Reinforcement learning meets minority game: Toward optimal resource allocation

Si-Ping Zhang,[1,2] Jia-Qi Dong,[2,3] Li Liu,[4] Zi-Gang Huang,[1,*] Liang Huang,[2] and Ying-Cheng Lai[5]

[1]*The Key Laboratory of Biomedical Information Engineering of Ministry of Education, The Key Laboratory of Neuro-informatics & Rehabilitation Engineering of Ministry of Civil Affairs, and Institute of Health and Rehabilitation Science, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*

[2]*Institute of Computational Physics and Complex Systems, Lanzhou University, Lanzhou 730000, China*

[3]*Institute of Theoretical Physics, Key Laboratory of Theoretical Physics, Chinese Academy of Sciences, P.O. Box 2735, Beijing 100190, China*

[4]*School of Software Engineering, Chongqing University, Chongqing 400044, People's Republic of China*

[5]*School of Electrical, Computer and Energy Engineering, Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Received 20 July 2018; published 6 March 2019)

The main point of this paper is to provide an affirmative answer through exploiting reinforcement learning (RL) in artificial intelligence (AI) for eliminating herding without any external control in complex resource allocation systems. In particular, we demonstrate that when agents are empowered with RL (e.g., the popular Q-learning algorithm in AI) in that they get familiar with the unknown game environment gradually and attempt to deliver the optimal actions to maximize the payoff, herding can effectively be eliminated. Furthermore, computations reveal the striking phenomenon that, regardless of the initial state, the system evolves persistently and relentlessly toward the optimal state in which all resources are used efficiently. However, the evolution process is not without interruptions: there are large fluctuations that occur but only intermittently in time. The statistical distribution of the time between two successive fluctuating events is found to depend on the parity of the evolution, i.e., whether the number of time steps in between is odd or even. We develop a physical analysis and derive mean-field equations to gain an understanding of these phenomena. Since AI is becoming increasingly widespread, we expect our RL empowered minority game system to have broad applications.

DOI: 10.1103/PhysRevE.99.032302

I. INTRODUCTION

The tremendous developments in information technology have made it possible for artificial intelligence (AI) to penetrate every aspect of human society. One of the fundamental traits of AI is decision making—individuals, organizations, and governmental agencies tend to rely more and more on AI to make all kinds of decisions based on vast available information in an increasingly complex environment. At present, whether a strong reliance on AI is beneficial or destructive to mankind is an issue of active debate that attracts a great deal of attention from all professions. In the vast field of AI-related research, a fundamental issue is how AI affects or harnesses the behaviors of complex dynamical systems. In this paper, we address this issue by focusing on complex resource allocation systems that incorporate AI in decision making at the individual agent level, and we demonstrate that AI can be quite advantageous for complex systems to reach their optimal states.

Resource allocation systems are ubiquitous and provide fundamental support for the modern economy and society, which are typically complex systems consisting of a large number of interacting elements. Examples include ecosystems of different sizes, various transportation systems (e.g., the Internet, urban traffic systems, and rail and flight networks), public service providers (e.g., marts, hospitals, and schools), as well as social and economic organizations (e.g., banks and financial markets). In a resource allocation system, a large number of components/agents compete for limited public resources in order to maximize payoff. The interactions among the agents can lead to extremely complex dynamical behaviors with negative impacts on the whole system, among which irrational herding is of great concern as it can cause certain resources to be overcrowded but leave others unused, and it has the potential to lead to a catastrophic collapse of the whole system in a relatively short time. A general paradigm to investigate the collective dynamics of resource allocation systems is complex adaptive systems theory [1–3]. At the microscopic level, multiagent models such as the minority game model [4] and interaction models based upon traditional game theory [5–7] have been proposed to account for the interactions among the individual agents.

The minority game is a paradigmatic model for resource allocation in a population. It was introduced in 1997 [4] to study quantitatively the classic El Farol bar-attendance problem first conceived by Arthur in 1994 [8]. In the past two decades, the minority game and its variants were studied extensively [9–35], where a central goal was to uncover the dynamical mechanisms responsible for the emergence of various collective behaviors. In the original minority game model, an individual's scheme for state updating (or decision making) is essentially a trial-and-error learning process based on the global historical winning information [4]. In other

*huangzg@xjtu.edu.cn

models, learning-mechanism-based local information from neighbors was proposed [11,12,16,17,25,28,31–35]. The issue of controlling and optimizing complex resource allocation systems was also investigated [32], e.g., utilizing pinning control to harness the herding behavior, where it was demonstrated that a small number of control points in the network can suppress or even eliminate herding. A theoretical framework for analyzing and predicting the efficiency of pinning control was developed [32], revealing that the connecting topology among the agents can play a significant role in the control outcome. Typically, control requires external interventions. A question is whether herding can be suppressed or even eliminated without any external control.

In this paper, we address the question of how AI can be exploited to harness undesired dynamical behaviors to greatly benefit the operation of the underlying complex system. More generally, we aim to study how reinforcement learning (RL) in AI affects the collective dynamics in complex systems. For this purpose, we introduce a minority game model incorporating RL at the individual agent level, where the agents participating in the game are "intelligent" in the sense that they are capable of RL [36], a powerful learning algorithm in AI. Empowered with RL, an agent is capable of executing an efficient learning path toward a predefined goal through a trial-and-error process in an unfamiliar game environment. Our model is constructed based on the interplay of a learning agent and the environment in terms of the states, actions, rewards, and decision making. In RL, the concepts of value and value functions are key to intelligent exploration, and there have been a number of RL algorithms, such as dynamic programming [36,37], the Monte Carlo method [36,37], temporal differences [36,38], Q-Learning [36,39,40], Sarsa [36], Dyna [36], etc. We focus on Q-learning, which was demonstrated previously to perform well for a small number of individuals in their interaction with an unknown environment [41–45]. In particular, it was demonstrated that incorporating Q-learning into minority game dynamics [46] can suppress herding. Distinct from previous work, here we study minority game dynamics with a large number of "intelligent" players, where Q-learning is adopted for state updating in a stochastic dynamical environment, without involving any other algorithm. The question is whether the multiagent RL minority game system can self-organize to generate optimal collective behaviors. Our main result is an affirmative answer to this question. In particular, we find that the population of RL-empowered agents can approach the optimal state of resource utilization through self-organization regardless of the initial state, effectively eliminating herding. However, the process of evolution toward the optimal state is typically disturbed by intermittent, large fluctuations (oscillations) that can be regarded as failure events. There can be two distinct types of statistical distributions of the "laminar" time intervals in which no failure occurs, depending on their parity, i.e., whether the number of time steps between two consecutive failures is odd or even. We develop a physical analysis and use the mean-field approximation to understand these phenomena. Our results indicate that Q-learning is generally powerful in optimally allocating resources to agents in a complex interacting environment.

