


## Understanding collective behaviors in reinforcement learning evolutionary games via a belief-based formalization

Ji-Qiang Zhang <sup>1,\*</sup> Si-Ping Zhang<sup>2,†</sup> Li Chen,<sup>3</sup> and Xu-Dong Liu<sup>4</sup>

<sup>1</sup>*Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China*

<sup>2</sup>*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*

<sup>3</sup>*School of Physics and Information Technology, Shaanxi Normal University, Xi'an, 710062, China*

<sup>4</sup>*Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China*



(Received 17 January 2020; accepted 24 February 2020; published 6 April 2020)

Collective behaviors by self-organization are ubiquitous in nature and human society and extensive efforts have been made to explore the mechanisms behind them. Artificial intelligence (AI) as a rapidly developing field is of great potential for these tasks. By combining reinforcement learning with evolutionary game (RLEG), we numerically discover a rich spectrum of collective behaviors—explosive events, oscillation, and stable states, etc., that are also often observed in the human society. In this work, we aim to provide a theoretical framework to investigate the RLEGs systematically. Specifically, we formalize AI-agents' learning processes in terms of belief switches and behavior modes defined as a series of actions following beliefs. Based on the preliminary results in the time-independent environment, we investigate the stability at the mixed equilibrium points in RLEGs generally, in which agents reside in one of the optimal behavior modes. Moreover, we adopt the maximum entropy principle to infer the composition of agents residing in each mode at a strictly stable point. When the theoretical analysis is applied to the  $2 \times 2$  game setting, we can explain the uncovered collective behaviors and are able to construct equivalent systems intuitively. Also, the inferred composition of different modes is consistent with simulations. Our work may be helpful to understand the related collective emergence in human society as well as behavioral patterns at the individual level and potentially facilitate human-computer interactions in the future.

DOI: [10.1103/PhysRevE.101.042402](https://doi.org/10.1103/PhysRevE.101.042402)

### I. INTRODUCTION

Emergence at the global scale from local interactions is widespread in biological systems and human society [1–6], which has attracted many researchers to comprehend from various perspectives. Many theories are developed to classify these collective behaviors, to make predictions [7–9], or to infer the local interaction rules based on the available data [10–12]. An outstanding one—the evolutionary game (EG) theory introduced in 1973 studies the destination of a given population in the ecosystem by incorporating the concept of evolution into the classic game theory [13–15]. Within this framework, various collective phenomena such as oscillating coexistence of species [16–18], rich patterns in structured population with different topologies [19–22], and so on, are well understood. One topic of particular interest is to explore the cooperation mechanism among unrelated individuals, where the population structures, self-adaptation, and social factors could all play a role [23–25].

In recent years, machine learning is becoming one of the most exciting fields [26–31] and is applicable to various tasks, like pattern recognition [28,32,33], disease prediction [34,35], decision-making, as well as human-level control [36,37], for instance. Not surprisingly, its boom has also permeated to the studies of collective behaviors—like the inference of statistical properties [38–40] and trend prediction [41–43]. While machine learning provides some statistical insights, it contributes little to the understanding of the collective behaviors from the individual level [44]. As a particularly suitable candidate, reinforcement learning (RL) is rooted in the psychological and neuroscience and is widely used for it is highly adaptable to quite different environments [29,45,46]. The marriage between RL and evolutionary game may be a proper choice to study complex behaviors. Yet there is a vacancy in between.

Inspired by this thought, we investigate the collective behaviors in the evolutionary games with the reinforcement learning (RLEGs) manner in terms of states, actions, rewards, and decision-making through exploratory trials. Specifically, our AI-agents play games with other agents and maximize their payoffs through Q-learning algorithm [47–49]. Our simulations together with our previous work [50] show that the

\*zhangjq13@lzu.edu.cn

†zhangsp15@lzu.edu.cn

reward gap between actions propels the action preferences in the population toward equilibrium points, in which agents get identical rewards. In addition, the cooperation preferences present various collective behaviors around the mixed equilibrium point, such as explosive events, oscillating or stable coexistence, etc.

A series of questions are naturally following: *Can we construct a theoretical framework for the RLEGs to explore the mechanism behind the phenomena? What is the essential difference between the RLEGs and traditional evolutionary game at the individual-level even though the collective behaviors are similar?* Addressing these questions is of paramount significance because establishing a proper framework is a critical step for a deep understanding of the systems of this sort and is the foundation for any possible application in the future, such as designing human-machine systems.

The paper is organized as follows. In Sec. II, we introduce the RLEG model by combining a reinforcement learning algorithm with the evolutionary game and provide numerical results in  $2 \times 2$  game setting. The theoretical framework is developed in Sec. III. First, the case of static environment is analyzed as the preliminary in Sec. III A; then we take a series of analyses on stability of the mixed equilibrium points and composition of AI agents at the stable points in Sec. III B. We apply the framework to  $2 \times 2$  game setting in Sec. III C and expound the various collective behaviors in simulations, and we also classify  $2 \times 2$  RLEGs according to the spectrum of collective behaviors. Furthermore, we provide the composition of agents and reveal that agents' actions are time-correlated and form various robust behavior modes. Finally, we provide a discussion and conclusion in Sec. IV.

## II. RESULTS

### A. The model for evolutionary game of reinforcement learning

In our model, the system consists of  $N$  agents empowered by the Q-learning algorithm and each is in one of the available states  $\mathcal{S} = \{s_1, \dots, s_{n_s}\}$ . For an arbitrary round  $\tau$ , a random agent  $i$  is chosen as the initiator to play a battery of pairwise games with the rest (participants). All players will take one action chosen from the action set  $\mathcal{A} = \{a_1, \dots, a_{n_a}\}$ . The payoff of the initiator  $i$  is determined by actions of its opponents and its own following the payoff matrix

$$\mathbf{\Pi} = \begin{pmatrix} \Pi_{a_1 a_1} & \cdots & \Pi_{a_1 a_{n_a}} \\ \vdots & \ddots & \vdots \\ \Pi_{a_{n_a} a_1} & \cdots & \Pi_{a_{n_a} a_{n_a}} \end{pmatrix},$$

in which subscripts denote the initiator's and its opponent's actions in games. For instance,  $i$  gets a payoff  $\Pi_{a_k a_l}$  when  $i$  and its opponent take action  $a_k$  and  $a_l$ , respectively. The average payoff for the initiator is  $\bar{\Pi}(\tau) = \sum_{j \in \Omega \setminus i} \Pi_{a_i a_j}(\tau) / (N - 1)$ , where  $\Omega \setminus i$  refers to all agents but excluding itself.

In the classical Q-learning algorithm, each agent is equipped with a time-dependent Q table based on the state (rows) and actions (columns):

$$\mathbf{Q}(\tau) = \begin{bmatrix} Q_{s_1 a_1}(\tau) & \cdots & Q_{s_1 a_{n_a}}(\tau) \\ \vdots & \ddots & \vdots \\ Q_{s_{n_s} a_1}(\tau) & \cdots & Q_{s_{n_s} a_{n_a}}(\tau) \end{bmatrix}.$$

In our model, agents seek the optimal action in the sense that it maximizes the expected reward by updating the Q table. In our setting, only the initiator updates both Q table and state at the end of each round. This setup considers the fact that initiators are actively engaged in their state and improve their wisdom (via Q table), while participants are only passively involved in the games proposed without the expectation. Note that this setting is just equivalent to asynchronous updating of Monte Carlo (MC) simulations [51,52].

During action decision stage in round  $\tau$ , if  $i$ 's current state is  $s$ , then it takes action following its Q table,

$$a(\tau) \rightarrow h[\mathbf{Q}(\tau), s(\tau)] = \arg \max_{a'} \{Q_{sa'}(\tau)\}, \quad a' \in \mathcal{A},$$

with probability  $1 - \epsilon$ , or chooses one random action with  $\epsilon$ . Here,  $\arg \max_{a'} [Q_{sa'}(\tau)]$  is the action with the maximum Q value given the current state  $s$ . For participants, they follow the same procedure by selecting actions with the largest value in the row of its current states.

In the learning process, the element  $Q_{sa}$  in  $i$ 's Q table is updated as follows:

$$Q_{sa}(\tau + 1) = Q_{sa}(\tau)(1 - \alpha) + \alpha[r(\tau) + \gamma Q_{s'a'}^{\max}(\tau)], \quad (1)$$

where  $\alpha \in (0, 1]$  is the learning rate reflecting the strength of memory effect and  $r(\tau) = \bar{\Pi}(\tau)$  is the reward received for its action. The parameter  $\gamma \in [0, 1)$  is the discount factor determining the importance of future rewards according to current Q table and  $Q_{s'a'}^{\max} = \max_{a'} (Q_{s'a'})$  is the maximum element in the row of future state  $s'$ . The evolving Q-function is  $\mathbf{Q}(\tau + 1) = g[\mathbf{Q}(\tau), r(\tau)]$ , i.e.,  $Q_{sa}(\tau)$  is replaced by  $Q_{sa}(\tau + 1)$  at the end of the round. Besides,  $i$ 's state is now replaced by its current action,  $s(\tau + 1) = a(\tau)$ . Given this fact, the round could also be called as  $i$ 's learning round.

In sum, the protocol of Q-learning in our RLEG is as Fig. 1 shows:

(1) Initialize all agents' matrix Q and state  $s$ .

(2) For each round, a randomly chosen initiator proposes a battery of games with the rest and takes the action  $a$  following  $h$  function with the largest value of  $Q_{sa'}(\tau)$  in the row of current state  $s$  with probability  $1 - \epsilon$ , or chooses a random action with probability  $\epsilon$ . The initiator then gets a reward according to its action and its opponents' actions. Meanwhile, each participant takes the action  $a$  with the largest value of  $Q_{sa'}(\tau)$  in the row of its current state  $s$  as the response.

(3) In the learning process, the initiator updates the value of  $Q_{sa}$  following Eq. (1), and the state is also updated as  $s(\tau + 1) = a(\tau)$ .

(4) Repeat the processes (2) and (3) until the system becomes statistically stable or evolves into the desired stage.

By default, we initialize the Q table and state for each agent to a null matrix and a random state, respectively.

Different from our previous work [50], we use a vector  $\mathbf{f}(\tau) = [f_{a_1}(\tau), f_{a_2}(\tau), \dots, f_{a_{n_a}}(\tau)]^T$  to character the composition of agents at a given round  $\tau$ . Here,  $f_{a_i}(\tau)$  is the preference for action  $a \in \mathcal{A}$  and is defined as

$$f_{a_i}(\tau) = \frac{\sum_{k=1}^N \delta(a_k(\tau) - a_i)}{N}, \quad (2)$$

where  $\delta = 1$  if agent  $k$ 's expected action  $a_k(\tau)$  is  $a_i$  following  $h$ -function at  $\tau$ , and  $\delta = 0$  otherwise.

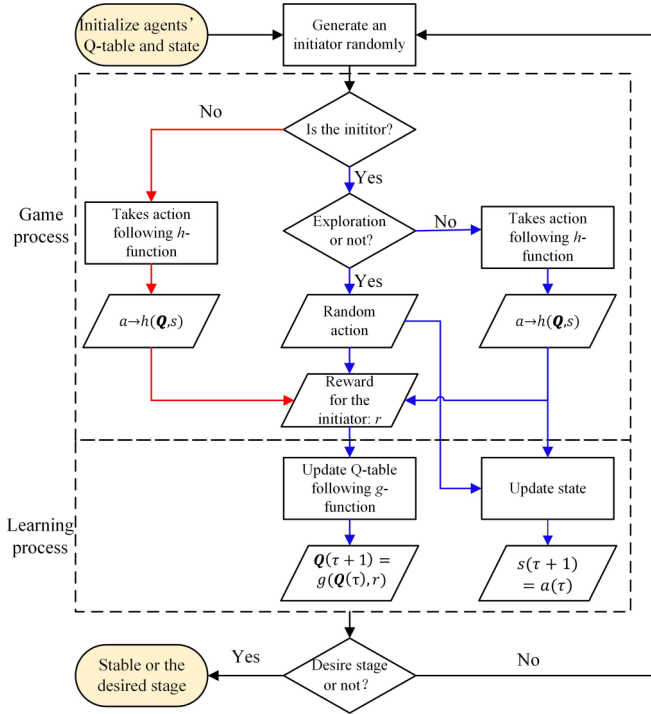


FIG. 1. The flow chart of protocol for reinforcement learning evolutionary games (RLEGs). The logic diagram for the initiator and participants are indicated by blue and red arrows, respectively.

### B. Simulation results in $2 \times 2$ RLEGs

In our simulations, we focus on the simplest RLEGs setting by adopting the  $2 \times 2$  games. For these RLEGs, each agent is equipped with a suite of state and action sets that are  $\mathcal{A} = \{C, D\}$  and  $\mathcal{S} = \{C, D\}$ , and a time-dependent  $Q$  table,

$$\mathbf{Q}(\tau) = \begin{bmatrix} Q_{cc}(\tau) & Q_{cd}(\tau) \\ Q_{dc}(\tau) & Q_{dd}(\tau) \end{bmatrix}, \quad (3)$$

which is a matrix on the Cartesian product for state (columns)–actions (rows). The reward is based on the following  $2 \times 2$  payoff matrix:

$$\mathbf{\Pi} = \begin{pmatrix} \Pi_{cc} & \Pi_{cd} \\ \Pi_{dc} & \Pi_{dd} \end{pmatrix}. \quad (4)$$

Hereafter,  $\mathbf{\Pi} = (6, b; 6 + b, 2)$  with a tunable parameter  $b$ . With this form, the game can be classified into four different classic categories by varying the value of  $b$ : (1) stag hunt (SH) for  $b \in (-\infty, 0)$ ; (2) prisoner's dilemma (PD) for  $b \in [0, 2]$ ; (3) snowdrift (SD) for  $b \in [2, 6]$ ; and (4) mixed stable (MS) for  $b \in (6, \infty)$ .

