4. Dynamics of the Bush-Mosteller Reinforcement Learning Algorithm in 2x2 Games[♣]

4.1. 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 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

[♣] Some parts of the material presented in this chapter are in press in Izquierdo, L.R., Izquierdo, S.S., Gotts, N.M. and Polhill, J.G. (2007), "Transient and asymptotic dynamics of reinforcement learning in games", *Games and Economic Behavior* , and others have been accepted for publication in the *Journal of Artificial Societies and Social Simulation*.

information beyond the payoff they receive and specifically may be unaware of the strategic nature of the situation (Mookherjee and Sopher, 1994; Roth and Erev, 1995; Bendor et al., 2001a; Camerer, 2003; Duffy, 2006). In the context of experimental game theory with human subjects, several authors have used simple models of reinforcement learning successfully to explain and predict behaviour in a wide range of games (McAllister, 1991; Mookherjee and Sopher, 1994; Roth and Erev, 1995; Mookherjee and Sopher, 1997; Chen and Tang, 1998; Erev and Roth, 1998; Erev et al., 1999; Erev and Roth, 2001). 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 behaviour in addition to approach behaviour (e.g. Bendor et al. (2001b) vs. Erev and 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, 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 favour 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 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 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; Beggs, 2005; Hopkins and Posch, 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 minimum 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 analysed 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 behaviour (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), and Bendor et al. (2001b)).

The model analysed 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 analysed 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 analysing 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 Hegselmann and Flache (2000), Macy and Flache (2002) and Flache and Macy (2002) in 2×2 social dilemmas using computer simulation. Here we formalize their analyses and extend their results to cover any 2×2 game.

In contrast to other reinforcement learning models in the literature, we show that, in general, the asymptotic behaviour of the BM model cannot be approximated using the continuous time limit version of its expected motion. Such an approximation may be valid over bounded time intervals but it can deteriorate 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 behaviour of the BM model is characterized in the present chapter using the theory of distance diminishing models (Norman, 1968, 1972). Börgers and Sarin (1997) also used this theory to analyse 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.

4.2. The BM model

The model we analyse here is an elaboration of a conventional Bush-Mosteller (Bush and 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 $Y_i = \{\text{C, D}\}$. Let u_i be the *payoff function* that gives 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, $u_i(\text{C, D})$ 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. 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 $p_{i,j} \in \text{R}$ 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, 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:

$$s_i(\boldsymbol{y}) = \frac{u_i(\boldsymbol{y}) - A_i}{\mathbf{sup}_{\boldsymbol{k}\in Y} \mid u_i(\boldsymbol{k}) - A_i \mid}$$

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 $\mathbf{sup}_{\boldsymbol{k}\in Y} \mid u_i(\boldsymbol{k}) - A_i \mid$. 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:

$$p_{i,y_i}^{n+1} = \begin{cases} p_{i,y_i}^n + l_i \cdot s_i(\boldsymbol{y}^n) \cdot (1 - p_{i,y_i}^n) & \text{if } s_i(\boldsymbol{y}^n) \geq 0 \\ p_{i,y_i}^n + l_i \cdot s_i(\boldsymbol{y}^n) \cdot p_{i,y_i}^n & \text{if } s_i(\boldsymbol{y}^n) < 0 \end{cases} \qquad [4\text{-}1]$$

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.

A 2×2 BM model parameterization requires specifying both players' payoff function u_i, aspiration level (A_i), and learning rate (l_i). Unless otherwise stated, the analysis conducted here is valid for any 2×2 game but, for illustrative purposes, we focus on 2×2 symmetric social dilemma games where both players are parameterised in exactly the same way (homogeneous models). A certain parameterisation of such a homogeneous model will be specified using the template [*Temptation* , *Reward* , *Punishment* , *Sucker* | *A* | *l*]^2.

The following notation will also be useful. 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, however, that there are only two independent variables in $\boldsymbol{p}$, so the state of the game can be determined using a two-dimensional vector [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$], where $p_{i,\mathrm{C}}$ is player i's probability to cooperate (the actual name of the action is irrelevant for the mathematical analysis). 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 $\{\boldsymbol{P}_n(\boldsymbol{S})\}_{n\geq 0}$ constitutes a discrete-time Markov process with potentially infinite transient states. In a slight abuse of notation we refer to such a process $\{\boldsymbol{P}_n(\boldsymbol{S})\}_{n\geq 0}$ as the BM process $\boldsymbol{P}_n$.

4.3. 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[11]. An SRE corresponds to a pair of pure strategies ($p_{i,j}$ 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 $[\, p_{1,\mathrm{C}} , p_{2,\mathrm{C}} \,] = [\, 1 , 1 \,]$ is an SRE if both players' aspiration levels are below their respective $Reward_i$. 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.

Flache and Macy (2002, p. 634) define SCEs in the following way: "The SCE obtains when the expected change of probabilities is zero and there is a positive probability of punishment as well as reward". In this context, punishment means negative stimulus while reward means positive stimulus; the expected change of probability for one player is defined as the sum of the possible changes in probability the player might experience weighted by the likelihood of such changes actually happening. As we show below, SCEs defined in this way are not necessarily attractors, but may be unstable saddle points where small

[11] The concept of SRE is extensively used by Macy and Flache but we have not found a clear definition in their papers (Flache and Macy, 2002; Macy and Flache, 2002). Sometimes their use of the word SRE seems to follow our definition (e.g. Macy and Flache, 2002, p. 7231), but often it seems to denote a mutually satisfactory outcome (e.g. Macy and Flache, 2002, p. 7231) or an infinite sequence of such outcomes (e.g. Macy and Flache, 2002, p. 7232).

perturbations can cause expected probabilities to move away from them. Figure 4-1 represents the expected movement after one time-step for different states of the system in a Stag Hunt game. The Expected Motion (**EM**) of a system $\boldsymbol{S}$ in state $\boldsymbol{p}$ for the following iteration is given by a function vector $\mathbf{EM}^{S}(\boldsymbol{p})$ whose components are, for each player, the expected change in the probabilities of undertaking each of the two possible actions. Mathematically,

$$\mathbf{EM}^{S}(\boldsymbol{p}) \equiv \mathbf{E}(\Delta \boldsymbol{P}_{n}(\boldsymbol{S}) \mid \boldsymbol{P}_{n}(\boldsymbol{S}) = \boldsymbol{p})$$