II. MODEL

Our minority game model with agents empowered by Q-learning can be described as follows. The system has N agents competing for two resources denoted by $r = +1$ and -1, and each agent chooses one resource during each round of the game. The resources have a finite capacity C_r, i.e., the maximum number of agents that each resource can accommodate. For simplicity, we set $C_r = N/2$. Let $A(t)$ denote the number of agents selecting the resource $r = +1$ at time step t. For $A(t) \leqslant C_r$, agents choosing the resource $+1$ belong to the minority group and win the game in this round. Conversely, for $A(t) > C_r$, the resource $+1$ is overcrowded, so the corresponding agents fail in this round.

The Q-learning adaptation mechanism [40] is incorporated into the model by assuming that the states of the agents are parametrized through Q functions that characterize the relative utility of a particular action. The Q functions are updated during the course of the agents' interaction with the environment. Actions that lead to a higher reward are reinforcement. To be concrete, in our model, agents are assumed to have four available actions, and we let $Q(s, a)$ be the Q value of the corresponding action at time t, where s and a denote the current state of agent and the action that the agent may take, respectively. A Q function can then be expressed in the following form:

$$\mathbf{Q} = \begin{bmatrix} Q(+1,+1) & Q(+1,-1) \\ Q(-1,+1) & Q(-1,-1) \end{bmatrix}.$$

For an agent in state s, after selecting a given action a, the corresponding Q value is updated according to the following rule:

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha\big[R_t(a) + \gamma Q_{t-1}^{\max}(s',a') - Q_{t-1}(s,a)\big], \tag{1}$$

where s denotes the current state of the agent, i.e., the agent's action in the last step, a denotes the action that the agent may take, $\alpha \in (0, 1]$ is the learning rate, and $R_t(a)$ is the reward from the corresponding action a at time t. The parameter $\gamma \in [0, 1)$ is the discount factor that determines the importance of future reward. Agents with $\gamma = 0$ are "short sighted" in that they consider only the current reward, while those with larger values of γ care about reward in the long run. The quantity $Q_{t-1}^{\max}(s', a')$ is the maximum element in the row of the s' state, which is the outcome of the action a based on s, that is, s' is equal to action a at the current time step. Equation (1) indicates that the matrix $\mathbf{Q}$ contains information about the accumulative experience from history, where the reward $R_t(a)$ (for action a from state s) and the expected best value $Q_{t-1}^{\max}(s', a')$ based on s' both contribute to the updated value $Q_t(s, a)$ with the weight α, and the previous value $Q_{t-1}(s, a)$ is also accumulated into $Q_t(s, a)$ with the weight $1 - \alpha$.

While agents select the action mostly through RL, certain randomness can be expected in decision making. We thus assume that a random action occurs with a small probability ϵ, and agents select the action with a large value of $Q(s, a)$ with probability $1 - \epsilon$. For a given setting of parameters α and γ, the Q-learning algorithm is carried out, as follows.

First, we initialize the matrix **Q** to zero to mimic the situation in which the agents are unaware of the game environment, and we initialize the state s of each agent randomly to $+1$ or -1. Next, for each round of the game, each agent chooses an action a with a larger value of $Q_t(s, a)$ in the row of its current state s with probability $1-\epsilon$, or chooses an action a randomly with probability ϵ. The $Q(s, a)$ value of the selected action is then updated according to Eq. (1). The action leading to the state s' identical to the current winning (minority) state has $R_t(a) = R = 1$, and the action leading to the failed (majority) state has $R_t(a) = 0$. Finally, we take the selected action a to update the state from s to s'.

Distinct from the standard supervised learning [47], agents adopting RL aim to understand the environment and maximize their rewards gradually through a trial-and-error process. The coupling or interaction among the agents is established through competing for limited resources. Our RL-based minority game model also differs from the previously studied game systems [32] in that our model takes into account agents' complicated memory and decision-making process. For our system, a key question is whether the resulting collective behaviors from RL may lead to high efficiency or optimal resource allocation in the sense that the number of agents that a resource accommodates is close to its capacity.

III. SELF-ORGANIZATION AND COMPETITION

In minority game dynamics, a common phenomenon is herding, in which a vast majority of the agents compete for a few resources, leaving other resources idle. The phenomenon emerges due to the feedback on historical information in the game system, i.e., the individuals rely on global or local historical information before making a decision. Herding is harmful and undesired as it can lead to starvation of certain resources and ineffective usage of others, greatly reducing the system's efficiency. Herding can even cause the whole system to collapse in a short time. In our system with RL, herding also occurs but, due to the intrinsic Q-learning mechanism, the behavior is spontaneously suppressed in a periodic fashion, as the periodic bursts of failures can lead to dramatic fluctuations in the utilization of the resources.

In the traditional minority game, the dynamical rules stipulate that competition and learning among agents can lead to the detrimental herding behavior to which game systems composed of less diversified agents are particularly susceptible [32–35]. In our minority game system of agents empowered with RL, herding is dramatically suppressed. To give a concrete example, we set the parameters for Q learning as follows: learning rate $\alpha = 0.9$, discount factor $\gamma = 0.9$, and exploration rate $\epsilon = 0.02$. Figure 1(a) shows the temporal evolution of the number $A(t)$ of agents choosing resource $+1$. The main features of the time series are the continuous oscillations of $A(t)$ about the capacity C_r of resources, convergence of the oscillation amplitude, and bursts of $A(t)$ that occur intermittently. As the oscillations converge to the optimal state, the two resources $r = +1$ and -1 play as the minority resource alternatively. The remarkable feature is that the agent population tends to self-organize into a nonequilibrium state with a certain temporal pattern in order to reach the highly efficient, optimal state, but the process is interrupted by large bursts (failures or fluctuations).

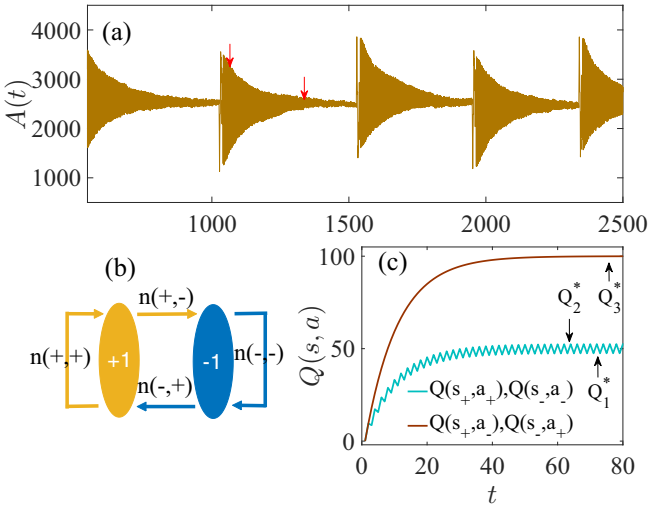

FIG. 1. Typical temporal evolutionary behavior of the proposed minority game system with RL empowered agents. (a) Time series of the attendance $A(t)$ of resource $+1$. Interactions among the agents make the system self-organize into a special temporal pattern with two main features: the convergence of regular oscillations toward the optimal value $C_r = N/2$, and intermittent bursts of failures in utilizing resources. (b) A schematic sketch of the state transitions of agents during the dynamical process. There are self-satisfied agents in a fixed state and speculative agents that continuously switch states between $+1$ and -1. (c) Time series of $Q(s, a)$ as the numerical solutions of Eqs. (2)–(4). The parameters are as follows: learning rate $\alpha = 0.9$, discount factor $\gamma = 0.9$, exploration rate $\epsilon = 0.02$, and system size $N = 5001$.

