11

Dynamics of the Bush-Mosteller Learning Algorithm in 2x2 Games

Luis R. Izquierdo[1] and Segismundo S. Izquierdo[2]
[1]*University of Burgos*
[2]*University of Valladolid*
Spain

1. Introduction

Reinforcement learners interact with their environment and use their experience to choose or avoid certain actions based on the observed consequences. Actions that led to satisfactory outcomes (i.e. outcomes that met or exceeded aspirations) in the past tend to be repeated in the future, whereas choices that led to unsatisfactory experiences are avoided. The empirical study of reinforcement learning dates back to Thorndike's animal experiments on instrumental learning at the end of the 19th century (Thorndike, 1898). The results of these experiments were formalised in the well known 'Law of Effect', which is nowadays one of the most robust properties of learning in the experimental psychology literature:

"*Of several responses made to the same situation those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections to the situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.*" (Thorndike, 1911, p. 244)

Nowadays there is little doubt that reinforcement learning is an important aspect of much learning in most animal species, including many phylogenetically very distant from vertebrates (e.g. earthworms (Maier & Schneirla, 1964) and fruit flies (Wustmann, 1996)). Thus, it is not surprising that reinforcement learning –being one of the most widespread adaptation mechanisms in nature– has attracted the attention of many scientists and engineers for decades. This interest has led to the formulation of various models of reinforcement learning and –when feasible– to the theoretical analysis of their dynamics. In particular, this chapter characterises the dynamics of one of the best known stochastic models of reinforcement learning (Bush & Mosteller, 1955) when applied to decision problems of strategy (i.e. games).

The following section is devoted to explaining in detail the context of application of our theoretical analysis, i.e. 2-player 2-strategy games. Section 3 is a brief review of various models of reinforcement learning that have been studied in strategic contexts. Section 4 presents the Bush-Mosteller reinforcement learning algorithm. Section 5 describes two types of critical points that are especially relevant for the dynamics of the process: self-reinforcing-equilibria (SREs) and self-correcting-equilibria (SCEs). Sections 6 and 7 detail the relevance

of these equilibria. Section 8 analyses the robustness of the model to "trembling-hands" noise and, finally, section 9 presents the conclusions of this chapter. The reader can replicate all the simulation runs reported in this chapter using an applet available at http://www.luis.izquierdo.name/papers/rl-book; we have also placed the source code used to create every figure in this chapter at the same web address.

2. Decision problems of strategy

At the heart of any learning algorithm we always find the problem of choice: learning is about making better decisions. At the most elementary level, decision problems can be classified according to the factors that may influence the outcome of the problem. Following that criterion we can distinguish, in ascending order of generality, the following categories (Colman, 1995):

1. Individual decision-making problems of *skill*. In this category there is no uncertainty involved: a single individual makes a decision, and the outcome of the problem depends solely on that decision (e.g. the problem of distributing a fixed production generated in various factories over several consumption centres, each with a given demand, in order to minimise transportation costs).
2. Individual decision-making problems under *risk*. In these problems, the solitary decision maker does not know with certainty the consequences of each of the possible options available to her, but she can meaningfully attach probabilities to each of the outcomes that may occur after each of her possible choices (e.g. the decision of buying a lottery ticket or not).
3. Individual decision-making problems under *uncertainty*. In this type of problem, as in the previous case, even though the consequences of a decision cannot be known with certainty at the time of making the decision, the range of possible consequences for each decision can be roughly identified in advance. However, unlike in decisions under risk, in decisions under uncertainty probabilities cannot be meaningfully attached to each of those consequences (e.g. deciding what to order in a new restaurant).
4. Decision problems of *strategy*. These problems involve many decision makers, and each of them has only partial control over which outcome out of a conceivable set of them will actually occur. Decision makers may have the ability to adapt to each other's decisions (e.g. setting prices in an oligopoly with the aim of maximising individual profit).
5. Decision problems under *ignorance*, or structural ignorance (Gilboa & Schmeidler, 1995 and 2001). This category is characterised by the fact that it is not possible to meaningfully anticipate the set of potential consequences that each of the possible choices may have (e.g. deciding whether to give the go-ahead to genetically modified crops).

Problems of skill have been extensively studied in several branches of mathematics. In decision-making under risk, compelling solutions have been derived using the theory of probability and expected utility theory. Expected utility theory, however, has not been so successful in the study of decision-making under uncertainty and strategic decision-making, which is the competence of game theory. Finally, understandably so, the formal study of decision problems under ignorance has not developed much.

In this chapter we formally study social interactions that can be meaningfully modelled as decision problems of strategy and, as such, using game theory as a framework. Game theory

is a branch of mathematics devoted to the formal analysis of decision making in social interactions where the outcome depends on the decisions made by potentially several individuals. A game is a mathematical abstraction of a social interaction where (Colman, 1995):

- there are two or more decision makers, called *players*;
- each player has a choice of two or more ways of acting, called actions or (*pure*) *strategies*, such that the outcome of the interaction depends on the strategy choices of all the players;
- the players have well-defined preferences among the possible outcomes (Hargreaves Heap & Varoufakis, 1995). Thus, *payoffs* reflecting these preferences can be assigned to all players for all outcomes. These payoffs are very often numerical (Fig. 1)

		Player 2	
		Player 2 chooses LEFT	Player 2 chooses RIGHT
Player 1	Player 1 chooses UP	3 , 3	0 , 4
	Player 1 chooses DOWN	4 , 0	1 , 1

Fig. 1. Normal form or payoff matrix of a 2-player, 2-strategy game.

A normal (or strategic form) game can be defined using a function that assigns a payoff to each player for every possible combination of actions. For games with only two players this function is commonly represented using a matrix (see Fig. 1). The example shown in Fig. 1 is a 2-player 2-strategy game: there are two players (player 1 and player 2), each of whom must select one out of two possible (pure) strategies. Player 1 can choose Up or Down, and player 2 simultaneously decides between Left or Right. The payoffs obtained by each player are represented in the corresponding cell of the matrix. Player 1 obtains the first payoff in the cell (coloured in red) and player 2 gets the second (coloured in blue). As an example, if player 1 selects Down and player 2 selects Left, then player 1 gets a payoff of 4 and player 2 obtains a payoff of 0. This chapter deals with 2×2 (2-player 2-strategy) games, which can be represented using a matrix like the one shown in Fig. 1.

Game theory is a useful framework to accurately and formally describe interdependent decision-making processes. Furthermore, it also provides a collection of solution concepts that narrow the set of expected outcomes in such processes. The most widespread solution concept in game theory is the Nash equilibrium, which is a set of strategies, one for each player, such that no player, knowing the strategy of the other(s), could improve her expected payoff by unilaterally changing her own strategy (e.g. the unique Nash equilibrium of the game represented in Fig. 1 is the combination of strategies Down-Right). The Nash equilibrium has been tremendously influential in the social sciences,

especially in economics, partly because it can be interpreted in a great number of meaningful and useful ways (Holt & Roth, 2004). Unfortunately, as a prediction tool, the concept is formally valid only when analysing games played by rational players with common knowledge of rationality[1] under the assumption of consistently aligned beliefs (Hargreaves Heap & Varoufakis, 1995). Such assumptions are clearly not appropriate in many social contexts, where it might not be clear at all that the outcome of the game should be a Nash equilibrium. In particular, if players are assumed to adapt their decisions using a reinforcement learning algorithm, it is often the case that the final outcome of their repeated interaction will not be a Nash equilibrium –as will be shown below.

3. Reinforcement learning in strategic contexts

In strategic contexts in general, empirical evidence suggests that reinforcement learning is most plausible in animals with imperfect reasoning abilities or in human subjects who have no information beyond the payoff they receive and may not even be aware of the strategic nature of the situation (Duffy, 2005; Camerer, 2003; Bendor et al., 2001a; Roth & Erev, 1995; Mookherjee & 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 behaviour in a wide range of games (McAllister, 1991; Roth & Erev, 1995; Mookherjee & Sopher, 1994; Mookherjee & Sopher, 1997; Chen & Tang, 1998; Erev & Roth, 1998; Erev et al., 1999). In general, the various models of reinforcement learning that have been applied to strategic contexts 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 & Roth (1998) vs. Börgers & Sarin (1997)).
- Whether the model allows for avoidance behaviour in addition to approach behaviour (e.g. Bendor et al. (2001b) vs. Erev & Roth (1998)). Approach behaviour is the tendency to repeat the associated choices after receiving a positive stimulus; avoidance behaviour 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; Bendor et al., 2001b).
- Whether "forgetting" is considered, i.e. whether recent observations weigh more than distant ones (Erev & Roth, 1998; Rustichini, 1999; Beggs, 2005).
- Whether the model imposes inertia – a positive bias in favour of the most recently selected action (Bendor et al., 2001a; Bendor et al., 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 behaviour 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

[1] Common knowledge of rationality means that every player assumes that all players are instrumentally rational, and that all players are aware of other players' rationality-related assumptions (this produces an infinite recursion of shared assumptions).

the theory of stochastic approximation (Benveniste et al., 1990; Kushner & Yin, 1997) and from the theory of urn models can often be applied (e.g. Ianni, 2001; Hopkins & Posch, 2005; Beggs, 2005), whereas if the learning rate is constant, results from the theory of distance diminishing models (Norman, 1968; Norman, 1972) tend to be more useful (e.g. Börgers & Sarin, 1997; Bendor et al., 2001b; Izquierdo et al., 2007). 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).