In the simulation, the systems evolves many Monte Carlo (MC) steps (denoted as  $t = 1, 2, 3, \dots$ ), and each MC step consists of  $N$  rounds. Bear in mind, for a single round, a randomly chosen initiator plays  $2 \times 2$  games with the rest, its state as well as  $Q$  table are updated. As  $f_c(t)$  denotes time series of the cooperation preference,  $\langle f_c \rangle$  is then the expected cooperation preference, which is computed after the transient and is the key quantity in previous works [14,16,20].

The simulations together with our previous work [50] show that features of the payoff matrix play a significant

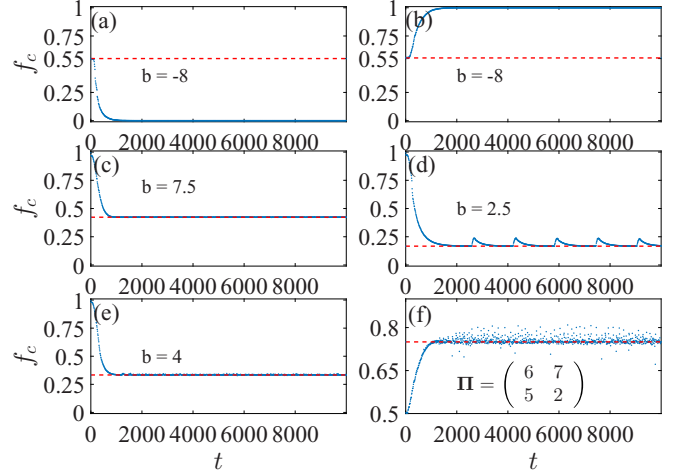


FIG. 2. The time series of  $f_c$  over MC steps in RLEGs for different game settings. Panels (a), (b) show the time series in the SH RLEGs from two different initial conditions. The time series in an MS RLEG is exhibited in panel (c), while those for SD RLEGs are provided in panels (d)–(f). The mixed and weak equilibrium point for cooperation is shown by a red dash line for each game setting. The learning parameters  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$  and the system size  $N = 10000$ .

role in the resulting cooperation preference (Fig. 2 and Appendix). Similar to the paradigmatic evolutionary game (EG), the signs of column differences  $\Delta\Pi_{:c} = \Pi_{cc} - \Pi_{dc}$  and  $\Delta\Pi_{:d} = \Pi_{dd} - \Pi_{dc}$  determine the final cooperation preference in RLEGs for most cases. For PD settings with  $\text{sgn}(\Delta\Pi_{:d}) = -\text{sgn}(\Delta\Pi_{:c}) = 1$ , defection are dominating in the system. By contrast, cooperators prevails for the Harmony game (HM) setting with  $\text{sgn}(\Delta\Pi_{:c}) = -\text{sgn}(\Delta\Pi_{:d}) = 1$  (see Fig. 10 in Appendix). For SH game settings with  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d}) = 1$ , the cooperation preference is bistable and is thus sensitive to the initial conditions as shown in Figs. 2(a) and 2(b). Interestingly, entirely different collective behaviors emerge for the MS and SD settings in RLEGs, although  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d}) = -1$  in both settings. For the former, cooperators and defectors coexist stably as in EGs [Fig. 2(c)]. For the latter, the cooperation preference  $f_c$  could present in the form of periodic oscillations. In each period, there is an explosive growth followed by a long quiescent stage [Fig. 2(d)]. But, the oscillation fades away and aperiodic oscillation appears with the increase of  $\Delta\Pi_{:d}/\Delta\Pi_{:c}$  [see Figs. 2(e) and 2(f)].

Further research shows that the low learning rate  $\alpha$  and high discounting factor  $\gamma$  promotes the increase of  $A$  and  $T$ . In addition, a high exploration rate  $\epsilon$  accelerates the increase of  $f_c$  in the explosive stage and contributes to the increase of amplitude  $A$ . Some more simulations are also conducted to study their impact of on amplitude  $A$  and period  $T$  in the periodic oscillation (see Fig. 11). For the default form in the SD area, the increase of  $b$  decreases both the amplitude and the period and makes the oscillation fade away. It shows the collective behavior is divided into oscillation and stable regions by a transition point  $b'$  (Figs. 3 and 12).

In sum, a rich spectrum of collective behaviors emerges in the RLEG by tuning the game parameters. The collective

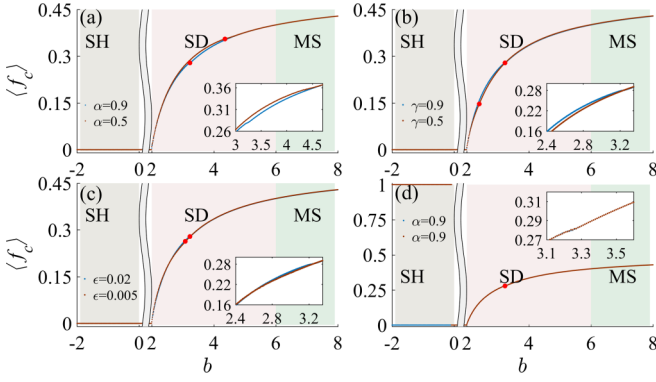


FIG. 3.  $\langle f_c \rangle$  as a function of game parameter  $b$  with different learning parameters. The transition points between periodically oscillating and stable areas are marked with a red dot for each subplot. The insets in subplots further manifest effects of parameters on the amplitude and transition point  $b'$  by zooming in. The learning parameters in each panel are as follows: (a)  $\gamma = 0.9$  and  $\epsilon = 0.02$ ; (b)  $\alpha = 0.9$  and  $\epsilon = 0.02$ ; (c)  $\alpha = \gamma = 0.9$ ; (d)  $\alpha = \gamma = 0.9$  and  $\epsilon = 0.02$ . In panel (d), a bias initialization replaces the standard initialization (orange dots). The system size  $N = 10\,000$  in all simulations.

behaviors are quite different for the SD and MS settings in RLEGs although that are similar for the traditional EGs. For the SD RLEGs, a series of behaviors appear, such as explosive events, periodic or aperiodic oscillation and stable coexistence. Moreover, those learning parameters that are uniquely present in RLEG seemly also influence the outcome considerably, which remain to be elucidated in the following.

### III. ANALYSIS

#### A. The formalization of learning dynamics in the static environment

To understand the dynamical process in RLEGs, we first focus on a simpler case where an agent lives in a static environment and formalize the corresponding learning dynamics as a preliminary step. As with agents in RLEGs, in this case each agent's state is within the state set  $\mathcal{S} = \{s_1, \dots, s_{n_s}\}$  and takes an action from the action set  $\mathcal{A} = \{a_1, \dots, a_{n_a}\}$ . The agent also maximizes its reward via reinforcement learning: updates both its  $Q$  table and its state in the process. But unlike RLEGs, the rewards for all actions are time-independent and are denoted as  $\mathcal{R} = (r_{a_1}, \dots, r_{a_{n_a}})$ .

As the protocol of RLEGs shows, the agent takes action either by following  $h$  function with probability  $1 - \epsilon$  or by choosing a random one with probability  $\epsilon$ . In our work, we term the update events if the agent adopts the action by following the former process as “freezing events” ( $f$  events), and as “melting events” ( $m$  events) if otherwise. Physically,  $m$  events can be regarded as perturbations to  $f$  events because  $(n_a - 1)\epsilon/n_a \ll 1$ . This probability is because there is a chance  $\epsilon/n_a$  for the latter scheme still taking the same actions as the former one and needs to be subtracted. Actually, the update evolution of  $Q$  table in  $f$  events and  $m$  events could be described by a double-layered directed graph, i.e., each layer of the directed graph in the case of two states and two actions is shown in Figs. 4(a) and 4(b)

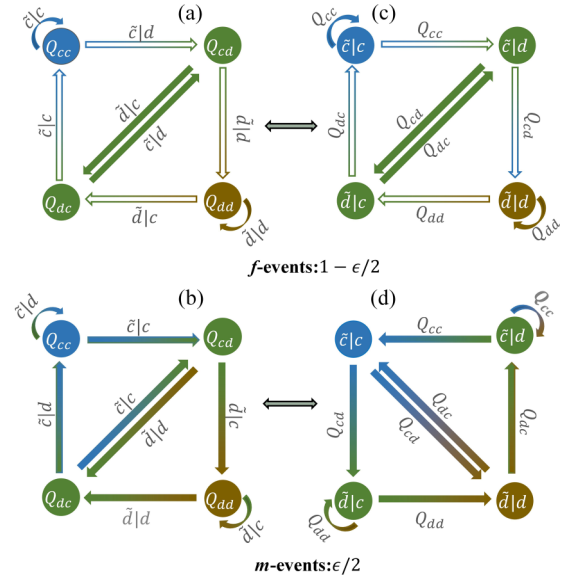


FIG. 4. The evolution scheme of  $Q$ -table update (a), (b) and the switching of visible beliefs (c), (d). These two types of schemes can be mutually derived by the interchange of their edges and vertices. In  $f$  events only, paths at the frozen point of each mode are connected with solid arrows, while paths between these modes are blocked and marked with unfilled arrows in panels (a) and (c). But these modes could be potentially melted and interchangeable in  $m$  events as panels (b) and (d) show.

In our formalization, the agent's  $Q$  table is regarded as its belief for optimal actions. The decision-making for actions indicates the robustness of the belief at a certain state  $s$  is increased with the gap between the largest and the second largest element in the row of  $s$ . Here, we formalize the agent's  $Q$  table with a time-dependent belief set  $\mathcal{B}(\tau) = \{s|a'(\tau) : s \in \mathcal{S}\}$  through a coarse method [see Fig. 5(a)]. Within the set, the element  $s|a'(\tau)$  means the agent's optimal action at state  $s$  is  $a'$  at  $\tau$ , i.e.,  $Q_{sa'}$  is the maximal element in the row of state  $s$ . However, only one belief in  $\mathcal{B}$  is visible in an  $f$  event and the rest are hidden since the agent can put itself in only one state in  $\mathcal{S}$  at any time. Thus, we denote the visible belief by  $\bar{s}|a(\tau)$ , which refers to the agent's state is  $s$  and the corresponding belief is  $s|a$  at  $\tau$ .

According to the belief set  $\mathcal{B}$ , the switch of the agents' visible belief forms a closed and uncontradicted path,  $\mathcal{B}_\mathcal{E} = (s'_1|a'_2, s'_2|a'_3, \dots, s'_{n_\mathcal{E}}|a'_1)$  with  $n_\mathcal{E} \in \mathbb{N}^*$ . Along the path, a sequence of circular and ordered actions— $\mathcal{M} = (a'_2, \dots, a'_{n_\mathcal{E}}, a'_1)$  constitute a behavior mode [see Figs. 5(a) and 5(b)]. Here, we appoint the agent is in the behavior mode  $\mathcal{M}$  when its current visible belief is in  $\mathcal{B}_\mathcal{E}$ . Besides, a mode is called an optimal mode if the reward for each action  $a'_k$  in  $\mathcal{M}$  is equal to the maximum  $r_{\max} = \max\{\mathcal{R}\}$ , otherwise, a nonoptimal mode.

For those  $f$  events, the agent's belief set and  $Q$  table will be “frozen” one after the other in such environment. The freezing rate is dependent on the learning rate  $\alpha$  and the discounting factor  $\gamma$ : a larger  $\alpha$  facilitates the freezing process, but  $\gamma$  does the opposite. We say the agent is at the frozen point of a mode when its  $Q$  table is frozen. These frozen points

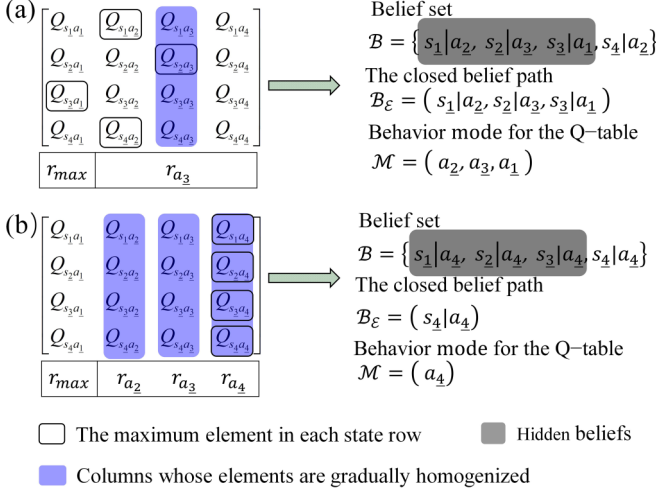


FIG. 5. The formalization of  $Q$  table under the four-states and four-actions setting. In panels (a) and (b) the agent investigated stays at state  $s_4$ . After formalization, we display the belief set  $B$ , its closed path  $B_E$ , its behavior mode  $M$  for the agent's current  $Q$  table, the table and the hidden beliefs in  $B$  with gray shadow. In panel (a), the agent is not in the behavior mode for its  $Q$  table because the *visible belief*  $s_4|a_2$  is not in  $B_E$ . In addition, the mode is unstable because the mode contains nonoptimal actions. In fact, there is unique stable mode  $M = (a_3)$  for  $r_{\max} = r_{a_3}$ . In panel (b), the agent is in the behavior mode for its current  $Q$  table but the beliefs in  $B_E$  become fragile even at the *frozen point*, because the *optimal mode* is not unique for  $r_{\max} = r_{a_2} = r_{a_3} = r_{a_4}$ .

are reminiscent of attractors in nonlinear dynamics, just the convergence rate towards these frozen points is determined by the learning parameters  $\alpha$ ,  $\gamma$  and  $\mathcal{R}$ .

