SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 495–514

INDIVIDUAL Q-LEARNING IN NORMAL FORM GAMES*

DAVID S. LESLIE† AND E. J. COLLINS‡

Abstract. The single-agent multi-armed bandit problem can be solved by an agent that learns the values of each action using reinforcement learning. However, the multi-agent version of the problem, the iterated normal form game, presents a more complex challenge, since the rewards available to each agent depend on the strategies of the others. We consider the behavior of value-based learning agents in this situation, and show that such agents cannot generally play at a Nash equilibrium, although if smooth best responses are used, a Nash distribution can be reached. We introduce a particular value-based learning algorithm, which we call individual Q-learning, and use stochastic approximation to study the asymptotic behavior, showing that strategies will converge to Nash distribution almost surely in 2-player zero-sum games and 2-player partnership games. Player-dependent learning rates are then considered, and it is shown that this extension converges in some games for which many algorithms, including the basic algorithm initially considered, fail to converge.

Key words. reinforcement learning, normal form games, stochastic approximation, multi-agent learning, player-dependent learning rates

AMS subject classifications. 93E35, 91A26, 68T05, 62L20

DOI. 10.1137/S0363012903437976

1. Introduction.

We will study value-based learning agents in the multi-agent multi-armed bandit problem (more commonly known as an iterated normal form game). These simple agents adapt in a very similar manner to the reinforcement learning algorithm used by Sutton and Barto (1998) in the single-agent case. However, the multi-agent setting presents a significantly more complex problem, since the rewards available to each agent depend on the strategies of all the other learners and hence are not stationary.

In this paper we will show how to study the asymptotic behavior of these agents using the ODE method of stochastic approximation. The resulting dynamical systems can be analyzed to prove that convergence to equilibrium must occur for 2-player zero-sum games and 2-player partnership games (Proposition 4.2). However, it is shown that convergence does not occur for some specific games known to cause difficulties for all learning algorithms; player-dependent learning rates are then introduced, and convergence can then be proved to occur in a larger class of games (Proposition 5.4 and Corollary 5.6).

One of the best studied models of adaptation in iterated normal form games is fictitious play (Brown (1951), Fudenberg and Levine (1998)). However, this paradigm requires each player to know her own reward function, to observe the actions of all players, and to calculate expected rewards from this information. While this is realistic in some situations, it is clear that players of a game such as the stock

*Received by the editors December 1, 2003; accepted for publication (in revised form) December 8, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/43797.html

†School of Economics, The University of New South Wales, Sydney, NSW 2052, Australia (d.leslie@unsw.edu.au). This research was completed while this author was at the University of Bristol, and was supported by CASE Research Studentship 00317214 from the UK Engineering and Physical Sciences Research Council in cooperation with BAE SYSTEMS.

‡Department of Mathematics, University Walk, Bristol BS8 1TW, UK (e.j.collins@bristol.ac.uk).

market, or animals involved in evolutionary games (Maynard Smith (1982)), do not know the relevant reward structures, while in potential applications of multi-agent learning (e.g., Boyan and Littman (1994), Singh and Bertsekas (1997), Crites and Barto (1998)) these requirements are also frequently not satisfied.

Therefore we study models of adaptation under which agents simply respond to observations of the rewards they receive for playing actions, as in the simple and successful approach used by Sutton and Barto (1998) to solve the single-agent multi-armed bandit problem. Agents estimate the value of each action, and play a strategy that is a simple function of these estimates; we will call such agents *value-based learners*. These learners take no account of the presence of other players of the game—the only information used is the reward received for each action played. Recent neurophysiological data from rhesus monkeys trying to perform a repetitive task (Glimcher, Dorris, and Bayer (2005)) suggests that value-based learning is actually employed in nature, adding extra interest to our investigations.

Other current approaches to learning in games include actor-critic learning, hypothesis-based learning, consistent learning, and alternative reinforcement learning models based on Roth and Erev (1995). Our value-based learning scheme compares favorably with these in terms of simplicity, form of convergence, and structure, respectively.

The Roth and Erev (1995) model of learning is the standard model of reinforcement learning in the game theory community. However, it has the disadvantages that it requires all rewards to be positive, and it uses rewards as "reinforcement signals" instead of information about the values of actions, despite the fact that strategies and values are "dimensionally different" quantities—strategies are probability distributions whereas rewards are arbitrary real numbers (the same criticism can be made of stimulus-response learning (Börgers and Sarin (1997))).

Actor-critic learning is a more sophisticated framework that explicitly links rewards and strategies. In this model, agents maintain separate value functions and strategies and map the value function to strategy space in order to update the strategy. It has been used in several recent approaches to learning in games (Borkar (2001), Bowling and Veloso (2002), Leslie and Collins (2003)), but, in contrast with the value-based approach considered in this paper, the extra complication involved in separating the values and the strategies is unnecessary and overly sophisticated when applied in a single-agent setting.

A different approach to the problem of learning in games has been to develop consistent (Hannan (1957)) reinforcement learning procedures (Baños (1968), Megido (1980), Auer et al. (1995), Hart and Mas-Colell (2001)). Hart and Mas-Colell (2000) show that the long-run average actions of players using a consistent algorithm will converge to a correlated equilibrium of the game (Aumann (1974)) (as opposed to a classical Nash equilibrium). However, it is often difficult to characterize the correlated equilibria (Fudenberg and Tirole (1991)), and the convergence is in the sense of the average action played, instead of convergence of actual play as developed in this paper.

An even more sophisticated framework is the hypothesis-testing formulation recently proposed by Foster and Young (2003). Here the players formulate hypotheses about opponent strategies and repeatedly test these hypotheses against observed play. It is shown that play is close to a Nash equilibrium of the repeated game formulation most of the time. However, the sophistication required from the players is greater than for virtually all other algorithms in the literature, and the scheme is clearly not applicable in the minimal information setting we consider here.

It is appropriate to comment here on a recent result of Hart and Mas-Colell (2003) showing that no "uncoupled" adaptive dynamics can converge to equilibrium in all games. The dynamics we consider in this paper are certainly uncoupled, and indeed the dynamics resulting from any system where players do not observe opponent payoffs must necessarily be uncoupled. However, as we shall see later, convergence to Nash distribution may be a more attainable goal than convergence to Nash equilibrium (see section 2 for details). Furthermore, the player-dependent learning rates introduced in section 5 are a further step away from the framework studied by Hart and Mas-Colell (2003), thus allowing convergence to be proved for a larger class of games than has been previously achieved.

This paper is therefore an analysis of a simple and intuitive learning procedure applied in the setting of normal form games with minimal information available to the players. In section 2 we discuss some difficulties faced by value-based learners in games, and propose the use of smooth best responses (also known as softmax action selection). This motivates our individual Q-learners, introduced in section 3, where we show how to characterize their behavior using stochastic approximation (Benaïm (1999)). The behavior of these learners in 2-player games is analyzed in section 4, where we show that strategy evolution is closely related to the smooth best response dynamics (Hofbauer and Hopkins (2005)); this is the same dynamical system that characterizes stochastic fictitious play (Benaïm and Hirsch (1999)), despite the fact that individual Q-learning uses significantly less information than stochastic fictitious play. However, previous work (Leslie and Collins (2003)) suggests that convergence to a fixed point can be proved to occur in a larger class of games if player-dependent learning rates are used to break symmetry between the players; this is studied in section 5. A problem with player-dependent learning rates is that the analytical methods so far developed do not apply for all games; section 6 uses graphical representations of games to investigate classes of games in which player-dependent learning rates can be analyzed.

2. Value-based players in games. In this section we will introduce our notation and discuss some problems faced by value-based players of games. In particular, we will show that value-based players cannot generally play at a Nash equilibrium, but if smooth best responses are used, equilibrium play becomes possible (although this will no longer be at the classical Nash equilibrium).

