Cycles of cooperation and defection in imperfect learning

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Cycles of cooperation and defection in imperfect learning

Tobias Galla

Theoretical Physics, School of Physics and Astronomy, University of Manchester, Manchester M13 9PL, UK
E-mail: Tobias.Galla@manchester.ac.uk

Received 12 April 2011
Accepted 7 July 2011
Published 8 August 2011

Online at stacks.iop.org/JSTAT/2011/P08007
doi:10.1088/1742-5468/2011/08/P08007

Abstract. We investigate a model of learning the iterated prisoner's dilemma game. Players have the choice between three strategies: always defect (ALLD), always cooperate (ALLC) and tit-for-tat (TFT). The only strict Nash equilibrium in this situation is ALLD. When players learn to play this game convergence to the equilibrium is not guaranteed, for example we find cooperative behaviour if players discount observations in the distant past. When agents use small samples of observed moves to estimate their opponent's strategy the learning process is stochastic, and sustained oscillations between cooperation and defection can emerge. These cycles are similar to those found in stochastic evolutionary processes, but the origin of the noise sustaining the oscillations is different and lies in the imperfect sampling of the opponent's strategy. Based on a systematic expansion technique, we are able to predict the properties of these learning cycles, providing an analytical tool with which the outcome of more general stochastic adaptation processes can be characterised.

Keywords: game-theory (theory), applications to game theory and mathematical economics, stochastic processes

ArXiv ePrint: 1101.4378

Contents

1. Introduction

The mathematical theory of games goes back to von Neumann and Morgenstern [1], and was initially concerned with the study of equilibrium points [2, 3]. The idea that players would be able to compute such equilibria requires severe assumptions, in particular perfect rationality and full knowledge of the game. Additionally each player has to assert that all other players are rational as well. Since the work of von Neumann and Morgenstern more than 70 years ago, several different routes have been taken to formulate a dynamical theory of games. Evolutionary game theory was launched by Maynard Smith in the 1970s and considers time-dependent dynamics of populations of players [4, 5]. Each individual in the population carries a pure strategy, inherited from its parent(s), and agents then reproduce and pass on their strategies to their offspring, with a reproduction rate depending on the performance in the game. The strategic content of the population evolves, with the concentration of successful strategies increasing over time, and those of less successful strategies being reduced. Techniques from the theory of nonlinear systems and from statistical physics are increasingly being used to understand the dynamics of evolutionary systems [6, 7]. Selection processes in finite populations in particular are stochastic in nature, and they can be seen as interacting many-body systems, making methods from statistical mechanics readily applicable. Evolutionary systems have, for example, been studied with methods from the theory of stochastic processes [8] or with techniques used to characterise pattern formation [9].

Evolutionary game theory has been used to model a vast number of phenomena in the social sciences and in economics [10]–[16]. These applications include in particular the study of the emergence of cooperation and altruism [17]. The evolution of cooperative

doi:10.1088/1742-5468/2011/08/P08007

J. Stat. Mech. (2011) P08007

behaviour under selection pressure constitutes a formidable puzzle. The dynamics of evolution is governed by a fierce competition between individuals, and only those who act in their own interest and who selfishly promote their own evolutionary success at the expense of their competitors should prevail in the long run. Nevertheless altruism and cooperative behaviour are found in a number of evolved systems, ranging from cooperating genes or cells to cooperating animals or humans in social contexts [18]–[20]. The question how cooperative behaviour has evolved under strictly competitive and selective dynamics is still unresolved, and has recently been listed as one of the 125 big open problems in science [21].

Our goal here is to address the emergence of cooperation in a third approach to game theory. We focus on adaptive learning processes of a small fixed set of individuals, who interact repeatedly in a game [22]–[27]. Players observe their opponents' actions and aim to react dynamically by adapting their own strategic propensities, learning from past experience. Such learning models are of particular importance for the understanding of experiments in behavioural game theory, where human subjects play a given game repeatedly under controlled conditions, see, e.g., [25]–[30]. *A priori* it is not clear whether adaptation will converge to Nash equilibria. Learning has, for example, been seen to fail to converge in games with cyclic payoff structures, and complex trajectories including limit cycles, quasiperiodic motion and Hamiltonian chaos have instead been identified [31]–[33]. Noise during learning has been shown to prevent convergence in certain circumstances, see [36] for an initial study. Its effects on fixation of game learning has recently been investigated in [37].

Mathematical models of cooperative behaviour are often based on stylised games played by a small number of interacting individuals, each choosing from a small number of strategies. The most basic set-up is the celebrated prisoner's dilemma, a game in which two players have the choice between cooperation and defection. Defection dominates cooperation in this game; no matter what the other player decides to do, either player will always do better defecting than cooperating. Fully rational players hence end up playing the only equilibrium strategy, defection, and have to put up with the a suboptimal payoff, when they could have scored higher had they both cooperated.

If the prisoner's dilemma is iterated, more complex behaviour is possible and the space of all strategies grows rapidly as the number of iterations is increased. In order to make progress it is therefore necessary to restrict the mathematical analysis to a subset of this space. We will focus on three strategies: always defect (ALLD), always cooperate (ALLC) and tit-for-tat (TFT). Players using the TFT strategy cooperate in the first iteration and then proceed by playing whatever the opponent played in the previous round. The replicator–mutator dynamics of populations of players engaging in this game have been studied in [38, 39]. ALLD has been identified as the deterministic replicator fixed point and mutation has been seen to move the attractor towards cooperation. Demographic noise in finite populations can alter the dynamics and can induce coherent evolutionary cycles between defection and cooperation.

As one main result we show that the effects of memory loss in the learning dynamics are very similar to those of mutation in evolutionary dynamics. While deterministic learning in the absence of memory loss converges to ALLD, this Nash equilibrium is no longer an attractor when players discount observations in the distant past, and a different fixed point, involving all three pure strategies, emerges. Secondly, we focus on

J. Stat. Mech. (2011) P08007

the effects of sampling noise during learning. Deterministic replicator-type equations are a faithful description of the learning process if and only if a large number of observations of the opponent's actions is made before players update their own strategic preferences. If, in contrast, adaptation occurs more frequently and is based only a small sample of observations, the dynamics becomes stochastic. The source of randomness lies in the imperfect sampling of the opponent's mixed strategy profile. When each player uses a small number of observed actions to estimate the opponent's mixed strategy, then the estimate will generally be subject to statistical errors. The observed actions were chosen according to the opponent's mixed strategy profile, but still they are random variables. This source of noise is different from the origin of demographic noise in the evolution of finite populations. Nevertheless the effects are similar: as our second main result we show that sustained cycles between cooperation and defection can emerge in stochastic learning, similar to those found in evolutionary scenarios of the iterated prisoner's dilemma game [38]. Using an expansion in the inverse noise strength during learning [36] we are able to predict the characteristic frequency and power spectra of these cycles analytically as a function of the parameters of the game and the learning dynamics.

2. Model

2.1. Set-up of the game

To define the iterated prisoner's dilemma we will follow the notation of [38]. Assuming that m iterations of the prisoner's dilemma are played in any one interaction of the two players, and that a complexity cost c is associated with playing TFT the payoff matrix is given by

	ALLC	ALLD	TFT
ALLC	R	S	R
ALLD	T	P	$\frac{T+P(m-1)}{m}$
TFT	$\frac{Rm-c}{m}$	$\frac{S+P(m-1)-c}{m}$	$\frac{Rm-c}{m}$

i.e. a player playing ALLC will, for example, receive a payoff of R (per round) when meeting another ALLC player, a payoff of S when playing against ALLD and a payoff of R upon encountering TFT. We will denote the payoff matrix elements as a_{ij}, where $i, j = 1, 2, 3$ label the strategies ALLC, ALLD and TFT, respectively. Throughout this paper we use $T = 5, R = 3, P = 1, S = 0.1, m = 10$ and $c = 0.8$ (the parameters chosen in [38]). We keep these fixed, as we are mainly interested in the outcome of learning as the parameters of the adaptation algorithm are varied, and not primarily as a function of the underlying game. Other choices of iterated prisoner's dilemma games are possible of course, but we expect the general mechanism of stochastically sustained oscillations also to hold in such cases, whenever deterministic learning approaches a suitable stable fixed point. The theoretical analysis and formalism provided below applies to general two-player games.

