The Annals of Applied Probability
2003, Vol. 13, No. 4, 1231–1251

CONVERGENT MULTIPLE-TIMESCALES REINFORCEMENT LEARNING ALGORITHMS IN NORMAL FORM GAMES

BY DAVID S. LESLIE[1] AND E. J. COLLINS

University of Bristol

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 the smooth fictitious play algorithm, in that 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 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. We show that this extension will converge for two-player zero-sum games and two-player partnership games, as well as 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 behavior 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 behavior of the stochastically evolving adaptive process.

One of the most generally applicable recent schemes for adaptive learning in games is smooth fictitious play, introduced by Fudenberg and Kreps [7] and studied in greater generality by Benaïm and Hirsch [2]. Players observe the actions played by their opponents and, by (incorrectly) assuming that opponent mixed strategies are not changing, estimate the current value of each of their actions based upon knowledge of the game; a mixed strategy based upon these estimates (a smooth best response) is then played. Benaïm and Hirsch used the theory of stochastic approximation to show that the asymptotic behavior of this process

Received February 2002; revised July 2002.

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

AMS 2000 subject classifications. 68T05, 91A20.

Key words and phrases. Stochastic approximation, reinforcement learning, repeated normal form games, best response dynamics.

is characterized by the smooth best response dynamics studied by Hopkins [11] and Hofbauer and Hopkins [10]. However, the smooth fictitious play algorithm assumes that all players observe the actions of all other players and also know the structure of the game (how many players are playing and the reward function). 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, to observe the opposition or 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 both algorithms are only as good as the convergence properties of the smooth best response dynamics. While Hofbauer and Hopkins [10] proved convergence to Nash distribution in rescaled two-player zero-sum and rescaled two-player partnership games, there are two *difficult* games for which there are nonconvergent trajectories of the smooth best response dynamics. Indeed most (if not all) proposed adaptive dynamics in games have nonconvergent trajectories for these two games: Jordan's matching pennies game [13] and Shapley's variant of rock–scissors–paper [19]. Section 3 generalizes a result of Borkar [4], allowing us to show, in Section 4, that a modification of our reinforcement learning algorithm approximates a singularly perturbed [12] 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 our two difficult special cases.

1.1. *Preliminaries.* We consider a game of N players, labelled $1, \dots, N$. Each player $i \in 1, \dots, N$ 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 $\underline{A} = A^1 \times \cdots \times A^N$. 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 $\pi = (\pi^1, \dots, \pi^N) \in \Delta(A^1) \times \cdots \times \Delta(A^N)$. There are unique multilinear extensions of the payoff functions to the mixed strategy space, and in standard abuse of notation we write

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

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. The standard solution concept of noncooperative game theory is the Nash equilibrium [17]—joint strategies $\tilde{\pi}$, where each player plays a best response to the opponent strategies, so that

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

As discussed in Chapter 4 of [8], the discontinuities inherent in this maximization present difficulties for adaptive processes. Instead, as in [10], we follow Fudenberg and Levine in assuming that players choose a strategy π^i to maximize

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

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. The conditions on v^i mean that, for each π^{-i}, there is a unique maximizing π^i, so we can define the *smooth best response function*

$$\begin{aligned}\beta^i(\pi^{-i}) &= \arg\max_{\pi^i} \{r^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i)\} \\ &= \arg\max_{\pi^i} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) r^i(a^i, \pi^{-i}) + \tau v^i(\pi^i) \right\}.\end{aligned}$$

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

$$\beta^i(Q) = \arg\max_{\pi^i} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) Q(a^i) + \tau v^i(\pi^i) \right\}. \tag{1}$$

It is clear that β^i will approximate the absolute best response as $\tau \to 0$ and will approximate $\arg\max_{\pi^i} v^i(\pi^i)$ 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.

Harsanyi [9] showed 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 [2].

However, the introduction of mixed strategies necessitates use of stochastic approximation theory. The approach we follow is that of Benaïm [1] (this is a development of the ordinary differential equation (ODE) approach to stochastic approximation originally proposed by Kushner and Clark [15]). This general theory considers equations of the form

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

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:

ASSUMPTION G1. F is a globally Lipschitz continuous vector field.

ASSUMPTION G2. The iterates θ_n are bounded, that is, $\sup_n \|\theta_n\| < \infty$.