In the context of 2×2 social dilemma games, the two independent components of the equation above can be rewritten as follows:

$$\mathrm{EM}^{S}_{i,\mathrm{C}}(\boldsymbol{p}) = \\ \Pr\{\mathrm{CC}\} \cdot \Delta p_{i,\mathrm{C}}\big|_{\mathrm{CC}} + \Pr\{\mathrm{CD}\} \cdot \Delta p_{i,\mathrm{C}}\big|_{\mathrm{CD}} + \Pr\{\mathrm{DC}\} \cdot \Delta p_{i,\mathrm{C}}\big|_{\mathrm{DC}} + \Pr\{\mathrm{DD}\} \cdot \Delta p_{i,\mathrm{C}}\big|_{\mathrm{DD}}$$

where $\mathrm{EM}^{S}_{i,\mathrm{C}}(\boldsymbol{p})$ is the expected change in player *i*'s probability to cooperate, and {CC, CD, DC, DD} represent the four possible outcomes that may occur. Note that in general the expected change will not reflect the actual change in a simulation run, and to make this explicit we have included the trace of a simulation run starting in state [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$] = [0.5 , 0.5] in Figure 4-1. The expected change – represented by the arrows in Figure 4-1 – is calculated considering the four possible changes that could occur (see equation above), whereas the actual change in a simulation run – represented by the numbered balls in Figure 4-1 – is only *one* of the four possible changes (*e.g.* $\Delta p_{i,\mathrm{C}}\big|_{\mathrm{CC}}$, if both agents happen to cooperate). The source code used to create every figure in this chapter is available in the Supporting Material.

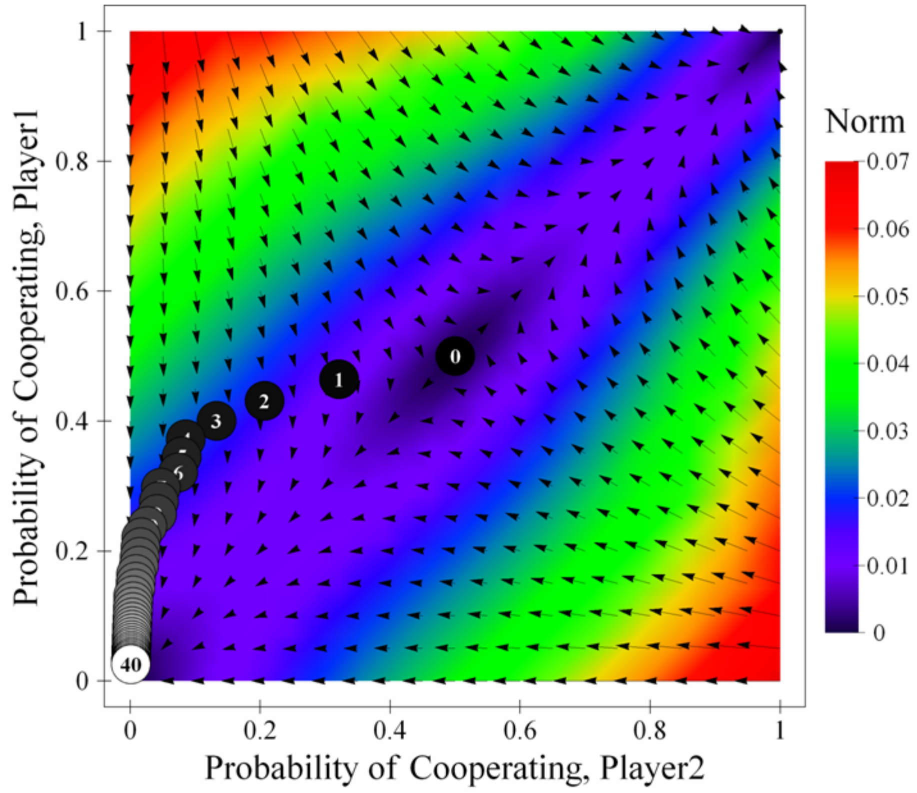

Figure 4-1. Expected motion of the system in a Stag Hunt game parameterised as [3 , 4 , 1 , 0 | 0.5 | 0.5 $]^2$, together with a sample simulation run (40 iterations). The arrows represent the expected motion for various states of the system; the numbered balls show the state of the system after the 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 would vary but their direction would be preserved.

The state [$p_{1,C}$, $p_{2,C}$] = [0.5 , 0.5] in Figure 4-1 is an example of a strategy profile that satisfies Flache and Macy's requirements for SCE, but where small deviations tend to lead the system away from it (saddle point). To avoid such undesirable situations where an SCE is not self-correcting, we redefine the concept of SCE in a more restrictive way: an SCE of a system ***S*** is an asymptotically stable critical point (Mohler, 1991) of differential equation [4-2] (the continuous time limit approximation of the system's expected motion).

$$\dot{\boldsymbol{f}} = \mathbf{EM}^{S}(\boldsymbol{f}) \qquad [4\text{-}2]$$

Roughly speaking this means that all trajectories in the phase plane of Eq. [4-2] 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 this definition, there could be a state of the system that is an SRE and an SCE at the same time (this is not possible using Flache and Macy's definitions of SRE and SCE).

Figure 4-2 shows several trajectories for the differential equation corresponding to the Stag Hunt game used in Figure 4-1. It can be clearly seen that state [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$] = [0.5 , 0.5] is not an SCE according to our definition, since there are trajectories that get arbitrarily close to it, but then escape from its neighbourhood.

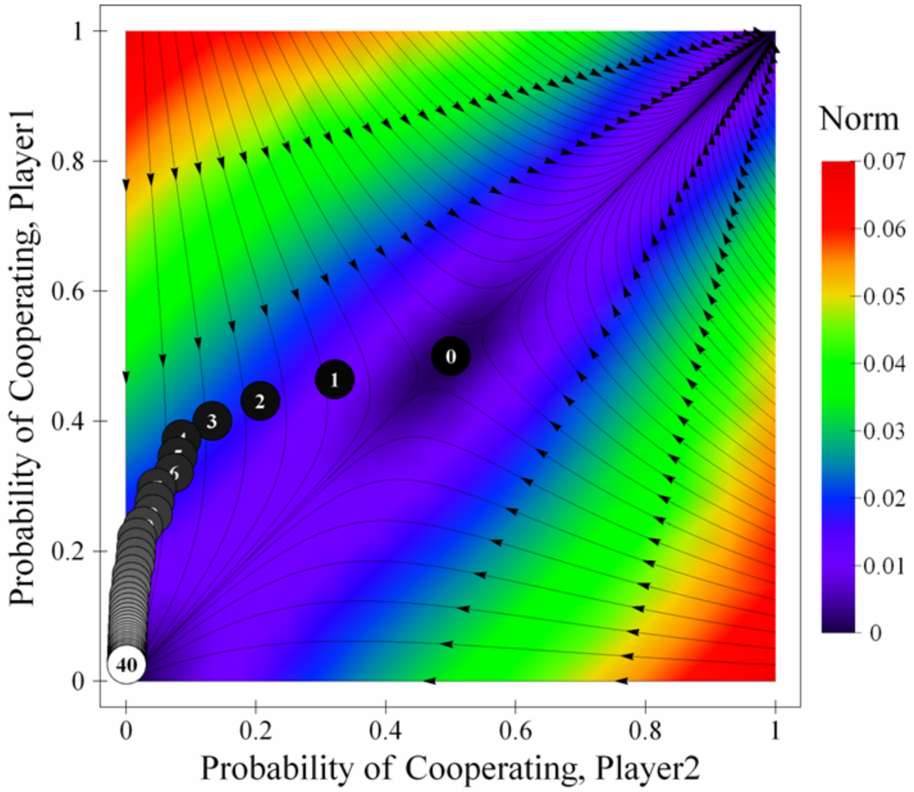

Figure 4-2. Trajectories in the phase plane of the differential equation corresponding to a Stag Hunt game parameterised as [3 , 4 , 1 , 0 | 0.5 | 0.5]^2, together with a sample simulation run (40 iterations). The background is coloured using the norm of the expected motion.

