Convergent Multiple-Timescales Reinforcement Learning Algorithms in Normal Form Games

David S. Leslie*and E. J. Collins,
University of Bristol,
University Walk, Bristol, BS8 1TW
Email: {d.s.leslie, e.j.collins}@bristol.ac.uk

February 4, 2002

Abstract

We consider reinforcement learning algorithms in normal form games. Using two-timescales stochastic approximation we introduce a model-free algorithm which is asymptotically equivalent to smooth fictitious play, since both result in asymptotic pseudotrajectories to the flow defined by the smooth best response dynamics. Both of these algorithms are shown to converge almost surely to the Nash distribution in two-player zero-sum games and N-player partnership games. However there are simple games for which these, and most other adaptive processes, fail to converge — in particular we consider the N-player matching pennies game and Shapley's variant of the rock-scissors-paper game. By extending stochastic approximation results to multiple timescales we can allow each player to learn at a different rate. This extension will converge for two-player zero-sum games and two-player partnership games. It will also converge for the two special cases we consider.

1 Introduction

Current work in the theory of multiagent reinforcement learning has provided renewed impetus for the study of adaptive processes which evolve to equilibrium in general classes of normal form games. Recent developments in this area have used the theory of stochastic approximation to study the long term behaviour of adaptive processes in which players repeatedly play a normal form game and adjust their mixed strategies in response to the observed outcomes. This theory uses results from the theory of deterministic dynamical systems to gain information about the asymptotic behaviour of the stochastically evolving adaptive process.

One of the most generally applicable recent schemes for adaptive learning in games is smooth fictitious play, studied by Benaïm and Hirsch [3]. At each play of the game, each player estimates the expected reward to be obtained from each of their actions using knowledge of the game and their observations of the actions played by the other players in the

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

past; a mixed strategy based upon these estimates (a smooth best response) is then played. This results in a stochastic approximation process and it follows that the appropriate deterministic dynamical system is the smooth best response dynamics studied by Hopkins [13] and Hofbauer and Hopkins [12]. However these dynamics are only known to converge for rescaled two-player zero-sum games and rescaled two-player partnership games, and so Benaïm and Hirsch's smooth fictitious play algorithm is only proven to converge in these cases. Further, there are two particular games — Jordan's matching pennies game [15] and Shapley's variant of rock–scissors–paper [24] — for which it is known that the smooth best response dynamics have a unique equilibrium which is linearly unstable for certain smooth best response functions. Therefore in these simple cases the smooth fictitious play algorithm will almost surely not converge to the equilibrium point.

There is also a more general issue regarding Benaïm and Hirsch's algorithm: all players must observe the actions played by all other players, and also know the structure of the game (how many players are playing and the reward function) in order to calculate the expected values of their actions. In Section 2 we present a model-free multiagent reinforcement learning algorithm which also approximates the smooth best response dynamics, and so has the same convergence properties as smooth fictitious play. For this new algorithm it is not necessary for players to know anything about the game being played, nor to observe the opposition, nor even to know that they are playing a game at all. All that is required is for each player to observe the reward they obtain at each play of the game.

However the convergence properties of this algorithm are only as good as the convergence properties of the smooth best response dynamics which it approximates, and, as stated above, there are simple games for which these dynamics fail to converge. The basic technique used in the first algorithm allows us to aproximate any of the standard dynamical systems of adaptive game theory which are based on the expected values of actions; but the search for a continuous time dynamical system which converges to Nash equilibrium in general games has been largely unsuccessful, with the two 'difficult' games already mentioned causing problems for most (if not all) of the proposed dynamics. Section 3 generalises a result of Borkar [6]. This allows us to show, in Section 4, that a modification of our reinforcement learning algorithm approximates a singularly perturbed [14] variant of the smooth best response dynamics. Consequently we see in Sections 5 and 6 that this modified algorithm must converge in the two classes of games for which the standard smooth best response dynamics converge (two-player zero-sum, and two-player partnership games) as well as converging for the two difficult special cases previously mentioned.

1.1 Related work

Claus and Boutilier [8] have considered reinforcement learning algorithms in normal form games where all players receive identical rewards (partnership games). In their algorithm players estimate the value of their actions using standard reinforcement learning (see the books by Sutton and Barto [25] and Bertsekas and Tsitsiklis [4] for an introduction to these techniques). Players then use these estimates to choose a smooth best response, just as in smooth fictitious play. Littman and Stone [19] point out several problems with this approach when it is applied to other types of game (partnership games are qualitatively easier since pure strategy equilibria must exist). 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 [5], which in turn is based upon earlier work in the field of learning automata (see Narendra and Thathachar's book [21] for an introduction). 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. This however creates the problem that the only randomness in the algorithm comes from the strategy which is currently being updated, so as this gets close to a pure strategy there is very little noise in the system. Therefore it is possible for the algorithm to converge to a pure strategy combination which is not an equilibrium.

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 adaptive learning algorithms of Claus and Boutilier [8] and Littman and Stone [19]. Algorithms based upon policy iteration are related to Börgers and Sarin's algorithm [5]. 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 [1, 27] and references therein), few theoretical results were available until recently. Konda and Borkar [16] use a two-timescales stochastic approximation method [6] and update the actor (the strategy) on a slower timescale than the critic (the value function). Two further papers, by Sutton *et al.* [26] and Konda and Tsitsiklis [17], use a functional approximation of both the value function and the strategy. This has proved theoretically tractable, and convergence to a local maximum is proved.

We adapt the approach of Konda and Borkar [16] to provide the algorithms presented in this paper. Borkar [7] has applied his two-timescales stochastic approximation theory in a similar manner to Markovian games, adapting the value functions and strategies on different timescales. He shows convergence to 'generalised Nash equilibrium' in these games.

1.2 Preliminaries

We consider a game of N players, labelled $1, \ldots, N$. Each player $i \in 1, \ldots, N$ has a set A^i of available actions, one of which must be chosen each time the game is played. Together these action sets form the joint action set $\underline{A} = A^1 \times \ldots \times A^N$. When the game is played, each player chooses an action $a^i \in A^i$, resulting in a joint action $\underline{a} \in \underline{A}$. Each player receives a subsequent reward $r^i(\underline{a})$, where $r^i : \underline{A} \to \mathbb{R}$ is the reward function of player i.

As is standard in game theory we consider mixed strategies π^i for each player i, where $\pi^i \in \Delta(A^i)$, the set of probability distributions over the set A^i; in abuse of notation we write $\pi^i(a^i)$ for the probability that action a^i is played in mixed strategy π^i. This gives rise to a joint mixed strategy $\pi = (\pi^1, \ldots, \pi^N) \in \Delta(A^1) \times \ldots \times \Delta(A^N)$. There are unique multilinear extensions of the payoff functions to the mixed strategy space, and in standard abuse of notation we write

$$r^i(\pi) = \mathbb{E}(r^i(\underline{a}) \mid a^j \sim \pi^j, j = 1, \ldots, N) = \sum_{\underline{a} \in \underline{A}} \left(\prod_{j=1}^{N} \pi^j(a^j) \right) r^i(\underline{a}).$$

Given a joint mixed strategy $\pi = (\pi^1, \ldots, \pi^N)$ we define the *opponent joint strategy* $\pi^{-i} = (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^N)$, and identify the pair (π^i, π^{-i}) with the joint mixed strategy $(\pi^1, \ldots, \pi^{i-1}, \pi^i, \pi^{i+1}, \ldots, \pi^N)$. Also, in further standard abuse of notation, we identify a^i with the mixed strategy π^i for which $\pi^i(a^i) = 1$; this allows us to write (a^i, π^{-i}) for the joint mixed strategy where all players other than i play as if joint mixed strategy π is played, and player i uses the pure strategy a^i.