We start by introducing our notation and presenting a familiar example. A normal form game consists of N players, where each player $i \in \{1, \dots, N\}$ has a finite set A^i of actions and a reward function $r^i : A^1 \times \cdots \times A^N \to \mathbb{R}$. When the game is played, each player $i \in \{1, \dots, N\}$ selects an action $a^i \in A^i$ and then receives a reward which has expected value $r^i(a^1, \dots, a^N)$; each player tries to maximize her expected reward. A traditional 2-player example is rock-scissors-paper, where the action set for each player is {Rock, Scissors, Paper}, and the reward functions are given in the payoff matrix

$$\begin{array}{c} \\ \text{Rock} \\ \text{Scissors} \\ \text{Paper} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Rock} & \text{Scissors} & \text{Paper} \end{array} \\ \begin{pmatrix} (0,0) & (1,-1) & (-1,1) \\ (-1,1) & (0,0) & (1,-1) \\ (1,-1) & (-1,1) & (0,0) \end{pmatrix} \end{array}, \tag{2.1}$$

where player 1's action determines the row, player 2's action determines the column, and an entry (x, y) means that player 1 receives reward x and player 2 receives reward y. Note that this is a 2-player zero-sum game, where for any joint action (a^1, a^2) the

rewards satisfy $r^1(a^1, a^2) + r^2(a^1, a^2) = 0$. We will also have occasion to consider partnership games, where for any joint action the reward given to each player is the same.

A mixed strategy for player i is an element $\pi^i \in \Delta^i$, where Δ^i is the set of probability distributions over the action space A^i; we will write $\pi^i(a^i)$ for the probability that player i selects action a^i when using strategy π^i. There are unique multilinear extensions of the reward functions, also denoted r^i, to the joint strategy space $\Delta = \Delta^1 \times \cdots \times \Delta^N$, and in further abuse of notation we will write $r^i(a^i, \pi^{-i})$ (resp., $r^i(\pi^i, \pi^{-i})$) for the expected reward that player i will receive if she plays action a^i (resp., strategy π^i) and the other players select actions according to the opponent mixed strategy $\pi^{-i} = (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^N)$.

Nash (1950) defined the classical solution concept for normal form games by observing that a rational player will not play a strategy π^i that does not maximize her expected reward given the opponent strategy π^{-i}. Thus a Nash equilibrium is a joint strategy $\tilde{\pi} \in \Delta$ satisfying, for each i,

$$r^i(\tilde{\pi}) \geq r^i(\pi^i, \tilde{\pi}^{-i}) \quad \text{for any } \pi^i \in \Delta^i.$$

Nash (1950) showed that at least one Nash equilibrium exists for any normal form game. In our rock-scissors-paper example, it is well known that there is a unique Nash equilibrium where each player plays the mixed strategy $(1/3, 1/3, 1/3)$, but in general there can be many Nash equilibria of a game. The problem of equilibrium selection (i.e., when faced with a game, how should players decide which Nash equilibrium strategy to play when many might exist) can be seen as a motivating factor for the study of learning in games (Fudenberg and Levine (1998)). An alternative perspective is to consider the Nash equilibria of a game as the only points to which a sensible learning procedure should converge—note that if only one player is present, then a Nash equilibrium is simply the action which returns the highest expected reward.

However, value-based approaches, as used by Sutton and Barto (1998) in the single-agent problem, encounter problems with Nash equilibria. At a Nash equilibrium, any action played by player i with positive probability will receive expected reward $\max_{a^i \in A^i} r^i(a^i, \tilde{\pi}^{-i})$, yet the equilibrium strategy might well require these maximizing actions to be played with specific and possibly unequal probabilities. For example, consider the game with payoff matrix

$$\begin{pmatrix} (2,0) & (0,1) \\ (0,2) & (1,0) \end{pmatrix},$$

which has a unique equilibrium where $\pi^1 = (2/3, 1/3)$ and $\pi^2 = (1/3, 2/3)$. At this equilibrium, both actions of both players get expected reward $2/3$, yet player 1 must favor action 1, and player 2 must favor action 2. Thus play at an equilibrium is not possible when the strategies played are restricted to be simple functions of the expected rewards (unless these functions are asymmetric under reordering of the actions).

A solution to this problem lies in using smooth best responses, which can also be considered as arising from a Bayesian uncertainty about the rewards (Harsanyi (1973)) and are closely related to the softmax exploration method of reinforcement learning (Sutton and Barto (1998)). If player i has estimates $Q^i(a^i)$ of the values of actions $a^i \in A^i$, then a smooth best response $\beta^i(Q^i) \in \Delta^i$ to these estimates is

given by

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

Here $\tau > 0$ is a temperature parameter, and $v^i : \Delta^i \to \mathbb{R}$ is a player-dependent smoothing function, which is a smooth, strictly differentiably concave function such that as π^i approaches the boundary of Δ^i, the slope of v^i becomes infinite (Fudenberg and Levine (1998)). As the temperature parameter $\tau \to 0$, $\beta^i(Q^i)$ approaches the set of best responses (i.e., strategies that select only actions a^i maximizing $Q^i(a^i)$). However, while there may be many best responses (e.g., suppose all the Q values are equal), the conditions on v^i imply that there is a unique smooth best response given τ and v^i (Fudenberg and Levine (1998)).

A familiar example of a smooth best response is Boltzmann action selection. Under this scheme, the smoothing functions are

$$v^i(\pi^i) = -\sum_{a^i \in A^i} \pi^i(a^i) \log \pi^i(a^i),$$

resulting in the smooth best response function

$$\beta^i(Q^i)(a^i) = \frac{\mathrm{e}^{Q^i(a^i)/\tau}}{\sum_{b^i \in A^i} \mathrm{e}^{Q^i(b^i)/\tau}}. \tag{2.2}$$

The use of smooth best responses means that, in general, Nash equilibria are no longer fixed points in strategy space, and an alternative equilibrium concept must be defined. (For example, consider a 1-player game with a unique optimal action; the conditions on v^1 mean that all actions are played with positive probability, which is clearly not a Nash equilibrium.) Given a set of smooth best response functions, β^i, we define a *Nash distribution* to be a joint mixed strategy $\pi \in \Delta$ such that, for each i,

$$\pi^i = \beta^i(r^i(\cdot, \pi^{-i})); \tag{2.3}$$

i.e., each player plays a smooth best response to the rewards arising from opponent play. Brouwer's fixed point theorem shows that such distributions must exist; Govindan, Reny, and Robson (2003) show that for small temperatures τ, the Nash equilibria of a game are approximated by Nash distributions (the proof of Harsanyi (1973) is insufficient in this case, because it relies on perturbations of the rewards having compact support). Note that the Nash distributions depend on the smooth best response functions β^i (through the particular choices of τ and v^i) as well as on the reward functions r^i—for the remainder of this paper we will assume that any particular player uses a fixed smooth best response function for all time.

Thus we have shown that a value-based approach cannot result in Nash equilibrium play in general games but can result in strategies that are close to a Nash equilibrium if players use smooth best responses. In the next section we will introduce a value-based learning algorithm incorporating this idea, which can therefore converge to a Nash distribution.

3. Individual *Q*-learning. Sutton and Barto (1998) show that a simple reinforcement learning scheme can be used to estimate action values in a single-agent task. The basic algorithm is given by

$$Q_{n+1}(a) = Q_n(a) + \lambda_{n+1} \mathbb{I}_{\{a_n = a\}} \{R_n - Q_n(a)\} \quad \text{for each } a \in A, \tag{3.1}$$

TABLE 1
Individual Q-learning.