A. Convergence of oscillations

1. Emergence of two types of agents

From numerical simulations of the RL minority game system, we find that, as the system self-organizes itself into patterns of regular oscillations, agents with two types of behaviors emerge. The first type is those agents who are "self-satisfied" in the sense that they remain in either the $s = +1$ state or the $s = -1$ state. Those agents win and lose the game alternatively as the system develops regular oscillations. The population sizes of the self-satisfied agents are denoted as $n(+1, +1)$ and $n(-1, -1)$, respectively. The second type of agents are the "speculative" agents, or speculators, who switch states at each time step between $s = +1$ and -1. These agents always win the game when the system exhibits regular oscillations. We denote the population sizes of the speculative agents as $n(+, -)$ and $n(-, +)$, which correspond to the two possibilities of switching: from $s = +1$ to -1 and vice versa, respectively.

Figure 1(b) shows the state transition paths induced by the self-satisfied agents and the speculative agents. The oscillations of $A(t)$ associated with the convergent process can be attributed to the state transition of the speculative agents between the states $+1$ and -1. This agrees with the intuition that, e.g., the investing behavior of speculators in a financial

market is always associated with high risks and large oscillations. Due to the decrease in the population of the speculative agents, the oscillation amplitude in any time interval between two successive failure events tends to decay with time.

2. *Stable state of Q table*

The oscillations of $A(t)$ mean that $r = +1$ and -1 act as the minority resource alternatively. For the self-satisfied agents, according to the Q-learning algorithm, the update of the element $Q(s_+, a_+)$ can be expressed as

$$\begin{aligned} Q_{t+1}(s_+, a_+) &= Q_t(s_+, a_+) \\ &\quad + \alpha[R + \gamma Q_t(s_+, a_+) - Q_t(s_+, a_+)], \\ Q_{t+2}(s_+, a_+) &= Q_{t+1}(s_+, a_+) \\ &\quad + \alpha[\gamma Q_{t+1}(s_+, a_+) - Q_{t+1}(s_+, a_+)], \end{aligned} \quad (2)$$

where $Q_t^{\max}(s', a') = Q_t(s_+, a_+)$ due to the inequality $Q(s_+, a_+) > Q(s_+, a_-)$. The update of the element $Q(s_-, a_-)$ is described by

$$\begin{aligned} Q_{t+1}(s_-, a_-) &= Q_t(s_-, a_-) \\ &\quad + \alpha[R + \gamma Q_t(s_-, a_-) - Q_t(s_-, a_-)], \\ Q_{t+2}(s_-, a_-) &= Q_{t+1}(s_-, a_-) \\ &\quad + \alpha[\gamma Q_{t+1}(s_-, a_-) - Q_{t+1}(s_-, a_-)], \end{aligned} \quad (3)$$

where $Q_t^{\max}(s', a') = Q_t(s_-, a_-)$ as a result of the inequality $Q(s_-, a_-) > Q(s_-, a_+)$.

For the speculative agents, the updating equations of elements $Q(s_+, a_-)$ and $Q(s_-, a_+)$ are

$$\begin{aligned} Q_{t+1}(s_+, a_-) &= Q_t(s_+, a_-) \\ &\quad + \alpha[R + \gamma Q_t(s_-, a_+) - Q_t(s_+, a_-)], \\ Q_{t+2}(s_-, a_+) &= Q_{t+1}(s_-, a_+) \\ &\quad + \alpha[R + \gamma Q_{t+1}(s_+, a_-) - Q_{t+1}(s_-, a_+)], \end{aligned} \quad (4)$$

where $Q_t^{\max}(s', a') = Q_t(s_-, a_+)$ or $Q_t(s_+, a_-)$, due to the inequalities $Q(s_+, a_+) < Q(s_+, a_-)$ and $Q(s_-, a_-) < Q(s_-, a_+)$.

Figure 1(c) shows numerically obtained time series of the elements of the matrix $\mathbf{Q}$ from Eqs. (2)–(4). For the self-satisfied agents, the values of $Q(s_+, a_+)$ and $Q(s_-, a_-)$ increase initially, followed by an oscillating solution between the two values Q_1^* and Q_2^*, where

$$Q_1^* = \frac{[1 + \alpha(\gamma - 1)]\alpha R}{1 - [1 + \alpha(\gamma - 1)]^2},$$

$$Q_2^* = \frac{\alpha R}{1 - [1 + \alpha(\gamma - 1)]^2}$$

are obtained from Eqs. (2) and (3). For the speculative agents, both $Q(s_+, a_-)$ and $Q(s_-, a_+)$ reach a single stable solution $Q_3^* = R/(1 - \gamma)$, which can be obtained by solving Eq. (4). The three relevant values have the relationship $Q_1^* < Q_2^* < Q_3^*$.

The emergence of the two types of agents can be understood from the following heuristic analysis. In the dynamical process, a speculative agent emerges when the element associated with an agent satisfies the inequalities $Q(s_+, a_-) > Q(s_+, a_+)$ and $Q(s_-, a_+) > Q(s_-, a_-)$ simultaneously. Initially, the agents attend both resources $+1$ and -1, with one group winning but the other losing. Only the group that always wins the game can reinforce themselves through further increment in $Q(s_+, a_-)$ and $Q(s_-, a_+)$. The stable group of speculative agents leads to regular oscillations of $A(t)$, because they switch states together between $+1$ and -1. An agent becomes self-satisfied when it is in the $+1$ state and the inequality $Q(s_+, a_+) > Q(s_+, a_-)$ holds, or in the -1 state and $Q(s_-, a_-) > Q(s_-, a_+)$ holds. The self-satisfied state can be strengthened following the evolution governed by Eqs. (2) and (3), with $Q(s_+, a_+)$ or $Q(s_-, a_-)$ reaching the oscillating state between Q_1^* and Q_2^*, as shown in Fig. 1(c). We see that the condition for an agent to become speculative is more strict than to be self-satisfied. Moreover, a speculative agent has a certain probability to become self-satisfied, as determined by the value of the exploration rate ϵ. As a result, the population of the speculative agents tends to shrink, leading to a decrease in the oscillation amplitude $|n(+, -) - n(-, +)|$ and convergence of $A(t)$ closer to the optimal state $N/2$.

