

# Evolution of cooperation in the public goods game with Q-learning

Guozhong Zheng,<sup>1</sup> Jiqiang Zhang,<sup>2</sup> Shengfeng Deng,<sup>1</sup> Weiran Cai,<sup>3</sup> and Li Chen<sup>1,\*</sup>

<sup>1</sup>*School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710061, P. R. China*

<sup>2</sup>*School of Physics, Ningxia University, Yinchuan 750021, P. R. China*

<sup>3</sup>*School of Computer Science, Soochow University, Suzhou 215006, P. R. China*

(Dated: July 30, 2024)

Recent paradigm shifts from imitation learning to reinforcement learning (RL) is shown to be productive in understanding human behaviors. In the RL paradigm, individuals search for optimal strategies through interaction with the environment to make decisions. This implies that gathering, processing, and utilizing information from their surroundings are crucial. However, existing studies typically study pairwise games such as the prisoners' dilemma and employ a self-regarding setup, where individuals play against one opponent based solely on their own strategies, neglecting the environmental information. In this work, we investigate the evolution of cooperation with the multiplayer game – the public goods game using the Q-learning algorithm by leveraging the environmental information. Specifically, the decision-making of players is based upon the cooperation information in their neighborhood. Our results show that cooperation is more likely to emerge compared to the case of imitation learning by using Fermi rule. Of particular interest is the observation of an anomalous non-monotonic dependence which is revealed when voluntary participation is further introduced. The analysis of the Q-table explains the mechanisms behind the cooperation evolution. Our findings indicate the fundamental role of environment information in the RL paradigm to understand the evolution of cooperation, and human behaviors in general.

## 1. INTRODUCTION

Cooperation is crucial for the sustainable development of humans from tackling climate warming [1], fighting global infectious diseases, and dealing with the melting of glaciers causing the accelerated extinction of species, where the web of lives that sustains us is being worn and torn apart. The solution to these issues depends on human cooperation [2]. Deciphering the cooperation mechanism is thus a question of paramount importance, and is crucial for addressing the challenges of these public tragedies [3].

To achieve this aim, researchers investigate the emergence of cooperation by adopting the framework of evolutionary game theory [4, 5], where the prisoners' dilemma (PD) game [6] serves as a typical model to study the evolution of cooperation behavior in pairwise interactions [7]. In the PD, there is a potential conflict for the two engaged players between the individual and their collective interests [6, 8]. Importantly, the public goods game (PGG) [9–11] extends the scenario to multilateral interactions, where the number of participants could be arbitrary. In PGG, cooperation contributes to the common goods, but the free-riding behaviors ruin the benefit of the collective and ultimately lead to the tragedy of commons if no countermeasure is engaged [12–14]. Nowadays, the PGG is widely adopted to discuss many issues we are facing, such as resource management, environmental protection, and the provision of public services [15].

After endeavors for several decades, various mechanisms for the emergence of cooperation in the PGG have been proposed, including voluntary participation [16–19], punishment [20–23] and reward [24–26], social diversity [27], noise [28], network heterogeneity [29], reputation and reciprocity [30], etc. These endeavors provide important insights into the mechanism of cooperation emergence. However,

most of these findings are mainly based on imitation learning (IL) [31, 32], where players imitate strategies of neighbors who are better off in terms of payoffs. In essence, IL can be taken as a simple version of social learning [33], where individuals learn from others in their socioeconomic activities, through observation or instruction, which may or may not involve direct experiences. As a different paradigm, reinforcement learning (RL) [34] provides a completely distinct framework for understanding human behaviors, which has garnered increasing attention in recent years, including deciphering the emergence mystery of cooperation [35–49], trust [50], resource allocation [51, 52], and other collective behaviors for humans [47, 53]. There players acquire information through interactions with the environment and adjust behaviors based on rewards provided by the environment, aiming to maximize their cumulative rewards.

Within the framework of RL, some new insights into the mechanism of cooperation emergence for the pairwise game have been provided. For example, by adopting the Bush-Mosteller model as a simple RL setup, Masuda *et al.* reproduce the failure of network reciprocity [54] and discuss the condition for cooperation emergence [55]. Jia *et al.* use the Monte Carlo method to study pattern formation and phase transitions towards cooperation in social dilemmas that are driven by RL, they show that global players play a decisive role in ensuring cooperation [56]. Song *et al.* combine the iterated prisoner's dilemma game with the Bush-Mosteller reinforcement learning model and show that there exists a moderate switching dynamics of the interaction intensity that is optimal for the evolution of cooperation [57]. In Ref. [41], when reinforcement learners playing PD game are subject to Lévy noise, a particularly positive role of Lévy noise is unveiled within this social dilemma. Ding *et al.* reveal that a strong memory and a long-sighted expectation yield the emergence of coordinated optimal policies, where both players act

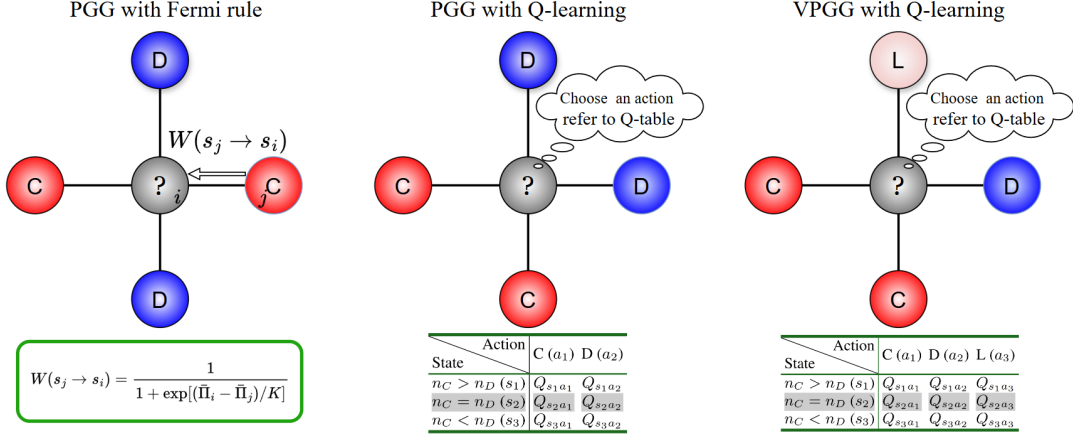


FIG. 1. **Three different setups for PGG.** (Left) PGG with Fermi rule: players engage in the PGG and update their strategies using the Fermi updating rule. (Middle) PGG with Q-learning algorithm: players engage in the PGG and update their strategies using the Q-learning algorithm. (Right) VP GG with Q-learning algorithm: players engage in the VP GG and update their strategies using the Q-learning algorithm. The grey circle represents the focal individual, and red, blue, and pink represent the cooperators, defectors, and loners in the neighborhood. At the bottom of each panel, the shown are the Fermi rule, and the corresponding Q-tables for the other two setups using Q-learning, respectively.

like “win stay, lose shift” to maintain a high level of cooperation [44].

Till now, these studies are primarily based on the PD where only two players are engaged in each single game. Only recently, few works have started to discuss the evolution of multiplayer game within the RL paradigm by adopting PGG. In Ref. [58], the researchers investigate how cooperative behavior evolves in a PGG on a network with empty nodes based on the Bush-Mosteller model; the results show that only the existence of appropriate empty nodes in the network can stimulate individual cooperative behavior. Wang *et al.* propose a framework that combines the PGG and adaptive reward mechanisms to better capture decision-making behaviors in multiplayer interactions, where they are mainly concentrating on the synergistic effects of them in RL [42]. Ref. [59] introduces the Q-learning algorithm into the voluntary PGG to explore the evolution of cooperation, and uncovers when the synergy factor is large and the adjust loner payoff’s multiply factor is smaller, the number of cooperators becomes gradually consistent, and the proportion of defectors evolves non-linearly for different synergy factors.

However, these works based on PGG consider external factors, such as empty nodes or rewards, it’s unclear what’s evolution of cooperation within the original PGG game. More importantly, these works only consider the individual’s own action information or are termed as a self-regarding setup, where the players neglect the information of surroundings. This setup simplifies the model but obviously contradicts our daily experiences, where our decision-making is based upon not only our own actions but also the environmental information (e.g. neighbors’ actions) we perceived. It is generally believed that once more information about their surroundings is incorporated, we could make better decisions to maximize the payoffs [34].