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, receives reward R_n^i, and then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}\mathbb{I}_{\{a_n^i = a^i\}}\frac{R_n^i - Q_n^i(a_n^i)}{\beta^i(Q_n^i)(a_n^i)} \quad \text{for each } a^i \in A^i, \tag{3.2}$$

where $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence of learning parameters satisfying

$$\sum_{n\geq 1}\lambda_n = \infty, \qquad \sum_{n\geq 1}(\lambda_n)^2 < \infty. \tag{3.3}$$

where a_n is the action selected at time n, and R_n is the subsequent reward. A similar scheme appears to fit recent neurophysiological data measured from the brains of rhesus monkeys learning to perform a repetitive task (Glimcher, Dorris, and Bayer (2005)). In applying (3.1), actions are selected in such a way that each action is played infinitely often, but as $n \to \infty$ the probability of playing any action which does not have a maximal Q value tends to 0. We will study an analogous system in the equivalent multi-agent task; N players are faced with a normal form game, which they play repeatedly, learning Q values by observing rewards. The algorithm we study is given in Table 1.

This scheme was originally suggested by Fudenberg and Levine (1998); it will be noticed that it is very similar to the standard reinforcement learning model (3.1). The first difference is that a player's strategy at any stage of the game is determined by the algorithm (as opposed to the single-agent case, where there is no need to be specific about action choices at any particular play). This is because the rewards observed by a particular player depend crucially on the strategies of the other players, so these strategies must be carefully specified. The second difference is that the reward prediction error $(R_n^i - Q_n^i(a_n^i))$ (Sutton and Barto (1998)) is divided by $\beta^i(Q_n^i)(a_n^i)$, the probability with which a_n^i was selected. This can be viewed as compensating for the fact that actions played with low probability do not receive frequent updates of their Q values, so when they are played any reward prediction error must have greater influence on the Q value than if frequent updates occur. Further, we will see in section 4 that this division by $\beta^i(Q_n^i)(a_n^i)$ results in a system that is closely related to the (well-studied) smooth best response dynamics (Hofbauer and Hopkins (2005)).

The conditions (3.3) on the learning parameters $\{\lambda_n\}_{n\geq 1}$ mean that standard theorems of stochastic approximation can be used (Benaïm (1999)). The interested reader may wish to consult Benaïm and Hirsch (1999) for an introduction to the ODE method of stochastic approximation in the context of game theory. Writing $Q_n = (Q_n^1, \ldots, Q_n^N)$, and writing $\beta^{-i}(Q_n)$ for the opponent mixed strategies resulting from the values Q_n, note that for each i and a^i

$$\begin{aligned}\mathbb{E}[Q_{n+1}^i(a^i) - Q_n^i(a^i) \,|\, Q_n] &= \lambda_{n+1} \times \beta^i(Q_n^i)(a^i) \times \frac{r^i(a^i, \beta^{-i}(Q_n)) - Q_n^i(a^i)}{\beta^i(Q_n^i)(a^i)} \\ &= \lambda_{n+1}\left\{r^i(a^i, \beta^{-i}(Q_n)) - Q_n^i(a^i)\right\}.\end{aligned}$$

The following lemma follows immediately from the results of Benaïm (1999).

LEMMA 3.1. *The values Q_n resulting from the individual Q-learning algorithm* (3.2) *converge almost surely to a connected internally chain-recurrent set*[1] *of the flow defined by the Q-learning ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t}q_t^i(a^i) = r^i(a^i, \beta^{-i}(q_t)) - q_t^i(a^i) \quad \textit{for each } i \textit{ and } a^i, \tag{3.4}$$

provided that the Q_n remain bounded for all n.

This assumption of boundedness could be dropped if we used either fixed truncation to a bounded space (Kushner and Yin (1997)) or randomly varying truncations (Chen and Zhu (1986)); the former will change the differential equations to be studied a little, while the latter will have no consequence on the analysis of the asymptotic behavior. However, the purpose of this paper is not to investigate such issues, and in the numerical experiments carried out there were no problems with values growing large. Therefore in the interests of space we will be content to keep this as an assumption throughout the rest of the paper.

The first thing to notice about this algorithm is that convergence to a point can occur only at Nash distribution values. This follows from what is essentially a law of large numbers result, saying that convergence to a point can occur only if the expected change at that point is zero. Thus such a point must satisfy $Q^i(a^i) = r^i(a^i, \beta^{-i}(Q))$ for each i and a^i. However, if we write $\pi^i = \beta^i(Q^i)$, this translates to $\pi^i = \beta^i(r^i(\cdot, \pi^{-i}))$, which is precisely the definition of a Nash distribution (2.3). The learning algorithm can therefore be applied blindly in any game, and if convergence occurs, a Nash distribution must have been reached.

Our next step in analyzing this system is to consider whether the values of the players are ever consistent with the structure of the game; for example, if in a zero-sum game the values ever actually sum to zero. Writing

$$\mathcal{B} = \{(r^1(\cdot, \pi^{-1}), \ldots, r^N(\cdot, \pi^{-N})) : \pi \in \Delta\}$$

for the set of values that could arise from a joint mixed strategy, we call values Q_n *asymptotically belief-based* if the limit set of the values is contained in $\mathcal{B}$.

LEMMA 3.2. *The values Q_n resulting from the individual Q-learning algorithm* (3.2) *are almost surely asymptotically belief-based, provided that the Q_n remain bounded for all time.*

Proof. We will rewrite the Q-learning ODE (3.4) to show that $\mathcal{B}$ is a global attractor of the resulting flow, which suffices to show that the values Q_n resulting from (3.2) are asymptotically belief-based (Benaïm (1999)). Start by writing $q_t = (q_t^1(\cdot), \ldots, q_t^N(\cdot))$ and $r_t = (r^1(\cdot, \beta^{-1}(q_t)), \ldots, r^N(\cdot, \beta^{-N}(q_t)))$, so that

$$\frac{\mathrm{d}}{\mathrm{d}t}q_t = r_t - q_t.$$

This can be rewritten as

$$q_t = \mathrm{e}^{-t}q_0 + (1 - \mathrm{e}^{-t})\bar{r}_t,$$

[1]The concept of a chain-recurrent set is central to the ODE method of stochastic approximation and is developed fully in Benaïm (1999). In the interests of space, this theory is therefore not repeated here. Loosely, an internally chain-recurrent set of a flow is an invariant set of the flow such that a trajectory starting at any point of the set can return to its start point without ever leaving the set, when small "jumps" are allowed to be made. It will suffice to know that a connected internally chain-recurrent set of a flow is an invariant set of the flow containing no proper attractors.

where $\bar{r}_t = (\mathrm{e}^t - 1)^{-1} \int_0^t \mathrm{e}^s r_s \, \mathrm{d}s$ is a weighted average of r_s, $0 \leq s \leq t$. Since $r_s \in \mathcal{B}$ for all s, it follows immediately that $\bar{r}_t \in \mathcal{B}$ for all t. Therefore $q_t \to \mathcal{B}$ for any q_0, and $\mathcal{B}$ is a global attractor of the flow defined by (3.4). □

As well as being interesting in itself, this is crucial to the analysis of the next section, where we relate this value-based learning to the smooth best response dynamics, usually considered to arise from models of learning in which players use significantly more information on the structure of the game and observations of opponent play than is necessary for individual Q-learning.

4. 2-player games. In this section we will relate the Q-learning ODE (3.4) to the smooth best response dynamics in 2-player games, as suggested by Fudenberg and Levine (1998). This will allow us to characterize the limiting behavior of the individual Q-learning algorithm in certain classes of games.

