Levine (1998, Ch4) point out, when opponents' actions are unobservable, fictitious players will have to use something like (3).

The basic reinforcement model, of course, defines a Markov process on $\mathbb{R}^N \times \mathbb{R}^M$, the state variable being the vector of propensities q_n. Note that if realised payoffs are negative, then this may lead to one of the propensities becoming negative and the probabilities x, y will no longer be defined. To ensure the problem does not arise we assume the support of payoffs is strictly positive. This has a consequence that for each player Q_n, the sum of the propensities, is strictly increasing with order n.

The real interest is in the evolution of x_n, that is, the players' mixed strategies. Now, if in period n, player A chooses action i, and receives a payoff of π_n^A,

$$x_{in+1} - x_{in} = \frac{(1 - x_{in})\pi_n^A}{Q_n^A + \pi_n^A}, \tag{5}$$

but if some other action j was chosen

$$x_{in+1} - x_{in} = \frac{-x_{in}\pi_n^A}{Q_n^A + \pi_n^A}. \tag{6}$$

Note that the rate of change of x_i is decreasing in Q_n^A and that Q_n^A increases each period by a stochastic increment equal to the realised payoff. We refer to $1/Q_n^k$ as the step size of player k's learning process.

In general, summing (5) over the N possible actions, using the fact that each action is taken with the associated probability x_i, and performing a similar operation for B, we can write the learning process for the two players as

$$\begin{aligned} x_{in+1} - x_{in} &= \tfrac{1}{Q_{n+1}^A} x_{in}[(Ay_n)_i - x_n \cdot Ay_n] + \tfrac{1}{Q_{n+1}^A}\eta_i^A(x_n, y_n) + \tfrac{1}{(Q_{n+1}^A)^2}\epsilon_i^A \\ y_{jn+1} - y_{jn} &= \tfrac{1}{Q_{n+1}^B} y_{jn}[(Bx_n)_j - y_n \cdot Bx_n] + \tfrac{1}{Q_{n+1}^B}\eta_j^B(x_n, y_n) + \tfrac{1}{(Q_{n+1}^B)^2}\epsilon_j^B \end{aligned}, \tag{7}$$

for $i = 1, ..., N$ and $j = 1, ..., M$. There are three terms on the righthand side of these equations. The first is a deterministic driving term which gives the expected change in each x_i, correct to order of $1/Q_n^2$. Any errors are subsumed into the last term ϵ. However accurate an estimate of the expected motion, it must not be forgotten that the actual change in x_n will depend on the realisation both of the agent's choice and of the subsequent payoff. This is captured by the stochastic term η which has expectation zero.

Note that the expected motion of x is driven by the difference between the expected return to the ith action and the expected return of the agent's current probability distribution over all actions. This is, of course, the same principle that lies behind the evolutionary replicator dynamics, which have been analysed extensively (Hofbauer and Sigmund, 1998; Weibull, 1995). However, the process (7) is in discrete time and stochastic, whilst the replicator dynamics are deterministic and are (typically) analysed in continuous time. The link is the theory of stochastic approximation, which will become apparent later in this paper.

Stochastic or smooth fictitious play is where the standard fictitious play updating rule is used but the deterministic choice of a best response is replaced by a stochastic choice rule. Remember that the fictitious play rule picks out the strategy with the highest propensity, which is equivalent to choosing the strategy with the highest historical payoff or choosing x to maximise $x \cdot Av_n$, where Av_n is the vector which describes A's historical payoffs given B's past choices v_n. Imagine this maximisation problem was subject to noise, so that instead A chose x to maximise

$$x \cdot Av_n + \lambda\phi(x),$$

where λ is a scaling factor for the perturbation $\phi(x)$.

1. $\phi(x)$ is strictly concave and ϕ'', the matrix of second derivatives of ϕ with respect to x, is negative definite.
2. $\lim_{x_i \to 0} \frac{\partial \phi}{\partial x_i} = \infty$ for all x_i.

Then it is certain there exists a solution to the following first order conditions for a maximum,

$$x = (\phi')^{-1}\left(-\frac{1}{\lambda}Av_n\right) = \overline{BR}(v_n). \tag{8}$$

$\overline{BR}$ is thus a perturbed best response function, with λ as a noise parameter. As it drops to zero, the above rule approaches the standard fictitious play rule (2) and will pick out the strategy with the highest expected return with probability one. However, high values of λ, that is, lots of noise, will mean the probability of a best response will be much decreased.

One example of this is where the deterministic choice of strategy for player A is replaced by,

$$x_{in} = \frac{\exp \beta(Av_n)_i}{\sum_{j=1}^{N} \exp \beta(Av_n)_j} = \overline{BR}_i^e(v_n), \tag{9}$$

where the "e" superscript is for exponential, and I have written $\beta = 1/\lambda$ for concision. The other particular functional form that has been popular in the literature is

$$x_{in} = \frac{(Av_n)_i^\beta}{\sum_{j=1}^{N} (Av_n)_j^\beta} = \overline{BR}_i^p(v_n), \tag{10}$$

where the "p" superscript is for power. If $\beta = 1$, then this rule is similar to the reinforcement learning choice rule. It is possible to fit this form to experimental data and use estimates of β to test between reinforcement learning and stochastic fictitious play.[3]

The above equation (8) combined with the updating rule (4) define a stochastic learning process. Rather than looking at the evolution of the choice probabilities (x_n, y_n), it is more usual with the historical frequencies of choices. More specifically, some calculation reveals,

$$E[u_n|u_{n-1}, v_{n-1}] - u_{n-1} = \frac{\overline{BR}(v_n) - u_n}{n},\ E[v_n|u_{n-1}, v_{n-1}] - v_{n-1} = \frac{\overline{BR}(u_n) - v_n}{n}. \quad (11)$$

There is no apparent or obvious connection between this process and that for reinforcement learning considered above. However, the as will be seen, the connection, though hidden, is very strong. First, some more theory is needed.

3 Stochastic Approximation

The textbook exposition, for example, Benveniste et al. (1990), of the theory of stochastic approximation assumes a discrete time stochastic process of the form

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) + \gamma_n^2 \epsilon(\theta_{n-1}, X_n). \quad (12)$$

The evolution of the parameter vector θ is determined by H and ϵ, an error term. $(\gamma_n)_{n\geq 0}$ is a sequence of "small" scalar gains, with $\gamma_n \geq 0$, and $\sum_n \gamma_n = \infty$. X_n is a sequence of random vectors. The following is referred to the mean or averaged ordinary differential equation (ODE) associated with (12),

$$\dot{\theta} = h(\theta), \quad (13)$$

where

$$h(\theta) = \lim_{n\to\infty} E[H(\theta, X_n)]$$

This is important in that recent results in the theory of stochastic approximation have shown the behaviour of this ODE (13) and of the stochastic process (12) are very closely linked. And indeed in this paper, the results obtained on the learning process will largely be obtained by analysis of the appropriate averaged ODE. We write the solution of the ODE $\theta(t)$, and note that whereas discrete time is indexed by n, the variable t refers to continuous time.

The step size, that is the magnitude of the change in θ, of the stochastic process is given by γ. If

$$\sum_n \gamma_n^\alpha < \infty \text{ for some } \alpha > 1$$

[3]There is the problem, however, that this rule could not arise as the result of the maximisation of a perturbed payoff function as considered here.

then we describe the algorithm as having *decreasing gain*. One obvious example of this is where $\gamma_n = 1/n$, and this is the most commonly analysed case.

Now, the astute reader will have noticed that if we set $\theta_n = (x_n, y_n)$ our equation (7) seems to fit the pattern (12) with, in particular, $h_i(x, y) = x_i((Ay_n)_i - x_n \cdot Ay_n)$ and $H = h(x_n, y_n) + \eta(x_n, y_n)$. However, the case here differs from the textbook model in two respects. Firstly, the step size is endogenous, being determined by the accumulation of payoffs. Second, it is not scalar. There are two step sizes, $1/Q^A$ and $1/Q^B$, one for each player.

To obtain the associated ODE for the reinforcement model, set θ equal to q. Then X_n is an indicator function giving the outcome (out of the $N \times M$ possible) of the two players' randomisations in period n. This implies the averaging that defines the ODE is over the possible payoff realisations so that

$$\lim_{n\to\infty} E[H_i(q, X_n)] = \frac{1}{Q^A} x_i[(Ay)_i - x \cdot Ay] + O\left(\frac{1}{(Q^A)^2}\right), \tag{14}$$

$$\lim_{n\to\infty} E[H_j(q, X_n)] = \frac{1}{Q^B} y_j[(Bx)_j - y \cdot Bx] + O\left(\frac{1}{(Q^B)^2}\right), \tag{15}$$

for $i = 1, ..., N$ and $j = 1, ..., M$.

If the step size is defined as $\gamma_n = 1/Q_n^A$, then he relative step size of the change in B's strategies will be given by a factor $\mu_n = Q_n^A/Q_n^B$. Note that as long as all payoffs are bounded and strictly positive, μ is also bounded and strictly positive even as n goes to infinity. Thus the associated ODE's will be the (modified) evolutionary replicator dynamics,

$$\dot{x}_i = x_i[(Ay)_i - x \cdot Ay], \ \dot{y}_j = \mu y_j[(Bx)_j - y \cdot Bx], \tag{16}$$

for $i = 1, ..., N$ and $j = 1, ..., M$. They differ from the standard replicator dynamics in that the equations for the second player are multiplied by the factor μ.

There are two possible approaches to take here. The first, taken by Posch (1997), is to make additional assumptions to bring reinforcement learning in line with standard stochastic approximation. The second is to persevere with the original model. This paper contains elements of both approaches. To be clear, the different assumptions are:

Assumptions A

1. Rebalancing. At each period, for each player k after propensities are updated by the addition of payoffs according to (3) every propensity is multiplied by an appropriate factor so that $Q_n^k = Q_0^k + n$, but leaving x_n, y_n unchanged. Hence, $\mu_n = \mu = 1$.

2. No rebalancing. For each player k, only the original updating rule (3) is used, so that $Q_n^k = Q_{n-1}^k + \pi_{n-1}^k$. The parameter μ is determined endogenously.