2.2. Learning dynamics

In our model the game is played repeatedly by two players Alice and Bob. We will assume that Alice carries a (time-dependent) mixed strategy profile $\mathbf{x}(t) = (x_1(t), x_2(t), x_3(t))$

J. Stat. Mech. (2011) P08007

and similarly Bob's mixed strategy profile at t is $\mathbf{y}(t) = (y_1(t), y_2(t), y_3(t))$. We will write $i(t)$ for Alice's action at time t, and $j(t)$ for Bob's action, i.e. $i(t), j(t) \in \{\text{ALLC, ALLD, TFT}\}$. Following [25]–[27] each player keeps attractions for each of the pure strategies. Alice's attractions at time t are labelled by $A_i(t)$ and Bob's attractions by $B_j(t)$. We will again follow [25]–[27] as well as [31]–[33] and assume that attractions determine choice probabilities through a logit rule, i.e. that the probabilities for Alice and Bob to play the different pure strategies at time t are given by

$$x_i(t) = \frac{e^{\beta A_i(t)}}{\sum_k e^{\beta A_k(t)}}, \qquad y_j(t) = \frac{e^{\beta B_j(t)}}{\sum_k e^{\beta B_k(t)}}. \tag{1}$$

The variable β is a model parameter and describes the intensity of selection or response sensitivity [27]. For $\beta \to \infty$ the players strictly choose the pure action with highest attraction, for $\beta = 0$ they play at random. We will here restrict the discussion to models in which both players use the same intensity of selection; generalisation to heterogeneous intensities is straightforward.

A simple reinforcement learning dynamics is then defined by the following update rules for the attractions [31]–[33]:

$$A_k(t+1) = (1-\lambda)A_k(t) + a_{k,j(t)}, \qquad B_k(t+1) = (1-\lambda)B_k(t) + a_{k,i(t)}. \tag{2}$$

Alice's attraction A_k is therefore reinforced by the payoff $a_{k,j(t)}$ she would have received at time t had she played action k, and similarly for Bob. The parameter λ indicates memory loss; observations in the distant past carry a lesser weight than more recent rounds. For $\lambda = 0$ the players have perfect memory of past play, and use the outcome of all past rounds with equal weight to determine their attractions. In particular, A_k, for example, is then the total payoff Alice would have received had she always played action $k \in \{\text{ALLC}, \text{ALLD}, \text{TFT}\}$, given Bob's moves. For $\lambda > 0$ experiences in the past are discounted exponentially. This may happen voluntarily as part of a learning mechanism or simply be due to fading memories. We will occasionally refer to λ as a memory-loss rate or discounting factor. We assume that both players learn at identical memory-loss rates; generalisation to heterogeneous learning rules ($\lambda_{\text{Alice}} \neq \lambda_{\text{Bob}}$) is straightforward. Up to relabelling this learning rule is a special case of experience-weighed attraction learning, as discussed in [26, 27]. We note that other choices of update rules are possible, for example those introducing a factor λ in front of the second term on the RHS of equation (2) (see, e.g., [34]).

2.3. Deterministic limit and relation to replicator dynamics

The process defined by equations (1) and (2) is intrinsically stochastic, the actions $i(t)$ and $j(t)$ are drawn from the mixed strategy profiles $\mathbf{x}(t)$ and $\mathbf{y}(t)$, respectively, and accordingly the attractions $A_k(t)$ and $B_k(t)$ are random variables as well. Simple averaging, taking into account that $i(t)$ takes the value $i(t) = \ell$ with probability $x_\ell(t)$ and that $j(t) = \ell$ with probability $y_\ell(t)$, results in the following average attraction update:

$$\begin{aligned} A_k(t+1) &= (1-\lambda)A_k(t) + \sum_{\ell=1}^{3} a_{k\ell} y_\ell(t), \\ B_k(t+1) &= (1-\lambda)B_k(t) + \sum_{\ell=1}^{3} a_{k\ell} x_\ell(t). \end{aligned} \tag{3}$$

J. Stat. Mech. (2011) P08007

Limiting dynamics of this type can provide insight into the expected outcome of learning. Deterministic learning has been shown to lead to modified replicator equations in a continuous-time limit [31]–[33]. Analyses of discrete-time deterministic learning can be found in [40, 41]. The derivation of the deterministic dynamics relies on an adiabatic approximation though. It is assumed that strategy updates occur on a much slower timescale than the actual play. In order to perform the update of equations (3) Alice has to have full knowledge of Bob's mixed strategy $\mathbf{y}(t)$, and Bob needs to be aware of Alice's strategy $\mathbf{x}(t)$. This will generally be very hard to achieve for the players. Equations (3) are therefore only an approximate description of the learning process, and can at best be expected to describe the average behaviour. Describing learning in terms of these deterministic equations is procedurally akin to describing the average behaviour of evolving populations by means of deterministic replicator equations.

Taking into account equations (1) and (3) one can write the update rule solely in terms of $\mathbf{x}$ and $\mathbf{y}$ and finds the following map [32]:

$$x_i(t+1) = \frac{x_i(t)^{1-\lambda} \mathrm{e}^{\beta \sum_j a_{ij} y_j(t)}}{\sum_k x_k(t)^{1-\lambda} \mathrm{e}^{\beta \sum_j a_{kj} y_j(t)}}, \qquad y_j(t+1) = \frac{y_j(t)^{1-\lambda} \mathrm{e}^{\beta \sum_i a_{ji} x_i(t)}}{\sum_k y_k(t)^{1-\lambda} \mathrm{e}^{\beta \sum_i a_{ki} x_i(t)}}. \tag{4}$$

Taking, on the other hand, a continuous-time limit of equations (3), as discussed in [33], one finds

$$\dot{A}_k = -\lambda A_k + \sum_j a_{kj} y_j, \qquad \dot{B}_k = -\lambda B_k + \sum_i a_{ki} x_i. \tag{5}$$

Using equations (1) it is then straightforward to derive deterministic continuous-time evolution equations for the frequencies $\{x_i(t), y_j(t)\}$ with which the pure strategies $i = 1, \ldots, S$ are played by the respective players. One finds [32]

$$\begin{aligned} \dot{x}_i &= x_i \beta \left(\sum_k a_{ik} y_k - \sum_{k\ell} x_k a_{k\ell} y_\ell \right) - \lambda x_i \left(\log x_i - \sum_k x_k \log x_k \right), \\ \dot{y}_j &= y_j \beta \left(\sum_k a_{jk} x_k - \sum_{k\ell} y_k a_{k\ell} x_\ell \right) - \lambda y_j \left(\log y_j - \sum_k y_k \log y_k \right). \end{aligned} \tag{6}$$

These equations are occasionally referred to as the Sato–Crutchfield equations, and it is worth pointing out that they reduce to the standard replicator equations for the case of learning without memory loss ($\lambda = 0$). Furthermore their behaviour is solely determined by the ratio λ/β. If this ratio is fixed then the role of the remaining parameter is merely to set the timescale. It is also easy to verify that the fixed points of equations (6) coincide with those of equations (4). The behaviour of these dynamics can be quite intricate, depending on the structure of the underlying game. Sato *et al* have, for example, identified chaotic motion in modified versions of the celebrated rock–paper–scissors game [31]–[33].