The smooth best response (SBR) dynamics are defined by the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^i = \beta^i(r^i(\cdot, \pi_t^{-i})) - \pi_t^i \quad \text{for each } i \tag{4.1}$$

and have been shown to characterize stochastic fictitious play, in the same sense that (3.4) characterizes individual Q-learning (Benaïm and Hirsch (1999)). Fudenberg and Levine (1998) observe that for 2-player games, if strategies evolve according to these dynamics, the resultant rewards evolve according to the Q-learning ODE (3.4). This will be used in the proof of Lemma 4.1, where we show that a connected internally chain-recurrent set of the flow defined by the Q-learning ODE (3.4) corresponds to a connected internally chain-recurrent set of the SBR dynamics. By Lemma 3.1, this relates the limiting behavior of individual Q-learning with that of stochastic fictitious play, despite the fact that individual Q-learners have no information on the structure of the game and do not observe opponent play, both of which are necessary for stochastic fictitious play.

LEMMA 4.1. *For 2-player games, any connected internally chain-recurrent set of the Q-learning ODE* (3.4) *is of the form*

$$r(\mathcal{C}) := \{(r^1(\cdot, \pi^{-1}), \ldots, r^N(\cdot, \pi^{-N})) : \pi \in \mathcal{C}\},$$

where $\mathcal{C} \subset \Delta$ is a connected internally chain-recurrent set of the flow defined by the SBR dynamics (4.1).

Proof. Let $\mathcal{D}$ denote an arbitrary connected internally chain-recurrent set of the Q-learning ODE (3.4). Benaïm (1999, Proposition 5.3) shows that a set is connected internally chain-recurrent if and only if it is a compact invariant set admitting no proper attractor. Therefore, since the set $\mathcal{B}$ of belief-based values is a global attractor we must have $\mathcal{D} \subset \mathcal{B}$, and $q \in \mathcal{D}$ means that there exists $\pi \in \Delta$ such that $q^i(a^i) = r^i(a^i, \pi^{-i})$ for each i and a^i.

However, suppose π evolves according to the SBR dynamics (4.1), so that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} r^i(a^i, \pi_t^{-i}) &= \sum_{a^{-i} \in A^{-i}} \frac{\partial r^i(a^i, \pi^{-i})}{\partial \pi^{-i}(a^{-i})} \frac{\mathrm{d}}{\mathrm{d}t} \pi_t^{-i}(a^{-i}) \\
&= \sum_{a^{-i} \in A^{-i}} r^i(a^i, a^{-i}) \{\beta^{-i}(r^{-i}(\cdot, \pi_t^i))(a^{-i}) - \pi_t^{-i}(a^{-i})\} \\
&= r^i(a^i, \beta^{-i}(r^{-i}(\cdot, \pi_t^i))) - r^i(a^i, \pi_t^{-i}),
\end{aligned}$$

and the $r^i(a^i, \pi^{-i})$ evolve according to the same ODE as the $q^i(a^i)$. Note that this calculation is valid only for 2-player games.

Therefore trajectories of the Q-learning ODE (3.4) in $\mathcal{B}$ correspond to trajectories of the SBR dynamics (4.1) in Δ, and the invariant set $\mathcal{D}$ of (3.4) must be of the form $r(\mathcal{C})$, where $\mathcal{C} \subset \Delta$ is an invariant set of (4.1). Now since $\mathcal{D}$ admits no proper attractor, it follows that $\mathcal{C}$ admits no proper attractor. $\mathcal{C}$ may consist of several connected components, but any one of them will be a connected internally chain-recurrent set of the flow defined by the SBR dynamics (4.1) with $\mathcal{D} = r(\mathcal{C})$. □

This result allows us to characterize the limit set of the individual Q-learning algorithm in terms of the chain-recurrent sets of the SBR dynamics. These chain-recurrent sets are well studied, since they are the limit set of a stochastic fictitious play process (Benaïm and Hirsch (1999)), and in particular Hofbauer and Hopkins (2005) provide Lyapunov functions for 2-player zero-sum games and 2-player partnership games. This allows us to prove our first main proposition, which gives convergence results for individual Q-learning in these situations.

PROPOSITION 4.2. *In either 2-player zero-sum games, or 2-player partnership games with countably many Nash distributions (given the smooth best responses β^i), strategies of players using the individual Q-learning algorithm* (3.2) *will converge almost surely to a Nash distribution.*

Proof. Lemma 3.1 shows that the Q values converge to a connected internally chain-recurrent set of the Q-learning ODE (3.4). From Lemma 4.1 we know that this is of the form $r(\mathcal{C})$, where $\mathcal{C} \subset \Delta$ is a connected internally chain-recurrent set of flow defined by the SBR dynamics (4.1). In the games we consider, Hofbauer and Hopkins (2005) provide Lyapunov functions for the set of Nash distributions under the SBR dynamics, and this set is isolated (by assumption in the case of partnership games and by a result of Hofbauer and Hopkins (2005) for zero-sum games). Therefore Benaïm (1999, Corollary 6.6) shows that we can assume $\mathcal{C} = \{\tilde{\pi}\}$ where $\tilde{\pi}$ is a Nash distribution. Therefore $Q_n^i \to r^i(\cdot, \tilde{\pi}^{-i})$ for each i, and so $\beta^i(Q_n^i) \to \beta^i(r^i(\cdot, \tilde{\pi}^{-i}))$ by continuity. But $\tilde{\pi}$ is a Nash distribution, so $\beta^i(r^i(\cdot, \tilde{\pi}^{-i})) = \tilde{\pi}^i$, and we have shown that the strategies converge to a Nash distribution. □

We illustrate this convergence with the rock-scissors-paper game (2.1). This is a 2-player zero-sum game, and therefore strategies converge to the unique Nash distribution where all actions are played with probability 1/3 (this is the same as the Nash equilibrium, although for general games the Nash equilibria and Nash distributions do not coincide). In Figure 1 we see that despite erratic initial strategy shifts, the strategy of player 1 appears to be converging to the Nash distribution.

However, this convergent behavior does not occur for all games. Two classic examples of games that cause problems for learning algorithms are Shapley's variant of rock-scissors-paper (Shapley (1964)) and Jordan's 3-player matching pennies game (Jordan (1993)). In both of these games, the SBR dynamics (4.1) admit a unique linearly unstable fixed point, and an asymptotically stable limit cycle, for certain smooth best response functions β^i (Cowan (1992), Benaïm and Hirsch (1999)). We shall not reproduce all of these results for the Q-learning ODE but will show that in Shapley's game, using Boltzmann action selection (2.2), a Hopf bifurcation occurs at the unique Nash distribution as the temperature parameter tends to 0. This shows that for sufficiently small τ the Nash distribution is linearly unstable and a periodic orbit is an attractor. We use the symmetric formulation of Shapley's game, with payoff matrix

$$\begin{pmatrix} (0,0) & (1,0) & (0,1) \\ (0,1) & (0,0) & (1,0) \\ (1,0) & (0,1) & (0,0) \end{pmatrix}, \tag{4.2}$$

