Available online at www.sciencedirect.com
ScienceDirect
Games and Economic Behavior 61 (2007) 259–276
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Transient and asymptotic dynamics of reinforcement learning in games

Luis R. Izquierdo [a,*], Segismundo S. Izquierdo [b], Nicholas M. Gotts [c], J. Gary Polhill [c]

[a] *Universidad de Burgos, Edificio la Milanera, C/Villadiego s/n, 09001, Burgos, Spain*
[b] *Department of Industrial Organization, University of Valladolid, 47011, Spain*
[c] *The Macaulay Institute, Craigiebuckler, AB15 8QH, Aberdeen, UK*

Received 25 May 2005

Available online 6 April 2007

Abstract

Reinforcement learners tend to repeat actions that led to satisfactory outcomes in the past, and avoid choices that resulted in unsatisfactory experiences. This behavior is one of the most widespread adaptation mechanisms in nature. In this paper we fully characterize the dynamics of one of the best known stochastic models of reinforcement learning [Bush, R., Mosteller, F., 1955. Stochastic Models of Learning. Wiley & Sons, New York] for 2-player 2-strategy games. We also provide some extensions for more general games and for a wider class of learning algorithms. Specifically, it is shown that the transient dynamics of Bush and Mosteller's model can be substantially different from its asymptotic behavior. It is also demonstrated that in general—and in sharp contrast to other reinforcement learning models in the literature—the asymptotic dynamics of Bush and Mosteller's model cannot be approximated using the continuous time limit version of its expected motion.
© 2007 Elsevier Inc. All rights reserved.

JEL classification: C73

Keywords: Reinforcement learning; Bush and Mosteller; Learning in games; Stochastic approximation; Slow learning; Distance diminishing

* Corresponding author. Fax: +34 947 258910.
E-mail addresses: luis@izquierdo.name, lrizquierdo@ubu.es (L.R. Izquierdo).

0899-8256/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.geb.2007.01.005

1. Introduction

Reinforcement learners use their experience to choose or avoid certain actions based on their consequences. Actions that led to satisfactory outcomes (i.e. outcomes that met or exceeded aspirations) will tend to be repeated in the future, whereas choices that led to unsatisfactory experiences are avoided. This paper characterizes the dynamics of a variant of Bush and Mosteller's (1955) linear stochastic model of reinforcement learning for 2×2 (i.e. 2-player 2-strategy) games.

The empirical study of reinforcement learning as a crucial facet of (human and non-human) animal behavior finds its roots in Thorndike's experiments on instrumental learning at the end of the 19th century (Thorndike, 1898). The results of these experiments were formalized in the 'law of effect', one of the most robust properties of learning in the experimental psychology literature. Reinforcement learning is an important aspect of much learning in almost all animals, including many phylogenetically very distant from vertebrates (e.g., earthworms, Maier and Schneirla, 1964, and fruit flies, Wustmann et al., 1996). In strategic contexts, empirical evidence for reinforcement learning is strongest in animals with limited reasoning abilities or in human subjects who have no information beyond the payoff they receive and specifically may be unaware of the strategic nature of the situation (Duffy, 2006; Camerer, 2003; Bendor et al., 2001a; Roth and Erev, 1995; Mookherjee and Sopher, 1994). In the context of experimental game theory with human subjects, several authors have used simple models of reinforcement learning to successfully explain and predict behavior in a wide range of games (McAllister, 1991; Roth and Erev, 1995; Mookherjee and Sopher, 1994, 1997; Chen and Tang, 1998; Erev and Roth 1998, 2001; Erev et al., 1999). Reinforcement models in the literature tend to differ in the following, somewhat interrelated, features:

- Whether learning slows down or not, i.e. whether the model accounts for the 'power law of practice' (e.g., Erev and Roth, 1998 vs. Börgers and Sarin, 1997).
- Whether the model allows for avoidance behavior in addition to approach behavior (e.g., Bendor et al., 2001b vs. Erev and Roth, 1998). Approach behavior is the tendency to repeat the associated choices after receiving a positive stimulus; avoidance behavior is the tendency to avoid the associated actions after receiving a negative stimulus (one that does not satisfy the player). Models that allow for negative stimuli tend to define an aspiration level against which achieved payoffs are evaluated. This aspiration level may be fixed or vary endogenously (Bendor et al., 2001a, 2001b).
- Whether "forgetting" is considered, i.e. whether recent observations weigh more than distant ones (Erev and Roth, 1998; Rustichini, 1999; Beggs, 2005).
- Whether the model imposes inertia—a positive bias in favor of the most recently selected action (Bendor et al., 2001a, 2001b).

Laslier et al. (2001) present a more formal comparison of various reinforcement learning models. Each of the features above can have important implications for the behavior of the particular model under consideration and for the mathematical methods that are adequate for its analysis. For example, when learning slows down, theoretical results from the theory of stochastic approximation (Benveniste et al., 1990; Kushner and Yin, 1997) and from the theory of urn models can often be applied (e.g., Ianni, 2001; Hopkins and Posch, 2005; Beggs, 2005), whereas if the learning rate is constant, results from the theory of distance diminishing models (Norman, 1968, 1972) tend to be more useful (e.g., Börgers and Sarin, 1997;

Bendor et al., 2001b). Similarly, imposing inertia facilitates the analysis to a great extent, since it often ensures that a positive stimulus will be followed by an increase in the probability weight on the most recently selected action at some minimal geometric rate (Bendor et al., 2001b).

A popular model of reinforcement learning in the game theory literature is the Erev–Roth (ER) model (Roth and Erev, 1995; Erev and Roth, 1998). Understanding of the ER model (also called cumulative proportional reinforcement model by Laslier et al., 2001 and Laslier and Walliser, 2005) and its relation with an adjusted version of the evolutionary replicator dynamics (Weibull, 1995) has been developed in papers by Laslier et al. (2001), Hopkins (2002), Laslier and Walliser (2005), Hopkins and Posch (2005), and Beggs (2005). An extension to the ER model covering both partial and full informational environments (in the latter, a player can observe the payoffs for actions not selected), as well as linear and exponential adjustment procedures, is analyzed for single person decision problems by Rustichini (1999).

Arthur (1991) proposed a model differing from the ER model only in that the step size of the learning process in ER is stochastic whereas it is deterministic in Arthur's model—but step sizes are of the same order in both (see Hopkins and Posch, 2005 for details). Theoretical results for Arthur's model in games and its relation with the ordinary evolutionary replicator dynamics are given by Posch (1997), Hopkins (2002), Hopkins and Posch (2005), and Beggs (2005): despite their similarity, the ER model and Arthur's model can have different asymptotic behavior (Hopkins and Posch, 2005).

Another important set of reinforcement models are the aspiration-based models, which allow for negative stimuli (see Bendor et al., 2001a for an overview). The implications of aspiration-based reinforcement learning in strategic contexts have been studied thoroughly by Karandikar et al. (1998) and Bendor et al. (2001b). This line of work tends to require very mild conditions on the way learning is conducted apart from the assumption of inertia. Assuming inertia greatly facilitates the mathematical analysis, enabling the derivation of sharp predictions for long-run outcomes in 2-player repeated games, even with evolving aspirations (see, e.g., Karandikar et al., 1998; Palomino and Vega-Redondo, 1999; Bendor et al., 2001b).

The model analyzed here is a variant of Bush and Mosteller's (1955) linear stochastic model of reinforcement learning (henceforth BM model). The BM model is an aspiration-based reinforcement learning model, but does not impose inertia. In contrast to the ER model and Arthur's model, it allows for negative stimuli and learning does not fade with time. A special case of the BM model where all stimuli are positive was originally considered by Cross (1973), and analyzed by Börgers and Sarin (1997), who also related it to the replicator dynamics. Börgers and Sarin (2000) studied an extension of the BM model where aspirations evolve simultaneously with choice probabilities in single person decision contexts. Here, we develop Börgers and Sarin's work by analyzing the dynamics of the BM model in 2×2 games where aspiration levels are fixed, but not necessarily below the lowest payoff, so negative stimuli are possible. These dynamics have been explored by Macy and Flache (2002) and Flache and Macy (2002) in 2×2 social dilemmas using computer simulation. Here we formalize their qualitative analysis and extend their results for any 2×2 game.