Using this notation we see that $r^i(a^i, \pi^{-i})$ is the expected reward to player i if action a^i is played against the opponent joint strategy arising from joint mixed strategy π. Nash [22] discusses equilibria of games — joint strategies where each player must play a strategy (pure or mixed) that is a best response to the opponent strategies. That is, at a Nash equilibrium $\tilde{\pi}$ we must have

$$r^i(\tilde{\pi}) = \max_{a^i \in A^i} r^i(a^i, \tilde{\pi}^{-i}) \quad \text{for each } i.$$

Nash shows [22] that every game must have an equilibrium. However difficulties arise when we try to use this maximisation in learning algorithms, including

1. At an equilibrium $\tilde{\pi}$, each action a^i for which $\tilde{\pi}^i(a^i) > 0$ receives the same expected reward $r^i(a^i, \tilde{\pi}^{-i})$. So if the value of actions is the only available information there is no motivation for a player to stay at the Nash equilibrium (as opposed to playing any of the pure strategies played with positive probability at the Nash equilibrium, or any other mixed strategy using only these actions).

2. The inherent discontinuity in taking a maximum means that strategies arising from sampling of observed rewards will rarely result in a mixed strategy being played. See [20] for further discussion of this issue.

To circumvent these difficulties we replace the absolute best response in the definition of a Nash equilibrium with a *smooth best response*. Here, instead of choosing actions to maximise $r^i(a^i, \pi^{-i})$, a distribution π^i is chosen to maximise

$$r^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i)$$

where $\tau > 0$ is a *temperature parameter* and $v^i : \Delta(A^i) \to \mathbb{R}$ is a player-dependent *smoothing function*, which is a smooth, strictly differentiable concave function such that as π^i approaches the boundary of $\Delta(A^i)$ the slope of v^i becomes infinite. This is the approach used by Fudenberg and Levine [10] and Hofbauer and Hopkins [12]. The conditions on v^i mean that there is a unique maximising π^i, so we can define the function

$$\beta^i(\pi^{-i}) = \operatorname*{argmax}_{\pi^i} \left\{ r^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i) \right\}.$$

Note that this can be written as

$$\beta^i(\pi^{-i}) = \operatorname*{argmax}_{\pi^i} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) r^i(a^i, \pi^{-i}) + \tau v^i(\pi^i) \right\},$$

and so the only use of the opponent joint strategy π^{-i} is in the assessment of the action values $r^i(a^i, \pi^{-i})$. Therefore if we have a vector Q of estimates of the action values we will often write

$$\beta^i(Q) = \underset{\pi^i}{\operatorname{argmax}} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) Q(a^i) + \tau v^i(\pi^i) \right\}, \tag{1}$$

and in this way

$$\beta^i(r^i(\cdot, \pi^{-i})) = \beta^i(\pi^{-i}).$$

It is clear that β^i will approximate the absolute best response as $\tau \to 0$, and β^i will approximate the uniform distribution over actions as $\tau \to \infty$.

Since absolute best responses are no longer relevant, the equilibria of a game under the assumption that players make smooth best responses are joint mixed strategies $\tilde{\pi}$ such that

$$\tilde{\pi}^i = \beta^i(\tilde{\pi}^{-i}).$$

Such points are called *Nash distributions*. Henceforth, 'convergence' of an algorithm is taken to mean convergence to Nash distribution under a fixed temperature parameter τ.

These smooth best responses can be viewed as a realisation of the incomplete games of Harsanyi [11]. He shows that the equilibria of a game are limit points of sequences of Nash distributions as the temperature parameter $\tau \to 0$. So when trying to learn the equilibria of a game it makes sense to consider smooth best responses with a small temperature parameter. This is the approach used by Benaïm and Hirsch when considering smooth fictitious play [3].

However the introduction of mixed strategies necessitates use of stochastic approximation theory. A good introduction to this area, and the approach we follow, is that of Benaïm [2] (this is a development of the ODE approach to stochastic approximation originally proposed by Kushner and Clark [18]). This general theory considers equations of the form

$$\theta_{n+1} = \theta_n + \lambda_n \left(F(\theta_n) + U_{n+1} \right), \tag{2}$$

where $\theta_n, U_{n+1} \in \mathbb{R}^m$, $F : \mathbb{R}^m \to \mathbb{R}^m$ and $\lambda_n \in \mathbb{R}_+$. We make the following generic assumptions throughout this paper:

G1 *F is a globally Lipschitz continuous vector field,*

G2 *the iterates θ_n are bounded, i.e.* $\sup_n \|\theta_n\| < \infty$,

G3 *the* learning parameters *decrease at a suitable rate:*

$$\textstyle\sum_{n \geq 0} \lambda_n = \infty, \quad \sum_{n \geq 0} {\lambda_n}^2 < \infty.$$

These assumptions naturally hold true in all of our applications, and are frequently necessary for the stochastic approximation theory to be valid. We use two main results from Benaïm's lecture notes [2], Propositions 4.1 and 4.2:

Proposition 1 (Benaïm) *Consider a stochastic approximation process (2). Let $t_n = \sum_{k=0}^{n-1} \lambda_k$ and define the interpolated process $\Theta : \mathbb{R}_+ \to \mathbb{R}^m$ by*

$$\Theta(t_n + s) = \theta_n + \frac{s}{t_{n+1} - t_n}(\theta_{n+1} - \theta_n) \quad \textit{for} \quad 0 \leq s < \lambda_n.$$

Assume that for all $T > 0$

$$\lim_{n\to\infty} \sup\left\{\left\|\sum_{l=n}^{k-1} \lambda_l U_{l+1}\right\| : k = n+1, \dots, m(t_n + T)\right\} = 0, \tag{3}$$

where $m(t) = \sup\{\kappa \geq 0 : t_\kappa \leq t\}$. Then Θ is an asymptotic pseudotrajectory of the flow φ induced by F.

Proposition 2 (Benaïm) *Consider a stochastic approximation process (2) for which*

1. *$\{\lambda_n\}_{n\geq 0}$ is a deterministic sequence,*
2. *$\{U_n\}_{n\geq 1}$ is adapted with respect to the σ-field $\mathcal{F}_n \overset{\text{def}}{=} \sigma(\theta_0, \dots, \theta_n)$,*
3. *$\mathbb{E}(U_{n+1} \mid \mathcal{F}_n) = 0$,*
4. *$\sup_n \mathbb{E}(\|U_{n+1}\|^2) < \infty$.*

Then the assumption (3) is true with probability 1.

