# Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems

Yuzuru Sato
James P. Crutchfield

SFI WORKING PAPER: 2002-04-017

SANTA FE INSTITUTE

# Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems

Yuzuru Sato[1, 2, *] and James P. Crutchfield[2, †]

[1] *Brain Science Institute, The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Saitama 351-0198, Japan*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

(Dated: April 24, 2002)

Starting with a group of reinforcement-learning agents we derive coupled replicator equations that describe the dynamics of collective learning in multiagent systems. We show that, although agents model their environment in a self-interested way without sharing knowledge, a game dynamics emerges naturally through the environment. As an application, with a rock-scissors-paper game interaction between agents, the collective learning dynamics exhibits a diversity of competitive and cooperative behaviors. These include quasiperiodicity, stable limit cycles, intermittency, and deterministic chaos—behaviors that are to be expected in the multiagent, heterogeneous setting described by the general replicator equations.

PACS numbers: 05.45.-a, 02.50.Le, 87.23.-n

Adaptive behavior in multiagent systems is an important interdisciplinary topic that appears in various guises in many fields, including biology [1], computer science [2], economics [3], and cognitive science [4]. One of the key common questions is how and whether a group of intelligent agents truly engages in collective behaviors that are more functional than individuals acting alone.

Suppose that many agents interact with an environment and each independently builds a model from its sensory stimuli. In this simple type of coupled multiagent system, collective learning (if it occurs) is a dynamical behavior driven by agents' environment-mediated interaction [5, 6]. Here we show that the collective dynamics in multiagent systems, in which agents use reinforcement learning [7], can be modeled using coupled replicator equations. While replicator dynamics were given originally in terms of evolutionary game theory [8], recently the relationship between reinforcement learning and replicator equations has been discussed [9]. Here, we show that game dynamics is introduced as a continuous-time limit in a multiagent reinforcement learning system.

Notably, in learning with memory, our model reduces to the form of a multipopulation replicator equation [10]. With memory loss, however, the dynamics become dissipative. As an application, we note that the dynamics of learning with memory in the rock-scissors-paper game exhibits Hamiltonian chaos, if it is a zero-sum interaction between two agents. With memory decay, the multiagent system becomes dissipative and displays the full range of nonlinear dynamical behaviors, including limit cycles, intermittency, and deterministic chaos.

Our multiagent model begins with standard reinforcement learning agents [7]. For simplicity, here we assume that there are two such agents $X$ and $Y$ and that at each time step each agent takes one of $N$ actions: $i = 1, 2, \ldots, N$. Let the probability for $X$ to chose action $i$ be $x_i(n)$ and $y_i(n)$ for $Y$, where $n$ is the number of the learning iterations from the initial state $x_i(0)$ and $y_i(0)$. The agents' state vectors at time $n$ are $\mathbf{x}(n) = (x_1(n), x_2(n), \ldots, x_N(n))$ and $\mathbf{y}(n) = (y_1(n), y_2(n), \ldots, y_N(n))$, where $\Sigma_i x_i(n) = \Sigma_i y_i(n) = 1$. Let $R_i^X(n)$ and $R_i^Y(n)$ denote the reward for $X$ or $Y$ taking action $i$ at step $n$, respectively. Then $X$'s and $Y$'s memories—denoted $Q_i^X(n)$ and $Q_i^Y(n)$, resp.—of the past benefits for action $i$ are governed by

$$\begin{aligned} \Delta Q_i^X(n+1) &= R_i^X(n) - \alpha_X Q_i^X(n) \text{ and} \\ \Delta Q_i^Y(n+1) &= R_i^Y(n) - \alpha_Y Q_i^Y(n) \ , \end{aligned} \tag{1}$$

where $\alpha_x, \alpha_y \in [0, 1)$ control each agent's memory decay rate.

The agents chose their next actions based on their memory and update their choice distributions—i.e., $\mathbf{x}$ and $\mathbf{y}$—as follows:

$$x_i(n) = \frac{e^{\beta_X Q_i^X(n)}}{\Sigma_j e^{\beta_X Q_j^X(n)}} \text{ and } y_i(n) = \frac{e^{\beta_Y Q_i^Y(n)}}{\Sigma_j e^{\beta_Y Q_j^Y(n)}}, \tag{2}$$

where $\beta_x, \beta_y \in [0, \infty]$ control the learning sensitivity: how much the current choice distributions are affected by past rewards. Using Eq. (2), the dynamic governing the change in state is given by:

$$x_i(n+1) = \frac{x_i(n) e^{\beta_X \Delta Q_i^X(n)}}{\Sigma_k x_j(n) e^{\beta_X \Delta Q_k^X(n)}} \ , \tag{3}$$

where $\Delta Q_i^X(n) = Q_i^X(n+1) - Q_i^X(n)$ and similarly for $y_i(n+1)$.