The advantage of the first assumption is that it allows easy application of existing results in the theory of stochastic approximation. It also captures much of the original model in that the sum of the propensities in both cases is of order n. This will have the effect that rate of change of x and y will slow over time. However, one cannot be sure that rebalancing does not alter the model significantly without analysing the model in its absence. Furthermore, it is possible to obtain results on Erev and Roth's actual model, that is, under Assumption A2, even though it does involve some extra work. However, I leave analysis of this more complex system to Section 10.

In the case of stochastic fictitious play, it is similarly possible to examine the discrete time stochastic process (11) in terms of an associated ODE. Notice that the step size is exactly $1/n$ for both players and so no correction is necessary. The associated ODE's are

$$\dot{u} = \overline{BR}(v) - u, \ \dot{v} = \overline{BR}(u) - v, \tag{17}$$

which I will refer to as the perturbed best response dynamics. Hopkins (1999b) and Hofbauer and Hopkins (1999) investigate the properties of these ODE's. Some of the results obtained are given in Section 6 below, where they are compared with the replicator dynamics arising from the reinforcement learning model.

Under some simple conditions the limit point of the discrete time stochastic process will be a limit point of the associated ODE. This implies that if the global behaviour of the ODE is known, this is sufficient to show convergence of the discrete time stochastic process. This need not be to a single point. The equilibria of 2×2 games with a unique mixed equilibrium are neutrally stable under the replicator dynamics (equivalent to (16) with μ fixed to unity). That is, the equilibrium point is surrounded by a continuum of closed orbits. Posch (1997) uses this fact to show that reinforcement learning may converge to one of these closed orbits and not to the mixed equilibrium. In Section 6, a more general result of this nature is presented. It uses the following proposition.

Proposition 1 *The limit set of a discrete time stochastic process with decreasing gain defined recursively in the manner (12), with probability one is a compact, connected attractor-free invariant set for the flow induced by the ODE (13).*

Proof: This follows from Theorems 1.2 and 2.1 of Benaïm (1996) (but see also Benaïm and Hirsch (1996) Theorem 3.3). ∎

The other type of result available from stochastic approximation theory is a negative one for points which are unstable equilibria of the associated ODE.

Proposition 2 *Let θ^* be a linearly unstable equilibrium point on the interior of $S_N \times S_M$ for the ODE associated either with the stochastic fictitious play process with decreasing gain.*

If the solution to this stochastic process is θ_n, then

$$\Pr\{\lim_{n\to\infty} \theta_n = \theta^*\} = 0.$$

(b) Similarly, the reinforcement learning process with decreasing gain will converge with probability zero to an equilibrium point on the interior of $S_N \times S_M$ which is linearly unstable with respect to the associated ODE.

Proof: Part (a) is Theorem 4.5 of Benaïm and Hirsch (1996). Part (b) follows in a straightforward manner. ∎

4 Best Response Dynamics as Noisy Replicator Dynamics

In this section, I show how in fact the two different models give rise to almost identical associated ODE's. In previous research, fictitious play has been analysed in terms of ODE's in historic frequencies, the perturbed best response dynamics (17) above. The key is to make the switch to looking at marginal frequencies. For example, it is possible to obtain the expected change in the probability A places on her kth strategy by summing over all ij possible events,

$$E[x_{kn+1}|x_n, y_n] - x_{kn} = \sum_{i=1}^{N} x_{in} \sum_{i=1}^{M} y_{jn} \left(\overline{BR}_k \left(f_{ij}(Av_n) \right) - \overline{BR}_k(Av_n) \right). \tag{18}$$

It will be useful to apply the following approximation:

$$E[x_{kn+1}|x_n, y_n] - x_{kn} = \sum_{i=1}^{N} x_{in} \sum_{j=1}^{M} y_{jn} \left(\frac{\partial \overline{BR}_k}{\partial Av_n} \cdot \left(f_{ij}(Av_n) - Av_n \right) \right) + O(\frac{1}{n^2}).$$

Note that from (11), $E[v_{n+1}|u_n, v_n] = (y_n - v_n)/(n+1)$ and that in the specific case of the exponential rule, from (9), $\partial \overline{BR}_i^e / \partial (Av_n)_i = \beta x_{in}(1 - x_{in})$ and $\partial \overline{BR}_i^e / \partial (Av_n)_j = -\beta x_{in} x_{jn}$. Using this, one can obtain

$$E[x_{kn+1}|x_n, y_n] - x_{kn} = \frac{\beta}{n+1} x_{kn}((A(y_n - v_n))_k - x_n \cdot A(y_n - v_n)) + O(\frac{1}{n^2}). \tag{19}$$

Thus, the expected motion of this process is very similar to the replicator dynamics produced by the "dumb" reinforcement model without optimisation. There are two differences. First, the replicator dynamic is not simply in terms of the instantaneous payoffs Ay_n but rather the instantaneous payoff with the historic payoffs Av_n subtracted. It is as though agents using this rule employed average past payoffs as an aspiration level or reference point. Second, the

version here is multiplied by the factor β. That is, the closer to optimisation agents get, the faster they learn.

The next step is to find an associated ODE entirely in terms of current choice probabilities. Note that from the first order conditions (8), one can substitute $-Av_n = \lambda\phi'(x)$. In the special case of the exponential choice rule where $\phi(x) = -\sum x_i \log x_i$, and $\phi' = (-\log x_1 - 1, ..., -\log x_N - 1)$. Therefore, the ODE for the first player can be written

$$\dot{x}_i = \beta\left(x_i((Ay)_i - x \cdot Ay) + \lambda x_i(-\log x_i + x \cdot \log x)\right), \tag{20}$$

for $i = 1, ..., N$. The first term in the above is exactly the replicator dynamics. The second term can be interpreted as a noise term, in that as λ goes to zero, the replicator term dominates. Furthermore, the ODE's associated with reinforcement learning are the replicator dynamics, this suggests that a perturbed version of reinforcement learning could generate this ODE. It will be seen that this is the case.

The following proposition summarises the material in this section in a somewhat more formal manner and includes other perturbed best response function besides the exponential form.

Proposition 3 *On the interior of $S_N \times S_M$ there is an ODE associated with the stochastic fictitious play process (11) which can be written*

$$\dot{x} = \beta Q(x)(Ay + \lambda\phi'(x)),\ \dot{y} = \beta Q(y)(Bx + \lambda\phi'(y)) \tag{21}$$

where $Q(.) = -(\phi'')^{-1}(.)$ is a symmetric positive semi-definite matrix function.

Q is a "PDA dynamic operator" in the terminology of Hopkins (1999b). Note that the replicator dynamics (16) can be written in vector form as $\dot{x} = R(x)Ay, \dot{y} = \mu R(y)Bx$, where $R(.)$ is one particular PDA operator which we can therefore refer to as the replicator operator. For the exponential choice rule where $\phi(x) = -\sum x_i \log x_i$, then $-(\phi'')^{-1}(x) = R(x)$. PDA dynamics generalise the replicator dynamics by replacing the replicator operator $R(.)$ by an arbitrary PDA operator $Q(.)$ but while retaining many of their properties (Hopkins, 1999a, b). So, it can be seen that the general version of the ODE associated with stochastic fictitious play (21), just as in the special case (20), is composed of a replicator-like term and a noise term. In both cases, the ODE is multiplied by the factor β which increases the speed of learning.

This result suggests that stochastic fictitious play and reinforcement learning will generate learning paths which are qualitatively similar. Both Erev and Roth (1998) and Camerer and Ho (1999) have used experimental data to test between stochastic fictitious play and reinforcement learning. What the above analysis suggests is that the only way that learning behaviour generated by the two models differ is in speed of passage along similar paths. It is only this difference in speed that the tests are picking up.

5 Noisy Reinforcement Learning

We have seen how in fact that with the introduction of noise, the expected motion of fictitious play becomes a form of noisy replicator dynamic. The introduction of noise to reinforcement learning, as we will now see, has a similar result. Erev and Roth (1998) introduce what they call experimentation to the basic reinforcement learning model by assuming there is some reinforcement for all propensities, not just for the one corresponding to the action taken. That is, when player k takes action i for some small $\lambda > 0$,

$$\begin{array}{l} q^k_{in+1} = q^k_{in} + (1-\lambda)\pi^k_n \\ q^k_{jn+1} = q^k_{jn} + \frac{1}{N-1}\lambda\pi^k_n \qquad \text{for all } j \neq i. \end{array} \tag{22}$$

This specification might capture the idea of experimentation in that even though a strategy is currently performing poorly, it still receives some reinforcement and will not be entirely forgotten. In any case, if all payoffs are positive, this noise and/or experimentation will prevent the probability of taking any action, even if dominated, falling to zero.

Given that our knowledge as to which learning model best describes human behaviour is limited, the question as to what form noise should take is even more murky. The specification chosen by Erev and Roth above is difficult to analyse for games larger than 2×2 because as the noise depends upon the payoff earned it depends on the actions of both players. The following simpler alternative to (22) is rather more tractable and no less plausible:

$$\begin{array}{l} q^k_{in+1} = q^k_{in} + \pi^k_n + \lambda \\ q^k_{jn+1} = q^k_{jn} + \lambda \qquad \text{for all } j \neq i. \end{array} \tag{23}$$

That is, all propensities are reinforced by a small amount.

Finally, as will be seen, much the same effect can be obtained from the following formulation, when player k takes action i for some small $\lambda > 0$,

$$q^k_{in+1} = q^k_{in} + \pi^k_n - \lambda \log\left(q^k_{in}/Q^k_n\right) \tag{24}$$

while the other propensities remain unchanged. Under this rule, if the probability of taking an action is low, then that action is strongly reinforced when taken. This specification is more in the spirit of the perturbation of payoffs found in the model of stochastic fictitious play. And as will be seen, it obtains an almost identical result.

