Population-level reinforcement learning resulting in smooth best response dynamics

David S. Leslie *,[1] E. J. Collins

University of Bristol

Abstract

Recent models of learning in games have attempted to produce individual-level learning algorithms that are asymptotically characterised by the replicator dynamics of evolutionary game theory. In contrast, we describe a population-level model which is characterised by the smooth best response dynamics, a system which is intrinsic to the theory of adaptive behaviour in individuals. This model is novel in that the population members are not required to make any game-theoretical calculations, and instead simply assess the values of actions based upon observed rewards. We prove that this process must converge to Nash distribution in several classes of games, including zero-sum games, games with an interior ESS, partnership games and supermodular games. A numerical example confirms the value of our approach for the Rock–Scissors–Paper game.

Key words: normal form games, population-level reinforcement learning, stochastic approximation, two-timescales

1 Introduction

Recent models of learning in games have attempted to produce individual-level learning algorithms that are asymptotically characterised by the replicator dynamics of evolutionary game theory (e.g. Börgers and Sarin, 1997). In contrast, we describe a population-level model which is characterised by the

* Corresponding author. Address: Department of Mathematics,University of Bristol, University Walk, Bristol, BS8 1TW, UK. Email: d.s.leslie@bristol.ac.uk. Telephone: +44 (0)117 954 6951.

[1] Research supported by CASE Research Studentship 00317214 from the UK Engineering and Physical Sciences Research Council in cooperation with BAE SYSTEMS.

smooth best response dynamics, a system which is intrinsic to the theory of adaptive behaviour in individuals.

Our results are based on the work of Hofbauer and Sandholm (2002), who study population level stochastic fictitious play. Their model requires all members of the population to make explicit game-theoretical calculations in order to update their strategy. In contrast, we introduce a model in which society learns the values of actions using simple reinforcement learning, then players update their strategies in a natural way based solely on these action values. By assuming that the rate at which the action values are learned is faster than the rate at which society adapts, we can use the two-timescales stochastic approximation results of Borkar (1997). We prove that this results in a system which can be characterised asymptotically by the single-population smooth best response dynamics, and so has the same convergence properties as stochastic fictitious play (Hofbauer and Sandholm, 2002). This complements recent work by Leslie and Collins (2002), who demonstrate a similar extension of individual-level ficitious play (Fudenberg and Kreps, 1993; Fudenberg and Levine, 1998; Benaïm and Hirsch, 1999).

1.1 Framework

We consider single population games (Hofbauer and Sigmund, 1998) which can be described by an $m \times m$ payoff matrix A. When two individuals meet each chooses an action from the set $\{1, \ldots, m\}$, and an individual choosing action a against an individual who chooses b receives reward $e_a^T A e_b$, where e_i is the unit vector with a 1 in the ith position.

Each individual plays a pure strategy, and the population state can be described by a unit vector π in which the ath element is the proportion of the population playing action a. Each time the game is played a proportion of the population change the action which they play, and so this population state evolves with time. In our model we consider the infinite population-size limit, so that π can take any value in the unit simplex Δ, consisting of all non-negative unit vectors with m elements. Given a population in state π, the expected reward to an individual playing action a against a random member of the population is $e_a^T A \pi$. The environment is such that at each discrete time step two individuals are picked at random from the population and play the game. All members of the population observe the outcome, i.e. which actions are played and the reward obtained as a result.

In stochastic fictitious play, these observations are used by the individuals to estimate the current state of the population, π_n. Each individual uses this estimate (which is the average of the observed past actions) to calculate the

expected reward, $e_a A\pi_n$, received for each action a against the current population state π_n, and updates their own strategy according to this information, resulting in a new population state π_{n+1}. This model therefore requires all individuals to have knowledge of the payoff matrix A, and of how to calculate the expected payoff for each action given an estimate of π_n.