For the special case of $\epsilon = 0$ [the gray regions in Figs. 2(a) and 2(b)], agents take action entirely based on historical experience $\mathbf{Q}$. In this case, the numbers of the self-satisfied and speculative agents become constant, and $A(t)$ no longer converges to that associated with the optimal state. In general, randomness in exploration can affect the convergence of the system dynamics toward the state in which the resources are optimally utilized. Specifically, random explorations can dramatically increase the number of possible evolutionary paths, while actions according to the Q-function restrict the evolution direction of the system toward the optimal reward-driven path. As a result, setting $\epsilon \neq 0$ can lead to an opti-

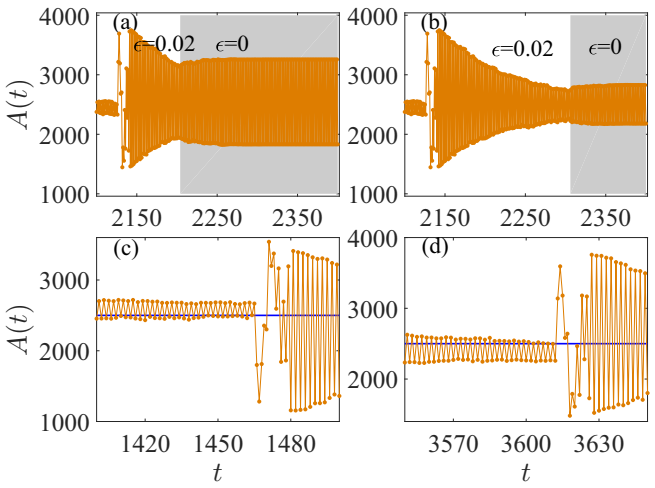

FIG. 2. Convergence of regular oscillations and bursts of failure. (a),(b) Convergence of the regular oscillation pattern depends on the exploration behavior of the agents as characterized by the rate ϵ. For $\epsilon = 0$ (gray region), the oscillation amplitude does not converge. (c),(d) Detailed processes for the bursts of failure. If the regular oscillations do not cross the line $C_r = N/2$ but behave either as (c) $A(t) > C_r$ and $A(t + 1) > C_r$ or as (d) $A(t) < C_r$ and $A(t + 1) < C_r$, the regular oscillations stop and a systematic failure burst emerges. The blue line specifies $A(t) = C_r$. The parameters are the same as in Fig. 1.

mal path of equilibrium attendance, effectively eliminating herding. Nonetheless, a bursting behavior can emerge in the long-term evolution of the system, reducing efficiency. Randomness in exploration thus plays the role of a double-edged sword: a larger value of ϵ can facilitate a system's settling into the optimal convergence path and suppressing herding, but it can lead to an undesired bursting behavior.

B. Intermittent failures in the RL empowered minority game system

The intermittent bursts of failure events in the whole system take place during the convergent process to the optimal state. An understanding of the mechanism of the failures can provide insights into the articulation of strategies to make the system more robust and resilient.

The criterion to determine if an agent selecting $+1$ wins the minority game is $A(t) < C_r = N/2$. If the event $A(t) < C_r$ [or $A(t) > C_r$] occurs twice in a row, the oscillation pattern will be broken. Since the agents are empowered with RL, two consecutive winnings of either resource -1 or resource $+1$ represent an unexpected event, and this would lead to cumulative errors in the Q table, triggering a burst of error in decision making and, consequently, leading to failures in utilizing the resources. To see this in a more concrete way, we note that a self-satisfied agent wins and fails alternatively following a regular oscillation pattern. If the agent fails twice in row, its confidence in preserving the current state is reduced. As a result, the event $Q(s_+, a_+) < Q(s_+, a_-)$ or $Q(s_-, a_-) < Q(s_-, a_+)$ would occur with a high probability, leading to a decrease in the populations $n(+,+)$ and $n(-,-)$ of the self-satisfied agents. The populations of the speculative agents, $n(+,-)$ and $n(-,+)$, are increased accordingly. These events collectively generate a bursting disturbance to the regular oscillation pattern of $A(t)$, terminating the system's convergence toward the optimal state, as shown in Figs. 2(a) and 2(b).

In general, the stability of the regular oscillations depends on two factors: the equilibrium position determined by the self-satisfied agents, and the random fluctuations introduced by agents' exploration behavior. For the first factor, the equilibrium position is given by $A_0 = n(+,+) + [n(-,+) - n(+,-)]/2$, which deviates from C_r due to the asymmetric distribution of the self-satisfied agents in the two distinct resources. Figures 2(c) and 2(d) show two examples with the equilibrium position A_0 larger or smaller than C_r (the blue solid line), respectively. We see that the converging process is terminated when either the upper or the lower envelope reaches C_r, i.e., when two consecutive steps of $A(t)$ stay on the same side of C_r in replacement of an oscillation about C_r. In the thermodynamic limit, for an infinitely large system with self-satisfied agents symmetrically distributed between $+1$ and -1 (so that the equilibrium position A_0 is at C_r), the oscillation would persist indefinitely and $A(t)$ approaches C_r asymptotically.

The second factor of random fluctuations in agents' exploratory behavior is caused by the finite system size, which affects the oscillation stability. As the populations $[n(+,-)$ and $n(-,+)]$ of the speculative agents decrease during the converging process, the amplitude of oscillation, $|n(+,-) - n(-,+)|$, becomes comparable to $\sqrt{\epsilon N}$, the level of random fluctuations in the system. The occurrence of two consecutive steps of $A(t) > C_r$ [or $A(t) < C_r$] as a result of the fluctuations will break the regular oscillation pattern. In the thermodynamical limit, the effects of the random fluctuations are negligible. We note that, while similar bursting behaviors were observed previously [48,49], the underlying mechanisms are different from ours. In particular, a finite memory was introduced into the standard or grand-canonical minority game models, where in the former the nonergodic behavior and sensitive dependence on initial conditions were suppressed, but in the latter large fluctuations arise. The main reason for the bursts is a finite score effect based on the memory factor λ, where speculators who may abstain from playing if the game was not profitable enough are allowed to participate in the game. That is, the long-term weighted historical memory is used in agents' decision making. In our work, the bursting phenomenon can be attributed to the interaction between the Q function and random explorations in systems of finite size, i.e., the historical information is recorded in the Q function through RL.