The effect of the different specifications is easiest to see if we look at the effect on the rate of change of x. In fact, the associated ODE's will be

$$\begin{array}{rcl} \dot{x}_i &=& x_i[(Ay)_i - x \cdot Ay] + \lambda g^A_i(x,y) \\ \dot{y}_j &=& \mu\left(y_j[(Bx)_j - y \cdot Bx] + \lambda g^B_j(x,y)\right) \end{array} \tag{25}$$

The expected motion of the stochastic process is no longer given simply by the replicator dynamics. Each equation now has an additional noise term depending on λ. We will refer

to the above equations as the *perturbed replicator dynamics*. Again under Assumption A1, the factor μ is fixed at 1.

Given (22), it can be calculated that,

$$g_i^A(x,y) = \frac{1}{N-1}(x \cdot Ay - Nx_i(Ay)_i)\,,\; g_j^B(x,y) = \frac{1}{M-1}(y \cdot Bx - My_j(Bx)_j) \tag{26}$$

Thus, for example, if $x_i = 0$, then the associated ODE will be $\dot{x}_i = \lambda x \cdot Ay/(N-1) > 0$. That is, the noise directs the system inward away from the boundary of $S_N \times S_M$. This has the consequence that every element of x will remain strictly positive. As for (23), it gives rise to,

$$g_i^A(x) = 1 - Nx_i,\; g_j^B(y) = 1 - My_j \tag{27}$$

This specification of noise has been used in Gale et al. (1995) combined with replicator dynamics in the same manner as (25). Finally, (24) leads to

$$g_i^A(x) = x_i(-\log x_i + x \cdot \log x),\; g_j^B(y) = y_j(-\log y_j + y \cdot \log y) \tag{28}$$

The class of 2×2 games that have a unique mixed strategy equilibrium has attracted particular interest (Roth and Erev, 1998; Posch, 1997; Benaïm and Hirsch, 1996; Fudenberg and Kreps, 1993). Because we can replace x_2 by $1 - x_1$ for player A, and similarly for B, the state of the system can be summarised by the vector (x_1, y_1). Or in other words the learning dynamics will take place on the unit square. Without loss of generality, we can write the payoff matrices for the two players as

$$A = \begin{pmatrix} 1-a+c & c \\ c & a+c \end{pmatrix},\; B = \begin{pmatrix} c & b+c \\ 1-b+c & c \end{pmatrix}, \tag{29}$$

where $1 > a, b > 0$ and $c > 0$. The latter constant ensures all payoffs are strictly positive. There is a unique mixed strategy equilibrium where $(x^*, y^*) = (b, a)$. However, the corresponding perturbed equilibrium of (25) is not in general at (x^*, y^*).

Note that for all the different specifications of noise, when $a = b = \frac{1}{2}$, that is when the mixed equilibrium is exactly in the middle of the unit square, the perturbed equilibrium $(\hat{x}, \hat{y})$ equals the Nash equilibrium (x^*, y^*). Otherwise the perturbed equilibrium may be some distance from the actual Nash equilibrium. Take one game investigated by Ochs (1995), which provides one of the data sets analysed by Erev and Roth. In this game, the Nash equilibrium was $\theta^* = (x_1^*, y_1^*) = (0.5, 0.1)$. This is illustrated in Figure 1. The arrows represent expected motion of learning and are generated under the simple assumption that a strategy whose expected return exceeds the other will grow in frequency. This is a property of many learning models, including the basic model considered here, that is, without experimentation.

However, with experimentation the equilibrium is no longer θ^* but a perturbed equilibrium. In their first paper Roth and Erev (1995) a value of λ of 0.05 was used. In Erev

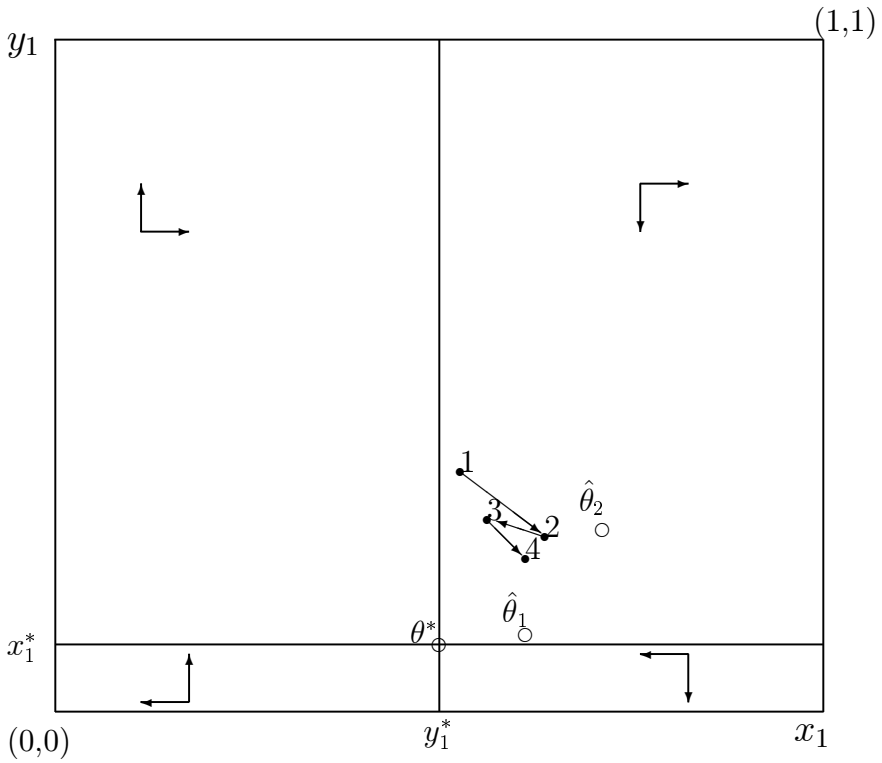

Figure 1: Perturbed equilibria and game dynamics.

and Roth (1998), the best fit of the data is obtained with a value of λ equal to 0.2. Given (26) and a value of $\lambda = 0.05$ then the fixed point of the ODE (25), solving these cubic equations numerically, will be $\hat{\theta}_1 = (0.6405, 0.1063)$. With a value of $\lambda = 0.2$ then $\hat{\theta}$ moves to $\hat{\theta}_2 = (0.7428, 0.2673)$, though, as λ increases and the noise dominates, the equilibrium will move toward (0.5,0.5). Points $\{1, 2, 3, 4\}$ represent aggregate data in blocks of 16 periods from the experiments run by Ochs (1995) as reported in McKelvey and Palfrey (1995).

Hence, if this model of reinforcement learning accurately describes subjects' learning behaviour then Nash equilibrium is not to be expected even in long run. Rather the perturbed equilibrium $(\hat{x}, \hat{y})$ is a limit point for the learning process. Note that the motion "away" from Nash equilibrium that Roth and Erev find in their data is toward this perturbed equilibrium. However, this model is not unique in possessing an equilibrium which is not identical to Nash.

This is also a characteristic of the stochastic version of fictitious play, (see Fudenberg and Kreps, 1993; Fudenberg and Levine, 1998; Benaïm and Hirsch, 1996). An equilibrium for this system of equations, like for the perturbed replicator dynamics, can be some distance away from the Nash equilibrium of the underlying game, the distance depending on the value of the noise parameter λ. Equally McKelvey and Palfrey (1995) propose a new equilibrium concept, "quantal response equilibrium" or QRE, which is based on perturbation of players' payoffs. This is very similar in motivation to the experimentation considered here. Furthermore, for the game considered here any QRE would like $(\hat{x}, \hat{y})$ be up and right of the Nash equilibrium. Indeed, McKelvey and Palfrey estimate the QRE from Ochs' experimental data at (0.649,0.254) which is roughly intermediate between $\hat{\theta}_1$ and $\hat{\theta}_2$. Note that a QRE can also be considered as an equilibrium point of the stochastic fictitious play process (11) above. Indeed, it is possible to show equivalence between all three types of equilibrium.[4]

Proposition 4 *Any equilibrium of the perturbed replicator dynamics (25) with noise specification (27) is an equilibrium point for the perturbed best response dynamics (17), which in turn is a quantal response equilibrium.*

Proof: In the Appendix. ∎

The intuition is simply that as the ODE associated with stochastic fictitious play can be written as a form of perturbed replicator dynamic, to match reinforcement learning with stochastic fictitious play, one merely needs to find the right noise function. For example, because the specification (28) has been very carefully chosen, equivalence can be taken a step further.

Proposition 5 *The ODE's associated with the perturbed reinforcement learning process with noise specification (28) are identical to a positive factor with the ODE's (20) associated*

[4]Roth and Erev's actual specification of noise, because there g^A, g^B depend on both x and y, does not fit this pattern, though as we have seen the perturbed equilibria it produces are qualitatively similar.

with exponential fictitious play. They therefore share equilibrium points, and the stability properties of these equilibria are identical under the two dynamics.

Proof: This follows from inspection of (20), (25) and (28). ∎

Fudenberg and Kreps (1993), Benaïm and Hirsch (1996) show that in 2×2 games with a unique mixed equilibrium stochastic fictitious play converges to the corresponding perturbed equilibrium. I now show a similar result for the model of reinforcement learning considered here.