Now an asymptotic pseudotrajectory of a flow φ in a metric space $(\mathcal{M}, d)$ is a continuous function X such that

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} d(X(t+h), \varphi_h(X(t))) = 0$$

for any $T > 0$. That is, the function X moves like φ, but is allowed an asymptotically vanishing amount of correction every T units of time. Adapting Proposition 5.3(iii) of [2] shows the following:

Proposition 3 (Benaïm) *If J is a global attractor for the flow φ and X is an asymptotic pseudotrajectory of φ then the limit set of X is contained in J.*

Further, define a point θ to be *attainable* if for each n and every open neighbourhood U of θ

$$\mathbb{P}(\exists m \geq n : \theta_m \in U) > 0.$$

Proposition 7.5 of [2] tells us:

Proposition 4 (Benaïm) *Consider a stochastic approximation process satisfying the conditions of Proposition 2, and let A be an attractor for the flow φ defined by the vector field F. Suppose the basin of attraction of A contains an attainable point. Then*

$$\mathbb{P}(\lim_{n\to\infty} d(\theta_n, A) = 0) > 0.$$

This shows that any attractor of the flow φ may contain the limit set of the stochastic approximation process, and a global attractor for the flow φ contains the limit set with probability 1. The following complementary result is provided by Pemantle [23]:

Theorem 5 (Pemantle) *Consider a stochastic approximation process (2) and let $p \in \mathbb{R}^m$ be a linearly unstable critical point of the vector field F. Let $\mathcal{N}$ be a neighbourhood of p and assume there are constants $\rho \in (1/2, 1]$ and $c_1, c_2, c_3, c_4 > 0$ such that whenever $\theta_n \in \mathcal{N}$*

1. *F has continuous first derivative,*
2. *$c_1/n^\rho \leq \lambda_n \leq c_2/n^\rho$,*
3. *$\|U_n\| \leq c_3$,*

4. $\mathbb{E}((U_{n+1} \cdot e)^+ | \mathcal{F}_n) \geq c_4$ *for every unit vector* $e \in \mathbb{R}^m$,

where $(U_n \cdot e)^+ = \max\{U_n \cdot e, 0\}$. *Then*

$$\mathbb{P}(\theta_n \to p) = 0.$$

Benaïm and Hirsch use these theorems to study smooth fictitious play. It turns out that the appropriate deterministic dynamical system is the smooth best response dynamics, given by

$$\dot{\pi}^i = \beta^i(\pi^{-i}) - \pi^i.$$

Thus the asymptotic behaviour of (stochastic) smooth fictitious play is closely related to the asymptotic behaviour of the (deterministic) smooth best response dynamics: attractors for these dynamics contain the limit set of the learning process with positive probability, and linearly unstable points contain the limit set with probability zero. It is this that shows that smooth fictitious play will not converge for certain combinations of temperature parameter and smoothing function — those combinations for which the unique Nash distribution is linearly unstable (as shown by Cowan [9] for Shapley's game and Benaïm and Hirsch [3] for Jordan's pennies game).

We are now in a position to extend these ideas to stochastic approximation algorithms with multiple timescales, and apply these extensions to develop model-free algorithms for learning in games.

2 A two-timescales learning algorithm

The motivation for this work is our observation that the only reason players need to know the structure of the game and observe opponent behaviour is so that they can estimate the expected value of each of their actions in order to calculate the smooth best response. Reinforcement learning is a model-free alternative for estimating expected values of a set of actions, although it relies on the fact that these expected values do not change with time.

Assume we have a stationary random environment where at each stage player i must choose an action a^i from a finite set A^i, and associated with each action $a^i \in A^i$ there is a random reward $R(a^i)$ which has a fixed distribution and bounded variation. Consider the learning scheme

$$Q_{n+1}(a^i) = Q_n(a^i) + \lambda_n I_{\{a^i_n = a^i\}}(R^i_n - Q_n(a^i)),$$

where a^i_n is the action chosen at stage n, R_n is the subsequent reward and $\{\lambda_n\}_{n\geq 0}$ is a deterministic sequence satisfying

$$\sum_{n\geq 0} \lambda_n = \infty, \quad \sum_{n \geq 0} {\lambda_n}^2 < \infty.$$

It is well-known in the reinforcement learning literature ([25], [4]) that, provided each action is chosen infinitely often, the Q values in this algorithm will converge almost surely to the expected action values, i.e.

$$Q_n(a^i) \to \mathbb{E}[R(a^i)] \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

However when we move to multiagent learning the players' strategies are all changing simultaneously as each player learns, and consequently the sampled rewards, R^i_n, do not come from a stationary distribution. So when learning $r^i(a^i, \pi^{-i})$ the standard results no longer apply. A solution is to be found in Borkar's two-timescales stochastic approximation [6]. We state a slight generalisation of Borkar's results which he obtains in the course of proving his main theorem.

Theorem 6 (Borkar) *Consider two coupled stochastic approximation processes*

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

where, for each i, $F^{(i)}$, $\theta^{(i)}_n$ and $\lambda^{(i)}_n$ satisfy the generic assumptions **G1–G3** *and $\sum_{n\geq 0} \lambda^{(i)}_n M^{(i)}_{n+1} < \infty$ almost surely. Further,*

$$\frac{\lambda^{(1)}_n}{\lambda^{(2)}_n} \to 0 \quad as \quad n \to \infty.$$

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

$$\dot{Y} = F^{(2)}(\theta^{(1)}, Y) \tag{4}$$

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

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

and a suitable continuous time interpolation of the process $\{\theta^{(1)}_n\}_{n\geq 0}$ is an asymptotic pseudotrajectory of the flow defined by the ODE

$$\dot{X} = F^{(1)}(X, \xi(X)). \tag{5}$$

This theorem says that if the 'fast' process, $\{\theta^{(2)}_n\}_{n\geq 0}$, converges to a unique limit point for any particular fixed value, $\theta^{(1)}$, of the 'slow' process, we can analyse the asymptotic behaviour of the algorithm as if the fast process is always fully 'calibrated' to the current value of the slow process. The "suitable continuous time interpolation" is given in the proof of the generalisation of this result in Section 3, but for application of this theorem it suffices to note that Proposition 3 tells us that the limit set of the stochastic approximation process $\theta^{(1)}_n$ is contained within any global attractor of the flow defined by (5).

Theorem 6 becomes very useful when we consider learning in games — provided the strategies change on a slower timescale than the timescale on which action values are learned then we can examine the asymptotic behaviour of the algorithm as if the action estimates are accurate. This is the basic technique which would, if required, allow us to approximate any of the standard dynamical systems of game theory which use estimates of action values. Our algorithm is as follows:

Two-timescales algorithm

For each player $i = 1, \ldots, N$,

$$\begin{aligned} \pi^i_{n+1} &= (1-\lambda_n)\pi^i_n + \lambda_n \beta^i(Q^i_n) \\ Q^i_{n+1}(a^i) &= Q^i_n(a^i) + \mu_n I_{\{a^i_n = a^i\}} \left\{ R^i_n - Q^i_n(a^i) \right\}. \end{aligned} \tag{6}$$