In contrast, we show in Section 2 how the same observations can be used by individuals in such a way that the state of the society evolves in a comparable manner to stochastic ficititious play without requiring any knowledge of the payoff matrix, A, or of how to calculate expected rewards. Section 3 briefly reviews convergence results for the smooth best response dynamics, which give an asymptotic characterisation of the behaviour of both stochastic fictitious play and our new algorithm. Using this characterisation, we can identify several classes of games for which our algorithm will converge almost surely to equilibrium. Finally, in Section 4 we provide a numerical example which suggests that there may be games for which features of our model are neccessary for the population to converge to a steady state.

1.2 Related work

The method we use to estimate the values of actions is in the spirit of machine learning, and in particular reinforcement learning (Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996). Various authors (Claus and Boutilier, 1998; Crites and Barto, 1998; Erev and Roth, 1998; Littman and Stone, 2001) have applied this model to learning in games. Claus and Boutilier (1998) restrict attention to partnership games, or games of identical interest, for which a pure strategy equilibrium must exist. Littman and Stone (2001) point out several problems with this approach when it is applied to other types of game, and especially in cases where all the equilibria are mixed. In essence the problems arise because a player's estimates of the values of actions can not change sufficiently quickly to keep up with changing opponent strategies.

A complementary approach is that used by Börgers and Sarin (1997), which in turn is based upon earlier work in the field of learning automata (Narendra and Thathachar, 1989). Here the reward at each stage is used to directly update the mixed strategy to be played, as opposed to maintaining an estimate of action values. However, Börgers and Sarin (1997) show that in the long run play will converge almost surely to a pure strategy combination, even for games in which no pure strategy equilbrium exists.

In the field of Markov decision processes, there is an analogous problem of how to deal with both action values and strategies. Algorithms based upon classical value iteration are the basis for the value-based learning algorithms of

Claus and Boutilier (1998) and Littman and Stone (2001). Algorithms based upon policy iteration are related to Börgers and Sarin's algorithm (Börgers and Sarin, 1997). A hybrid scheme has proved popular in the field of Markov decision processes: actor–critic algorithms maintain separate estimates of the action values and the current optimal policy, using the former to update the latter towards optimality. Although these have been successfully used in empirical approaches for some time (see Barto *et al.*, 1983; Williams and Baird, 1990, and references therein), few theoretical results were available until recently. Konda and Borkar (2000) use a two-timescales stochastic approximation method (Borkar, 1997) and update the actor (the strategy) on a slower timescale than the critic (the value function), and prove convergence to an optimal strategy. We use a similar approach in this paper, resulting in a two-timescales process similar in flavour to the algorithm proposed by Borkar (2001) for learning in Markovian games.

2 A two-timescales algorithm

As stated previously, current models of learning in games frequently assume that the individuals explicitly calculate the expected values of actions, by observing the past behaviour of individuals in order to estimate the current population state and using this estimate to calculate the expected action values using the game matrix. A simple modification is to estimate the action values directly, instead of estimating the current population state. A standard method for this is reinforcement learning, or stimulus–response learning, originally proposed by Thorndike (1898) as a model of animal learning.

We assume that each individual forms identical beliefs about the value of the actions (since each population member observes the same information, namely the outcomes of previous games). The current estimates of action values are stored in a vector Q, which at stage n is updated after the game is played. Suppose two individuals are chosen to play the game; they choose actions a and b and receive rewards $R_n(a)$ and $R_n(b)$ respectively. The vector of Q values is updated according to the rule

$$Q_{n+1} = Q_n + \lambda_n \Big[e_a \left\{ R_n(a) - Q_n(a) \right\} + e_b \left\{ R_n(b) - Q_n(b) \right\} \Big], \qquad (1)$$

where the λ_n are a decreasing positive sequence of learning parameters. It is well known in the reinforcement learning literature that, under conditions on the decay rate of the λ_n, this scheme will necessarily converge in the limit to the correct action values if the population state π_n remains fixed and all actions are played with positive probability (i.e. $\pi_n = \pi \in \mathrm{int}(\Delta)\ \forall n$, where $\mathrm{int}(\Delta)$ is the interior of the simplex Δ).