2.4. Batch learning

We now address the differences in the dynamical behaviours of the stochastic dynamics, equation (2), and the deterministic limit, equations (3), both in discrete time. In order to understand the nature of the approximation underlying the deterministic limit it is instructive to interpolate between the deterministic average process and the actual

J. Stat. Mech. (2011) P08007

stochastic dynamics. We here consider a batch learning process, in which each player samples N actions of their respective opponent, and then updates their attractions. The above 'adiabatic' approximation consists in assuming stationarity of the mixed strategy profiles between attraction updates. Specifically we introduce the following process:

$$A_k(\tau+1) = (1-\lambda)A_k(\tau) + \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,i_\alpha(\tau)}, \qquad B_k(\tau+1) = (1-\lambda)B_k(\tau) + \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,j_\alpha(\tau)}. \tag{7}$$

The interpretation of these update rules is as follows: at time τ Alice independently selects N actions $i_\alpha(\tau)$ ($\alpha = 1,\ldots,N$) following her mixed strategy profile $\mathbf{x}(\tau)$ at that time. I.e. the $\{i_\alpha(\tau)\}$ are independent random variables, and for each α one has $i_\alpha(\tau) = \ell$ with probability $x_\ell(\tau)$. Bob draws his actions $j_\alpha(\tau)$ in a similar manner, using his mixed strategy $\mathbf{y}(\tau)$.These actions represent the moves made by the two players in N successive rounds of the game: the mixed strategies $\mathbf{x}(\tau)$ and $\mathbf{y}(\tau)$ are kept fixed during the course of these rounds. At the end of the batch of N rounds both Alice and Bob update their attractions based on equation (7) and then adapt their mixed strategy profiles using equation (1) (with t replaced by τ). We have intentionally used the notation τ rather than t to denote time steps of this batch dynamics. One unit of time τ corresponds to N repetitions of the game, i.e. to N units of time t. We will refer to N, the number of observations made in between updates of the attractions, as the batch size, following the language of machine learning [42]. Small batch sizes N correspond to fast adaptation. If $N = 1$ we recover the original dynamics of equation (2) where strategy updates are performed after every single round of the game. Large N, on the other hand, indicate infrequent adaptation; the limit of infinite batches leads to the deterministic update rule equations (3). This limit is based on the assumption that the mixed strategy profiles $\mathbf{x}(\tau)$ and $\mathbf{y}(\tau)$ are stationary during each batch of N repetitions of the game. This assumption will be irrelevant at small batch sizes N, but more severe in the limit of large N. Taking the limit $N \to \infty$ to derive the deterministic learning rule is analogous to the procedure leading to a description of evolving populations in terms of deterministic replicator equations. In evolutionary systems these descriptions are accurate for populations with an infinite number of individuals. Stochastic corrections cannot be neglected in finite populations though, and the resulting noise has been seen to alter the dynamics substantially, see, e.g., [6, 35] for general cases, and [38, 39] for the iterated prisoner's dilemma. Similarly, real-world players do not operate adiabatic learning dynamics, but instead small batch sizes N are probably more appropriate to describe experiments in behavioural economics. It is therefore important to go beyond the deterministic limit of equations (3) and to study stochastic effects at finite batch sizes. First steps have been taken in [36], and it is one of the main purposes of this work to apply these ideas to the iterated prisoner's dilemma game.

3. Results

3.1. Deterministic dynamics in continuous time

We illustrate the outcome of the continuous-time deterministic learning for the iterated prisoner's dilemma game in figure 1. Data has been obtained from a numerical integration

J. Stat. Mech. (2011) P08007

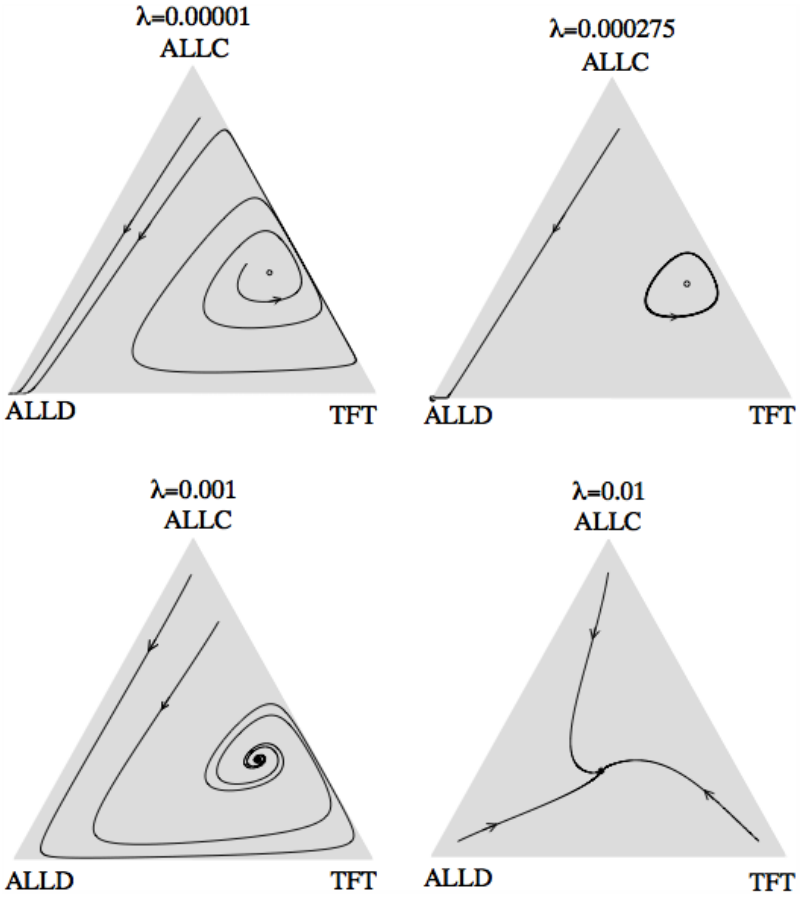

Figure 1. Illustration of the behaviour of the deterministic continuous-time learning at different memory-loss rates λ. Intensity of selection is $\beta = 0.01$.

of equations (6), using an Euler-forward scheme. At low memory-loss rates the dynamics is essentially governed by the standard replicator equations, and the system has a single stable fixed point near ALLD, similar to what is reported for low mutation rates in evolutionary systems [38]. As the memory-loss rate is increased ALLD remains a stable attractor, but cyclic attractors around an unstable fixed point emerge (top-right panel of figure 1)[1]. At even higher memory loss this second fixed point becomes a stable spiral. Provided players do not discount past play too strongly this spiral fixed point is located in the vicinity of the ALLC/TFT edge of the strategy simplex, and we conclude that moderate memory loss may enhance cooperative behaviour. When the memory becomes even shorter the fixed point moves towards the centre of the simplex. In the extreme case of full memory loss $\lambda = 1$ players ignore the past history beyond the last iteration entirely. Depending on the response sensitivity both players play essentially at random: the three strategies are used with very similar frequencies. It is interesting to note that the outcome

[1] We point out that it is hard to accurately determine the shape of cyclic attractors such as the one in the top-right panel of figure 1, even when integrating the dynamics up to large times of up to 5×10^6 and/or at small time stepping ($dt \approx 10^{-3}$). The cycle in figure 1 should therefore be understood as an illustration, rather than as a quantitative characterisation of the attractor.