Figure 4-3 shows some trajectories of the differential equation corresponding to the Prisoner's Dilemma parameterised as [4 , 3 , 1 , 0 | 2 | l]^2. This system exhibits a unique SCE at [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$] = [0.37 , 0.37] and a unique SRE at [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$] = [1 , 1]. The two independent components of the function **EM(*p*)** for this system can be written as follows:

$$[\mathrm{EM}^S_{1,\mathrm{C}}(\boldsymbol{p}),\mathrm{EM}^S_{2,\mathrm{C}}(\boldsymbol{p})] = l\left[p_{1,\mathrm{C}}p_{2,\mathrm{C}} \quad p_{1,\mathrm{C}}(1-p_{2,\mathrm{C}}) \quad (1-p_{1,\mathrm{C}})p_{2,\mathrm{C}} \quad (1-p_{1,\mathrm{C}})(1-p_{2,\mathrm{C}})\right]\cdot \begin{bmatrix} (1-p_{1,\mathrm{C}})/2 & (1-p_{2,\mathrm{C}})/2 \\ -p_{1,\mathrm{C}} & -p_{2,\mathrm{C}} \\ -p_{1,\mathrm{C}} & -p_{2,\mathrm{C}} \\ (1-p_{1,\mathrm{C}})/2 & (1-p_{2,\mathrm{C}})/2 \end{bmatrix}$$

And the associated differential equation is

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

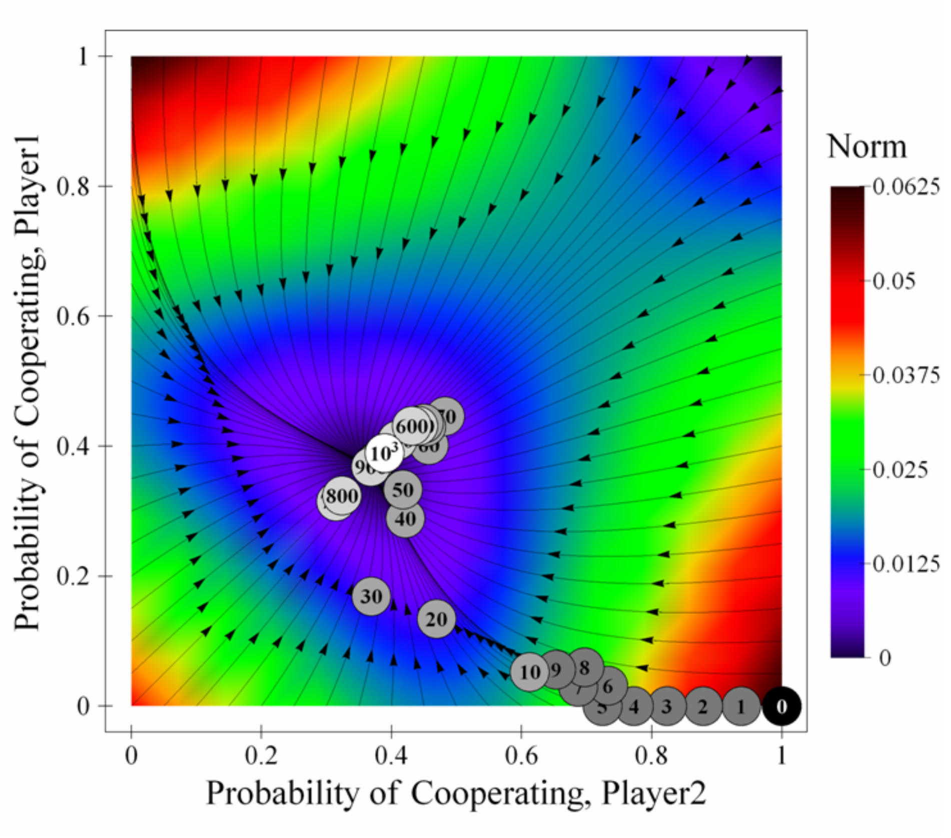

Figure 4-3. Trajectories in the phase plane of the differential equation corresponding to the Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l]^2, together with a sample simulation run ($l = 2^{-4}$). This system has a SCE at [$p_{1,\mathrm{C}}$, $p_{2,\mathrm{C}}$] = [0.37 , 0.37]. The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

Let $\boldsymbol{f_x}(t)$ denote the solution of the differential equation [4-2] for some initial state $\boldsymbol{x}$. As an example, Figure 4-4 shows $\boldsymbol{f_x}(t)$ for the Prisoner's Dilemma game parameterised as $[\ 4\ ,\ 3\ ,\ 1\ ,\ 0\ |\ 2\ |\ l\ ]^2$ for different (and symmetric) initial conditions $[\ p_{1,\mathrm{C}}\ ,\ p_{2,\mathrm{C}}\ ]\ = [\ x_0\ ,\ x_0\ ]$. For this particular case and settings, the two independent components of $\boldsymbol{f_x}(t)$ corresponding to each player's probability to cooperate – denoted $f_{i,x}(t)$ – take the same value at any given t, so the representation in Figure 4-4 corresponds to both these independent components. Convergence to the SCE at [0.37 , 0.37] can be clearly observed for every initial condition $[\ x_0\ ,\ x_0\ ]$, except for $[\ x_0\ ,\ x_0\ ] = [1, 1]$, which is the SRE.