Next, we consider the continuous-time limit of this system, which corresponds to the agents performing a large number of learning updates—iterations of Eqs. (3)—for each memory update—iteration of Eqs. (2). Thus, in the continuous-time limit $X$ behaves as if it knows $\mathbf{y}$ (the distribution of Y's choices) and $Y$ behaves similarly. Going from time step $n\delta$ to $n\delta + \delta$ the continuous-learning rule for agent $X$ is

$$x_i(n\delta+\delta) \; - \; x_i(n\delta) = \frac{x_i(n\delta)}{\Sigma_j x_j(n\delta) e^{\beta_X (Q_j^X(n\delta+\delta)-Q_j^X(n\delta))}} \times \Big[ e^{\beta_X (Q_i^X(n\delta+\delta)-Q_i^X(n\delta))} - \Sigma_j x_j(n\delta) e^{\beta_X (Q_j^X(n\delta+\delta)-Q_j^X(n\delta))} \Big] , \quad (4)$$

based on Eq. (3). In the limit $\delta \to 0$ with $t = n\delta$, Eq. (4) reduces to

$$\dot{x}_i = \beta_X x_i (\dot{Q}_i^X - \Sigma_j \dot{Q}_j^X x_j). \quad (5)$$

For the dynamic governing memory updates, we have

$$\dot{Q}_i^X = R_i^X - \alpha_X Q_i^X. \quad (6)$$

Putting together Eqs. (2), (5), and (6) one obtains

$$\frac{\dot{x}_i}{x_i} = \beta_X [R_i^X - \Sigma_j x_{ij} R_j^X] + \alpha_X I(x_i) \ , \quad (7)$$

where $I(x_i) = \Sigma_j x_j \log(x_j/x_i)$. The continuous-time dynamics of $Y$ follows in a similar manner.

Simplifying again, consider a fixed linear relationship between rewards and actions:

$$R_i^X = \Sigma_j a_{ij} y_j \text{ and } R_i^Y = \Sigma_j b_{ij} x_j \ . \quad (8)$$

In this special case, the continuous dynamics is given by:

$$\begin{aligned} \frac{\dot{x}_i}{x_i} &= \beta_X [(A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}] + \alpha_X I(x_i) \ , \\ \frac{\dot{y}_i}{y_i} &= \beta_Y [(B\mathbf{x})_i - \mathbf{y} \cdot B\mathbf{x}] + \alpha_Y I(y_i) \ , \end{aligned} \quad (9)$$

where $(A)_{ij} = a_{ij}$ and $(B)_{ij} = b_{ij}$; $(A\mathbf{x})_i$ is the $i$th element of the vector $A\mathbf{x}$; and $I(x_i)$ and $I(y_i)$ represent the effect of memory with decay parameters $\alpha_X$ and $\alpha_Y$. $\beta_X$ and $\beta_Y$ control the time-scale of each agent's learning. We can regard $A$ and $B$ as $X$'s and $Y$'s game-theoretic payoff matrices for action $i$ against opponent's action $j$ [18]. Note that the development of our model begins with selfish-learning agents with no knowledge of a "game" in which they are playing. Nonetheless, a game dynamics emerges—via $R^X$ and $R^Y$ in Eq. (7)—as a description of the collective's global behavior. That is, the agents' mutual adaptation induces a game at the collective level.