However we also assume that simultaneously a fraction μ_n of the population are 'restless' and wish to change their strategy. The vector Q_n provides an estimate of the current expected value of each action, and it would seem reasonable to expect an individual to select the action which gives the maximal expected value, i.e. to choose a to maximise $Q_n(a)$. However it is now fairly standard, based upon the work of Harsanyi (1973) and McFadden (1973), not to model action choice in this way, since individuals may make mistakes, may decide they don't 'believe' what the Q vector tells them, or may deliberately try to second guess the evolutionary process. There are also technical difficulties involved in choosing a maximal action here, including:

(1) If the population is at an equilibrium $\tilde{\pi}$, each action a for which $\tilde{\pi}(a) > 0$ receives the same expected reward. Thus an individual will be equally happy to play any of these actions and the specific distribution resulting in equilibrium is unstable.
(2) The discontinuity inherent in choosing the maximal action means that the population does not adapt in a smooth manner.

To circumvent these difficulties, we therefore assume that each restless individual applies an independent random additive perturbation to each $Q_n(a)$, independently of the rest of the population; the new action is chosen to maximise the perturbed Q value. Under weak conditions on the distributions of these perturbations, Hofbauer and Sandholm (2002) show that the probability a restless player selects action a is $\beta(Q)(a)$, where $\beta(Q) \in \text{int}(\Delta)$ is the vector which maximises the quantity

$$\pi \cdot Q + v(\pi)$$

for some smoothing function v. Here v is a smooth, strictly differentiable, strictly concave function such that as π approaches the boundary of Δ the slope of v becomes infinite. A commonly used smoothing function is $v(\pi) = \tau \sum_a \pi(a) \log(\pi(a))$, which arises from perturbations with the extreme value distribution function $F(x) = \exp(-\exp(-\tau^{-1}x - \gamma))$, where γ is Euler's constant (Hofbauer and Sandholm, 2002). Using this form of perturbations,

$$\beta(Q)(a) \propto \exp(Q(a)/\tau), \tag{2}$$

giving the Boltzmann distribution.

So we see that the change in population state, given the current values Q_n, population state π_n, and that a fraction μ_n of the population switch action, is given by

$$\pi_{n+1} - \pi_n = \mu_n(\beta(Q_n) - \pi_n). \tag{3}$$

Note that since we assume an infinite population size a law of large numbers means that this update is essentially deterministic. We have two coupled

processes — the (stochastic) Q process (1) and the (deterministic) π process (3). Although theoretically possible to examine these processes using the ODE method of stochastic approximation (Kushner and Clark, 1978; Benaïm, 1999), the resultant dynamical system is difficult to analyse. A more direct approach is to use Borkar's theory of two-timescales stochastic approximation (Borkar, 1997):

Theorem 1 (Borkar) *Consider two coupled stochastic approximation processes*

$$\begin{aligned}\theta^{(1)}_{n+1} &= \theta^{(1)}_n + \lambda_n \left\{ F^{(1)}(\theta^{(1)}_n, \theta^{(2)}_n) + M^{(1)}_{n+1} \right\} \\ \theta^{(2)}_{n+1} &= \theta^{(2)}_n + \mu_n \left\{ F^{(2)}(\theta^{(1)}_n, \theta^{(2)}_n) + M^{(2)}_{n+1} \right\}\end{aligned}$$

where, for each i, $F^{(i)}$ is a globally Lipshitz continuous vector field, $\theta^{(i)}_n$ is bounded almost surely, and the partial sums $\left\{\sum_{n=0}^k \lambda_n M^{(1)}_{n+1}\right\}_k$ and $\left\{\sum_{n=0}^k \mu_n M^{(2)}_{n+1}\right\}_k$ converge almost surely. Suppose also that

$$\begin{aligned}&\sum_{n\geq 0} \lambda_n = \infty, \quad \sum_{n\geq 0} \lambda_n^2 < \infty, \\ &\sum_{n\geq 0} \mu_n = \infty, \quad \sum_{n\geq 0} \mu_n^2 < \infty, \\ &\mu_n/\lambda_n \to 0 \ \textit{as}\ n \to \infty,\end{aligned} \tag{4}$$