In contrast to other reinforcement learning models in the literature, we show that the asymptotic behavior of the BM model cannot be approximated using the continuous time limit version of its expected motion. Such an approximation can be valid over bounded time intervals but deteriorates as the time horizon increases. This important point—originally emphasized by Boylan (1992, 1995) in a somewhat different context—was already noted by Börgers and Sarin (1997) in the BM model for strictly positive stimuli, and has also been found in other models since then (Beggs, 2002). The asymptotic behavior of the BM model is characterized in the present paper

using the theory of distance diminishing models (Norman, 1968, 1972). Börgers and Sarin (1997) also used this theory to analyze the case where aspirations are below the minimum payoff; here we extend their results for 2×2 games where aspiration levels can have any fixed value.

2. The BM model

Consider a normal-form 2×2 game with *set of players* $I = \{1, 2\}$, pure-strategy space D_i for each player $i \in I$, and *payoff functions* u_i that give player i's payoff for each profile $\boldsymbol{d} = (d_1, d_2)$ of pure strategies, where $d_i \in D_i$ is a pure strategy for player i. Let $D = \times_{i \in I} D_i$ be the space of pure-strategy profiles, or possible outcomes of the game. We can represent any mixed strategy for player i as a *vector* $\boldsymbol{p}_i$ in the *unit simplex* Δ^1, where the jth coordinate $p_{i,j} \in \mathbb{R}$ of the vector $\boldsymbol{p}_i$ is the probability assigned by $\boldsymbol{p}_i$ to player i's jth pure strategy. A *mixed-strategy profile* is a vector $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2)$, where each component $\boldsymbol{p}_i \in \Delta^1$ represents a mixed strategy for player $i \in I$.

In the BM model, after each player i has selected an action according to their mixed strategy $\boldsymbol{p}_i$, strategy revision follows a reinforcement learning approach: players increase the probability of actions leading to payoffs above their aspiration level, and decrease the probability of actions that led to payoffs below aspiration level. Players use only information concerning their own past choices and payoffs, ignoring the payoffs and choices of their counterparts. More precisely, strategy updating takes place in two steps. First, after outcome $\boldsymbol{d}^n = (d_1^n, d_2^n)$ in time-step n, each player i calculates her stimulus $s_i(\boldsymbol{d}^n)$ for the action just chosen d_i^n according to the following formula:

$$s_i(\boldsymbol{d}) = \frac{u_i(\boldsymbol{d}) - A_i}{\sup_{\boldsymbol{k} \in D} |u_i(\boldsymbol{k}) - A_i|},$$

where A_i is player i's aspiration level. Hence the stimulus is always a number in the interval $[-1, 1]$. Note that players are assumed to know $\sup_{\boldsymbol{k} \in D} |u_i(\boldsymbol{k}) - A_i|$. Secondly, having calculated their stimulus $s_i(\boldsymbol{d}^n)$ after the outcome $\boldsymbol{d}^n$, each player i updates her probability p_{i,d_i} of undertaking the selected action d_i as follows

$$p_{i,d_i}^{n+1} = \begin{cases} p_{i,d_i}^n + l_i \cdot s_i(\boldsymbol{d}^n) \cdot (1 - p_{i,d_i}^n) & \text{if } s_i(\boldsymbol{d}^n) \geqslant 0, \\ p_{i,d_i}^n + l_i \cdot s_i(\boldsymbol{d}^n) \cdot p_{i,d_i}^n & \text{if } s_i(\boldsymbol{d}^n) < 0, \end{cases} \tag{1}$$

where p_{i,d_i}^n is player i's probability of undertaking action d_i in time-step n, and l_i is player i's learning rate ($0 < l_i < 1$). Thus, the higher the stimulus magnitude (or the learning rate), the larger the change in probability. The updated probability for the action not selected derives from the constraint that probabilities must add up to one.

A 2×2 BM model parameterization requires specifying both players' payoff function u_i, aspiration level (A_i), and learning rate (l_i). A parameterized model will be denoted $\boldsymbol{S}$ (for System). Since the state of any particular system can be fully characterized by the strategy profile $\boldsymbol{p}$, $\boldsymbol{p}$ will also be named *state of the system*. Note that there are only two independent variables in $\boldsymbol{p}$. Let $\boldsymbol{P}_n(\boldsymbol{S})$ be the state of a system $\boldsymbol{S}$ in time-step n. Note that $\boldsymbol{P}_n(\boldsymbol{S})$ is a random variable and $\boldsymbol{p}$ is a particular value of that variable; the sequence of random variables $\{\boldsymbol{P}_n(\boldsymbol{S})\}_{n \geqslant 0}$ constitutes a discrete-time Markov process with potentially infinite transient states. In a slight abuse of notation we refer to such a process $\{\boldsymbol{P}_n(\boldsymbol{S})\}_{n \geqslant 0}$ as the BM process $\boldsymbol{P}_n$.

3. Attractors in the dynamics of the system

Using computer simulation, Macy and Flache (2002) described two types of learning-theoretic equilibria that govern the dynamics of the BM model: self-reinforcing equilibria (SRE), and self-correcting equilibria (SCE). These are not static equilibria, but strategy profiles which act as attractors in the sense that, under certain conditions, the system will tend to approach them or linger around them. Here, we formalize these two concepts.

We define an SRE as an absorbing state of the system (i.e. a state $\boldsymbol{p}$ that cannot be abandoned) where both players receive a positive stimulus. An SRE corresponds to a pair of pure strategies such that its certain associated outcome gives a strictly positive stimulus to both players (henceforth a *mutually satisfactory outcome*). Escape from an SRE is impossible since no player will change her strategy. Moreover, SREs act as attractors in the sense that if the system is near to an SRE, then it will probably move further towards it. This is because near an SRE there is a high probability that its associated mutually satisfactory outcome will occur, and this brings the system even closer to the SRE. The number of SREs in a system is the number of outcomes where both players obtain payoffs above their respective aspiration levels.

To formalize the concept of SCE, we need to consider the *expected* motion of the system and its associated ordinary differential equation (ODE). The expected motion (EM) of a system $\boldsymbol{S}$ in state $\boldsymbol{p}$ is given by a vector function $\mathrm{EM}^{\boldsymbol{S}}(\boldsymbol{p})$ whose components are, for each player, the expected change in the probabilities of undertaking each of the two possible actions. Mathematically,

$$\mathrm{EM}^{\boldsymbol{S}}(\boldsymbol{p}) \equiv \mathrm{E}(\Delta \boldsymbol{P}_n \mid \boldsymbol{P}_n = \boldsymbol{p}).$$

Hence, the ODE formed by taking the expected motion of the stochastic process $\boldsymbol{P}_n$ is

$$\dot{\boldsymbol{f}} = \mathrm{EM}^{\boldsymbol{S}}(\boldsymbol{f}). \tag{2}$$

We define an SCE of a system $\boldsymbol{S}$ as an asymptotically stable critical point of the continuous time limit approximation of its expected motion (given by Eq. (2)). Note that a particular state of the system could be an SRE and an SCE at the same time. An illustration of the phase plane of the ODE corresponding to a system where two players with the same learning rate l and aspirations $A_i = 2$ are playing a symmetric Prisoner's Dilemma with payoffs $[4, 3, 1, 0]$ is shown in Fig. 1.

A crucial question to characterize the dynamics of learning models, and one to which stochastic approximation theory (Benveniste et al., 1990; Kushner and Yin, 1997) is devoted, is whether the *expected* and *actual* motion of the system should become arbitrarily close in the long run. This is generally true for processes whose motion slows down at an appropriate rate (as explained by Hopkins and Posch, 2005 when studying the ER model), but not necessarily so in other cases. We show in the next sections that the BM model's *asymptotic* behavior can be dramatically different from that suggested by its associated ODE, which is, however, very relevant for characterizing the *transient* dynamics of the system, particularly with small learning rates.