Two of the simplest and most popular models of reinforcement learning in the game theory literature are the Erev-Roth (ER) model (Roth & Erev, 1995; Erev & Roth, 1998) and the Bush-Mosteller (BM) model (Bush & Mosteller, 1955). Both models are stochastic: players' strategies are probabilities or propensities to take each of their possible actions. In the ER model, playing one action always increases the probability of playing that action again (i.e. only positive stimulus are considered), and the sensitivity of players' strategies to a new outcome decreases as the game advances (Power Law of Practice). On the other hand, the BM model is an aspiration-based reinforcement learning model where negative stimuli are possible 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 analysed by Börgers & Sarin (1997). In this chapter we characterise the dynamics of the BM model in 2×2 games where aspiration levels are fixed, but not necessarily below the lowest payoff (i.e. negative stimuli are possible). The dynamics of this model were initially explored by Macy & Flache (2002) and Flache & Macy (2002) in 2×2 social dilemma games using computer simulation, and their work was formalised and extended for general 2×2 games by Izquierdo et al. (2007). This chapter follows closely the work conducted by Izquierdo et al. (in press), who analysed the BM model using a combination of computer simulation experiments and theoretical results. Most of the theoretical results used in this chapter derive from Izquierdo et al. (2007).

4. The BM reinforcement learning algorithm

The model we analyse here is an elaboration of a conventional Bush-Mosteller (1955) stochastic learning model for binary choice. In this model, players decide what action to select stochastically: each player's strategy is defined by the probability of undertaking each of the two actions available to them. After every player has selected an action according to their probabilities, every player receives the corresponding payoff and revises her strategy. The revision of strategies takes place following a reinforcement learning approach: players increase their probability of undertaking a certain action if it led to payoffs above their aspiration level, and decrease this probability otherwise. When learning, players in the BM model use only information concerning their own past choices and payoffs, and ignore all the information regarding the payoffs and choices of their counterparts.

More precisely, let $I = \{1, 2\}$ be the *set of players* in the game, and let Y_i be the *pure-strategy space* for each player $i \in I$. For convenience, and without loss of generality, later we will call the actions available to each of the players C (for Cooperate) and D (for Defect). Thus $Y_i = \{C, D\}$. Let u_i be the *payoff functions* u_i that give player i's payoff for each profile $\boldsymbol{y} = (y_1, y_2)$ of pure strategies, where $y_i \in Y_i$ is a pure strategy for player i. As

an example, u_i(C, D) denotes the payoff obtained by player i when player 1 cooperates and player 2 defects. Let $Y = \times_{i \in I} Y_i$ be the space of pure-strategy profiles, or possible outcomes of the game. Finally, let p_{i,y_i} denote player i's probability of undertaking action y_i.

In the BM model, strategy updating takes place in two steps. First, after outcome $\boldsymbol{y}^n = (y_1^n, y_2^n)$ in time-step n, each player i calculates her stimulus $s_i(\boldsymbol{y}^n)$ for the action just chosen y_i^n according to the following formula:

$$s_i(\boldsymbol{y}) = \frac{u_i(\boldsymbol{y}) - A_i}{\sup_{\boldsymbol{k} \in Y} | u_i(\boldsymbol{k}) - A_i |}$$

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 $\sup_{\boldsymbol{k} \in Y} | u_i(\boldsymbol{k}) - A_i |$. Secondly, having calculated their stimulus $s_i(\boldsymbol{y}^n)$ after the outcome $\boldsymbol{y}^n$, each player i updates her probability p_{i,y_i} of undertaking the selected action y_i as follows:

$$p_{i,y_i}^{n+1} = \begin{cases} p_{i,y_i}^n + l_i \cdot s_i(\boldsymbol{y}^n) \cdot (1 - p_{i,y_i}^n) & \text{if } s_i(\boldsymbol{y}^n) \geq 0 \\ p_{i,y_i}^n + l_i \cdot s_i(\boldsymbol{y}^n) \cdot p_{i,y_i}^n & \text{if } s_i(\boldsymbol{y}^n) < 0 \end{cases}$$

where p_{i,y_i}^n is player i's probability of undertaking action y_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. Note that the state of the game can be fully characterized by a two-dimensional vector $\boldsymbol{p}$ = [p_1 , p_2], where p_i is player i's probability to cooperate (i.e. $p_i = p_{i,C}$). We will refer to such vector $\boldsymbol{p}$ as a *strategy profile*, or a *state of the system*.

In the general case, a 2×2 BM model parameterisation requires specifying both players' payoff function u_i, aspiration level A_i, and learning rate l_i. Our analysis is based on the theoretical results derived by Izquierdo et al. (2007), which are valid for any 2×2 game, but – for illustrative purposes– we focus here on systems where two players parameterised in exactly the same way ($A_i = A$ and $l_i = l$) play a symmetric Prisoner's Dilemma game. The Prisoner's Dilemma is a two-person game where each player can either cooperate or defect. For each player i, the payoff when they both cooperate (u_i(C, C) = R_i, for *Reward*) is greater than the payoff obtained when they both defect (u_i(D, D) = P_i, for *Punishment*); when one cooperates and the other defects, the cooperator obtains S_i (*Sucker*), whereas the defector receives T_i (*Temptation*). The dilemma comes from the fact that, individually, each player is better off defecting given any of her counterpart's choices ($T_i > R_i$ and $P_i > S_i$; i = 1, 2), but they both obtain a greater payoff when they both cooperate than when they both defect ($R_i > P_i$; i = 1, 2). Symmetry implies that $T_i = T$, $R_i = R$, $P_i = P$ and $S_i = S$. Figure 1 shows an example of a symmetric Prisoner's Dilemma. A certain parameterisation of this type of system will be specified using the template [T , R , P , S | A | l][2].

The following notation will be useful: A parameterised model will be denoted **S**, for System. Let $\boldsymbol{P}_n(\mathbf{S})$ be the state of a system **S** in time-step n. Note that $\boldsymbol{P}_n(\mathbf{S})$ is a random variable and a strategy profile $\boldsymbol{p}$ is a particular value of that variable. The sequence of random variables

$\{\boldsymbol{P}_n(\mathbf{S})\}_{n\geq 0}$ constitutes a discrete-time Markov process with potentially infinite transient states.

5. Attractors in the dynamics of the system

Macy & Flache (2002) observed and described two types of attractors that govern the dynamics of the BM model: self-reinforcing equilibria (SRE), and self-correcting equilibria (SCE). These two concepts are not equilibria in the static sense of the word, but strategy profiles which act as attractors that pull the dynamics of the simulation towards them. The original concepts of SRE and SCE were later formalised and refined by Izquierdo et al. (2007).

SREs are absorbing states of the system (i.e. states $\boldsymbol{p}$ that cannot be abandoned) where both players receive a positive stimulus (Izquierdo et al., 2007). An SRE corresponds to a pair of pure strategies (p_i is either 0 or 1) such that its certain associated outcome gives a strictly positive stimulus to both players (henceforth a *mutually satisfactory outcome*). For example, the strategy profile [1 , 1] is an SRE if both players' aspiration levels are below their respective $R_i = u_i$(C, C). Escape from an SRE is impossible since no player will change her strategy. More importantly, SREs act as attractors: near an SRE, there is a high chance that the system will move towards it, because 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.

The definition of the other type of attractor, namely the SCE, is related to the expected motion function of the system. The Expected Motion (**EM**) of a system **S** in state $\boldsymbol{p}$ for the following iteration is given by a function vector $\mathbf{EM}^{\mathbf{S}}(\boldsymbol{p})$ whose components are the expected change in the probabilities to cooperate for each player. Mathematically,

$$\mathbf{EM}^{\mathbf{S}}(\boldsymbol{p}) \equiv [\, \mathrm{EM}_1^{\mathbf{S}}(\boldsymbol{p}) \,,\, \mathrm{EM}_2^{\mathbf{S}}(\boldsymbol{p}) \,] \equiv \mathbf{E}(\Delta \boldsymbol{P}_n(\mathbf{S}) \mid \boldsymbol{P}_n(\mathbf{S}) = \boldsymbol{p})$$

$$\mathrm{EM}_i^{\mathbf{S}}(\boldsymbol{p}) = \Pr\{\mathrm{CC}\} \cdot \Delta p_i\big|_{\mathrm{CC}} + \Pr\{\mathrm{CD}\} \cdot \Delta p_i\big|_{\mathrm{CD}} + \Pr\{\mathrm{DC}\} \cdot \Delta p_i\big|_{\mathrm{DC}} + \Pr\{\mathrm{DD}\} \cdot \Delta p_i\big|_{\mathrm{DD}}$$

where {CC, CD, DC, DD} represent the four possible outcomes that may occur.

For instance, for a Prisoner's Dilemma parameterised as [4 , 3 , 1 , 0 | 2 | l][2], the function $\mathbf{EM}(\boldsymbol{p})$ is

$$\left[\mathbf{EM}_1(\boldsymbol{p}), \mathbf{EM}_2(\boldsymbol{p})\right] = l \begin{bmatrix} p_1 p_2 & p_1(1-p_2) & (1-p_1)p_2 & (1-p_1)(1-p_2) \end{bmatrix} \begin{bmatrix} (1-p_1)/2 & (1-p_2)/2 \\ -p_1 & -p_2 \\ -p_1 & -p_2 \\ (1-p_1)/2 & (1-p_2)/2 \end{bmatrix}$$

This Expected Motion function is represented by the arrows shown in figure 2.