Here R_n^i is the reward obtained by player i at step n, and $\beta^i(Q_n^i)$ is the smooth best response (see equation (1)) given the value estimates Q_n^i. The sequences $\{\lambda_n\}_{n\geq 0}$ and $\{\mu_n\}_{n\geq 0}$ are each chosen to satisfy the condition **G3**, and the additional condition

$$\frac{\lambda_n}{\mu_n} \to 0 \quad \text{as} \quad n \to \infty.$$

Defining

$$\begin{aligned}
F^{(1)}(\pi, Q) &= \mathbb{E}((Q_{n+1} - Q_n)/\mu_n \,|\, \pi_n = \pi, Q_n = Q), \\
F^{(2)}(\pi, Q) &= \mathbb{E}((\pi_{n+1} - \pi_n)/\lambda_n \,|\, \pi_n = \pi, Q_n = Q) \\
&= \beta(Q_n) - \pi_n,
\end{aligned}$$

we apply Theorem 6. The implicitly defined $M_n^{(i)}$ of that theorem are martingale difference sequences, and so the condition $\sum_{n\geq 0} \lambda_n^{(i)} M_{n+1}^{(i)} < \infty$ follows immediately. We have already observed that the $Q_n^i(a^i)$ processes will converge to the true values of $r^i(a^i, \pi^{-i})$ if the strategies π^{-i} are fixed; indeed the ODE corresponding to (4) is simply

$$\dot{Q}^i(a^i) = \pi^i(a^i)(r^i(a^i, \pi^{-i}) - Q^i(a^i)),$$

which clearly has a globally asymptotically stable fixed point for fixed π so long as no strategy a^i has zero probability of being played. The other conditions of Theorem 6 are clearly met, and so we get the following:

Theorem 7 *For the two-timescales algorithm (6),*

$$\|Q_n^i(a^i) - r^i(a^i, \pi_n^{-i})\| \to 0 \quad \textit{as} \quad n \to \infty \quad \textit{a.s.}$$

and a suitable interpolation of the π_n^i processes will almost surely be an asymptotic pseudotrajectory of the flow defined by the smooth best response dynamics

$$\dot{\pi}^i = \beta^i(\pi^{-i}) - \pi^i.$$

So the asymptotic behaviour of the two-timescales algorithm (6) is characterised by the same dynamical system as characterises smooth fictitious play. Hofbauer and Hopkins [12] have studied these dynamics; they give a Lyapunov function for two-player zero-sum games, and also for two-player partnership games, hence showing that the set of Nash distributions is a global attractor in each case. Indeed, the Lyapunov function they give for partnership games is easily extended to N-player partnership games (in this case the function is $r(\pi) + \tau \sum_{i=1}^N v^i(\pi^i)$). So the smooth best response dynamics are also globally convergent for general partnership games. Hence we have shown the following:

Theorem 8 *The two-timescales algorithm (6) applied in* either

1. *a two-player zero-sum game,* or
2. *an N-player partnership game*

will converge with probability 1 to a Nash distribution.

On the other hand, Benaïm and Hirsch [3] show that for the 3-player matching pennies game [15] and certain values of the smoothing parameter τ the unique equilibrium is linearly unstable, and there exists a periodic orbit which is an attractor. Similarly, Cowan [9] shows that for the Shapley game [24] the smooth best response dynamics with Boltzmann smoothing admit a Hopf bifurcation as the parameter τ goes to zero, so that for small values of τ a limit cycle is again asymptotically stable and the unique equilibrium is unstable. It seems reasonable that an analogous result to Pemantle's (Theorem 5 above) should hold in this case, since there is noise present in the system. However the noise is only present on the fast timescale, so is of vanishing size with respect to the slow process where the instability of the equilibrium exists, and so the probabilistic estimates used by Pemantle are not valid in this case. The presence of an attracting orbit however means that by an extension of Proposition 4 (condition (24) of [2] is easily verified) the probability of convergence to the equilibrium is less then 1.

Despite these non-convergence results, the following is true:

Theorem 9 *If the two-timescales algorithm (6) converges to a fixed point*

$$(Q_n, \pi_n) \to (Q, \pi) \quad as \quad n \to \infty$$

then $Q^i(a^i) = r^i(a^i, \pi^{-i})$ *and* π *is a Nash distribution.*

Proof It is a basic result of stochastic approximation theory that if convergence occurs then the limit point must be a zero of the associated ODE. It follows immediately that $Q^i(a^i) = r^i(a^i, \pi^{-i})$, and $\beta^i(Q^i) = \pi^i$. Therefore $\pi^i = \beta^i(\pi^{-i})$.

3 Borkar's result extended to multiple timescales

The non-convergence of the two-timescales algorithm (6) in certain games motivates a further extension. Littman and Stone's work [19] suggests the consideration of players that learn at different rates. To consider this possibility we must extend Borkar's result [6] beyond two timescales.

Consider N interdependent stochastic approximation processes $\theta_n^{(1)}$, ..., $\theta_n^{(N)}$, which are updated according to the rules

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} + \lambda_n^{(i)} \left\{ F^{(i)} \left(\theta_n^{(1)}, \ldots, \theta_n^{(N)} \right) + M_{n+1}^{(i)} \right\}, \tag{7}$$

where, for each i, $F^{(i)}$, $\theta_n^{(i)}$ and $\lambda_n^{(i)}$ satisfy the generic assumptions **G1**–**G3** and $\sum_{n \geq 0} \lambda_n^{(i)} M_{n+1}^{(i)} < \infty$ almost surely. In addition we assume that

$$\frac{\lambda_n^{(i)}}{\lambda_n^{(j)}} \to 0 \quad \text{as} \quad n \to \infty \quad \text{whenever} \quad i < j.$$

This final assumption is what makes the algorithm multiple-timescale. Write $\theta_n = (\theta_n^{(1)}, \ldots, \theta_n^{(N)})$; in the sequel it will also be convenient to write $\theta_n^{(<i)}$ for the vector $(\theta^{(1)}, \ldots, \theta^{(i-1)})$.

As before, we define a timescale on which to interpolate the approximation processes. However we now follow Borkar [6] in establishing a different timescale corresponding to each process. For $i, j \in 1, \ldots, N$ let

$$t_n^{(j)} = \sum_{k=0}^{n-1} \lambda_k^{(j)},$$

let $\Theta^{(i,j)}(t)$ be the interpolation of the process $\theta_n^{(i)}$ on the jth timescale, i.e.

$$\Theta^{(i,j)}(t_n^{(j)} + s) = \theta_n^{(i)} + \frac{s}{t_{n+1}^{(j)} - t_n^{(j)}} (\theta_{n+1}^{(i)} - \theta_n^{(i)}) \quad \text{for} \quad 0 \leq s \leq \lambda_n^{(j)},$$

and let

$$m^{(j)}(t) = \sup\{\kappa \geq 0 : t_\kappa^{(j)} \leq t\}.$$

We start by considering the Nth timescale, and the interpolations on this timescale $\Theta^{(i,N)}(t)$. Rewrite the stochastic approximation processes (7) in the form

$$\begin{aligned}
\theta_{n+1}^{(i)} &= \theta_n^{(i)} + \lambda_n^{(N)} U_{n+1}^{(i,N)} \quad \text{for} \quad i < N, \\
\theta_{n+1}^{(N)} &= \theta_n^{(N)} + \lambda_n^{(N)} \left\{ F^{(N)}(\theta_n) + M_{n+1}^{(N)} \right\},
\end{aligned}$$

where for $i < N$ we have implicitly defined

$$U_{n+1}^{(i,N)} = \frac{\lambda_n^{(i)}}{\lambda_n^{(N)}} \left\{ F^{(i)}(\theta_n) + M_{n+1}^{(i)} \right\}.$$

