Evolutionary dynamics under combined regret-based learning and random exploration

Mengfan Zhu[a], Jianhua Xu[a], Xiaojie Chen[a,*], Attila Szolnoki[b]

[a]School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China
[b]Institute of Technical Physics and Materials Science, Centre for Energy Research, P.O. Box 49, H-1525 Budapest, Hungary

Abstract

Individual decisions in population games are often grounded on various motivations, including regret-based learning or random exploration of the strategy space. Despite the intensively growing interest in these microscopic rules, their simultaneous consequences on strategic behaviors remain largely unexplored. Motivated by this shortage, here we propose a novel protocol that combines regret-based learning and random exploration into a general two-player, two-strategy game between individuals and their neighbors arranged in a network. During the evolutionary process, agents randomly explore alternative strategies with a certain probability or employ regret-based learning using a Boltzmann-type regret function. We derive an analytical condition under which a strategy can prevail and find that the condition depends solely on the game parameters, and is independent of the regret function, random exploration rate, and network structure under weak regret strength. On the other hand, the chance of a random exploration weakens the evolutionary advantage of the dominant strategy and enhances the fitness of the less-favored one. Furthermore, we reveal through computer simulations that increasing the regret strength enhances the position of more dominant strategy, while the evolutionary chances is suppressed when it is not favored.

Keywords: Evolutionary dynamics, population games, regret-based learning, random exploration, strategy evolution

1. Introduction

Understanding and predicting the evolutionary dynamics of emergent behavior in a large population of individuals engaged in strategic interactions has became an intensively studied research field in recent years [1, 2, 3, 4, 5]. A focal question is how microscopic strategy updating procedures influence the final evolutionary outcomes, which can provide new insights into understanding, analysis, and design of multi-agent learning algorithms [2]. During the past decades, several update rules have been developed to mimic how agents make decisions in gaming environments [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. It is important to note that these protocols are primarily motivated by the exploitation or exploration principles [19, 20].

Indeed, the integration of exploitation and exploration has been proved to be particularly important for the efficiency with which protocols lead to different evolutionary patterns of collective behavior in complex networks [21, 22, 23, 24, 25, 26, 27]. It is worth mentioning that the regret value is often considered as a key exploitation metric. Essentially, the regret value is determined by the difference between the maximum attainable payoff and the actual payoff received by an agent [28, 29]. In the context of exploration, random exploration is a frequently considered approach. The actual protocols used in this regard are softmax and ε-greedy exploration [19]. Recent studies have focused on proposing multi-agent learning algorithms that incorporate random exploration and regret minimization [30, 31]. The rationale of such algorithms is

*Corresponding author
Email address: xiaojiechen@uestc.edu.cn (Xiaojie Chen)

reflected by the phenomena in human and artificial societies that agents may make decisions based on the regret values of different strategic actions, but they may also take actions through random exploration of the strategy space.

However, the majority of related works focused on investigating the separate effects of regret-based learning or random exploration on the evolutionary outcomes of collective actions, in which the results are mainly obtained by computer simulations. The systematic theoretical analysis is lacking, which would be essential to get a deeper understanding about their simultaneous consequences in realistic scenarios of game interactions.

To address this shortage, our present goal is to theoretically investigate the impact of regret-based learning and random exploration on evolutionary dynamics in network games. In this work, we propose a combined learning protocol based on the regret value and random exploration into a population where agents play a general two-player, two-strategy non-cooperative game on any connected graphs. We assume that each individual has two strategies A or B to choose. During the evolutionary process, individuals may select a randomly chosen strategy with a certain probability. Otherwise, they update their strategy using the Boltzmann exploration mechanism involving the information of regret values. Additionally, for convenience, we take the average frequency of strategy A as a key quantity to depict the evolutionary outcome. We then focus on how the average frequency of strategy A changes with this protocol. We derive the theoretical condition for strategy A to be favored on any connected network structure by employing the Markov chain and matrix theory under weak regret strength. We find that the strategy success condition depends only on the elements of the game payoff matrix, and is independent of the regret values of strategies and random exploration rate. Additionally, we find that whether strategy A prevails over strategy B under the learning procedure is independent of the network structure and regret function form. We find that the average frequency of strategy A depends on the regret function and random exploration rate. When strategy A is more favored over strategy B, the average frequency of strategy A decreases as the random exploration rate increases. Otherwise, the average frequency of strategy A increases as the random exploration rate increases. In addition, we perform computer simulations and confirm that these results are valid on different representative network structures and various forms of regret functions. Furthermore, we show through computer simulations that increasing the regret strength enhances the dominance of strategy A when it is more favored, but suppresses its evolution when it is not favored. This result indicates the dynamic of "the strong get stronger, the weak get weaker" when the regret strength is not weak.