4. Three dynamic regimes

In the general case, the dynamics of the BM model may exhibit three different regimes: medium run, long run, and ultralong run. This terminology is borrowed from Binmore and Samuelson (1993) and Binmore et al. (1995, p. 10), who reserve the term short run for the initial conditions. The medium run is '*the time intermediate between the short run* [i.e. initial conditions] *and the long run, during which the adjustment to equilibrium is occurring*.' The long

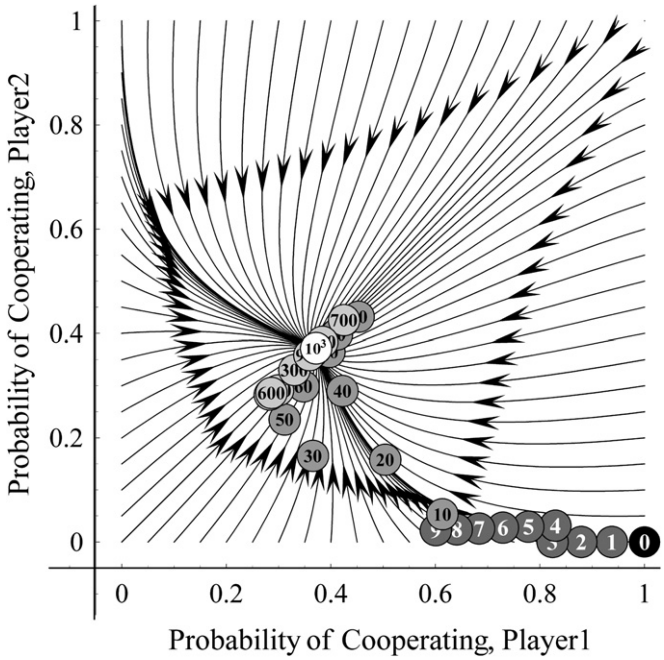

Fig. 1. Trajectories in the phase plane of the ODE associated with a system where two players with the same learning rate $l_i = l$ and aspirations $A_i = 2$ are playing a symmetric Prisoner's Dilemma with payoffs [4, 3, 1, 0]. This system has a unique SCE at $[p_{1,C}, p_{2,C}] = [0.37, 0.37]$ and a unique SRE at $[p_{1,C}, p_{2,C}] = [1, 1]$. The numbered balls show the state of the system after the indicated number of iterations in a sample simulation run ($l_i = 2^{-4}$; initial state $[p_{1,C}, p_{2,C}] = [1, 0]$).

run is '*the time span needed for the system to reach the vicinity of the first equilibrium in whose neighborhood it will linger for some time*.' Finally, the ultralong run is '*a period of time long enough for the asymptotic distribution to be a good description of the behavior of the system*.'

Binmore et al.'s terminology is particularly useful for our analysis because it is often the case in the BM model that the transient (i.e. medium and long run) dynamics of the system are dramatically different from its asymptotic (i.e. ultralong run) behavior. Whether the three different regimes are clearly distinguishable strongly depends on the players' learning rates. For high learning rates the system quickly reaches its asymptotic behavior and the distinction between the different regimes is not particularly useful. For small learning rates, however, the three different regimes can be clearly observed.

In brief, it is shown in this paper that with sufficiently small learning rates l_i and number of iterations n not too large ($n \cdot l_i$ bounded), the medium run dynamics of the system are best characterized by the trajectories in the phase plane of Eq. (2). Under these conditions, SCEs constitute the '*the first equilibrium in whose neighborhood it* [the system] *will linger for some time*' and, as such, they usefully characterize the long run dynamics of the system. After a potentially very lengthy long-run regime in the neighborhood of an SCE, the system will eventually reach its ultralong run behavior, which in most BM systems consists in approaching an SRE asymptotically (see formal analysis below).

For an illustration of the different regimes, consider a system where two players with aspirations $A_i = 2$ and learning rates $l_i = l$ are playing a symmetric Prisoner's Dilemma with payoffs

[4, 3, 1, 0]. This system has a unique SCE at $[p_{1,C}, p_{2,C}] = [0.37, 0.37]$ and a unique SRE at $[p_{1,C}, p_{2,C}] = [1, 1]$, where $p_{i,C}$ denotes player i's probability to cooperate. It is shown below that this system asymptotically converges to its unique SRE with probability 1 regardless of the value of l. The evolution of the probability to cooperate (which is identical for both players) is represented in the rows of Fig. 2 for different values of l. Initially, both players' probability to cooperate is 0.5. The phase plane of the corresponding ODE is shown in Fig. 1.

With high learning rates (e.g., see top row in Fig. 2) the ultralong run behavior is quickly reached. On the other hand, with low learning rates the two transient regimes become apparent; these are closely related to the trajectories of Eq. (2), which follow paths substantially apart from the end-state of the system (Fig. 1). Note that asymptotic convergence to mutual cooperation is guaranteed in the seven systems represented in the rows of Fig. 2, but this cannot be appreciated in the bottom rows, which do not go beyond their long run behavior.

The following sections are devoted to the formal analysis of the transient and asymptotic dynamics of the BM model. The proofs of every proposition in the paper are included in Appendix A.

5. Transient dynamics

As mentioned above, when learning takes place by large steps the system quickly reaches its asymptotic behavior, and no clear (transient) patterns are observed before it does so (see top row of Fig. 2). With small learning rates, however, the two transient regimes, which may be significantly different from the asymptotic regime, are clearly distinguishable. This section shows that SCEs are powerful attractors of the *actual* dynamics of the system when learning occurs by small steps. Specifically, it is demonstrated that the BM process $\boldsymbol{P}_n$ follows the trajectories of its associated ODE with probability approaching 1 as learning rates decrease and n is kept within certain limits.

Consider a family of BM systems $\boldsymbol{S}^l$ whose members, indexed in $l = l_1$, only differ in both players' learning rates, and such that l_1/l_2 is a fixed constant for every model in the family. Let $\boldsymbol{P}^l_n = \boldsymbol{P}_n(\boldsymbol{S}^l)$ be the family of stochastic processes associated with such a family of systems $\boldsymbol{S}^l$. As an example, note that Fig. 2 shows simulation runs of seven stochastic processes $(\boldsymbol{P}_n(\boldsymbol{F}^{0.5}), \boldsymbol{P}_n(\boldsymbol{F}^{0.25}), \ldots)$ belonging to one particular family $\boldsymbol{F}^l$. Consider the ODE given by Eq. (3) below, and let $\boldsymbol{f}_{\boldsymbol{x}}(t)$ be the trajectory of this ODE with initial state $\boldsymbol{x}$.

$$\dot{\boldsymbol{f}} = \frac{1}{l}\mathrm{EM}^{\boldsymbol{S}^l}(\boldsymbol{f}). \tag{3}$$

The ODE in Eq. (3) is common to every member of a given family, and its solution trajectories $\boldsymbol{f}_{\boldsymbol{x}}(t)$ only differ from those given by Eq. (2) (which determines a different ODE for each member) in the time scale, i.e. the representation of the trajectories of ODEs (2) and (3) in the phase plane is identical: the learning rate determines how quickly the path is walked, but the path is the same for every model of a family. Similarly, SCEs and SREs are common to every model in a family. The following proposition characterizes the medium-run (statements (i) and (ii)) and the long-run (statement (iii)) dynamics of the BM model when l is small. No conditions are imposed on players' aspirations.