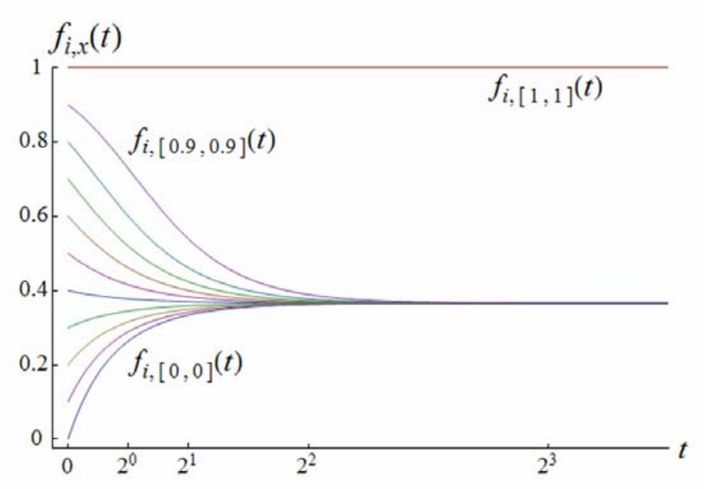

Figure 4-4. Solutions of differential equation [4-2] for the Prisoner's Dilemma game parameterised as $[\ 4\ ,\ 3\ ,\ 1\ ,\ 0\ |\ 2\ |\ l\ ]^2$ with different (and symmetric) initial conditions $[\ p_{1,\mathrm{C}}\ ,\ p_{2,\mathrm{C}}\ ]\ = [x_0\ ,\ x_0]$. This system has a unique SCE at $[\ p_{1,\mathrm{C}}\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

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. 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 and Izquierdo (2005); Edwards et al. (2003); Castellano, Marsili, and Vespignani (2000)). Note, however, that such approaches are approximations whose validity may be constrained to specific conditions: as we can see in Figure 4-3, simulation runs and trajectories will not coincide in general. 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* behaviour 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. From now on we will use our definitions of SRE and SCE.

4.4. Attractiveness of SREs

Macy & Flache's experiments (Flache and Macy, 2002; Macy and Flache, 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. $A_i < Reward_i = u_i$(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)[12]. To investigate this apparent 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 4-5 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 [$p_{1,C}$, $p_{2,C}$] = [0.37 , 0.37] and a unique SRE at [$p_{1,C}$, $p_{2,C}$] = [1 , 1]. Figure 4-5 also includes the trace of a sample simulation run. Note that the only difference between the

[12] 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 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).

parameterisation of the system shown in Figure 4-3 and that shown in Figure 4-5 is the value of the learning rate.

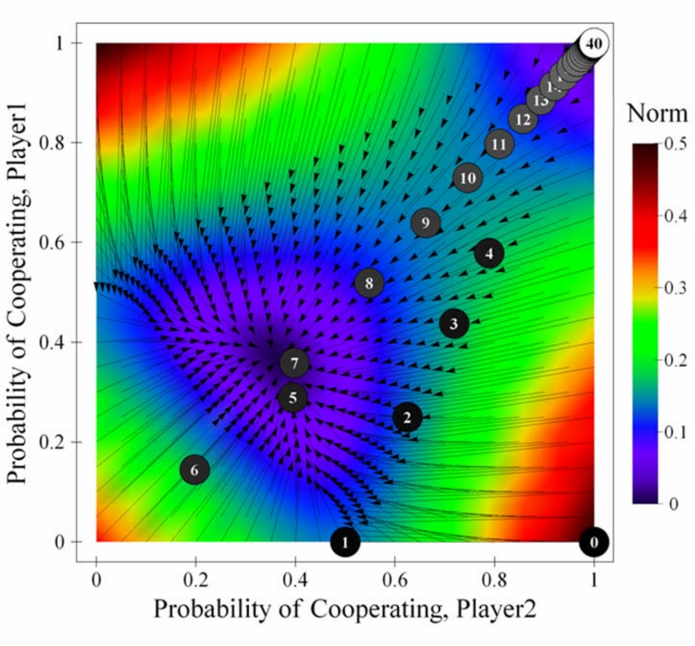

Figure 4-5. Expected motion of the system in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.5]^2, with a sample simulation run.

Figure 4-5 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 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 4-6.

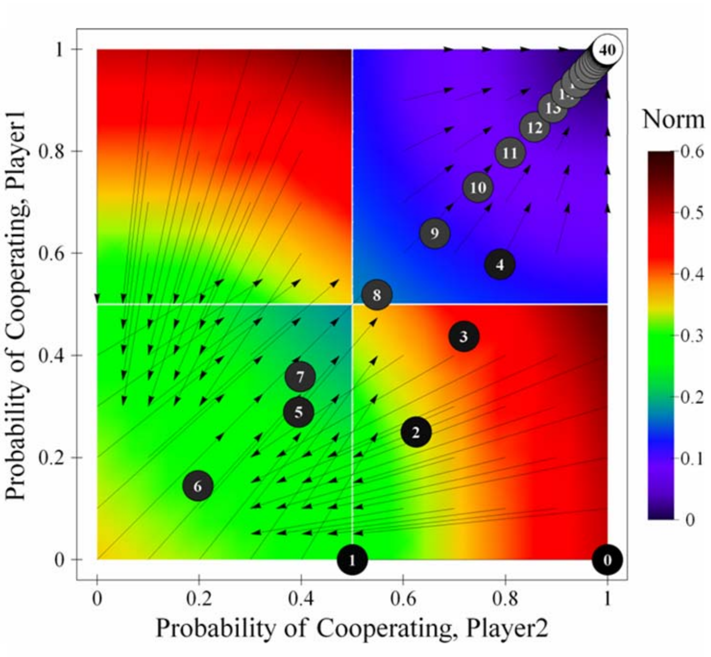

Figure 4-6 Figure showing the most likely movements at some states of the system in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0.5 $]^2$, with a sample simulation run. The background is coloured using the norm of the most likely movement.

Figure 4-6 differs significantly from Figure 4-5; 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 characterized 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 4-5).

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. This latter probability can be calculated using a result derived by Professor Jörgen W. Weibull (see Appendix A). The probability of starting an infinite chain of CC events depends largely on the value of the learning rate *l*. Figure 4-7 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%).

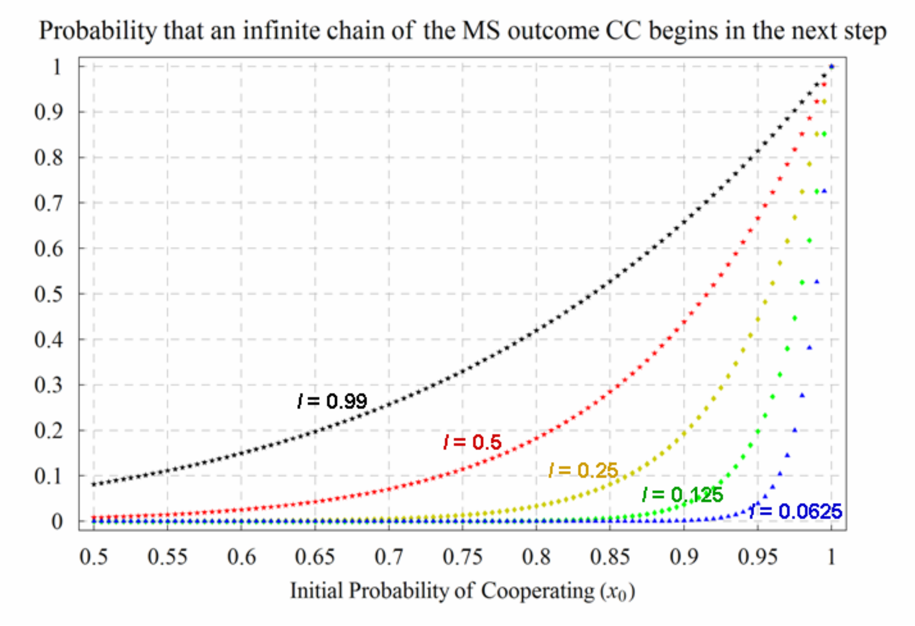

Figure 4-7. Probability of starting an infinite chain of the Mutually Satisfactory (MS) outcome CC in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | l $]^2$. The 5 different (coloured) series correspond to different learning rates l. The variable x_0, represented in the horizontal axis,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

In summary, assuming that aspirations are different from payoffs, 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.