Given the basic equations of motion for the reinforcement-learning multiagent system (Eq. (9)), one becomes interested in, on the one hand, the time evolution of each agent's state vector in the simplices $\mathbf{x} \in \Delta_x$ and $\mathbf{y} \in \Delta_y$ and, on the other, the dynamics in the higher-dimensional simplex $(\mathbf{x}, \mathbf{y}) \in \Delta_x \times \Delta_y$ of the collective. Transforming from $(\mathbf{x}, \mathbf{y}) \in \Delta_x \times \Delta_y$ to $\mathbf{U} = (\mathbf{u}, \mathbf{v}) \in \mathbf{R}^{2(N-1)}$ with $\mathbf{u} = (u_1, u_2, \ldots, u_{N-1})$ and $\mathbf{v} = (v_1, v_2, \ldots, v_{N-1})$ where $u_i = \log \frac{x_{i+1}}{x_1}$ and $v_i = \log \frac{y_{i+1}}{y_1}$, $(i = 1, 2, \ldots, N-1)$ , we have a simplified version of our model (Eqs. (9))

$$\begin{aligned} \dot{u}_i &= \beta_X \frac{\Sigma_j \tilde{a}_{ij} e^{v_j} + \tilde{a}_{i1}}{1 + \Sigma_j e^{v_j}} - \alpha_X u_i \text{ and} \\ \dot{v}_i &= \beta_Y \frac{\Sigma_j \tilde{b}_{ij} e^{u_j} + \tilde{b}_{i1}}{1 + \Sigma_j e^{u_j}} - \alpha_Y v_i \ , \end{aligned} \quad (10)$$

where $\tilde{a}_{ij} = a_{i+1,j} - a_{1,j}$ and $\tilde{b}_{ij} = b_{i+1,j} - b_{1,j}$ [11]. Since the dissipation rate $\gamma$ in $\mathbf{U}$ is

$$\gamma = \Sigma_i \frac{\partial \dot{u}_i}{\partial u_i} + \Sigma_j \frac{\partial \dot{v}_j}{\partial v_j} = -(N-1)(\alpha_X + \alpha_Y), \quad (11)$$

Eqs. (9) are conservative when $\alpha_X = \alpha_Y = 0$ and the time average of a trajectory is the Nash equilibrium of the game specified by $(A, B)$, if a limit set exists in the interior of simplex [19]. Moreover, if the game is zero-sum, the dynamics are Hamiltonian in $\mathbf{U}$ with

$$H = -(\Sigma_j x_j^* u_j + \Sigma_j y_j^* v_j) + \log(1 + \Sigma_j e^{u_j}) + \log(1 + \Sigma_j e^{v_j}) \quad (12)$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ is an interior Nash equilibrium [11].

To illustrate the dynamics of learning in multiagent systems using the above developments, we now analyze the behavior of the two-person rock-scissors-paper interaction. This familiar game describes a three-sided competition: rock beats scissors, scissors beats paper, and paper beats rock. The payoff matrices are:

$$A = \begin{bmatrix} \epsilon_X & 1 & -1 \\ -1 & \epsilon_X & 1 \\ 1 & -1 & \epsilon_X \end{bmatrix} \text{ and } B = \begin{bmatrix} \epsilon_Y & 1 & -1 \\ -1 & \epsilon_Y & 1 \\ 1 & -1 & \epsilon_Y \end{bmatrix} \ , \quad (13)$$

where $\epsilon_X, \epsilon_Y \in [-1.0, 1.0]$ are the payoffs for ties. The mixed Nash equilibrium is $x_i^* = y_i^* = 1/3, (i = 1, 2, 3)$—the center of the simplices. Note that if $\epsilon_X = -\epsilon_Y$, the game is zero-sum.

In the special case without memory loss ($\alpha_X = \alpha_Y = 0$) and with large and equal learning sensitivity ($\beta_X = \beta_Y = 1$), the linear version (Eqs. (9)) of our model (Eqs. (7)) reduces to a multipopulation replicator equation [10]:

$$\frac{\dot{x}_i}{x_i} = [(A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}] \text{ and } \frac{\dot{y}_i}{y_i} = [(B\mathbf{x})_i - \mathbf{y} \cdot B\mathbf{x}] \ . \quad (14)$$

The game-theoretic behavior in the case with rock-scissors-paper interactions (Eqs. (13)) was investigated in [13]. In the zero-sum case ($\epsilon_X = -\epsilon_Y$), it was noted there that a Hamiltonian form of the equations of motion exists. Here, by way of contrast to our more general setting, we briefly recall the behavior in these special cases, noting several additional results.