Proposition 1. *Consider the family of stochastic processes* $\{\boldsymbol{P}^{l,\boldsymbol{x}}_n\}_{n\geqslant 0}$ *with initial state* $\boldsymbol{P}^l_0 = \boldsymbol{x}$ *for every* l. *Let* K *be an arbitrary constant. For learning by small steps* $(l \to 0)$ *and transient behavior* $(n \cdot l \leqslant K < \infty)$, *we have*:

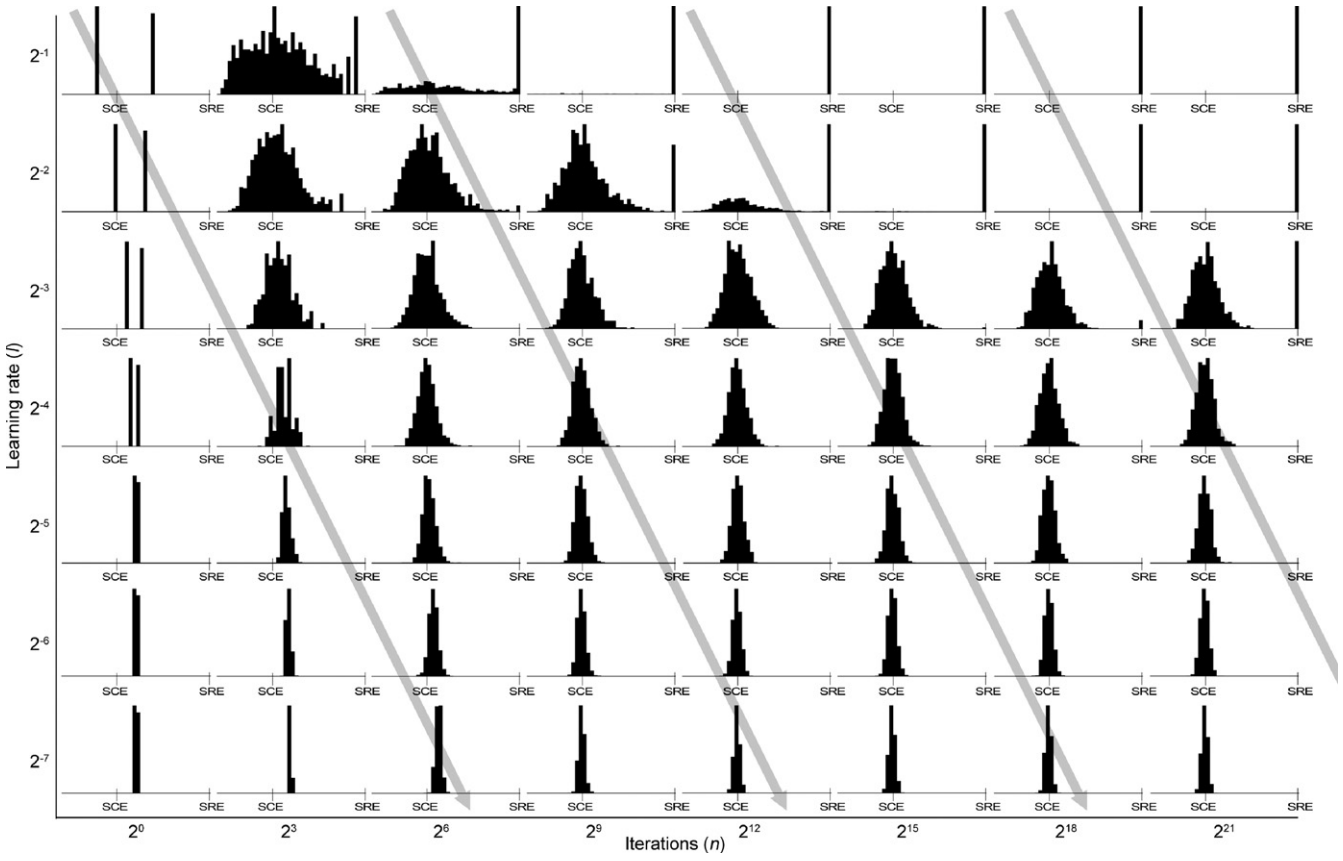

Fig. 2. Histograms representing the probability to cooperate for one player (both players' probabilities are identical) after n iterations, for different learning rates $l_i = l$, with $A_i = 2$, in a symmetric Prisoner's Dilemma with payoffs [4, 3, 1, 0]. Each histogram has been calculated over 1000 simulation runs. The initial probability for both players is 0.5. The significance of the gray arrows will be explained later in the text.

(i) *For fixed* $\varepsilon > 0$ *and* l *sufficiently small,*

$$\Pr\left\{\max_{n\leqslant(K/l)}\left\|\boldsymbol{P}_n^{l,\boldsymbol{x}} - \boldsymbol{f}_{\boldsymbol{x}}(n\cdot l)\right\| > \varepsilon\right\} \leqslant C(l,K),$$

where, for fixed $K < \infty$, $C(l,K) \to 0$ *as* $l \to 0$. *Thus, for transient behavior and learning by small steps, we have uniform convergence in probability of* $\boldsymbol{P}_n^{l,\boldsymbol{x}}$ *to the trajectory* $\boldsymbol{f}_{\boldsymbol{x}}$ *of the ODE in* (3).

(ii) *The distribution of the variable* $(\boldsymbol{P}_n^{l,\boldsymbol{x}} - \boldsymbol{f}_{\boldsymbol{x}}(n\cdot l))/\sqrt{l}$ *converges to a normal distribution with mean* 0 *and variance independent of* l *as* $l \to 0$ *and* $n\cdot l \to K < \infty$.

(iii) *Let* $\mathrm{L}_{\boldsymbol{x}}$ *be the limit set of the trajectory* $\boldsymbol{f}_{\boldsymbol{x}}(t)$. *For* $n = 0, 1, \ldots, N < \infty$, *and for any* $\delta > 0$, *the proportion of values of* $\boldsymbol{P}_n^{l,\boldsymbol{x}}$ *within a neighborhood* $\mathrm{B}_\delta(\mathrm{L}_{\boldsymbol{x}})$ *of* $\mathrm{L}_{\boldsymbol{x}}$ *goes to* 1 (*in probability*) *as* $l \to 0$ *and* $N\cdot l \to \infty$.

To see an application of Proposition 1, consider the particular family $\boldsymbol{F}^l$ (Fig. 2). Statement (i) says that when n is not too large ($n\cdot l$ bounded), with probability increasingly close to 1 as l decreases, the process $\boldsymbol{P}_n^{\boldsymbol{x}}(\boldsymbol{F}^l)$ with initial state $\boldsymbol{P}_0(\boldsymbol{F}^l) = \boldsymbol{x}$ follows the trajectory $\boldsymbol{f}_{\boldsymbol{x}}(n\cdot l)$ of the ODE in (3) within a distance never greater than some arbitrary, a priori fixed, $\varepsilon > 0$. (This proves the conjecture put forward by Börgers and Sarin (1997) in Remark 2.) The trajectories corresponding to $\boldsymbol{P}_n(\boldsymbol{F}^l)$ are displayed in Fig. 1, and the convergence of the processes to the appropriate point in the trajectory $\boldsymbol{f}_{\boldsymbol{x}}(n\cdot l)$ as $l \to 0$ can be appreciated following the gray arrows (which join histograms for which $n\cdot l$ is constant) in Fig. 2. These arrows also illustrate statement (ii): the distribution of $\boldsymbol{P}_n^{\boldsymbol{x}}(\boldsymbol{F}^l)$ approaches normality with decreasing variance as $l \to 0$, keeping $n\cdot l$ constant. The fact that the trajectory $\boldsymbol{f}_{\boldsymbol{x}}$ is a good approximation for the medium-run dynamics of the system for slow learning shows the importance of SCEs as attractors of the actual dynamics of the system. To illustrate this, consider family $\boldsymbol{F}^l$ again. It can be shown using the square of the Euclidean distance to the SCE as a Liapunov function that every trajectory starting in any state different from the SRE $[p_{1,C}, p_{2,C}] = [1, 1]$ will end up in the SCE $[p_{1,C}, p_{2,C}] = [0.37, 0.37]$—i.e. the limit set $\mathrm{L}_{\boldsymbol{x}}$ is formed exclusively by the SCE for any $\boldsymbol{x} \neq$ SRE (see Fig. 1). This means that starting from any initial state $\boldsymbol{x} \neq$ SRE, if K is sufficiently large and $n < K/l$ (i.e. if in Fig. 2 we consider the region to the left of a gray arrow that is sufficiently to the right), the distribution of $\boldsymbol{P}_n^{\boldsymbol{x}}(\boldsymbol{F}^l)$ will be tightly clustered around the SCE [0.37, 0.37] and will approach normality as n increases. Furthermore, statement (iii) says that, for any $\boldsymbol{x} \neq$ SRE, any $\delta > 0$, and $n = 0, 1, \ldots, N < \infty$, the proportion of values of $\boldsymbol{P}_n^{\boldsymbol{x}}(\boldsymbol{F}^l)$ within a neighborhood B_δ (SCE) of the SCE goes to 1 (in probability) as $l \to 0$ and $N\cdot l \to \infty$. This is the long run. Remember, however, that given any l, $\boldsymbol{P}_n^{\boldsymbol{x}}(\boldsymbol{F}^l)$ will eventually converge to the unique SRE [1, 1] in the ultralong run ($n \to \infty$). This is proved in the following section.