Proposition 6 *The perturbed equilibrium $(\hat{x}, \hat{y})$ of the game (29) is globally asymptotically stable under the perturbed replicator dynamics (25), for noise specification (26), (27) and (28).*

Proof: For dynamics with noise form (28), this result follows from Proposition 11 below. Note that for noise of form (26) and (27), $\frac{\partial g_i^A}{\partial x_i} < 0$, $\frac{\partial g_j^B}{\partial y_j} < 0$ for all i, j,. This implies that the divergence, defined as $\sum_i \frac{\partial g_i^A}{\partial x_i} + \mu \sum_j \frac{\partial g_j^B}{\partial y_j}$ is negative. Negative divergence is sufficient in a 2 dimensional system to rule out cycles or other exotic behaviour. The replicator dynamics (16) have zero divergence on $S_2 \times S_2$, (Hofbauer and Sigmund, 1998, pp132-3). To see this, consider the modification of (16) $\dot{x}/P$, $\dot{y}/P$ where $P(x, y) = x_1 x_2 y_1 y_2$. Note that a positive transformation, such as division by P, only changes the velocity not the orbits of a dynamical system. Thus, as the modified system has zero divergence, so does (16). Hence, as the perturbed replicator dynamics (25) are simply (16) plus a term with negative divergence, (25) will have negative divergence. Thus, the flow of the perturbed dynamics (25) must be volume contracting on the whole of $S_2 \times S_2$ and converge to the unique equilibrium point. ∎

This result together with the stochastic approximation results of Section 3 of course implies the following:

Proposition 7 *In game (29), the stochastic process defined by reinforcement learning with noise specification (26) or (27) or (28) converges with probability one to the perturbed equilibrium $(\hat{x}, \hat{y})$.*

Proof: This follows from Proposition 1 and the global stability of $(\hat{x}, \hat{y})$ under the associated ODE established in Proposition 6. ∎

Erev and Roth (1998) use data from experimental play of 2×2 games with a unique mixed strategy equilibrium to test between reinforcement learning and stochastic fictitious play. However, the result here is that in these games noisy reinforcement learning converges to a perturbed equilibrium, just as Fudenberg and Kreps (1993), and Benaïm and Hirsch (1996) have shown for stochastic fictitious play. That is, the two rival models have the same long run properties. The next section shows how this result extends to more general games.

6 Some General Results

In this section, several results on how the properties of the two models coincide either globally or in the neighbourhood of equilibrium are presented. The implication of Proposition 1 is that any stochastic learning process must be convergent if the associated ODE is globally stable. This might not seem very useful in that global stability results are hard to come by. However, there are three classes of games, dominance solvable games, rescaled zero sum games and rescaled partnership games, for which suitable results are obtainable, for both the perturbed replicator dynamics, associated with reinforcement learning, and the perturbed best response dynamics associated with stochastic fictitious play. This implies that for these games the two different models of learning will have the same long-run properties.

So, what are rescaled zero sum and partnership games? In zero sum games, players' interests are diametrically opposed. In contrast, partnership games are games of coordination and also have been called games of identical interest. There are many other games with the same fundamental structure without falling into these narrow categories. For example, whilst the game (29) is not zero-sum, it is a rescaled zero-sum game. Intuitively, this is because in this game the two players are in direct competition, A only gets a good payoff if B gets a bad one. The idea of rescaling is due to Hofbauer and Sigmund (1998, p128). (A', B') is a rescaling of (A, B) if there exist constants c_j, d_i and $\alpha > 0, \beta > 0$ such that

$$a'_{ij} = \alpha a_{ij} + c_j, b'_{ji} = \beta b_{ji} + d_i. \tag{30}$$

Then (A, B) is a rescaled zero sum game if there exists a rescaling such that $B' = -(A')^T$ and a rescaled partnership game if $B' = (A')^T$. Rescaling a game does not change its equilibria. That is, if (x^*, y^*) is a Nash equilibrium for (A, B), it is also for (A', B').

Each rescaled partnership game has a potential, equal to the players' payoffs, which is at a (local) minimum at any mixed equilibrium and at a (local) maximum at any pure equilibrium. The potential is always rising both under the perturbed best response dynamics and the perturbed replicator dynamic. This enables the following result.[5]

Proposition 8 *For any rescaled partnership game, both (a) the smoothed best response dynamics (17) and (b) the perturbed replicator dynamics (25) with noise specification (27) or (28) converge to a perturbed equilibrium. In generic games, and for generic initial conditions, this limit point will be a perturbed strict equilibrium.*

Proof: (a) Hofbauer and Hopkins (1999), (b) in the Appendix. ∎

[5]The genericity referred to in the following proposition is intended to rule out games with connected components of Nash equilibria which indeed are non-generic in the strategic form.

In contrast, rescaled zero sum games have a unique perturbed equilibrium, possibly corresponding to a mixed equilibrium, which is globally stable for the perturbed best response dynamics.

Proposition 9 *Rescaled zero sum games have a unique equilibrium for the perturbed best reply dynamics (17), which is globally asymptotically stable.*

Proof: Hofbauer and Hopkins (1999). ∎

Moving on to reinforcement learning, it is possible to obtain some further results for the unperturbed case investigated by Posch (1997). First, the proof of Proposition 8b is also valid for the case where $\lambda = 0$. That is, unperturbed reinforcement learning will converge in general rescaled partnership games. Next, consider mixed equilibria of rescaled zero sum games of arbitrary size. A fully mixed isolated equilibrium is only possible in the case where $N = M$, the case which is addressed in the following proposition.

Proposition 10 *In any rescaled zero sum game, any fully mixed equilibrium is neutrally stable under the replicator dynamics (16). Any solution (x_n, y_n) of the reinforcement learning process (7) that does not converge to the boundary or to the equilibrium point, converges to a closed orbit of the ODE's.*

Proof: In the Appendix. ∎

Staying with rescaled zero sum games, we can see that noisy reinforcement learning behaves differently. Global convergence results are obtainable with the perturbation function (28) because of the link with the perturbed best response dynamics. However, in the case of specification (27) global convergence can only be shown for equilibria located at the centre point of $S_N \times S_M$. In the 2×2 case, covered by Proposition 6, the negative divergence of the ODE's guaranteed global stability by ruling out cycles. But for general games, when a mixed equilibrium is not centrally located, interaction between the noise term which pushes toward the centre point of $S_N \times S_M$ and the replicator dynamics which are centered on the mixed equilibrium may produce limit cycles (even though the perturbed equilibrium must be locally asymptotically stable, see Proposition 13 below).

Proposition 11 *Any perturbed equilibrium of a rescaled zero sum game is globally stable under the perturbed replicator dynamics with noise (28). Any fully mixed equilibrium of a rescaled zero sum game located centrally, that is $x^* = y^* = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$, is globally stable under the perturbed replicator dynamics (25) with noise specification (27).*

Proof: In the Appendix. ∎

There is one other class of game where global convergence results are possible, games which are dominance solvable. They are included here for completeness and because one of the experimental games analysed by Camerer and Ho (1999) was of this type. It is well known that the unperturbed best response and replicator dynamics delete dominated strategies (see for example, Fudenberg and Levine, 1998, Ch3). That is, the area of the state space containing only undominated strategies is globally attracting, a property which will survive perturbation. However, the introduction of noise does prevent the share of any strategy falling to zero, so the following proposition contains the appropriately modified claim.

Proposition 12 *From any fully mixed initial state, under either the perturbed replicator or perturbed best response dynamics, the share x_i or y_j of any iteratively strictly dominated strategy as $t \to \infty$ is less than or equal to $C(\lambda) \geq 0$, where C is a constant and* $\lim_{\lambda \to 0} C(\lambda) = 0$.

Global results have been hard to come by. It is possible, however, to gain a quite general result on local stability.

Proposition 13 *If a perturbed equilibrium is asymptotically (un)stable for all perturbed best response dynamics (17) then it is asymptotically (un)stable for noisy replicator dynamics, specification (27) or (28).*

Proof: In the Appendix. ∎

There are some implications of this result that need to be highlighted. First, this result combined with Proposition 9 means that the perturbed equilibrium of a rescaled zero sum game is always at least locally asymptotically stable for noisy reinforcement learning. Second, unstable perturbed equilibria are only possible for sufficiently small λ. As λ becomes large, there will be a unique globally stable equilibrium corresponding to the maximum of the noise function ϕ. Finally, unstable perturbed equilibria are rare even for small λ. Some Nash equilibria which are not fully mixed and which are not asymptotically stable under the unperturbed dynamics, will have no corresponding perturbed equilibrium. Which equilibria survive depends on the type of perturbation, the level of perturbation λ, and the form of the original dynamics (see Binmore and Samuelson, 1999, for an analysis of this issue). But because of Propositions 4 and 5, we know that if an equilibrium survives for the perturbed replicator dynamics, it survives for the perturbed best response dynamics, and if it disappears, it disappears for both.

It is possible to combine the results of this section and those of Section 3.

Proposition 14 *Both stochastic fictitious play and noisy reinforcement learning converge with probability one in generic rescaled partnership games to a perturbed equilibrium corresponding to a pure equilibrium of the unperturbed game. In rescaled zero sum games both stochastic fictitious play and noisy reinforcement learning, specification (28), converge with probability one to the unique perturbed equilibrium. If this equilibrium is centrally located, noisy reinforcement learning, specification (27), also converges with probability one.*

Proposition 15 *The limit point of stochastic fictitious play and noisy reinforcement learning with probability one places a share of at most $C(\lambda)$, as defined in Proposition 12, on any strategy which is iteratively strictly dominated.*

A negative result is also possible.

Proposition 16 *If a perturbed equilibrium is unstable for all perturbed best response dynamics of form (17), then this is the limit point of either stochastic fictitious play or noisy reinforcement learning, specification (27) or (28), with probability zero.*

7 Transition Dynamics and the Role of Information

In the long run we are all dead. Those running experiments or analysing experimental data, where there is at most a couple of hundred of rounds of play and usually much less, may react the same way to the asymptotic results presented up till now. In this section, I try to draw out the conclusions of the preceding theoretical analysis for the behaviour of learning in finite time, in particular differences in relative speed. Second, I try to isolate the effect of agents' use of information. The conclusions are surprising. Hypothetical reinforcement does not make learning faster. It actually worsens the performance of learning when the standard reinforcement learning choice rule is used.

Stochastic approximation techniques are also relevant for finite time horizons. Solutions of the discrete time stochastic process with high probability will remain "close" to the solution of the associated ODE with the same initial conditions (see for example, Theorem 1, Benveniste et al., 1990, p43). Given that each pair of players playing a series of games offers an independent realisation, data averaged over p pairs, as p grows large, should approximate the solution of the ODE. Here I present numerical solutions for the two sets of ODE's, (20), associated with the exponential version of stochastic fictitious play and (25) with noise (28), associated with perturbed reinforcement learning. Now, the first set of ODE's is simply β times the second. Consequently, they have the same qualitative behaviour. For the game (29) for example, for parameter values $a = 0.1, b = 0.5, \beta = 5$, they both will have a globally stable perturbed equilibrium $(\hat{x}_1, \hat{y}_1) = (0.6997, 0.2692)$. However, one can see that the speed of convergence is quite different.