Figure 1 shows Poincaré sections of Eqs. (14)'s trajectories on the hyperplane $\dot{u}_1 = 0$, $\dot{v}_1 > 0$ and representative trajectories in the individual agent simplices $\Delta_X$ and $\Delta_Y$. When $\epsilon_X = -\epsilon_Y = 0.0$, we expect the system

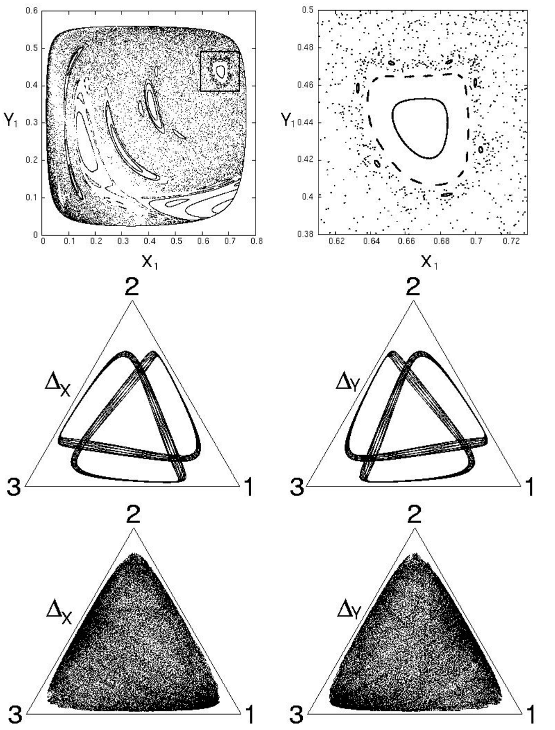

FIG. 1: Quasiperiodic tori and chaos: $\epsilon_X = -\epsilon_Y = 0.5$, $\alpha_X = \alpha_Y = 0$, and $\beta_X = \beta_Y = 1$. We give a Poincaré section (top) on the hyperplane defined by $\dot{u}_1 = 0$ and $\dot{v}_1 > 0$; that is, in the $(\mathbf{x}, \mathbf{y})$ space: $(3 + \epsilon_X)y_1 + (3 - \epsilon_X)y_2 - 2 = 0$ and $(3 + \epsilon_Y)x_1 + (3 - \epsilon_Y)x_2 - 2 < 0$. There are 23 randomly selected initial conditions with energies $H = -1/3(u_1 + u_2 + v_1 + v_2) + \log(1 + e^{u_1} + e^{u_2}) + \log(1 + e^{v_1} + e^{v_2}) = 2.941693$, which surface forms the outer border of $H \leq 2.941693$. Two rows (bottom): Representative trajectories, simulated with a 4th-order symplectic integrator [12], starting from initial conditions within the Poincaré section. The upper simplices show a torus in the section's upper right corner; see the enlarged section at the upper right. The initial condition is $(\mathbf{x}, \mathbf{y}) = (0.3, 0.054196, 0.645804, 0.1, 0.2, 0.7)$. The lower simplices are an example of a chaotic trajectory passing through the regions in the section that are a scatter of dots; the initial condition is $(\mathbf{x}, \mathbf{y}) = (0.05, 0.35, 0.6, 0.1, 0.2, 0.7)$.

to be integrable and only quasiperiodic tori would exist. Otherwise, $\epsilon_X = -\epsilon_Y > 0.0$, Hamiltonian chaos can occur with positive-negative pairwise Lyapunov exponents [13]. The dynamics is very rich, there are infinitely many distinct behaviors near the unstable fixed point at the center—the classical Nash equilibrium—and a periodic orbit arbitrarily close to any chaotic one. Moreover, when the game is not zero-sum ($\epsilon_X \neq \epsilon_Y$), transients to heteroclinic cycles are observed [13]: On the one hand, there are intermittent behaviors in which the time spent near pure strategies (the simplicial vertices) linearly increases with $\epsilon_X + \epsilon_Y < 0$ and, on the other hand, $\epsilon_X + \epsilon_Y > 0$, for which chaotic transients persist [14].