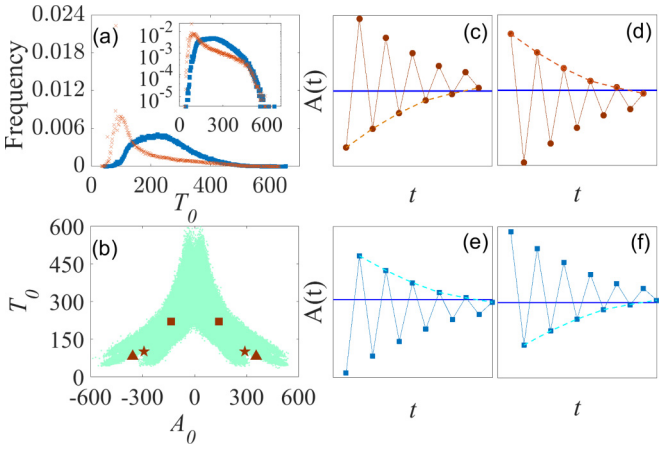

FIG. 3. Statistical distributions of the time interval T_0 between two successive bursts of failure. (a) The distributions obtained from one realization of the system dynamics, where those of the odd T_0 values (red crosses) and even T_0 values (blue squares) are remarkably distinct. (b) T_0 vs the deviation $A_0 - C_r$ of the equilibrium position from the resource capacity. The solid squares (triangles) denote the most probable value of the set of even (odd) T_0 values. The parameters are $\alpha = 0.9$, $\gamma = 0.9$, and $N = 5001$. (c)–(f) Schematic illustration of four cases associated with the regular oscillations of $A(t)$, where cases (c),(d) lead to odd intervals T_0 while cases (e),(f) lead to the even values of T_0. The dashed curves represent the envelopes that cross the capacity value C_r (solid blue lines), which triggers a failure burst.

C. Time intervals between failure bursts

The dynamical evolution of the system can be described as random failure bursts superimposed on regular oscillations with decreasing amplitude. The intermittent failures can be characterized by the statistical distribution of the time interval T_0 between two successive bursting events. Figure 3(a) shows

a representative histogram of T_0 obtained from a single statistical realization of the system dynamics (the inset showing the same data but on a semilogarithmic scale). A remarkable feature is that the distributions of the odd (red crosses) and even values of T_0 (blue squares) are characteristically distinct. In particular, the odd values of T_0 emerge with a smaller probability and the corresponding distribution has a smaller most probable value as compared with that for the even values of T_0. A possible explanation lies in the existence of two intrinsically distinct processes.

Our computation and analysis indicate that the regular oscillation processes can be classified into two categories, as shown in Figs. 3(c)–3(f), leading to insights into the mechanism for the two distinct types of statistical distributions in T_0. In Fig. 3(c), $A(t)$ starts from a value below $C_r = N/2$ and terminates at a value above C_r, due to the two consecutive values above C_r as the lower envelope of $A(t)$ crosses C_r. Similarly, in Fig. 3(d), $A(t)$ starts from a value above C_r and terminates at a value below C_r, with the upper envelope of $A(t)$ crossing C_r. In Fig. 3(e), $A(t)$ starts from a value below C_r and terminates at a value below C_r. In Fig. 3(f), $A(t)$ starts from a value above C_r and terminates at a value above C_r. In Figs. 3(c) and 3(d), odd intervals are generated, while in Figs. 3(e) and 3(f) the intervals are even. Between the cases in the same category [e.g., (c),(d) or (e),(f)], there is little difference in the statistical distribution of T_0, especially in the long-time limit.

We have seen that the equilibrium position A_0 plays an important role in terminating the regular oscillations, which can be calculated as $A_0 = \langle A(t)\rangle_t$, where $\langle\cdot\rangle_t$ denotes the average over time. From Fig. 3(b) where the time interval T_0 is displayed as a function of the quantity $A_0 - C_r$, we see that the values of A_0 closer to the capacity C_r lead to regular oscillations with larger values of T_0. The most probable values of the distributions of the even (squares) and odd (stars) T_0 values are also indicated in Fig. 3(b).

D. Mean-field theory

We develop a mean-field analysis to capture the main features of the dynamical evolution of the multiagent RL minority game system. We assume that the agents empowered with RL are identical and share the same matrix $\mathbf{Q}$. The dynamical evolution of $A(t)$ can be described by the following equation:

$$\frac{dA(t)}{dt} = \epsilon\frac{N}{2} + (1-\epsilon)\{A(t)\Theta(X_1) + [N - A(t)]\Theta(X_2)\} - A(t), \tag{5}$$

where the first item $\epsilon N/2$ is the number of agents that act randomly with probability ϵ, half of which select $+1$. The second item indicates the number of agents that act based on the matrix $\mathbf{Q}$ with probability $1-\epsilon$, which include agents that stay in the $+1$ state and those that transition from -1 to $+1$. $\Theta(X)$ denotes the step function: $\Theta(X) = 0$ for $X < 0$, $\Theta(X) = 1/2$ for $X = 0$, and $\Theta(X) = 1$ for $X > 0$. The quantities X_1 and X_2 are defined as $X_1 \equiv Q_t(s_+, a_+) - Q_t(s_+, a_-)$ and $X_2 \equiv Q_t(s_-, a_+) - Q_t(s_-, a_-)$. The elements of the matrix $\mathbf{Q}$ are updated according to the following rules:

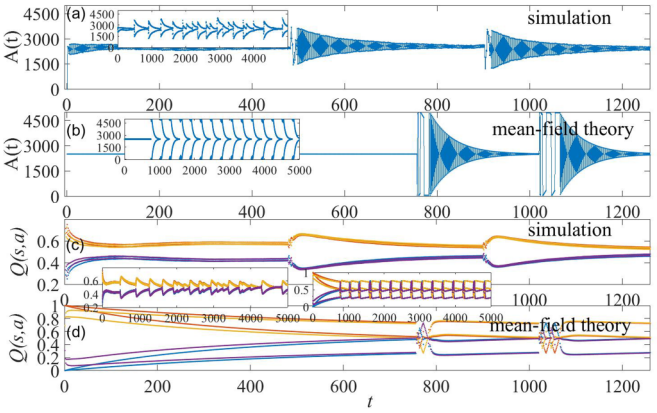

FIG. 4. Comparison of dynamical evolution of the system obtained from simulation and mean-field theory. The attendance $A(t)$ obtained from (a) multiagent simulation and (b) numerical solution of Eqs. (4)–(9). (c), (d) The corresponding results of the elements of $\mathbf{Q}$ from multiagent simulation and from numerical solution, respectively. The insets in (a)–(d) show the corresponding time series of $A(t)$ and $\mathbf{Q}$ in a large time regime. The parameters are $\alpha = 0.9$, $\gamma = 0.9$, $\epsilon = 0.02$, and $N = 5001$.