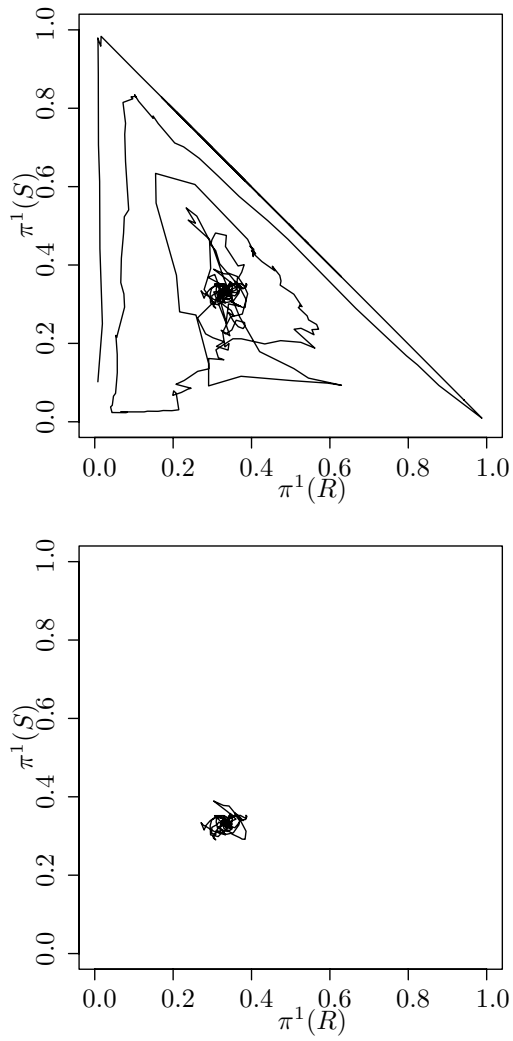


FIG. 1. *Strategies of player* 1 *in the rock-scissors-paper game* (2.1) *with Boltzmann action selection* ($\tau = 0.1$) *over* 5×10^5 *iterations of basic individual Q-learning* (3.2) *with* $\lambda_n = (n{+}100)^{-0.9}$. *The top diagram shows the entire learning run, whereas the bottom diagram omits the first* 10^4 *iterations, showing that strategies are converging to the unique Nash distribution.*

and therefore, for $i = 1, 2$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}Q^i(R) &= \pi^{-i}(S) - Q^i(R),\\
\frac{\mathrm{d}}{\mathrm{d}t}Q^i(S) &= \pi^{-i}(P) - Q^i(S),\\
\frac{\mathrm{d}}{\mathrm{d}t}Q^i(P) &= \pi^{-i}(R) - Q^i(P).
\end{aligned}$$

The Jacobian for this system, evaluated at the unique Nash distribution where all Q

values have the value $1/3$, is given by

$$\begin{pmatrix} -1 & 0 & 0 & \frac{-1}{9\tau} & \frac{2}{9\tau} & \frac{-1}{9\tau} \\ 0 & -1 & 0 & \frac{-1}{9\tau} & \frac{-1}{9\tau} & \frac{2}{9\tau} \\ 0 & 0 & -1 & \frac{2}{9\tau} & \frac{-1}{9\tau} & \frac{-1}{9\tau} \\ \frac{-1}{9\tau} & \frac{2}{9\tau} & \frac{-1}{9\tau} & -1 & 0 & 0 \\ \frac{-1}{9\tau} & \frac{-1}{9\tau} & \frac{2}{9\tau} & 0 & -1 & 0 \\ \frac{2}{9\tau} & \frac{-1}{9\tau} & \frac{-1}{9\tau} & 0 & 0 & -1 \end{pmatrix},$$

which has eigenvalues

$$\frac{1}{6\tau}(1 - 6\tau \pm \sqrt{3}i), \quad \frac{1}{6\tau}(-1 - 6\tau \pm \sqrt{3}i), \quad -1, \quad -1.$$

As $\tau \to 0$, the real part of the first conjugate pair crosses the imaginary axis from the negative half-plane to the positive half-plane, resulting in a Hopf bifurcation, so for sufficiently small τ the fixed point is linearly unstable. Therefore by the results of Pemantle (1990), convergence to this fixed point is a probability zero event; in Figure 2 we see that in fact play cycles, as it would under the SBR dynamics, and as implied by the Hopf bifurcation. Previous work (Leslie and Collins (2003)) suggests that using player-dependent learning rates helps to break the symmetry that allows this cycling to occur. We will apply this idea to individual Q-learning in the next section.

5. Player-dependent learning rates. Returning to general N-player games, Leslie and Collins (2003) introduce player-dependent learning rates (PDLR) to break the symmetry that allows strategies to cycle under the SBR dynamics (4.1). Under this paradigm, each player's learning parameters decay to zero at different rates, resulting in a process which is a stochastic approximation of a singularly perturbed dynamical system. The algorithm we study is shown in Table 2; it is a simple extension of individual Q-learning (3.2) which incorporates PDLR. In fact the only difference between this and the individual Q-learning algorithm (3.2) is that each player uses her own sequence of learning parameters $\{\lambda_n^i\}_{n\geq 1}$.

Note that condition (5.2) is used for ease of exposition but is equivalent to the condition that either $\lambda_n^i/\lambda_n^j \to 0$ or $\lambda_n^j/\lambda_n^i \to 0$ whenever $i \neq j$, since if this latter condition is true we can assume that the players are indexed in such a way that (5.2) is true. A suitable choice of learning parameters would be to choose $\lambda_n^i = (n+C)^{-\rho^i}$, where the rate $\rho^i \in (0.5, 1]$ is chosen differently for each player; indeed if players are thought of as selecting their own learning rate ρ^i independently using any continuous distribution on $(0.5, 1]$, then the necessary conditions will be met with probability 1.

The more slowly a player's learning parameters decrease to 0, the more "responsive" that player will be, since greater emphasis is placed on recent observations when estimating an action's value. In contrast, players with more rapidly decreasing learning parameters are more "cautious," since their value estimates take greater account of the entire history of observed rewards. Condition (5.2) means that players with higher indices i are more responsive (and hence less cautious) than those with lower indices.

As observed in Leslie and Collins (2003), in order to analyze an algorithm incorporating PDLR theoretically, we need to make an assumption about what would happen to the more responsive players if the strategies of the $i-1$ most cautious players were fixed.

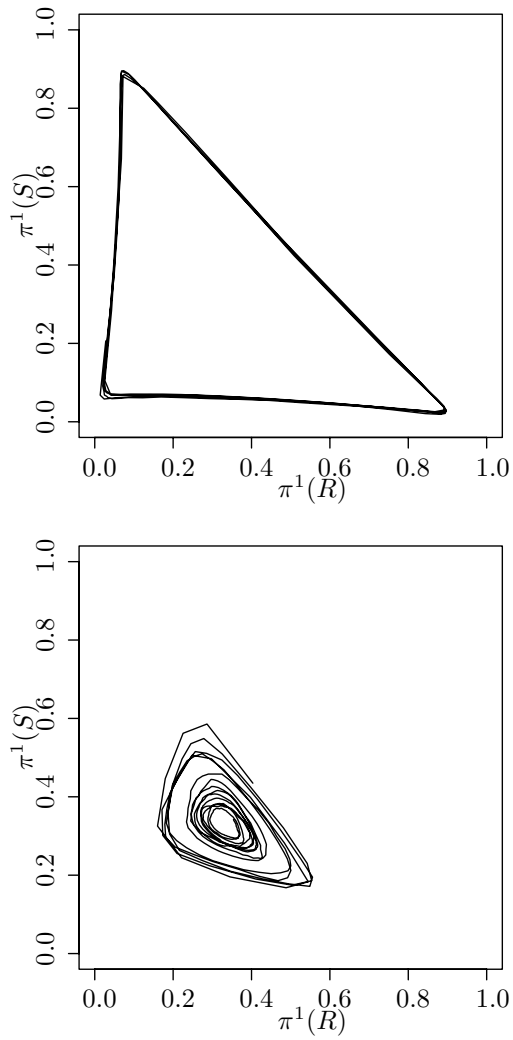


FIG. 2. *Strategies of player* 1 *in Shapley's game* (4.2) *with Boltzmann action selection* ($\tau = 0.1$) *over* 5×10^5 *iterations of basic individual Q-learning* (3.2) *with* $\lambda_n = (n + 100)^{-0.9}$ *(top) and of individual Q-learning with PDLR* (5.1)*, with* $\lambda_n^1 = (n + 100)^{-0.9}$ *and* $\lambda_n^2 = (n + 100)^{-0.7}$ *(bottom). The first* 1×10^4 *iterations are omitted in each case. For basic individual Q-learning* (3.2) *the strategies follow a limit cycle, while for individual Q-learning with PDLR* (5.1) *the strategies are spiraling anticlockwise toward the unique Nash distribution.*

ASSUMPTION 5.1. *For each* $i \in \{2, \ldots, N\}$ *there exists a function* $\tilde{q}^i : \Delta^1 \times \cdots \times \Delta^{i-1} \to \mathbb{R}^{|A^i|}$ *such that, for arbitrary fixed* $(Q^1, \ldots, Q^{i-1})$*, the ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^i(a^i) = r^i(a^i, [\pi^{(<i)}, B^{(>i)}[\pi^{(<i)}, \beta^i(q_t^i)]]) - q_t^i(a^i) \quad \textit{for each } a^i \in A^i$$

has the globally attracting fixed point $\tilde{q}^i(\pi^{(<i)})$*, where*

$$\pi^{(<i)} = (\beta^1(Q^1), \ldots, \beta^{i-1}(Q^{i-1}))$$

TABLE 2
Individual Q-learning with PDLR.