Different from the noise in nonlinear dynamics, however,  $m$  events not only perturb the agent's actions but may also "melt" the agent's belief and mode at its frozen point. A frozen point is stable if all beliefs in  $B_E$  and behavior modes  $M$  are robust to any  $m$  events. However, for a given  $\mathcal{R}$ , elements in each action column tend to be exactly the same under the melting effect. Moreover, the elements in different action columns also become homogeneous over time if rewards for these actions are identical. This means the elements in the column of action  $a_i$  only depends on  $r_{a_i}$  and  $r_{\max}$ . With these, two propositions are proposed with proofs provided in Appendix A2: (1) A frozen point for a nonoptimal mode is unstable; (2) for a stable frozen point, beliefs in  $B_E$  become fragile if the optimal mode is not unique in the environment. Here it is necessary to point out that the switch of *visible belief* in  $f$  events and  $m$  events can also be described as a double-layered directed graph analogous to the update of  $Q$  table. The key is to make the interchange of edges and vertices, just as the transformation in Fig. 4

## B. The analysis in the general RLEGs

### 1. The stability for the general RLEGs

Different from the static environment, the reward for action in RLEGs is now time-dependent. As  $N \rightarrow \infty$ , the reward for action  $a_i$  is  $r_{a_i} = \sum_{a_j \in \mathcal{A}} f_{a_j}(\tau) \Pi_{a_i a_j}$  in RLEGs. For those

randomly chosen initiators, their rewards are symmetrical because the reward only depends on the current action and regardless of who takes it. Because of this symmetry and the motivation for seeking the maximal reward as shown in Sec. III A and Appendix A 2, this will narrow the rewards difference among agents gradually and push action preferences to an equilibrium point,  $\mathbf{f}^* = (f_{a_1}^*, \dots, f_{a_{n_a}}^*)^T$ . At this point, the rewards for all agents are identical in  $f$  events and no agent is able to explore a better mode to replace the current one by  $m$  events. Here, we say an equilibrium point  $\mathbf{f}^*$  is *trend stable* if the reward gap to explore the optimal mode will push action preference  $\mathbf{f}$  in its neighborhood towards it.

Given the above analysis in the static environment, we classify the equilibrium  $\mathbf{f}^*$  into two categories based on the features of optimal modes: (1) pure and (2) mixed. For the former, all agents are in the unique optimal mode consisting of the only action with the maximum reward, which indicates  $B = \{s_j|a_i : s_j \in \mathcal{S}\}$  for all agents and  $f_{a_i}^* = 1$ . For the latter, agents are in the different optimal modes consisting of one or several actions with the maximum reward. Here, one learns that  $f_{a_i}^* \in (0, 1)$  if  $a_i \in \mathcal{A}^*$  and  $f_{a_i}^* = 0$  otherwise, where  $\mathcal{A}^*$  is the set of optimal actions at  $\mathbf{f}^*$ . Furthermore, we take an equilibrium point  $\mathbf{f}^*$  is strictly stable if the system is able to suppress fluctuations at the point without delay.

We now focus on the stability of the mixed equilibrium points rather than the pure ones in RLEGs since each pure point is a fixed point and is always strictly stable [50]. To determine the properties of a mixed  $\mathbf{f}^*$ , we assume it is strictly stable firstly and then investigate whether this assumption is self-consistent. Under the assumption, our analysis in Sec. III A indicates that agents' belief set  $B$  are frozen gradually but beliefs in  $B_E$  become fragile at the frozen points. So, agents' belief is likely to be undermined by fluctuations around  $\mathbf{f}^*$ . Meanwhile, the switches of beliefs also react to these fluctuations that either amplify or suppress them. Therefore, we next pay attention to the interaction between fluctuations and *frozen* agents' beliefs through rewards in the game.

Without loss of generality, we assume that there is a fluctuation  $\delta \mathbf{f}$  with  $\delta f_{a_k} = 0, \forall a_k \notin \mathcal{A}^*$  in the neighborhood of  $\mathbf{f}^*$  at  $\tau$  (see Appendix A 3). In the round, the closed beliefs path and state for the initiator are denoted in their general form,  $B_E(\tau) = (s'_1|a'_2, s'_2|a'_3, \dots, s'_{n_E}|a'_1)$  and  $s'_i$ , in which  $n_E \in \mathbb{N}^*$  and  $s'_i|a'_j$  is in  $B_E$ . Here, the beliefs in  $B_E$  are fragile because of the narrowing of the gaps  $\delta Q = \{Q_{s'_i a'_j} - Q_{s'_i a'_l} : s'_i|a'_j \in B_E, s'_i|a'_l \notin B_E, a'_l \in \mathcal{A}^*\}$  at  $\mathbf{f}^*$ . Thus, the belief  $s'_i|a'_j$  in  $B_E$  will be undermined and replaced by a new one once the reward change by fluctuations causes any difference in  $\delta Q_{s'_i} = \{Q_{s'_i a'_j} - Q_{s'_i a'_l} : s'_i|a'_l \notin B_E, a'_l \in \mathcal{A}^*\}$  that is less than zero in the current update event.

In an  $f$  event, the update for the initiator's  $Q$  table is that

$$\begin{aligned} Q_{s'_i a'_j}(\tau + 1) &= Q_{s'_i a'_j}(\tau) + \alpha \delta r_{a'_j} \\ &= Q_{s'_i a'_j}(\tau) + \alpha \sum_{a_k \in \mathcal{A}^*} \delta f_{a_k}(\tau) \Pi_{a'_j a_k}. \end{aligned}$$

So, the probability of  $s'_i|a'_j$  being replaced by a new one increases with the reduction of reward for  $a'_j$ ,  $-\delta r_{a'_j}(\tau) = r_{a'_j}^* -$

$r_{a'_j}(\tau)$ . While for an  $m$  event, the update for the initiator's  $Q$  table is

$$\begin{aligned} Q_{s'_l a_l}(\tau + 1) &\approx Q_{s'_l a_l}(\tau) + \alpha \delta r_{a_l} \\ &= Q_{s'_l a_l}(\tau) + \alpha \sum_{a_k \in \mathcal{A}^*} \delta f_{a_k}(\tau) \Pi_{a_l a_k}, \end{aligned}$$

as the initiator's action  $a_l \in \mathcal{A}^*$ . Therefore, the transition probability of belief from  $s'_l | a'_j$  to  $s'_l | a_l$  increases with the increment of the reward for  $a_l$ ,  $\delta r_{a_l} \approx \delta r_{a_l}(\tau) - r_{a_l}^* > 0$ . As mentioned above,  $f_{a_j}$  increases but  $f_{a_l}$  decreases potentially if  $s'_l | a'_j$  is replaced by  $s'_l | a_l$  in the update events.

Based on these analysis, we reach a set of ordinary differential equations  $d\delta\mathbf{f}/d\tau = F(\delta\mathbf{f})$  with

$$\begin{aligned} \frac{d\delta f_{a_l}}{d\tau} &= \left[ \frac{1}{|\mathcal{A}^*| - 1} \cdot \sum_{a_i \in \mathcal{A}^* \setminus a_l} f_{a_i}^* \psi_{a_i}^f(-\delta r_{a_i}; \alpha, \gamma, \mathbf{\Pi}) - f_{a_l}^* \psi_{a_l}^f(-\delta r_{a_l}; \alpha, \gamma, \mathbf{\Pi}) \right] \cdot \left[ 1 - \frac{(n_a - 1)\epsilon}{n_a} \right] \\ &+ \left[ \sum_{a_i \in \mathcal{A}^* \setminus a_l} (f_{a_i}^* \psi_{a_i}^m(\delta r_{a_i}; \alpha, \gamma, \mathbf{\Pi}) - f_{a_l}^* \times \psi_{a_l}^m(\delta r_{a_i}; \alpha, \gamma, \mathbf{\Pi})) \right] \cdot \frac{\epsilon}{n_a}, \end{aligned} \quad (5)$$

where  $\psi_a^f(x; \alpha, \gamma, \mathbf{\Pi})$  and  $\psi_a^m(x; \alpha, \gamma, \mathbf{\Pi})$  are the change rates of beliefs from  $\{s_l | a : s_l \in \mathcal{S}\}$  to  $\{s_l | a' : s_l \in \mathcal{S}\}$  ( $a \neq a'$ ) in  $f$  events and  $m$  events, respectively. The analysis of belief change for the initiator indicates  $\psi_a^\mu(x; \alpha, \gamma, \mathbf{\Pi})$  increases with  $x$  if  $\text{sgn}(x) = 1$ , and zero otherwise. In addition, the change of reward  $\delta r_a$  for  $a \in \mathcal{A}^*$  is the function of the fluctuation  $\delta\mathbf{f}$ . Notice that, the first and last terms are flows between beliefs in  $f$  events and  $m$  events, respectively. The learning parameters and the payoff matrix determine initiators' feedback to fluctuations and the distribution of  $\delta Q$ .

According to Eq. (5), we can now judge the assumption is self-consistent if the dominant eigenvalue of  $DF(\delta\mathbf{f})$  at  $\mathbf{0}$  is less than zero, and not self-consistent otherwise. In other words, whether a mixed equilibrium is strictly stable in RLEG can be determined with the help of the semi-analytic equation. However, it is difficult or even impossible to get accurate  $\psi$ . Fortunately, knowing the main properties of  $\psi$  is sometimes sufficient to determine the strict stability of  $\mathbf{f}^*$  to understand qualitatively the dynamics. In fact, for an RLEG at a strictly stable mixed equilibrium point, we may get more detailed information, such as the composition of agents residing in different modes, as shown in the next subsection.

## 2. The analysis at the strictly stable equilibrium point in the general RLEGs

Based on agents' *visible belief*, we further divide agents into  $n_s \times n_a$  types,  $\Xi = \{\tilde{s} | a' : s \in \mathcal{S}, a' \in \mathcal{A}\}$ , and we use  $\mathbf{f}^\Xi = (f_{\tilde{s}_1 | a_1}, f_{\tilde{s}_1 | a_2}, \dots, f_{\tilde{s}_{n_s} | a_{n_a}})^T$  to denote the fraction of agents in these types. One can see the connection between the beliefs and behaviors is  $f_{a_l} = \sum_{s_j \in \mathcal{S}} f_{\tilde{s}_j | a_l}$ . The evolution of  $\mathbf{f}^\Xi$  is

$$\begin{aligned} \frac{d\mathbf{f}^\Xi}{d\tau} &= W(\tau) \cdot \mathbf{f}^\Xi \\ &= \left[ \left(1 - \frac{\epsilon}{n}\right) W^f(\tau) + \frac{\epsilon}{n} W^m(\tau) - I \right] \cdot \mathbf{f}^\Xi, \end{aligned} \quad (6)$$

where  $W^f$  and  $W^m$  represent transition rates due to the switches of initiator's *visible belief* in  $f$  events and  $m$  events,

respectively. For  $W^f$  and  $W^m$ , we have  $w_{\tilde{s}_i | a_l \rightarrow \tilde{s}_k | a_l}^f(\tau) = 0$  for  $s_k \neq s_j$  and  $w_{\tilde{s}_i | a_j \rightarrow \tilde{s}_j | a_k}^m(\tau) = 0$ .

For a strictly stable mixed equilibrium point  $\mathbf{f}^*$ ,  $\mathbf{f}^\Xi$  can be denoted as  $\mathbf{f}^{*\Xi}$  and reasonably be assumed fixed. But the matrix  $W$  is not fixed at the point because the degree of freedom for  $W$  is higher than the one for  $\mathbf{f}^{*\Xi}$ , i.e.,  $W$  keeping  $\mathbf{f}^\Xi$  at  $\mathbf{f}^{*\Xi}$  is not unique. To infer  $\mathbf{f}^{*\Xi}$  at  $\mathbf{f}^*$ , we define a Shannon entropy

$$S(\tau) = \sum_{\tilde{s}_i | a_j \in \Xi} f_{\tilde{s}_i | a_j}(\tau) \log [f_{\tilde{s}_i | a_j}(\tau)], \quad (7)$$

which characterizes the degree of disorder of agents' *visible belief*. Thus,  $\mathbf{f}^{*\Xi}$  could be inferred with the help of the maximal entropy principle

$$\begin{aligned} &\max\{S(W)\} \\ &\text{s.t. } \sum_{s_l \in \mathcal{S}} f_{\tilde{s}_l | a_j}^* = f_{a_j}^*, \quad \sum_{\tilde{s}_i | a_j \in \Xi} f_{\tilde{s}_i | a_j}^* = 1, \\ &\sum_{a_k \in \mathcal{A}} w_{\tilde{s}_i | a_j \rightarrow \tilde{s}_j | a_k}^{*f} = 1, \quad \sum_{\substack{s_k \in \mathcal{S} \setminus s_j \\ a_l \in \mathcal{A}}} w_{\tilde{s}_i | a_j \rightarrow \tilde{s}_k | a_l}^{*m} = 1, \\ &0 \leq f_{\tilde{s}_i | a_j}^* \leq 1, \quad 0 \leq w_{\tilde{s}_i | a_j \rightarrow \tilde{s}_k | a_l}^* \leq 1, \\ &\frac{d\mathbf{f}^\Xi}{d\tau} |_{\mathbf{f}^\Xi = \mathbf{f}^{*\Xi}} = 0. \end{aligned} \quad (8)$$

The information  $d\mathbf{f}^\Xi/d\tau = 0$  at  $\mathbf{f}^{*\Xi}$  derives from the assumption that  $\mathbf{f}^\Xi$  is fixed at  $\mathbf{f}^*$ . In specific cases, we can then obtain the proportion of agents residing in different modes.