Consider now differential equation (1), which is the continuous time limit approximation of the system's expected motion:

$$\dot{f} = \mathbf{EM}^{\mathbf{S}}(f) \tag{1}$$

or, equivalently,

$$\left.\begin{aligned} \frac{df_1(t)}{dt} &= \mathrm{EM}_1^{\mathbf{S}}(f(t)) \\ \frac{df_2(t)}{dt} &= \mathrm{EM}_2^{\mathbf{S}}(f(t)) \end{aligned}\right\}$$

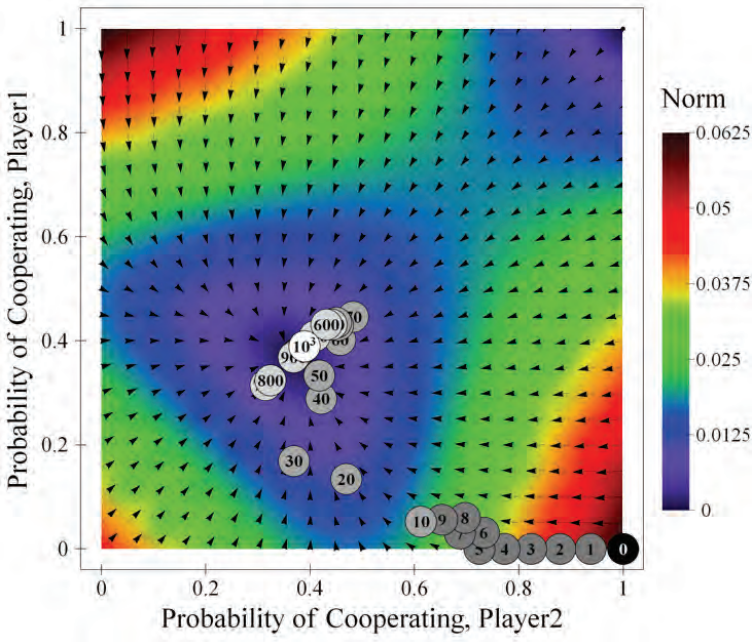

Fig. 2. Expected motion of the system in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 2^{-4}]², together with a sample simulation run (1000 iterations). The arrows represent the expected motion for various states of the system; the numbered balls show the state of the system after the indicated number of iterations in the sample run. The background is coloured using the norm of the expected motion. For any other learning rate the size of the arrows (i.e. the norm of the expected motion) would vary but their direction would be preserved.

Thus, for the Prisoner's Dilemma parameterised as [4 , 3 , 1 , 0 | 2 | l]², the associated differential equation is

$$\left[\frac{df_1}{dt}, \frac{df_2}{dt}\right] = l\left[f_1 f_2 \quad f_1(1-f_2) \quad (1-f_1)f_2 \quad (1-f_1)(1-f_2)\right]\begin{bmatrix} (1-f_1)/2 & (1-f_2)/2 \\ -f_1 & -f_2 \\ -f_1 & -f_2 \\ (1-f_1)/2 & (1-f_2)/2 \end{bmatrix}$$

Some trajectories of this differential equation are shown in figure 3. The expected motion at any point $\boldsymbol{p}$ in the phase plane is a vector tangent to the unique trajectory to which that point belongs. Having explained the expected motion of the system and its continuous time limit approximation we can now formally define SCEs.

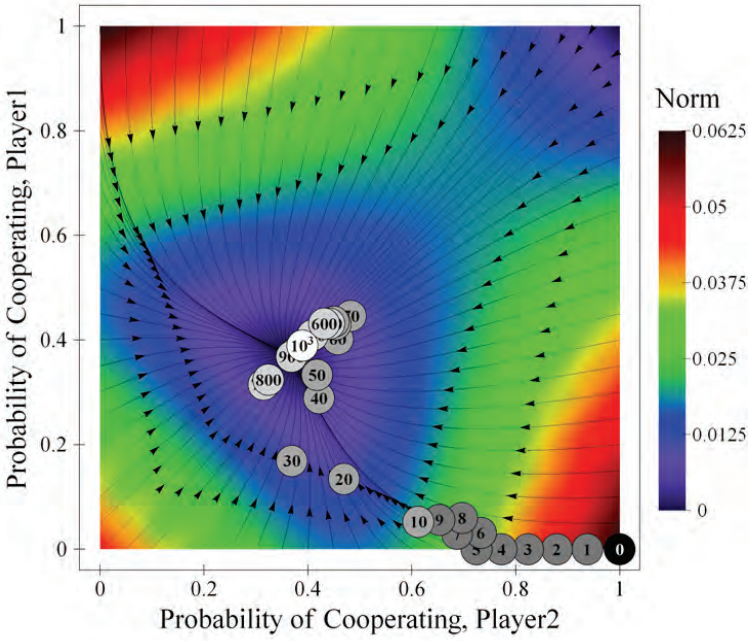

Fig. 3. Trajectories in the phase plane of the differential equation corresponding to the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]^2, together with a sample simulation run ($l = 2^{-4}$). The background is coloured using the norm of the expected motion. This system has an SCE at [0.37 , 0.37].

An SCE of a system **S** is an asymptotically stable critical point (Mohler, 1991) of differential equation (1) (Izquierdo et al., 2007). Roughly speaking this means that all trajectories in the phase plane of Eq. (1) that at some instant are sufficiently close to the SCE will approach the SCE as the parameter t (time) approaches infinity and remain close to it at all future times. Note that, with these definitions, there could be a state of the system that is an SRE and an SCE at the same time. Note also that $\mathbf{EM}^{\mathbf{S}}(\boldsymbol{SCE}) = \mathbf{0}$ and $\mathbf{EM}^{\mathbf{S}}(\boldsymbol{SRE}) = \mathbf{0}$. In particular, the Prisoner's Dilemma represented in figure 3 exhibits a unique SCE at [0.37 , 0.37] and a unique SRE at [1 , 1].

Let $f_x(t)$ denote the solution of differential equation (1) for some initial state $\boldsymbol{x}$. Figure 4 shows $f_x(t)$ for the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]^2 for different (and symmetric) initial conditions $\boldsymbol{x} = [x_0 , x_0]$. For this particular case and settings, the two components of $f_x(t) = [f_{1,x}(t) , f_{2,x}(t)]$ take the same value at any given t, so the representation in figure 4 corresponds to both components of $f_x(t)$. Convergence to the SCE

at [0.37 , 0.37] can be clearly observed for every initial condition $\boldsymbol{x}$, except for $\boldsymbol{x}$ = [1, 1], which is the SRE.

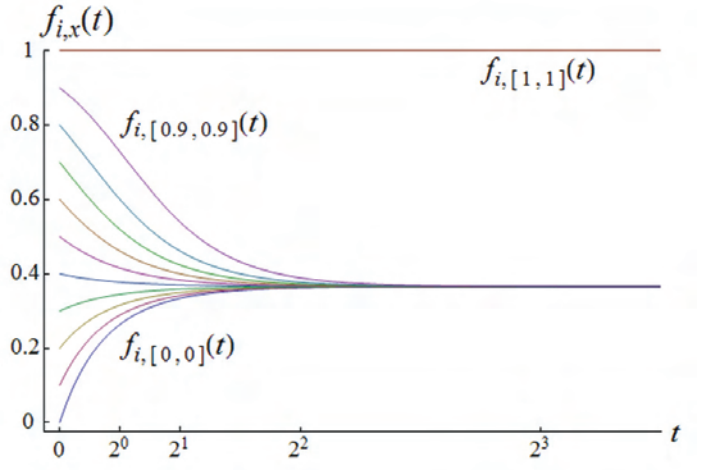

Fig. 4. Solutions of differential equation (1) for the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l][2] with different (and symmetric) initial conditions $\boldsymbol{x}$ = [x_0 , x_0]. This system has a unique SCE at [0.37 , 0.37] and a unique SRE at [1 , 1].

The use of expected motion (or mean-field) approximations to understand simulation models and to design interesting experiments has already proven to be very useful in the literature (e.g. Huet et al., 2007; Galán & Izquierdo, 2005; Edwards et al., 2003; Castellano et al., 2000). Note, however, that such approaches are approximations whose validity may be constrained to specific conditions: as we can see in Figure 3, simulation runs and trajectories will not coincide in general. Later in this chapter we show that trajectories and SCEs are especially relevant for the transient dynamics of the system, particularly with small learning rates, but, on the other hand, the mean-field approximation can be misleading when studying the asymptotic behaviour of the model.

6. Attractiveness of SREs

Macy and Flache's experiments (Macy & Flache, 2002; Flache & Macy, 2002) with the BM model showed a puzzling phenomenon. A significant part of their analysis consisted in studying, in a Prisoner's Dilemma in which mutual cooperation was mutually satisfactory (i.e. $A_i < R_i = u_i$(C, C)), the proportion of simulation runs that "locked" into mutual cooperation. Such "lock-in rates" were reported to be as high as 1 in some experiments. However, starting from an initial state which is not an SRE, the BM model specifications guarantee that after any finite number of iterations any outcome has a positive probability of occurring (i.e. strictly speaking, lock-in is impossible)[2]. To investigate this apparent

[2] The specification of the model is such that probabilities cannot reach the extreme values of 0 or 1 starting from any other intermediate value. Therefore if we find a simulation run that

contradiction we conducted some qualitative analyses that we present here to familiarise the reader with the complex dynamics of this model. Our first qualitative analysis consisted in studying the expected dynamics of the model. Figure 5 illustrates the expected motion of a system extensively studied by Macy & Flache: the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.5]². As we saw before, this system features a unique SCE at [0.37 , 0.37] and a unique SRE at [1 , 1]. Figure 5 also includes the trace of a sample simulation run. Note that the only difference between the parameterisation of the system shown in figure 2 and that shown in figure 5 is the value of the learning rate.

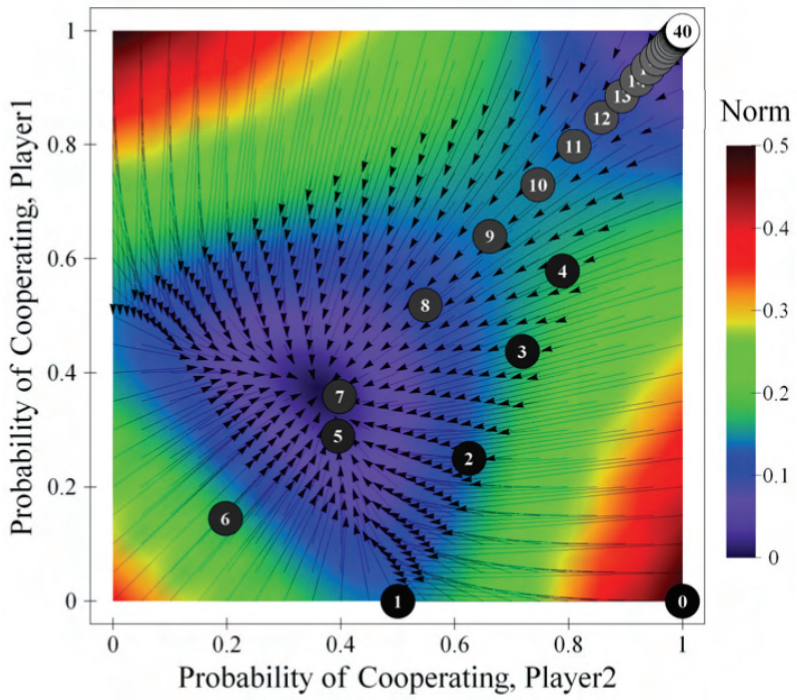

Fig. 5. Expected motion of the system in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.5]², with a sample simulation run.