and that for each $\theta^{(1)}$ the ODE

$$\dot{Y} = F^{(2)}(\theta^{(1)}, Y) \tag{5}$$

has a unique globally asymptotically stable equilibrium point $\xi(\theta^{(1)})$ such that ξ is Lipschitz. Then, almost surely,

$$\|\theta^{(2)}_n - \xi(\theta^{(1)}_n)\| \to 0 \quad \textit{as} \quad n \to \infty,$$

and a suitable continuous time interpolation of $\{\theta^{(1)}_n\}_{\{n\geq 0\}}$ is an asymptotic pseudotrajectory (Benaïm, 1999) of the dynamical system

$$\dot{X} = F^{(1)}(X, \xi(X)). \tag{6}$$

Thus we can consider this as a 'fast' system (5) which sees the 'slow' system (6) as fixed, whereas the 'slow' system sees the 'fast' system as being fully calibrated to the unique fixed point $\xi(X)$ at all times. If the sequences $M^{(i)}_n$ are bounded martingale differences, i.e. $\mathbb{E}(M^{(i)}_{n+1} \mid \theta^{(1)}_n, \theta^{(2)}_n) = 0$ for $i = 1, 2$, then the partial sums $\left\{\sum_{n=0}^k \lambda_n M^{(1)}_{n+1}\right\}_k$ and $\left\{\sum_{n=0}^k \mu_n M^{(2)}_{n+1}\right\}_k$ will converge almost surely from the theory of martingales.

Benaïm (1999) shows that the limit set of an asymptotic pseudotrajectory of a flow is a connected internally chain recurrent set of the flow (see Benaïm (1999) for the definition of these concepts, but in Section 3 we identify classes of games for which any such set consists of the Nash distributions). We can therefore prove the following:

Theorem 2 *Assume that the perturbations applied to the payoffs have sufficiently smooth distribution so that β is globally Lipschitz continuous; also that the learning parameters λ_n and proportions of restless individuals μ_n satisfy (4). Then, almost surely as $n \to \infty$,*

$$\|Q_n - A\pi_n\| \to 0, \quad \textit{and} \tag{7}$$
$$\|\pi_n - \mathrm{CR}\| \to 0, \tag{8}$$

where CR *is the chain recurrent set of the single population smooth best response dynamics*

$$\dot{\pi} = \beta(A\pi) - \pi. \tag{9}$$

Proof Note that

$$\mathbb{E}(Q_{n+1} - Q_n) = 2\lambda_n \pi_n \cdot (A\pi - Q)\,,$$

and recall that

$$\pi_{n+1} - \pi_n = \mu_n(\beta(Q_n) - \pi_n).$$

Therefore taking

$$F^{(1)}(Q, \pi) = 2\pi \cdot (A\pi - Q)\,,$$
$$F^{(2)}(Q, \pi) = \beta(Q) - \pi,$$

we can apply Theorem 1 with bounded martingale differences $M_n^{(i)}$. The resulting differential equations are

$$\dot{Q} = 2\pi \cdot (A\pi - Q), \tag{10}$$
$$\dot{\pi} = \beta(Q) - \pi. \tag{11}$$

Now for fixed value of the slow parameter, π, the fast equation (10) has a globally asymptotically stable fixed point where $Q = A\pi$, i.e. the Q values converge to the true expected action values (note that this proves the "well known result" from reinforcement learning mentioned above). Thus (7) follows from Theorem 1, which also tells us that a suitable continuous time interpolation of the π_n is an asymptotic pseudotrajectory of the flow defined by (9).

However, results from Benaïm (1999) show that the the limit set of an asymptotic pseudotrajectory of a flow is precisely the chain recurrent set, CR, of the flow. Since the asymptotic pseudotrajectory is an interpolation of the discrete process, it is clear that the limit set of the discrete process is contained in that of the asymptotic pseudotrajectory, completing the proof.