		ODE (20) Stochastic Fictitious Play		ODE (25) Reinforcement Learning	
n	t_n	x_1	y_1	x_1	y_1
1	1	0.5	0.5	0.5	0.5
100	5.18	0.6977	0.2657	0.7281	0.389
200	5.88	0.6978	0.2685	0.7384	0.367
500	6.79	0.699	0.2696	0.7457	0.3418
1000	7.49	0.6996	0.2696	0.7479	0.325
5000	9.09	0.6698	0.2692	0.7455	0.295
∞	∞	0.6997	0.2692	0.6997	0.2692

Table 1: Speed of convergence to perturbed equilibrium

Note that the number n of discrete periods can be converted to a continuous time analogue by the following formula

$$t_n = \sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} \frac{1}{i}$$

and some values are recorded are reported in the first two columns of the following table. This conversion means that though one ODE is running five times slower in continuous time, the difference in speed in terms of discrete periods is dramatically greater. The initial conditions are given in the first row of data. This suggests, and simulations of the stochastic model confirm, that it may take as many as 10,000 periods for reinforcement learning to approach the perturbed equilibrium in this simple game, a feat that stochastic fictitious play can achieve in about 100.[6] The latter is sufficiently quick that asymptotic predictions may be of relevance in experimental time frames. Note that the difference in speed is entirely driven by the parameter β, which reflects the degree of optimisation. As will be shown, it does not result from the two models' different use of information.

The reinforcement learning model is adapted to the situation where there is no hypothetical reinforcement. As Rustichini (1998) notes, the introduction of hypothetical reinforcement worsens the performance of reinforcement learning in single agent decision problems. A similar effect can be seen in games. If propensities are reinforced according to the standard fictitious play rule (4), that is, without any noise or experimentation, then the approximation $q_{in}^A/n = (Av_n)_i$ can be used. The standard reinforcement learning choice rule given by (1) becomes simply the stochastic fictitious play process (10) with $\beta = 1$. That is, reinforcement learning with hypothetical reinforcement is a stochastic fictitious play process. However, it is a very noisy one, even though there is no noise in the reinforcement process. The interaction of hypothetical reinforcement and the basic reinforcement choice rule has

[6]Reinforcement learning can be quicker than this. The introduction of a forgetting parameter by placing a lower bound on the step size of the learning process will speed convergence. See Section 8 below.

introduced bias into the learning process.

In contrast, fictitious play is adapted to hypothetical reinforcement and may perform suboptimally in its absence. However, if the actions of one's opponent are unobservable, then hypothetical reinforcement becomes very difficult. Sarin and Vahid (1998) argue that agents should still maximise given the information they possess. What can be said about models of this type? Given that in this case opponents' actions are assumed unobservable, it makes no sense to use the standard fictitious play choice rule which is in terms of historical frequencies of opponents' play. Suppose instead we imagine that each player has propensities q_n^k for each of her strategies and player A chooses a strategy x, to maximise $x \cdot q_n^A/n + \lambda\phi(x)$, the solution will be $x = \overline{BR}(q^A)$. So assume $x_n = \overline{BR}(q_n^A)$ and $y_n = \overline{BR}(q_n^B)$. For example in the exponential case

$$x_{in} = \frac{\exp \beta q_{in}^A/n}{\sum_{j=1}^N \exp \beta q_{jn}^A/n} = \overline{BR}_i^e(q_n^A). \tag{31}$$

This type of model is potentially problematic. In the extreme case of classical fictitious play the strategy with the highest propensity is played with probability one. With no hypothetical reinforcement, only this strategy will be reinforced. In these circumstances a player will be locked into a single strategy, simply because it was initially preferred. Fudenberg and Levine (1998, Ch4) suggest as a correction that a propensity for an action should be reinforced in inverse proportion to the probability of that action being taken. That is, actions which are ascribed a low probability are reinforced more strongly when taken. If A chooses action i and B chooses j, then A's ith propensity is updated in this way

$$q_{in+1}^A = f_{ij}(q_{kn}^A) = q_{kn}^A + \frac{a_{ij}}{x_{in}}. \tag{32}$$

We now have a choice rule and a method of updating propensities. To construct an ODE in terms of (x, y), exactly the same method can be used as in Section 4. It can easily be checked that there will be exactly the same result.

Proposition 17 *The ODE associated with the stochastic learning process defined by (32) is the same as that (21) associated with standard stochastic fictitious play.*

Thus, for example in the case of the exponential choice rule, the same ODE (20) is associated with the process and so the numerical results in Table 7 are equally applicable. That is, stochastic fictitious play can be just as fast without any hypothetical reinforcement. It is difficult therefore to find any effect from the differing use of information! However, note that in the absence of hypothetical reinforcement, in each period there are $N \times M$ possible realisations for changes in the propensities of each player. With hypothetical reinforcement, there are only M ways that A's propensities can be reinforced with those M outcomes being convex combinations of the $N \times M$. Thus, I conjecture that the only result of varying the

amount of information would be to affect the variance of the sample paths of the learning process around the solution of the ODE. Learning without hypothetical reinforcement would be more noisy.

It is possible to make a similar but opposite correction for reinforcement learning. Suppose the basic reinforcement learning choice rule is used but with the following hypothetical reinforcement learning rule,

$$q^A_{kn+1} = f_{ij}(q^A_{kn}) = q^A_{kn} + x_{kn}(a_{kj} - \log x_{kn}) \text{ for } k = 1, ..., N. \tag{33}$$

Then, the ODE associated with the learning process would be the perturbed replicator dynamics (25) with noise (28) just as in the absence of hypothetical reinforcement. Note that the correction employed is not throwing away any information. It simply removes the bias that hypothetical reinforcement can give to reinforcement learning. Every propensity is updated every period. However, learning is no faster than for standard reinforcement learning.

The conclusion is that learning can be just as fast and accurate without hypothetical reinforcement as with it. Speed more depends on which functional form is used for the choice rule, the exponential form being fast, the standard reinforcement rule being slow.

8 Forgetting

Up to now, it has been assumed that all observations are given equal weight. This has the effect that as the number of periods progresses, the marginal impact of new experiences upon behaviour decreases, asymptotically approaching zero. As we have seen this has a certain mathematical convenience. However, both Erev and Roth (1998) and Camerer and Ho (1999) find that experimental data seems to support the hypothesis that agents discount previous experience, which implies that learning will not come to a complete halt even asymptotically (as Cheung and Friedman (1997) point out, disaggregated data indicates that the level of discounting varies enormously between individuals). It has been hypothesised that this form of learning would be useful in non-stationary environments but this claim has received little analysis. Rustichini (1998), however, casts doubt that using such rules in single agent decision problems aids optimal decisions. It remains to see what effect discounting previous experience has on the way people learn to play games.

The easiest way of modelling this form of behaviour is to introduce a "forgetting" or "recency" parameter denoted here by $0 < \delta < 1$. Every period, all propensities are multiplied by δ. That is, when event ij occurs,

$$q^k_{n+1} = \delta q^k_n + \Pi^k_{ij}, \tag{34}$$

where Π^k_{ij} is a vector of payoffs. Under reinforcement learning, using the experimentation rule (22), Π^k_{ij} is the vector with $\lambda\pi^k_n/(N-1)$ in every position except the ith, which is equal

to $(1-\lambda)\pi_n^k$. Under the alternative rule (23), Π_n^k is the vector with λ in every position except the ith which is equal to $\pi_n^k + \lambda$.

Under fictitious play Π_{ij}^A is the jth column of A. Under discounting, beliefs about one's opponent's play are not identical to historical frequencies. In this section, the variables (u, v) are redefined as what one could call "discounted historical frequencies". If in period n, player A chooses her first strategy, then $u_{1n+1} = 1-\delta+\delta u_{1n}$. The rule (34) above is then equivalent to the following rule,

$$E[u_{n+1}|u_n, v_n]-u_n = (1-\delta)(\overline{BR}(v_n)-u_n),\ E[v_{n+1}|u_n, v_n]-v_n = (1-\delta)(\overline{BR}(u_n)-v_n). \quad (35)$$

Note that, for example, it is now true that $Av_n \approx q_n^A(1-\delta)$.

Some quick calculation will reveal that even with the introduction of forgetting the stochastic learning process can still be written in the form (12). The difference is only in evolution of the step size, the expected motion of the system is unchanged. Consequently, whether we are considering fictitious play or reinforcement learning, we have the same appropriate associated ODE's. This implies that in the cases we have identified as globally stable, the stochastic process will still tend to move toward the equilibrium point of the ODE, however, with an important difference. Up to now, what has brought much of the results has been the decreasing step size of the algorithm.

Dealing with models of this sort where randomness is persistent is problematic. The obvious and most common approach to such a problem is to demonstrate the existence of an invariant distribution. This I do here. However, the task then is to characterise it. Except in the simplest examples, this typically can only be done by taking some form of limit. This is done in the stochastic approximation literature by looking at the asymptotic normality of the deviations of the stochastic process from the solution of the ODE. This approach is followed here.

Some results follow based on techniques developed by Norman (1968). The focus is now on the underlying Markov process defined by the appropriate choice rule and the updating rule (34). The state of the process at any time can be given by $q_n \in I\!R_+^N \times I\!R_+^M = S$, that is, a vector of weights for each player. This obviously evolves according to the actions chosen by the two players. If the first player chooses action i, and the second j, then denote that event as ij and event operator f_{ij}. Norman (1968) defines a Markov process on a metric space with metric d to be strictly "distance diminishing" if $\rho(f_{ij}) < 1$ for all ij where

$$\rho(f) = \sup_{q \neq q'} \frac{d(f(q), f(q'))}{d(q, q')}.$$

Lemma 1 *The Markov process defined by reinforcement learning or stochastic fictitious play with forgetting is distance diminishing with respect to the standard Euclidean metric.*

Proof: Given arbitrary states q, q', $f_{ij}(q) = (\delta q^A + \Pi^A, \delta q^B + \Pi^B)$ and $f_{ij}(q') = (\delta q'^A + \Pi^A, \delta q'^B + \Pi^B)$. It is easy to show therefore that $d(f_{ij}(q), f_{ij}(q')) = \delta d(q, q')$ and $\rho(f_{ij}) = \delta$ for all possible events. ∎

Let $T_n(q)$ be the set of states reached with positive probability in n steps if we start at q. Let $d(A, B)$ be distance between two subsets A and B of the state space. That is,

$$d(A, B) = \inf_{q \in A, q' \in B} d(q, q')$$

Then Norman (1968, Theorem 2.2, p66) is able to show that if the following condition holds

$$\lim_{n \to \infty} d(T_n(q), T_n(q')) = 0 \text{ for all } q, q' \in S$$

then a distance diminishing Markov process is ergodic. That is, its limit distribution is independent of initial conditions.