In this work, we investigate the evolution of cooperation of the original PGG within the paradigm of RL where the environmental information is incorporated. Specifically, we empower players with the Q-learning algorithm to revise their strategies, where each player is guided by a Q-table. The environmental information is incorporated into the state of the Q-table in a way that keeps the Q-table simple enough. Our simulation shows that the incorporation of environmental information is able to promote the cooperation prevalence, and cooperation is more likely to emerge than the case of IL. Besides, for the PGG with voluntary participation, the cyclic dominance in the previous studies [16] within IL disappears; instead, a non-monotonic cooperation dependence of the game parameter is observed, which is counter-intuitive.

## 2. MODEL

Let’s consider a square lattice with periodic boundary condition, where each node represents an individual engaging in the public goods game. Specifically, we investigate two models with the Q-learning algorithm: the public goods game (PGG) and the voluntary public goods game (VP GG). In the first model, the focal player initiates a public goods game with four of its nearest neighbors, where each player can either choose cooperation (C) or defection (D). The choice of C means that an individual invests a certain amount (set to 1) in the public pool, while D represents a free-riding behavior without any contribution. Different from PGG, participation in VP GG is voluntary, players can choose to be “loner” (L) as the third strategy, if they do not want to play the game at all, and receive a small fixed income denoted as  $\sigma$  (fixed at 1 if not stated otherwise). It is noteworthy that Ref. [59] mainly focus on the impact of the loner strategy with self-regarding

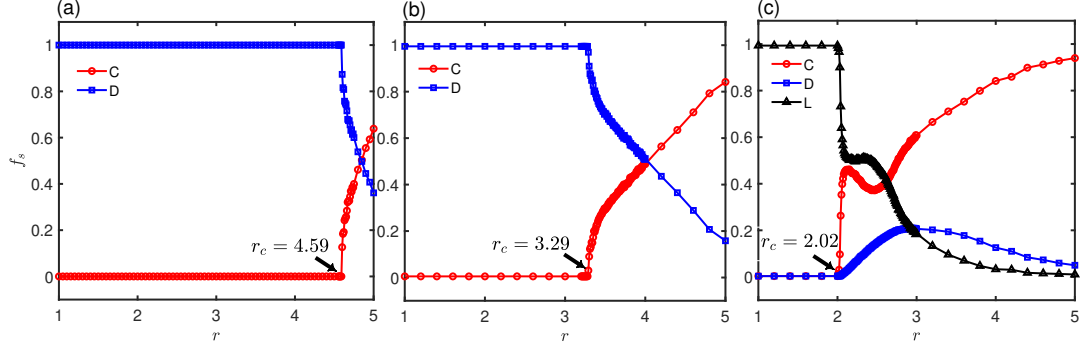


FIG. 2. (Color online) **Cooperation phase transition versus the gain factor  $r$  for three different models.** (a) Players update their strategies using the Fermi rule in PGG with  $K=0.1$ , and the threshold for cooperation emergence is around  $r_c=4.59$ . (b) Players update their strategies using the Q-learning algorithm in PGG, and the threshold for cooperation emergence is around  $r_c=3.29$ . (c) Players update their strategies using the Q-learning algorithm in VPGG, and the threshold for cooperation emergence is around  $r_c=2.02$ . Other parameters:  $\epsilon=0.01$ ,  $\alpha=0.1$ ,  $\gamma=0.9$ ,  $\sigma=1.0$ , and the population size  $N = 100 \times 100$ .

setup on the evolution of cooperation, without considering the environmental information of individuals and using adjustable payoffs for loners. Here, we maintain consistency with previous studies on IL [16] by fixing the loner's payoff at 1, while incorporating individuals' environmental information into the Q-table setup to study the evolution of cooperation within the RL framework. This allows for a better comparison with the results of IL.

Suppose now a player  $i$  initiates a PGG, five participants ( $n_i = 5$ ) are involved, while the number of participants in VPGG needs to subtract the number of loners  $n_L$  in their neighborhood. After the investment, the amount of money in the public pool is multiplied by a gain/synergy factor  $r$ , and then evenly divided among all participating players. The expected payoffs for different types of players  $i$  are

$$\pi_i = \begin{cases} \frac{r n_C}{n_C + n_D} - 1, & \text{if } s_i = C \\ \frac{r n_C}{n_C + n_D}, & \text{if } s_i = D \\ \sigma, & \text{if } s_i = L \end{cases} \quad (1)$$

where  $s_i$  represents the strategy of the player  $i$ ,  $n_{C,D}$  represents the number of cooperators and defectors, respectively. By convention, in the case of  $r \in (1, n_i)$  or  $r \in (1, n_i - n_L)$ , social dilemmas arise where the choice D brings a higher payoff than C.

In our models, individuals update their strategies according to the Q-learning algorithm [60, 61], where each has a Q-table to guide their decision-making. The Q-table is a two-dimensional table expanded by states and actions. The action set consists of all available strategies, which is  $\mathbb{A} = \{C, D\}$  in PGG, and  $\mathbb{A} = \{C, D, L\}$  in VPGG. The state set serves to characterize the environment perceived by individuals. In the previous self-regarding setup, the state only records the action adopted by the player itself in the last round. Here, the environment information is captured in a simple manner by comparing the number of cooperators  $n_C$  and defectors

$n_D$  among the player's four nearest neighbors. Therefore, there are three possible states  $s_{1,2,3}$  defined by the condition of  $n_C > n_D$ ,  $n_C = n_D$ ,  $n_C < n_D$ , respectively. We use the same state set in both models  $\mathbb{S} = \{s_1, s_2, s_3\}$ . The two Q-tables are shown at the bottom of Fig. 1. The idea of Q-table is score to the value of each action  $a$  within each state  $s$ . A larger value of  $Q_{s,a}$  within a given row is supposed to be more valuable within that state, which means the action  $a$  is more likely to be chosen when within the state  $s$ .

---

**Algorithm 1:** Pseudocode of PGG/VPGG with Q-learning

---

**Input:**  $\alpha, \gamma, \epsilon$   
1 **Initialization;**  
2 **for each agent do**  
3   Pick an action randomly from  $\mathbb{A}$ ;  
4   Create a Q-table with each item in the matrix near zero;  
5 **for each agent do**  
6   Generate state:  $s$ ;  
7 **repeat**  
8   **for each agent do**  
9     **if**  $\text{rand}() < \epsilon$  **then**  
10       Pick an action randomly from  $\mathbb{A}$ ;  
11     **else**  
12       Takes action according to the Q-table;  
13        $a = \arg \max_{a_i} Q_{s,a_i}$ ;  
14     **for each agent do**  
15       Compute rewards according to Eq.(1);  
16       Update state and Q-table according to Eq.(2);  
17 **until the termination condition is met;**

---

Without loss of generality, each player is randomly assigned with an action from the action set  $\mathbb{A}$ , and the elements  $Q_{s,a}$  in the Q-tables randomly assigned a value between [0, 1) independently. In our Monte Carlo simulations, the evolution follows a synchronous updating procedure. In each round  $t$ , with a probability  $\epsilon$ , they opt for a random action  $a \in \mathbb{A}$ ;

with probability  $1 - \epsilon$ , they strictly follow the guidance of their Q-tables by choosing the action with the largest Q-value given their current states. After all players make their moves, they receive the reward  $\pi_i$  following Eq. (1), and are able to compute the new state  $s'$ . As the next step, each player  $i$  tries to draw lessons by revising the corresponding  $Q_{s,a}^i$  they just adopted in this round, according to the Bellman equation [34]:

$$Q_{s,a}^i(t+1) = (1 - \alpha)Q_{s,a}^i(t) + \alpha \left( \pi_i(t) + \gamma \max_{a'} Q_{s',a'}^i(t) \right), \quad (2)$$

where  $s$  and  $a$  represent the current state and action of the focal individual  $i$ , and  $s'$  is the new state at  $t + 1$ .  $\alpha \in (0, 1]$  is the learning rate determining how fast of the Q-value is revised,  $\gamma \in [0, 1]$  is the discount factor, defining the significance of future rewards. Notice that,  $\pi_i(t)$  is the payoff for player  $i$  when she/he acts as the focal player initiating a PGG game, not all payoffs received. Because the item  $Q_{s,a}^i$  reflects the value of action the player  $i$  actively made [40]. When all players update their Q-tables, a single MC round is completed. The evolution consists of many MC steps and is terminated until the system reaches equilibrium or the desired duration is reached, the pseudocode is provided in Algorithm 1.