For any n,

$$\begin{aligned}
&\sup \left\{ \left\| \sum_{l=n}^{k-1} \lambda_l^{(N)} U_{l+1}^{(i,N)} \right\| : k = n+1, \ldots, m^{(N)}(t_n^{(N)} + T) \right\} \qquad (8) \\
&\leq \sup_k \left\{ \left(\sum_{l=n+1}^{m^{(N)}(t_n^{(N)}+T)} \lambda_{l-1}^{(N)} \right) \left(\frac{\lambda_k^{(i)}}{\lambda_k^{(N)}} \right) F^{(i)}(\theta_k) + \left\| \sum_{l=n}^{k-1} \lambda_l^{(i)} M_{l+1}^{(i)} \right\| \right\}.
\end{aligned}$$

As $n \to \infty$ the second term converges to zero, by assumption. Also $\lambda_k^{(i)}/\lambda_k^{(N)} \to 0$ while the $F^{(i)}(\theta_k)$ are bounded and, from the definitions of $t_n^{(N)}$ and $m^{(N)}$, it should be clear that

$$\sum_{l=n+1}^{m^{(N)}(t_n^{(N)}+T)} \lambda_{l-1}^{(N)} \approx T.$$

Therefore the limit of the quantity (8) as $n \to \infty$ must be zero.

Taking $U_n^{(N,N)} = M_n^{(N)}$ we see that the equivalent limit in this case is also zero, and so we can use Theorem 1 to show that on this timescale the interpolated processes $\Theta^{(\cdot,N)}(t)$ are asymptotic pseudotrajectories for the flow defined by the differential equations

$$\begin{aligned}
\dot{X}^{(i)} &= 0 \quad \text{for} \quad i < N \qquad (9) \\
\dot{X}^{(N)} &= F^{(N)}(X) \qquad (10)
\end{aligned}$$

At this point we need to make the following assumption:

A(N) *There exists a Lipschitz continuous function $\xi^{(N)}(\theta^{(<N)})$ such that, for any $\theta^{(N)}$, solutions of the differential equations (9)–(10) converge to the point $(\theta^{(<N)}, \xi^{(N)}(\theta^{(<N)}))$ given initial conditions $(\theta^{(<N)}, \theta^{(N)})$.*

It therefore follows from Proposition 3 that the possible limit points of an asymptotic pseudotrajectory to the flow defined by equations (9)–(10) are the set of all points

$$(\theta^{(<N)}, \xi^{(N)}(\theta^{(<N)})),$$

where $\theta^{(<N)}$ can take any value. In other words

$$\left\| \theta_n - (\theta_n^{(<N)}, \xi^{(N)}(\theta_n^{(<N)})) \right\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

Now consider the timescale $t^{(N-1)}$, and the interpolations $\Theta^{(i,N-1)}(t)$ for $i < N$. Rewrite the stochastic approximation processes (7) in the form

$$\begin{aligned}
\theta_{n+1}^{(i)} &= \theta_n^{(i)} + \lambda_{n+1}^{(N-1)} U_{n+1}^{(i,N-1)} \quad \text{for} \quad i < N-1 \\
\theta_{n+1}^{(N-1)} &= \theta_n^{(N-1)} \\
&\quad + \lambda_{n+1}^{(N-1)} \left\{ F^{(N-1)}(\theta_n^{(<N)}, \xi^{(N)}(\theta^{(<N)})) + U_{n+1}^{(N-1,N-1)} \right\}
\end{aligned}$$

The implicit definition of $U_{n+1}^{(i,N-1)}$ for $i < N-1$ is equivalent to that of $U_{n+1}^{(i,N)}$, and so we can proceed as before. On the other hand we have implicitly defined

$$U_{n+1}^{(N-1,N-1)} = F^{(N-1)}(\theta_n) - F^{(N-1)}(\theta_n^{(<N)}, \xi^{(N)}(\theta^{(<N)})) + M_{n+1}^{(N-1)}.$$

However, we have already shown that as $n \to \infty$

$$\left\| \theta_n - (\theta_n^{(<N)}, \xi^{(N)}(\theta_n^{(<N)})) \right\| \to 0,$$

and we have assumed that $F^{(N-1)}$ is continuous, so

$$\left\| F^{(N-1)}(\theta_n) - F^{(N-1)}(\theta_n^{(<N)}, \xi^{(N)}(\theta_n^{(<N)})) \right\| \to 0.$$

Therefore when we take the sums $\sum_l \lambda_l^{(N-1)} U_{l+1}^{(N-1,N-1)}$ these terms will vanish as $n \to \infty$, as will the term $\sum_l \lambda_l^{(N-1)} M_{l+1}^{(N-1)}$, and so we see that on the $t^{(N-1)}$ timescale the interpolated processes $\Theta^{(<N,N-1)}(t)$ are an asymptotic pseudotrajectory of the flow defined by the differential equations

$$\begin{aligned}
\dot{X}^{(i)} &= 0 \quad \text{for} \quad i < N-1 && (11) \\
\dot{X}^{(N-1)} &= F^{(N-1)}(X^{(<N)}, \xi^{(N)}(X^{(<N)})) && (12)
\end{aligned}$$

We need to make an assumption analogous to **A(N)** above:

A(N-1) *There exists a Lipschitz continuous function $\xi^{(N-1)}(\theta^{(<N-1)})$ such that, for any $\theta^{(\geq N-1)}$, solutions of the differential equations (11)–(12) converge to the point $(\theta^{(<N-1)}, \xi^{(N-1)}(\theta^{(<N-1)}))$ given initial conditions $(\theta^{(<N-1)}, \theta^{(\geq N-1)})$.*

Defining

$$\Xi^{(\geq N-1)}(\theta^{(<N-1)}) = (\xi^{(N-1)}(\theta^{(<N-1)}), \xi^{(N)}(\theta^{(<N-1)}, \xi^{(N-1)}(\theta^{(<N-1)}))),$$

it follows that

$$\left\| \theta_n - (\theta_n^{(<N-1)}, \Xi^{(\geq N-1)}(\theta_n^{(<N-1)})) \right\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

We proceed recursively for each $j \geq 2$, noting that the interpolated processes $\Theta^{(\leq j,j)}$ are asymptotic pseudotrajectories of the flow defined by

$$\begin{aligned}
\dot{X}^{(i)} &= 0 \quad \text{for} \quad i < j && (13) \\
\dot{X}^{(j)} &= F^{(j)}(X^{(\leq j)}, \Xi^{(\geq j+1)}(X^{(\leq j)})) && (14)
\end{aligned}$$

For each $j \geq 2$ we need to make the assumption

A(j) *There exists a Lipschitz continuous function $\xi^{(j)}(\theta^{(<j)})$ such that, for any $\theta^{(\geq j)}$, solutions of the differential equations (13)–(14) converge to the point $(\theta^{(<j)}, \xi^{(j)}(\theta^{(<j)}))$ given initial conditions $(\theta^{(<j)}, \theta^{(\geq j)})$.*

Then defining

$$\Xi^{(\geq j)}(\theta^{(<j)}) = (\xi^{(j)}(\theta^{(<j)}), \Xi^{(\geq j+1)}(\theta^{(<j)}, \xi^{(j)}(\theta^{(<j)}))),$$

it follows that for $2 \leq j \leq N$

$$\left\| \theta_n - (\theta_n^{(<j)}, \Xi^{(\geq j)}(\theta_n^{(<j)})) \right\| \to 0 \quad \text{as} \quad n \to \infty \quad \text{a.s.}$$