4.5. 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 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 dynamics of the system are dramatically different from its asymptotic behaviour. Whether the three different regimes (i.e. medium, long, and ultralong run) are clearly distinguishable strongly depends on the players' learning rates. For high learning rates the system quickly approaches its asymptotic behaviour 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 the following section 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 characterised by the trajectories in the phase plane of eq. [4-2]. 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 behaviour, which in most BM systems consists in approaching an SRE asymptotically (see formal analysis below).

For an illustration of the different regimes, consider once again the Prisoner's Dilemma game parameterised as $[\ 4\ ,\ 3\ ,\ 1\ ,\ 0 \mid 2 \mid l\ ]^2$. 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 with initial state $[p_{1,C}\ ,\ p_{2,C}] = [\ 0.5\ ,\ 0.5\ ]$ (with these settings the probability is identical for both players) is represented in the rows of Figure 4-8 for different learning rates l.

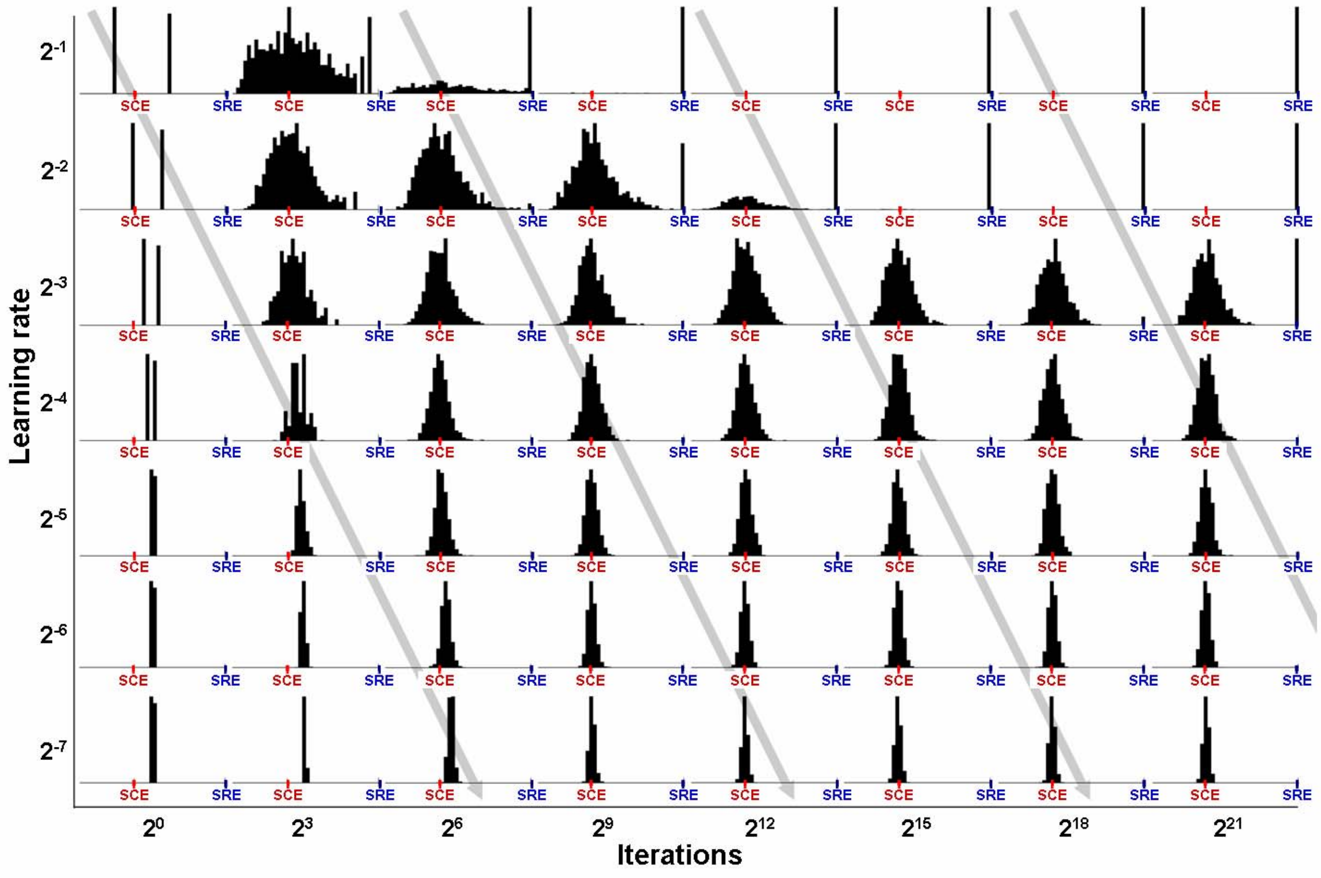

Figure 4-8. Histograms representing the probability to cooperate for one player (both players' probabilities are identical) after *n* iterations, for different learning rates $l_i = l$, with $A_i = 2$, in a symmetric Prisoner's Dilemma with payoffs [4 , 3 , 1 , 0]. Each histogram has been calculated over 1,000 simulation runs. The initial probability for both players is 0.5. The significance of the gray arrows will be explained later in the text.

For l = 0.5 (see top row in Figure 4-8), 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 following sections are devoted to the formal analysis of the transient and asymptotic dynamics of the BM model. The proofs of every proposition in this chapter are included in Appendix A.