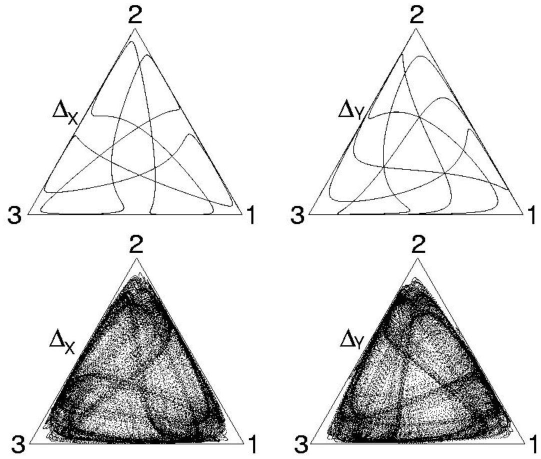

FIG. 2: Limit cycle (top: $\epsilon_Y = 0.025$) and chaotic attractors (bottom: $\epsilon_Y = -0.365$), with $\epsilon_X = 0.5$, $\alpha_X = \alpha_y = 0.01$, and $\beta_X = \beta_Y = 1.0$.

Our model goes beyond these special cases and, generally, beyond the standard multipopulation replicator equations (Eqs. (14)) due to its accounting for the effects of individual and collective learning. For example, if the memory decay rates ($\alpha_X$ and $\alpha_Y$) are positive, the system becomes dissipative and exhibits limit cycles and chaotic attractors; see Fig. 2. Figure 3 (top) shows a diverse range of bifurcations as a function of $\epsilon_Y$: dynamics on the hyperplane ($\dot{u}_1 = 0$, $\dot{v}_1 > 0$) projected onto $y_1$. When the game is nearly zero-sum, agents can reach the stable Nash equilibrium, but chaos can also occur, when $\epsilon_X + \epsilon_Y > 0$. Figure 3 (bottom) shows that the largest Lyapunov exponent is positive across a significant fraction of parameter space; indicating that chaos is common. The dual aspects of chaos, irregularity and coherence, imply that agents may behave cooperatively or competitively (or dynamically switch between both) in the collective dynamics. As noted above, this derives directly from individual self-interested learning.

Within this framework a number of extensions suggest themselves as ways to investigate the emergence of collective behaviors in multiagent systems. The most obvious is the generalization to an arbitrary number of agents with an arbitrary number of strategies and the analysis of behaviors in thermodynamic limit. It is relatively straightforward to develop an extension to the linear-reward version (Eqs. (9)) of our model. For example, in the case of three agents $X$, $Y$, and $Z$, one obtains for the learning dynamics in $\Delta_X \times \Delta_Y \times \Delta_Z$:

$$\frac{\dot{x}_i}{x_i} = \beta_X [\Sigma_{j,k} a_{ijk} y_j z_k - \Sigma_{j,k,l} a_{jkl} x_j y_k z_l] - \alpha_X I(x_i) \ , \qquad (15)$$

with tensor $(A)_{ijk} = a_{ijk}$, and similarly for $Y$ and $Z$. Not surprisingly, this is also a conservative system when $\alpha_X = \alpha_Y = \alpha_Z = 0$. However, the extension to multiple agents for the full nonlinear collective learning equations

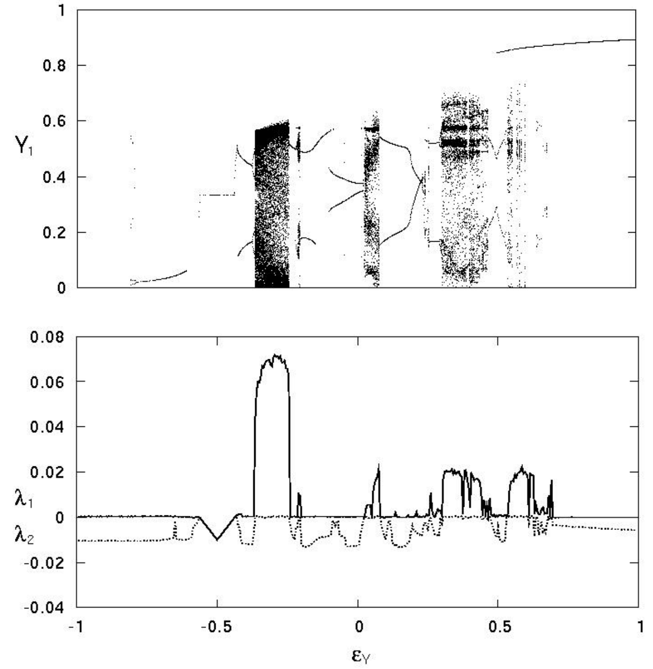

FIG. 3: Bifurcation diagram (top) of dissipative (learning with memory loss) dynamics projected onto coordinate $y_1$ from the Poincaré section hyperplane ($\dot{u}_1 = 0$, $\dot{v}_1 > 0$) and the largest two Lyapunov exponents $\lambda_1$ and $\lambda_2$ (bottom) as a function of $\epsilon_Y \in [-1, 1]$. Here with $\epsilon_X = 0.5$, $\alpha_X = \alpha_Y = 0.01$, and $\beta_X = \beta_Y = 1.0$. Simulations show that $\lambda_3$ and $\lambda_4$ are always negative.