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, receives reward R_n^i, and then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}^i \mathbb{I}_{\{a_n^i = a^i\}} \frac{R_n^i - Q_n^i(a_n^i)}{\beta^i(Q_n^i)(a_n^i)} \quad \text{for each } a^i \in A^i, \tag{5.1}$$

where for each i, $\{\lambda_n^i\}_{n\geq 1}$ is a deterministic sequence of learning parameters satisfying the conditions (3.3), and additionally

$$\lambda_n^i / \lambda_n^{i+1} \longrightarrow 0 \quad \text{as } n \longrightarrow \infty. \tag{5.2}$$

and $B^{(>i)} : \Delta^1 \times \cdots \times \Delta^i \to \Delta^{i+1} \times \cdots \times \Delta^N$ *is defined recursively by*

$$B^{(>(N-1))}(\pi^{(<N)}) = \beta^N(\tilde{q}^N(\pi^{(<N)})),$$
$$B^{(>(i-1))}(\pi^{(<i)}) = (\beta^i(\tilde{q}^i(\pi^{(<i)})), B^{(>i)}[\pi^{(<i)}, \beta^i(\tilde{q}^i(\pi^{(<i)}))]).$$

Although this looks technical, it can be expressed relatively simply: for any i, if the values (and hence strategies) of players $(1, \ldots, i-1)$ were fixed, the strategies of the more responsive players $(i, \ldots, N)$ would converge to a unique fixed point determined by the functions $\tilde{q}^i$ and $B^{(>i)}$. This assumption is satisfied for any 2-player game: for Q^1 fixed, player 2 simply faces a multi-armed bandit problem. However, it is clearly not always satisfied for general N-player games: in a 3-player partnership game, for Q^1 fixed the other two players still face a partnership game, which might well have more than one Nash distribution, and therefore more than one potential limit point for the more responsive players. We will investigate when Assumption 5.1 is satisfied using a graphical analysis in section 6, but in this section we will retain it as an assumption.

As with the basic individual Q-learning algorithm of section 3, it is immediate that convergence of algorithm (5.1) to a fixed point can occur only at Nash distribution values. Also analogously, we can prove the following lemma.

LEMMA 5.2. *Under Assumption* 5.1*, the values* Q_n^1 *resulting from individual* Q*-learning with PDLR* (5.1) *converge almost surely to a connected internally chain-recurrent set of the flow defined by the singularly perturbed* Q*-learning ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^1(a^1) = r^1(a^1, B^{(>1)}[\beta^1(q_t^1)]) - q_t^1(a^1) \quad \textit{for each } a^1 \in A^1, \tag{5.3}$$

provided that the Q_n *remain bounded for all time, where* $B^{(>1)}$ *is the function defined in Assumption* 5.1*. Additionally,*

$$\|Q_n^i - r^i(\cdot, [\beta^1(Q_n^1), B^{>1}[\beta^1(Q_n^1)]])\|_\infty \longrightarrow 0$$
$$\textit{almost surely for each } i > 1 \textit{ as } n \longrightarrow \infty.$$

Proof. This is immediate from the results of Leslie and Collins (2003) and Assumption 5.1. □

Note that Lemma 5.2 tells us we can analyze the asymptotic behavior of the algorithm as if the values of the remaining players have all converged to the fixed

point determined by the current strategy of the most cautious player, despite the fact that in reality all players adjust their values after every play of the game.

We will proceed to analyze the dynamical system (5.3) in two different ways: in section 5.1 we proceed as in section 4 and relate (5.3) to the singularly perturbed SBR dynamics (Leslie and Collins (2003)), while in section 5.2 we perform a direct analysis in a more restricted class of games.

5.1. Relating the singularly perturbed *Q*-learning ODE to the singularly perturbed SBR dynamics. Leslie and Collins (2003) study the singularly perturbed SBR dynamics, defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^1 = \beta^1[r^1(\cdot, B^{(>1)}[\pi_t^1])] - \pi_t^1. \tag{5.4}$$

As in section 4, we will relate the singularly perturbed Q-learning ODE (5.3) to the singularly perturbed SBR dynamics (5.4).

LEMMA 5.3. *Any connected internally chain-recurrent set of the singularly perturbed Q-learning ODE* (5.3) *is of the form*

$$r^1(\mathcal{C}^1) := \{r^1(\cdot, B^{(>1)}(\pi^1)) : \pi^1 \in \mathcal{C}^1\},$$

where $\mathcal{C}^1 \subset \Delta^1$ is a connected internally chain-recurrent set of flow defined by the singularly perturbed SBR dynamics (5.4)*, and $B^{(>1)}$ is the function defined in Assumption* 5.1.

Proof. This lemma's proof is identical to that of Lemma 4.1, and will not be repeated here. □

As in section 4 this enables us to use previous results (Leslie and Collins (2003)) on more sophisticated forms of learning to prove convergence of the individual Q-learning algorithm with PDLR (5.1) to Nash distribution in certain classes of games.

PROPOSITION 5.4. *The strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *will converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time, in the following games:*

(i) 2*-player zero-sum games,*

(ii) 2*-player partnership games,*

(iii) *Shapley's game* (4.2) *(if Boltzmann action selection is used), and*

(iv) *the N-player matching pennies game (Leslie and Collins* (2003)*) (if the smooth best responses are symmetric under a reordering of the actions).*

Proof. Leslie and Collins (2003) show that Assumption 5.1 holds for these games. The result therefore follows immediately from Lemmas 5.2 and 5.3 and the results of Leslie and Collins (2003) on the connected internally chain-recurrent sets of the singularly perturbed SBR dynamics. □

Thus the individual Q-learning algorithm with PDLR (5.1) is proved to converge in the same games as basic individual Q-learning (3.2). In addition, we have proved that individual Q-learning with PDLR will converge in two games (Shapley's game and the N-player matching pennies game) for which most learning algorithms fail to converge. Indeed the authors know of no adaptive algorithm that converges to either Nash equilibrium or Nash distribution in these games without using PDLR.

We illustrate these results with some numerical experiments using Shapley's game (4.2). As observed in section 4, the basic individual Q-learning algorithm (3.2) will cycle in this game. On the other hand, we have shown that the individual Q-learning algorithm with PDLR (5.1) will converge to the unique Nash distribution values. This is confirmed in Figure 2.

5.2. Direct analysis of the singularly perturbed Q-learning ODE. We now change our emphasis away from the SBR dynamics and instead analyze the singularly perturbed Q-learning ODE (5.3) directly in games where player 1 has 2 actions. We show here that convergence to Nash distribution occurs if Boltzmann action selection is used.