ASSUMPTION G3. The *learning parameters* decrease at a suitable rate:

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

These assumptions naturally hold true in all of our applications and are frequently necessary for the stochastic approximation theory to be valid. Benaïm [1] related the asymptotic behavior of such processes to that of deterministic dynamical systems and, in particular, to asymptotic pseudotrajectories:

DEFINITION 1. Let $\varphi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ be a flow on the metric space $(\mathcal{M}, d)$. An *asymptotic pseudotrajectory* to φ is a continuous function $X : \mathbb{R} \to \mathcal{M}$ such that

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

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. The following result, Proposition 4.1 of Benaïm [1], provides the main link between stochastic processes (2) and deterministic dynamical systems:

PROPOSITION 2 (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*

$$\Theta(t_n + s) = \theta_n + \frac{s}{t_{n+1} - t_n}(\theta_{n+1} - \theta_n) \qquad \text{for } 0 \le s < \lambda_n.$$

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

Benaïm gave easily verified sufficient conditions for (3) to be satisfied. Under these conditions, any limit set of the stochastic process (2) will be contained in the limit set of the asymptotic pseudotrajectory defined in the proposition. We use the following concepts from the theory of dynamical systems:

DEFINITION 3. Let $\varphi : \mathbb{R} \times \mathcal{M} \to \mathcal{M}$ be a flow on the metric space $(\mathcal{M}, d)$.

1. A fixed point $x^\star$ of φ is *globally attracting* if $\varphi_t(x) \to x^\star$ as $t \to \infty$ for all x.
2. A continuous function $V : \mathcal{M} \to \mathbb{R}$ is *a strict Lyapunov function* for the flow φ if $V(\varphi_t(x))$ is strictly increasing in t whenever x is not a fixed point of φ.

Corollaries 5.4 and 6.6 of [1] give two situations under which the limit sets of asymptotic pseudotractories to a flow, and hence the limit sets of stochastic approximation processes, will coincide with the set of fixed points. We restate these results here:

PROPOSITION 4 (Benaïm). *Let* φ *be a flow with set of fixed points* L *and let* X *be an asymptotic pseudotrajectory of* φ. *Let* Λ *denote the limit set of the asymptotic pseudotrajectory* X.

1. *If* $L = \{x^\star\}$ *and* $x^\star$ *is a globally attracting fixed point, then* $\Lambda = \{x^\star\}$.
2. *If the set* L *of fixed points of* φ *is countable and there exists a strict Lyapunov function for the flow, then* $\Lambda \in L$.

Further results exist in the literature, which are relevant when the behavior of the flow φ arising from a stochastic approximation process (2) is more complex. Define a nonempty, compact, invariant set A to be an attractor if it has a neighborhood N such that $\lim_{t\to\infty} d(\varphi_t(x), A) = 0$ for $x \in N$; Benaïm [1] showed (under mild conditions) that, for an attractor A, $\mathbb{P}(\lim_{n\to\infty} d(\theta_n, A) = 0) > 0$. Conversely, Pemantle [18] showed (under more severe conditions) that for a linearly unstable fixed point $\theta^\star$ of the flow, $\mathbb{P}(\theta_n \to \theta^\star) = 0$. Benaïm and

Hirsch [2] used these results to study smooth fictitious play. They showed 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 behavior of (stochastic) smooth fictitious play is closely related to the asymptotic behavior 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 0. It follows 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 [6] for Shapley's game and Benaïm and Hirsch [2] 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 behavior is so they can estimate the expected value of each of their actions and thus 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

$$Q_{n+1}(a^i) = Q_n(a^i) + \lambda_n I_{\{a_n^i = a^i\}}\big(R_n^i - Q_n(a^i)\big),$$

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

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

It is well known in reinforcement learning [3, 20], that, provided each action is chosen infinitely often, the Q values in this algorithm will converge almost surely to the expected action values, that is,

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

However, when we move to multiagent learning the players' strategies are all changing simultaneously as each player learns and, consequently, the sampled

rewards, R_n^i, 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 [4], and its use by Konda and Borkar [14] and Borkar [5]. We state a slight generalization of Borkar's results which he obtained in the course of proving his main theorem.