Besides the two models using Q-learning proposed above, we also adopt the Fermi updating rule [62] in the traditional framework of IL as the baseline model for comparison. Here, players make decisions by comparing utility with their neighbors and imitate strategies of those who are better off in the game. Since this setup has been extensively studied in previous game-theoretic works, we refer to Ref. [16, 63] for the simulation details. The three involved model setups are intuitively illustrated in Fig. 1.

### 3. RESULTS

We first report the fraction of different strategies as a function of the gain factor  $r$  in Fig. 2 for the three models. Fig. 2(a) shows the results with the Fermi updating rule, where we observe that cooperation emerges as the gain factor is larger than a critical threshold  $r > r_c \approx 4.59$ , in line with previous studies [63]. When the gain factor  $r$  is high enough, a cooperation phase transition is seen as shown here. Notice that, the model with a strictly synchronous updating exhibits similar phase transition, with the threshold being  $r_c \approx 4.80$  (data now shown), showing a bit more difficult for the cooperation to emerge. Fig. 2(b) shows the results for PGG with the Q-learning algorithm. Compared to Fig. 2(a), the cooperation phase transition is qualitatively the same, but with a reduced threshold  $r_c \approx 3.29$ . This means that cooperation is more likely to emerge within the reinforcement learning paradigm than in the case of imitation learning.

When voluntary participation is introduced, Fig. 2(c) shows that the threshold is further reduced to  $r_c \approx 2.02$  where cooperation is most likely to emerge among the three models we studied here. Detailed examination indicates that the introduction of “loner” is an efficient manner to considerably in-

hibit the growth of defectors. For a lower gain factor  $r < r_c$ , the fraction of loners dominate. Defectors appear only when cooperators survive as  $r > r_c$ , but as the gain factor further increases cooperators dominates. The fraction of defectors remains  $f_D \lesssim 0.2$  for the whole range. This overall picture aligns with findings from previous work [16] by IL. However, there is a non-monotonic dependence region for  $r \in (2.13, 2.50)$  where the increase in  $r$  leads to the reduction in  $f_C$  accompanied by the increase in  $f_D$ . Notice that, the non-vanishing  $f_L$  at the end of  $r \rightarrow 5$  effectively inhibits the invasion of defectors, guaranteeing a higher level of cooperation compared to the case without loner [see Fig. 2(b)]. A systematic investigation of the impact of the two learning parameters is given in Appendix A. It shows that when individuals both focus on historical experience and have a long-term vision, cooperation is likely to emerge. Otherwise, either a forgetful property (a large  $\alpha$ ) and/or a short-term vision (a small  $\gamma$ ) lead to the failure of its emergence.

To develop some intuition for the evolution process for the three scenarios, we present some typical spatio-temporal snapshots of the system. Starting from random initial conditions, Figs. 3(a1-a4) show that some cooperation clusters are formed. This is a typical illustration of network reciprocity [31], where the structured population enables cooperators to persist by forming clusters, thereby avoiding the exploitation by defectors. For the given parameter  $r = 4.60$ , being slightly larger than the threshold, the resulting prevalence of cooperation is stabilized around 0.13, while the cooperation clusters are not frozen but fluctuate all the time.

The PGG with the Q-learning algorithm is illustrated in Fig. 3(b1-b4), where  $r$  is also chosen slightly larger than the threshold. As can be seen, it's evident that the patterns are quantitatively distinct from Figs. 3(a1-a4), where there are no apparent cooperator clusters observed. Over time, the number of cooperators gradually declines, and upon reaching a steady state, these cooperators are evenly scattered across the whole domain. Additionally, the time-scale reaching a steady state seems longer in this case than in Figs. 3(a1-a4).

Finally, we showcase the pattern evolution for the VPGG with Q-learning, with identical parameter  $r = 3.32$  as adopted in Fig. 3(b1-b4) for comparison. Similar to the PGG, no obvious cooperation clusters are seen in Fig. 3(c1-c4). As time goes by, however, both the numbers of loners and defectors decrease, and cooperators dominate in the end, with a notably higher cooperation prevalence compared to Fig. 3(b1-b4). Furthermore, the patterns depicted do not display the waves with cyclic dominance as observed in IL [16]. This suggests a different mechanism working here.

### 4. MECHANISM ANALYSIS

To understand how cooperation evolves when people play PGG with Q-learning algorithm. In this section, we move to the mechanism analysis of the latter two cases playing with Q-learning.



#### 4.1 Mechanism analysis of PGG with Q-learning

We commence our analysis by delving into PGG with Q-learning, where the cooperation phase transition and pattern evolution have been shown above. To further observe the temporal evolution, we present time series in Fig. 4(a), with identical parameters adopted in Fig. 3(b1- b4). Two strategy fractions remain at the initial level around 0.5 and begin to decline later on and eventually stabilize. This observation is fundamentally different from the “first down and later up” trend often regarded as a sign of network reciprocity in IL [64, 65].

To understand this trend, we need to take into account the surrounding information and the resulting payoffs. As depicted in Fig. 4(b), with random initial conditions, the proportions of the three states  $s_{1,2,3}$  are approximately identical. Fig. 4(c) shows that initially free-riders (D) consistently yield higher payoffs than cooperators (C) since approximately half of the individuals choose to cooperate. As a result, players tend to defect by learning, leading to an increasing proportion of the state  $s_3$  ( $n_C < n_D$ ), and accordingly the other two state fractions of  $s_{1,2}$  decrease. Specifically, the state  $s_3$  includes two different scenarios (see Fig. 5): (a) The focal individual is surrounded by four defectors, in which case choosing C yields a negative payoff, which is disadvantageous. (b) There is one cooperator present in the neighborhood around the focal individual, choosing C will result in a positive payoff, leading to an increase in the Q-value in this state. While opting for strategy D yields a higher payoff but this also regresses the environment back to scenario (a).

In the process of decision-making using the Q-learning algorithm, individuals focus on both historical experience and long-term rewards, therefore they try to avoid falling into scenario (a). Consequently, by factoring in the low payoff in scenario (a) and the potential for higher payoffs in the new state  $s_2$  ( $n_C = n_D$ ), the preference of choosing C in state  $s_3$  arises, leading to the emergence of cooperation. In fact, the choice of C indeed brings higher payoffs on average as shown by the crossover in Fig. 4(c). In Fig. 4(d-e), we show the evolution of action preferences in three different states. It can be observed that the emergence of cooperation primarily occurs mostly in the state  $s_3$ .

To further confirm, we monitor the preference evolution captured by the Q-values of a typical individual within state  $s_3$ , as shown in Fig. 6. It's convenient to define  $\Delta Q = Q_C - Q_D$  as the cooperation preference, a positive value corresponds to the case that cooperation is preferred by learning and vice versa. With this in mind, we observe that the value of  $\Delta Q$  fluctuates up and down around 0. Most of the time,  $\Delta Q < 0$  with intermittent occurrences of  $\Delta Q > 0$ . This means that D is the preferred mostly but C is also selected from time to time. The two actions compete and reach a balance as seen in Fig. 4(f). As the gain factor  $r$  increases, the payoff advantage of defectors over cooperators diminishes, indicating an increased likelihood of cooperation emergence. This is validated in the inset of Fig. 6, which shows an in-

creased frequency of  $\Delta Q > 0$  when  $r$  increases to 3.4.

The above analysis indicates that self-reflective decision-making within Q-learning does not rely on clusters of cooperators to resist defectors. Instead, it possesses the potential to escape from the most adverse environments of full defection, thereby leading to the emergence of cooperation.

#### 4.2 Mechanism analysis of VPGG with Q-learning

We next turn to the mechanism in the VPGG with the Q-learning, especially the impact of voluntary participation, shown in Fig. 7. Starting from random initial conditions, the fraction of loners (L) starts to increase, and both defectors and cooperators decrease. But once the fraction of D stabilizes near zero, cooperators start to increase and loners decrease, and eventually, the two fractions become stabilized with approximately equal size for the given parameters, shown in Fig. 7(a). This evolution is also captured by the fraction of the three states  $s_{1,2,3}$ , where the fraction of state  $s_1$  dominates and the fraction of  $s_3$  diminishes.