So we see that the population state π_n will asymptote to the chain recurrent set of the smooth best response dynamics (9). In the following section we review results on these dynamics to give several classes of games for which the algorithm will converge to Nash distribution.

3 Single population smooth best response dynamics

Hofbauer and Sandholm (2002) provide relevant results for several classes of games:

Zero-sum games For these games the sum of the rewards to the two players is zero. In particular, given the symmetry of the games we consider, we require that A is skew-symmetric ($A^T = -A$).

Games with an interior ESS The concept of an evolutionarily stable strategy (ESS) originates in the work of Maynard Smith (1982), and is central to evolutionary game theory. In essence, an ESS is an equilibrium at which a small invasion of mutants will do less well than the current population. In terms of the payoff matrix A, an equilibrium $\pi^\star$ is an ESS if $\pi^{\star T} A\pi > \pi^T A\pi$ for all mixed strategies in a neighbourhood of $\pi^\star$. We call an ESS $\pi^\star$ interior if $\pi^\star \in \text{int}(\Delta)$.

Partnership games These are also sometimes called *potential games* or *games with identical interests*. Both players receive identical rewards — in our setting this requires $A^T = A$.

Supermodular games This is a class of games where players' actions are complementary. Assume there is an ordering on the actions, and that for $a' > a$, $b' > b$ we have $e_{a'}^T A e_{b'} - e_a^T A e_{b'} > e_{a'}^T A e_b - e_a^T A e_b$. That is, when one player moves to a 'higher' action the incentive for the other to also switch to a higher action increases.

For these classes of games the following is true:

Theorem 3 (Hofbauer and Sandholm) *Consider the smooth best response dynamics (9). The following is true:*

(1) for zero sum games there is a unique globally asymptotically stable fixed point π^* of (9), and $\text{CR} = \{\pi^*\}$,
(2) for games with an interior ESS there is a unique globally asymptotically stable fixed point π^* of (9), and $\text{CR} = \{\pi^*\}$,
(3) for partnership games the set E of fixed points of (9) is globally asymptotically stable, and if the fixed points are isolated then $\text{CR} = E$.
(4) for supermodular games there is an open dense set of initial conditions from which solution trajectories converge to the set of fixed points, with the remaining initial conditions contained in a finite or countable union

of invariant manifolds with codimension 1 (and so measure 0); if there is a unique fixed point π^* of (9) then $\mathrm{CR} = \{\pi^*\}$.

In fact it is shown, also by Hofbauer (2000), that the first two classes are special cases of the class of games for which

$$\xi^T A \xi \leq 0 \quad \forall \xi \in \mathbb{R}_0^m = \{\xi \in \mathbb{R}^n : \sum_i \xi_i = 0\}, \tag{12}$$

and that the conclusion for those two classes continues to hold for this wider class.

Since we know that the population state will converge to CR, the following is clear:

Theorem 4 *For the process defined in Section 2, the population state will converge almost surely to the set of Nash distributions in the following four classes of games:*

(1) zero-sum games,
(2) games with an internal ESS,
(3) partnership games with isolated Nash distributions,
(4) supermodular games with a unique Nash distribution.

Proof For all of these games, CR consists of fixed points of the dynamical system (9). But these fixed points are clearly just the points at which

$$\pi = \beta(A\pi),$$

i.e. the Nash distributions of the game. Therefore the result follows immediately from Theorem 2.