6. Asymptotic behavior

This section presents theoretical results on the asymptotic (i.e. ultralong run) behavior of the BM system. Note that with low learning rates the system may take an extraordinarily long time to reach its ultralong-run behavior (e.g., see bottom row in Fig. 2).

Proposition 2. *In any* 2×2 *game, assuming players' aspirations are different from their respective payoffs* $(u_i(\boldsymbol{d}) \neq A_i$ *for all* i *and* $\boldsymbol{d})$ *and below their respective maximin,*[1] *the BM process*

[1] Maximin is the largest possible payoff players can guarantee themselves in a single-stage game using pure strategies.

$\boldsymbol{P}_n$ converges to an SRE with probability 1 (*the set formed by all SREs is reached with probability* 1). *If the initial state is completely mixed, then every SRE can be reached with positive probability.*

Proposition 3. *In any* 2×2 *game, assuming players' aspirations are different from their respective payoffs and above their respective maximin*:

(i) *If there is any SRE then the BM process $\boldsymbol{P}_n$ converges to an SRE with probability* 1 (*the set formed by all SREs is reached with probability* 1). *If the initial state is completely mixed, then every SRE can be reached with positive probability.*
(ii) *If there is no SRE then the BM process $\boldsymbol{P}_n$ is ergodic*[2] *with no absorbing state.*

Corollary to Proposition 3. Consider any of the three 2×2 social dilemma games: Prisoner's Dilemma, Chicken, and Stag Hunt (Macy and Flache, 2002). Assuming players' aspirations are different from their respective payoffs and above their respective *maximin*:

(i) The BM process $\boldsymbol{P}_n$ is ergodic.
(ii) There is an SRE if and only if mutual cooperation is satisfactory for both players. In that case, the process converges to the unique SRE (i.e. certain mutual cooperation) with probability 1.

Since most BM systems end up converging to an SRE in the ultralong run, but their transient dynamics with slow learning are governed by their associated ODE, mathematical results that relate SREs with the solutions of the ODE can be particularly useful. The following proposition shows that the Nash equilibrium concept is key to determining the stability of SREs under the associated ODE.

Proposition 4. *Consider the BM process $\boldsymbol{P}_n$ and its associated ODE* (*Eq.* (2) *or* (3)) *in any* 2×2 *game*:

(i) *All SREs whose associated outcome is not a Nash equilibrium are unstable.*
(ii) *All SREs whose associated outcome is a strict Nash equilibrium where at least one unilateral deviation leads to a satisfactory outcome for the non-deviating player are asymptotically stable* (*i.e. they are SCEs too*).

Thus, our analysis adds to the growing body of work in learning game theory that supports the general principle that to assess the stability of *outcomes* in games, it is important to consider not only how unilateral deviations affect the deviator, but also how they affect the non-deviators. Outcomes where unilateral deviations hurt the deviator (strict Nash) but not the non-deviators (protected) tend to be the most stable. In the particular case of reinforcement learning with fixed aspirations, an additional necessary condition for the stability of an outcome is, of course, that every player finds the outcome satisfactory. Remark: Proposition 4 can be strengthened for the special case where all stimuli are positive (Sastry et al., 1994; Phansalkar et al., 1994).

[2] Following Norman (1968, p. 67), by 'ergodic' we mean that the sequence of stochastic kernels defined by the n-step transition probabilities of the Markov process associated with the BM system converges uniformly to a unique limiting kernel independent of the initial state. Intuitively, this means that the asymptotic probability distribution over the states of the system (i.e. the distribution of $\boldsymbol{P}_n$ when $n \to \infty$) is unique and does not depend on the initial state.

7. Extensions

The theoretical results on asymptotic behavior presented in this paper derive from the theory of distance diminishing models developed by Norman (1968; 1972), which can also be applied to 2-player games with any finite number of strategies without losing much generality. The results on transient behavior when learning takes place by small steps (which derive from the theory of stochastic approximation, Benveniste et al., 1990; Kushner and Yin, 1997, and the theory of slow learning, Norman, 1972) and Proposition 4 (which derives from Sastry et al., 1994) can be easily extended to any finite game.

More immediately, every proposition in this paper can be directly applied to finite populations from which two players are randomly[3] drawn repeatedly to play a 2×2 game. Indications on how to prove this are given in Appendix A. As an example, assume that there is a finite population of BM reinforcement learners with aspirations above their respective *maximin* and below their payoff for mutual cooperation, who meet randomly to play a 2×2 social dilemma game (Macy and Flache, 2002). Then, every player in the group will end up cooperating with probability 1 in the ultralong term. The more players in the group, the longer it takes the group to reach universal cooperation.

Another possible extension to the BM model consists in allowing players to suffer from 'trembling hands' (Selten, 1975): after deciding which action to undertake, each player i might select the wrong action with a small probability $\varepsilon_i > 0$ in each iteration. This feature generates a new stochastic process which is ergodic in any 2×2 game.[4] Proposition 1 applies to this extension too.

As for the general existence of SREs and SCEs in games with any finite number of players and strategies, note that both solution concepts require that the expected change in every player's strategy is zero—i.e. they are both critical points of the expected motion of the system. This is an important property since if any system converges to a state, that state must be a critical point of its expected motion. The following shows that every game has at least one such critical point for a very wide range of models. Consider the extensive set of models of normal-form games where every player's strategy is determined at any time-step by the probability of undertaking each of their possible actions. Assume that, after any given outcome $\boldsymbol{d}$ in time step n, every player i ($i = 1, 2, \ldots, m$) updates her strategy $\boldsymbol{p}_i$ using an adaptation rule $\boldsymbol{p}_i^{n+1} = \boldsymbol{r}_i^d(\boldsymbol{p}^n)$, where $\boldsymbol{r}_i^d(\boldsymbol{p}^n)$ is continuous for every $\boldsymbol{d}$ and every i. Let us call such adaptation rules continuous. Note that BM adaptation rules are continuous, and consider the following proposition.

Proposition 5. *Assuming that players' adaptation rules after every possible outcome of the game are continuous, every finite normal-form game has at least one critical point* (*a strategy profile where the expected change of every player's strategy is zero*).

8. Conclusions

This paper has focused on the study of games played by individuals who use one of the most widespread forms of learning in nature: reinforcement learning. This analysis (and related

[3] The important point here is that, at any time, every player must have a positive probability of being selected to play the game.

[4] This statement can be proved using Theorem 2.2 in Norman (1968, p. 67). We exclude here the meaningless case where the payoffs for some player are all the same and equal to her aspiration.

literature cited before) has shown that the outcome of games played by reinforcement learners can be substantially different from the expected outcomes when the game is played by perfectly rational individuals with common knowledge of rationality. As an example, cooperation in the repeated Prisoner's Dilemma is not only feasible but also the unique asymptotic outcome in many cases. More generally, outcomes where players select dominated strategies can emerge through social interaction and persist through time.