## C. The analysis in the general RLEGs specific to 2 × 2 game setting

### 1. The stability of mixed equilibrium point

In this part, we apply the above analysis to the simplest case—2 × 2 game setting. The rewards for actions are  $r_c(\tau) = f_c(\tau)\Delta\Pi_c + \Pi_{cd}$  and  $r_d(\tau) = -f_c(\tau)\Delta\Pi_d + \Pi_{dd}$ , where  $\Delta\Pi_c = \Pi_{cc} - \Pi_{cd}$  and  $\Delta\Pi_d = \Pi_{dd} - \Pi_{dc}$ . Besides, there is at most one mixed equilibrium point  $\mathbf{f}^* = (f_c^*, f_d^*)^T =$

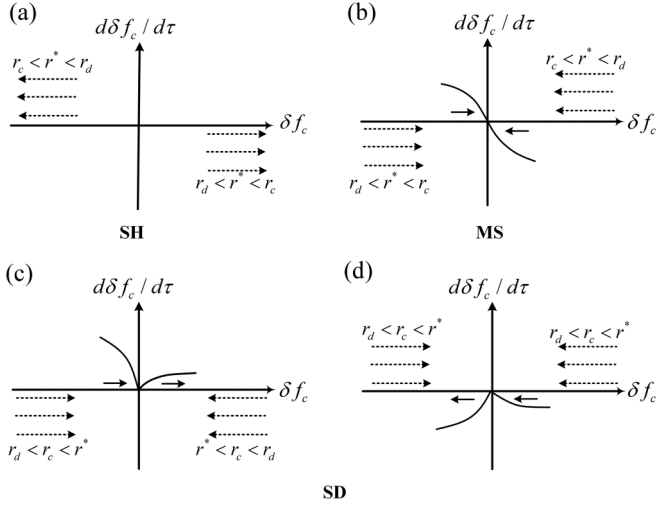


FIG. 6. The mechanism diagram of the stability for the mixed equilibrium points in  $2 \times 2$  RLEGs. In panels (a)–(d), the feedback to reward gap would push  $f_c$  toward or away from the mixed equilibrium point  $f_c^*$  (dotted arrows), which determines the trend stability of  $f_c^*$ . The feedback to different fluctuations around a trend stable  $f_c^*$  would take a driving force in the same or opposite direction (solid arrows), which decides the strict stability of  $f_c^*$ . In panel (a), the mixed equilibrium point is unstable in trend for SH RLEGs since the reward gap always drives cooperation preference away from it. Panel (b) shows the mixed equilibrium point are both trend stable and strictly stable for MS RLEGs. In panels (c), (d), we show two cases that the mixed equilibrium point is trend stable but not strictly stable. In case (c),  $\Delta\Pi_{:d}/\Delta\Pi_{:c} < \psi_d^m/\psi_c^m$ , the point is strictly stable in the left neighborhood but not in the right neighborhood. A periodic oscillation emerges in the right neighborhood. For case (d),  $\Delta\Pi_{:d}/\Delta\Pi_{:c} > \psi_d^f/\psi_c^f$ , the point is strictly stable in the right neighborhood but not in the left neighborhood. There, the equilibrium always breaks down from the left neighborhood and an aperiodic oscillation arises.

$(\frac{\Delta\Pi_{:d}}{\Delta\Pi_{:c} + \Delta\Pi_{:d}}, \frac{\Delta\Pi_{:c}}{\Delta\Pi_{:c} + \Delta\Pi_{:d}})^T$ , at which the rewards for cooperators and defectors are identical. Thus, the existence condition for the point is that  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d})$  because  $f_c^* \in (0, 1)$ . As mentioned before, the reward gap between agents pushes  $\mathbf{f}$  toward one of equilibrium points, and the gap is erased. In addition, the gap also drives agents to take superior action or mode by learning. In the following, we first analyze the trend stability of mixed equilibrium points.

Based on the property of columns in the payoff matrix, we divide settings with a mixed equilibrium point  $f_c^*$  into two classes:  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d}) = 1$  and  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d}) = -1$ . For the former, the environment favors cooperators rather than defectors under  $f_c \in (f_c^*, 1)$  because  $r_c > r_d$ . As a result, the reward gap will promote the cooperator prevalence further. In the opposite case of  $r_d > r_c$  under  $f_c \in [0, f_c^*)$ , defectors are favored and cooperators decreases. Accordingly, the reward gap always drives  $f_c$  away from  $f_c^*$ . Therefore, the mixed equilibrium point is trend unstable, such as SH game settings [see Figs. 2(a), 2(b), and 6(a)]. Comparatively for the later case, the environment favors defectors over cooperators when  $f_c \in (f_c^*, 1)$ , and the cooperators are favored when  $f_c \in [0, f_c^*)$ . It suggests the reward gap pushes  $\mathbf{f}$  toward the mixed equilibrium point, and the mixed

equilibrium point is trend stable, such as MS and SD game settings [see Figs. 2(c)–2(f) and Figs. 6(b)–6(d)].

But a mixed equilibrium point could be not strictly stable. Here, we investigate the strict stability of  $\mathbf{f}^*$  by examining the dynamics of fluctuations in Eq. (5). By normalizing  $\mathbf{f}$ , the evolution of  $\delta f_c$  is

$$\begin{aligned} \frac{d\delta f_c}{d\tau} = & \frac{\epsilon}{2} [f_d^* \psi_d^m(\delta r_c; \alpha, \gamma, \mathbf{\Pi}) - f_c^* \psi_c^m(\delta r_d; \alpha, \beta, \mathbf{\Pi})] \\ & + \left(1 - \frac{\epsilon}{2}\right) [f_d^* \psi_d^f(-\delta r_d; \alpha, \beta, \mathbf{\Pi}) \\ & - f_c^* \psi_c^f(-\delta r_c; \alpha, \beta, \mathbf{\Pi})], \end{aligned} \quad (9)$$

in which  $\delta r_c = \delta f_c \cdot \Delta\Pi_{:c}$  and  $\delta r_d = -\delta f_c \cdot \Delta\Pi_{:d}$ . Besides,  $\psi$  subject to the distribution of  $\delta Q$  with  $\psi_c^m = \psi_d^f$  and  $\psi_c^f = \psi_d^m$  (Figs. 13–15 in Appendix A 3). In most cases, its distribution is approximately symmetric at the point, especially when  $\Delta\Pi_{:c} = \Delta\Pi_{:d}$ .

Equation (9) shows that the properties of rows for the payoff matrix pose great influence on the strict stability. For the MS game with  $\text{sgn}(\Delta\Pi_{:c}) = \text{sgn}(\Delta\Pi_{:d}) = -1$ , the agents' belief change will suppress fluctuations in the neighborhood of  $\mathbf{f}^*$  without delay since the signs of  $\delta f_c$  and  $d\delta f_c/d\tau$  are opposite in Eq. (9) [Figs. 6(b), 13(a), 13(b), and 14]. Thus, the mixed  $\mathbf{f}^*$  in RLEGs for the MS game setting is both stable in trend and strictly stable [Figs. 2(a) and (3)].

For SD games with  $\text{sgn}(-\Delta\Pi_{:d}) = \text{sgn}(\Delta\Pi_{:c}) = 1$ , we focus on two specific cases: (i)  $f_c^*/f_d^* = \Delta\Pi_{:d}/\Delta\Pi_{:c} < \psi_d^m/\psi_c^m$  as  $\delta f_c > 0$  and (ii)  $f_c^*/f_d^* > \psi_d^f/\psi_c^f$  as  $\delta f_c < 0$ . For case (i), the fluctuation  $\delta f_c \rightarrow 0_-$  will be suppressed without delay because  $d\delta f_c/d\tau > 0$  [see Fig. 6(c)]. Therefore, the mixed equilibrium point is strictly stable in the left neighborhood of  $f_c^*$ . But the change of frozen initiators' belief caused by the fluctuation  $\delta f_c \rightarrow 0_+$  is to amplify it further since  $\delta f_c$  and  $d\delta f_c/d\tau$  are of the same sign [see Figs. 13(c) and 15]. The enhancement of the fluctuations will cause more agents to change their fragile beliefs and the system goes into an “explosive stage” by cascades. In this stage, the increasing rate of cooperators increases with  $\epsilon/2$  as Eq. (9) shows. But the increase of  $r_d - r_c$  in the meanwhile will push  $\mathbf{f}$  back to  $\mathbf{f}^*$  and enter the “quiescent stage.” Again, the agents' beliefs become fragile gradually, later on, a periodic oscillation is formed [Fig. 2(d)]. The above analysis also explains why the periodic oscillation fades away with the increase of  $b$  (Figs. 3 and 12) and the presence of the point  $b'$  separating oscillation from nonoscillation areas for the SD RLEGs (Fig. 3).

For case (ii), a fluctuation  $\delta f_c \rightarrow 0_+$  will be suppressed without delayed time because  $\delta f_c$  and  $d\delta f_c/d\tau$  are of opposite sign as Fig. 6(d) shows. Thus, the mixed equilibrium point is stable in the right neighborhood of  $f_c^*$ . However, a fluctuation  $\delta f_c \rightarrow 0_-$  could be increased further since  $\delta f_c$  and  $d\delta f_c/d\tau$  are of the same sign [see Fig. 13(d)]. Analogously, the increase of the fluctuation cause more agents to change their beliefs in this delayed stage, and the cooperators decrease rapidly because  $1 - \epsilon/2 \gg \epsilon/2$ . Meanwhile, there is a quite high increase of  $r_c - r_d$ , which result in the a drastic promotion of cooperation again after this stage. As  $f_c > f_c^*$ , the system enters the second delayed stage because  $r_d > r_c$ . Finally, the reward difference between actions pushes  $f_c$  back to  $f_c^*$ . In brief, the fluctuation goes through two delayed

TABLE I. The classification of  $2 \times 2$  RLEs according to properties of payoff matrix and the spectrum of collective behaviors. In the table, the properties of the payoff matrices in RLEs are shown in the first two rows, while game types, equilibrium points, and the spectrum of collective behaviors in the following rows, respectively. The connection between the spectrum and payoff matrix demonstrates that the columns determine the position of equilibrium points and their trend stability, while their strict stabilities are up to the rows.

$\text{sgn}(\Delta\Pi_{:c} \cdot \Delta\Pi_{:d}) = -1$			$\text{sgn}(\Delta\Pi_{:c} \cdot \Delta\Pi_{:d}) = 1$		
			$\text{sgn}(\Delta\Pi_{:c}) = -1$		
$\text{sgn}(\Delta\Pi_{:c}) = \text{lsgn}(\Delta\Pi_{:c}) = -\text{lsgn}(\Delta\Pi_{:c}) = 1$			$\text{sgn}(\Delta\Pi_{:c}) = 1, \text{sgn}(\Delta\Pi_{:d}) = -1$		$\text{sgn}(\Delta\Pi_{:c}) = -1, \text{sgn}(\Delta\Pi_{:d}) = -1$
HM	PD	SH	SD		MS
$f_c^* = 1$	$f_c^* = 0$	$cf_c^* = 0, 1$	$f_c^* = \frac{\Delta\Pi_{:d}}{\Delta\Pi_{:d} + \Delta\Pi_{:c}}$		
Stable	Stable	or $\frac{\Delta\Pi_{:d}}{\Delta\Pi_{:d} + \Delta\Pi_{:c}}$	$\frac{\Delta\Pi_{:d}}{\Delta\Pi_{:c}} < \frac{\psi_d^m}{\psi_c^m}$	$\frac{\psi_d^m}{\psi_c^m} < \frac{\Delta\Pi_{:d}}{\Delta\Pi_{:c}} < \frac{\psi_d^f}{\psi_c^f}$	$\frac{\Delta\Pi_{:d}}{\Delta\Pi_{:c}} > \frac{\psi_d^f}{\psi_c^f}$
		Bistable	Periodic Oscillation	Stable	Aperiodic Oscillation
					$f_c^* = \frac{\Delta\Pi_{:d}}{\Delta\Pi_{:d} + \Delta\Pi_{:c}}$
					Stable

stages before returning to  $f_c^*$  and an aperiodic oscillation emerges [Fig. 2(f)]. According to above analysis, we learn that there is a nonoscillating area between (i) and (ii), in which  $\psi_d^m/\psi_c^m < f_c^*/f_d^* < \psi_d^f/\psi_c^f$  [see Fig. 2(e)]. Summarizing our analysis and previous work [50], we further classify RLEs for the  $2 \times 2$  game settings into several classes according to the spectrum of collective behaviors and the properties of payoff matrix as shown by Table I.

To explore the impact of learning parameters on the amplitude and the period, it is necessary to discuss what causes the delay effects and their influence. In fact, after the cooperation preference relaxes to the mixed equilibrium point  $f_c^*$ , nearly all AI agents approach  $r_c^* = r_d^* = r^*$ , where a fluctuation  $\delta f_c$  can change both  $r_c$  and  $r_d$  by then. But, the initiator is unable to get the reward change for both actions in its learning round because it's impossible to get the reward change for both actions since the initiator can only take either cooperation or defection. As a result, it will take a delay effect on exploration of optimal beliefs for agents if  $\text{sgn}(r_c - r_d) = -\text{sgn}(r_c - r_d^*)$  or  $\text{sgn}(r_d - r_c) = -\text{sgn}(r_d - r_c^*)$ , such as SD RLEs [Figs. 6(c) and 6(d)]. Therefore, long memory and long-term vision for agents magnify the impact of past  $Q$  table so as to strengthen the delay effect confronting the environmental change. It suggests that a low  $\alpha$  or a high  $\gamma$  increases the amplitude and period as Fig. 11 shows. In contrast, there is no delay effect under  $\text{sgn}(r_c - r_d) = \text{sgn}(r_c - r_d^*)$  or  $\text{sgn}(r_d - r_c) = \text{sgn}(r_d - r_c^*)$  and the mixed equilibrium point is strictly stable [see Fig. 6(b)], such as MS RLEs.

$r_c) = \text{sgn}(r_d - r_c^*)$  and the mixed equilibrium point is strictly stable [see Fig. 6(b)], such as MS RLEs.