J. Stat. Mech. (2011) P08007

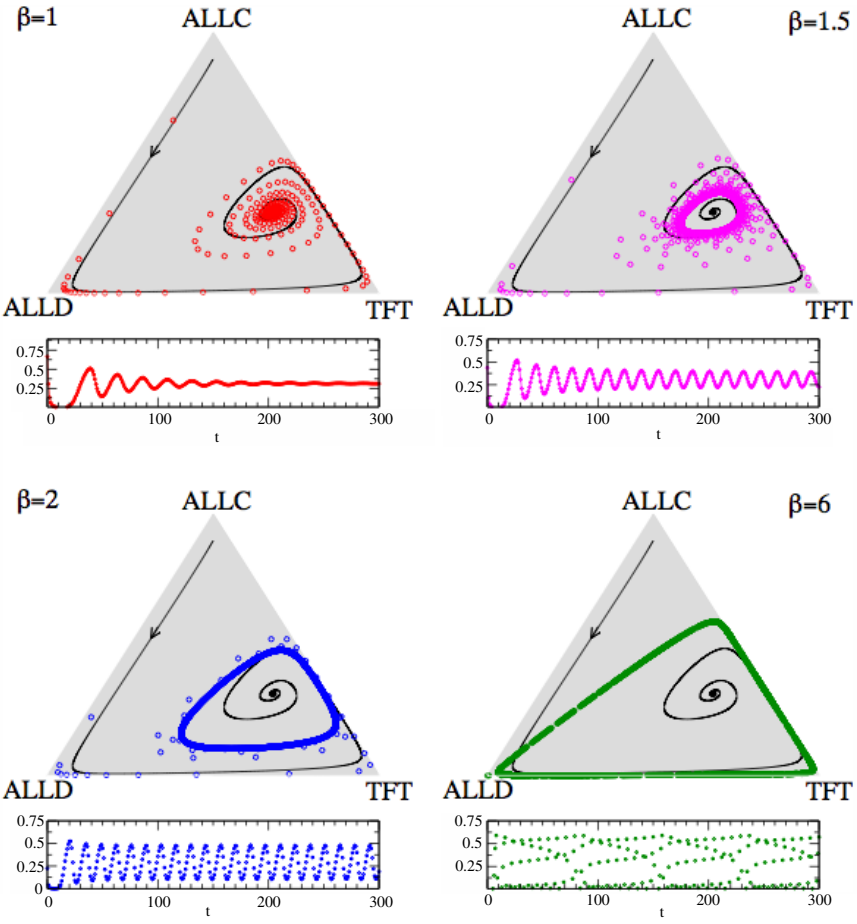

Figure 2. Comparison of discrete and continuous-time deterministic learning. We show trajectories of the dynamics at fixed $\lambda/\beta = 0.1$, started from homogeneous initial conditions, $\mathbf{x}(t=0) = \mathbf{y}(t=0)$. The black line in each simplex is obtained from $\beta = 0.01$ and represents the continuous-time limit. The symbols are for $\beta = 1$ (upper left panel), $\beta = 1.5$ (upper right), $\beta = 2$ (lower left) and $\beta = 6$ (lower right). In each panel we show the trajectory in the strategy simplex, as well as the corresponding time series of the propensity, $x_1(t)$, of playing ALLC.

of deterministic learning with memory loss resembles the behaviour of replicator–mutator dynamics of this game [38]. Discounting past experience in learning and mutation in evolution both promote cooperation when they are moderate in strength. The attractors of learning with quick memory loss, on the other hand, are similar to those of evolutionary systems in which mutation dominates selection.

3.2. Comparison of continuous-time and discrete-time deterministic dynamics

The modified replicator equations (6) are differential equations and, as such, describe a continuous-time learning process. This approximation is valid for $\beta \ll 1$. The behaviour of discrete-time deterministic learning can, however, be quite different from this continuous-time limit, as illustrated in figure 2. We here fix the ratio λ/β and consider the behaviour

J. Stat. Mech. (2011) P08007

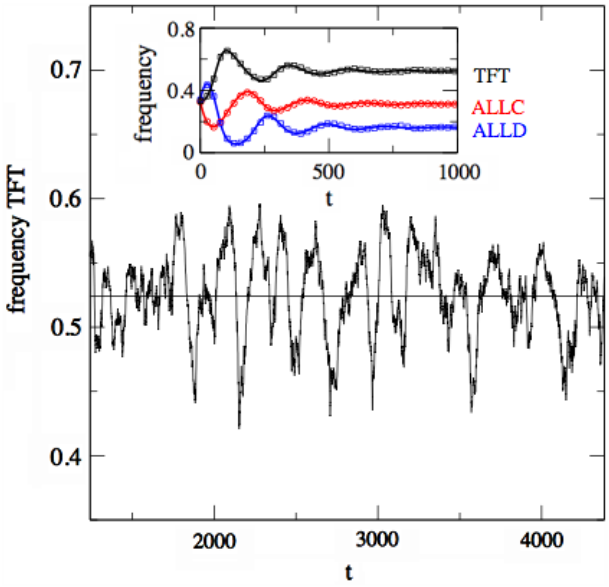

Figure 3. Sustained oscillations in the stochastic dynamics. Frequency with which TFT is played by Alice as a function of time at $N = 10$ observations between adaptation events. The horizontal line is the fixed point of deterministic learning. The inset shows the frequencies of ALLC, ALLD and TFT in the initial phase of the dynamics. Solid lines are the outcome of deterministic learning and symbols show data from an average over 100 independent runs of stochastic learning at $N = 10$. Model parameters are $\beta = 0.1, \lambda = 0.01$.

at different values of β. For small β the discrete-time maps behave essentially like the continuous dynamics, and have a stable spiral fixed point (for the value of λ chosen in the figure). As β is increased, however, this fixed point becomes unstable and a cyclic attractor develops[2]. Further increasing β enlarges the cycle, until its the attractor finally becomes a rather large triangular-shaped object as depicted in the lower-right panel of figure 2. It is here important to note that, even though the attractor set looks smooth, the dynamics does not revolve around the attractor in a continuous motion. We expect that more complicated behaviour, such as chaotic attractors, will in principle be possible, even though we have not observed them for the present game and the present learning dynamics. Other learning rules in similar games have, however, been shown to admit chaotic motion, see [40, 41].

3.3. Effects of noise

We will now move to learning at finite batch sizes N. Players are then no longer able to obtain a perfect sample of their opponents' mixed strategy profile before updating their own strategic propensities, and the dynamics becomes stochastic. Results of numerical simulations are shown in figure 3. We here focus on a regime in which deterministic learning approaches a fixed point. Stochastic learning at the same discounting rate and

[2] While the attractor of the dynamics appears to be a closed cyclic object, it is hard to determine numerically whether the trajectory is actually periodic, as we cannot exclude small drifts. The attractors plotted in the figure may therefore be invariant curves of the map, rather than actual cycles.

J. Stat. Mech. (2011) P08007

intensity of selection results in sustained cycles between cooperation and defection. The amplitude of these cycles is found to scale as $N^{-1/2}$ in the batch size, but the coefficient multiplying $N^{-1/2}$ can be substantial (see the appendix for analytical results) so that the oscillations can have a significant amplitude. The inset of figure 3 confirms that the average of several independent runs of the stochastic dynamics is accurately described by the deterministic update rules of equations (3).

The cycling behaviour of the stochastic learning process can be understood as the result of an amplification mechanism, which turns intrinsic white noise into coherent oscillations [43]. The intuitive picture is here as follows: at the memory-loss rate chosen in figure 3 the deterministic dynamics spirals into a stable fixed point asymptotically. The relevant eigenvalue of the dynamics is complex. If an instantaneous perturbation were applied to the deterministic system at the fixed point, the dynamics would return to the fixed point following a trajectory of damped oscillations. At finite batch sizes, however, the dynamics is subject to persistent random fluctuations, constantly driving the system away from the fixed point. The combination of this permanent 'excitation' and the oscillatory relaxation results in a coherently maintained cyclic pattern.

Similar noise-induced oscillation phenomena have been observed in various individual-based models of population dynamics, evolutionary game theory, epidemics and biochemical reactions, see, e.g., [35, 39], [43]–[49]. While the mechanism of resonant amplification in stochastic learning is analogous to the one observed in population-based models, the origin of the noise is different. In the individual-based models, large but finite populations are considered. Deterministic mean-field equations can then be derived in the limit of infinite populations. In finite populations, the dynamics remains stochastic, due to the random nature of the interactions on the microscopic level. The resulting noise scales with the inverse square root of the system size and has been termed 'demographic stochasticity' [43, 50]. In the learning model the source of the noise is the inaccuracy with which players sample their opponent's strategy profile at finite batch sizes, and the amplitude of the noise and of the resulting quasi-cycles is proportional to the inverse square root of the batch size N.