4 A numerical example

In this section we illustrate the need for the separation of the learning parameters, using a simple numerical example. The game considered is a simple roxk–scissors–paper game, with payoff matrix

$$\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

Boltzmann smoothing was used, with temperature parameter 0.2 ($\tau = 0.2$ in (2)). In each case the experiment was run for 5×10^5 iterations, starting at a

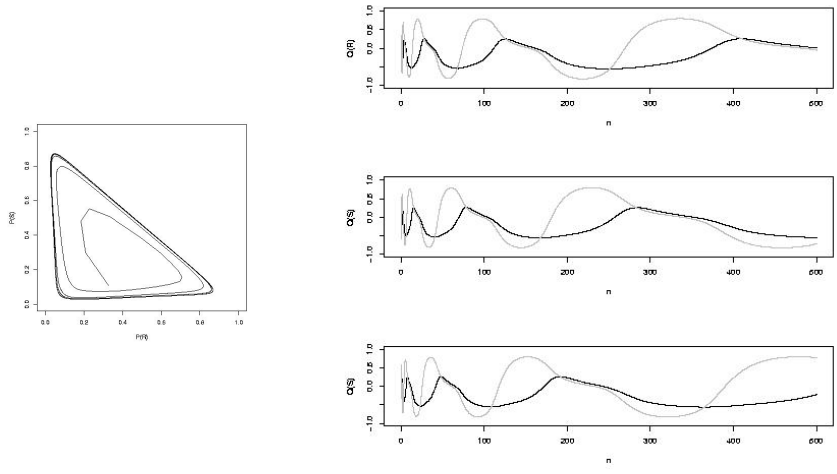

Fig. 1. Continuous clockwise cycling of strategies (left) and nonconvergence of Q values to $A\pi$ (right), using $\lambda_n = \mu_n = n^{-0.8}$.

random start point. Points were plotted every 50 iterations.

In the first experiment, the same decay rate was used for both λ_n and μ_n, resulting in a single-timescale stochastic approximation. The resuts are shown in Fig. 1. It is clear from the left hand diagram that the strategies cycle continuously in this case. Of more interest are the diagrams on the right, comparing the current estimates of the values, Q, with the calculated current value of the actions, $A\pi$. In each diagram the Q value is plotted in black while the calculated value is in grey. It is clear that the Q values cannot 'keep up' with the calculated values, which adjust at the same rate as the population state π. The 'stretching' of the process as n increases in these plots is due to the fact that as n increases the learning parameters decrease, and so the processes adjust more slowly.

For the second experiment we we introduce our second timescale, so that the Q values update on a faster timescale than the population state. There is a significant change in the diagrams. The left hand diagram in Fig. 2 shows a convergent trajectory (although after 5×10^5 iterations it has not yet fully converged). Again, of much greater interest is the right hand set of diagrams. In each case it is clear that the Q value tracks the calculated value $A\pi$ very closely

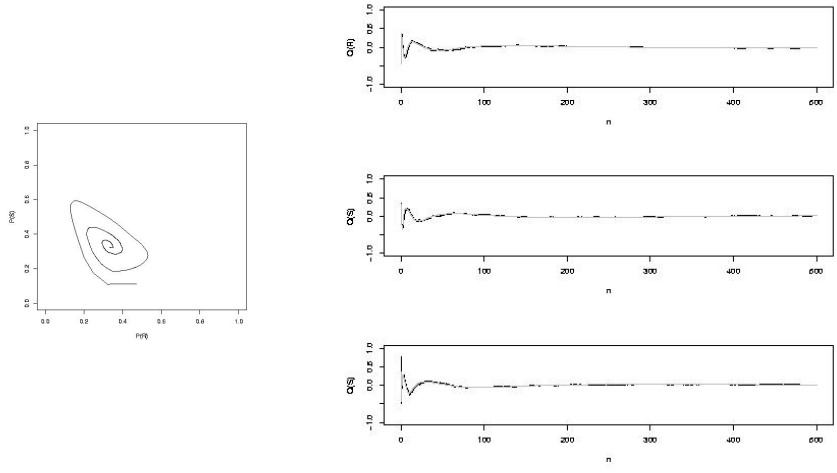

Fig. 2. Convergence of strategies in a clockwise spiral (left) and convergence of Q values to $A\pi$ (right), using $\lambda_n = n^{-0.7}$, $\mu_n = n^{-1}$.

indeed (again the Q value is plotted in black whereas the calculated value is in grey). This corresponds to the result (7), saying that $\|Q_n - A\pi_n\| \to 0$ as $n \to \infty$.