In the paradigmatic EGs, the evolution of cooperation preference only depends on the ratio  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$  and sign of  $\Delta\Pi_{:a_i} (\forall a_i \in \mathcal{A})$  [13,14]. Our simulations and analysis, however, indicate that the evolution of RLEs is not only subject to the columns of  $\mathbf{\Pi}$  but also to its rows: the ratio  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$  and the sign of  $\Delta\Pi_{:a_i}$  determine the position of equilibrium points and their trend stability, while the properties of rows affect the agents' response to fluctuations. According to Eq. (9), we conjecture that the response depends on the ratio  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$  and the signs of  $\Delta\Pi_{:s_i} (s_i \in \mathcal{S})$  analogously. If this is true, then any two systems that have identical  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$ ,  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$ ,  $\text{sgn}(\Delta\Pi_{:a_i})$ , and  $\text{sgn}(\Delta\Pi_{:s_i})$  should be qualitatively equivalent. To check the conjecture, we compute the ensemble average of time series of  $w_{\tilde{s}|a \rightarrow \tilde{s}'|a'}^\mu$  and  $f_{\tilde{s}|a}$  with  $\tilde{s}|a \in \Xi$  in a series equivalent MS and SD RLEs by simulations, in which the superscript  $\mu \in \{f, m\}$  refers to  $f$  events or  $m$  events. The results support our conjecture since they exhibit the same dynamics (see Figs. 7 and 17).

## 2. The analysis for MS RLEs

Here, we apply the beliefs formalization in Sec. III B 2 to MS RLEs to infer  $\mathbf{f}^{\Xi}$  with the help of the maximum entropy principle and the known information as Eq. (8) shows. In this case, the master equation of  $\mathbf{f}^{\Xi}$  is expressed as

$$\frac{d\mathbf{f}^{\Xi}}{d\tau} = W(\tau) \cdot \mathbf{f}^{\Xi} = \left[ \left(1 - \frac{\epsilon}{2}\right) W^f(\tau) + \frac{\epsilon}{2} W^m(\tau) - I \right] \cdot \mathbf{f}^{\Xi} \quad (10)$$

$$= \begin{pmatrix} \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{c}|c \rightarrow \tilde{c}|c}^f(\tau) - 1 & \frac{\epsilon}{2} w_{\tilde{c}|d \rightarrow \tilde{c}|c}^m(\tau) & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{d}|c \rightarrow \tilde{c}|c}^f(\tau) & \frac{\epsilon}{2} w_{\tilde{d}|d \rightarrow \tilde{c}|c}^m(\tau) \\ \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{c}|c \rightarrow \tilde{c}|d}^f(\tau) & \frac{\epsilon}{2} w_{\tilde{c}|d \rightarrow \tilde{c}|d}^m(\tau) - 1 & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{d}|c \rightarrow \tilde{c}|d}^f(\tau) & \frac{\epsilon}{2} w_{\tilde{d}|d \rightarrow \tilde{c}|d}^m(\tau) \\ \frac{\epsilon}{2} w_{\tilde{c}|c \rightarrow \tilde{d}|c}^m(\tau) & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{c}|d \rightarrow \tilde{d}|c}^f(\tau) & \frac{\epsilon}{2} w_{\tilde{d}|c \rightarrow \tilde{d}|c}^m(\tau) - 1 & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{d}|d \rightarrow \tilde{d}|c}^f(\tau) \\ \frac{\epsilon}{2} w_{\tilde{c}|c \rightarrow \tilde{d}|d}^m(\tau) & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{c}|d \rightarrow \tilde{d}|d}^f(\tau) & \frac{\epsilon}{2} w_{\tilde{d}|c \rightarrow \tilde{d}|d}^m(\tau) & \left(1 - \frac{\epsilon}{2}\right) w_{\tilde{d}|d \rightarrow \tilde{d}|d}^f(\tau) - 1 \end{pmatrix} \cdot \begin{pmatrix} f_{\tilde{c}|c} \\ f_{\tilde{c}|d} \\ f_{\tilde{d}|c} \\ f_{\tilde{d}|d} \end{pmatrix},$$

in which elements in  $W^f$  and  $W^m$  are transition rate because of the switches of the *visible belief* in update events (Fig. 4). To proceed, it is necessary to compare  $\mathbf{f}^{\Xi}$  and  $\mathcal{S}$  by inference with those by the simulations. Interestingly, an identical result is shown by the method for different payoff matrices but with

an identical equilibrium point. Since Eq. (9) shows that the difference between  $\psi_c$  and  $\psi_d$  is increased with the difference between  $\Delta\Pi_{:c}$  and  $\Delta\Pi_{:d}$  (see Fig. 14), we therefore focus on the case of  $\Delta\Pi_{:c} = \Delta\Pi_{:d}$  first. The comparison between ensemble average of time series for  $f_{\tilde{s}|a}$  as well as  $\mathcal{S}$  and those



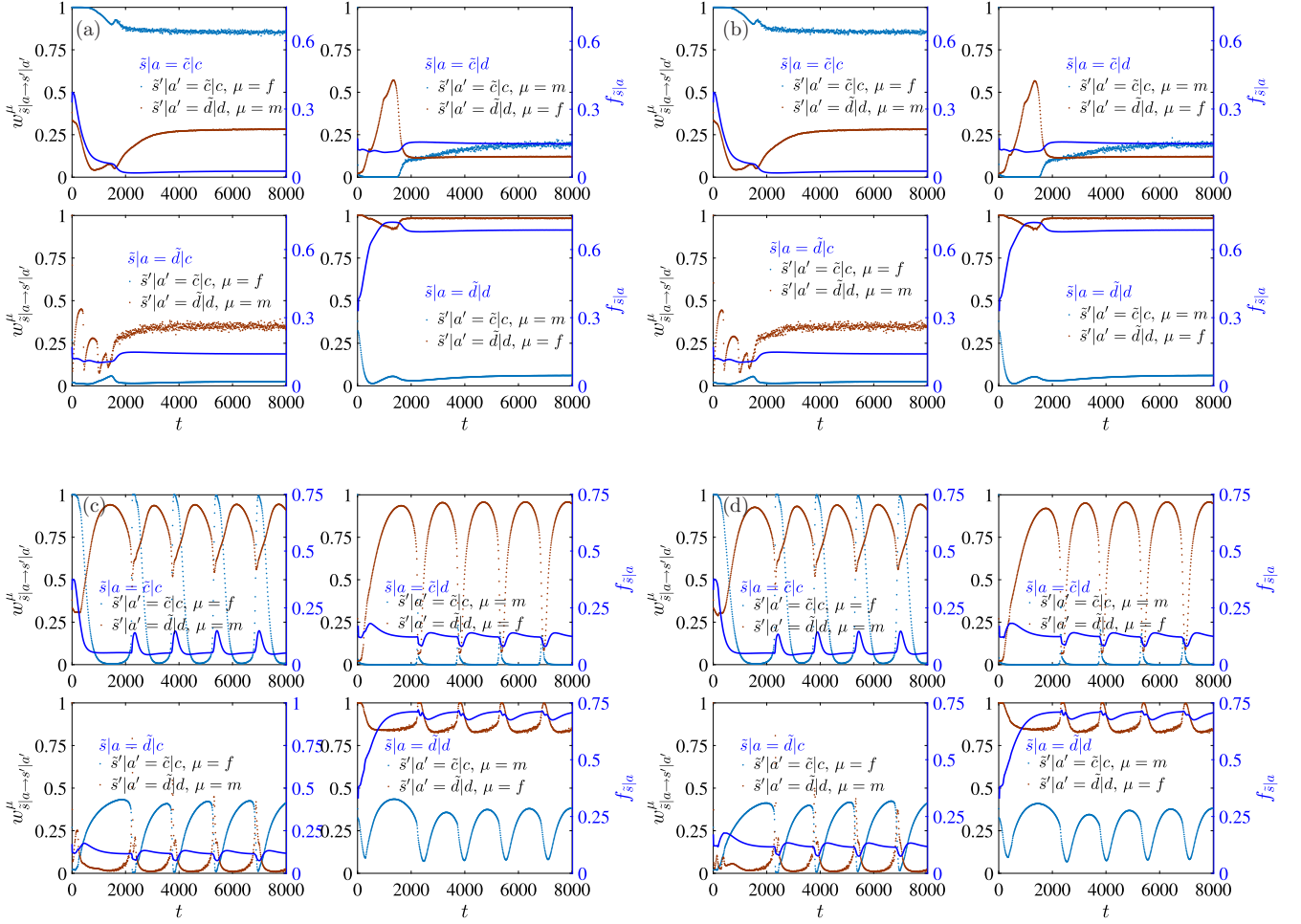


FIG. 7. The ensemble averaged time series of  $W_{s|a \rightarrow s'|a'}^\mu$  and  $f_{s|a}$  in equivalent MS and SD RLEGs. In panels (a) and (b), the payoff matrices for the equivalent MS RLEGs are (1, 4; 6, 3) and (5, 14; 20, 11), respectively. In panels (c) and (d), the payoff matrices for SD game settings take default form (6,  $b$ ; 6 +  $b$ , 2) with  $b = 2.5$  and (21, 14; 26, 13), respectively. Time series of  $w_{s|a \rightarrow \bar{c}|c}^\mu$ ,  $W_{s|a \rightarrow \bar{d}|d}^\mu$ , and  $f_{s|a}$  are shown in panels (a)–(d). Other parameters:  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$ , and the system size  $N = 10000$ . In panels (a)–(d), the ensemble average is over 1000 realizations.

by inference at  $f_c^*$  shows that the performance by our method is decent under various payoff matrices and exploration rates as Figs. 8(a)–8(c) show. However, the results by inference deviate from the simulations slightly when  $\Delta \Pi_c \neq \Delta \Pi_d$ : as Fig. 8(d) shows. The reason for the mismatch may be that the constraints on  $W$  demand the feedback is symmetric in our method.

At the mixed point of an MS RLEG, agents reside one of following three behavior modes: frozen cooperation in the form of C-C mode (CCM), frozen defection in the form of D-D mode (DDM), and cyclic mode in the form of cyclic C-D mode (CDM). Since  $\mathbf{f}^{*\varepsilon}$  reflects the fraction of agents in different modes, the fraction of agents in CCM and DDM equal to  $f_{\bar{c}|c}$  and  $f_{\bar{d}|d}$ , and the fraction in CDM approaches to the sum of  $f_{\bar{c}|d}$  and  $f_{\bar{d}|c}$ . In addition,  $W^f$  and  $W^m$  are correlated with the average residence time of agents in modes at the equilibrium point. Unfortunately, the matrices are not unique and hard to infer via our method (see Fig. 16). Therefore, we calculate the agent's mode residence over its learning round by simulations and compute the migrating rates between different modes at  $\mathbf{f}^*$  (Fig. 9 shows). Very different from

the disordered behaviors for MS EGs at the fixed point, the agents' action exhibit a remarkably high time correlation via behavior modes [see Fig. 9(a)]. In addition, the migration rate indicates each mode is robust in  $f$  events even the mode is rare, such as CCM [see Fig. 9(b)].

#### IV. DISCUSSION AND CONCLUSION

In this work, we developed a theoretical framework to understand the collective behaviors in the reinforcement learning evolutionary games (RLEGs). Within this framework, we formalize each agent's  $Q$  table in the learning as beliefs of the optimal action at different states. The series of agents' actions following their beliefs form different behavior modes. As a preliminary step, we investigate a single agent in a time-independent environment and find two useful propositions: (i) each nonoptimal mode is unstable and (ii) an optimal mode becomes fragile if the optimal is not unique.

Along with the above clues, we reveal that the reward gap in RLEGs between actions drive the action preferences toward one of equilibrium points, where the reward gap between

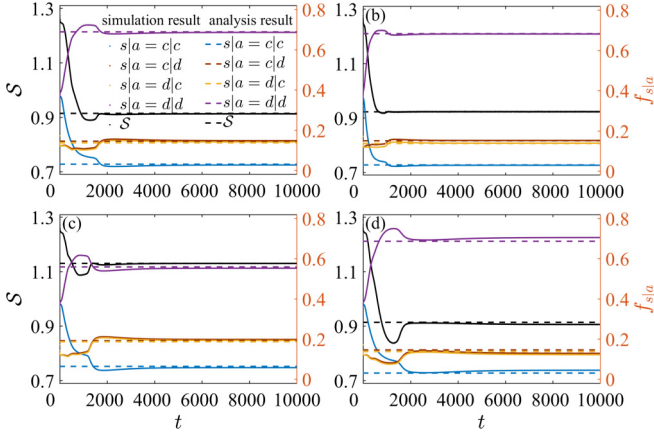


FIG. 8. The comparison between  $f_{s|a}^*$  as well as  $S^*$  in inference and ensemble average of time series of  $f_{s|a}$  as well as  $S$  in simulation. In panel (a), the payoff matrix  $\Pi = (0, 3; 5, 2)$  and learning parameters  $\alpha = \gamma = 0.9$  as well as  $\epsilon = 0.02$ . Compare with panel (a), the difference is that exploration rate is replaced with  $\epsilon = 0.04$  in panel (b), while alternative payoff matrices are  $(1, 5; 7, 3)$  and  $(0, 2.5; 2.5, 2)$  in panels (c) and (d). In the settings, the ratio  $\Delta\Pi_{c:}/\Delta\Pi_{d:} = 1$  is the same but the equilibrium point is different in panels (a) and (c). On the contrary, there is an identical equilibrium point but different  $\Delta\Pi_{c:}/\Delta\Pi_{d:}$  in panels (a) and (d). The scale of each system is  $N = 10000$  and the ensemble average is over 100 realizations in the simulation.

agents disappears. Furthermore, we uncover that the reward gap in the neighbourhood of an equilibrium point determines its trend stability. We also provide a semianalytic equation to analyze whether a mixed equilibrium point is strictly stable or not when it is trend stable. This equation helps understand the various collective behaviors in the simulation, such as

explosive events, different oscillation, and bistable states, etc. As the equilibrium point is strictly stable, we combine the maximum entropy principle with dynamics to infer the composition of agents from the perspective of modes. When applied to  $2 \times 2$  game settings, we find that the columns of payoff matrix in the game setting determine the position of equilibrium points and their trend stability; The rows determine the instantaneous response of fluctuations for agents thus influence the strict stability. Inspired by the above analysis, we propose an intuitive method to construct equivalent  $2 \times 2$  RLEs, with which the inferred fraction of agents residing in different modes is verified by numerical experiments. The series of modes residence for individuals indicates its actions are correlated in time, which is significantly different from the disordered behaviors in traditional evolutionary games.