THEOREM 5 (Borkar). *Consider two coupled stochastic approximation processes*

$$\theta_{n+1}^{(1)} = \theta_n^{(1)} + \lambda_n^{(1)}\{F^{(1)}(\theta_n^{(1)}, \theta_n^{(2)}) + M_{n+1}^{(1)}\},$$
$$\theta_{n+1}^{(2)} = \theta_n^{(2)} + \lambda_n^{(2)}\{F^{(2)}(\theta_n^{(1)}, \theta_n^{(2)}) + M_{n+1}^{(2)}\},$$

where, for each i, $F^{(i)}$, $\theta_n^{(i)}$ *and* $\lambda_n^{(i)}$ *satisfy the generic assumptions* G1–G3 *and the sequence* $\{\sum_{n=0}^k \lambda_n^{(i)} M_{n+1}^{(i)}\}_k$ *converges almost surely. And where, the* $\lambda_n^{(i)}$ *are chosen so that*

$$\frac{\lambda_n^{(1)}}{\lambda_n^{(2)}} \to 0 \qquad \text{as } n \to \infty.$$

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,*

$$\|\theta_n^{(2)} - \xi(\theta_n^{(1)})\| \to 0 \qquad \text{as } n \to \infty$$

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*

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

This theorem says that if the *fast* process, $\{\theta_n^{(2)}\}_{n\geq 0}$, converges to a unique limit point for any particular fixed value, $\theta^{(1)}$, of the *slow* process, we can analyze the asymptotic behavior 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 generalization of this result in Section 3, but for application of this theorem it suffices to note that Proposition 4 identifies conditions under which the limit set of the stochastic approximation process $\theta_n^{(1)}$ is contained within the set of fixed points of the flow defined by (5).

Theorem 5 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 behavior of the algorithm

as if the action estimates are accurate. This is a 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 $i = 1, \dots, N$,

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

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

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

Defining

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

we apply Theorem 5. The implicitly defined $M^{(i)}_n$ of that theorem are bounded martingale difference sequences, and so the convergence of the sequence $\{\sum_{n=0}^k \lambda^{(i)}_n M^{(i)}_{n+1}\}_k$ follows immediately. We have already observed that the $Q^i_n(a^i)$ 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

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

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 5 are clearly met and so we get the following:

THEOREM 6. *For the two-timescales algorithm* (6),

$$\|Q^i_n(a^i) - r^i(a^i, \pi^{-i}_n)\| \to 0 \qquad \text{as } n \to \infty \text{ a.s.}$$

and a suitable interpolation of the π^i_n 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 behavior of the two-timescales algorithm (6) is characterized by the same dynamical system as characterizes smooth fictitious play. Hofbauer and Hopkins [10] have studied these dynamics; they gave a Lyapunov function for two-player zero-sum games and also for two-player partnership games. Indeed, the Lyapunov function they gave for partnership games is easily extended to N-player partnership games—in this case the function is $r(\pi)+\tau\sum_{i=1}^{N} v^i(\pi^i)$. Hence, using Proposition 4, we have shown the following:

THEOREM 7. *Assume there are only finitely or countably many Nash distributions in a game which is either of the following*:

1. *A two-player zero-sum game.*
2. *An N-player partnership game.*

Then the two-timescales algorithm (6) *applied in that game will converge with probability* 1 *to a Nash distribution.*

On the other hand, Benaïm and Hirsch [2] showed that for the three-player matching pennies game [13] 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 [6] showed that for the Shapley game [19] the smooth best response dynamics with Boltzmann smoothing admit a Hopf bifurcation as the parameter τ goes to 0, so that for small values of τ, a limit cycle is again asymptotically stable and the unique equilibrium is unstable. It seems reasonable that a nonconvergence result analogous to Pemantle's [18] should hold in this case, since there is noise present in the system. However, the noise is present only 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 a simple extension of Benaïm's results [1], the probability of convergence to the equilibrium is less than 1.