We have used a systematic expansion in the inverse batch size to characterise these cycles further. These methods are similar to system-size expansions widely used in population-based models [8], even though the expansion parameter in the learning dynamics is the inverse batch size, not the size of the population. This process has been applied to other games in [36]. Full details of the calculation can be found in the appendix: we here only sketch the main steps. In essence, for large but finite batch sizes N, the stochastic analogue of equations (4) can be approximated as

$$\begin{aligned} x_i(t+1) &= \frac{x_i(t)^{1-\lambda} \mathrm{e}^{\beta[\sum_j a_{ij} y_j(t) + N^{-1/2}\xi_i(t)]}}{\sum_k x_k(t)^{1-\lambda} \mathrm{e}^{\beta[\sum_j a_{kj} y_j(t) + N^{-1/2}\xi_k(t)]}}, \\ y_j(t+1) &= \frac{y_j(t)^{1-\lambda} \mathrm{e}^{\beta[\sum_i a_{ji} x_i(t) + N^{-1/2}\eta_j(t)]}}{\sum_k y_k(t)^{1-\lambda} \mathrm{e}^{\beta[\sum_j a_{kj} x_j(t) + N^{-1/2}\eta_k(t)]}}, \end{aligned} \tag{8}$$

where the noise variables $\xi_k(t)$ and $\eta_k(t)$ reflect the intrinsic stochasticity of the dynamics. In the limit of large N their covariance properties can be worked out analytically, as explained in detail in the appendix. Given these noise terms, the mixed strategy profiles $\{x_i(t), y_j(t)\}$ will be stochastic variables themselves. The next step is to self-consistently

J. Stat. Mech. (2011) P08007

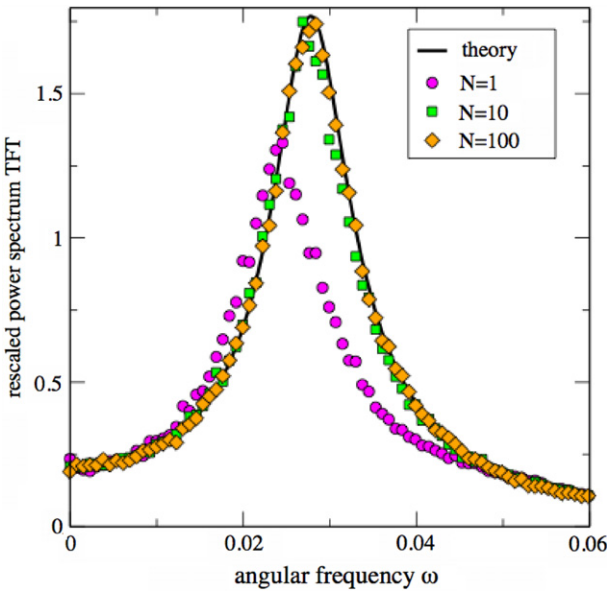

Figure 4. Power spectrum of the frequency with which TFT is played. Horizontal axis shows the angular frequency ω, while the vertical axis shows the spectrum of fluctuations about the deterministic fixed point. Results from numerical simulations of the stochastic dynamics are shown (markers) along with the curve predicted by the theory in the limit of large, but finite, batch size. Power spectra have been rescaled by the inverse batch size, see the appendix. Model parameters are $\beta = 0.1, \lambda = 0.01$. Simulations are averaged over 1000 runs.

separate deterministic from stochastic contributions, and to derive a closed set of equations describing the evolution of fluctuations about the deterministic limit. To this end we write

$$x_i(t) = \bar{x}_i(t) + \frac{1}{\sqrt{N}}\tilde{x}_i(t), \qquad y_j(t) = \bar{y}_j(t) + \frac{1}{\sqrt{N}}\tilde{y}_j(t), \tag{9}$$

where the quantities with overbars refer to the deterministic trajectory. This is then inserted into equation (8), and after a systematic expansion in powers of $N^{-1/2}$ one obtains a set of closed linear equations for the variables $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. It is then straightforward to compute, for example, correlation functions or power spectra of the fluctuations about the deterministic trajectory. Again we relegate the mathematical details to the appendix.

As seen in figure 4 the power spectrum of the coherent oscillations can be predicted analytically with great accuracy for moderate and high batch sizes N. The agreement for batch sizes of $N = 10$ is still reasonable; systematic deviations are only found if the number of observations between strategy updates is reduced further.

To characterise the outcome of the stochastic learning process in more detail we show the resulting stationary distributions in strategy space in figure 5. The panels in the upper row correspond to a memory-loss parameter for which the deterministic dynamics has a cyclic attractor. At small batch sizes stochastic learning essentially covers the entire strategy simplex, with the exception of the region near the ALLD/ALLC edge. Surprisingly, the most frequently visited points in strategy space are found along the ALLD/TFT edge; the Nash strategy ALLD is played only very rarely. At larger batch

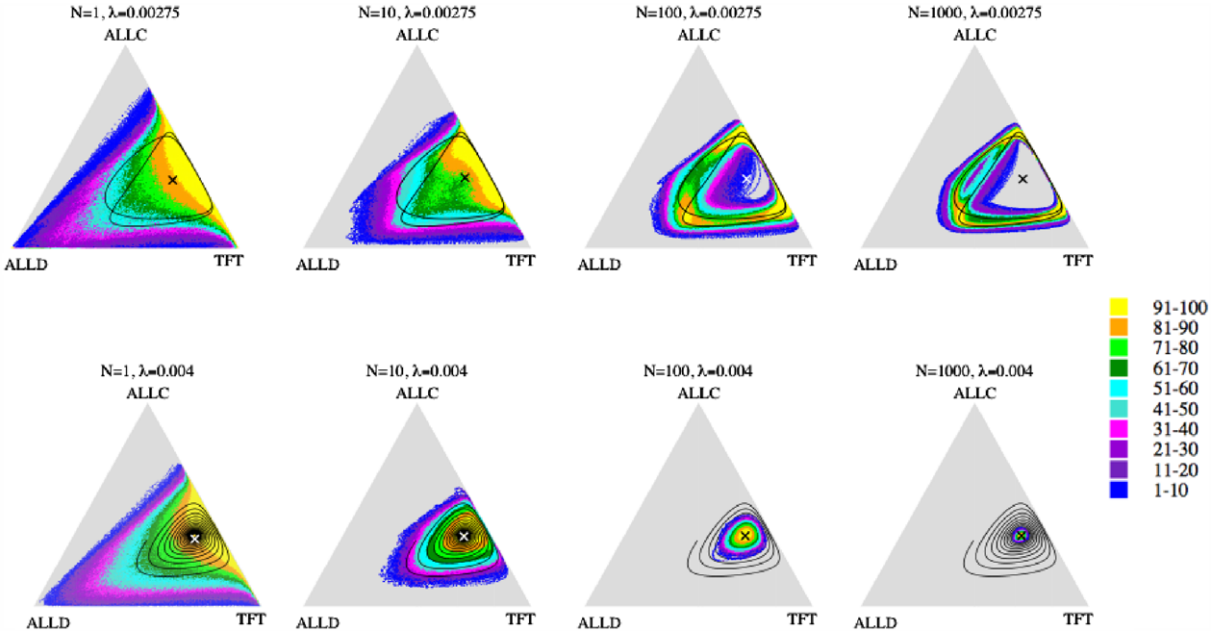

Figure 5. Frequencies of visits of the stochastic learning dynamics. Crosses mark the time average of the stochastic dynamics, while black lines are the trajectory or attractor of the deterministic discrete-time map. Data is obtained from 100 runs of the stochastic process at an intensity of selection $\beta = 0.1$. Colours indicate the frequency with which different regions are visited, a binning of strategy space is performed and, for example, yellow stands for the 10% most visited bins, orange for the next 10% and so on (see legend). Grey areas are not visited by the dynamics at all in our simulations.

sizes the dynamics concentrates in a region about the deterministic cycle. At fast memory loss (lower row of figure 5) deterministic learning has a fixed point. Again, the stochastic dynamics reaches almost the full strategy space for small batch sizes, but more and more concentration on the deterministic attractor is found as the frequency of adaptation is lowered (i.e. when the batch size is increased). In all cases shown in figure 5 the time average of learning is found near TFT; defection occurs only rarely. This is similar to what was observed in [38] in an evolutionary context. The precise mechanism for the concentration of probability near TFT is not fully clear at this point. We speculate that it can be traced back to a combination of the underlying deterministic dynamics (e.g. speed with which the system moves along the limit cycle) and the relative proximity of the deterministic trajectory to the edge of the simplex, effectively truncating fluctuations and 'reflecting' them back into the simplex, or concentrating them near the ALLC/TFT edge. Analytical results for the corresponding evolutionary case have been discussed in [38].