Our work could provide a theoretic foundation for further systematic investigation of evolutionary games from the perspective of machine learning. It may also help us further understand explosive events and the related human behavior patterns in the real world since the reinforcement learning mimics the introspectiveness of human. Obviously, there still many open questions remain. Since human strategies are a mixture of several pure strategies, it would be very interesting to extend our theory of reinforcement learning evolution games to continuous game settings. Besides, how to utilize our theory to detect explosive events in the society, how to catalyze the beneficial ones among them, on one hand, and how to suppress the harmful ones, on the other hand.

## ACKNOWLEDGMENTS

L.C. is supported by the National Natural Science Foundation of China under Grant No. 61703257 and by the Fundamental Research Funds for the Central Universities Grant No. GK201903012. J.-Q.Z. and S.-P.Z. are supported by the National Natural Science Foundation of China under Grant No. 61703257 and Grants No. 61977012, No. 11775101, respectively.

## APPENDIX

This Appendix contains additional remarks and supporting materials for the main texts. To further manifest the spectrum of collective behaviors in  $2 \times 2$  RLEs, Appendix appends the time series of the cooperation preference for various game settings with different learning parameters. In Appendix A 1, we provide the proofs for the propositions (i) and (ii) of analysis regarding learning dynamics in the static environment (see Sec. III A). Furthermore, Appendix A 3 as additional remarks in Secs. III B 1 and III C 1 expounds the connection between function  $\psi$  and distribution of  $\delta Q$ . In Appendix A 4, we provide more simulations to support our remarks or conjecture about transition matrix  $W$  in Secs. III B 2 and III C 2.

### 1. More simulations for $2 \times 2$ RLEs

Here, we first provide more times series of cooperation preference in RLEs for various  $2 \times 2$  game settings for Sec. II B. Figure 10 shows that defection is dominant for the prisoner's dilemma (PD) game setting, while cooperation wins over for the harmony (HM) game setting. These trivial

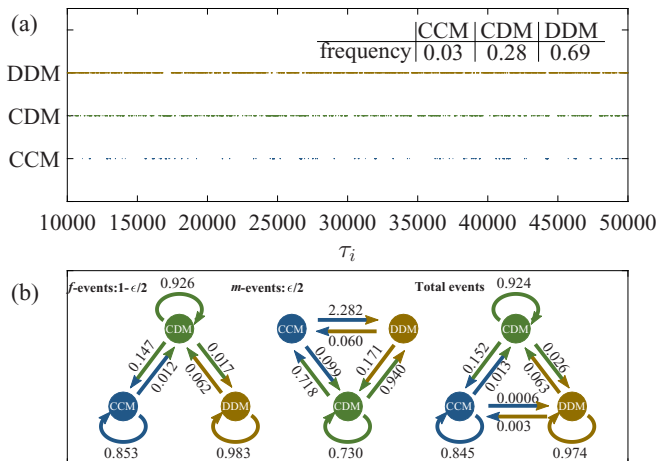


FIG. 9. The series of modes residence for a focused agent over its learning round and migration rates between various modes at the equilibrium point for an MS RLEG. In panel (a), we provide the series and frequencies for the focused agent in different modes over update events. In the figure,  $\tau_i$  is the focus agent's learning round. The migration rate between the modes in  $f$  events,  $m$  events, and total events are shown in panel (b). The simulation shares the payoff matrix and learning parameters with Fig. 8(a).

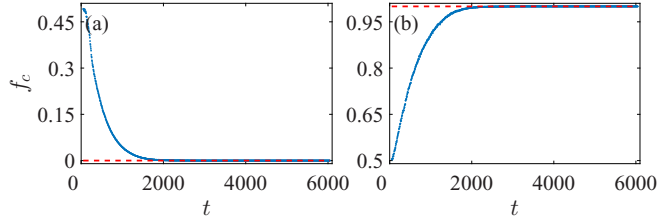


FIG. 10. The time series of cooperation preference over MC steps in reinforcement learning evolutionary games (RLEGs) for PD and HM game setting. The payoff matrices are  $\Pi = (6, b; 6 + b, 2)$  with  $b = 1.5$ , and  $(5, 2; 4, 1)$  in panels (a) and (b), respectively. The learning parameters  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$  and the system size  $N = 10000$ .

results show that the collective behaviors in RLEGs are the same as the cases in traditional evolutionary game (EG) if there is a single pure equilibrium point for the game setting. These points are the globally stable fixed points, where all agents'  $Q$  table are exactly the same and thus of less interest.

Different from the PD and HM game cases, the collective behaviors are complex and diverse in the  $2 \times 2$  RLEGs with a mixed equilibrium, especially for the snowdrift (SD) games (see Figs. 2 and 3). Figure 2 shows that three kind of collective behaviors around the equilibrium point for SD game settings, periodic oscillation, aperiodic oscillation, and stable coexistence. Here, we make more simulations to investigate how the learning parameters affect the properties of periodic oscillation (see Fig. 11). The results show that a lower learning rate (long memory effect) or a higher discounting factor (long-sight) increases the period and the amplitude of the oscillation. Therefore, the transition point between periodic oscillation and stable coexistence may increase with  $\alpha$ , while it decreases with the increasing  $\gamma$  (see Fig. 3). In addition, the increasing rate of cooperators during the *explosive stage* as well as the amplitude  $A$  increases with exploration rate  $\epsilon$ , but the period is reduced.

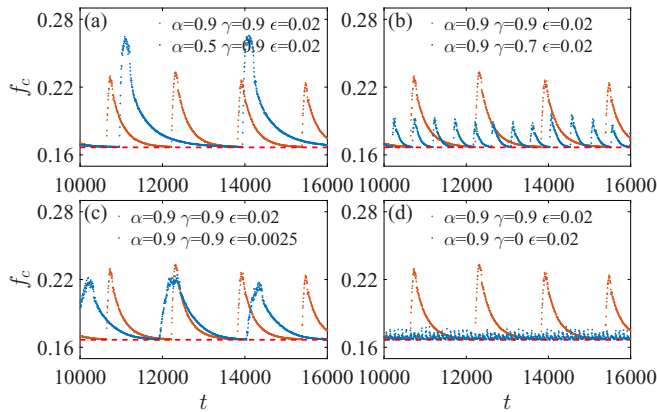


FIG. 11. The comparison of time series of  $f_c$  over MC steps in RLEGs for a SD game setting using different learning parameters. In the SD game, the payoff matrix  $\Pi = (6, b; 6 + b, 2)$  with  $b = 2.5$ . The learning parameters are shown in each subfigure and  $N = 10000$ . The red dashed lines are the fraction of cooperators at the fixed point in the traditional EGs.

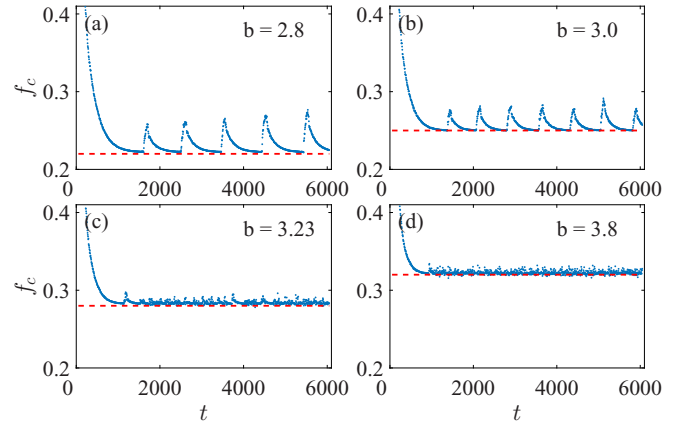


FIG. 12. Time series of  $f_c$  over MC steps in RLEGs for SD game setting with the increase of  $b$ . In the simulations, the payoff matrix take the default form,  $\Pi = (6, b; 6 + b, 2)$ , with  $b$  being shown in each panel. The learning parameters  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$ , and  $N = 10000$ . The red dashed line is the fraction of cooperators at the fixed point in the corresponding traditional EGs.

To check whether the RLEGs for default SD game settings are indeed divided into two types—periodic oscillation and stable coexistence, we provide some further time series of  $f_c$  with the increase of  $b$  in Fig. 12. The results display that the periodic oscillation fades away and turns into more or less stable as  $b$  increases. This thus supports our above two-type conjecture, and they are separated by a transition point  $b'$ . Similar to the period  $T$  and the amplitude  $A$ , the transition point is also influenced by learning parameters: higher  $\alpha$ , lower  $\gamma$  and  $\epsilon$  make RLEGs enter stable area early with the increase of  $b$ .

## 2. The analysis of learning dynamics in a static environment within belief formalization

Here, we give the proofs to the two propositions in Sec. III A: (i) *A frozen point for a nonoptimal mode is unstable;* (ii) *for a stable frozen point, beliefs in  $\mathcal{B}_{\mathcal{E}}$  become fragile if the optimal mode is not unique in the environment.*

To certify the proposition (i) with the reduction to absurdity, we assume the frozen point for a behavior mode  $\mathcal{M} = (a'_2, \dots, a'_{n_{\mathcal{E}}}, a'_1)$  is stable and the rewards for actions in  $\mathcal{M}$  meet one of following conditions: (a) there are  $a'_i$  and  $a'_j$  having  $r_{a'_j} < r_{a'_i} = r_{\max}$  or (b) any  $a'_i$  in  $\mathcal{M}$  with  $r_{a'_i} < r_{\max}$ . According to  $\mathcal{B}_{\mathcal{E}}$  for the mode  $\mathcal{M}$ , one learns that elements in the agent's  $Q$  table have

$$\begin{pmatrix} Q_{s'_1 a'_2} \\ Q_{s'_2 a'_2} \\ \vdots \\ Q_{s'_{n_{\mathcal{E}}} a'_1} \end{pmatrix} = \begin{pmatrix} 1 - \alpha & \alpha\gamma & \cdots & 0 \\ 0 & 1 - \alpha & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha\gamma & 0 & \cdots & 1 - \alpha \end{pmatrix} \cdot \begin{pmatrix} Q_{s'_1 a'_2} \\ Q_{s'_2 a'_2} \\ \vdots \\ Q_{s'_{n_{\mathcal{E}}} a'_1} \end{pmatrix} + \alpha \begin{pmatrix} r_{a'_2} \\ r_{a'_2} \\ \vdots \\ r_{a'_1} \end{pmatrix} \quad (\text{A1})$$

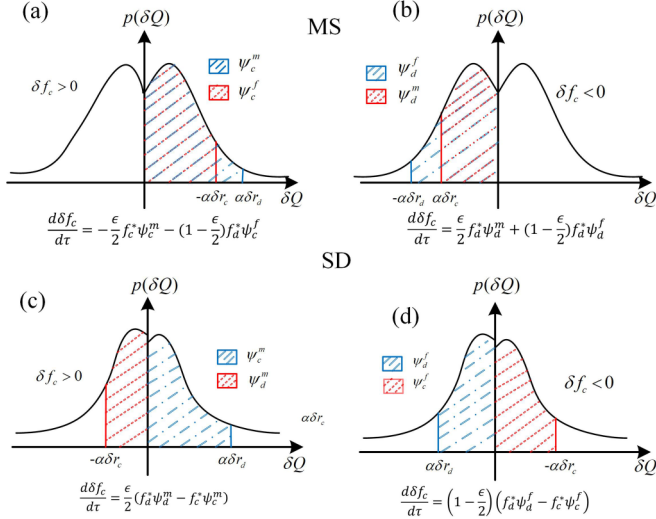


FIG. 13. The schematic of the connection between  $\psi$  and  $\delta Q$  for the SD and MS RLEGs under various fluctuations. Panels (a, b) show connection between  $\psi$  and  $\delta Q$  for the MS RLEGs under  $\delta f_c > 0$  and  $\delta f_c < 0$ , respectively, while panels (b, c) are for the SD RLEGs.

at the point. The maximum element in the row of  $s'_{k-1}$  is

$$Q_{s'_{k-1}a'_k} = \frac{\sum_{i=k}^{n_\varepsilon} r_{a'_i} \gamma^{i-k} + \sum_{i=1}^{k-1} r_{a'_i} \gamma^{n_\varepsilon - k + i}}{1 - \gamma^{n_\varepsilon}}. \quad (\text{A2})$$

By default,  $Q_{s'_{k-1}a'_k}$  refers to  $Q_{s'_{n_\varepsilon}a'_1}$  when  $s'_{k-1} = s'_{n_\varepsilon}$  in our work.

However,  $Q_{s'_k a'_k}$  along the paths passing through  $s'_k | a'_k$  in  $m$  events tends to

$$Q_{s'_k a'_k}(\infty) = \gamma \tilde{Q}_{s'_k a'_{k+1}} + r_{a'_k}$$

under the assumption that the frozen point is stable. Thus, we get the following result:

$$\sum_{k=1}^{n_\varepsilon} \Delta_k = \sum_{k=1}^{n_\varepsilon} (Q_{s'_k a'_k}(\infty) - Q_{s'_k a'_{k+1}}) = 0, \quad (\text{A3})$$

which suggests  $Q_{s'_k a'_k}(\infty)$  is greater than  $Q_{s'_k a'_{k+1}}$  in at least one state row unless  $\Delta_k = 0$  for all  $k \in \{1, \dots, n_\varepsilon\}$ , which requires  $r_{a'_i} = r_{a'_j}$  for any  $a'_i, a'_j$  in the mode  $\mathcal{M}$ . However, the necessary condition to keep the stability of the frozen point is in contradiction with our assumption  $r_{a'_j}$  in case (a). So, the frozen point is unstable for the case (a).