4.6. Transient Dynamics

As mentioned above, when learning takes place by large steps the system quickly approaches its asymptotic behaviour, and no clear (transient) patterns are observed before it does so (see top row in Figure 4-8). 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 $\boldsymbol{P}_n^l = \boldsymbol{P}_n(\boldsymbol{S}^l)$ be the family of stochastic processes associated with such a family of systems $\boldsymbol{S}^l$. As an example, note that Figure 4-8 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. [4-3] below, and let $\boldsymbol{f}_{\boldsymbol{x}}(t)$ be the trajectory of this ODE with initial state $\boldsymbol{x}$.

$$\dot{\boldsymbol{f}} = \frac{1}{l}\mathbf{EM}^{S^l}(\boldsymbol{f}) \qquad [4\text{-}3]$$

The ODE in eq. [4-3] 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. [4-2] (which determines a different ODE for each member) in the time scale, *i.e.* the representation of the trajectories of ODEs [4-2] and [4-3] 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 4-1: Consider the family of stochastic processes $\{\boldsymbol{P}_n^{l,x}\}_{n\geq 0}$ with initial state $\boldsymbol{P}_0^l = \boldsymbol{x}$ for every l. Let K be an arbitrary constant. For learning by small steps ($l \to 0$) and transient behaviour ($n{\cdot}l \leq K < \infty$), we have:

i. For fixed $\varepsilon > 0$ and l sufficiently small,

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

where, for fixed $K < \infty$, $C(l, K) \to 0$ as $l \to 0$. Thus, for transient behaviour and learning by small steps, we have uniform convergence in probability of $\boldsymbol{P}_n^{l,x}$ to the trajectory $\boldsymbol{f}_{\boldsymbol{x}}$ of the ODE in [4-3].

ii. The distribution of the variable $\dfrac{\boldsymbol{P}_n^{l,x} - \boldsymbol{f}_x(n\cdot l)}{\sqrt{l}}$ 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 $\boldsymbol{P}_n^{l,x}$ 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 4-1, consider the particular family $\boldsymbol{F}^l$ (Figure 4-8). 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 $\boldsymbol{P}_n^x(\boldsymbol{F}^l)$ 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 [4-3] 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 Figure 4-3, and the convergence of the processes to the appropriate point in the trajectory $\boldsymbol{f}_{\mathbf{x}}(n{\cdot}l)$ as $l \to 0$ can be appreciated following the gray arrows (which join histograms for which $n{\cdot}l$ is constant) in Figure 4-8. Figure 4-9 illustrates this convergence in the phase plane. The grey arrows in Figure 4-8 also illustrate statement (ii): the distribution of $\boldsymbol{P}_n^x(\boldsymbol{F}^l)$ approaches normality with decreasing variance as $l \to 0$, keeping $n{\cdot}l$ constant.

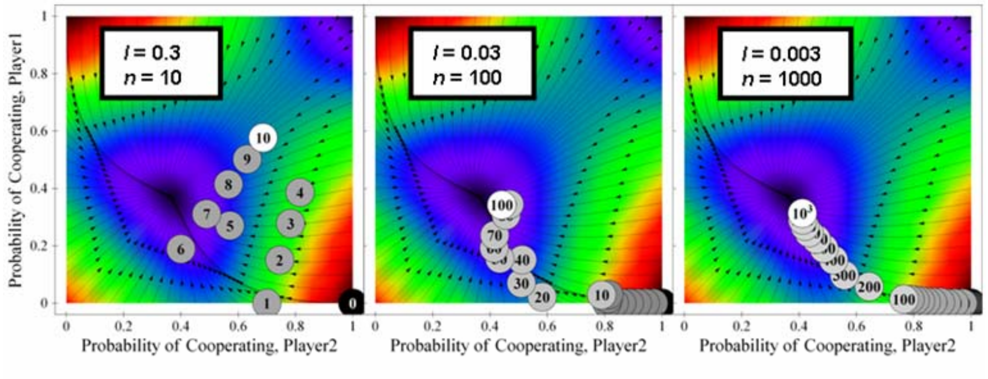

Figure 4-9. Three sample runs of a system parameterised as [4 , 3 , 1 , 0 | 2 | l]^2 for different values of n and l. The product $n{\cdot}l$ is the same for the three simulations; therefore, for low values of l, the state of the system at the end of the simulations tends to concentrate around the same point.

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 Figure 4-3). 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 Figure 4-8 we consider the region to the left of a grey arrow that is sufficiently to the right), the distribution of $\boldsymbol{P}_n^x(\boldsymbol{F}^l)$ 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 $\boldsymbol{P}_n^x(\boldsymbol{F}^l)$ within a neighbourhood 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*, $\boldsymbol{P}_n^x(\boldsymbol{F}^l)$ will eventually converge to the unique SRE [1, 1] in the ultralong run ($n \rightarrow \infty$). This is proved in the following section.

4.7. Asymptotic Behaviour

This section presents theoretical results on the asymptotic (i.e. ultralong run) behaviour of the BM system. Note that with low learning rates the system may take an extraordinarily long time to reach its ultralong-run behaviour (e.g. see bottom row in Figure 4-8).

Proposition 4-2: 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*[13], the BM process $\boldsymbol{P}_n$ converges to an SRE with probability 1 (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.

Proposition 4-3: 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 asymptotically reached with probability 1). If the initial state is completely mixed, then every SRE can be asymptotically reached with positive probability.
ii. If there is no SRE then the BM process $\boldsymbol{P}_n$ is ergodic[14] with no absorbing state.

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

[14] 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 $n \rightarrow \infty$) is unique and does not depend on the initial state.

Corollary to Proposition 4-3: Consider any of the three 2×2 social dilemma games: Prisoner's Dilemma, Chicken, and Stag Hunt (see section 3.1). 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 4-4: Consider the BM process $\boldsymbol{P}_n$ and its associated ODE (eq. [4-2] or [4-3]) 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[15]) 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 4-4 can be

[15] An outcome is protected if unilateral deviations by any player do not hurt any of the other players (Bendor et al., 2001b).

strengthened for the special case where all stimuli are positive (Phansalkar et al., 1994; Sastry et al., 1994).

4.8. 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 ***prop*** evolves after any particular outcome following the rules given by eq. [4-1]. Theorem 2.2 in Norman (1968, p. 67) can be used to prove that this noisy process is ergodic in any 2×2 game[16]. Proposition 4-1 applies to this extension too.

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 neighborhood of an SREUP (see Figure 4-10). 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 4-10).

[16] We exclude here the meaningless case where the payoffs for some player are all the same and equal to her aspiration.

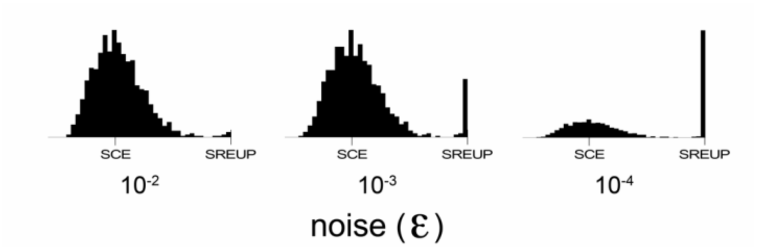

Figure 4-10. Histograms representing the propensity to cooperate for one player (both players' propensities are identical) after 1,000,000 iterations (when the distribution is stable) for different levels of noise ($\varepsilon_i = \varepsilon$) in a Prisoner's Dilemma game parameterised as [4 , 3 ,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

Figure 4-11 and Figure 4-12, 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.