The present paper has characterized not only the asymptotic behavior of the Bush–Mosteller model of reinforcement learning, but also its transient dynamics. The study of the transient dynamics of learning algorithms has been neglected until recently due to the complexity of its formal analysis. Thus, most of the literature in learning game theory focuses on asymptotic equilibria. This may be insufficient since, as this paper has illustrated, the transient dynamics of learning algorithms may be substantially different from their asymptotic behavior. In broader terms, the importance of understanding the transient dynamics of formal models of social interactions is clear: social systems tend to exhibit an impressive ability to adapt and reorganize themselves structurally, meaning that most likely it is not asymptotic behavior that we observe in the real world.

Acknowledgments

This work has been generously supported by the Scottish Executive Environment and Rural Affairs Department. We would also like to thank Jörgen W. Weibull, Bruce Edmonds, Dale Rothman, and a very helpful anonymous reviewer for several useful comments.

Appendix A. Proofs

Notation. Since most of the proofs follow Norman (1968) we adopt his notation. The state of the system in iteration n, characterized in the BM model by the mixed-strategy profile in iteration n, is denoted S_n. The set of possible states is called the *state space* and denoted S. The realization of both players' decisions in iteration n is referred to as an event and denoted E_n. The set of possible events is called the *event space* and denoted E. S_n and E_n are to be considered random variables. In general, s and e denote elements of the state and event spaces, respectively. The function of S into S that maps S_n into S_{n+1} after the occurrence of event e is denoted $f_e(\cdot)$. Thus, if $E_n = e$ and $S_n = s$, then $S_{n+1} = f_e(s)$. Let $T_n(s)$ be the set of values that S_{n+1} takes on with positive probability when $S_1 = s$. Let us say that a state s is associated with an event e if s is a pure state (where all probabilities are either 0 or 1) and the occurrence of e pushes the system towards s from any other state. In any system, only one state is associated with a certain event, but the same state may be associated with several events. Finally, use $d(A, B)$ for the minimum Euclidean distance between two subsets A and B of S

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

Lemma 1. *Assuming players' aspiration levels are different from their respective payoffs, the 2-player 2-strategy BM model can be formulated as a strictly distance diminishing model* (*Norman, 1968, p. 64*).

Proof. Proving that the BM model can be formulated as a strictly distance diminishing model involves checking that hypotheses H1 to H8 in Norman (1968) hold. Define the state of the system S_n in iteration n in the BM model as the mixed-strategy profile in iteration n. The state space

is then the mixed-strategy space of the game, and the event space E is the space of pure-strategy profiles, or possible outcomes of the game; consider also the Euclidean distance in S. Having stated that, hypotheses H1 to H6 are almost immediate. H7 for strictly distance diminishing models reads

$$\text{H7.} \quad \sup_{s \neq s'} \frac{d(f_e(s), f_e(s'))}{d(s, s')} < 1 \quad \text{for all } e \in E.$$

Given that learning rates are strictly within 0 and 1 and stimuli are always non-zero numbers between -1 and 1 (since players' aspiration levels are different from their respective payoffs by assumption), it can easily be checked that H7 holds. The intuitive idea is that after any event e, the distance from any state s to the pure state s_e associated with event e is reduced by a fixed proportion in each of the components of s which is not already equal to the corresponding component in s_e. For the strict inequality in H7 to hold, it is instrumental that every state of the system (except at most one for each event) changes after any given event occurs (i.e. $f_e(s) \neq s$ for all $s \neq s_e$). The assumption that players' aspiration levels are different from their respective payoffs guarantees such a requirement. Without that assumption, H7 does not necessarily hold in its strict form. Finally, H8 reads:

H8. For any $s \in S$ there is a positive integer k and there are k events $e_1, \ldots, e_k$ such that

$$\sup_{s \neq s'} \frac{d(f_{e_1,\ldots,e_n}(s), f_{e_1,\ldots,e_n}(s'))}{d(s, s')} < 1 \quad \text{and} \quad P(E_j = e_j,\ 1 \leqslant j \leqslant n \mid S_1 = s) > 0,$$

where $f_{e_1,\ldots,e_n}(s) = f_{e_n}(f_{e_{n-1}}(\ldots(f_{e_1}(s))))$.

H8 is immediate having proved H7 in its strict form, since at least one event is possible in any state. □

Lemma 2. *Consider any 2-player 2-strategy BM system where players' aspiration levels differ from all their respective payoffs. Let s_e be the state associated with event e. If e may occur when the system is in state s* ($\Pr\{E_n = e \mid S_n = s\} > 0$)*, then*

$$\lim_{n \to \infty} d\big(T_n(s), s_e\big) = 0.$$

Proof. The BM model specifications guarantee that if event e may occur when the system is in state s, then it will also have a positive probability of happening in any subsequent state. Mathematically,

$$\Pr\{E_n = e \mid S_n = s\} > 0 \ \rightarrow\ \Pr\{E_{n+t} = e \mid S_n = s\} > 0 \quad \text{for any } t \geqslant 0.$$

This means that any finite sequence of events $\{e, e, \ldots, e\}$ has positive probability of happening. Note now that if the system is in state $s \neq s_e$ and event e occurs, the distance from s to s_e is reduced by a fixed proportion in each of the components of s which is not already equal to the corresponding component in s_e. This proportion of reduction is, for each player, the product of the player's absolute stimulus magnitude generated after e and the player's learning rate. Both proportions are strictly between 0 and 1 since players' aspiration levels are different from their respective payoffs by assumption. Let k be the minimum of those two proportions. Imagine then that event e keeps occurring, and note the following bound:

$$d\big(T_n(s), s_e\big) \leqslant (1 - k)^n \cdot d(s, s_e).$$

The proof is completed taking limits in the expression above

$$0 \leqslant \lim_{n\to\infty} d\big(T_n(s), s_e\big) \leqslant \lim_{n\to\infty} (1-k)^n \cdot d(s, s_e) = 0. \qquad \square$$

Proof of Proposition 1. Statement (i) is an application of Theorem 1 in Chap. 2 of Benveniste et al. (1990, p. 43). Statement (ii) follows from Norman (1972, Theorem 8.1.1, p. 118). The assumptions to apply this theorem are listed in Norman (1972, p. 117). Here we show that with the hypotheses in Proposition 1, all these assumptions hold. In this section, following Norman (1972), the state of the system in iteration n is denoted X_n, and the letter θ denotes the learning rate. Since the state space $I_\theta = I$ is independent of θ, (a.1) is satisfied. $H_n^\theta = \Delta X_n^\theta/\theta$ does not depend on θ, so (a.2) and (a.3) hold. All components of the functions $w(x) = \mathrm{E}(H_n^\theta \mid X_n^\theta = x)$ and $s(x) = \mathrm{E}((H_n^\theta - w(x))^2 \mid X_n^\theta = x)$ are polynomials, so every assumption (b) is satisfied. Finally, since H_n^θ does not depend on θ the supremum over θ can be omitted in (c), and also the module of each of the components of H_n^θ is bounded by the maximum learning rate, so (c) is also satisfied. Thus Theorem 8.1.1 is applicable. Finally, Statement (iii) is an application of Theorem 4.1 in Chap. 8 of Kushner and Yin (1997). $\square$

Proof of Proposition 2. Proposition 2 follows from Norman (1968, Theorem 2.3, p. 67), which requires the model to be distance-diminishing and one extra assumption H10.

H10. There are a finite number of absorbing states $a_1, \ldots, a_N$, such that, for any $s \in S$, there is some $a_{j(s)}$ for which

$$\lim_{n\to\infty} d\big(T_n(s), a_{j(s)}\big) = 0.$$

Given the assumptions of Proposition 2, Lemma 1 can be used to assert that the BM model is distance diminishing, with associated stochastic processes $\boldsymbol{S}_n$ and $\boldsymbol{E}_n$. Proving that H10 prevails will then complete the proof. The proof of H10 rests on the following three points:

(a) If in state s there is a positive probability of an event e occurring, then, applying Lemma 2:

$$\lim_{n\to\infty} d\big(T_n(s), s_e\big) = 0,$$

where s_e is the state associated with the event e.