For case (b), we should further assume that  $r_{a'_i} = r_{a'_k} < r_{\max}$  for any  $a'_i, a'_k$  in  $\mathcal{M}$  at the frozen point according to the necessary condition in above analysis. Along with the paths passing through  $s_l | a'_k$ , we have

$$Q_{s_l a'_k} \rightarrow \gamma Q_{s'_k a'_{k+1}} + r_{a'_{k+1}} = Q_{s'_k a'_{k+1}} = Q_{s_{k-1} a'_k}$$

for all  $s_l | a'_k \notin \mathcal{B}_\varepsilon$  if beliefs in  $\mathcal{B}_\varepsilon$  are stable, i.e., the elements in the column of action  $a'_k$  in  $\mathcal{M}$  are homogeneous. Thus, as  $r_{a_l} > r_{a'_k}$ ,  $Q_{s'_k a'_k}$  will be greater than  $Q_{s'_k a'_{k+1}}$  along with the paths passing through  $s'_k | a_l$  in  $m$  events finally because the maximum element in row  $s_l$  is no less than  $Q_{s_l a'_k}$ . Therefore, the frozen point is unstable as  $s_l | a'_k \notin \mathcal{B}_\varepsilon$  is undermined. To

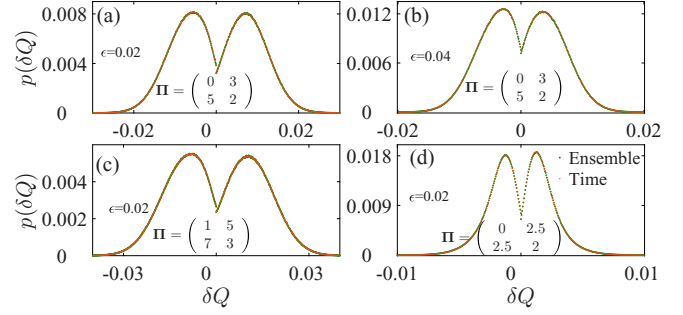


FIG. 14. The distribution of  $\delta Q$  in various MS RLEGs. The learning parameters  $\alpha = \gamma = 0.9$  and  $N = 10\,000$ .

sum up, for both the cases (a) and (b), the proposition (I) is certified.

For proposition (II), the assumption about reward for the action  $a'_i$  in  $\mathcal{M}$  is that  $r_{a'_i} = r_{\max}$ . In this case, the above analysis shows that  $m$  events homogenize elements in the column of any action  $a'_i$  in  $\mathcal{M}$ . It suggests that the difference between the largest element and the second largest is narrowing with update events. So, beliefs in  $\mathcal{B}_\varepsilon$  become fragile as long as  $n_\varepsilon > 1$ . In the case  $n_\varepsilon = 1$  with the mode  $\mathcal{M} = (a'_i)$ , the analysis shows the element in the column  $a'_i$  become homogeneous gradually with  $m$  events,

$$Q_{s'_l a'_i} \rightarrow Q_{s'_l a'_i} = \frac{r_{a'_i}}{1 - \gamma}, \quad \forall s'_l \in \mathcal{S}. \quad (\text{A4})$$

Thus, the elements in an arbitrary column of action  $a'_j$

$$Q_{s'_k a'_j} \rightarrow \gamma Q_{s'_k a'_j} + r_{a'_j} = \frac{r_{a'_j} + \gamma(r_{a'_i} - r_{a'_j})}{1 - \gamma}, \quad \forall s'_k \in \mathcal{S}. \quad (\text{A5})$$

The result indicates the elements in the column  $a'_j$  only depend on  $r_{a'_i}$  and  $r_{\max}$ . Therefore, the elements in different columns of actions also tend to be exactly the same as in  $m$  event as long as the rewards for the actions are identical. Thus, the nonmaximum elements in the column that have the maximum reward approach the maximum gradually so that the beliefs in  $\mathcal{B}_\varepsilon$  become fragile, such as elements in columns of  $a_2, a_3$ , and  $a_4$  in Fig. 5(b).

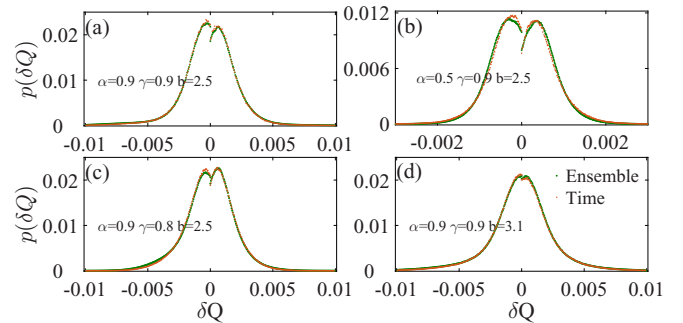


FIG. 15. The distribution of  $\delta Q$  in SD RLEGs under various learning parameters. The exploration rate and the scale of the systems are  $\epsilon = 0.02$  and  $N = 10\,000$ .

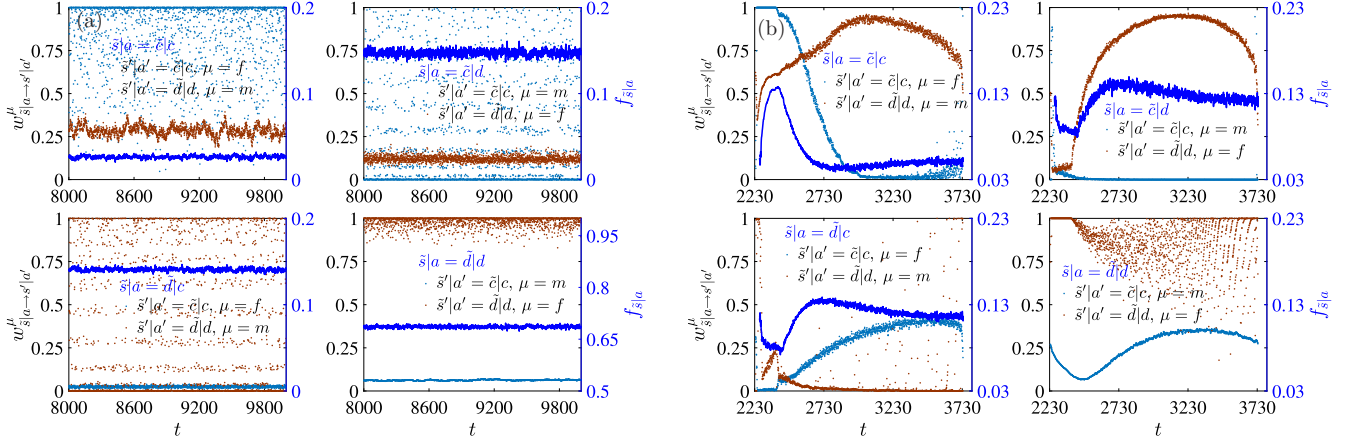


FIG. 16. The time series of  $W_{s_a \rightarrow s'_a}^\mu$  and  $f_{s|a}$  in MS RLEG and SD RLEG. In panels (a) and (b), the payoff matrices for the MS and SD RLEGs are (1, 4; 6, 3) and (6,  $b$ ; 6 +  $b$ , 2) with  $b = 2.5$ , respectively. Other parameters:  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$ ,  $N = 10\,000$ .

### 3. The connection between function $\psi$ and distribution of $\delta Q$

In this part, we first interpret our assumption in Sec. III B 1 that a general fluctuation  $\delta \mathbf{f}$  meets  $\delta f_{a_k} = 0$  for all  $a_k \notin \mathcal{A}^*$  at a mixed equilibrium point  $\mathbf{f}^*$ . As mentioned in Sec. III B 1, the optimal actions at each state for any agent is one action

in  $\mathcal{A}^*$  at the point. Therefore,  $f_{a_k}^* = 0$  for  $a_k \notin \mathcal{A}^*$ . In fact, fluctuations come from the random selection for initiators in RLEGs. For example, the fraction  $f_{a_j}$  decreases if the *visible belief* for initiators in successive rounds is  $\tilde{s}_i|a_j$  with  $s_i \neq s_j$  at  $\mathbf{f}^*$ . Therefore, the assumption makes sense because  $\tilde{s}_i|a_k$  for

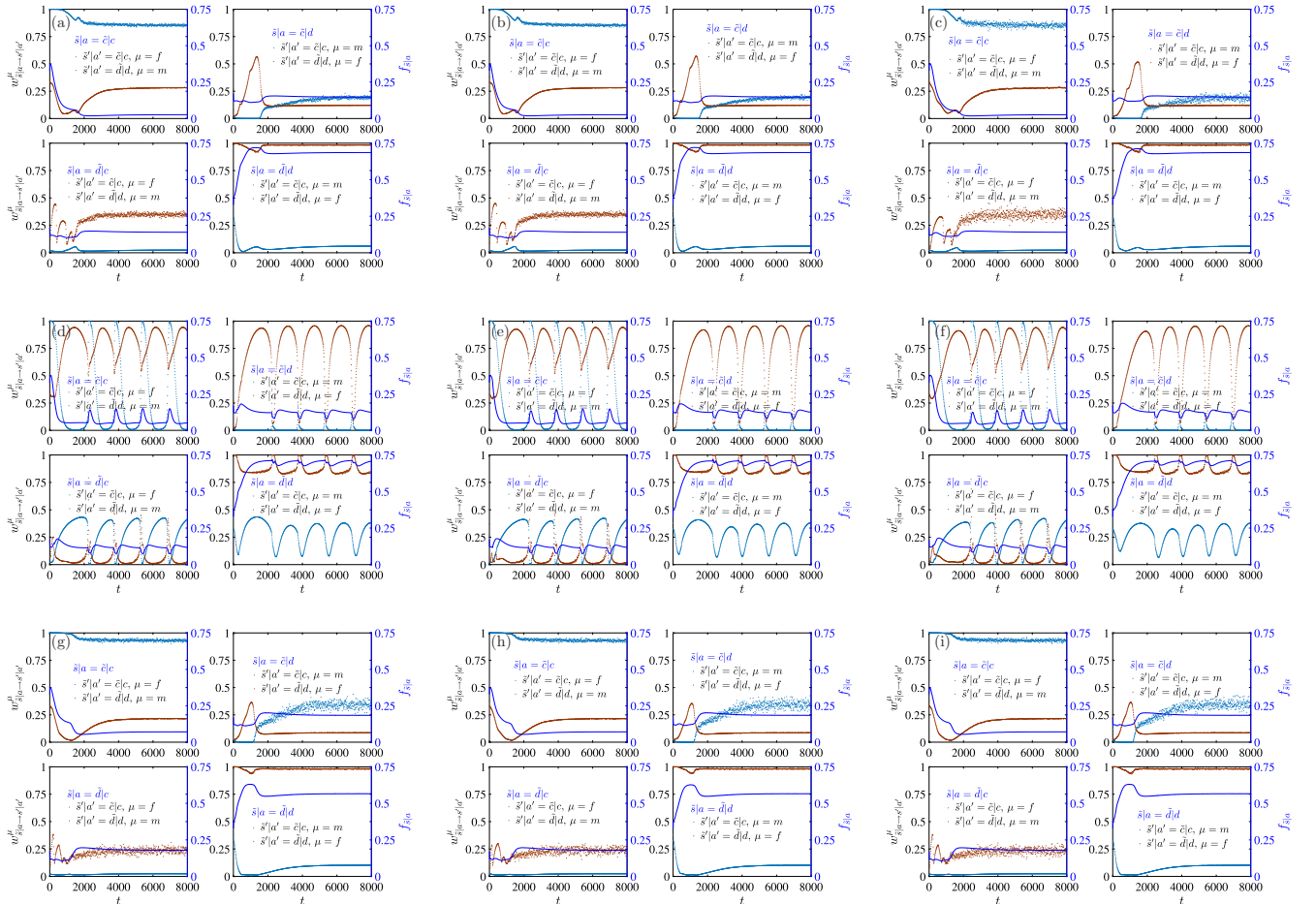


FIG. 17. The time series of  $W_{s_a \rightarrow s'_a}^\mu$  and  $f_{s|a}$  in MS RLEG as well as SD RLEG. The payoff matrices in the three groups are shown in Table II. The rest parameters are  $\alpha = \gamma = 0.9$ ,  $\epsilon = 0.02$  and scale of system is  $N = 10\,000$ . The ensemble average is over 1 000 realizations.

TABLE II. The payoff matrices of RLEGs in Fig. 17. The RLEGs are divided into three groups: (a), (c), (b), (f), and (g), (i), and the RLEGs in the same group are equivalent systems.

	Group 1 (MS)			Group 2 (SD)			Group 3 (MS)		
$\Delta\Pi_{c:}/\Delta\Pi_{d:}$	1			-7/13			2		
$\Delta\Pi_{:c}/\Delta\Pi_{:d}$	5			-5			-3		
$\text{sgn}(\Delta\Pi_{c:})$	-1			1			-1		
$\text{sgn}(\Delta\Pi_{:c})$	-1			-1			-1		
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
$\Pi_{cc}$	1	5	10	6	21	47	3	7	9
$\Pi_{cd}$	4	14	13	2.5	14	40	11	15	33
$\Pi_{dc}$	6	20	15	8.5	26	52	12	16	36
$\Pi_{dd}$	3	11	12	2	13	39	8	12	24

an initiator is impossible after  $\mathbf{f}$  has stayed at  $\mathbf{f}^*$  for a long time.