Fig. 7(c) shows that the payoff of the loner is much larger than the other two fractions at the initial stage, which explains the number increases in loners. Figs. 7(d-f), respectively, show the evolution of action fraction within the three states  $s_{1,2,3}$ , where the action preference differs significantly. To specifically understand the non-trivial evolution trend in Fig. 7(a), we need to examine the action prevalence in the three states  $s_{1,2,3}$  by incorporating the state fractions shown in Fig. 7(b). At the first stage ( $t \lesssim 10^4$ ), the state fraction of  $s_3$  is slightly higher than the other two, where the action fraction of loner is highest ( $10^2 \lesssim t \lesssim 10^4$ ), the three action fractions in  $s_{1,2}$  are approximately identical; these observations then explain the increasing trend in loners and decreasing trend of both C and D.

At around  $t = 10^4$ , the action fraction of L becomes largest in all three  $s_{1,2,3}$ , and loner becomes the most preferred choice in all cases by learning at this moment. This change then leads to the peak of fraction in L in Fig. 7(a). As  $f_D \rightarrow 0$ , this change provides an ideal surrounding for players to choose cooperation, which brings higher rewards than the case of being loners (see Fig. 7(c)). Therefore, the state  $s_1$  dominates afterward, and the action of C becomes a more preferred action than L, which is supported by the payoff comparison provided in Table I in Appendix B. That's the reason for the crossover for loners and cooperators as seen in later evolution in Fig. 7(d). Though, within state  $s_2$ , being a loner is a better choice than being a cooperator (see Table II in Appendix B). Within state  $s_3$ , being a loner is still the best choice, and being a cooperator is worst, though the fraction of  $s_3$  is too small.

Taken together Fig. 7(a-c), players opt to be loners within  $s_{2,3}$ , and cooperator is more preferred within  $s_1$ . This is the reason why only cooperators and loners survive and coexist. The overall picture is that individuals in this setup choose strategy L from time to time to inhibit the growth of defectors, laying the groundwork for cooperators to thrive in a sea

of loners.

By this point, one might question the coexistence stability of L and C, and whether the strategy L can entirely be transformed into C or vice versa. The answer is no. An all-C system is unstable and susceptible to the invasion of D. Similarly, an all-L system is also unstable and can be invaded by C, and an all-D system is susceptible to the invasion of L. This cyclic dominance determines that none could dominate, they have to coexist in the form of a subtle balance, even though the fraction of D is very small. An example is shown in Fig. 8, where we present the time series of Q values for all three strategies in a stable state, randomly selecting an individual. We see that  $Q_C$  fluctuates up and down around  $Q_L$ , while  $Q_D$  shows a dependence on  $Q_C$ , their orders are reshuffled consistently. This observation demonstrates that the three strategies are in dynamic equilibrium.

### 4.3 Analysis of non-monotonic variation

Till now, the non-monotonic dependence of cooperation level for  $r \in (2.13, 2.50)$  in Fig. 2(c) remains unexplained. In this interval, the fraction of the strategy L remains nearly constant around 0.5, and the cooperation prevalence decreases as the gain factor  $r$  is increased.

To proceed, we estimate the tendency to choose strategy L in theory by computing the proportion of players where the condition  $\pi_C < \pi_L$  and  $\pi_D < \pi_L$  are both satisfied within all possible states. Here,  $\pi_C$ ,  $\pi_D$ , and  $\pi_L$  represent the payoffs of choosing C, D, and L, respectively. This reveals the tendency of individuals to choose the strategy L as a function of  $r$ , as depicted in Fig. 9(a). As shown in Fig. 9(a), the tendency to choose strategy L exhibits a stage-wise plateau in the interval  $r \in (2.15, 2.49)$ . Within this range, the fraction of loners remains fixed, while the fraction of cooperators shows a decreasing trend. As  $r$  further increases to  $r \geq 2.5$ , the tendency of loners continues to decrease, approaching a new plateau, from the results of numerical simulations, we notice a reduction in the fraction of loners and a simultaneous increase in the fraction of cooperators. Notice that, these stage-wise plateaus correspond to the present rewards that are immediately received at present, where the impact of expected payoffs in the future is not included.

The reason why the plateau well matches the non-monotonic interval is exactly because the impact of expected payoffs in the future is weakened. As depicted in Fig. 9(b), three typical cases within the non-monotonic region  $r \in (2.13, 2.50)$  are shown, where we observe that the dependence of the cooperation prevalence on the discount factor  $\gamma$  is non-monotonic, there is an optimal value of  $\gamma$  in each case yielding the highest cooperation prevalence. This is in sharp contrast with the results obtained from previous observations [44, 50], where a higher expectation (i.e. a larger  $\gamma$ ) always leads to a better outcome of evolution; two such examples outside of the non-monotonic region ( $r = 2.05$  and  $2.66$ ) are shown in Fig. 12 in Appendix C 17. This difference implies that,

within this non-monotonic interval, a high emphasis on future rewards is less favorable for the emergence of cooperation, suggesting potential challenges in maintaining long-term cooperation stability, the underlying cause of this result is the uncertainty in the environment due to changes in the strategies of other individuals.

Specifically, as discussed in the above section, L and C are more preferred than D within the three states  $s_{1,2,3}$  in typical scenarios. But this is not always the case, the inset in Fig. 9(a) shows that the Q-value for L is no longer larger than the value for D, instead their difference  $\Delta Q = Q_D - Q_L$  becomes positive and is enlarged as  $r$  increases within this region. This means that players hesitate in the choice of action D or L, leading to increasing uncertainty of the environment. This explains the non-monotonic trend observed in Fig. 9(b), where the power of long-term vision doesn't work anymore and players prioritize their immediate rewards.

Outside of this non-monotonic region, players no longer get vacillated between L and D as the environmental information becomes certain, and the expectation factor for the future again plays its role in the emergence of cooperation. By further increasing in  $r$ , the advantage of D over L becomes more obvious, and the same is also true for  $Q_C > Q_D$ , leading to the decreasing fraction of L and increasing trend in C.

## 5. CONCLUSION AND DISCUSSION

In summary, we investigate the original PGG and PGG with voluntary participation (VPGG) within the paradigm of reinforcement learning, where each player acts following a Q-learning algorithm. Different from most previous works where the self-regarding Q-learning algorithm is employed, we incorporate environmental information into the state by comparing the numbers of cooperators and defectors in their neighborhoods. We observe that the thresholds for cooperation emergence in both cases occur are much reduced compared to the case in the IL framework. The cooperation emergence is found to be most likely to occur in VPGG. Unexpectedly, a non-monotonic dependence of cooperation prevalence on the gain factor is revealed in the VPGG. By the analysis of the evolution of state, action preference, rewards, and the evolution of the Q-table, we clarify the mechanism of cooperation in both cases and the reason behind the non-monotonic dependence in VPGG.

Our research reveals significant differences in the mechanisms of cooperation emergence between the traditional IL and our RL. Within RL, the emergence of cooperation does not rely on the cooperator clustering, the so-called network reciprocity [31], instead by incorporating the historical experiences, present rewards, and the expected payoffs in the future, players avoid complete defection and cooperators emerge in the form of scattered distribution. Furthermore, in the VPGG, the choice of the loner (non-participating players) for the individuals significantly inhibits the survival of defectors, and players turn to cooperators when the loners are dominating.

Unexpectedly, there is a non-monotonic interval where the number of cooperators decreases with the gain factors  $r$ . Our analysis reveals that this is due to the enhanced environmental uncertainty, reducing the expectation of future rewards, and cooperation is thus degraded as players put more weight on the present rewards. This uncertainty induced risk aversion disappears as the gain factor  $r$  is further increased where the environment becomes less uncertain, and the cooperation prevalence continues to increase with increasing  $r$ .

Till now, reinforcement learning has been employed to explain various socio-economic activities beyond cooperation, including trust [50], resource allocation [51, 52], and other human behaviors [40, 47, 53]. In fact, there is abundant experimental evidence in neuroscience [66, 67] indicating that many decision-making processes of humans fall into RL paradigm. Notice that, we do not devalue social learning [68]. We conjecture that, in our daily lives both paradigms are relevant, and each works in some specific scenarios and together they underpin most of our decision-makings. Of course, this speculation requires further validation through behavioral experiments. We believe that the research conducted in this work is crucial for elucidating some basic questions along this research line.

#### ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [Grants Nos. 12075144, 12165014], and Fundamental Research Funds for the Central Universities (GK202401002). ZGZ is supported by Excellent Graduate Training Program of Shaanxi Normal University [Grants No. LHRCTS23064].

#### Appendix A: Phase diagrams of the learning parameters

To systematically examine the impact of the learning parameters on the cooperation evolution, we provide phase diagrams for both PGG and VPGG with Q-learning algorithm.

In PGG with Q-learning, we find a prominent emergence of cooperation is observed in a physically meaningful region, see the phase diagram Fig. 10, it reports the final fractions of cooperators in the learning parameter domain  $(\gamma, \alpha)$ . As shown, there is a white region where the cooperation emerges, where the learning rate  $\alpha$  is small and the discount factor  $\gamma$  is large. This observation means that when individuals focus on both historical experience and long-term vision, cooperation is likely to emerge. Otherwise, either a forgetful property (a large  $\alpha$ ) and/or a short-term vision (a small  $\gamma$ ) lead to the failure of its emergence.

Similar results can also be observed in the VPGG, as shown in Fig. 11. The emergence of cooperators fails with the decrease in  $\gamma$  and/or the increase in  $\alpha$ . However, the fraction of defectors remains small across the entire parameter space.

This result is easily understandable because, with the introduction of the loner strategy, individuals tend to choose strategy L with a higher payoff rather than the all-defect strategy with zero payoffs as we discuss above.

These observations are in line with our previous findings revealed in the emergence of cooperation in the prisoners' dilemma [44] or the donation game [48], and the emergence of trust [50].

#### Appendix B: Payoff analysis in VPGG

For a more intuitive understanding of the transition in individual action preferences from the perspective of payoffs, we provide payoffs for different strategies in three states  $s_{1,2,3}$ , as shown in table I to table III.

Table I shows payoffs for individuals in state  $s_1$  where  $n_C > n_D$ . Only in scenario  $(nC, 0D)$  does choosing cooperation yields a higher payoff than choosing the loner. This implies that in an environment where defectors exist, cooperators are prone to shifting toward being loners.

Table II displays payoffs for individuals in state  $s_2$  where  $n_C = n_D$ . Choosing L consistently yields the highest payoff, except in the scenario  $(0C, 0D)$ , where all individuals around are loners, facilitating the emergence of cooperation.

Table III presents payoffs for individuals in state  $s_3$  where  $n_C < n_D$ . Choosing L consistently yields the highest payoff, suggesting individuals typically opt for higher payoff strategies when in this state.

$s_1$	Action		
	C ( $a_1$ )	D ( $a_2$ )	L ( $a_3$ )
1C, 0D	1.11	1.055	1
2C, 0D	1.11	1.406	1
2C, 1D	0.5825	1.055	1
3C, 0D	1.11	1.5825	1
3C, 1D	0.688	1.266	1
4C, 0D	1.11	1.688	1

TABLE I. In the VPGG, the average payoffs when choosing different actions in state  $s_1$ , where  $n_C > n_D$ . Parameter:  $r = 2.11$ .

$s_2$	Action		
	C ( $a_1$ )	D ( $a_2$ )	L ( $a_3$ )
0C, 0D	1.11	0	1
1C, 1D	0.406	0.703	1
2C, 2D	0.266	0.844	1

TABLE II. In the VPGG, the average payoffs when choosing different actions in state  $s_2$ , where  $n_C = n_D$ . Parameter:  $r = 2.11$ .

Action $s_3$	C ( $a_1$ )	D ( $a_2$ )	L ( $a_3$ )
0C, 4D	-0.578	0	1
0C, 3D	-0.4725	0	1
0C, 2D	-0.296	0	1
0C, 1D	0.055	0	1
1C, 2D	0.055	0.5275	1
1C, 3D	-0.156	0.422	1

TABLE III. In the VPGG, the average payoffs when choosing different actions in state  $s_3$ , where  $n_C < n_D$ . Parameter:  $r = 2.11$ .

### Appendix C: Dependence on $\gamma$ outside of the non-monotonic interval

As a comparison, here we present the fraction of cooperators outside of the non-monotonic interval (at  $r = 2.05$  and  $r = 2.66$ ) as the discount factor  $\gamma$  varies, as shown in Fig. 12. We can observe that the fraction of cooperators continues to increase with the increasing  $\gamma$ , quantitatively different from the observations made in Fig. 9(b) in the non-monotonic interval. This indicates that individuals valuing future rewards are conducive to the emergence of cooperation, which aligns with our previous findings [44, 50], where a higher expectation (i.e. a larger  $\gamma$ ) always leads to a better outcome of evolution.

\* Email address: chenl@snnu.edu.cn

- [1] M. Milinski, D. Semmann, H.-J. Krambeck, and J. Marotzke, *Proceedings of the National Academy of Sciences* **103**, 3994 (2006).
- [2] P. Kollock, *Annual Review of Sociology* **24**, 183 (1998).
- [3] G. Hardin, *Science* **162**, 1243 (1968).
- [4] M. A. Nowak and K. Sigmund, *Science* **303**, 793 (2004).
- [5] M. A. Nowak, *Science* **314**, 1560 (2006).
- [6] R. Axelrod and W. D. Hamilton, *Science* **211**, 1390 (1981).
- [7] M. Doebeli and C. Hauert, *Ecology Letters* **8**, 748 (2005).
- [8] A. Rapoport and A. M. Chammah, *Prisoner's dilemma: A study in conflict and cooperation*, Vol. 165 (University of Michigan press, 1965).
- [9] C. Hauert and G. Szabó, *Complexity* **8**, 31 (2003).
- [10] M. Perc, J. Gómez-Gardeñes, A. Szolnoki, L. M. Floría, and Y. Moreno, *Journal of The Royal Society Interface* **10**, 20120997 (2013).
- [11] J. Tanimoto, *Physical Review E* **87**, 062136 (2013).
- [12] R. M. Dawes, *Annual Review of Psychology* **31**, 169 (1980).
- [13] U. Fischbacher, S. Gächter, and E. Fehr, *Economics Letters* **71**, 397 (2001).
- [14] J. Wang, F. Fu, T. Wu, and L. Wang, *Physical Review E* **80**, 016101 (2009).
- [15] H. Gintis, S. Bowles, R. Boyd, and E. Fehr, *Evolution and Human Behavior* **24**, 153 (2003).
- [16] G. Szabó and C. Hauert, *Physical Review Letters* **89**, 118101 (2002).
- [17] C. Hauert, S. D. Monte, J. Hofbauer, and K. Sigmund, *Science* **296**, 1129 (2002).
- [18] C. Hauert, S. D. Monte, J. Hofbauer, and K. Sigmund, *Journal of Theoretical Biology* **218**, 187 (2002).
- [19] D. Semmann, H.-J. Krambeck, and M. Milinski, *Nature* **425**, 390 (2003).
- [20] E. Fehr and S. Gächter, *Nature* **415**, 137 (2002).
- [21] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson, *Proceedings of the National Academy of Sciences* **100**, 3531 (2003).
- [22] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, *Physics Reports* **687**, 1 (2017).
- [23] B. Yang, M. Yang, W. Liu, and X. Chen, *Acta Physica Sinica* **66**, 196401 (2017).
- [24] K. Sigmund, C. Hauert, and M. A. Nowak, *Proceedings of the National Academy of Sciences* **98**, 10757 (2001).
- [25] A. Szolnoki and M. Perc, *Europhysics Letters*.
- [26] Y. Fang, T. P. Benko, M. Perc, H. Xu, and Q. Tan, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **475**, 20190349 (2019).
- [27] F. C. Santos, M. D. Santos, and J. M. Pacheco, *Nature* **454**, 213 (2008).
- [28] J. Guan, Z. Wu, and Y. Wang, *Physical Review E* **76**, 056101 (2007).
- [29] Z. Rong and Z. Wu, *Europhysics Letters* **87**, 30001 (2009).
- [30] C. Xia, J. Wang, M. Perc, and Z. Wang, *Physics of Life Reviews* **46**, 8 (2023).
- [31] M. A. Nowak and R. M. May, *Nature* **359**, 826 (1992).
- [32] C. P. Roca, J. A. Cuesta, and A. Sánchez, *Physics of Life Reviews* **6**, 208 (2009).
- [33] A. Bandura and R. H. Walters, *Social Learning Theory*, Vol. 1 (Englewood cliffs Prentice Hall, 1977).
- [34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT press, 2018).
- [35] S. Tanabe and N. Masuda, *Journal of Theoretical Biology* **293**, 151 (2012).
- [36] T. Ezaki, Y. Horita, M. Takezawa, and N. Masuda, *PLoS Computational Biology* **12**, e1005034 (2016).
- [37] Y. Horita, M. Takezawa, K. Inukai, T. Kita, and N. Masuda, *Scientific Reports* **7**, 39275 (2017).
- [38] H. Ding, G. Zhang, S. Wang, J. Li, and Z. Wang, *Physica A: statistical mechanics and its applications* **536**, 122551 (2019).
- [39] L. Fan, Z. Song, L. Wang, Y. Liu, and Z. Wang, *Chaos: An Interdisciplinary Journal of Nonlinear Science* **32**, 123140 (2022).
- [40] S. Zhang, J. Zhang, L. Chen, and X. Liu, *Nonlinear Dynamics* **99**, 3301 (2020).
- [41] L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, and Z. Wang, *Nonlinear Dynamics* **108**, 1837 (2022).
- [42] L. Wang, L. Fan, L. Zhang, R. Zou, and Z. Wang, *New Journal of Physics* **25**, 073008 (2023).
- [43] Z. He, Y. Geng, C. Du, L. Shi, and Z. Wang, *New Journal of Physics* **24**, 123038 (2022).
- [44] Z. Ding, G. Zheng, C. Cai, W. Cai, L. Chen, J. Zhang, and X. Wang, *Chaos, Solitons & Fractals* **175**, 114032 (2023).
- [45] Y. Geng, Y. Liu, Y. Lu, C. Shen, and L. Shi, *Applied Mathematics and Computation* **427**, 127182 (2022).
- [46] Z. Yang, L. Zheng, M. Perc, and Y. Li, *Applied Mathematics and Computation* **463**, 128364 (2024).
- [47] Y. Shi and Z. Rong, *IEEE Transactions on Circuits and Systems II: Express Briefs* **69**, 2463 (2022).
- [48] J. Zhang, Z. Rong, G. Zheng, J. Zhang, and L. Chen, *Journal of Physics: Complexity* **5**, 025006 (2024).
- [49] C. Zhao, G. Zheng, C. Zhang, J. Zhang, and L. Chen, *Chaos: An Interdisciplinary Journal of Nonlinear Science* **34**, 073123 (2024).
- [50] G. Zheng, J. Zhang, J. Zhang, W. Cai, and L. Chen, *New Journal of Physics* **26**, 053041 (2024).
- [51] M. Andreucut and M. Ali, *Physical Review E* **64**, 067103 (2001).