Figure 5 shows that the expected movement from any state is towards the SCE, except for the only SRE, which is an absorbing state. In particular, near the SRE, where both probabilities are high but different from 1, the distribution of possible movements is very peculiar: there is a very high chance that both agents will cooperate and consequently move

---

has actually ended up in an SRE starting from any other state, we know for sure that such simulation run did not follow the specifications of the model (e.g. perhaps because of floating-point errors). For a detailed analysis of the effects of floating point errors in computer simulations, with applications to this model in particular, see Izquierdo and Polhill (2006), Polhill and Izquierdo (2005), Polhill et al. (2006), Polhill et al. (2005).

a small distance towards the SRE, but there is also a positive chance, tiny as it may be, that one of the agents will defect, causing both agents to jump away from the SRE towards the SCE. The improbable, yet possible, leap away from the SRE is of such magnitude that the resulting expected movement is biased towards the SCE despite the unlikelihood of such an event actually occurring. The dynamics of the system can be further explored analysing the most likely movement from any given state, which is represented in Figure 6.

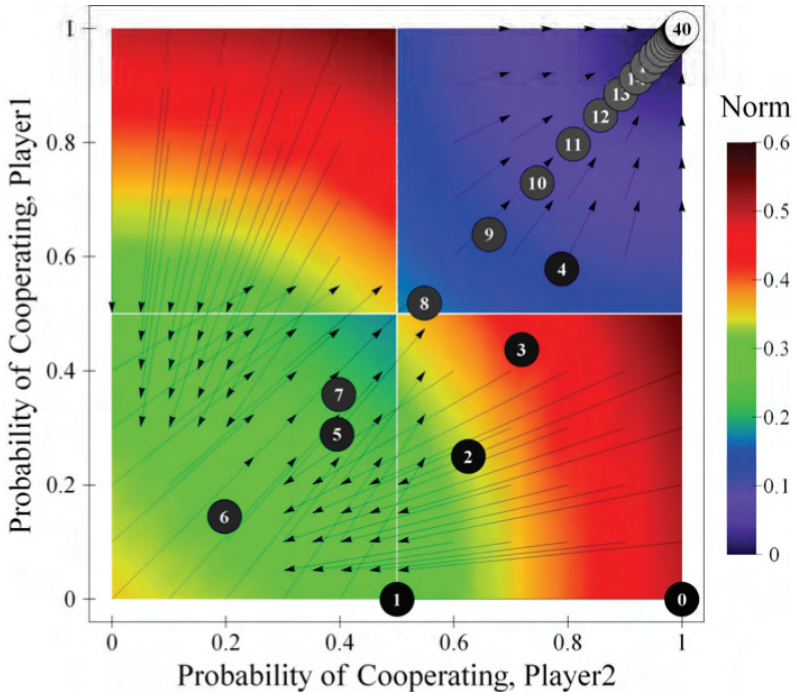

Fig. 6. Figure showing the most likely movements at some states of the system in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.5][2], with a sample simulation run. The background is coloured using the norm of the most likely movement.

Figure 6 differs significantly from Figure 5; it shows that the most likely movement in the upper-right quadrant of the state space is towards the SRE. Thus, the walk towards the SRE is characterised by a fascinating puzzle: on the one hand, the most likely movement leads the system towards the SRE, which is even more likely to be approached the closer we get to it; on the other hand, the SRE cannot be reached in any finite number of steps and the expected movement as defined above is to walk away from it (see figure 5).

It is also interesting to note in this game that, starting from any mixed (interior) state, both players have a positive probability of selecting action D in any future time-step, but there is also a positive probability that both players will engage in an infinite chain of the mutually satisfactory event CC forever, i.e., that neither player will ever take action D from then onwards (see Izquierdo et al., in press).

The probability of starting an infinite chain of CC events depends largely on the value of the learning rate l. Figure 7 shows the probability of starting an infinite chain of the mutually satisfactory outcome CC in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]², for different learning rates l, and different initial probabilities to cooperate x_0 (the same probability for both players). For some values, the probability of immediately starting an infinite chain of mutual cooperation can be surprisingly high (e.g. for l = 0.5 and initial conditions [x_0 , x_0] = [0.9 , 0.9] such probability is approximately 44%).

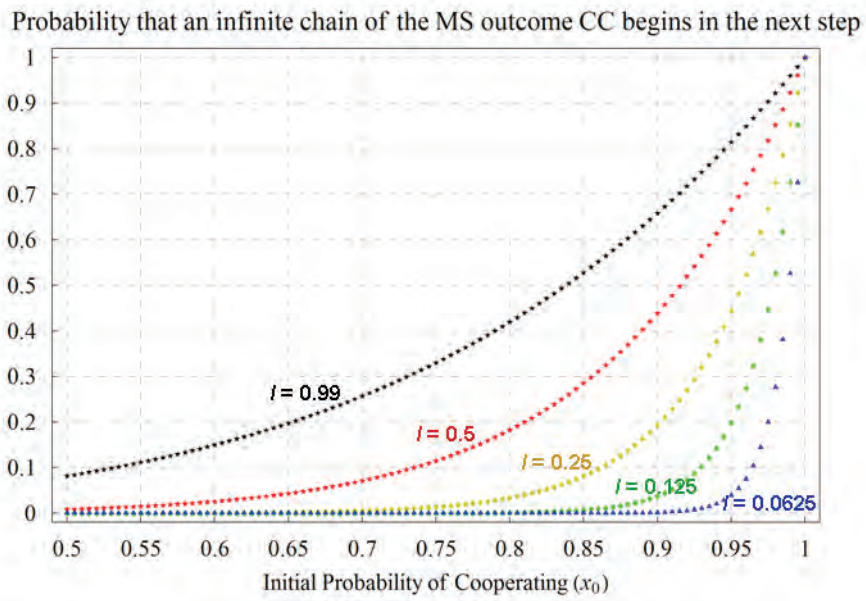

Fig. 7. Probability of starting an infinite chain of the Mutually Satisfactory (MS) outcome CC in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]². The 5 different (coloured) series correspond to different learning rates l. The variable x_0, represented in the horizontal axis, is the initial probability of cooperating for both players.

In summary, assuming that aspirations are different from payoffs (see Izquierdo et al., 2007), a BM process that starts in an initial state different from an SRE will never reach an SRE in finite time, and there is always a positive probability that the process leaves the proximity of an SRE. However, if there is some SRE, there is also a positive probability that the system will approach it indefinitely (i.e. forever) through an infinite chain of the mutually satisfactory outcome associated to the SRE.

7. Different regimes in the dynamics of the system

This section illustrates the dynamics of the BM model for different learning rates. Most of the theoretical results that we apply and summarise in this section are valid for any 2×2 game and can be found in Izquierdo et al. (2007). The analysis is presented here in a

somewhat qualitative fashion for the sake of clarity and comprehensibility, and illustrates the behaviour of the BM model using the Prisoner's Dilemma shown in figure 1.

In the general case, the dynamics of the BM model may exhibit three different regimes: medium run, long run, and ultralong run. The terminology used here is borrowed from Binmore & Samuelson (1993) and Binmore et al. (1995), who reserve the term short run for the initial conditions.

"*By the ultralong run, we mean a period of time long enough for the asymptotic distribution to be a good description of the behavior of the system. The long run refers to the time span needed for the system to reach the vicinity of the first equilibrium in whose neighborhood it will linger for some time. We speak of the medium run as the time intermediate between the short run* [i.e. initial conditions] *and the long run, during which the adjustment to equilibrium is occurring.*" (Binmore et al., 1995, p. 10)

Binmore et al.'s terminology is particularly useful for our analysis because it is often the case in the BM model that the "*first equilibrium in whose neighborhood it* [the system] *will linger for some time*", i.e. the long run, is significantly different from the asymptotic dynamics of the system. Whether the three different regimes (medium, long, and ultralong run) are clearly distinguishable in the BM model strongly depends on the players' learning rates. For high learning rates the system quickly approaches its asymptotic behaviour (the ultralong run) and the distinction between the different regimes is not particularly useful. For small learning rates, however, the three different regimes can be clearly observed. Since the ultralong run is the only regime that is (finally) observed in every system, we start our description of the dynamics of the BM model characterising such regime (for details see Propositions 2 and 3 in Izquierdo et al., 2007). Assuming players' aspirations are different from their respective payoffs ($u_i(\boldsymbol{y}) \neq A_i$ for all i and $\boldsymbol{y}$):

- If players' aspirations are below their respective *maximin*[3], the BM system converges to an SRE with probability 1 (i.e. the set formed by all SREs is asymptotically reached with probability 1). If the initial state is completely mixed, then every SRE can be asymptotically reached with positive probability.
- If players' aspirations are above their respective *maximin*:
  - if there is any SRE then the BM system converges to an SRE with probability 1. If the initial state is completely mixed, then every SRE can be asymptotically reached with positive probability.
  - If there are no SREs then the process is ergodic, so the states of the system present an asymptotic distribution which is independent of the initial conditions.