$$\frac{dQ_t(s_+, a_+)}{dt} = \alpha\left[R\Theta(X_3) + \gamma Q_t^{\max} - Q_t(s_+, a_+)\right] \times \left[(1-\epsilon)\Theta(X_1) + \frac{1}{2}\epsilon\right], \tag{6}$$

$$\frac{dQ_t(s_+, a_-)}{dt} = \alpha\left[R\Theta(-X_3) + \gamma Q_t^{\max} - Q_t(s_+, a_-)\right] \times \left[(1-\epsilon)\Theta(-X_1) + \frac{1}{2}\epsilon\right], \tag{7}$$

$$\frac{dQ_t(s_-, a_+)}{dt} = \alpha\left[R\Theta(X_3) + \gamma Q_t^{\max} - Q_t(s_-, a_+)\right] \times \left[(1-\epsilon)\Theta(X_2) + \frac{1}{2}\epsilon\right], \tag{8}$$

$$\frac{dQ_t(s_-, a_-)}{dt} = \alpha\left[R\Theta(-X_3) + \gamma Q_t^{\max} - Q_t(s_-, a_-)\right] \times \left[(1-\epsilon)\Theta(-X_2) + \frac{1}{2}\epsilon\right], \tag{9}$$

where $X_3 \equiv N - 2A(t)$, the step function $\Theta(X_3)$ indicates whether or not the agents gain a reward, and $Q_t^{\max}$ is the expected value after action. Specifically, we have $Q_t^{\max} = \max[Q_t(s_+, a_+), Q_t(s_+, a_-)]$ in Eqs. (6) and (8) for the agents who take action to transition to $+1$. Similarly, $Q_t^{\max} = \max[Q_t(s_-, a_+), Q_t(s_-, a_-)]$ in Eqs. (7) and (9) is for agents taking action to transition to the state -1.

The dynamical evolution of the system can thus be assessed either through simulation, as presented in Figs. 4(a) and 4(c), or through the mean-field equations (4)–(9), as shown in Figs. 4(b) and 4(d). A comparison between these results indicates that the mean-field equations (4)–(9) capture the main features of the collective dynamics of the RL minority game system, which are regular oscillations with converging amplitude and intermittent bursts of failure. Due to the

approximate nature of the mean-field analysis, its predictions tend to deviate slightly from the simulation results. In particular, the analysis predicts bursts of sizes somewhat larger than those from simulations. The reason is that, under the mean-field approximation, the dynamical behaviors of the agents are determined by a **Q** table. As a result, a burst characteristic of system failure involves the whole population of agents, making the size of the burst larger than that from simulation. In addition, the mean-field analysis shows that the period between adjacent bursts is approximately constant, while simulation reveals more variations in the period. The discrepancy can be attributed to the randomness in the size of the bursts in simulation. The state of the system after the last burst serves effectively as the initial condition of the process leading to a convergent solution, the length of which is affected by the burst size and randomness.

IV. DISCUSSION

Complex resource allocation systems with a large number of interacting components are ubiquitous in modern society. Optimal performance of such a system is typically measured by uniform and even utilization of all available resources by the users. Often this is not possible due to the phenomenon of herding that can emerge spontaneously in the evolution of the system, in which most agents utilize only a few resources, leaving the vast majority of the remaining resources little exploited [11,16,17,31,32,34,35,50–54]. The herding behavior can propagate through the system, as the few heavily used resources would be depleted quickly, directing most agents to another possibly small set of resources, which would be depleted as well, and so on. A final outcome is the total collapse of the entire system. An important goal in managing a complex resource allocation system is to devise effective strategies to prevent herding behavior from occurring. We note that similar behaviors occur in economics [55–58]. Thus any effective methods to achieve optimal performance of resource allocation systems can potentially be generalized to a broader context.

Mathematically, a paradigm to describe and study the dynamics of complex resource allocation is minority games, in which a large number of agents are driven to seek the less used resources based on available information to maximize payoff. In the minority game framework, a recent work addressed the problem of controlling herding [33] using the pinning method that had been studied in controlling collective dynamics such as synchronization in complex networks [32,59–66], where the dynamics of a small number of nodes are "pinned" to some desired behavior. In developing a pinning control scheme, the fraction of agents chosen to hold a fixed state and the structure of the pinned agents are key issues. For the minority game system, during the time evolution, fluctuations that contain characteristically distinct components can arise: intrinsic and systematic, and this allows one to design a control method based on separated control variables [33]. One finding was that a biased pinning control pattern can lead to an optimal pinning fraction that minimizes the system fluctuations, and this holds regardless of the network topologies.

Any control-based method aiming to suppress or eliminate herding requires external input. The question we address in this paper is whether it would be possible to design a "smart" type of resource allocation system that can sense the potential emergence of herding and adjust the game strategy accordingly to achieve the same goal but without any external intervention. Our answer is affirmative. In particular, we introduce RL from AI into the minority game system in which the agents are "intelligent" and empowered with RL. Exploiting a popular learning algorithm in AI, Q-learning, we find that the collective dynamics can evolve to the optimal state in a self-organized fashion, which is effectively immune from any herding behavior. Due to the complex dynamics, the evolution toward the optimal state is not uninterrupted: there can be intermittent bursts of failures. However, because of the power of self-learning, once a failure event has occurred, the system can self-repair or self-adjust to start a new process of evolution toward the optimal state, free of herding. One finding is that two distinct types of the probability distribution of the intervals of free evolution (the time interval between two successive failure events) arise, depending on the parity of the system state. We provide a physical analysis and derive mean-field equations to understand these behaviors. AI has become increasingly important and has been universally applied to all aspects of modern society. Our work demonstrates that the marriage of AI with complex systems can generate optimal performance to a certain extent, without the need for external control or intervention.

ACKNOWLEDGMENTS

We thank Ji-Qiang Zhang, Zhi-Xi Wu, and Ri-Chong Zhang for helpful discussions. This work was partially supported by the NSF of China under Grants No. 11275003, No. 11775101, No. 11647052, No. 61431012, and the Young Talent fund of University Association for Science and Technology in Shaanxi, China (Program No. 20170606). Z.-G.H. acknowledges the support of K. C. Wong Education Foundation, the Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning, and the Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2016-123. Y.-C.L. would like to acknowledge support from the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineering and funded by the Office of Naval Research through Grant No. N00014-16-1-2828.

APPENDIX

Convergence mechanism of $A(t)$

Typically, after a failure burst, $A(t)$ will converge to the value corresponding to the optimal system state. The mechanism of convergence can be understood as follows. The essential dynamical event responsible for the convergence is the change of agents from being speculative to being self-satisfied within the training time. If the inequalities $Q(s_+, a_+) < Q(s_+, a_-)$ and $Q(s_-, a_-) < Q(s_-, a_+)$ hold, the agent is speculative and wins the game all the time as a result of the state transition. Otherwise, for $Q(s_+, a_+) > Q(s_+, a_-)$ and $Q(s_-, a_-) > Q(s_-, a_+)$, the agent is self-satisfied and wins and loses the game alternatively.