Despite these nonconvergence results, the following is true:

THEOREM 8. *If the two-timescales algorithm* (6) *converges to a fixed point*

$$(Q_n, \pi_n) \to (Q, \pi) \qquad \text{as } n \to \infty,$$

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

PROOF. A basic result of stochastic approximation theory is that if convergence occurs, then the limit point must be a fixed point 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, $\pi^i = \beta^i(\pi^{-i})$. $\square$

3. Borkar's result extended to multiple timescales. The nonconvergence of the two-timescales algorithm (6) in certain games motivates a further extension. Littman and Stone's work [16] suggests consideration of players that learn at different rates. To consider this possibility we must extend Borkar's result [4] beyond two timescales.

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

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

where, for each i, $F^{(i)}$, $\theta_n^{(i)}$ and $\lambda_n^{(i)}$ satisfy the generic Assumptions G1–G3 and the sequence $\{\sum_{n=0}^{k} \lambda_n^{(i)} M_{n+1}^{(i)}\}_k$ converges almost surely. In addition, we assume that

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

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

As before, we define a timescale on which to interpolate the approximation processes. However, we now follow Borkar [4] and establish a different timescale to correspond to each process. For $i, j \in 1, \dots, N$, 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, that is,

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

and let

$$m^{(j)}(t) = \sup\big\{\kappa \ge 0 : t_\kappa^{(j)} \le t\big\}.$$

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

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} + \lambda_n^{(N)} U_{n+1}^{(i,N)} \qquad \text{for } i < N,$$
$$\theta_{n+1}^{(N)} = \theta_n^{(N)} + \lambda_n^{(N)}\{F^{(N)}(\theta_n) + M_{n+1}^{(N)}\},$$

where for $i < N$ we have implicitly defined

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

For any n,

$$\sup\left\{\left\|\sum_{l=n}^{k-1}\lambda_l^{(N)}U_{l+1}^{(i,N)}\right\| : k=n+1,\dots,m^{(N)}\bigl(t_n^{(N)}+T\bigr)\right\}$$

(8)

$$\leq \sup_k\left\{\left(\sum_{l=n+1}^{m^{(N)}(t_n^{(N)}+T)}\lambda_{l-1}^{(N)}\right)\left(\frac{\lambda_k^{(i)}}{\lambda_k^{(N)}}\right)F^{(i)}(\theta_k)+\left\|\sum_{l=n}^{k-1}\lambda_l^{(i)}M_{l+1}^{(i)}\right\|\right\}.$$

As $n\to\infty$, the second term converges to 0, 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 0.

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 Proposition 2 to show that on this timescale the interpolated processes $\Theta^{(\cdot,N)}(t)$ are asymptotic pseudotrajectories for the flow defined by the differential equations

$$\dot{X}^{(i)}=0 \qquad \text{for } i<N, \tag{9}$$

$$\dot{X}^{(N)}=F^{(N)}(X). \tag{10}$$

At this point we need to make the following assumption:

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) and (10) converge to the point $(\theta^{(<N)},\xi^{(N)}(\theta^{(<N)}))$ given initial conditions $(\theta^{(<N)},\theta^{(N)})$.

It therefore follows from Proposition 4 that the possible limit points of an asymptotic pseudotrajectory to the flow defined by (9) and (10) are the set of all points

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

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

$$\bigl\|\theta_n-\bigl(\theta_n^{(<N)},\xi^{(N)}(\theta_n^{(<N)})\bigr)\bigr\|\to 0 \qquad \text{as } n\to\infty \text{ a.s.}$$

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

$$\theta_{n+1}^{(i)}=\theta_n^{(i)}+\lambda_{n+1}^{(N-1)}U_{n+1}^{(i,N-1)} \qquad \text{for } i<N-1,$$

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

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

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

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

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

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

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

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

$$\dot{X}^{(i)} = 0 \qquad \text{for } i < N-1, \tag{11}$$

$$\dot{X}^{(N-1)} = F^{(N-1)}\big(X^{(<N)}, \xi^{(N)}(X^{(<N)})\big). \tag{12}$$

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

ASSUMPTION 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) and (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

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

it follows that

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

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

$$\dot{X}^{(i)} = 0 \qquad \text{for } i < j, \tag{13}$$

$$\dot{X}^{(j)} = F^{(j)}\big(X^{(\leq j)}, \Xi^{(\geq j+1)}(X^{(\leq j)})\big). \tag{14}$$

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

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) and (14) converge to the point $(\theta^{(<j)}, \xi^{(j)}(\theta^{(<j)}))$ given initial conditions $(\theta^{(<j)}, \theta^{(\geq j)})$.