- [52] S. Zhang, J. Dong, L. Liu, Z. Huang, L. Huang, and Y.-C. Lai, [Physical Review E \*\*99\*\*, 032302 \(2019\)](#).
- [53] M. S. Tomov, E. Schulz, and S. J. Gershman, [Nature Human Behaviour \*\*5\*\*, 764 \(2021\)](#).
- [54] T. Ezaki and N. Masuda, [PloS one \*\*12\*\*, e0189220 \(2017\)](#).
- [55] N. Masuda and M. Nakamura, [Journal of Theoretical Biology \*\*278\*\*, 55 \(2011\)](#).
- [56] D. Jia, H. Guo, Z. Song, L. Shi, X. Deng, M. Perc, and Z. Wang, [New Journal of Physics \*\*23\*\*, 083020 \(2021\)](#).
- [57] Z. Song, H. Guo, D. Jia, M. Perc, X. Li, and Z. Wang, [Neurocomputing \*\*513\*\*, 104 \(2022\)](#).
- [58] D. Jia, T. Li, Y. Zhao, X. Zhang, and Z. Wang, [Applied Mathematics and Computation \*\*413\*\*, 126658 \(2022\)](#).
- [59] H. Zhang, T. An, P. Yan, K. Hu, J. An, L. Shi, J. Zhao, and J. Wang, [Chaos, Solitons & Fractals \*\*178\*\*, 114358 \(2024\)](#).
- [60] C. J. C. H. Watkins, *Learning from delayed rewards (Ph.D. thesis)*, Ph.D. thesis (1989).
- [61] P. Watkins, Christopher J. C. H. and Dayan, [Machine Learning \*\*8\*\*, 279 \(1992\)](#).
- [62] G. Szabó and C. Tóke, [Phys. Rev. E \*\*58\*\*, 69 \(1998\)](#).
- [63] L. S. Flores, M. A. Amaral, M. H. Vainstein, and H. C. Fernandes, [Chaos, Solitons & Fractals \*\*164\*\*, 112744 \(2022\)](#).
- [64] M. Perc, A. Szolnoki, and G. Szabó, [Physical Review E \*\*78\*\*, 066101 \(2008\)](#).
- [65] A. Szolnoki and M. Perc, [The European Physical Journal B \*\*67\*\*, 337 \(2009\)](#).
- [66] D. Lee, H. Seo, and M. W. Jung, [Annual Review of Neuroscience \*\*35\*\*, 287 \(2012\)](#).
- [67] A. Rangel, C. Camerer, and P. Montague, [Nature Reviews. Neuroscience \*\*9\*\*, 545 \(2008\)](#).
- [68] A. Olsson, E. Knapska, and B. Lindström, [Nature Reviews Neuroscience \*\*21\*\*, 197 \(2020\)](#).

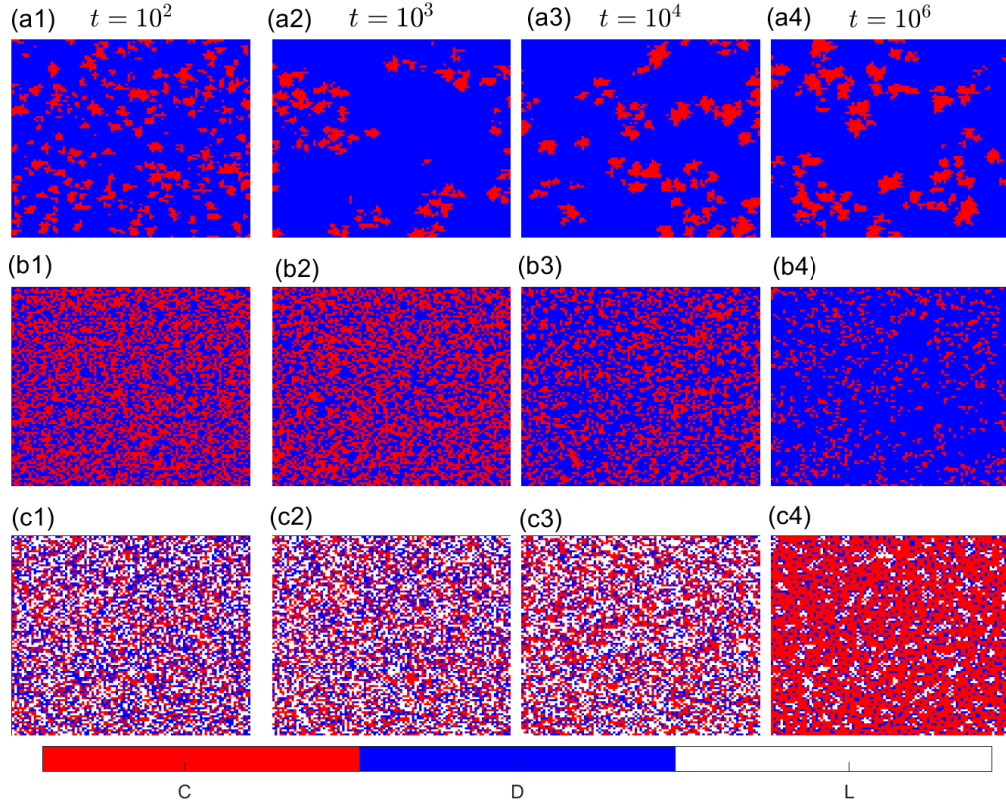


FIG. 3. (Color online) **Typical cooperation snapshots for the three setups.** In (a1-a4), PGG with Fermi rule, and cooperation clusters are formed to resist the exploitation by defectors. The gain factor  $r = 4.60$ . In (b1-b4), PGG with the Q-learning algorithm, no apparent clusters of cooperators are seen instead, and the evolution time takes longer. In (c1-c4), VPGG with the Q-learning algorithm, there are also no obvious cooperation clusters, but the resulting cooperation prevalence  $f_c$  in panel (c4) is much higher than the panel (b4).  $\epsilon = 0.01$ ,  $\alpha=0.1$ ,  $\gamma=0.9$ ,  $\sigma=1.0$ ,  $r = 3.32$  are used in both (b1-b4) and (c1-c4). Other parameters: the population size  $N = 100 \times 100$ .