In the context of the Prisoner's dilemma game described above, this implies that if players' aspirations are above the payoff they receive when they both defect ($A_i > u_i(\mathrm{D}, \mathrm{D}) = P_i$), which is their *maximin*, then the ultralong run is independent of the initial state. Under such conditions, there is an SRE if and only if mutual cooperation is satisfactory for both players (i.e. $A_i < u_i(\mathrm{C}, \mathrm{C}) = R_i$) and, if that is the case, the process converges to certain mutual cooperation (i.e. the unique SRE) with probability 1. As an example, note that the ultralong-run behaviour of the systems shown in figures 2, 3, 5 and 6 is certain mutual cooperation.

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

7.1 Learning by large steps (fast adaptation)

As mentioned above, when learning takes place by large steps, the system quickly reaches its ultralong-run behaviour. To explain why this is the case we distinguish between two possible classes of systems:

- In systems where there is at least one SRE, the asymptotic behaviour is quickly approached because SREs are powerful attractors (e.g. see figures 5 and 6). The reason for this is that, if an SRE exists, the chances of a mutually satisfactory outcome not occurring for a long time are low, since players update their strategies to a large extent to avoid unsatisfactory outcomes. Whenever a mutually satisfactory outcome occurs, players update their strategy so the chances of repeating such a mutually satisfactory outcome increase. Since learning rates are high, the movement towards the SRE associated with such a mutually satisfactory outcome takes place by large steps, so only a few coordinated moves are sufficient to approach the SRE so much that escape from its neighbourhood becomes very unlikely. In other words, with fast learning the system quickly approaches an SRE, and is likely to keep approaching that SRE forever (this is the system's ultralong-run behaviour). As an example, consider figure 7 again: starting from any initial probability to cooperate x_0, the occurrence of a mutually satisfactory outcome CC would increase both players' probability to cooperate (the updated probability can be seen as the following period's x_0), which in turn would increase the probability of never defecting (i.e. the probability of starting an infinite chain of CC). Thus, if learning rates are large, a few CC events are enough to take the state of the system into areas where the probability of never defecting again is large.
- In the absence of SREs, the fact that any outcome is unsatisfactory for at least one of the players[4] and the fact that strategy changes are substantial, together imply that at least one player will switch between actions very frequently –i.e. the system will indefinitely move rapidly and widely around a large area of the state space.

7.2 Learning by small steps (slow adaptation)

The behaviour of the BM process with low learning rates is characterised by the following features (Izquierdo et al., 2007; Proposition 1):

- For low enough learning rates, the BM process with initial state $\boldsymbol{x}$ tends to follow the trajectory $\boldsymbol{f}_x(t)$ in the phase plane of Eq. (1), i.e. the trajectory that corresponds to $\boldsymbol{f}(0) = \boldsymbol{x}$ (e.g. see figure 3).
- For low enough learning rates l, the BM process in time-step n tends to be concentrated around a particular point of the mentioned trajectory: the point $\boldsymbol{f}_x(n \cdot l)$ (e.g. see figure 4).
- If trajectories get close to an SCE (as t increases), then, for low learning rates, the BM process will tend to approach and linger around the SCE; the lower the learning rate, the greater the number of periods that the process will tend to stay around the SCE.
- Eventually the system will approach its asymptotic behaviour, which –as explained above– is best characterised by the SREs of the system.

When learning takes place by small steps the transient regimes (i.e. the medium and the long run) can be clearly observed, and these transient dynamics can be substantially different from the ultralong-run behaviour of the system. For sufficiently small learning

[4] Recall that each player's aspiration level is assumed to be different from every payoff the player may receive.

rates and number of iterations n not too large ($n \cdot l$ bounded), the medium-run dynamics of the system are best characterised by the trajectories in the phase plane of Eq. (1), which can follow paths substantially apart from the end-states of the system (see figure 8, where the end-state is [1 , 1]). Under such conditions, the expected state of the system after n iterations can be estimated by substituting the value $n \cdot l$ in the trajectory that commences at the initial conditions (see figure 4). The lower the learning rates, the better the estimate, i.e. the more tightly clustered the dynamics will be around the corresponding trajectory in the phase plane (see figure 8).

When trajectories finish in an SCE, the system will approach the SCE and spend a significant amount of time in its neighbourhood if learning rates are low enough and the number of iterations n is large enough (and finite)[5]. This latter regime is the long run. The fact that trajectories are good approximations for the transient dynamics of the system for slow learning shows the importance of SCEs –points that "attract" trajectories within their neighbourhood– as attractors of the actual dynamics of the system. This is particularly so when, as in most 2×2 games, there are very few asymptotically stable critical points and they have very wide domains of attraction.

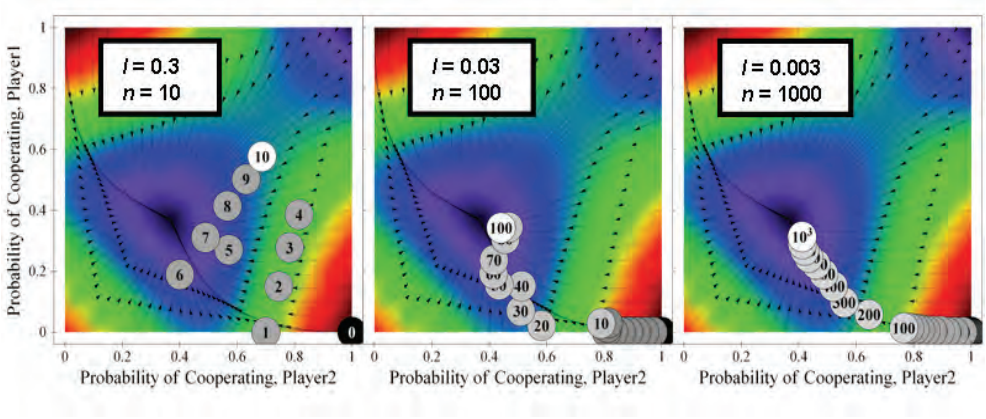

Fig. 8. Three sample runs of a system parameterised as [4 , 3 , 1 , 0 | 2 | l]² for different values of n and l. The product $n \cdot l$ is the same for the three simulations; therefore, for low values of l, the state of the system at the end of the simulations tends to concentrate around the same point.

Remember, however, that the system will eventually approach its asymptotic behaviour, which in the systems shown in figures 2, 3, 4, 5, 6, 7 and 8 is certain mutual cooperation. Having said that, as Binmore et al., (1995) point out, approaching the asymptotic behaviour may require an extraordinarily long time, much longer than is often meant by long run, hence the term ultralong run.

To illustrate how learning rates affect the speed of convergence to asymptotic behaviour, consider once again the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]², a system extensively studied by Macy & Flache (2002). The evolution of the probability to cooperate with initial state [0.5 , 0.5] (with these settings the probability is identical for both

[5] Excluded here is the trivial case where the initial state is an SRE.

players) is represented in the rows of figure 9 for different learning rates l. The top row shows the evolution for $l = 0.5$, and the bottom row shows the evolution for $l = 2^{-7}$.
For $l = 0.5$, after only $2^9 = 512$ iterations, the probability that both players will be almost certain to cooperate is very close to 1, and it remains so thereafter. For $l = 2^{-4}$ and lower learning rates, however, the distribution is still clustered around the SCE even after $2^{21} = 2097152$ iterations. With low learning rates, the chain of events that is required to escape from the neighbourhood of the SCE is extremely unlikely, and therefore this long run regime seems to persist indefinitely. However, given sufficient time, such a chain of coordinated moves will occur, and the system will eventually reach its ultralong-run regime, i.e. almost-certain mutual cooperation. The convergence of the processes to the appropriate point in the trajectory $f_x(n \cdot l)$ as $l \to 0$ and $n \cdot l$ is kept bounded can be appreciated following the grey arrows (which join histograms for which $n \cdot l$ is constant).

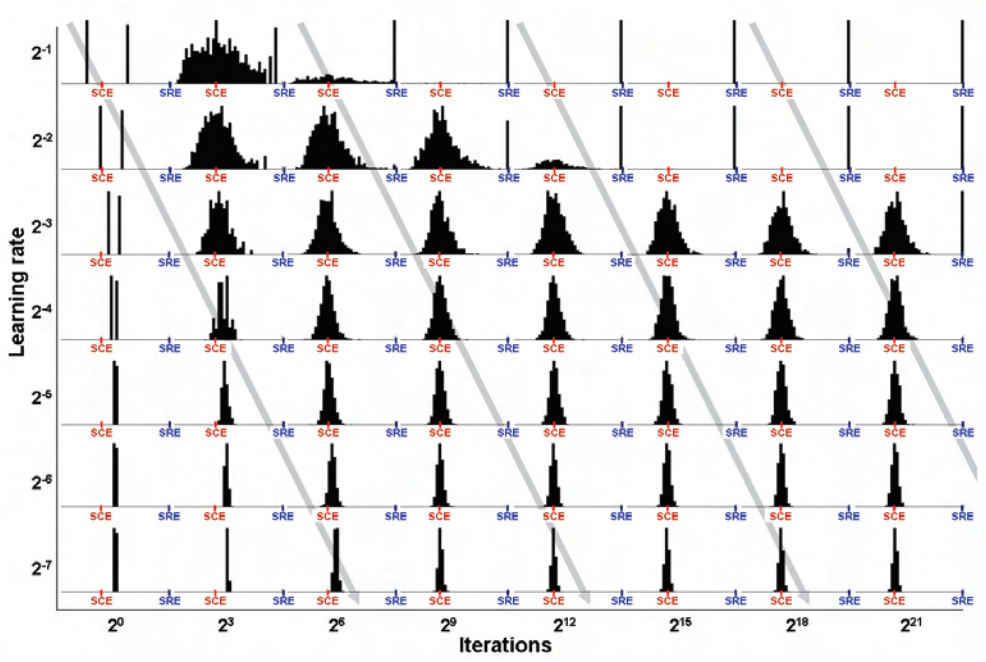

Fig. 9. Histograms representing the probability of cooperating for one player (both players' probabilities are identical) after n iterations for different learning rates l in a Prisoner's Dilemma game parameterised as $[\,4\,,\,3\,,\,1\,,\,0\mid 2\mid l\,]^2$, each calculated over 1,000 simulation runs. The initial probability for both players is 0.5. The grey arrows join histograms for which $n \cdot l$ is constant.