Then defining

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

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

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

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

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

We have therefore proved the following theorem:

THEOREM 9. *Consider a multiple-timescales stochastic approximation process* (7). *If Assumptions* A(2)–A(N) *hold, then almost surely*

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

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*

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

4. A multiple-timescales learning algorithm. Theorem 9 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 $i = 1, \ldots, N$,

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

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 Assumption G3 and the additional conditions

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

This last condition says that each player is adapting his or her 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 10. *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 9 we need to check that Assumptions A(2)–A(N) are satisfied. We start by noting that the ODE

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

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

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

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

where we recursively define

$$B^{>(N-1)}(\pi^{\leq(N-1)}) = \beta^N(\pi^{\leq(N-1)}),$$
$$B^{>i}(\pi^{\leq i}) = \big(b^{i+1}(\pi^{\leq i}), B^{>(i+1)}\big(\pi^{\leq i}, b^{i+1}(\pi^{\leq i})\big)\big).$$

Effectively this says that, for any i, if we fix the strategies for players $1, \ldots, i$, 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 and 6). It allows us to use Theorem 9 to characterize the asymptotic behavior of the algorithm (15).

THEOREM 11. *For the multiple-timescales algorithm* (15) *under Assumption* C,

$$\big\|(\pi_n^2, \ldots, \pi_n^N) - B^{>1}(\pi_n^1)\big\| \to 0 \qquad \text{as } n \to \infty \text{ a.s.}$$

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

$$\dot{\pi}^1 = \beta^1\big(B^{>1}(\pi^1)\big) - \pi^1.$$

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 Assumption C. □

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

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

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

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

THEOREM 12. *For both two-player zero-sum games and two-player partnership games the ODE* (16) *admits a strict Lyapunov function. Fixed points of the ODE* (16) *are Nash distributions of the game.*

PROOF. For zero-sum games the function

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

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

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

is a Lyapunov function. □

This gives rise to the following immediate corollary.

COROLLARY 13. *Assume there are only finitely or countably many Nash distributions in a game which is either of the following*:

1. *A two-player zero-sum game.*
2. *A two-player partnership game.*

Then the multiple-timescales algorithm (15) *applied in that game will converge with probability* 1 *to a Nash distribution.*

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 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 [19], introduced in 1964 to show that classical fictitious play need not always converge, and the three-player matching pennies game, a remarkably simple game introduced by Jordan [13] 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 generalization of the latter game, then show convergence of our algorithm for the Shapley game.

6.1. *N-player matching pennies.* Our generalization of Jordan's game [13] 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

$$r^i(\underline{a}) = I_{\{a^i = a^{i+1}\}} \qquad \text{for } i = 1, \dots, N-1,$$
$$r^N(\underline{a}) = I_{\{a^N \neq a^1\}}.$$

The cyclical nature of this game allows easy verification of 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 depends only on π^1. Similarly, under the assumption that player 1 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 11 it suffices to consider the ODE

$$\dot{\pi}^1 = \beta^1\big(\beta^2\big(\cdots\big(\beta^N(\pi^1)\big)\cdots\big)\big) - \pi^1.$$

We assume that the smooth best responses are monotonic in the payoffs, that is, $r^i(a^i) > r^i(b^i) \Rightarrow \beta^i(r^i)(a^i) > \beta^i(r^i)(b^i)$. 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,

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

for each $i = 1, \dots, N$. 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 globally attracting fixed point.

We have shown that the multiple-timescales 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

$$\begin{pmatrix} (0,0) & (1,0) & (0,1) \\ (0,1) & (0,0) & (1,0) \\ (1,0) & (0,1) & (0,0) \end{pmatrix}.$$

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 Assumption C is irrelevant (as observed previously). So we simply need to analyze the ODE

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

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 $\rho(a) = (\pi^1(a) - \pi^1(3))/\tau$ for $a = 1, 2$ it is clear that