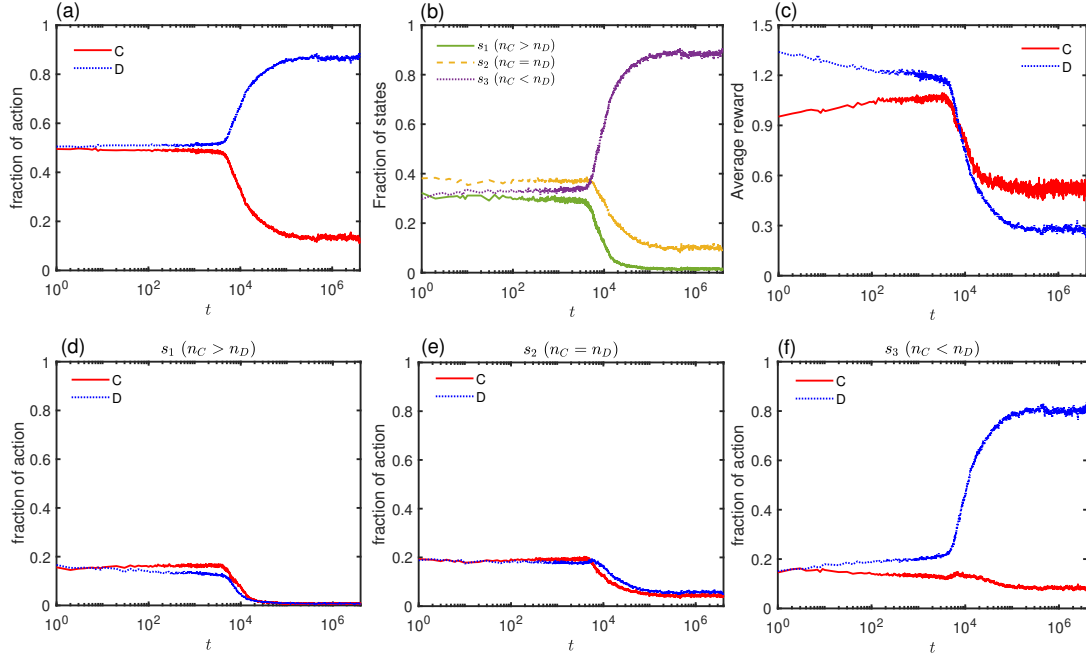


FIG. 4. (Color online) **Typical time series in PGG with Q-learning.** (a) Time evolution of two strategies, which differs from the typical “first down and later up” trend often observed as a sign of network reciprocity in the framework of imitation learning. (b) The temporal evolution of the fractions of the three states  $s_{1,2,3}$ . After the transient period,  $s_3$  dominates. (c) The evolution of average reward for two different strategies, where there is a crossover observed. (d-f) Time evolution of action preference in three different states. The emergence of cooperation primarily occurs in state  $s_1$  ( $n_C < n_D$ ). Other parameters:  $\epsilon = 0.01$ ,  $\alpha=0.1$ ,  $\gamma=0.9$ ,  $\sigma=1.0$ ,  $r=3.32$ , and the population size  $N = 100 \times 100$ .

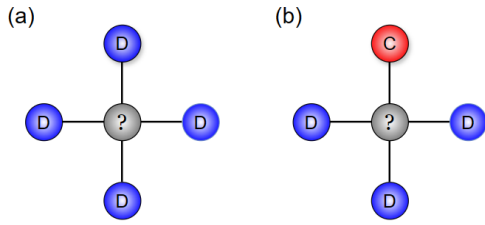


FIG. 5. (Color online) **Two scenarios within state  $s_3$  where  $n_C < n_D$ .** The gray circle represents the focal individual, who is either to cooperate or to defect. Red and blue circles represent the cooperators and defectors in neighbors, respectively. Cooperation is almost impossible in scenario (a), while it may occur in scenario (b).

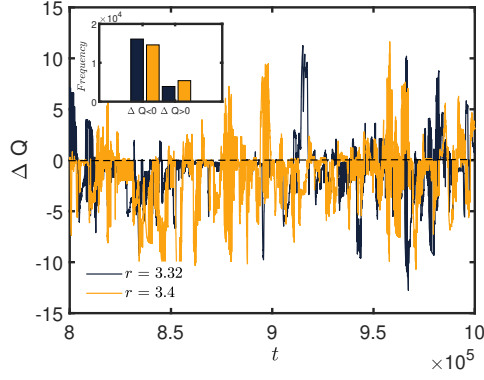


FIG. 6. (Color online) **Time evolution of  $\Delta Q$  after transient within state  $s_3$ .** Typical time evolution of  $\Delta Q = Q_{s_3,C} - Q_{s_3,D}$  in state  $s_3$  ( $n_C < n_D$ ). The inset shows the frequencies of occurrences when  $\Delta Q > 0$  and  $\Delta Q < 0$  at the same time intervals. Two parameters  $r = 3.32$  (black) and  $r = 3.4$  (orange) are chosen here for comparison. Other parameters:  $\epsilon = 0.01$ ,  $\alpha = 0.1$ ,  $\gamma = 0.9$ , and the population size  $N = 100 \times 100$ .

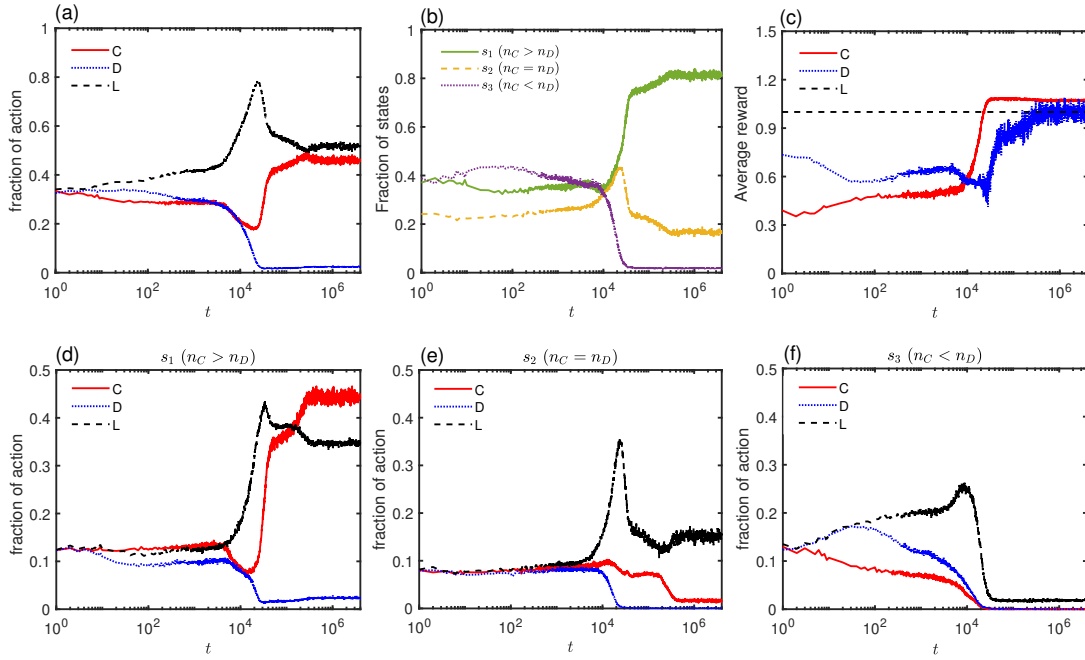


FIG. 7. (Color online) **Typical time series in VPGG with Q-learning.** (a) Time evolution of three strategies. As the defectors begin to decline, the fraction of cooperators shifts from a decreasing to an increasing trend, while the corresponding loners shift from an increasing to a decreasing trend, indicating the emergence of cooperators within clusters of loners. (b) Time evolution of the fractions of the three states. After a sufficiently large number of Monte Carlo steps, it can be observed that the system is most likely in the state of  $s_1$ , while there exists a certain proportion of states  $s_2$  and  $s_3$  in dynamic equilibrium. (c) The evolution of average reward for three different strategies. In the early stages, loners consistently generate higher average payoffs compared to the other strategies. (d-f) Time evolution of action preference in three states  $s_{1,2,3}$ . The emergence of cooperation primarily occurs in state  $s_1$ . Other parameters:  $\epsilon = 0.01$ ,  $\alpha = 0.1$ ,  $\gamma = 0.9$ ,  $r = 2.11$  and the population size  $N = 100 \times 100$ .