Finally, it follows that on the slowest timescale the interpolated process $\Theta^{(1,1)}(t)$ is an asymptotic pseudotrajectory to the flow defined by

$$\dot{X}^{(1)} = F^{(1)} \left(X^{(1)}, \Xi^{(\geq 2)}(X^{(1)}) \right)$$

We have therefore proved the following theorem:

Theorem 10 *Consider a multiple-timescales stoschastic approximation process (7). If assumptions* **A(2)**–**A(N)** *hold then almost surely*

$$\|\theta_n^{(>1)} - \Xi^{(\geq 2)}(\theta_n^{(1)})\| \to 0 \quad \text{as} \quad n \to \infty$$

and a suitable continuous time interpolation of the process $\{\theta_n^{(1)}\}_{n \geq 0}$ *is an asymptotic pseudotrajectory of the flow defined by the ODE*

$$\dot{X} = F^{(1)}(X, \Xi^{(\geq 2)}(X))$$

4 A multiple-timescales learning algorithm

Theorem 10 allows us to consider a learning algorithm where the players learn at different rates. In fact we assume that all players update their strategies on strictly different timescales, and all of these timescales are slower than the rate at which the Q values are learned. The algorithm is as follows:

Multiple-timescales algorithm

For each player $i = 1, \ldots, N$,

$$\begin{aligned} \pi_{n+1}^i &= (1 - \lambda_n^i)\pi_n^i + \lambda_n^i \beta^i(Q_n^i), \\ Q_{n+1}^i(a^i) &= Q_n^i(a^i) + \mu_n I_{\{a_n^i = a^i\}} \left\{ R_n^i - Q_n^i(a^i) \right\}. \end{aligned} \tag{15}$$

As before, R_n^i is the reward obtained by player i at step n, and $\beta^i(Q_n^i)$ is the smooth best response given the value estimates Q_n^i. The sequences $\{\lambda_n^i\}_{n \geq 0}$ and $\{\mu_n\}_{n \geq 0}$ are each chosen to satisfy condition **G3**, and the additional conditions

$$\begin{aligned} \lambda_n^i / \mu_n &\to 0 \quad \text{as} \quad n \to \infty, \\ \lambda_n^i / \lambda_n^j &\to 0 \quad \text{as} \quad n \to \infty \quad \text{for} \quad i < j. \end{aligned}$$

This last condition says that each player is adapting their strategy on a different timescale (although all players still learn the Q values at the same fast timescale).

The first thing to note about this algorithm is that the same argument as for the two-timescales algorithm will suffice to show the following.

Theorem 11 *If the multiple-timescales algorithm (15) converges to a fixed point*

$$(Q_n, \pi_n) \to (Q, \pi)$$

then $Q^i(a^i) = r^i(a^i, \pi^{-i})$ and π is a Nash distribution.

However to use Theorem 10 we need to check that assumptions **A(2)**–**A(N)** are satisfied. We start by noting that the ODE

$$\dot{\pi}^N = \beta^N(\pi^1, \dots, \pi^{N-1}) - \pi^N$$

for fixed $(\pi^1, \dots, \pi^{N-1})$ has a globally attracting point, $\beta^N(\pi^{<N})$, so these assumptions may fail only for intermediate players that are not the fastest or slowest (no assumption need be made about the slowest timescale). We must make the following assumption about the behaviour of the ODEs for these intermediate timescales:

C *For each $i = 2, \dots, N-1$ there exists a Lipschitz function b^i such that $b^i(\pi^1, \dots, \pi^{i-1})$ is the globally asymptotically stable equilibrium point of the ODE*

$$\dot{\pi}^i = \beta^i\left(\pi^{<i}, B^{>i}(\pi^{\leq i})\right) - \pi^i$$

where we recursively define

$$\begin{aligned} B^{>(N-1)}(\pi^{\leq(N-1)}) &= \beta^N(\pi^{\leq(N-1)}) \\ B^{>i}(\pi^{\leq i}) &= (b^{i+1}(\pi^{\leq i}), B^{>(i+1)}(\pi^{\leq i}, b^{i+1}(\pi^{\leq i}))) \end{aligned}$$

Effectively this says that, for any i, if we fix the strategies for players $1, \dots, i$ then almost surely

$$\pi_n^{>i} \to B^{>i}(\pi^{\leq i}).$$

This convergence assumption is fairly restrictive, although it does not prevent the application of this algorithm to several different games (see Sections 5–6 below). It allows us to use Theorem 10 to characterise the asymptotic behaviour of the algorithm (15).

Theorem 12 *For the multiple-timescales algorithm (15) under the convergence assumption* **C***,*

$$\left\|(\pi_n^2, \dots, \pi_n^N) - B^{>1}(\pi_n^1)\right\| \to 0 \quad as \quad n \to \infty \quad a.s.$$

and a suitable continuous interpolation of the π_n^1 is an asymptotic pseudotrajectory of the flow defined by the ODE

$$\dot{\pi}^1 = \beta^1\left(B^{>1}(\pi^1)\right) - \pi^1$$

Proof Since the $Q_n^i(a^i) \to r^i(a^i, \pi^{-i})$ whenever π is fixed, the proof is immediate from our extension of Borkar's result to multiple-timescales and the assumption **C**.

This result means that to analyse the multiple-timescales algorithm in a particular game, or class of games, it suffices to show that our assumption **C** is satisfied and to analyse the behaviour of the slowest player under the assumption that all other players play the strategy dictated by the function $B^{>1}$.

Note that we can consider this system as relating to a multiple-timescales singular perturbation of the smooth best response dynamics:

$$\begin{aligned}
\dot{\pi}^1 &= \beta^1(\pi^{-1}) - \pi^1, \\
\dot{\pi}^2 &= \epsilon^{(2)}\left(\beta^2(\pi^{-2}) - \pi^2\right), \\
&\vdots \\
\dot{\pi}^N &= \epsilon^{(N)}\left(\beta^N(\pi^{-N}) - \pi^N\right),
\end{aligned}$$

with $\epsilon^{(i+1)} = o(\epsilon^{(i)})$ as $\epsilon^{(i)} \to 0$. Consideration of this system may indicate how to relax the convergence assumption **C**.

5 Two-player games

It is easy to see that for two-player games the assumption **C** is vacuous, since there are no intermediate players (each player is either the fastest or the slowest). Thus it is sufficient to analyse the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\pi^1)) - \pi^1 \tag{16}$$

We have a positive convergence theorem for two major classes of two-player games: zero-sum games and partnership games.

Theorem 13 *For both two-player zero-sum games and two-player partnership games the ODE (16) has a globally asymtotically stable attractor – the set of Nash distributions of the game.*

Proof For zero-sum games the function

$$U = r^1(\pi^1, \beta^2(\pi^1)) + \tau v^1(\pi^1) - \tau v^2(\beta^2(\pi^1))$$

is a Lyapunov function for the ODE (16).
For partnership games the function

$$V = r(\pi^1, \beta^2(\pi^1)) + \tau v^1(\pi^1) + \tau v^2(\beta^2(\pi^1))$$

is a Lyapunov function.

This gives rise to the following immediate corollary.