8. Trembling hands process

To study the robustness of the previous asymptotic results we consider an extension of the BM model where players suffer from 'trembling hands' (Selten, 1975): after having decided

which action to undertake, each player i may select the wrong action with some probability $\varepsilon_i > 0$ in each iteration. This noisy feature generates a new stochastic process, namely the *noisy process* $\boldsymbol{N}_n$, which can also be fully characterized by a 2-dimensional vector $\boldsymbol{prop} = [prop_1 , prop_2]$ of *propensities* (rather than probabilities) to cooperate. Player i's actual probability to cooperate is now $(1 - \varepsilon_i) \cdot prop_i + \varepsilon_i \cdot (1 - prop_i)$, and the profile of propensities $\boldsymbol{prop}$ evolves after any particular outcome following the rules given in section 4. Izquierdo et al. (2007) prove that the noisy process $\boldsymbol{N}_n$ is ergodic in any 2×2 game[6]. Ergodicity implies that the state of the process presents an asymptotic probability distribution that does not depend on the initial state.

The noisy process has no absorbing states (i.e. SREs) except in the trivial case where both players find one of their actions always satisfactory and the other action always unsatisfactory – thus, for example, in the Prisoner's Dilemma the inclusion of noise precludes the system from convergence to a single state. However, even though noisy processes have no SREs in general, the SREs of the associated unperturbed process (SREUPs, which correspond to mutually satisfactory outcomes) do still act as attractors whose attractive power depends on the magnitude of the noise: *ceteris paribus* the lower the noise the higher the long run chances of finding the system in the neighbourhood of an SREUP (see Figure 10). This is so because in the proximity of an SREUP, if ε_i are low enough, the SREUP's associated mutually satisfactory outcome will probably occur, and this brings the system even closer to the SREUP. The dynamics of the noisy system will generally be governed also by the other type of attractor, the SCE (see figure 10).

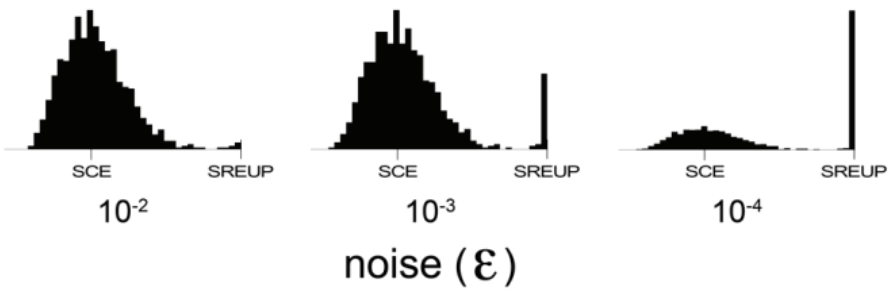

Fig. 10. Histograms representing the propensity to cooperate for one player (both players' propensities are identical) after 1,000,000 iterations (when the distribution is stable) for different levels of noise ($\varepsilon_i = \varepsilon$) in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.25][2]. Each histogram has been calculated over 1,000 simulation runs.

Figures 11 and 12, which correspond to a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l][2], show that the presence of noise can greatly damage the stability of the (unique) SREUP associated to the event CC. Note that the inclusion of noise implies that the probability of an infinite chain of the mutually satisfactory event CC becomes zero.

[6] We exclude here the meaningless case where the payoffs for some player are all the same and equal to her aspiration ($T_i = R_i = P_i = S_i = A_i$ for some i).

The systems represented on the left-hand side of figure 11, corresponding to a learning rate l = 0.5, show a tendency to be quickly attracted to the state [1 , 1], but the presence of noise breaks the chains of mutually satisfactory CC events from time to time (see the series on the bottom-left corner); unilateral defections make the system escape from the area of the SREUP before going back towards it again and again. The systems represented on the right-hand side of figure 11, corresponding to a lower learning rate (l = 0.25) than those on the left, show a tendency to be lingering around the SCE for longer. In these cases, when a unilateral defection breaks a chain of mutually satisfactory events CC and the system leaves the proximity of the state [1 , 1], it usually takes a large number of periods to go back into that area again.

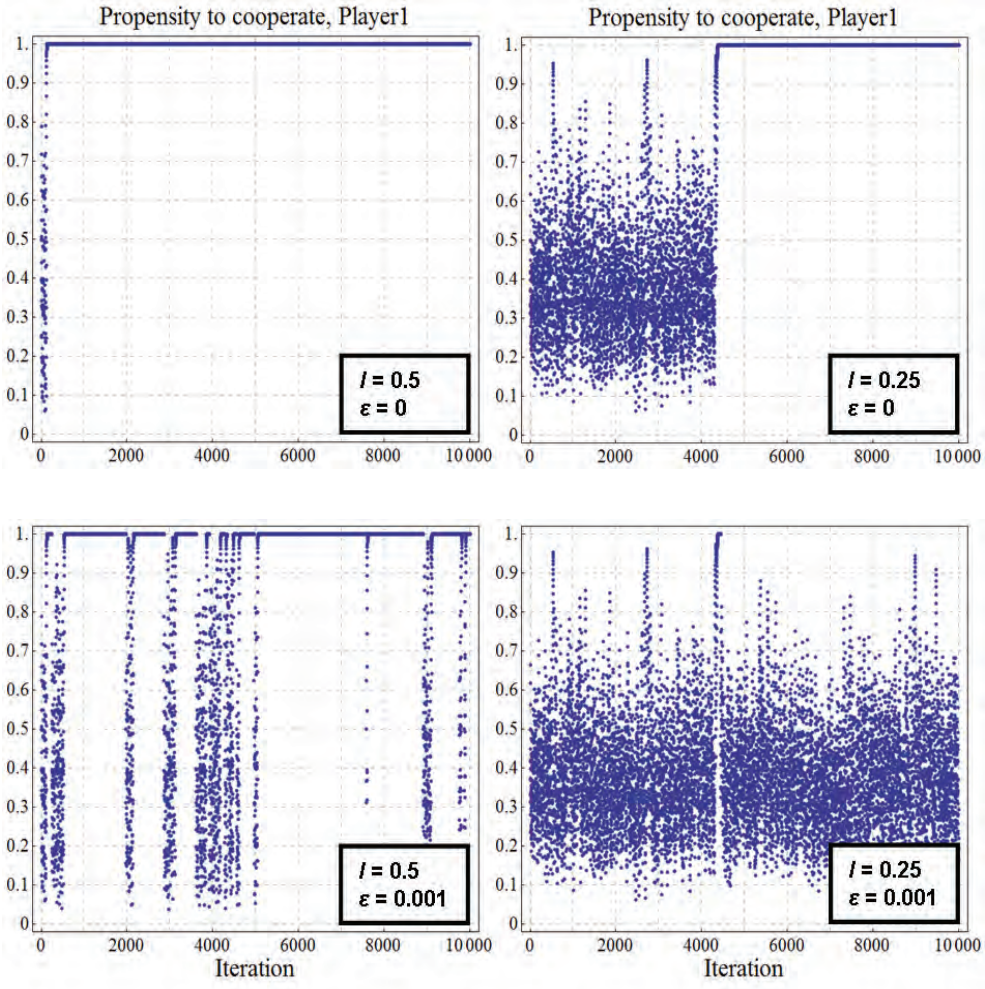

Fig. 11. Representative time series of player 1's propensity to cooperate over time for the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 |0.5]² (left) and [4 , 3 , 1 , 0 | 2

|0.25]² (right), with initial conditions [x_0 , x_0] = [0.5 , 0.5], both without noise (top) and with noise level $\varepsilon_i = 10^{-3}$ (bottom).

Figure 12 shows that a greater level of noise implies higher destabilisation of the SREUP. This is so because, even in the proximity of the SREUP, the long chains of reinforced CC events needed to stabilise the SREUP become highly unlikely when there are high levels of noise, and unilateral defections (whose probability increases with noise in the proximity of the SREUP) break the stability of the SREUP.

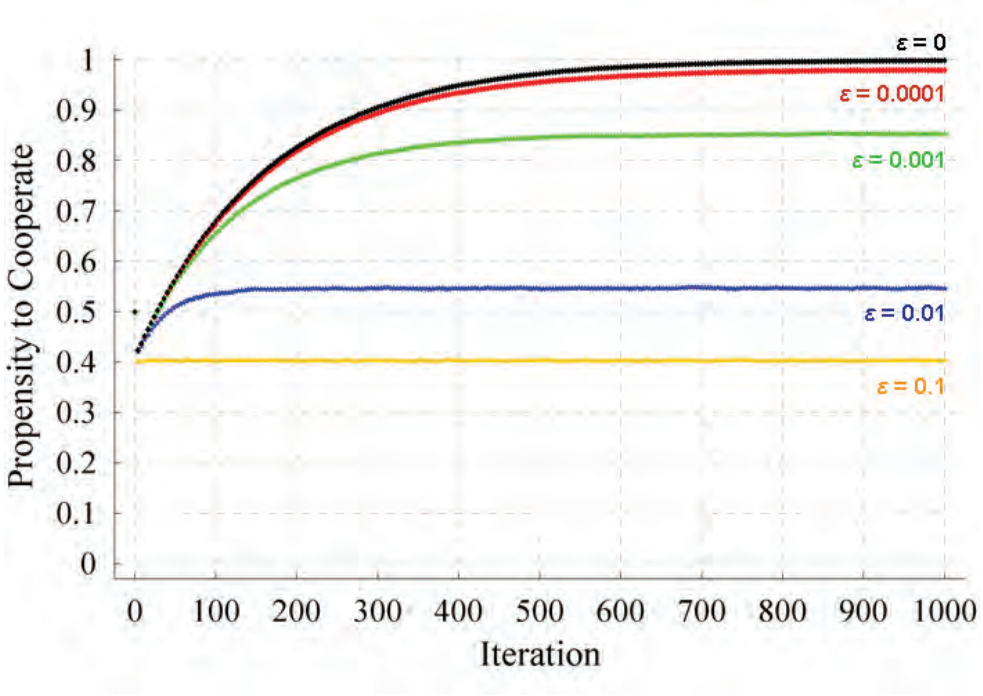

Fig. 12. Evolution of the average probability / propensity to cooperate of one of the players in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0. 5]² with initial state [0.5, 0.5], for different levels of noise ($\varepsilon_i = \varepsilon$). Each series has been calculated averaging over 100,000 simulation runs. The standard error of the represented averages is lower than $3 \cdot 10^{-3}$ in every case.