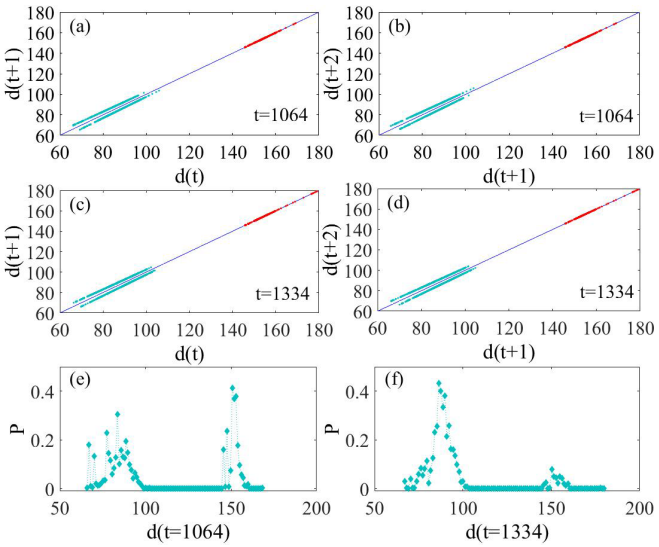

FIG. 5. Matrix norm d for all agents empowered with RL. The norms are indicated by the two red arrow positions in Fig. 1(a). (a)–(d) Evolution of the matrix norm d at three adjacent time steps: $t+1$ vs t and $t+2$ vs $t+1$. The blue line is $x=y$. Panels (a),(b) correspond to the left arrow, and (c),(d) to the right arrow. (e),(f) The corresponding distributions of the matrix norm.

Consider a speculative agent. Assume that its state is $r=+1$ at the current time step. The agent selects $r=+1$ with the probability $\epsilon/2$ and updates $Q(s_+,a_+)$ with reward or selects $r=-1$ with the probability $\epsilon/2+(1-\epsilon)$ and updates $Q(s_+,a_-)$ without reward. If the agent selects $r=+1$, the game will be lost, but the value of $Q(s_+,a_+)$ can increase. At the next time step, the agent selects $r=-1$ and loses the game, and $Q(s_+,a_-)$ will decrease. As a result, the inequality $Q(s_+,a_+)>Q(s_+,a_-)$ holds with the probability $\epsilon/2$. That is, the probability that a speculative agent changes to a self-satisfied one is approximately $\epsilon/2$.

Now consider a self-satisfied agent in the $r=+1$ state at the current time step. The agent selects $r=+1$ with the probability $\epsilon/2+(1-\epsilon)$ and updates $Q(s_+,a_+)$ with two stable solutions (Q_1^* and Q_2^*), or the agent selects $r=-1$ with the probability $\epsilon/2$. The agent selects $r=-1$ from the two stable solutions Q_1^* or Q_2^* with the respective probability $1/2$. If the agent is associated with the smaller stable solution Q_2^*, then $Q(s_+,a_-)$ will decrease. As a result, the agent remains self-satisfied. If the agent is associated with the larger stable solution Q_1^*, then $Q(s_+,a_-)$ will increase due to reward, and the inequality $Q(s_+,a_-)>Q(s_+,a_+)$ holds with the probability $1/2$. At the same time, if $Q(s_-,a_+)>Q(s_-,a_-)$, the probability is approximately equal to $1/2$, and the self-satisfied agent successfully becomes a speculative agent. Otherwise, the self-satisfied agent remains self-satisfied. That is, the probability that a self-satisfied agent changes to being speculative is approximately $\epsilon/16 \ll \epsilon/2$. As a result, $A(t)$ will converge to C_r asymptotically.

Two types of agents in phase space

For the RL minority game system, we can construct the phase space in which the two types of agents can be distinguished. We define the matrix norm $d=\|\mathbf{Q}\|$ for the $\mathbf{Q}$ matrix of each agent as the square root of the sum of all the matrix elements. For the two positions indicated by the red arrows in Fig. 1(a), Figs. 5(a)–5(d) show the relationship of matrix norm d at three adjacent time steps. We see that the agents can be distinguished and classified into two categories through the matrix norm d, where the self-satisfied and the speculative agents correspond to the top and bottom sides of the line $x=y$ and on the line $x=y$, respectively. The reason that the speculative agents change their state while the self-satisfied agents remain in their state lies in the property of the elements of the $\mathbf{Q}$ matrix. In particular, after the system reaches a steady state after training, for the speculative agents, the following inequalities hold: $Q(s_+,a_+)<Q(s_+,a_-)$ and $Q(s_-,a_-)<Q(s_-,a_+)$, while for the self-satisfied agents, the inequalities are $Q(s_+,a_+)>Q(s_+,a_-)$ and $Q(s_-,a_-)>Q(s_-,a_+)$. Since the values of the matrix elements $Q(s_+,a_+)$ and $Q(s_-,a_-)$ associated with the self-satisfied agents are between Q_1^* and Q_2^*, the matrix norm d of these agents rolls over on the line $x=y$ at the adjacent time. However, the elements $Q(s_+,a_-)$ and $Q(s_-,a_+)$ associated with the speculative agents reach only the stable solution Q_3^*. As a result, the matrix norm d of these agents remains unchanged. Figures 5(e) and 5(f) show that the matrix norms for the agents display a two-peak distribution, corresponding to the two types. The peak height on the left-hand side increases with time, while that on the right-hand side decreases.

[1] S. A. Kauffman, *The Origins of Order: Self-organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).
[2] S. A. Levin, Ecosys. **1**, 431 (1998).
[3] W. B. Arthur, S. N. Durlauf, and D. A. Lane, *The Economy as an Evolving Complex System II* (Addison-Wesley, Reading, MA, 1997), Vol. 28.
[4] D. Challet and Y.-C. Zhang, Physica A **246**, 407 (1997).
[5] M. A. Nowak, K. M. Page, and K. Sigmund, Science **289**, 1773 (2000).
[6] C. P. Roca, J. A. Cuesta, and A. Sánchez, Phys. Rev. E **80**, 046106 (2009).
[7] W. H. Press and F. J. Dyson, Proc. Natl. Acad. Sci. (USA) **109**, 10409 (2012).
[8] W. B. Arthur, Am. Econ. Rev. **84**, 406 (1994).
[9] D. Challet and M. Marsili, Phys. Rev. E **60**, R6271 (1999).
[10] R. Savit, R. Manuca, and R. Riolo, Phys. Rev. Lett. **82**, 2203 (1999).
[11] M. Paczuski, K. E. Bassler, and Á. Corral, Phys. Rev. Lett. **84**, 3185 (2000).
[12] D. Challet, A. D. Martino, and M. Marsili, J. Stat. Mech. Theor. Exp. (2008) L04004.