These two experiments suggest that for this particular game the use of two timescales is necessary to allow the Q values to successfully estimate $A\pi$. For simpler games, such as the Hawk–Dove game, without an inherent cycling of the strategies under the smooth best response dynamics, the use of two timescales proves to be unnecessary.

5 Conclusion

It seems unreasonable to expect agents in a population to perform explicit calculations in order to adapt their behaviour to that of the population. However most adaptive models assume that this is indeed the case.

We have used Borkar's two-timescales stochastic approximation theory to demonstrate a process which is asymptotically equivalent to the symmetric

stochastic fictitious play process studied by Hofbauer and Sandholm (2002).

Using the results from that paper on the smooth best response dynamics (9), we show that our process converges to Nash distribution in zero-sum games, games with an internal ESS, partnership games with isolated Nash distributions, and supermodular games with a unique Nash distribution.

Further, we have used a numerical experiment to show that, in the Rock–Scissors–Paper game, for the Q values to be a useful estimate of $A\pi$ requires the use of two-timescales stochastic approximation.

References

Barto, A. G., Sutton, R. S. and Anderson, C. W. (1983). "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. Systems Man Cybernet.* **SMC-13**, 834–846.

Benaïm, M. (1999). "Dynamics of stochastic approximation algorithms," in *Le Séminaire de Probabilités XXXIII*, Lecture Notes in Math. 1709, pp. 1–68. Berlin: Springer.

Benaïm, M. and Hirsch, M. W. (1999). "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games," *Games Econom. Behav.* **29**, 36–72.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming.* Belmont, MA: Athena Scientific.

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

Borkar, V. S. (1997). "Stochastic approximation with two time scales," *Systems Control Lett.* **29**, 291–294.

Borkar, V. S. (2002). Reinforcement learning in Markovian evolutionary games. Available at `http://www.tcs.tifr.res.in/~borkar/games.ps`.

Claus, C. and Boutilier, C. (1998). "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI-98 / IAAI-98 Proceedings*, 746–752. Cambridge, MA: AAAI Press.

Crites, R. H. and Barto, A. G. (1998). "Elevator group control using multiple reinforcement learning agents," *Machine Learning* **33**, 235–262.

Erev, I. and Roth, A. E. (1998). "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, **88**, 848–881.

Fudenberg, D. and Kreps, D. M. (1993). "Learning mixed equilibria," *Games Econom. Behav.* **5**, 320–367.

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

Harsanyi, J. (1973). "Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points," *Internat. J. Game Theory* **2**, 1–23.

Hofbauer, J. (2000). "From Nash and Brown to Maynard Smith: equilibria, dynamics, and ESS," *Selection* **1**, 81–88.

Hofbauer, J. and Sandholm, W. H. (2002). "On the global convergence of stochastic fictitious play," *Econometrica*, (forthcoming).

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

Konda, V. R. and Borkar, V. S. (2000). "Actor-critic-type learning algorithms for Markov decision process," *SIAM J. Control Opt.* **38**, 94–123.

Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer–Verlag.

Leslie, D. S. and Collins, E. J. (2002). "Convergent multiple-timescales reinforcement learning algorithms in normal form games," submitted to *Ann. Appl. Probab.*, Feb '02.

Littman, M. and Stone, P. (2001). "Leading best-response strategies in repeated games," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI '01)*. Morgan Kaufmann.

McFadden, D. (1973). "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics* (P. Zarembka, Ed.), pp. 105–142. New York: Academic Press.

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

Narendra, K. S. and Thathachar, M. A. L. (1989). *Learning Automata: An Introduction*. Englewood Cliffs, NJ: Prentice–Hall.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thorndike, E. (1898). "Some experiments on animal intelligence," *Science* **7**, 818–824.

Williams, R. J. and Baird, L. C. (1990). "A mathematical analysis of actor–critic architectures for learning optimal controls through incremental dynamic programming," *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*.