The remainder of this paper is organized as follows. In Section 2, we propose the strategy learning protocol that incorporates the regret factor and random exploration into network games. In Section 3, we present a detailed theoretical analysis under weak regret strength and present the mathematical condition for the evolutionary success of strategy A. In Section 4, we provide numerical and simulation results to verify our theoretical predictions. Finally, our conclusion and discussion are presented in Section 5.

2. Model

We consider a structured population of size N ($N \geq 2$). The population structure is described by a weighted network in which each node represents an agent, and the edges represent interactions between agents. Specifically, the edge weight w_{ij} between agent i and j is denoted as w_{ij}. Here $w_{ij} = 0$ implies that there is no connection between agent i and j, while $w_{ij} > 0$ means that the two agents can interact with each other. Accordingly, the weighted degree of node i is $d_i = \sum_{j=1}^{N} w_{ij}$. Here we consider an undirected network, i.e., $w_{ij} = w_{ji}$.

We consider that the two-player, two-strategy game is played on the network. In this game, there are two strategies and each agent can choose strategy A or B. Specifically, when A meets A, both gain the payoff a. When A meets B, the former gets the payoff b and the latter gets the payoff c. When B meets B, both gain the payoff d. On the network, each individual l plays the games with all its neighbors and collects the degree-weighted average payoff p_l [32]. On the other hand, each individual can have the maximum possible payoff by playing the games with neighbors according to the strategy pattern in the neighborhood. We therefore assume that an agent l has the personalized possible maximum payoff, defined by p_l^*. Based on

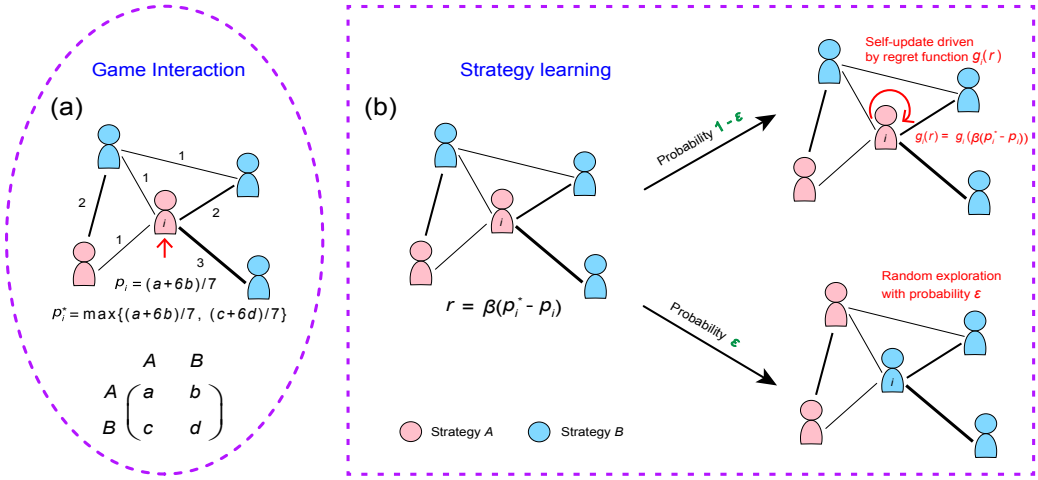

Fig. 1. Illustration of game interactions and strategy learning on graphs. (a) Individuals engage in a general two-player, two-strategy game with neighbors on a connected, weighted graph. The payoff matrix of general two-person game is given at the bottom of panel (a). Pink (blue) color represents strategy A (B). (b) Individuals update their strategies based on the regret function and random exploration. When an agent i interacts with its neighbors, it receives the actual average payoff p_i and identifies the maximum possible average payoff p_i^* according to the strategy pattern and the weighted connections in its neighborhood. Consequently, the regret value of agent i is denoted as $r = \beta(p_i^* - p_i)$, where β is the regret strength. The right part in panel (b) shows that agent i chooses a random exploration of strategies with probability ε, otherwise it updates its strategy based on the regret function associated with the defined regret value.

the actual payoff and the maximum attainable payoff, we define the $r \geq 0$ regret value for individual l as

$$r = \beta(p_l^* - p_l), \tag{1}$$

where the coefficient β ($\beta > 0$) is the strength of regret. In particular, $\beta = 0$ means neutral drift, implying that individual l does not distinguish the regret values for choosing strategy A or B. $\beta \to 0$ represents a weak regret strength, indicating that individual l has no obvious preference to distinguish the merits of regret in choosing different strategies. Furthermore, we consider that each individual l has its regret function $g_l(r)$ with respect to the regret value r, so that we have a generalized form of regret value for each individual. Furthermore, we assume that the regret function has the following properties:

- $g_l(r)$ is a strictly increasing function, i.e., $g_l'(r) > 0$, ensuring that individuals with higher regret values should have higher generalized regret merit;
- $g_l(0) > 0$ is assumed to prevent individual strategies from not being updated at the neutral drift.