Corollary 14 *For both two-player zero-sum games and two-player partnership games the multiple-timescales algorithm (15) will converge a.s. to the set of Nash distributions.*

So we have asymptotic convergence results which are comparable to those for smooth fictitious play, and for our two-timescales algorithm (6). However a proof of convergence for general N-player partnership games is not available, since in this framework it is likely that for a fixed strategy of the slow players there will be several equilibria to which the fast players may converge, and so our assumption **C** will not be satisfied.

6 Some difficult games

There are some games which have consistently confounded attempts to learn the equilibrium. The two classic examples are the Shapley game [24], introduced in 1964 to show that classical fictitious play need not always converge, and the 3-player matching pennies game, a remarkably simple game introduced by Jordan [15] to show that, even with heavy prior assumptions focusing on the equilibrium point, a limit cycle could occur using simple learning. We start by proving convergence of our algorithm in a generalisation of the latter game, then show convergence of our algorithm for the Shapley game.

6.1 N-player matching pennies

Our generalisation of Jordan's game [15] is the N-player matching pennies game, in which each player can choose to play 'heads' (H) or 'tails' (T) and the reward to player i depends only on the actions a^i and a^{i+1}, where $i+1$ is calculated modulo N. The reward structure is

$$\begin{aligned} r^i(\underline{a}) &= I_{\{a^i = a^{i+1}\}} \quad \text{for} \quad i = 1, \ldots, N-1, \\ r^N(\underline{a}) &= I_{\{a^N \neq a^1\}}. \end{aligned}$$

The cyclical nature of this game allows the easy verification of our assumption **C**. As long as player 1's strategy is fixed then player N's strategy will converge to $\beta^N(\pi^{-N})$ since this only depends on π^1. Similarly, under the assumption that player one is fixed and player N has calibrated, it is clear that player $(N-1)$'s strategy will converge to $\beta^{N-1}(\pi^{-(N-1)})$, since this depends only on $\pi^N = \beta^N(\pi^{-N})$ which is fixed. This is repeated, so that whenever player 1's strategy is fixed the strategies of the faster players must converge to the unique best responses. By Theorem 12 it suffices to consider the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\ldots(\beta^N(\pi^1))\ldots)).$$

We assume that the smooth best responses are monotonic in the payoffs i.e. $r^i(a^i) > r^i(b^i) \Rightarrow \beta^i(r^i)(a^i) > \beta^i(r^i)(b^i)$. A sufficient condition for this to be the case is for each smoothing function v^i to be invariant under permutations of the actions. Thus if $\pi^1(H) > 1/2$ we must have $\beta^N(\pi^1)(H) < 1/2$ and so, in turn,

$$\beta^i(\beta^{i+1}(\ldots(\beta^N(\pi^1))\ldots))(H) < 1/2$$

for each $i = 1, \ldots, N$. So for $\pi^1(H) > 1/2$ it is the case that $\dot{\pi}^1(H) < 0$. Similarly if $\pi^1(H) < 1/2$ then $\dot{\pi}^1(H) > 0$, and so it follows that the Nash distribution $\pi^i(H) = 1/2$ is a global attractor.

We have shown that the multiscale algorithm (15) will converge almost surely to the Nash distribution of the matching pennies game provided that the players are ordered in the same way for the game as for the learning rates. In fact it is not difficult to see that this specific ordering is unnecessary, and any ordering of the players will suffice.

6.2 The Shapley game

This game is a variant of the traditional rock–scissors–paper game. It is a two-player game with three actions available to each player; the payoff matrix is

$$\left(\begin{array}{ccc} (0,0) & (1,0) & (0,1) \\ (0,1) & (0,0) & (1,0) \\ (1,0) & (0,1) & (0,0) \end{array}\right).$$

Thus each player gets a point if their opponent plays an action 1 greater (modulo 3) and gets no point otherwise. Without loss of generality (due to the symmetry of the game) we assume that player 1 is the slower, and since it is a two-player game our assumption **C** is irrelevant (as observed previously). So we simply need to analyse the ODE

$$\dot{\pi}^1 = \beta^1(\beta^2(\pi^1)) - \pi^1. \tag{17}$$

Note $\pi^1(3) = 1 - \pi^1(1) - \pi^2(2)$, so that this defines a planar flow. Therefore we calculate the divergence of the flow in $(\pi^1(1), \pi^1(2))$-space — if this is negative then the solutions of the ODE must converge to equilibrium.

For simplicity we assume that both smooth best responses are defined by the Boltzmann distribution, where we take as our smoothing function

$$v^i(\pi^i) = -\sum_{a^i} \pi^i(a^i) \log \pi^i(a^i).$$

Consequently

$$\beta^i(r^i)(a^i) = \frac{e^{r^i(a^i)/\tau}}{\sum_{b^i \in A^i} e^{r^i(b^i)/\tau}}.$$

For this game, dropping the superscripts on actions, $r^i(a) = \pi^{-i}(a+1)$ and so for any opponent distribution π^{-i} it follows that

$$\beta^i(\pi^{-i})(a) = \frac{e^{\pi^{-i}(a+1)/\tau}}{\sum_{a' \in A} e^{\pi^{-i}(a')/\tau}}.$$

We can assume that $\pi^2 = \beta^2(\pi^1)$, so defining $\rho(a) = \left(\pi^1(a) - \pi^1(3)\right)/\tau$ for $a = 1, 2$ it is clear that

$$\pi^2 = \frac{1}{1 + e^{\rho(1)} + e^{\rho(2)}} (e^{\rho(2)}, 1, e^{\rho(1)}). \tag{18}$$

By the chain rule applied to (17),

$$\text{Div} = \sum_{a=1}^{2} \frac{\partial \dot{\pi}^1(a)}{\partial \pi^1(a)} = \sum_{a=1}^{2} \sum_{a'=1}^{3} \sum_{b=1}^{2} \frac{\partial \beta^1(\pi^2)(a)}{\partial \pi^2(a')} \frac{\partial \pi^2(a')}{\partial \rho(b)} \frac{\partial \rho(b)}{\partial \pi^1(a)} - 2,$$

so to calculate the value of this sum we first calculate the component partial derivatives:

$$\begin{aligned} \frac{\partial \beta^1(\pi^2)(a)}{\partial \pi^2(a')} &= \frac{e^{\pi^2(a')/\tau} \left(I_{\{a'=a+1\}} \sum_{b' \in A} e^{\pi^2(b')/\tau} - 1 \right)}{\tau \left(\sum_{b' \in A} e^{\pi^2(b')/\tau} \right)^2}, \\ \frac{\partial \pi^2}{\partial \rho(1)} &= \frac{e^{\rho(1)}}{\left(1 + e^{\rho(1)} + e^{\rho(2)}\right)^2} \left(-e^{\rho(2)}, \quad -1, \; 1 + e^{\rho(2)} \right), \\ \frac{\partial \pi^2}{\partial \rho(2)} &= \frac{e^{\rho(2)}}{\left(1 + e^{\rho(1)} + e^{\rho(2)}\right)^2} \left(1 + e^{\rho(1)}, \; -1, \quad -e^{\rho(1)} \right), \\ \frac{\partial \rho(b)}{\partial \pi^1(a)} &= (1 + I_{\{a=b\}})/\tau, \end{aligned}$$

where the last derives from the fact that $\pi^1(3) = 1 - \pi^1(1) - \pi^1(2)$ and so

$$\rho(1) = (2\pi^1(1) + \pi^1(2) - 1)/\tau, \quad \rho(2) = (\pi^1(1) + 2\pi^1(2) - 1)/\tau.$$