(b) The state s_e associated with a mutually satisfactory (MS) event e is absorbing. Note also that there are at most four absorbing states.

(c) From any state there is a positive probability of playing a MS event within three iterations.

Points (a) and (b) are straightforward. To prove (c) we define strictly mixed strategies as those that assign positive probability to both actions, and mixed states as states where both players' strategies are strictly mixed. Note that after an unsatisfactory event, every player modifies her strategy so the updated strategy is strictly mixed, and that strictly mixed strategies will always remain so.

Since players' aspiration levels are below their respective *maximin* by assumption, there is at least one MS event. Hence from any mixed state there is a positive probability for a MS event to happen. We focus then on non-mixed states where no MS event can occur in the first iteration. This implies that the event in the first iteration is unsatisfactory for at least one player, so at least one player will have a strictly mixed strategy in the second iteration. Without loss of generality,

let us say that player 1 has a strictly mixed strategy in the second iteration. If player 2's strategy were also strictly mixed, then the state in the second iteration would be mixed, and a MS event could occur. Imagine then that the state in the second iteration is not mixed. Given that player 1's aspiration is below its *maximin*, there is a positive probability that the event in iteration 2 will be satisfactory for player 1. If such a possible event is also satisfactory for player 2, an MS event has occurred. If not, then both players will have a strictly mixed strategy in iteration 3, so a MS event could happen in iteration 3. This finishes the proof of point (c).

The proof of the fact that every SRE can be reached with positive probability if the initial state is completely mixed rests on two arguments: (a) there is a strictly positive probability that an infinite sequence of any given MS event e takes place (this can be proved using Theorem 52 in Hyslop (1965, p. 94)), and (b) such an infinite run would imply convergence to the associated (SRE) state s_e. We also provide here a theoretical result to estimate with arbitrary precision the probability L_∞ that an infinite sequence of a MS event $\boldsymbol{e} = (d_1, d_2)$ begins when the system is in mixed state $\boldsymbol{p} = (p_{1,d_1}, p_{2,d_2})$.

$$L_\infty = \lim_{n\to\infty} \prod_{t=0}^{n} \left[1 - (1 - p_{1,d_1})\left(1 - l_1 s_1(d_1)\right)^t\right]\left[1 - (1 - p_{2,d_2})\left(1 - l_2 s_2(d_2)\right)^t\right].$$

The following result can be used to estimate L_∞ with arbitrary precision.
Let $P_k = \prod_{t=0}^{k-1}(1 - xy^t)$ and let $P_\infty = \lim_{k\to\infty} P_k$. Then, for $x, y \in (0, 1)$,

$$P_k > P_\infty > P_k\left(1 - \frac{xy^k}{1-y}\right).$$

We are indebted to Professor Jörgen W. Weibull for discovering and providing the lower bound in this result. □

Proof of Proposition 3. Each statement of Proposition 3 will be proved separately.

Statement (i) is an immediate application of Theorem 2.3 in Norman (1968, p. 67), which requires the model to be distance-diminishing and the extra assumption H10 (see proof of Proposition 2). Having proved in Lemma 1 that the model is distance-diminishing, we prove here that H10 holds. The proof of H10 rests on the same three points (a)–(c) exposed in the proof of Proposition 2. The terminology defined there is also used here. Points (a) and (b) are straightforward. To prove (c), remember that after an unsatisfactory event, every player modifies her strategy so the updated strategy is strictly mixed, and that strictly mixed strategies always remain so. By assumption, there is at least one absorbing state, which means that there must be at least one MS event. This implies that from any mixed state there is a positive probability for a MS event to happen.

Since players' aspirations are above their respective *maximin*, given any action for player i, there is always an action for her opponent such that the resulting event would be unsatisfactory for player i. In other words, if one of the players has a strictly mixed strategy, then there is a positive chance that the system will be in a mixed state in the next iteration. We focus then on states where no player has strictly mixed strategies and a MS event cannot occur in the first iteration. This implies that the event in the first iteration is unsatisfactory for at least one player, who will have a strictly mixed strategy in the second iteration and, as just shown, this implies a positive probability that the system will be in a mixed state in the third iteration. The proof of statement (i) is then finished.

Statement (ii) follows from Theorem 2.2 in Norman (1968, p. 66), which requires the model to be distance-diminishing and one extra assumption H9.

H9. $\lim_{n\to\infty} d\big(T_n(s), T_n(s')\big) = 0 \quad \text{for all } s, s' \in S.$

Having proved in Lemma 1 that the model is distance-diminishing, we prove here that H9 holds. Since, by assumption, there are no absorbing states, there cannot be MS events. This implies that the event in the first iteration is unsatisfactory for at least one player, who will have a strictly mixed strategy in the second iteration. As argued in the proof of statement (i), this implies a positive probability that the system will be in a mixed state in the third iteration. Therefore at the third iteration any event has a positive probability of happening, so we can select any one of them, the state s_e associated with it, and then, by Lemma 2, we know that $\lim_{n\to\infty} d(T_n(s), s_e) = 0$ for any state s, so H9 holds. □

Proof of Proposition 4. The reasoning behind this proof follows Sastry et al. (1994). Statement (i) can be proved considering one player i who benefits by deviating from the SRE by increasing her probability $p_{i,q}$ to conduct action q. The expected change in probability $p_{i,q}$ can then be shown to be strictly positive for all $p_{i,q} > 0$ while keeping the other player's strategy unchanged. Statement (ii) can be proved considering the Jacobian of the linearization of ODE (2). Without loss of generality, assume that $D_i = \{A, B\}$ and the certain outcome at the SRE is $\boldsymbol{d}_{\text{SRE}} = (A, A)$. Choose $p_{1,B}$ and $p_{2,B}$ as the two independent components of the system, so the SRE is $[p_{1,B}, p_{2,B}] = [0, 0]$. The Jacobian J at the SRE is then as follows:

$$J = \begin{bmatrix} l_1(\delta(s_1(\mathrm{B},\mathrm{A})) - s_1(\mathrm{A},\mathrm{A})) & l_1 \cdot \delta(-s_1(\mathrm{A},\mathrm{B})) \\ l_2 \cdot \delta(-s_2(\mathrm{B},\mathrm{A})) & l_2(\delta(s_2(\mathrm{A},\mathrm{B})) - s_2(\mathrm{A},\mathrm{A})) \end{bmatrix},$$

$$\text{where } \delta(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } x \geqslant 0. \end{cases}$$

It is then straightforward that if $\boldsymbol{d}_{\text{SRE}} = (A, A)$ is a mutually satisfactory ($s_i(A, A) > 0$) strict Nash equilibrium ($s_1(A, A) > s_1(B, A)$; $s_2(A, A) > s_2(A, B)$) and at least one unilateral deviation leads to a satisfactory outcome for the non-deviating player ($s_1(A, B) \geqslant 0$ or $s_2(B, A) \geqslant 0$), then the two eigenvalues of J are negative real, so the SRE is asymptotically stable. □

Notes to extend the theoretical results to populations of players. All the lemmas and propositions in this paper can be easily extended to finite populations from which two players are randomly drawn to play a 2×2 game taking into account the following points: (1) the state of the system S_n in iteration n is the mixed-strategy profile of the whole population. (2) An event E_n in iteration n comprises an identification of the two players who have played the game in iteration n and their decisions. (3) Pure states are now associated (in the sense given in the notation of Appendix A) with *chains* of events, rather than with single events. A pure state s is associated with a finite chain of events c (where every player must play the game at least once) if the occurrence of c pushes the system towards s from any other state.