The systems represented on the left-hand side of Figure 4-11, 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 4-11, 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.

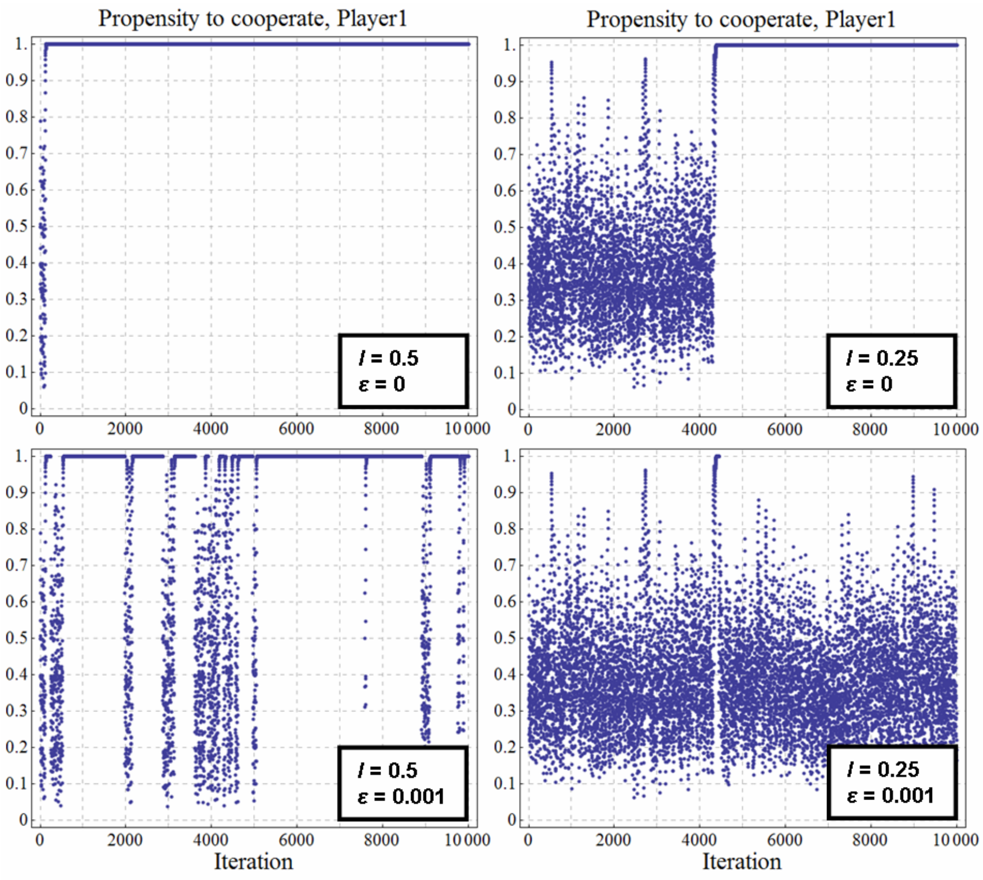

Figure 4-11. Representative time series of player 1's propensity to cooperate over time for the Prisoner's Dilemma game parameterised as $[4\ ,3\ ,1\ ,0\ |\ 2\ |\ 0.5\ ]^2$ (left) and $[4\ ,3\ ,1\ ,0\ |\ 2\ |\ 0.25\ ]^2$ (right), with initial conditions $[\ x_0\ ,x_0\ ] = [\ 0.5\ ,0.5\ ]$, both without noise (top) and with noise level $\varepsilon_i = 10^{-3}$ (bottom).

Figure 4-12 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.

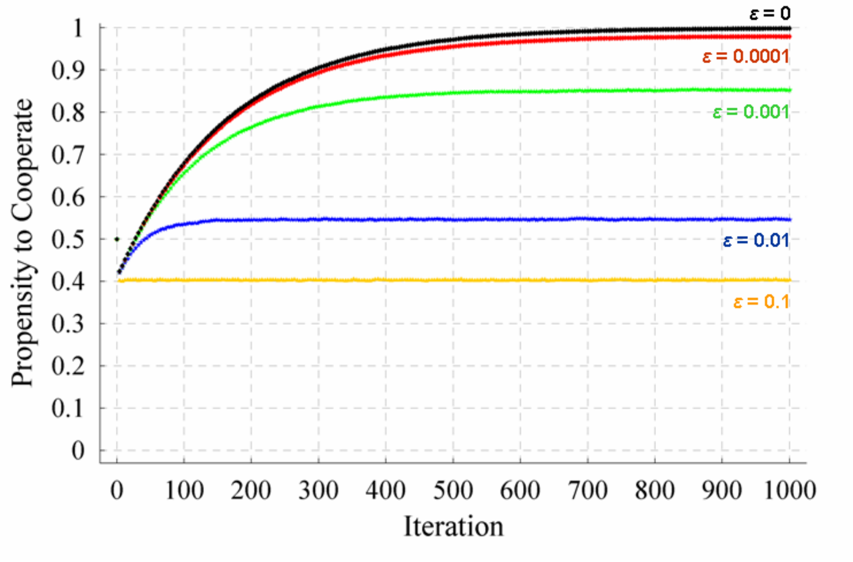

Figure 4-12. Evolution of the average probability / propensity to cooperate of one of the players in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 2 | 0. 5 $]^2$ with initial state [0.5 , 0.5], for different levels of noise (ε_i = ε). Each series has been calculated averaging over 100,000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

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: [$p_{1,C}$, $p_{2,C}$] = [1 , 1] and [$p_{1,C}$, $p_{2,C}$] = [0 , 0]. Using Proposition 4-2 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, in almost 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 4-13 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 and Young, 1990; Young, 1993; Ellison, 2000) and will be discussed in detail in chapter 5. As an example, consider the simulation results shown in Figure 4-13, 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.