Proposition 18 *The Markov process defined by reinforcement learning or stochastic fictitious play with forgetting is ergodic, with an invariant distribution on* $\mathbb{R}^N \times \mathbb{R}^M$.

Proof: From an arbitrary initial state q_0, each element of which is strictly positive, there is a positive probability that both players continue to choose their first action for an indefinite number of periods. As this run of play continues, q_n will approach the state $(\Pi^A_{11}, \Pi^B_{11})/(1 - \delta)$. This state is therefore accessible from any initial state and from the theorem of Norman the Markov process is ergodic. ∎

The task now is to characterise the unique limiting distribution. This I do for the cases identified as having a unique globally stable equilibrium. It is important to realise that the theory of stochastic approximation still has a lot to say when γ_n is non-decreasing, provided it remains "small". In the model considered here, this is equivalent to δ being close to one.

In the case of reinforcement learning, again for simplicity we rebalance the step size.

Assumption B: *Rebalancing.* At each period, for each player k, after updating by payoffs, every propensity is multiplied by an appropriate factor so that $Q^k_n = \delta Q^k_{n-1} + 1 = \delta^n Q_0 + (1 - \delta^n)/(1 - \delta)$, but leaving x_n, y_n unchanged.

In this case, for both players Q_n smoothly approaches $1/(1 - \delta)$. That is, even in the limit the sum of the propensities remains finite, and the step size is bounded away from zero. One important characteristic is that if the initial value of Q is lower than its long-run value, learning will initially be relatively quick, but will slow as time progresses. Unlike in the analysis of previous sections, learning will never come to a complete halt. Under stochastic fictitious play, the updating rule (35) implies that the step size is exactly $1 - \delta$. This is of course the asymptotic value of the step size for reinforcement learning under Assumption B.

Propositions 19 and 20 which follow assume a constant step size. Given that these results are on the asymptotic behaviour of learning, in what follows, for the constant "γ", one can read the asymptotic value of γ_n.

The first conclusion we can draw from the theory of stochastic approximation is that the limit distribution will be in the following sense clustered around the perturbed equilibrium $(\hat{x}, \hat{y})$.

Proposition 19 *For $\gamma > 0$ sufficiently small, for all $\varepsilon > 0$, if $(\hat{x}, \hat{y})$ is a globally stable equilibrium point for the ODE associated with the learning model, there exists a constant $C(\gamma)$ such that*

$$\limsup_{n \to \infty} \Pr\{||(x_n, y_n) - (\hat{x}, \hat{y})|| > \varepsilon\} \leq C(\gamma)$$

where $C(\gamma)$ tends to zero as γ tends to 0.

Proof: Benveniste et al. (1990, Ch2, Theorem 3). ∎

The second type of result is on the asymptotic normality of the deviations from the solution of the ODE. Thus, if the ODE converges to a perturbed equilibrium $(\hat{x}, \hat{y})$, the time average of the stochastic process converges too. Let $(x(t), y(t))$ be an orbit of the ODE associated with the discrete stochastic process. Then one can define the distance between the solutions of the stochastic and deterministic processes, suitably rescaled, as

$$(\tilde{x}_{t_n}, \tilde{y}_{t_n}) = \gamma^{-\frac{1}{2}}(x_n - x(t_n), y_n - y(t_n)). \tag{36}$$

Again a constant γ is assumed so that $t_n = n\gamma$.

Proposition 20 *If $(\hat{x}, \hat{y})$ is a globally stable rest point for the associated ODE's and all the eigenvalues of the matrix $h_{x,y}(\hat{x}, \hat{y})$ have strictly negative real parts, as $\gamma \to 0$ and $t_n \to \infty$, $\tilde{x}_{t_n}$, $\tilde{y}_{t_n}$ converge to normally distributed random variables with zero mean.*

Proof: Benveniste et al. (1990, Ch3, Theorem 2). ∎

This in turn implies,

Proposition 21 *In 2×2 games with a unique mixed equilibrium, under stochastic fictitious play and noisy reinforcement learning, and in rescaled zero sum games under stochastic fictitious play and noisy reinforcement learning with specification (28), as $\delta \to 1$ and $n \to \infty$ the time average of play $(\frac{1}{n}\sum x_n, \frac{1}{n}\sum y_n)$ converges to the unique equilibrium.*

	ODE (25) $\delta = 0.99$			ODE (25) $\delta = 0.9$		
n	t_n	x_1	y_1	t_n	x_1	y_1
1	1	0.5	0.5	1	0.5	0.5
100	5.70	0.7451	0.3444	12.71	0.7195	0.2609
200	7.01	0.7481	0.3143	22.71	0.6975	0.2673
500	10.15	0.7349	0.2727	52.71	0.6997	0.2692
∞	∞	0.6997	0.2692	∞	0.6997	0.2692

Table 2: Speed of convergence to perturbed equilibrium with forgetting

Proof: This follows from an application of Propositions 6, 9 and 20. Proposition 13 established the final condition that $h_{x,y}$ has negative eigenvalues. ∎

What this implies is that for values of δ close to one, and hence when (the asymptotic value of) γ is close to zero, the long term time average of (x_n, y_n) will be close to $(\hat{x}, \hat{y})$. At this point, it is worth comparing the results of this section with what has been previously obtained. In particular, will the introduction of forgetting increase the speed of reinforcement learning to levels achievable by stochastic fictitious play? An indication is given in Table 8, which presents more numerical solutions to the ODE (25) with noise (28). Here t_n is calculated as $\sum_{i=1}^{n} \gamma_i$, where, under Assumption B, $\gamma_n = 1/Q_n = (1-\delta)/(1-\delta^{n+1})$. One can see that if δ is as low as 0.9 then reinforcement learning seems to be about as fast as stochastic fictitious play. However, these results should be treated with caution. As δ moves away from one, the level of variance of the stochastic process around the solution of the ODE increases. Consequently, individual realisations of the stochastic learning process may spend a long time at a significant distance from the perturbed equilibrium.

9 Endogenous Aspirations

One somewhat unsatisfactory characteristic of the basic stimulus response model is that every payoff is a reinforcement. That is, given that all payoffs are assumed positive, once any action is taken the probability of doing so again rises. Hence an agent taking some action that pays \$1 will have that action reinforced even if all other actions pay \$1000. A very simple idea is that there should be an aspiration level or reference point, s_n. This aspiration level could vary over time according to an agent's payoff history. If an action is taken resulting in a payoff which is less than the aspiration level then the probability of repeating that action will decrease.

Of course, this problem does not arise under fictitious play, where hypothetical reinforce-

ment means that all payoffs received are effectively judged against the reference point of what a player would have received had she played some other strategy. But if the situation is such that hypothetical reinforcement is difficult, because for example, opponents' actions are unobservable, a variable reference point might be a good substitute. The question here is, therefore, does its introduction make reinforcement learning more like fictitious play?

To concentrate on the effect of the introduction of an aspiration level, in this section we revert to the "basic" model, that is, without experimentation or forgetting. One approach as suggested by Erev and Roth (1998) and implemented by Erev and Rapoport (1998) is simply to subtract the aspiration level from the realised payoff. If a player k takes an action i, the change in the propensity is,

$$q^k_{in+1} = q^k_{in} + \pi^k_n - s^k_n. \tag{37}$$

Thus reinforcement from choosing action i would be negative if $\pi^k_n < s^k_n$ and therefore the associated probability x_{in} or y_{in} of performing that action would decrease. As we will show, this is has little effect on the expected motion of the system, but it will affect the step size of the algorithm.

The principal disadvantage of this method is the possibility that a propensity, q_i, could become negative and therefore the associated probability x_i would no longer be defined. Erev and Rapoport (1998) therefore introduce a lower bound $\bar{q} > 0$ on all the propensities to ensure they remain positive. That is, (37) becomes

$$q^k_{in+1} = \max[q^k_{in} + \pi^k_n - s^k_n, \bar{q}]. \tag{38}$$

Note that as long as this bound is not hit, that is, as long as the system does not approach the boundary of $S_N \times S_M$, the expected motion of x is unchanged. This is because the expected change in the probability of playing each strategy is given by expected payoff of that strategy less average expected payoffs. Now, endogenous aspirations will reduce the expected payoff to all strategies by s_n, so the difference between the return to one strategy and the average return is unchanged.

However, this model of endogenous aspiration levels will affect the step size of the stochastic process. Let us further assume that aspirations are updated by the simple formula

$$s^k_{n+1} - s^k_n = \beta(\pi^k_n - s^k_n) \tag{39}$$

where $0 < \beta < 1$ is a constant. Note that the aspiration level will tend to catch up with current expected payoffs so that the expected growth in the sum of all the propensities is zero. The step size, which is the inverse of the sum of the propensities, Q_n, would then follow a random walk. Under the present assumptions Q_n would be bounded below by $N\bar{q}$.

In summary, therefore, the only effect of the variable reference point is on the step size of the algorithm. Again, like the introduction of forgetting, it prevents the step size of the

learning process falling to zero. It therefore may make learning run a little faster. Whereas the introduction of endogenous aspirations might have seemed to add psychological realism and lead to a better fit to experimental data, it in fact adds relatively little.[7]

10 Reinforcement Learning without Rebalancing

The purpose of this section is to investigate the behaviour of reinforcement learning without rebalancing, that is under Assumption A2. In this case, the step size of the learning process is endogenous, and each player has a different step size. How does this affect the learning process? It is certainly much more difficult to obtain results on this learning model. However, it does seem that behaviour is broadly similar to when propensities are rebalanced.