Proof of Proposition 5. Let Θ be the *mixed-strategy space* of the finite normal-form game. The proof consists in applying Brouwer's Fixed Point theorem to the function $\boldsymbol{W}(\boldsymbol{p}) \equiv E(\boldsymbol{P}_{n+1} \mid \boldsymbol{P}_n = \boldsymbol{p})$ that maps the mixed-strategy profile $\boldsymbol{p} \in \Theta$ to the *expected* mixed-strategy profile $\boldsymbol{W}(\boldsymbol{p})$ after the game has been played once and each player has updated her strategy $\boldsymbol{p}_i$ accordingly. Since the mixed-strategy space Θ is a non-empty, compact, and convex set, it only remains to

show that $W : \Theta \to \Theta$ is a continuous function. Let $\boldsymbol{w}_i(\boldsymbol{p})$ be the ith component of $\boldsymbol{W}(\boldsymbol{p})$, which represents player i's expected strategy for the following iteration. Therefore

$$\boldsymbol{w}_i(\boldsymbol{p}) = \sum_{\boldsymbol{d}\in D} \Pr\{\boldsymbol{d}\} \cdot \boldsymbol{r}_i^{\boldsymbol{d}}(\boldsymbol{p}) = \sum_{\boldsymbol{d}\in D} \left(\prod_{i\in I} p_{i,d_i} \right) \cdot \boldsymbol{r}_i^{\boldsymbol{d}}(\boldsymbol{p}).$$

Since all $\boldsymbol{r}_i^{\boldsymbol{d}}(\boldsymbol{p})$ are continuous for every $\boldsymbol{d}$ and every i by hypothesis, $\boldsymbol{W}(\boldsymbol{p})$ is also continuous. Thus, applying Brouwer's fixed-point theorem, we can state that there is at least one $\boldsymbol{p}^* \in \Theta$ such that $\boldsymbol{W}(\boldsymbol{p}^*) = \boldsymbol{p}^*$. This means that the *expected change* in all $(p_{i,j})^*$ (probability of player i following her jth pure strategy) is zero. □

References

Arthur, W.B., 1991. Designing economic agents that act like human agents: A behavioral approach to bounded rationality. Amer. Econ. Rev. 81 (2), 353–359.

Beggs, A., 2002. Stochastic evolution with slow learning. Econ. Theory 19, 379–405.

Beggs, A.W., 2005. On the convergence of reinforcement learning. J. Econ. Theory 122, 1–36.

Bendor, J., Mookherjee, D., Ray, D., 2001a. Aspiration-based reinforcement learning in repeated interaction games: An overview. Int. Game Theory Rev. 3 (2–3), 159–174.

Bendor, J., Mookherjee, D., Ray, D., 2001b. Reinforcement learning in repeated interaction games. Adv. Theor. Econ. 1 (1). Article 3.

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

Binmore, K., Samuelson, L., 1993. An economist's perspective on the evolution of norms. J. Institutional Theoret. Econ. 150, 45–63.

Binmore, K., Samuelson, L., Vaughan, R., 1995. Musical chairs: Modeling noisy evolution. Games Econ. Behav. 11 (1), 1–35.

Börgers, T., Sarin, R., 1997. Learning through reinforcement and replicator dynamics. J. Econ. Theory 77, 1–14.

Börgers, T., Sarin, R., 2000. Naive reinforcement learning with endogenous aspirations. Int. Econ. Rev. 41, 921–950.

Boylan, R.T., 1992. Laws of large numbers for dynamical systems with randomly matched individuals. J. Econ. Theory 57, 473–504.

Boylan, R.T., 1995. Continuous approximation of dynamical systems with randomly matched individuals. J. Econ. Theory 66, 615–625.

Bush, R., Mosteller, F., 1955. Stochastic Models of Learning. Wiley & Sons, New York.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Russell Sage Foundation, New York.

Chen, Y., Tang, F., 1998. Learning and incentive-compatible mechanisms for public goods provision: An experimental study. J. Polit. Economy 106, 633–662.

Cross, J.G., 1973. A stochastic learning model of economic behavior. Quart. J. Econ. 87, 239–266.

Duffy, J., 2006. Agent-based models and human subject experiments. In: Tesfatsion, L., Judd, K.L. (Eds.), Handbook of Computational Economics II: Agent-Based Computational Economics. Elsevier, North-Holland, pp. 949–1011 (Chapter 19).

Erev, I., Roth, A.E., 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. Amer. Econ. Rev. 88 (4), 848–881.

Erev, I., Roth, A.E., 2001. Simple reinforcement learning models and reciprocation in the Prisoner's Dilemma game. In: Gigerenzer, G., Selten, R. (Eds.), Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge, MA, pp. 216–231 (Chapter 12).

Erev, I., Bereby-Meyer, Y., Roth, A.E., 1999. The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. J. Econ. Behav. Organ. 39 (1), 111–128.

Flache, A., Macy, M.W., 2002. Stochastic collusion and the power law of learning. J. Conflict Resolution 46 (5), 629–653.

Hopkins, E., 2002. Two competing models of how people learn in games. Econometrica 70, 2141–2166.

Hopkins, E., Posch, M., 2005. Attainability of boundary points under reinforcement learning. Games Econ. Behav. 53 (1), 110–125.

Hyslop, J.M., 1965. Infinite Series, fifth ed. Oliver & Boyd, Edinburgh.

Ianni, A., 2001. Reinforcement learning and the power law of practice. Mimeo. University of Southampton.
Karandikar, R., Mookherjee, D., Ray, D., Vega-Redondo, F., 1998. Evolving aspirations and cooperation. J. Econ. Theory 80, 292–331.
Kushner, H.J., Yin, G.G., 1997. Stochastic Approximation Algorithms and Applications. Springer-Verlag, New York.
Laslier, J., Walliser, B., 2005. A reinforcement learning process in extensive form games. Int. J. Game Theory 33, 219–227.
Laslier, J., Topol, R., Walliser, B., 2001. A behavioral learning process in games. Games Econ. Behav. 37, 340–366.
Macy, M.W., Flache, A., 2002. Learning dynamics in social dilemmas. Proc. Natl. Acad. Sci. USA 99 (3), 7229–7236.
Maier, N.R.F., Schneirla, T.C., 1964. Principles of Animal Psychology. Dover Publications, New York.
McAllister, P.H., 1991. Adaptive approaches to stochastic programming. Ann. Oper. Res. 30, 45–62.
Mookherjee, D., Sopher, B., 1994. Learning behavior in an experimental matching pennies game. Games Econ. Behav. 7, 62–91.
Mookherjee, D., Sopher, B., 1997. Learning and decision costs in experimental constant sum games. Games Econ. Behav. 19, 97–132.
Norman, M.F., 1968. Some convergence theorems for stochastic learning models with distance diminishing operators. J. Math. Psychol. 5, 61–101.
Norman, M.F., 1972. Markov Processes and Learning Models. Academic Press, New York.
Palomino, F., Vega-Redondo, F., 1999. Convergence of aspirations and (partial) cooperation in the Prisoner's Dilemma. Int. J. Game Theory 28 (4), 465–488.
Phansalkar, V.V., Sastry, P.S., Thathachar, M.A.L., 1994. Absolutely expedient algorithms for learning Nash equilibria. Proc. Indian Acad. Sci. Math. Sci. 104 (1), 279–294.
Posch, M., 1997. Cycling in a stochastic learning algorithm for normal form games. J. Evolutionary Econ. 7, 193–207.
Roth, A.E., Erev, I., 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. Games Econ. Behav. 8, 164–212.
Rustichini, A., 1999. Optimal properties of stimulus-response learning models. Games Econ. Behav. 29, 244–273.
Sastry, P.S., Phansalkar, V.V., Thathachar, M.A.L., 1994. Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information. IEEE Trans. Syst. Man Cybernet. 24 (5), 769–777.
Selten, R., 1975. Re-examination of the perfectness concept for equilibrium points in extensive games. Int. J. Game Theory 4, 25–55.
Thorndike, E.L., 1898. Animal Intelligence: An Experimental Study of the Associative Processes in Animals. Psychological Review, Monograph Supplements, vol. 8. MacMillan, New York.
Weibull, J.W., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.
Wustmann, G., Rein, K., Wolf, R., Heisenberg, M., 1996. A new paradigm for operant conditioning of drosophila melanogaster. J. Comp. Physiol. (A) 179, 429–436.