8.1 Stochastic stability

Importantly, not all the SREs of the unperturbed process are equally robust to noise. Consider, for instance, the system [4 , 3 , 1 , 0 | 0.5 | 0. 5]², which has two SRES: [1 , 1] and [0 , 0]. Using the results outlined in section 7 we know that the set formed by the two SREs is asymptotically reached with probability 1; the probability of the process converging to one particular SRE depends on the initial state; and if the initial state is completely mixed, then the process may converge to either SRE. Simulations of this process show that, almost in every case, the system quickly approaches one of the SREs and then remains in its close

vicinity. Looking at the line labelled "ε = 0" in figure 13 we can see that this system with initial state [0.9 , 0.9] has a probability of converging to its SRE at [1 , 1] approximately equal to 0.7, and a probability of converging to its SRE at [0 , 0] approximately equal to 0.3. However, the inclusion of (even tiny levels of) noise may alter the dynamics of the system dramatically. In general, for low enough levels of "trembling hands" noise we find an ultralong-run (invariant) distribution concentrated on neighbourhoods of SREUPs. The lower the noise, the higher the concentration around SREUPs. If there are several SREUPs, the invariant distribution may concentrate on some of these SREUPs much more than on others. In the limit as the noise goes to zero, it is often the case that only some of the SREUPs remain points of concentration. These are called stochastically stable equilibria (Foster & Young, 1990; Young, 1993; Ellison, 2000). As an example, consider the simulation results shown in figure 13, which clearly suggest that the SRE at [0 , 0] is the only stochastically stable equilibrium even though the unperturbed process converges to the other SRE more frequently with initial conditions [0.9 , 0.9]. Note that whether an equilibrium is stochastically stable or not is independent on the initial conditions.

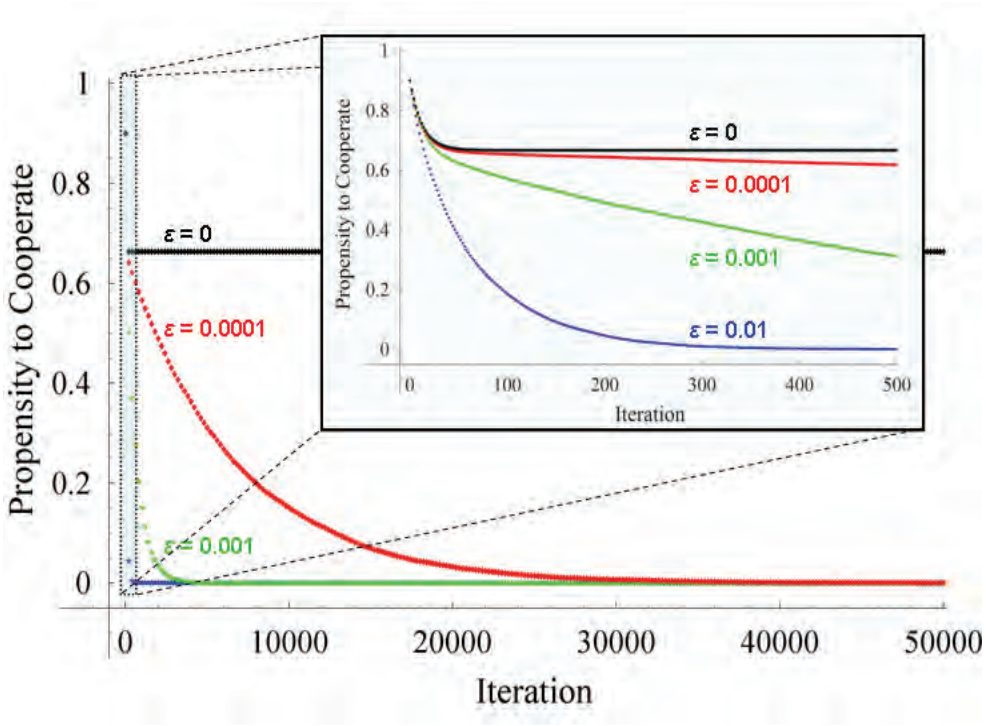

Fig. 13. Evolution of the average probability / propensity to cooperate of one of the players in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 0.5 | 0. 5]² with initial state [0.9 , 0.9], for different levels of noise ($\varepsilon_i = \varepsilon$). Each series has been calculated averaging over 10,000 simulation runs. The inset graph is a magnification of the first 500 iterations. The standard error of the represented averages is lower than 0.01 in every case.

Intuitively, note that in the system shown in figure 13, in the proximities of the SRE at [1 , 1], one single (possibly mistaken) defection is enough to lead the system away from it. On the other hand, near the SRE at [0 , 0] one single (possibly mistaken) cooperation will make the system approach this SRE at [0 , 0] even more closely. Only a coordinated mutual cooperation (which is highly unlikely near the SRE at [0 , 0]) will make the system move away from this SRE. This makes the SRE at [0 , 0] much more robust to occasional mistakes made by the players when selecting their strategies than the SRE at [1, 1], as illustrated in figures 14 and 15.

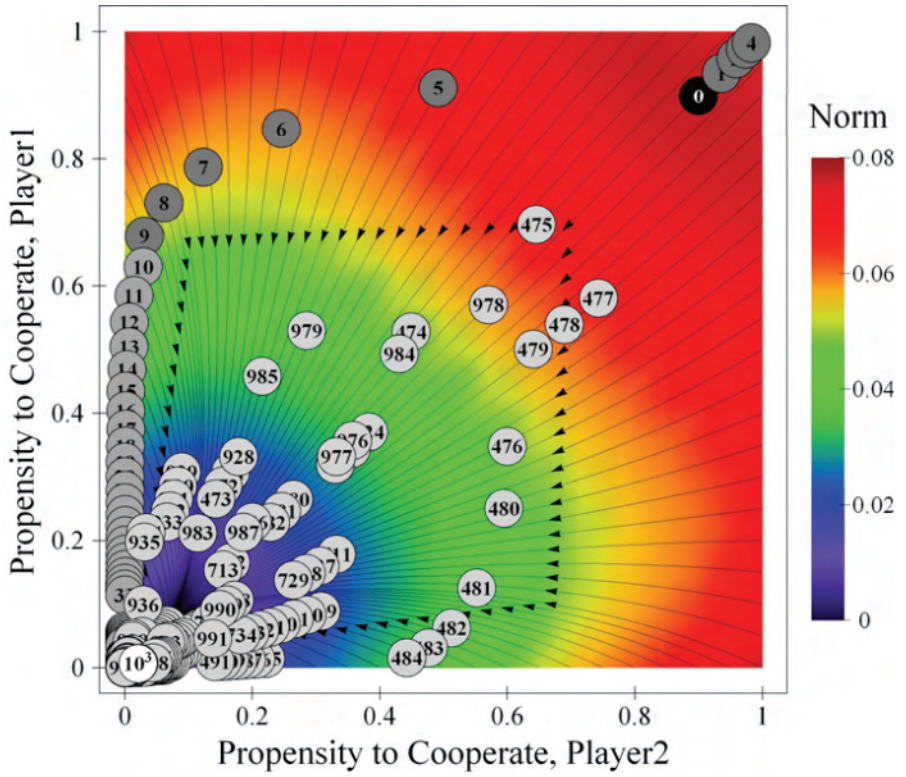

Fig. 14. One representative run of the system parameterised as [4 , 3 , 1 , 0 | 0.5 | 0. 5]^2 with initial state [0.9 , 0.9], and noise $\varepsilon_i = \varepsilon = 0.1$. This figure shows the evolution of the system in the phase plane of propensities to cooperate, while figure 15 below shows the evolution of player 1's propensity to cooperate over time for the same simulation run.

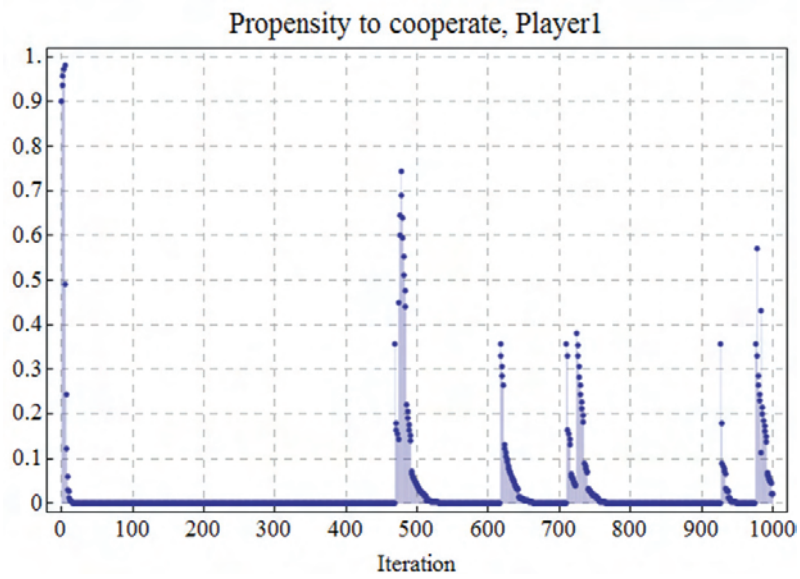

Fig. 15. Time series of player 1's propensity to cooperate over time for the same simulation run displayed in figure 14.