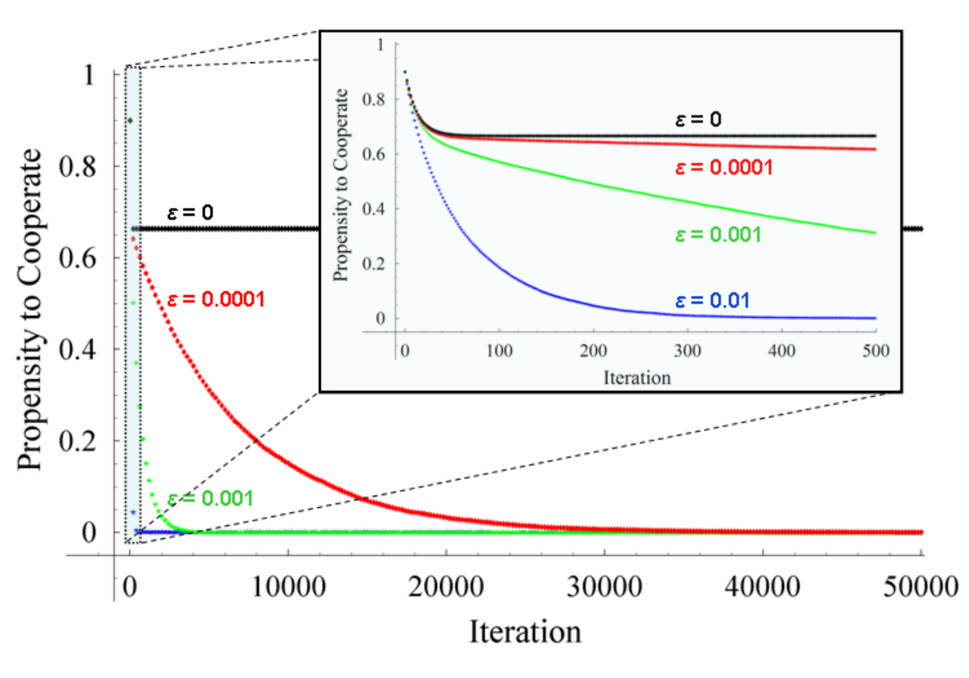

Figure 4-13. Evolution of the average probability / propensity to cooperate of one of the players in a Prisoner's Dilemma game parameterised as [4 , 3 , 1 , 0 | 0.5 | 0. 5]^2 with initial state [0.9 , 0.9], for different levels of noise ($\varepsilon_i = \varepsilon$). Each series has been calculated 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 4-13, 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 Figure 4-14 and Figure 4-15.

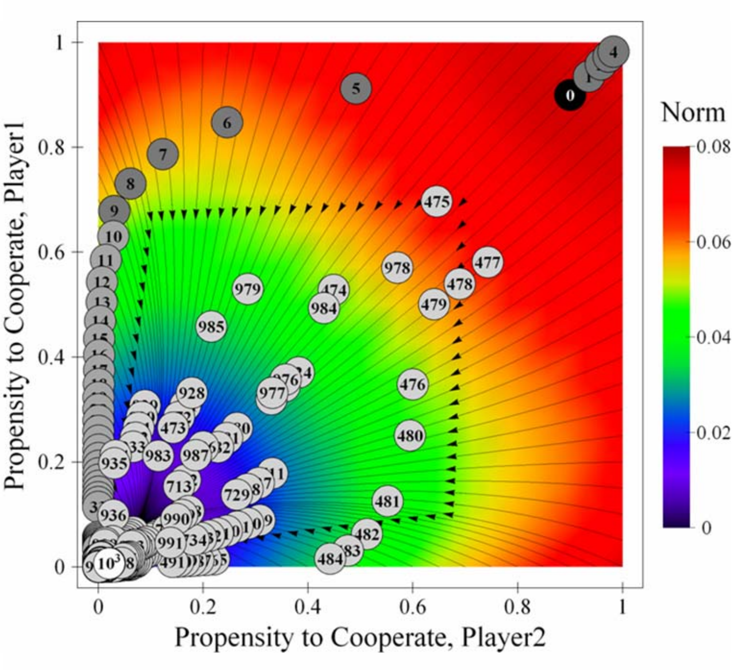

Figure 4-14. One representative run of the system parameterised as [4 , 3 , 1 , 0 | 0.5 | 0. 5 $]^2$ with initial state [0.9 , 0.9], and noise $\varepsilon_i = \varepsilon = 0.1$. This figure shows the evolution of the system in the phase plane of propensities to cooperate, while figure 15 below shows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

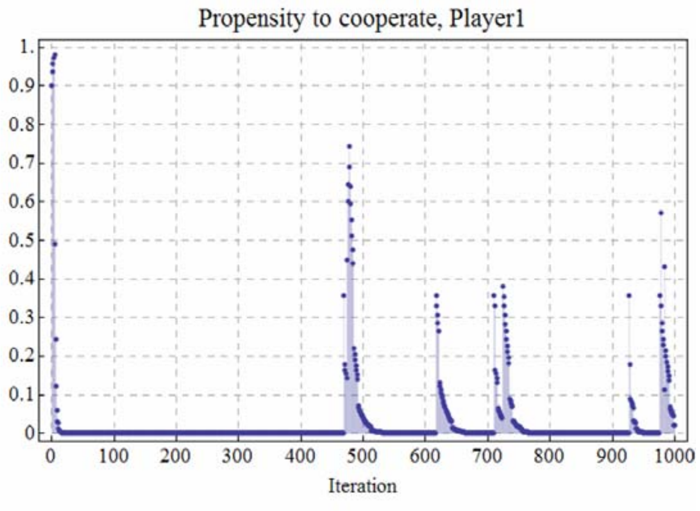

Figure 4-15. Time series of player 1's propensity to cooperate over time for the same simulation run displayed in Figure 4-14.

4.9. Extensions

The theoretical results on asymptotic behaviour presented in this chapter 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 behaviour 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 4-4 (which derives from Sastry et al. (1994)) can be easily extended to any finite game.

More immediately, every proposition in this chapter can be directly applied to finite populations from which two players are randomly[17] 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 run. The more players in the group, the longer it takes the group to reach universal cooperation.

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 $\boldsymbol{y}$ in time step n, every player i (i = 1, 2…m) updates her strategy $\boldsymbol{p_i}$ using an adaptation rule $\boldsymbol{p}_i^{n+1} = \boldsymbol{r}_i^{y}(\boldsymbol{p}^n)$, where $\boldsymbol{r}_i^{y}(\boldsymbol{p}^n)$ is continuous for every $\boldsymbol{y}$

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

and every *i*. Let us call such adaptation rules continuous. Note that BM adaptation rules are continuous, and consider the following proposition.

Proposition 4-5: 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).

4.10. Conclusions of this chapter

This chapter 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 literature cited in section 4.1) has shown that the outcome of games played by reinforcement learners can be substantially different from the expected outcomes when the game is played among 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.

This chapter in particular has characterised the dynamics of the Bush-Mosteller (Bush and Mosteller, 1955) aspiration-based reinforcement learning model in 2x2 games. These dynamics have been shown to 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.

An extension of the Bush-Mosteller model where players suffer from trembling hands has also been explored. It has been shown that the inclusion of small quantities of noise in the original Bush-Mosteller 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).

This chapter has characterised not only the asymptotic behaviour 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 chapter has illustrated, the transient dynamics of learning algorithms may be substantially different from their asymptotic behaviour. 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 behaviour that we observe in the real world.