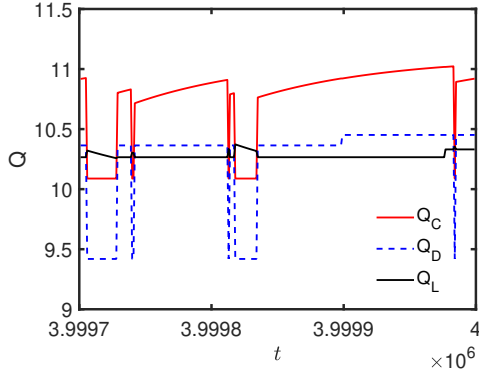


FIG. 8. (Color online) **Time evolution of Q value in a stable state.** In VPGG, time series of Q-value for three different strategies in a stable state. Other parameters:  $\epsilon = 0.01$ ,  $\alpha = 0.1$ ,  $\gamma = 0.9$ , and the population size  $N = 100 \times 100$ .

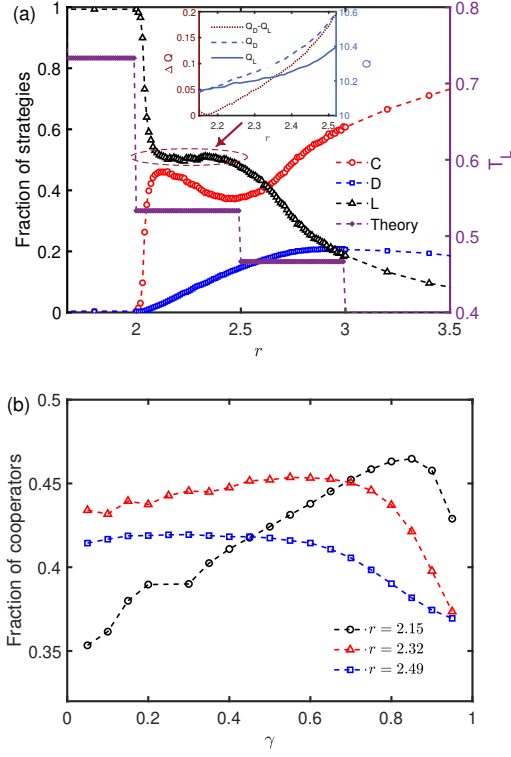


FIG. 9. (Color online) **The analysis of non-monotonic interval in VPGG.** (a) The left side of the y-axis shows the fraction of strategies as a function of the gain factor  $r$ . The right side shows the tendency of choosing strategy L based on theoretical analysis, characterized by the proportion of players where both conditions of  $\pi_L > \pi_C$  and  $\pi_L > \pi_D$  are satisfied. The inset shows the corresponding values of  $\Delta Q$  and  $Q_{L,D}$ . (b) The fraction of cooperators as the discount factor  $\gamma$  varies for three different  $r$  values. Other parameters:  $\epsilon = 0.01$ ,  $\alpha = 0.1$ ,  $\gamma = 0.9$ , and the population size  $N = 100 \times 100$ .

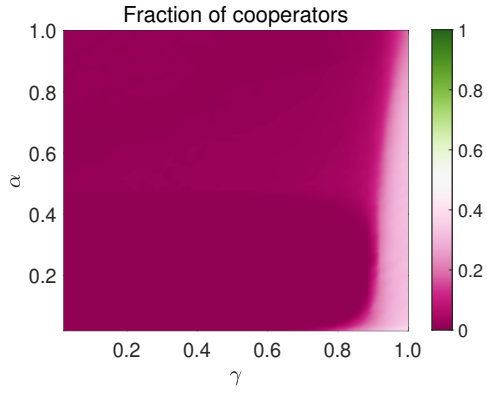


FIG. 10. (Color online) The color-coded stationary fraction of cooperators in the domain  $(\gamma, \alpha)$  for PGG with the Q-learning algorithm. Cooperation emerges at the corner of large  $\gamma$  and small  $\alpha$ , indicating that when individuals focus on both historical experience and the long-term vision, cooperation emerges. Each data is averaged 100 times after a transient of  $5 \times 10^5$ . Other parameters:  $\epsilon = 0.01$ ,  $r = 3.32$ , and the population size  $N = 100 \times 100$ .

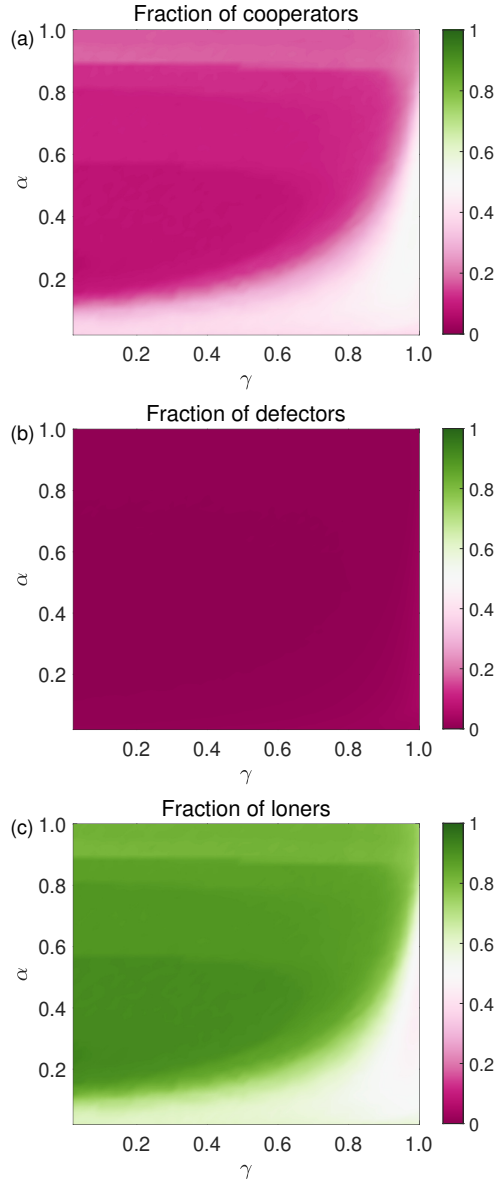


FIG. 11. (Color online) The color-coded stationary fraction of the three strategies in the domain  $(\gamma, \alpha)$  for VPGG with the Q-learning algorithm. Cooperation emerges at the corner of large  $\gamma$  and small  $\alpha$ , showing that when individuals focus on both historical experience and the long-term vision, cooperation emerges. Each data is averaged 100 times after a transient of  $5 \times 10^5$ . Other parameters:  $\epsilon = 0.01$ ,  $r = 2.11$ , and the population size  $N = 100 \times 100$ .

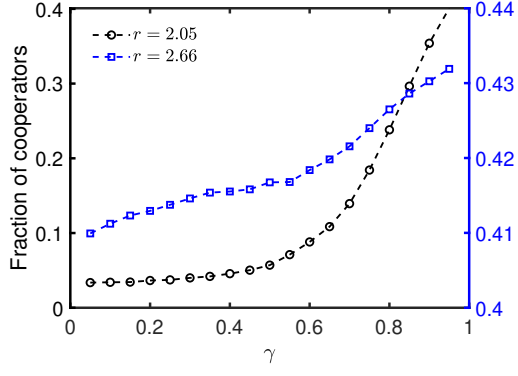


FIG. 12. (Color online) The fraction of cooperators outside of the non-monotonic interval. We can see that the fraction of cooperation increases with the increase of  $\gamma$ . Other parameters:  $\epsilon = 0.01$ ,  $\alpha = 0.1$ , and the population size  $N = 100 \times 100$ .