4. Summary

To summarise we have here analysed in detail the learning dynamics of two fixed players interacting in a repeated prisoner's dilemma game. We find that discounting past experience in a deterministic learning produces behaviour very similar to the dynamics

J. Stat. Mech. (2011) P08007

found in evolutionary replicator–mutator systems [38]. Memory loss removes the stability of the ALLD fixed point and leads to attractors near the ALLC/TFT edge of strategy space. In order to go beyond the adiabatic assumption underlying purely deterministic adaptation models, we have also addressed more realistic stochastic learning. Here, players update their strategic propensities more frequently, relying on an imperfect sampling of their opponent's strategy. We then observe persistent stochastic cycles, with a time average concentrated near TFT, paralleling earlier observations in finite-population evolutionary dynamics [38]. Based on a systematic expansion technique we have characterised these cycles analytically. This method is applicable very generally, and can be used to study the effects of stochasticity in other learning models [22, 26], in machine learning problems and in algorithmic game theory [51], or indeed to other games, including other variations of iterated prisoner's dilemma games. Work to apply these ideas to network games is in progress [52], see also [37, 55, 56] for existing work on the effects of memory in network games. Cyclic behaviour has been reported in experimental studies of multi-player learning [28]. We expect that the techniques we have introduced will be helpful in formulating and calibrating theoretical learning models describing these real-world laboratory experiments.

Acknowledgments

The author would like to thank J D Farmer and Y Sato for useful discussions. This work is supported by a Research Councils UK Fellowship (RCUK reference EP/E500048/1) and by The Engineering and Physical Sciences Research Council (grant EP/I005765/1).

Appendix. Analytical characterisation of stochastic cycles

In order to compute the spectrum of the oscillations between cooperation and defection in the limit of large, but finite, batch sizes N we start from the dynamics of equations (3):

$$\begin{aligned} A_k(t) &= (1-\lambda)A_k(t) + \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,i_\alpha(t)}, \\ B_k(t) &= (1-\lambda)B_k(t) + \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,j_\alpha(t)}, \end{aligned} \tag{A.1}$$

and note that the expression $(1/N)\sum_{\alpha=1}^{N} a_{k,i_\alpha(t)}$ on the right-hand side is a random variable at finite batch size N. The same is true for the analogous expression in the update rule for B_k. The mean value of $(1/N)\sum_{\alpha=1}^{N} a_{k,i_\alpha(t)}$ is given by $\mu_k(t) = \sum_j a_{kj}y_j(t)$, given that the $j_\alpha(t)$ are drawn from the mixed strategy profile $\mathbf{y}(t)$, i.e. action $\ell \in \{\mathrm{ALLC}, \mathrm{ALLD}, \mathrm{TFT}\}$ occurs with frequency $y_\ell(t)$ on average. Similarly $(1/N)\sum_{\alpha=1}^{N} a_{k,j_\alpha(t)}$ has an average of $\nu_k(t) = \sum_i a_{ki}x_i(t)$. Separating off fluctuations, and anticipating their scaling with N, we write

$$\begin{aligned} \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,i_\alpha(t)} &= \sum_j a_{kj}y_j(t) + \frac{1}{\sqrt{N}}\xi_k(t), \\ \frac{1}{N}\sum_{\alpha=1}^{N} a_{k,j_\alpha(t)} &= \sum_i a_{ki}x_i(t) + \frac{1}{\sqrt{N}}\eta_k(t). \end{aligned} \tag{A.2}$$

J. Stat. Mech. (2011) P08007

By means of the central limit theorem $\xi_k(t)$ and $\eta_k(t)$ can, in the limit of large but finite N, be approximated as Gaussian noise variables of mean zero and with the following correlations:

$$\begin{aligned}\langle \xi_k(t)\xi_\ell(t')\rangle &= \delta_{tt'}\sum_j \{y_j(t)[a_{kj}-\mu_k(t)][a_{\ell j}-\mu_\ell(t)]\},\\ \langle \eta_k(t)\eta_\ell(t')\rangle &= \delta_{tt'}\sum_i \{x_i(t)[a_{ki}-\nu_k(t)][a_{\ell i}-\nu_\ell(t)]\},\\ \langle \xi_k(t)\eta_\ell(t')\rangle &= 0.\end{aligned} \tag{A.3}$$

Here $\delta_{tt'}=1$ for $t=t'$ and $\delta_{tt'}=0$ otherwise. These expressions are obtained, for example, by writing

$$\xi_k(t)=\sqrt{N}\left[\frac{1}{N}\sum_{\alpha=1}^{N} a_{k,i_\alpha(t)}-\mu_k(t)\right], \tag{A.4}$$

followed by a straightforward evaluation of the above correlators to the appropriate order in $N^{-1/2}$, and taking into account the statistics of $i_\alpha(t)$.

We can now proceed to insert these expressions into the map (4) and find

$$\begin{aligned}x_i(t+1)&=\frac{x_i(t)^{1-\lambda}\mathrm{e}^{\beta[\sum_j a_{ij}y_j(t)+N^{-1/2}\xi_i(t)]}}{\sum_k x_k(t)^{1-\lambda}\mathrm{e}^{\beta[\sum_j a_{kj}y_j(t)+N^{-1/2}\xi_k(t)]}},\\ y_j(t+1)&=\frac{y_j(t)^{1-\lambda}\mathrm{e}^{\beta[\sum_i a_{ji}x_i(t)+N^{-1/2}\eta_j(t)]}}{\sum_k y_k(t)^{1-\lambda}\mathrm{e}^{\beta[\sum_j a_{kj}x_j(t)+N^{-1/2}\eta_k(t)]}}.\end{aligned} \tag{A.5}$$

Given the presence of the noise terms $\xi_k(t)$ and $\eta_k(t)$, the mixed strategy profiles $\{x_i(t),y_j(t)\}$ will be stochastic variables themselves. The next step is to self-consistently separate deterministic from stochastic contributions, and to derive a closed set of equations describing the evolution of fluctuations about the deterministic limit. To this end we write

$$x_i(t)=\bar{x}_i(t)+\frac{1}{\sqrt{N}}\tilde{x}_i(t),\qquad y_j(t)=\bar{y}_j(t)+\frac{1}{\sqrt{N}}\tilde{y}_j(t), \tag{A.6}$$

where the quantities with overbars represent the deterministic contributions and quantities with tildes are stochastic fluctuations. Equations (A.5) can be written in the form

$$x_i(t+1)=f_i(\mathbf{x}(t),\mathbf{y}(t),\boldsymbol{\xi}(t)),\qquad y_j(t+1)=g_j(\mathbf{x}(t),\mathbf{y}(t),\boldsymbol{\eta}(t)), \tag{A.7}$$

with suitable functions $\{f_i,g_j\}$. One proceeds by substituting (A.6) on both sides of equation (A.7), followed by a systematic expansion in powers of $N^{-1/2}$. To lowest order one finds

$$\bar{x}_i(t+1)=f_i(\bar{\mathbf{x}}(t),\bar{\mathbf{y}}(t),0),\qquad \bar{y}_j(t+1)=g_j(\bar{\mathbf{x}}(t),\bar{\mathbf{y}}(t),0), \tag{A.8}$$

i.e. one recovers the deterministic map (4).