9. Conclusions

This chapter has characterised the behaviour of the Bush-Mosteller (Bush & Mosteller, 1955) aspiration-based reinforcement learning model in 2x2 games. The dynamics of this process depend mainly on three features:

- The speed of learning.
- The existence of self-reinforcing equilibria (SREs). SREs are states which are particularly relevant for the ultralong-run or asymptotic behaviour of the process.
- The existence of self-correcting equilibria (SCEs). SCEs are states which are particularly relevant for the transient behaviour of the process with low learning rates.

With high learning rates, the model approaches its asymptotic behaviour fairly quickly. If there are SREs, such asymptotic dynamics are concentrated on the SREs of the system. With low learning rates, two transient distinct regimes (medium run and long run) can usually be distinguished before the system approaches its asymptotic regime. Such transient dynamics are strongly linked to the solutions of the continuous time limit approximation of the system's expected motion.

The inclusion of small quantities of noise in the model can change its dynamics quite dramatically. Some states of the system that are asymptotically reached with high probability in the unperturbed model (i.e. some SREs) can effectively lose all their attractiveness when players make occasional mistakes in selecting their actions. A field for further research is the analytical identification of the asymptotic equilibria of the unperturbed process that are robust to small trembles (i.e. the set of stochastically stable equilibria).

10. Acknowledgements

We are grateful to the Spanish Ministry of Education and Science (Projects DPI2004-06590 and DPI2005-05676 –SIGAME-) and the University of Burgos for providing financial support to conduct this piece of research. We are also very grateful to Jörgen W. Weibull for deriving a mathematical result that we used to produce figure 7, and to Alexandre Eudes, Alexis Revue and Vincent Barra, for helping to implement the applet provided at http://www.luis.izquierdo.name/papers/rl-book.

11. References

Beggs, AW. (2005). On the Convergence of Reinforcement Learning. *Journal of Economic Theory* 122, 1-36.

Bendor, J.; Mookherjee, D. & Ray, D. (2001a). Aspiration-Based Reinforcement Learning In Repeated Interaction Games: An Overview. *International Game Theory Review* 3(2-3), 159-174.

Bendor, J.; Mookherjee, D. & Ray, D. (2001b). Reinforcement Learning in Repeated Interaction Games. *Advances in Theoretical Economics* 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. *Journal of Institutional and Theoretical Economics* 150, 45-63.

Binmore, K.; Samuelson, L. & Vaughan, R. (1995) Musical Chairs: Modeling Noisy Evolution. *Games and Economic Behavior* 11(1), 1-35.

Börgers, T. & Sarin, R. (1997). Learning through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* 77, 1-14.

Bush, R. & Mosteller, F. (1955). *Stochastic Models of Learning*. John Wiliey & Son, New York.

Camerer, CF. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Russell Sage Foundation, New York.

Castellano, C.; Marsili, M. & Vespignani, A. (2000). Nonequilibrium phase transition in a model for social influence. *Physical Review Letters*, 85(16), pp. 3536-3539.

Chen, Y. & Tang, F. (1998). Learning and Incentive-Compatible Mechanisms for Public Goods Provision: An Experimental Study. *Journal of Political Economy* 106, 633-662.

Colman, AM. (1995). *Game theory and its applications in the social and biological sciences*. 2nd edition. Butterworth-Heinemann, Oxford, UK

Cross, JG. (1973). A Stochastic Learning Model of Economic Behavior. *Quartertly Journal of Economics* 87, 239-266.

Duffy, J. (2005). Agent-Based Models and Human Subject Experiments. In: Judd, K. L., Tesfatsion, L. (Eds.), *Handbook of Computational Economics II: Agent-Based Computational Economics*. Amsterdam: Elsevier.

Edwards, M.; Huet, S.; Goreaud, F. & Deffuant, G. (2003). Comparing an individual-based model of behaviour diffusion with its mean field aggregate approximation. *Journal of Artificial Societies and Social Simulation*, 6(4)9 http://jasss.soc.surrey.ac.uk/6/4/9.html.

Ellison, G. (2000). Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution. *Review of Economic Studies*, 67, 17-45.

Erev, I. & Roth, AE. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review* 88(4), 848-881.

Erev, I.; Bereby-Meyer, Y. & Roth, AE. (1999). The effect of adding a constant to all payoffs: experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior and Organization* 39(1), 111-128.

Flache, A. & Macy, MW. (2002). Stochastic Collusion and the Power Law of Learning. *Journal of Conflict Resolution* 46(5), 629-653.

Foster, D. & Young, HP. (1990). Stochastic Evolutionary Game Dynamics. *Theoretical Population Biology*, 38, 219-232.

Galán, JM. & Izquierdo, LR. (2005). Appearances Can Be Deceiving: Lessons Learned Re-Implementing Axelrod's 'Evolutionary Approach to Norms'. *Journal of Artificial Societies and Social Simulation*, 8(3)2. http://jasss.soc.surrey.ac.uk/8/3/2.html.

Gilboa I, Schmeidler D (1995) Case-Based Decision Theory. *The Quarterly Journal of Economics*, 110, 605-639.

Gilboa I, Schmeidler D (2001) *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, UK

Hargreaves Heap, SP. & Varoufakis, Y. (1995). *Game theory: a critical introduction*. Routledge, London

Holt, CA. & Roth, AE. (2004). The Nash equilibrium: A perspective. *Proceedings of the National Academy of Sciences USA* 101(12), 3999–4002.

Hopkins, E. & Posch, M. (2005). Attainability of Boundary Points under Reinforcement Learning. *Games and Economic Behavior* 53 (1), 110-125.

Huet, S.; Edwards, M. & Deffuant, G. (2007). Taking into Account the Variations of Neighbourhood Sizes in the Mean-Field Approximation of the Threshold Model on a Random Network. *Journal of Artificial Societies and Social Simulation* 10(1)10 http://jasss.soc.surrey.ac.uk/10/1/10.html.

Ianni, A. (2001). Reinforcement Learning and the Power Law of Practice. *Mimeo*, University of Southampton.

Izquierdo, LR.; Izquierdo, SS.; Gotts, NM. & Polhill, JG. (2007). Transient and Asymptotic Dynamics of Reinforcement Learning in Games. *Games and Economic Behavior* 61(2), 259–276.

Izquierdo, SS.; Izquierdo, LR. & Gotts, NM. (in press). Reinforcement Learning Dynamics in Social Dilemmas. *Journal of Artificial Societies and Social Simulation*.

Izquierdo, LR. & Polhill, JG. (2006). Is Your Model Susceptible to Floating-Point Errors?. *Journal of Artificial Societies and Social Simulation* 9(4)4, http://jasss.soc.surrey.ac.uk/9/4/4.html.

Kushner, HJ. & Yin, GG. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag. New York.

Laslier J.; Topol, R. & Walliser, B. (2001). A Behavioral Learning Process in Games. *Games and Economic Behavior* 37, 340-366.

Macy, MW. & Flache, A. (2002). Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences USA* 99(3), 7229-7236.

Maier, NRF. & Schneirla, TC. (1964). *Principles of Animal Psychology*. Dover Publications, New York.

McAllister, PH. (1991). Adaptive approaches to stochastic programming. *Annals of Operations Research* 30, 45-62.

Mohler, RR. (1991). *Nonlinear Systems, Volume I: Dynamics and Control*. Prentice Hall, Englewood Cliffs.

Mookherjee, D. & Sopher, B. (1994). Learning Behavior in an Experimental Matching Pennies Game. *Games and Economic Behavior* 7, 62-91.

Mookherjee, D. & Sopher, B. (1997). Learning and Decision Costs in Experimental Constant Sum Games. *Games and Economic Behavior* 19, 97-132

Norman, MF. (1968). Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators. *Journal of Mathematical Psychology* 5, 61–101.

Norman, MF. (1972). *Markov Processes and Learning Models*. Academic Press, New York.

Polhill, JG. & Izquierdo, LR. (2005). Lessons learned from converting the artificial stock market to interval arithmetic. *Journal of Artificial Societies and Social Simulation*, 8 (2)2, http://jasss.soc.surrey.ac.uk/8/2/2.html

Polhill, JG.; Izquierdo, LR. & Gotts, NM. (2005). The ghost in the model (and other effects of floating point arithmetic). *Journal of Artificial Societies and Social Simulation*, 8 (1)5, http://jasss.soc.surrey.ac.uk/8/1/5.html

Polhill, JG.; Izquierdo, LR. & Gotts, NM. (2006). What every agent based modeller should know about floating point arithmetic. *Environmental Modelling and Software*, 21 (3), March 2006. pp. 283-309.

Roth, AE. & Erev, I. (1995). Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior* 8, 164-212.

Rustichini, A. (1999). Optimal Properties of Stimulus–Response Learning Models. *Games and Economic Behavior* 29, 244-273.

Selten, R. (1975). Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory* 4, 25-55.

Thorndike, EL. (1898). *Animal Intelligence: An Experimental Study of the Associative Processes in Animals* (Psychological Review, Monograph Supplements, No. 8). MacMillan, New York.

Thorndike, EL. (1911). *Animal Intelligence*. New York: The Macmillan Company.

Young, HP. (1993). The evolution of conventions. *Econometrica*, 61(1): 57–84

Wustmann G.; Rein K.; Wolf R. & Heisenberg M. (1996). A New Paradigm for Operant Conditioning of Drosophila Melanogaster. *Journal of Comparative Physiology [A]* 179, 429-436.