Define H_μ such that

$$\mu_n = \mu_{n-1} + \frac{1}{Q_n^A} H_\mu(q, X_n) + O(\frac{1}{n^2}) \tag{40}$$

Then

$$\lim_{n\to\infty} H_\mu(q, X_n) = \frac{1}{Q^A}\mu(x \cdot Ay - \mu y \cdot Bx) + O\left(\frac{1}{(Q^B)^2}\right).$$

Therefore the behaviour of the learning process under Assumption A2 can be understood by augmenting the replicator dynamics (16) with an additional ODE,

$$\dot{\mu} = \mu(x \cdot Ay - \mu y \cdot Bx). \tag{41}$$

First, note that this specification will have exactly the same asymptotic behaviour as under Assumption A1 in dominance solvable games and in rescaled partnership games. The weight placed on dominated strategies will clearly go towards zero irrespective of the relative speed of the two players' learning. The potential of a rescaled partnership game, as given in the Proof of Proposition 8 is not a function of μ and will always be growing out of equilibrium irrespective of the value of μ. As for (41), in the special case of a partnership game where $A = B^T$ then, $\dot{\mu} = \mu(1-\mu)x \cdot Ay$. Clearly, in this case μ converges to one.

The behaviour of this learning scheme in other games is potentially somewhat more complex. Nonetheless it is possible to establish the same general result on local stability as for noisy reinforcement learning with rebalancing.

Proposition 22 *If a perturbed equilibrium is asymptotically (un)stable for all perturbed best response dynamics (17) then it is asymptotically (un)stable for noisy replicator dynamics, specification (27) or (28), when μ is determined by the additional ODE (41).*

[7] Börgers and Sarin (1996) take a different approach to modelling aspirations in reinforcement learning, which leads to markedly different results.

Proof: In the Appendix. ∎

11 Conclusions

This paper extends the analytic results on stochastic learning models beyond 2×2 games. In doing so, it shows that there a pure reinforcement model can generate exactly the same results as stochastic fictitious play, albeit with much less speed. Furthermore, in games where there is a strong structure, of competition or coordination or dominance solvability, there is a more general equivalence between stochastic fictitious play and reinforcement learning. This implies that existing tests between them using data from games in this class will have low power. The link between the two models is even stronger, if attention is confined to local stability. Together this implies somewhat surprisingly that the possession of or the lack of information about opponents' actions has only a marginal effect on the outcome of the learning process

This is not to make the claim that the two models are identical. There may well be more complex games where the greater sophistication of stochastic fictitious play leads to substantially different outcomes. In particular, one can hypothesise that games with a non-trivial extensive form and hence where the manipulation of information and the forming of hypotheticals are more important, might be such a case. This is the rationale for the experiments reported in Feltovich (1999) on a constant sum game with asymmetric information. However, he reports that the even here "two models yield qualitatively similar patterns of behavior". Of course, the results in this paper suggest why this may have been the case. There remains a potential research agenda, to find games for which the two models predict outcomes which are quite different. If these games can be run experimentally, then which model is better at predicting how people learn can really be tested.

Lastly, though it seems that information on opponents' moves has very little effect on the asymptotic properties of learning, this paper does not claim that any and all information that agents might receive is irrelevant. Rather, it seems that the debate over fictitious play and reinforcement learning about whether or not agents use hypothetical reinforcement has been a debate about the wrong type of information. These models are concerned with pairs of agents playing in isolation. In contrast, there have been experiments, for example, Huck et al. (1999) and Duffy and Feltovich (1999) where some subjects are informed about the behaviour of other subjects in a way that permits learning by imitation. Here, it seems that whether this information is provided or withheld can have significant effects on play.

Appendix

As a preliminary to the following proofs, define $I\!R_0^N = \{x \in I\!R^N : \sum x = 0\}$ and $I\!R_1^N = \{x \in I\!R^N : x_i = x_j \text{ for all } i, j\}$. These two subspaces are orthogonal. After a rescaling of the form

(30), one can write $A' = \alpha A + D$, where D is a matrix where each column is an element of $\mathbb{R}_1^N$. Hence $\xi A' = \alpha \xi A$ for any $\xi \in \mathbb{R}_0^N$ and $\eta B' = \beta \eta B$ for any $\eta \in \mathbb{R}_0^M$.

Second, note that the perturbed replicator dynamics with noise (27) can be written

$$\dot{x} = R(x)Ay + \lambda(u - Nx),\ \dot{y} = \mu\left(R(y)Bx + \lambda(u - My)\right). \tag{42}$$

where u is a vector of ones, and $R(x)$ and $R(y)$ are replicator operators, that is, matrices such that $R_{ii}(x) = x_i(1 - x_i)$ and $R_{ij}(x) = -x_i x_j$ and similarly for $R(y)$. These matrices are positive definite with respect to $\mathbb{R}_0^N$ and $\mathbb{R}_0^M$ respectively.

Proof of Proposition 4 The functions, $\phi(x) = \sum_{i=1}^N \log x_i$, $\phi(y) = \sum_{i=1}^M \log y_i$ satisfy the 2 conditions set out in Section 2 to act as suitable perturbations to construct a perturbed best response function. Second note that

$$\frac{d\phi(x)}{dx} \cdot R(x)Ay = \sum (Ay)_i - Nx \cdot Ay = (u - Nx)Ay.$$

That is, $d\phi(x)/dx = (u - Nx)R^{-1}$. Combining this with (42), the perturbed replicator dynamics can be rewritten as

$$\dot{x} = R(x)\left(Ay + \lambda\phi'(x)\right),\ \dot{y} = \mu R(y)\left(Bx + \lambda\phi'(y)\right). \tag{43}$$

Given that $R(x)$ and $R(y)$ are positive definite, the fixed points of the dynamic must satisfy the simultaneous equations

$$Ay + \lambda\phi'(x) = 0,\ Bx + \lambda\phi'(y) = 0. \tag{44}$$

But these are exactly the first order conditions that define a fixed point for the perturbed best response dynamics. The definition of a QRE (McKelvey and Palfrey, 1995) is the solution of a perturbed maximisation problem which includes the current specification as a special case. ∎

Proof of Proposition 8. The perturbed replicator dynamics with noise either of type (27) or (28) can be written in the form (43), with in the first case $\phi(x) = \sum \log x_i$ and in the second case $\phi(x) = -\sum x_i \log x_i$. Define

$$V(x, y) = x \cdot A'y + \lambda\phi(x) + \beta\lambda\phi(y),$$

where A' is the rescaling of form (30) of A such that $(A')^T = B'$. Without loss of generality, in the rescaling set $\alpha = 1$ and $\beta > 0$. Then,

$$\frac{\partial V}{\partial x} = A'y + \lambda\phi'(x),$$

and

$$\frac{\partial V}{\partial y} = xA' + \beta\lambda\phi'(y) = \beta(B'x + \lambda\phi'(y)).$$

Under the constraint that $x \in S_N$ and $y \in S_M$ the above conditions are identical to (44) above. Thus, V has turning points that correspond to perturbed equilibria. Furthermore,

$$\begin{aligned}\frac{\partial V}{\partial x} \cdot \dot{x} &= (A'y + \lambda\phi'(x)) \cdot R(x) \, (Ay + \lambda\phi'(x)) \\ &= (Ay + \lambda\phi'(x)) \cdot R(x) \, (Ay + \lambda\phi'(x)) \geq 0.\end{aligned} \tag{45}$$

This uses the fact that as $\dot{x} \in I\!R_0^N$, and therefore as noted above, $(A'y) \cdot \xi = (Ay) \cdot \xi$ for any $\xi \in I\!R_0^N$. Similarly

$$\frac{\partial V}{\partial y}\dot{y} = \mu\beta \, (Bx + \lambda\phi'(y)) \cdot R(y) \, (Bx + \lambda\phi'(y)) \geq 0. \tag{46}$$

Note that $\dot{V}$ is equal to the sum of (45) and (46), and hence is always positive when not in equilibrium. Thus, there is convergence to a (local) maximum of V. ∎

Proof of Proposition 10. Let there be a mixed equilibrium (x^*, y^*). Define

$$V(x, y) = \mu \sum_{i=1}^{N} x_i^* \log x_i + \beta \sum_{i=1}^{N} y_i^* \log y_i.$$

It is known (Hofbauer and Sigmund, 1998, p130), first that this function has a maximum at (x^*, y^*) and second that this function is a constant of motion for the dynamics (16), that is, $\dot{V} = V_x \cdot \dot{x} + V_y \cdot \dot{y} = 0$. To see this, first, note that if a game is rescaled zero sum, we can find a rescaling with $\alpha = 1$, and some $\beta > 0$, such that $A' = -(B')^T$. Then

$$V_x \cdot \dot{x} = \mu(x^* - x)R^{-1}(x) \cdot R(x)Ay = \mu(x^* - x)A'y = \mu N(x^* - x)A'(y - y^*)$$

and that

$$V_y \cdot \dot{y} = \beta(y^* - y)R^{-1}(y) \cdot \mu R(y)Bx = \mu N(y^* - y)B'(x - x^*).$$

These two expressions clearly sum to zero. Therefore the equilibrium will be neutrally stable under the replicator dynamics and is surrounded by a continuum of closed orbits which are the level curves of $V(x, y)$. The result then follows from Proposition 1. ∎