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

By the chain rule applied to (17),

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

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

$$\frac{\partial\beta^1(\pi^2)(a)}{\partial\pi^2(a')} = \frac{e^{\pi^2(a')/\tau}(I_{\{a'=a+1\}}\sum_{b'\in A} e^{\pi^2(b')/\tau} - 1)}{\tau(\sum_{b'\in A} e^{\pi^2(b')/\tau})^2},$$

$$\frac{\partial\pi^2}{\partial\rho(1)} = \frac{e^{\rho(1)}}{(1+e^{\rho(1)}+e^{\rho(2)})^2}\big(-e^{\rho(2)}, -1, 1+e^{\rho(2)}\big),$$

$$\frac{\partial\pi^2}{\partial\rho(2)} = \frac{e^{\rho(2)}}{(1+e^{\rho(1)}+e^{\rho(2)})^2}\big(1+e^{\rho(1)}, -1, -e^{\rho(1)}\big),$$

$$\frac{\partial\rho(b)}{\partial\pi^1(a)} = \frac{(1+I_{\{a=b\}})}{\tau},$$

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

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

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

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

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

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

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

$$e^{\pi^2(1)/\tau} e^{\pi^2(3)/\tau} \{\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3)\}$$
$$\leq e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \{\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2) - 2\pi^2(1)\pi^2(3)\},$$

and the expression in (19) is bounded above by

$$e^{\pi^2(1)/\tau} e^{\pi^2(2)/\tau} \{\pi^2(1)\pi^2(3) - 2\pi^2(2)\pi^2(3) - 2\pi^2(1)\pi^2(2)\}$$
$$+ e^{\pi^2(2)/\tau} e^{\pi^2(3)/\tau} \{-\pi^2(1)\pi^2(2) - \pi^2(2)\pi^2(3) - 4\pi^2(1)\pi^2(3)\},$$

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

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

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 [6], so this point must be globally attracting. Therefore, from Theorem 11 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 [2], since the asymptotic behavior of both algorithms can be shown to be characterized by the asymptotic behavior 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 there exists a strict Lyapunov function in these cases.

By extending Borkar's stochastic approximation result to multiple timescales, we have presented a new learning algorithm in which 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 rely on Assumption C, that faster players will converge to a unique fixed point for any fixed strategy of the slower players. 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.

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 [19] and the N-player matching pennies game—these latter two games having caused problems for all algorithms previously known to us.

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 toward 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 11 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.

Acknowledgment. The authors thank Andy Wright (BAE SYSTEMS) for several useful discussions during the preparation of this article, during his tenure on a Royal Society Industrial Fellowship at the University of Bristol.

REFERENCES

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

[2] 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.

[3] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[4] BORKAR, V. S. (1997). Stochastic approximation with two timescales. *Systems Control Lett.* **29** 291–294.

[5] BORKAR, V. S. (2002). Reinforcement learning in Markovian evolutionary games. Available at www.tcs.tifr.res.in/∼borkar/games.ps.

[6] COWAN, S. (1992). Dynamical systems arising from game theory. Ph.D. dissertation, Univ. California, Berkeley.

[7] FUDENBERG, D. and KREPS, D. M. (1993). Learning mixed equilibria. *Games Econom. Behav.* **5** 320–367.

[8] FUDENBERG, D. and LEVINE, D. K. (1998). *The Theory of Learning in Games*. MIT Press, Cambridge, MA.

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

[10] HOFBAUER, J. and HOPKINS, E. (2002). Learning in perturbed asymmetric games. Available at www.econ.ed.ac.uk/pdf/perturb.pdf.

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

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

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

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

[15] KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer, New York.

[16] LITTMAN, M. and STONE, P. (2001). Implicit negotiation in repeated games. *Intelligent agents VIII*: *Agent Theories, Architectures and Languages. Lecture Notes in Comput. Sci.* **2333** 393–404. Springer, Berlin.

[17] NASH, J. (1951). Non-cooperative games. *Ann. Math.* **54** 286–295.

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

[19] SHAPLEY, L. S. (1964). Some topics in two person games. In *Advances in Game Theory* (M. Dresher, L. S. Shapley and A. W. Tucker, eds.) 1–28. Princeton Univ. Press.

[20] SUTTON, R. S. and BARTO, A. G. (1998). *Reinforcement Learning*: *An Introduction*. MIT Press.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BRISTOL
UNIVERSITY WALK
BRISTOL BS8 1TW
UNITED KINGDOM
E-MAIL: d.s.leslie@bristol.ac.uk
e.j.collins@bristol.ac.uk