PROPOSITION 5.5. *Suppose player* 1 *has* 2 *actions and uses Boltzmann action selection* (2.2). *Suppose further that the game and smooth best responses are such that there are countably many Nash distributions. Then, under Assumption* 5.1, *the strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time.*

Proof. By Lemma 5.2, the Q^1 values converge to a connected internally chain-recurrent set of the singularly perturbed Q-learning ODE (5.3). To ease notation, for this proof we will write

$$\pi_t(1) = \beta^1(q_t^1)(1) = \frac{e^{q_t^1(1)/\tau}}{e^{q_t^1(1)/\tau} + e^{q_t^1(2)/\tau}} = (1 + e^{\{q_t^1(2) - q_t^1(1)\}/\tau})^{-1} = 1 - \pi_t(2),$$
$$\hat{r}_t(a) = r^1(a, B^{>1}[\beta^1(q_t^1)]), \quad a = 1, 2.$$

Since $\beta^1(q_t^1) = (\pi_t(1), 1 - \pi_t(1))$, $\hat{r}_t(a)$ is a function of the scalar variable $\pi_t(1)$, which is in turn a function of $q_t^1(1)$ and $q_t^1(2)$. Hence

$$\begin{aligned}\frac{d\hat{r}_t(a)}{dt} &= \frac{d\hat{r}_t(a)}{d\pi_t(1)} \left\{ \frac{\partial \pi_t(1)}{\partial q_t^1(1)} \frac{dq_t^1(1)}{dt} + \frac{\partial \pi_t(1)}{\partial q_t^1(2)} \frac{dq_t^1(2)}{dt} \right\} \\ &= \frac{d\hat{r}_t(a)}{d\pi_t(1)} \pi_t(1)\pi_t(2) \frac{d}{dt}\{q_t^1(1) - q_t^1(2)\}.\end{aligned}$$

From this, it follows that

$$\begin{aligned}\frac{d^2}{dt^2}\{q_t^1(1) - q_t^1(2)\} &= \frac{d}{dt}\{\hat{r}_t(1) - \hat{r}_t(2) - q_t^1(1) + q_t^1(2)\} \\ &= \left[\pi_t(1)\pi_t(2) \left\{ \frac{d\hat{r}_t(1)}{d\pi_t(1)} - \frac{d\hat{r}_t(2)}{d\pi_t(1)} \right\} - 1 \right] \frac{d}{dt}\{q_t^1(1) - q_t^1(2)\}.\end{aligned}$$

Therefore $\frac{d}{dt}\{q_t^1(1) - q_t^1(2)\}$ does not change sign, and $\{q_t^1(1) - q_t^1(2)\}$ acts as a Lyapunov function. The result follows from Benaïm (1999, Corollary 6.6). □

Note that Proposition 5.5 provides an independent proof of the convergence of the individual Q-learning algorithm with PDLR (5.1) for the N-player matching pennies game that does not rely on the smooth best responses being symmetric under a reordering of the actions. Furthermore the following immediate corollary shows that the result can be applied directly in a wide class of games.

COROLLARY 5.6. *In a* 2*-player game where player* 1 *has* 2 *actions and uses Boltzmann action selection* (2.2), *the strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time.*

Proof. As already noted, Assumption 5.1 holds automatically in 2-player games. Hence this is immediate from Proposition 5.5. □

This is comparable with a recent result (Berger (2005)) showing that fictitious play approaches equilibrium in nondegenerate $2 \times n$ games, although clearly more information is required for the players in a fictitious play process.

5.3. PDLR and Stackelberg equilibria. Here we discuss a point raised by an anonymous referee, concerning the relationship between PDLR and Stackelberg equilibria (equilibria in which one player selects her strategy first, in the knowledge that the other player will play an optimal action in response). If players used individual Q-learning with PDLR, it might be anticipated that a cautious player could gain an advantage by initially selecting a strategy and then relying on the fact that the responsive player will adapt quickly to this strategy. For example, in the battle of the sexes game with payoff matrix

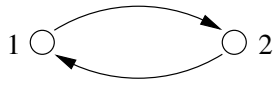
$$\begin{pmatrix} (10,1) & (0,0) \\ (0,0) & (1,10) \end{pmatrix},$$

player 1 could initially select row 1 with high probability, forcing a responsive player 2 to learn to play column 1; player 1 is acting as a Stackelberg leader.

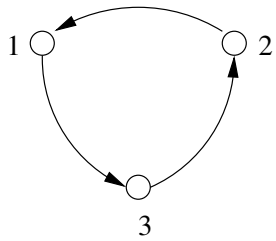
However, in the minimal information situation we consider here, player 1 cannot know in advance which strategy to select to achieve this effect, since the payoff matrix is not known. In the battle of the sexes game, ignoring stochastic effects, if player 1's initial strategy has $\pi^1(2) > 1/11$, then a responsive player 2 will play a strategy in which $\pi^2(2)$ is very nearly 1; the cautious player 1 will then slowly increase $\pi^1(2)$ until joint play reaches the Nash distribution where both players select action 2 with high probability. (Note that this is actually the equilibrium that would have been selected if player 2 was a Stackelberg leader.) Since $\pi^1(2) > 1/11$ with high probability, under reasonable assumptions about the initial Q values of the players, there is a high probability that play will converge to a Nash distribution different from that suggested by the Stackelberg theory. Thus the relationship between PDLR and Stackelberg equilibria is much less clear than might be expected.

6. Graphical analysis. The results of section 5 on individual Q-learning with PDLR rely on Assumption 5.1, which states that there is a unique limit point of the strategies of the more responsive players for any fixed values of the more cautious players $(1, \ldots, i-1)$. Although we have observed that this is always true for 2-player games, in this section we develop a graphical approach to analyzing when this is satisfied in general N-player games, building on previous graphical representations of games (Littman, Kearns, and Singh (2001), Koller and Milch (2003)).

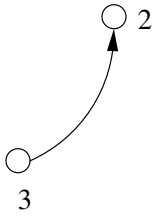
Given a game, we construct a graph by taking a node for each player and drawing a directed arc $\vec{ij}$ if the actions of player i directly affect the rewards of player j. Thus the graph of a (generic) 2-player game is given by



whereas the graph of the 3-player matching pennies game (Jordan (1993)), in which the rewards of player i are affected only by the actions of player $i+1$ (modulo 3), is given by



These graphs can be used to investigate Assumption 5.1 by removing node 1 (corresponding to the most cautious player) from the graph, then considering the behavior of the other players. The intuition behind this is that the fixed strategy of the most cautious player in Assumption 5.1 results in the other players facing a reduced game. For example, in the 3-player matching pennies game, removal of node 1 (and connected arcs) results in the graph



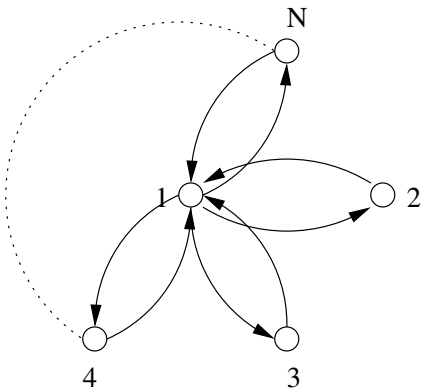
Thus for fixed Q^1, player 3's rewards are not affected by any other player, so Q_n^3 converges to unique values. But because of this, player 2's values must also converge to unique values, and so for fixed Q^1, the values (Q_n^2, Q_n^3) converge to a unique point, as required by Assumption 5.1.

In fact, this example generalizes very simply. If, after removal of a node, a graph has no directed cycle, then the players corresponding to nodes left in the graph will converge to a unique fixed point.