Proof of Proposition 11. Convergence of the perturbed replicator dynamics with noise (28) follows from Propositions 5 and 9. For the second part of the proof, take the function $V(x, y)$ defined in the proof of Proposition 10 above, and assume that the equilibrium is centrally located, that is, $x^* = y^* = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$. When we replace the replicator dynamics with the perturbed version, in the case of noise (27), $\dot{V}$ now becomes

$$\dot{V} = \lambda\mu(u - Nx) \cdot R^{-1}(u - Nx) + \lambda\mu\beta(u - Ny) \cdot R^{-1}(u - Ny) \geq 0,$$

and the result follows. ∎

Proof of Proposition 13. The local stability of an equilibrium of any dynamic invariant on $S_N \times S_M$ will be determined by the $N - 1$, $M - 1$ eigenvalues corresponding to $I\!R_0^N$, $I\!R_0^M$

respectively. This analysis, therefore, is confined to the subspace $\mathbb{R}_0^N \times \mathbb{R}_0^M$. To construct the linearisation of perturbed best response dynamics (21) at a perturbed equilibrium point note that

$$\frac{d\dot{x}}{dx} = \beta \left(Q'(Ay + \lambda\phi'(x)) + Q\lambda\phi'' \right) = -I$$

given that $Q(x) = -(\phi'')^{-1}$. Equally,

$$\frac{d\dot{x}}{dy} = \beta Q(x) A.$$

Thus the linearisation can be written

$$\beta \begin{pmatrix} Q(x) & 0 \\ 0 & Q(y) \end{pmatrix} \begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} - I = \beta Q(\theta) C - I$$

where again θ is used as an abbreviation for (x, y). As noted above, the perturbed replicator dynamics with noise either of type (27) or (28) can be written in the form (43). At a perturbed equilibrium $\hat{\theta}$ satisfying the conditions (44), the linearisation of these dynamics can be written

$$J = R(\hat{\theta})C + \lambda R(\hat{\theta})\Phi(\hat{\theta}) = \begin{pmatrix} 0 & R(\hat{x})A \\ \mu R(\hat{y})B & 0 \end{pmatrix} + \lambda \begin{pmatrix} R(\hat{x})\phi''(\hat{x}) & 0 \\ 0 & \mu R(\hat{y})\phi''(\hat{y}) \end{pmatrix}. \quad (47)$$

Now when $\phi(x) = -\sum x_i \log x_i$ as it does for (28) and indeed for exponential fictitious play, then $R(x)\phi''(x) = -I$ and similarly $R(y)\phi''(y) = -I$. When $\phi(x) = \sum \log x_i$, it is still true that because ϕ'' is a negative definite matrix, $R(x)\phi''(x)$ has all eigenvalues with negative real part (see eg Hopkins, 1999a, Lemma 2), and thus so does $R(\hat{\theta})\Phi(\hat{\theta})$.

Now, the aim is to show that $\beta Q(\theta)C - I$, the linearisation of the perturbed best response dynamics, and J, the linearisation of the perturbed replicator dynamics, have the same sign pattern at any perturbed equilibrium. If the perturbed equilibrium $\hat{\theta}$ is unstable under all perturbed best response dynamics, then $Q(\hat{\theta})C$ has at least one positive eigenvalue for all suitable Q. Hence $R(\hat{\theta})C$ has at least one positive eigenvalue and so does J for small enough λ. Because QC has a zero trace, it has either both positive and negative eigenvalues or all with zero real part. Hence, an equilibrium can only be asymptotically stable for all perturbed best response dynamics for small λ if $Q(\hat{\theta})C$ has all eigenvalues with real part zero. This can be for three reasons. One, either A or B is zero, then the eigenvalues of $R(C + \lambda\Phi)$ are identical to the eigenvalues of $\lambda R\Phi$. Two, QC has all eigenvalues with zero real part for for all nonzero suitable Q, and nonzero C, if and only if (A, B) is rescaled zero sum (Hofbauer and Hopkins, 1999). Then $\xi \cdot A\eta + c\eta \cdot B\xi = 0$ for some $c > 0$ and for any $\xi \in \mathbb{R}_0^N$ and $\eta \in \mathbb{R}_0^M$ (Hofbauer and Sigmund, 1998, p128-9). Note that if we multiply B and $\phi''(y)$ by the appropriate positive constant, c, and divide $R(y)$ by c, J is unchanged. However, now after this rescaling $C + C^T = 0$ and as Φ is still negative definite, $C + \lambda\Phi$ is negative definite and $R(C + \lambda\Phi)$ has all eigenvalues with real part negative.

The third case occurs if the perturbed equilibrium corresponds to an equilibrium which is pure for at least one of the players. Note that

$$\lim_{x_i \to 0} Q_{ij} = 0 \text{ for all } j.$$

This is a consequence of property 2 in the definition of ϕ in Section 2 and the definition $Q(x) = -(\phi'')^{-1}$. Suppose the perturbed equilibrium is located near enough a vertex of either S_N or S_M or both, (in which case either $Q(\hat{x})$ or $Q(\hat{y})$ or both are approximately zero) such that $\beta Q(\hat{\theta})C - I$ has all eigenvalues negative. Then the largest eigenvalue of QC, and indeed RC must be less than λ. If $\phi(x) = -\sum x_i \log x_i$ then the eigenvalues of $\lambda R\Phi$ are exactly $-\lambda$. In the case where $\phi(x) = \sum \log x_i$, as a vertex of S_N is approached then $R(x)\phi''(x)$ becomes arbitrarily large. ∎

Proof of Proposition 22. The dynamics are on $S_N \times S_M \times I\!R_+$. Taking the linearisation at a perturbed equilibrium point $(\hat{x}, \hat{y})$, which gives an equilibrium value $\hat{\mu} = \hat{x} \cdot A\hat{y} / \hat{y} \cdot B\hat{x}$, one obtains

$$K = \left(\begin{array}{cc|c} & J & \\ - - - - - - & - - - - - - & - - - - - - \\ \mu(Ay - \mu yB) & \mu(xA - \mu Bx) & -x \cdot Ay \end{array} \right),$$

where J is the Jacobian matrix derived in (47) above. Note that, writing $\theta = (x, y)$, the eigenvalue equation for the above matrix, that is, $K(\theta, \mu) = \chi(\theta, \mu)$, for some eigenvalue χ, can be decomposed into two separate equations, $J\theta = \chi\theta$, and $\mu(Ay - \mu yB)x + \mu(xA - \mu Bx) - x \cdot Ay\mu = \chi\mu$. Hence $N - 1 + M - 1$ of the eigenvalues of K are the eigenvalues of the matrix J. The remaining eigenvalue is therefore $-x \cdot Ay$. In conclusion, without rebalancing, at any perturbed equilibrium there is an additional negative eigenvalue. If J has any positive eigenvalues, K has too, and the perturbed equilibrium is unstable. If J has all negative eigenvalues, so does K. ∎

References

Benaïm, M. (1996) "A dynamical systems approach to stochastic approximations," *SIAM J. Control and Optimization*, **34**, 437-472.

Benaïm, M., Hirsch, M.W. (1996) "Learning processes, mixed equilibria and dynamical systems arising from repeated games," forthcoming, *Games Econ. Behav.*

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

Binmore, K., Samuelson, L. (1999) "Evolutionary drift and equilibrium selection," *Rev. Econ. Studies*, **66**, 363-393.

Börgers, T., Sarin, R. (1997) "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, **77**, 1-14.

Börgers, T., Sarin, R. (1996) "Naive reinforcement learning with endogenous aspirations," Working Paper, University College, London.

Camerer, C., Ho, T-H. (1999) "Experience-weighted attraction learning in normal form games," *Econometrica*, **67**, 827-874.

Cheung, Y-W., Friedman, D. (1997) "Individual learning in normal form games: some laboratory results," *Games Econ. Behav.*, **19**, 46-76.

Duffy J., Feltovich, N. (1999). "Does observation of others affect learning in strategic environments? An experimental study," *Int. J. Game Theory*, **28**, 131-152.

Erev, I., Rapoport, A. (1998). "Coordination, "magic", and reinforcement learning in a market entry game," *Games Econ. Behav.*, **23**, 146-175.

Erev, I., Roth, A.E. (1998). "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *Amer. Econ. Rev.*, **88**, 848-881.

Feltovich, N. (1999). "Reinforcement-base vs. belief-based learning models in experimental asymmetric-information games," forthcoming *Econometrica*.

Fudenberg, D., Kreps D. (1993). "Learning mixed equilibria," *Games Econ. Behav.*, **5**, 320-367.

Fudenberg, D., Levine D. (1998). *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gale, J., Binmore, K., Samuelson, L. (1995). "Learning to be imperfect: the ultimatum game," *Games Econ. Behav.*, **8**, 56-90.

Hofbauer, J., Hopkins, E. (1999). "Learning in perturbed asymmetric games," working paper.

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

Hopkins, E. (1999a). "Learning, matching and aggregation," *Games Econ. Behav.*, **26**, 79-110.

Hopkins, E. (1999b). "A note on best response dynamics," forthcoming *Games Econ. Behav.*

Huck S., Norman, H., Oechssler, J. (1999). "Learning in Cournot oligopoly–an experiment," *Econ. J.*, **109**, pp. C80-95.

McKelvey, R.D., Palfrey, T.R. (1995). "Quantal response equilibria for normal form games," *Games Econ. Behav.*, **10**, 6-38.

Norman, M.F (1968). "Some convergence theorems for stochastic learning models with distance diminishing operators," *J. Math. Psych.*, **5**, 61-101.

Ochs, J. (1995). "Simple games with unique mixed strategy equilibrium: an experimental study," *Games Econ. Behav.*, **10**, 202-217.

Posch, M. (1997). "Cycling in a stochastic learning algorithm for normal form games," *J. Evol. Econ.*, **7**, 193-207.

Roth, A.E., Erev, I. (1995). "Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term," *Games Econ. Behav.*, **8**, 164-212.

Rustichini, A. (1998). "Optimal properties of stimulus-response learning models," forthcoming *Games Econ. Behav.*.

Salmon, T. (1999). "An evaluation of econometric learning models of adaptive learning," working paper, California Institute of Technology.

Sarin, R., Vahid, F. (1998). "Predicting how people play games: a procedurally rational model of choice," Working Paper, Texas A&M University.

Van Huyck, J.B., Battalio, R.C., Rankin, F.W. (1997). "On the origin of convention: evidence from coordination games," *Econ. J.*, **107**, 576-596.

Vriend, N.J. (1997). "Will reasoning improve learning?" *Econ. Letters*, **55**, 9-18.