While the calculation up to now applies to any deterministic trajectory, we will from now on restrict the discussion to an asymptotic regime and assume that the deterministic dynamics has reached a fixed point $\bar{\mathbf{z}}^*=(\bar{\mathbf{x}}^*,\bar{\mathbf{y}}^*)$. This is appropriate in the context of the present investigation, as we are interested in stochastic quasi-cycles about deterministic

J. Stat. Mech. (2011) P08007

fixed points. Based on the restriction to deterministic fixed points further analytical progress is relatively straightforward[3].

In next-to-leading order of the expansion in powers of $N^{-1/2}$ one has

$$\begin{aligned}\tilde{x}_i(t+1) &= \sum_k \left(\left.\frac{\partial f_i(\mathbf{x},\mathbf{y},\boldsymbol{\xi})}{\partial x_k}\right|_{(\mathbf{x}^*,\mathbf{y}^*,0)} \tilde{x}_k(t) + \left.\frac{\partial f_i(\mathbf{x},\mathbf{y},\boldsymbol{\xi})}{\partial y_k}\right|_{(\bar{\mathbf{x}}^*,\bar{\mathbf{y}}^*,0)} \tilde{y}_k(t) \right) + \kappa_i(t) \\ \tilde{y}_j(t+1) &= \sum_k \left(\left.\frac{\partial g_j(\mathbf{x},\mathbf{y},\boldsymbol{\xi})}{\partial x_k}\right|_{(\mathbf{x}^*,\mathbf{y}^*,0)} \tilde{x}_k(t) + \left.\frac{\partial g_j(\mathbf{x},\mathbf{y},\boldsymbol{\xi})}{\partial y_k}\right|_{(\bar{\mathbf{x}}^*,\bar{\mathbf{y}}^*,0)} \tilde{y}_k(t) \right) + \rho_j(t)\end{aligned} \tag{A.9}$$

where

$$\kappa_i(t) = \beta \left(\bar{x}_i^* \xi_i(t) - \bar{x}_i^* \sum_k \bar{x}_k^* \xi_k(t) \right), \qquad \rho_j(t) = \beta \left(\bar{y}_j^* \eta_j(t) - \bar{y}_j^* \sum_k \bar{y}_k^* \eta_k(t) \right). \tag{A.10}$$

Writing $\mathbf{z} \equiv (z_1, \ldots, z_6) = (x_1, x_2, x_3, y_1, y_2, y_3)$, and using the notation $\mathbf{z}(t) = \bar{\mathbf{z}}^* + N^{-1/2}\boldsymbol{\zeta}(t)$ to separate deterministic from stochastic contributions ($\boldsymbol{\zeta} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$) one has

$$\boldsymbol{\zeta}(t+1) = \mathbb{J}^* \boldsymbol{\zeta}(t) + \boldsymbol{\phi}(t), \tag{A.11}$$

where $\mathbb{J}^*$ is the 6×6 Jacobian matrix of the deterministic equations (4), evaluated at the fixed point $\bar{\mathbf{z}}^* = (\bar{\mathbf{x}}^*, \bar{\mathbf{y}}^*)$. The variable $\boldsymbol{\phi} = (\varphi_1, \ldots, \varphi_6) = (\kappa_1, \kappa_2, \kappa_3, \rho_1, \rho_2, \rho_3)$ represents Gaussian noise, uncorrelated in time, but with cross-correlations between the different components:

$$\langle \varphi_a(t) \varphi_b(t') \rangle = \delta_{tt'} D_{ab}^*. \tag{A.12}$$

The elements of $\mathbb{D}^*$ can be expressed in terms of the deterministic variables $\bar{\mathbf{z}}$. More precisely one has, using equations (A.10),

$$D_{ij} = \beta^2 \Big[\bar{x}_i^* \bar{x}_j^* \langle \xi_i \xi_j \rangle - \bar{x}_i^* \bar{x}_j^* \sum_{k=1}^3 \bar{x}_k^* \langle \xi_i \xi_k \rangle - \bar{x}_j^* \bar{x}_i^* \sum_{k=1}^3 \bar{x}_k^* \langle \xi_j \xi_k \rangle + \bar{x}_i^* \bar{x}_j^* \sum_{k=1}^3 \sum_{\ell=1}^3 \bar{x}_k^* \bar{x}_\ell^* \langle \xi_k \xi_\ell \rangle \Big] \tag{A.13}$$

and

$$\begin{aligned} D_{i+3,j+3} = \beta^2 \Big[& \bar{y}_i^* \bar{y}_j^* \langle \eta_i \eta_j \rangle - \bar{y}_i^* \bar{y}_j^* \sum_{k=1}^3 \bar{y}_k^* \langle \eta_i \eta_k \rangle - \bar{y}_j^* \bar{y}_i^* \sum_{k=1}^3 \bar{y}_k^* \langle \eta_j \eta_k \rangle \\ & + \bar{y}_i^* \bar{y}_j^* \sum_{k=1}^3 \sum_{\ell=1}^3 \bar{y}_k^* \bar{y}_\ell^* \langle \eta_k \eta_\ell \rangle \Big] \end{aligned} \tag{A.14}$$

for $i, j \in \{1, 2, 3\}$. The noise variables φ_a with $a \in \{1, 2, 3\}$ are uncorrelated from those with $a \in \{4, 5, 6\}$ so that the matrix $\mathbb{D}^*$ is block diagonal (D_{ab}^* and D_{ba}^* both vanish if

[3] A full analytical characterisation of stochastic effects is possible also for periodic attractors of the deterministic dynamics. This has been discussed in the context of chemical reaction systems in [53] and [54]. Such approaches are based on Floquet theory and we expected that they are applicable also in the learning scenario (with suitable modifications to accommodate the discrete-time dynamics). This is beyond the scope of the work presented in this paper. We point out, however, that all equations up to (A.15) are valid for any deterministic trajectory, provided the fixed point values $\bar{\mathbf{z}}^*$ in the relevant expressions are replaced by their time-dependent counterparts.

J. Stat. Mech. (2011) P08007

$a \in \{1,2,3\}$ and $b \in \{4,5,6\}$). The covariances of the noise variables $\{\xi_k\}$ and $\{\eta_k\}$ are given by (A.3). One further potentially subtle point deserves some attention here. The covariance elements of the noise variables $\{\xi_k\}$ and $\{\eta_k\}$ as given in (A.3) depend on the variables $\mathbf{z}(t) = (x_1(t), x_2(t), x(t), y_1(t), y_2(t), y_3(t))$. These in turn have deterministic and stochastic contributions, $\mathbf{z} = \bar{\mathbf{z}}^* + N^{-1/2}\boldsymbol{\zeta}$. Within our expansion in powers of $N^{-1/2}$ it is justified to self-consistently suppress the stochastic contributions $N^{-1/2}\boldsymbol{\zeta}$ to the variables $\mathbf{z}$ in equation (A.3), as these contributions would not affect results of the order of $N^{-1/2}$ we are working on. For the purposes of equations (A.13) and (A.14) we therefore use

$$\begin{aligned}\langle \xi_k(t)\xi_\ell(t')\rangle &= \delta_{tt'}\sum_j \{\bar{y}_j^*[a_{kj} - \bar{\mu}_k^*][a_{\ell j} - \bar{\mu}_\ell^*]\}\\ \langle \eta_k(t)\eta_\ell(t')\rangle &= \delta_{tt'}\sum_i \{\bar{x}_i^*[a_{ki} - \bar{\nu}_k^*][a_{\ell i} - \bar{\nu}_\ell^*]\},\end{aligned} \tag{A.15}$$

where $\bar{\mu}_i^* = \sum_j a_{ij}\bar{y}_j^*$ and $\bar{\nu}_j^* = \sum_i a_{ji}\bar{y}_i^*$.