(Eqs. (7)) is more challenging.

Another key generalization will be to go beyond the limited adaptive dynamics of reinforcement learning agents to agents that actively build and interpret structural models of their environment; using, for example, online $\epsilon$-machine reconstruction [15]. To be relevant to applications, one will also need to develop a statistical dynamics generalization [16] of the deterministic equations of motion to account for finite and fluctuating numbers of agents and also finite histories used in learning.

Finally, another direction, especially useful if one attempts to quantify collective function in large multiagent systems, will be structural and information-theoretic analyses [17] of local and global learning behaviors and, importantly, their differences. Analyzing the stored information in each agent versus that in the collective, the causal architecture of information flow between an individual agent and the group, and how individual and global memories are processed to sustain collective function are projects for the future.

We presented a dynamical-systems model of collective learning in multiagent systems, which starts with reinforcement learning agents and reduces to coupled replicator equations, demonstrated that individual-agent learning induces a global game dynamics, and investigated some of the resulting periodic, intermittent, and chaotic behaviors in the rock-scissors-papers game interaction. Our model gives a macroscopic description of a network of learning agents that can be straightforwardly extended to model a large number of heterogeneous agents in fluctuating environments. Since deterministic chaos occurs even in this simple setting, one expects that in high-dimensional and heterogeneous populations typical of multiagent systems intrinsic unpredictability will become a dominant collective behavior. Sustaining useful collective function in multiagent systems becomes an even more compelling question in light of these results.

The authors thank J. D. Farmer, R. Shaw, E. Akiyama, P. Patelli, and C. Shalizi. This work was supported at the Santa Fe Institute under the Network Dynamics Program by Intel Corporation and under DARPA agreement F30602-00-2-0583. YS's participation was supported by the Postdoctoral Researchers Program at RIKEN.

* Electronic address: ysato@bdc.brain.riken.go.jp

† Electronic address: chaos@santafe.edu

[1] S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau, eds., *Self-Organization in Biological Systems* (Princeton University Press, Princeton, 2001).

[2] H. A. Simon, *The Sciences of the Artificial*, Karl Taylor Compton Lectures (MIT Press, Cambridge, 1996), first edition 1969.

[3] H. P. Young, *Individual strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton University Press, Princeton, 1998).

[4] E. Hutchins, *Cognition in the Wild* (MIT Press, Cambridge, 1996).

[5] O. E. Rossler, Ann. NY Acad, Sci. **504**, 229 (1987).

[6] M. Taiji and T. Ikegami, Physica **D134**, 253 (1999).

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT Press, 1998).

[8] P. Taylor and L. Jonker, Math. Bio. **40**, 145 (1978).

[9] T. Borgers and R. Sarin, J. Econ. Th. **77**, 1 (1997).

[10] P. Taylor, J. Appl. Prob. **16**, 76 (1979).

[11] J. Hofbauer, J. Math. Biol. **34**, 675 (1996).

[12] H. Yoshida, Phys. Lett. **A150**, 262 (1990).

[13] Y. Sato, E. Akiyama, and J. D. Farmer, Proc. Natl. Acad. Sci. USA **99**, 4748 (2002).

[14] T. Chawanya, Prog. Theo. Phys. **94**, 163 (1995).

[15] K. L. Shalizi, C. R. Shalizi, and J. P. Crutchfield, J. Mach. Learn. Res. (2002), submitted.

[16] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, Theoret. Comp. Sci. **229**, 41 (1999).

[17] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989), see also, J. P. Crutchfield and D. P. Feldman, arXiv.org/abs/cond-mat/0102181 (2001).

[18] Eqs. (8) specify the von Neumann-Morgenstern utility (J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, (Princeton University Press, Princeton, 1944)).

[19] The proof follows P. Schuster et al, Biol. Cybern. **40**, 1 (1981).