PROPOSITION 6.1. *Suppose that a game is such that removal of node* 1 *from the game graph results in a subgraph containing no directed cycles. Then Assumption* 5.1 *holds for this game.*

Proof. If the subgraph remaining after removal of node 1 has no directed cycle, then there exists at least one node with in-degree 0 (i.e., no arcs terminate at this node). The values of a player associated with such a node do not depend on the strategies of any other players (except perhaps the fixed strategy of player 1), and so the Q values of that player will converge almost surely to a unique point. Since the rewards, and hence strategies, of any player corresponding to a node with in-degree 0 are uniquely determined, given the fixed strategy of player 1, these nodes can be removed from the graph. This again results in a graph with no directed cycles, and we can proceed recursively to show that the rewards of all the players are uniquely determined by the values of player 1, as required in Assumption 5.1. □

From this proposition, it is immediate that Assumption 5.1 holds in games which have a graph with no directed cycle even before removal of a node, and games with a single directed cycle will clearly become acyclic if the node to be removed is part of that cycle. Consider also games with a star graph:



Here there is a distinguished player (the hub) connected to all others, but all non-distinguished players are disconnected from each other. This graph will become completely disconnected (and so obviously will have no directed cycle) if the node corresponding to the distinguished player is removed. A game such as this could have applications in computing, for example, where the distinguished player is a central resource, such as a server or router, and the other nodes correspond to users of the resource.

Although useful, this graphical approach is not sufficient in all situations. For example, consider a 3-player game in which the rewards of players 2 and 3 always sum to 0 for any fixed Q^1. Thus, the graph after removal of node 1 is the same as that of a 2-player game, and so (generically) contains a directed cycle. However, the resulting game is a 2-player zero-sum game, in which the players converge to a unique Nash distribution (Proposition 5.4 and Hofbauer and Hopkins (2005)). Therefore Assumption 5.1 is satisfied, even though the conditions of Proposition 6.1 are not.

7. Conclusion. We have shown that value-based learning agents cannot generally converge to a Nash equilibrium of a game, but if smooth best responses are used, a Nash distribution can be reached. Although Nash distributions are not generally the same as Nash equilibria, they are close if the temperature parameter of the smooth best responses is sufficiently small, and therefore we proposed that value-based learning agents should use smooth best responses to allow equilibrium play, even if this is not a classical Nash equilibrium.

Our value-based learning algorithm, individual Q-learning (3.2), is very similar to the simple and successful algorithm used by Sutton and Barto (1998) in the single-agent multi-armed bandit problem. We showed that convergence to a point that is not a Nash distribution is not possible and that the value estimates are asymptotically belief-based. Further, by relating the limiting behavior of the individual Q-learning algorithm to the SBR dynamics (a system previously used to characterize more sophisticated models of learning), it was shown that strategies of players converge almost surely to a Nash distribution for 2-player zero-sum games and 2-player partnership games.

The nonconvergence of strategies for certain games motivates the introduction of individual Q-learning with PDLR (5.1), resulting in cautious and responsive players. This modified algorithm converges to Nash distribution for the same games as basic individual Q-learning (3.2), and also for Shapley's game and the N-player matching pennies game. Moreover, convergence was shown to occur for any game in which player 1, the most cautious, has only 2 actions.

Finally, since the results on PDLR rely on an assumption about the behavior of the more responsive players if the values of player 1 were fixed, a simple graphical method was introduced to help determine when this assumption holds.

Acknowledgments. The authors thank three anonymous referees for helpful comments and Prof. Josef Hofbauer and Dr. Andy Wright for helpful discussions in the course of the research leading to this paper.

REFERENCES

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995), *Gambling in a rigged casino: The adversarial multi-armed bandit problem*, in 36th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, pp. 322–331.

R. J. Aumann (1974), *Subjectivity and correlation in randomized strategies*, J. Math. Econom., 1, pp. 67–96.

A. Baños (1968), *On pseudo-games*, Ann. Math. Statist., 39, pp. 1932–1945.

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

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

U. Berger (2005), *Fictitious play in $2 \times n$ games*, J. Econom. Theory, 120, pp. 139–154.

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

V. S. Borkar (2001), *Reinforcement learning in Markovian evolutionary games*, available at http://www.tcs.tifr.res.in/∼borkar/game.ps.

M. Bowling and M. Veloso (2002), *Multiagent learning using a variable learning rate*, Artificial Intelligence, 136, pp. 215–250.

J. A. Boyan and M. L. Littman (1994), *Packet routing in dynamically changing networks: A reinforcement learning approach*, in Advances in Neural Information Processing Systems, Vol. 6, J. D. Cowan, G. Tesauro, and J. Alspector, eds., Morgan Kaufmann, San Francisco, pp. 671–678.

G. W. Brown (1951), *Iterative solution of games by fictitious play*, in Activity Analysis of Production and Allocation, T. C. Koopmans, ed., John Wiley & Sons, New York, pp. 374–376.

H.-F. Chen and Y.-M. Zhu (1986), *Stochastic approximation procedures with randomly varying truncations*, Sci. China Ser. A, 29, pp. 914–926.

S. Cowan (1992), *Dynamical Systems Arising from Game Theory*, Ph.D. thesis, University of California, Berkeley.

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

D. P. Foster and H. P. Young (2003), *Learning, hypothesis testing, and Nash equilibrium*, Games Econom. Behav., 45, pp. 73–96.

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

D. Fudenberg and J. Tirole (1991), *Game Theory*, MIT Press, Cambridge, MA.

P. W. Glimcher, M. C. Dorris, and H. M. Bayer (2005), *Physiological utility theory and the neuroeconomics of choice*, Games Econom. Behav., to appear.

S. Govindan, P. J. Reny, and A. J. Robson (2003), *A short proof of Harsanyi's purification theorem*, Games Econom. Behav., 45, pp. 369–374.

J. Hannan (1957), *Approximation to Bayes risk in repeated play*, in Contributions to the Theory of Games, Vol. 3, Ann. of Math. Stud. 39, M. Drescher, A. W. Tucker, and P. Wolfe, eds., Princeton University Press, Princeton, NJ, pp. 97–139.

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

S. Hart and A. Mas-Colell (2000), *A simple adaptive procedure leading to correlated equilibrium*, Econometrica, 68, pp. 1127–1150.

S. Hart and A. Mas-Colell (2001), *A reinforcement procedure leading to correlated equilibrium*, in Economic Essays: A Festschrift for Werner Hildenbrand, W. N. G. Debreu and W. Trockel, eds., Springer-Verlag, Berlin, pp. 181–200.

S. Hart and A. Mas-Colell (2003), *Uncoupled dynamics cannot lead to Nash equilibrium*, Amer. Econom. Rev., 93, pp. 1830–1836.

J. Hofbauer and E. Hopkins (2005), *Learning in perturbed asymmetric games*, Games Econom. Behav., 52, pp. 133–152.

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

D. Koller and B. Milch (2003), *Multi-agent influence diagrams for representing and solving games*, Games Econom. Behav., 45, pp. 181–221.

H. J. Kushner and G. G. Yin (1997), *Stochastic Approximation Algorithms and Applications*, Appl. Math. 35, Springer-Verlag, New York.

D. S. Leslie and E. J. Collins (2003), *Convergent multiple-timescales reinforcement learning algorithms in normal form games*, Ann. Appl. Probab., 13, pp. 1231–1251.

M. L. Littman, M. Kearns, and S. Singh (2001), *An efficient, exact algorithm for solving tree-structured graphical games*, in Advances in Neural Information Processing Systems, Vol. 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, Cambridge, MA.

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

N. Megiddo (1980), *On repeated games with incomplete information played by non-Bayesian players*, Internat. J. Game Theory, 9, pp. 157–167.

J. Nash (1950), *Equilibrium points in n-person games*, Proc. Natl. Acad. Sci. USA, 36, pp. 48–49.

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

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

L. S. Shapley (1964), *Some topics in two person games*, in Advances in Game Theory, M. Drescher, L. S. Shapley, and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, pp. 1–28.

S. Singh and D. Bertsekas (1997), *Reinforcement learning for dynamic channel allocation in cellular telephone systems*, in Advances in Neural Information Processing Systems, Vol. 9, M. C. Mozer, M. I. Jordan, and T. Petsche, eds., MIT Press, Cambridge, MA, p. 974.

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