Under the normalization of  $\mathbf{f}$  in a  $2 \times 2$  RLEG, the evolution for a general fluctuation  $\delta f_c$  is

$$\begin{aligned} \frac{d\delta f_c}{d\tau} = & \frac{\epsilon}{2} [f_d^* \psi_d^m(\delta r_{c:}; \alpha, \gamma, \mathbf{\Pi}) - f_c^* \psi_c^m(\delta r_{d:}; \alpha, \beta, \mathbf{\Pi})] \\ & + \left(1 - \frac{\epsilon}{2}\right) [f_d^* \psi_d^f(-\delta r_{d:}; \alpha, \beta, \mathbf{\Pi}) \\ & - f_c^* \psi_c^f(-\delta r_{c:}; \alpha, \beta, \mathbf{\Pi})]. \end{aligned} \quad (\text{A6})$$

Here,  $\delta r_c = \delta f_c \Delta\Pi_{c:}$  and  $\delta r_d = -\delta f_c \Delta\Pi_{d:}$  with  $\Delta\Pi_{c:} = \Pi_{cc} - \Pi_{cd}$  and  $\Delta\Pi_{d:} = \Pi_{dd} - \Pi_{dc}$ . In Eq. (A6),  $\psi_d^m$  ( $\psi_c^m$ ) is the rate of belief change from  $\{s|d : s \in \mathcal{S}\}$  ( $\{s|c : s \in \mathcal{S}\}$ ) to  $\{s|c : s \in \mathcal{S}\}$  ( $\{s|d : s \in \mathcal{S}\}$ ) in  $m$  events, while  $\psi_d^f$  ( $\psi_c^f$ ) is the rate of belief change from  $\{s|d : s \in \mathcal{S}\}$  ( $\{s|c : s \in \mathcal{S}\}$ ) to  $\{s|c : s \in \mathcal{S}\}$  ( $\{s|d : s \in \mathcal{S}\}$ ) in  $f$  events. Sections III B 1 and III C 1 show that the increase of reward for cooperation in  $m$  events and decrease of reward for defection in  $f$  events could cause belief  $\{s|d : s \in \mathcal{S}\}$  change to  $\{s|c : s \in \mathcal{S}\}$ , while the increase of reward for defection in  $m$  events and decrease of reward for cooperation in  $f$  events could cause belief  $\{s|c : s \in \mathcal{S}\}$  change to  $\{s|d : s \in \mathcal{S}\}$ . Based on the properties of  $\psi$ , one learns that there are only two nonzero terms in the equation for a given fluctuation (see Fig. 13). Here, we employ  $\delta Q = Q_{sc} - Q_{sd}$  rather than  $\delta Q = Q_{sd}^{\max} - Q_{sa}$  for the sake of simplicity, in which  $s$  refers to focusing on agent's state. So,  $\delta Q > 0$  for cooperators while  $\delta Q < 0$  for defectors in the population. We normalize the distribution of  $\delta Q$  for cooperators and defectors, i.e.,  $\int_{-\infty}^0 p(\delta Q) d\delta Q = 1$  and  $\int_0^{\infty} p(\delta Q) d\delta Q = 1$ . In Sec. III B 1, the analysis shows  $\psi$  subject to the distribution of  $\delta Q$  at the equilibrium point  $\mathbf{f}^*$ .

We next show the relationship between  $\psi$  and  $\delta Q$  in MS and SD RLEGs as examples under fluctuations. For the MS RLEGs, the nonzero terms are

$$\psi_c^m \approx \int_0^{\alpha\delta r_d} p(\delta Q) d\delta Q, \quad \text{with } \delta r_d > 0,$$

and

$$\psi_c^f = \int_0^{-\alpha\delta r_c} p(\delta Q) d\delta Q, \quad \text{with } -\delta r_c > 0,$$

in the case  $\delta f_c > 0$  as Fig. 13(a) and Eq. (A6) show, while

$$\psi_d^f = \int_{\alpha\delta r_d}^0 p(\delta Q) d\delta Q, \quad \text{with } \delta r_d < 0,$$

and

$$\psi_d^m \approx \int_{-\alpha\delta r_c}^0 p(\delta Q) d\delta Q, \quad \text{with } -\delta r_c < 0,$$

in the case  $\delta f_c < 0$  as Fig. 13(b) shows. Here,  $f_c^*$  is stable because the sign of  $\delta f_c$  is always opposite to the one of  $d\delta f_c/d\tau$ . The simulations show that  $p(\delta Q)$  is symmetric in the case  $\Delta\Pi_{c:} = \Delta\Pi_{d:}$ . However, the symmetry is broken if  $\Delta\Pi_{c:} \neq \Delta\Pi_{d:}$  [see Figs. 14(a)–14(d)]. It makes our inferences decent in the case of  $\Delta\Pi_{c:} = \Delta\Pi_{d:}$ , but they slightly deviate for  $\Delta\Pi_{c:} \neq \Delta\Pi_{d:}$  (see Fig. 8).

For the SD RLEGs, the nonzero terms are

$$\psi_c^m \approx \int_0^{\alpha\delta r_d} p(\delta Q) d\delta Q, \quad \text{with } \delta r_d > 0,$$

and

$$\psi_d^m \approx \int_{-\alpha\delta r_c}^0 p(\delta Q) d\delta Q, \quad \text{with } \delta r_c > 0,$$

in the case  $\delta f_c > 0$  as Fig. 13(c) shows, while

$$\psi_d^f = \int_{\alpha\delta r_d}^0 p(\delta Q) d\delta Q, \quad \text{with } \delta r_d < 0,$$

and

$$\psi_c^f = \int_0^{-\alpha\delta r_c} p(\delta Q) d\delta Q, \quad \text{with } \delta r_c < 0,$$

in the case  $\delta f_c < 0$  as Fig. 13(d) shows. Through the simulations in Sec. III C, we learn that the collective behaviors in SD RLEGs are divided into three kinds, periodic oscillation ( $\Delta\Pi_{:d}/\Delta\Pi_{:c} < \psi_d^m/\psi_c^m$ ), stable coexistence ( $\psi_d^m/\psi_c^m < \Delta\Pi_{:d}/\Delta\Pi_{:c} < \psi_d^f/\psi_c^f$ ), and aperiodic oscillation ( $\psi_d^m/\psi_c^m < \Delta\Pi_{:d}/\Delta\Pi_{:c}$ ). For the default payoff matrix, the increase of  $b$  results in an increase of  $\psi_c^m$  but a decrease of  $\psi_d^m$  as Fig. 13(c) shows. Therefore, the periodic oscillation fades away with the increase of  $b$  in the simulation. In Fig. 15, we further investigate the impact of the learning parameters and the payoff matrix on  $p(\delta Q)$ . For fluctuations around  $f_c^*$ , the result shows the ratio  $\psi_d^m/\psi_c^m$  increases with increasing  $\gamma$ , while decreases with the increase of  $\alpha$  [see Figs. 15(a)–15(c)]. Therefore, the transition point  $b'$  becomes larger with  $\gamma$ , while it is reduced with the increase of  $\alpha$  (see Fig. 3). The result also manifests that  $\psi_c$  gets close to  $\psi_d$  gradually as  $b$  approaches  $b'$  (see Fig. 13).

#### 4. The transition matrix and equivalent systems

The analysis in Sec. III C 2 indicates the transition matrix  $W$  is not unique for a given  $\mathbf{f}^{*\pm}$  at  $\mathbf{f}^*$  in RLEGs for the MS game setting. In Fig. 16(a), we give the time series of  $W$  in an MS RLEG rather than the average in an ensemble to check it further. The result shows the  $W$  is not fixed but indicates some elements in  $W^f$  fall into several attractors, such as  $w_{\tilde{c}|d \rightarrow \tilde{c}|c}^m$  and  $w_{\tilde{c}|d \rightarrow \tilde{c}|c}^m$ . Furthermore, the time series of  $W$  in an SD RLEG also are displayed in Fig. 16(b). It indicates that

the element  $w_{\vec{d}|d \rightarrow \vec{d}|d}^f$  also falls into several attractors during the *quiescent stage* but the attractors are evolving over time and get together finally. The results above indicate that the transition matrix  $W$  is a better quantity than  $f_c$  to measure whether two systems are equivalent or not. Through the

average  $W$  in an ensemble, we give more simulations to check our conjecture that systems are equivalent if  $\Delta\Pi_{c:}/\Delta\Pi_{d:}$ ,  $\Delta\Pi_{:c}/\Delta\Pi_{:d}$ ,  $\text{sgn}(\Delta\Pi_{c:})$ , and  $\text{sgn}(\Delta\Pi_{:c})$  are identical in the systems. Due to the normalization, only part of elements in  $W^f$  and  $W^m$  are shown in Fig. 17. The results further confirm our conjecture in Sec. III C 1.

- 
- [1] R. H. Turner, L. M. Killian *et al.*, *Collective Behavior*, Vol. 3 (Prentice-Hall, Englewood Cliffs, NJ, 1957).
- [2] T. Vicsek and A. Zafeiris, *Phys. Rep.* **517**, 71 (2012).
- [3] I. D. Couzin and J. Krause, *Adv. Study Behav.* **32**, 10 (2013).
- [4] D. J. Sumpter, *Collective Animal Behavior* (Princeton University Press, Princeton, NJ, 2010).
- [5] S. Iwanaga and A. Namatame, *Procedia Comput. Sci.* **24**, 217 (2013).
- [6] S. Gavrilits and L. Fortunato, *Nat. Commun.* **5**, 3526 (2014).
- [7] A. Cuesta, O. Abreu, and D. Alvear, *Safety Sci.* **88**, 54 (2016).
- [8] G. Tkačik, O. Marre, T. Mora, D. Amodei, M. J. Berry II, and W. Bialek, *J. Stat. Mech.* (2013) P03011.
- [9] P. Gopikrishnan, B. Rosenow, V. Plerou, and H. E. Stanley, *Phys. Rev. E* **64**, 035106(R) (2001).
- [10] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, *Proc. Natl. Acad. Sci. USA* **107**, 12576 (2010).
- [11] J. E. Herbert-Read, A. Perna, R. P. Mann, T. M. Schaefer, D. J. Sumpter, and A. J. Ward, *Proc. Natl. Acad. Sci. USA* **108**, 18726 (2011).
- [12] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardinà, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini *et al.*, *Proc. Natl. Acad. Sci. USA* **105**, 1232 (2008).
- [13] J. M. Smith, *Evolution and the Theory of Games* (Cambridge University Press, Cambridge, UK, 1982).
- [14] M. A. Nowak, *Evolutionary Dynamics* (Harvard University Press, Cambridge, MA, 2006).
- [15] J. Newton, *Games* **9**, 31 (2018).
- [16] L. A. Imhof, D. Fudenberg, and M. A. Nowak, *Proc. Natl. Acad. Sci. USA* **102**, 10797 (2005).
- [17] B. Sinervo, E. Svensson, and T. Comendant, *Nature* **406**, 985 (2000).
- [18] A. Szolnoki and M. Perc, *New J. Phys.* **17**, 113033 (2015).
- [19] G. Szabó and C. Hauert, *Phys. Rev. Lett.* **89**, 118101 (2002).
- [20] W.-X. Wang, J. Ren, G. Chen, and B.-H. Wang, *Phys. Rev. E* **74**, 056113 (2006).
- [21] M. Nanda and R. Durrett, *Proc. Natl. Acad. Sci. USA* **114**, 6046 (2017).
- [22] Q. Su, L. Wang, and H. E. Stanley, *New J. Phys.* **20**, 103030 (2018).
- [23] M. Van Veelen, J. García, D. G. Rand, and M. A. Nowak, *Proc. Natl. Acad. Sci. USA* **109**, 9929 (2012).
- [24] K. Panchanathan and R. Boyd, *Nature* **432**, 499 (2004).
- [25] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, *Nature* **441**, 502 (2006).
- [26] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach* (Springer Science & Business Media, Berlin, 2013).
- [27] Y. LeCun, Y. Bengio, and G. Hinton, *Nature* **521**, 436 (2015).
- [28] J. Tubiana and R. Monasson, *Phys. Rev. Lett.* **118**, 138301 (2017).
- [29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).
- [30] C. Wang and H. Zhai, *Front. Phys.* **13**, 130507 (2018).
- [31] B.-J. Lin, X.-R. Li, and W.-L. Yu, *Front. Phys.* **15**, 24602 (2020).
- [32] N. M. Nasrabadi, *J. Electron. Imag.* **16**, 049901 (2007).
- [33] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, *IEEE Trans. Pattern. Anal. Mach. Intell.* **35**, 1915 (2013).
- [34] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2014), pp. 1799–1807.
- [35] J. A. Cruz and D. S. Wishart, *Cancer Info.* **2**, 117693510600200030 (2006).
- [36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, *Nature* **518**, 529 (2015).
- [37] N. Brown and T. Sandholm, *Science* **359**, 418 (2018).
- [38] J. J. Valletta, C. Torney, M. Kings, A. Thornton, and J. Madden, *Anim. Behav.* **124**, 203 (2017).
- [39] N. Abaid, E. Bollt, and M. Porfiri, *Phys. Rev. E* **85**, 041907 (2012).
- [40] P. DeLellis, M. Porfiri, and E. M. Bollt, *Phys. Rev. E* **87**, 022818 (2013).
- [41] J. Hagenauer and M. Helbich, *Expert Syst. Appl.* **78**, 273 (2017).
- [42] Z. Qin, T. Wan, Y. Dong, and Y. Du, *Appl. Soft Comput.* **26**, 368 (2015).
- [43] Z. Qin, F. Khawar, and T. Wan, *Neurocomputing* **194**, 74 (2016).
- [44] K. Ried, T. Müller, and H. J. Briegel, *PLoS One* **14**, e0212044 (2019).
- [45] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, *Science* **362**, 1140 (2018).
- [46] A. Potapov and M. K. Ali, *Phys. Rev. E* **67**, 026706 (2003).
- [47] C. J. C. H. Watkins and P. Dayan, *Mach. Learn.* **8**, 279 (1992).
- [48] H. Van Hasselt, A. Guez, and D. Silver, in *AAAI*, Vol. 2 (Phoenix, AZ, 2016), p. 5.
- [49] H. V. Hasselt, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2010), pp. 2613–2621.
- [50] S.-P. Zhang, J.-Q. Zhang, L. Chen, and X.-D. Liu, *Nonlinear Dynamics* (2020), doi: 10.1007/s11071-019-05398-4.
- [51] M. Cao, A. S. Morse, and B. D. Anderson, *IFAC Proc.* **38**, 17 (2005).
- [52] H.-L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi, *Phys. Rev. Lett.* **110**, 210601 (2013).