Starting from the linear equation (A.11) we now move to Fourier space and write $\hat{\zeta}_a(\omega)$ for the Fourier transform of $\zeta_a(t)$ and similarly for the noise components φ_a ($a = 1, \ldots, 6$). One then has

$$\hat{\boldsymbol{\zeta}}(\omega) = \mathbb{M}^{-1}\hat{\boldsymbol{\phi}}(\omega), \tag{A.16}$$

where $\mathbb{M} = \mathrm{e}^{\mathrm{i}\omega}\mathbb{I} - \mathbb{J}^*$. The notation $\mathbb{I}$ here indicates the 6×6 identity matrix. The power spectra of the components of $\boldsymbol{\zeta}$ can then be obtained from equation (A.16), taking into account (A.12), i.e. the fact that $\langle \hat{\varphi}_a(\omega)\hat{\varphi}_b(\omega')\rangle = \delta(\omega + \omega')\mathbb{D}^*_{ab}$. One then has

$$P_{aa}(\omega) = \left\langle |\hat{\zeta}_a(\omega)|^2 \right\rangle = \sum_{bc} (\mathbb{M}^{-1})_{ab}\mathbb{D}^*_{bc}(\mathbb{M}^{\dagger -1})_{ca}. \tag{A.17}$$

The right-hand side can be evaluated numerically using the explicit form of the Jacobian $\mathbb{J}^*$ and of the noise covariance matrix $\mathbb{D}^*$. These quantities only depend on the fixed point $\bar{\mathbf{z}}^*$ of the deterministic dynamics, which again can be obtained by numerical iteration of the map (4), or as a numerical solution of the corresponding fixed point relations.

References

[1] von Neumann J and and Morgenstern O, 1944 *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press)
[2] Nash J, 1950 *Proc. Nat. Acad. Sci.* **36** 48
[3] Nash J, 1951 *Ann. Math.* **54** 286
[4] Price G and Maynard Smith J, 1973 *Nature* **246** 15
[5] Maynard Smith J, 1998 *Evolution and The Theory of Games* (Cambridge: Cambridge University Press)
[6] Traulsen A and Hauert C, *Stochastic evolutionary game dynamics*, 2009 *Reviews of Nonlinear Dynamics and Complexity* vol 2, ed H-G Schuster (New York: Wiley-VCH)
[7] Szabo G and Fath G, 2007 *Phys. Rep.* **446** 97
[8] van Kampen N G, 1992 *Stochastic Processes in Physics and Chemistry* (New York: Elsevier)
[9] Reichenbach T, Mobilia M and Frey E, 2007 *Nature* **448** 1046
[10] Nowak M A, Komarova N L and Niyogi P, 2001 *Science* **291** 114
[11] Hens T and Schenk-Hoppe K, 2005 *J. Math. Econ.* **41** 43
[12] Myerson R B, 1991 *Game Theory: Analysis of Conflict* (Cambridge, MA: Harvard University Press)
[13] Gintis H, 2000 *Game Theory Evolving* 1st edn (Princeton, NJ: Princeton University Press)
[14] Hofbauer J and Sigmund K, 1998 *Evolutionary Games and Population Dynamics* (Cambridge: Cambridge University Press)
[15] Vega-Redondo F, 2003 *Economics and The Theory of Games* (Cambridge: Cambridge University Press)

doi:10.1088/1742-5468/2011/08/P08007

J. Stat. Mech. (2011) P08007

[16] Nowak M A, 2006 *Evolutionary Dynamics* (Cambridge, MA: Harvard University Press)
[17] Sigmund K, 2010 *The Calculus of Selfishness* (Princeton, NJ: Princeton University Press)
[18] Axelrod R and Hamilton W D, *The evolution of cooperation*, 1981 *Science* **211** 1390
[19] Axelrod R, 2006 *The Evolution of Cooperation* Revised edn (New York: Perseus Books) ISBN 0-465-00564-0
[20] Nowak M A, 2006 *Science* **314** 1560
[21] Pennisi E, 2005 *Science* **309** 93
[22] Fudenberg D and Levine D K, 1989 *The Theory of Learning in Games* (Cambridge, MA: MIT Press)
[23] Lugosi G and Cesa-Bianchi N, 2006 *Prediction, Learning, and Games* (New York: Cambridge University Press)
[24] Young H P, 2004 *Strategic Learning and its Limits* (Oxford: Oxford University Press)
[25] Camerer C F and Ho T H, 1999 *Econometrica* **67** 827
[26] Camerer C F, 2003 *Behavioural Game Theory—Experiments in Strategic Interaction* (Princeton, NJ: Princeton University Press)
[27] Chong J K, Ho T H and Camerer C F, 2007 *J. Econ. Theory* **133** 177
[28] Semmann D, Krambeck H J and Milinski M, 2003 *Nature* **425** 390
[29] Traulsen A, Semmann D, Sommerfeld R D, Krambeck H J and Milinski M, 2010 *Proc. Nat. Acad. Sci.* **107** 2962
[30] Henrich J *et al* (ed), 2004 *Foundations of Human Sociality* (Oxford: Oxford University Press)
[31] Sato Y, Akiyama E and Farmer J D, 2002 *Proc. Nat. Acad. Sci.* **99** 4748
[32] Sato Y and Crutchfied J P, 2003 *Phys. Rev.* E **67** 015206(R)
[33] Sato Y, Akiyama E and Crutchfield J P, 2005 *Physica* D **210** 21
[34] Sutton R S and Barto A G, 1988 *Reinforcement Learning—An Introduction* (Cambridge, MA: MIT Press)
[35] Traulsen A, Claussen J C and Hauert C, 2005 *Phys. Rev. Lett.* **95** 238701
[36] Galla T, 2009 *Phys. Rev. Lett.* **103** 198702
[37] Realpe Gomez J, Szczesny B, Dall'Asta L and Galla T, 2011 preprint http://arxiv.org/abs/1102.0876
[38] Imhof L A, Fudenberg D and Nowak M A, 2005 *Proc. Nat. Acad. Sci.* **102** 10797
[39] Bladon A J, Galla T and McKane A J, 2010 *Phys. Rev.* E **81** 066122
[40] Ochea M, 2010 *PhD Thesis* University of Amsterdam, Academic Publishing Services, ISBN 978 90 5170 688 8
[41] Vilone D, Robledo A and Sanchez A, 2011 preprint http://arxiv.org/abs/1103.1484
[42] Saad D (ed), 1998 *On-line Learning in Neural Networks* (Cambridge: Cambridge University Press)
[43] McKane A J and Newman T J, 2005 *Phys. Rev. Lett.* **94** 218102
[44] Alonso D, McKane A J and Pascual M, 2007 *J. R. Soc. Interface* **4** 575
[45] Kuske R, Gordillo L F and Greenwood P, *J. Theor. Biol.* **245** 459
[46] Reichenbach T, Mobilia M and Frey E, 2006 *Phys. Rev.* E **74** 051907
[47] Mobilia M, 2010 *J. Theor. Biol.* **264** 1
[48] Simoes M, Telo da Gama M M and Nunes A, 2008 *J. R. Soc. Interface* **5** 555
[49] Pineda-Krch M, Blok H J, Dieckmann U and Doebeli M, 2007 *Oikos* **116** 53
[50] Nisbet R and Gurney W, 1982 *Modelling Fluctuating Populations* (New York: Wiley)
[51] Nisan N, Roughgarden T, Tardos E and Vazirani V V, 2007 *Algorithmic Game Theory* (Cambridge: Cambridge University Press)
[52] Bladon A and Galla T, 2011 preprint http://arxiv.org/abs/1107.0878
[53] Boland R P, Galla T and McKane A J, 2008 *J. Stat. Mech.* P09001
[54] Boland R P, Galla T and McKane A J, 2009 *Phys. Rev.* E **79** 051131
[55] Wang W-X, Ren J, Chen G and Wang B-H, 2006 *Phys. Rev.* E **74** 056113
[56] Qin S-M, Chen Y, Zhao X-Y and Shi J, 2009 *Phys. Rev.* E **78** 041129

J. Stat. Mech. (2011) P08007