Substituting all of these into the expression for the divergence, we get that

$$\begin{aligned} &\tau^2 \left(\sum_{a=1}^{3} e^{\pi^2(a)/\tau} \right)^2 \left(1 + e^{\rho(1)} + e^{\rho(2)}\right)^2 \times (\mathrm{Div} + 2) \\ = \quad & e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} (e^{\rho(1)} e^{\rho(2)} - 2e^{\rho(1)} - 2e^{\rho(2)}) \\ & + e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} (e^{\rho(2)} - 2e^{\rho(1)} - 2e^{\rho(1)} e^{\rho(2)}) \\ & + e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} (e^{\rho(1)} - 2e^{\rho(2)} - 2e^{\rho(1)} e^{\rho(2)}) \end{aligned}$$

Recalling the expression (18) for π^2, this shows that

$$\begin{aligned} &\tau^2 \left(\sum_{a=1}^{3} e^{\pi^2(a)/\tau} \right)^2 \times (\mathrm{Div} + 2) \\ =& e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} \left\{ \pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) \right\} \qquad (19) \\ &+ e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(1)\pi^2(2) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(3) \right\} \\ &+ e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\} \end{aligned}$$

This expression is invariant under the permutation of actions $(1, 2, 3) \to (3, 1, 2)$, so without loss of generality we can assume $\pi^2(1) \leq \pi^2(3)$ and $\pi^2(2) \leq \pi^2(3)$. Initially we assume further that $\pi^2(1) \leq \pi^2(2) \leq \pi^2(3)$, so that

$$\begin{aligned} \pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) &< 0, \\ \pi^2(1)\pi^2(2) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(3) &< 0. \end{aligned}$$

If $\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) < 0$ we are done. Otherwise

$$\begin{aligned} & e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\} \\ \leq \quad & e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ \pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3) \right\}, \end{aligned}$$

and the expression in (19) is bounded above by

$$\begin{aligned} & e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} \left\{ \pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) \right\} \\ & \quad + e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \left\{ -\pi^2(1)\pi^2(2) - \pi^2(2)\pi^2(3) - 4\pi^2(1)\pi^2(3) \right\}, \end{aligned}$$

which is clearly negative. A similar argument works with the assumption $\pi^2(2) \leq \pi^2(1) \leq \pi^2(3)$, and so the expression in (19) is always negative. This shows that

$$\mathrm{Div} = \sum_{a=1}^{2} \frac{\partial \dot{\pi}^1(a)}{\partial \pi^1(a)} \leq -2.$$

Since we have a planar flow with negative divergence the system must converge to a fixed point; there is a unique fixed point, at the Nash distribution [9], so this point must be globally attracting. Therefore from Theorem 12 it follows that the learning algorithm (15) will converge with probability 1 to the Nash distribution of the Shapley game.

7 Conclusion

Using Borkar's theory of two-timescales stochastic approximation, we have demonstrated a model-free multiagent reinforcement learning algorithm which will converge with probability 1 in repeated normal form games whenever the same claim can be made of smooth fictitious play [3]. This is because the asymptotic behaviour of both algorithms can be shown to be characterised by the asymptotic behaviour of the flow induced by the smooth best response dynamics. In particular both algorithms will converge with probability 1 for two-player zero-sum games and for N-player partnership games, since for these classes of games the set of Nash distributions is a global attractor for these dynamics.

However there are simple games for which the smooth best response dynamics have attractors outwith the set of Nash distributions, with the unique Nash distribution being linearly unstable. For these games convergence to Nash distribution does not necessarily occur and so an improvement can be gained by extending Borkar's stochastic approximation results to give an algorithm where all players learn at a different rate. Although we showed that if the algorithm converges in any game then it must have converged to a Nash distribution, further theoretical convergence results for this algorithm only apply for games in which our convergence assumption **C** holds. This assumption is true for all two-player games and for cyclical games such as the N-player matching pennies game, but fails when we consider N-player partnership games (since faster players may have several possible attracting points for fixed strategies of the slower players).

The multiple-timescales algorithm has been proven to converge to Nash distribution with probability 1 for two-player zero-sum games and two-player partnership games, as well as for the Shapley game [24] and the N-player matching pennies game — these latter two games having caused problems for all algorithms previously known to the authors. However a general convergence theorem for two-player games has proved elusive, despite the general applicability of the multiscale algorithm for these games.

In fact it is easy to see that a further extension of our algorithm is asymptotically equivalent to the original. In this extension we additionally allow each player to learn their Q^i values at a different rate. All that is required is that no player is 'reckless', in that each must learn the values Q^i on a faster timescale than they adjust towards the smooth best response $\beta^i(Q^i)$ to these values. Since each player's values Q^i only directly affect their own strategy, π^i, the assumptions of Theorem 10 continue to hold in the same cases as when all players learn their values at the same rate, and the algorithm will behave exactly as before. So a collection of players need have no communication before interacting, so long as none are reckless and all have different rates with which to adjust their strategy (achievable, for instance, by insisting that players choose a decay rate $\rho^i \in (0.5, 1]$ using an atomless distribution and setting $\lambda^i_n = n^{-\rho^i}$).

Acknowledgements

The authors would like to thank Andy Wright (BAE SYSTEMS) for several useful discussions in the preparation of this paper, during his tenure of a Royal Society Industrial Fellowship at the University of Bristol.

References

[1] 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.

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

[3] 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.

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

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

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

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

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

[9] Cowan, S. (1992). *Dynamical Systems Arising from Game Theory.* PhD thesis, University of California, Berkeley.

[10] Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games.* MIT Press.

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

[12] Hofbauer, J. and Hopkins, E. (2002). Learning in perturbed asymmetric games. Available at `http://www.econ.ed.ac.uk/pdf/perturb.pdf`.

[13] Hopkins, E. (1999). A note on best response dynamics. *Games Econom. Behav.* **29**: 138–150.

[14] Jones, C. K. R. T. (1995). Geometric singular perturbation theory. *Dynamical Systems*: 44–118. Lecture Notes in Math. 1609. Springer, Berlin.

[15] Jordan, J. S. (1993). Three problems in learning mixed strategy equilibria. *Games Econom. Behav.* **5**: 368–386.

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

[17] Konda, V. R. and Tsitsiklis, J. N. (2001) Actor–critic algorithms. Submitted to *SIAM J. Control Opt.* February 2001.

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

[19] 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)*, 2001.

[20] McNamara, J. M., Webb, J. N., Collins, E. J., Székely, T. and Houston, A. I. (1997). A general technique for computing evolutionarily stable strategies based on errors in decision-making. *J. Theoret. Biol.* **189**: 211–225.

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

[22] Nash, J. (1951). Non-cooperative games. *Ann. of Math.* **54**: 286–295

[23] Pemantle, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.* **18**: 698–712.

[24] Shapley, L. S. (1964) Some topics in two person games. *Advances in Game Theory*: 1–28. Princeton University Press, Princeton, NJ.

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

[26] Sutton, R. S., McAllester, D., Singh, S. and Mansour, Y. (2000) Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems 12 (Proceedings of the 1999 conference)*: 1057–1063.

[27] 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*.