[13] E. Moro, The minority games: An introductory guide, *Advances in Condensed Matter and Statistical Physics* (Nova Science, Hauppauge, New York, 2004).

[14] D. Challet, M. Marsili, and Y.-C. Zhang, *Minority Games*, Oxford Finance (Oxford University Press, Oxford, 2005).

[15] C. H. Yeung and Y.-C. Zhang, in *Encyclopedia of Complexity and Systems Science*, edited by R. A. Meyers (Springer New York, 2009), pp. 5588–5604.

[16] T. Zhou, B.-H. Wang, P.-L. Zhou, C.-X. Yang, and J. Liu, Phys. Rev. E **72**, 046139 (2005).

[17] V. M. Eguiluz and M. G. Zimmermann, Phys. Rev. Lett. **85**, 5659 (2000).

[18] T. S. Lo, K. P. Chan, P. M. Hui, and N. F. Johnson, Phys. Rev. E **71**, 050101 (2005).

[19] N. F. Johnson, M. Hart, and P. M. Hui, Physica A **269**, 1 (1999).

[20] M. Hart, P. Jefferies, N. F. Johnson, and P. M. Hui, Physica A **298**, 537 (2001).

[21] M. Marsili, Physica A **299**, 93 (2001).

[22] G. Bianconi, A. D. Martino, F. F. Ferreira, and M. Marsili, Quant. Financ. **8**, 225 (2008).

[23] Y. B. Xie, C.-K. Hu, B. H. Wang, and T. Zhou, Eur. Phys. J. B **47**, 587 (2005).

[24] L.-X. Zhong, D.-F. Zheng, B. Zheng, and P. M. Hui, Phys. Rev. E **72**, 026134 (2005).

[25] M. Anghel, Z. Toroczkai, K. E. Bassler, and G. Korniss, Phys. Rev. Lett. **92**, 058701 (2004).

[26] T. S. Lo, H. Y. Chan, P. M. Hui, and N. F. Johnson, Phys. Rev. E **70**, 056102 (2004).

[27] F. Slanina, Physica A **299**, 334 (2001).

[28] T. Kalinowski, H.-J. Schulz, and M. Birese, Physica A **277**, 502 (2000).

[29] A. D. Martino, M. Marsili, and R. Mulet, Europhys. Lett. **65**, 283 (2004).

[30] C. Borghesi, M. Marsili, and S. Micciché, Phys. Rev. E **76**, 026104 (2007).

[31] A. Galstyan and K. Lerman, Phys. Rev. E **66**, 015103 (2002).

[32] J.-Q. Zhang, Z.-G. Huang, J.-Q. Dong, L. Huang, and Y.-C. Lai, Phys. Rev. E **87**, 052808 (2013).

[33] J.-Q. Zhang, Z.-G. Huang, Z.-X. Wu, R. Su, and Y.-C. Lai, Sci. Rep. **6**, 20925 (2016).

[34] Z.-G. Huang, J.-Q. Zhang, J.-Q. Dong, L. Huang, and Y.-C. Lai, Sci. Rep. **2**, 703 (2012).

[35] J.-Q. Dong, Z.-G. Huang, L. Huang, and Y.-C. Lai, Phys. Rev. E **90**, 062917 (2014).

[36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998), Vol. 21.

[37] R. E. Bellman, *Dynamic Programming* (Princeton University Press, Princeton, NJ, 1957).

[38] R. S. Sutton, Mach. Learning **3**, 9 (1998).

[39] C. J. C. Watkins, Ph.D. thesis, Cambridge University, 1989.

[40] C. J. C. H. Watkins and P. Dayan, Mach. Learning **8**, 279 (1992).

[41] A. Kianercy and A. Galstyan, Phys. Rev. E **85**, 041145 (2012).

[42] S. P. Zhang, J. Q. Zhang, Z. G. Huang, B. H. Guo, Z. X. Wu, and J. Wang, Nonlin. Dyn. (2019), doi: 10.1007/s11071-018-4649-4.

[43] A. Potapov and M. K. Ali, Phys. Rev. E **67**, 026706 (2003).

[44] Y. Sato and J. P. Crutchfield, Phys. Rev. E **67**, 015206 (2003).

[45] A. Kianercy and A. Galstyan, Phys. Rev. E **88**, 012815 (2013).

[46] M. Andrecut and M. K. Ali, Phys. Rev. E **64**, 067103 (2001).

[47] R. Das and D. J. Wales, Phys. Rev. E **93**, 063310 (2016).

[48] D. Challet, M. Marsili, and A. D. Martino, Physica A **338**, 143 (2004).

[49] D. Challet, A. De Martino, M. Marsili, and I. P. Castillo, J. Stat. Mech. Theor. Exp. (2006) P03004.

[50] A. Vázquez, Phys. Rev. E **62**, R4497 (2000).

[51] S. Lee and Y. Kim, J. Kor. Phys. Soc. **44**, 672 (2004).

[52] J. Wang, C.-X. Yang, P.-L. Zhou, Y.-D. Jin, T. Zhou, and B.-H. Wang, Physica A **354**, 505 (2005).

[53] P.-L. Zhou, C.-X. Yang, T. Zhou, M. Xu, J. Liu, and B.-H. Wang, New Math. Nat. Comp. **1**, 275 (2005).

[54] Z.-G. Huang, Z.-X. Wu, J.-Y. Guan, and Y.-H. Wang, Chin. Phys. Lett. **23**, 3119 (2006).

[55] A. V. Banerjee, Q. J. Econ. **107**, 797 (1992).

[56] R. Cont and J.-P. Bouchaud, Macroecon. Dyn. **4**, 170 (2000).

[57] S. N. Ali and N. Kartik, Econ. Theor. **51**, 601 (2012).

[58] A. Morone and E. Samanidou, J. Evol. Econ. **18**, 639 (2008).

[59] X. F. Wang and G. Chen, Physica A **310**, 521 (2002).

[60] X. Li, X. Wang, and G. Chen, IEEE Trans. Circ. Syst. **51**, 2074 (2004).

[61] T. Chen, X. Liu, and W. Lu, IEEE Trans. Circ. Syst. **54**, 1317 (2007).

[62] L. Xiang, Z. Liu, Z. Chen, F. Chen, and Z. Yuan, Physica A **379**, 298 (2007).

[63] Y. Tang, Z. Wang, and J.-a. Fang, Chaos **19**, 013112 (2009).

[64] M. Porfiri and F. Fiorilli, Chaos **19**, 013122 (2009).

[65] W. Yu, G. Chen, and J. Lü, Automatica **45**, 429 (2009).

[66] N. Yao, Z. G. Huang, H. P. Ren, C. Grebogi, and Y. C. Lai, Phys. Rev. E **99**, 010201 (2019).