Subsequently, individuals will update their strategies based on the proposed regret functions and random exploration. To be more specific, at each time step, an individual is randomly chosen and has the opportunity to update its strategy. Without losing generality, we assume that an individual l is chosen who updates its current strategy choice with probability ε, where $\varepsilon \geq 0$ is the random exploration rate. Otherwise, individual l updates the strategy choice based on the regret function according to the Boltzmann exploration mechanism [19]. Individual l will switch strategy A to strategy B with probability

$$P(A \to B) = \frac{g_l(r_A)}{\sum_s g_l(r_s)}, \tag{2}$$

otherwise, it keeps the original strategy A. Alternatively, an individual l having strategy B switches to strategy A with probability

$$P(B \to A) = \frac{g_l(r_B)}{\sum_s g_l(r_s)}, \tag{3}$$

otherwise, it maintains strategy B. Here r_s represents the regret value of individual l given that it interacts with neighbors with strategy s and $s \in \{A, B\}$. The details of our model are summarized in Fig. 1.

Based on the above description, we aim to explore how the proposed learning protocol affects individuals' strategy choices in the population. The evolutionary outcome is characterized by the average frequency of strategy A. We focus on the effects of random exploration and regret-based learning.

3. Theoretical Results

In this section, we present theoretical analysis for the evolutionary dynamics of strategies and derive the mathematical condition under which strategy A can prevail over strategy B in the competitive gaming environment, that is, the average frequency of strategy A, $\langle x_A \rangle > 1/2$.

To identify this condition, we first derive the average frequency of strategy A, $\langle x_A \rangle$. We define s_l as the strategy employed by individual l and $s_l \in \{0, 1\}$. $s_l = 1$ denotes that individual l chooses strategy A and $s_l = 0$ means that individual l chooses strategy B. The strategy choice of the population can be represented by a column vector $\mathbf{s} = (s_1, s_2, ..., s_l, ..., s_N)^T$. Therefore, the number of all the states in the network game is $Z = 2^N$, and the state space can be described by $\{\mathbf{s}_1, ..., \mathbf{s}_Z\}$. In state $\mathbf{s}$, the frequency of strategy A can be expressed as $x_A(\mathbf{s}) = \frac{1}{N}\sum_{l=1}^{N} s_l$. We can then set a column vector $\mathbf{x} = (x_A(\mathbf{s}_1), ..., x_A(\mathbf{s}_Z))^T$ to represent the frequency of strategy A at different states. We note that the evolutionary dynamics of strategy can be depicted by a Markov chain with the state space $\{\mathbf{s}_1, ..., \mathbf{s}_Z\}$ and the Markov chain is irreducible and aperiodic. Hence, there exists a unique stationary distribution $\mathbf{u}$. Furthermore, the transition probabilities between all Z states can be characterized by a $Z \times Z$ Markov matrix $\mathbf{P} = [p_{ij}]_{Z\times Z}$. Accordingly, we can calculate the stationary distribution $\mathbf{u}$ for the Markov matrix and hence the average frequency of strategy A is $\langle x_A \rangle = \mathbf{u} \cdot \mathbf{x}$.

We now calculate the elements of Markov matrix, so that we can obtain the stationary distribution $\mathbf{u}$. We note that the evolutionary process of strategies driven by individual regret values and random exploration can be characterized by a Markov matrix with 2^N states. At each time step, an individual l is randomly chosen from the population to update the strategy. For state $\mathbf{s}_i$, we then define the adjacent state of $\mathbf{s}_i$ as $\mathscr{N}_i$ when the strategy of individual l changes from A to B. The adjacent state of $\mathbf{s}_i$ is denoted as $\mathscr{M}_i$ when the strategy of the individual l switches from B to A. Hence, the above mentioned transition probability p_{ij} for individual l can be written as

$$p_{ij} = \begin{cases} \dfrac{1}{N} \cdot G_{ij}(r_A), & \text{if } j \in \mathcal{N}_i, \\ 1 - \dfrac{1}{N}\left[\displaystyle\sum_{k\in\mathcal{N}_i} G_{ik}(r_A) - \sum_{k\in\mathcal{M}_i} G_{ik}(r_B)\right], & \text{if } j = i, \\ \dfrac{1}{N} \cdot G_{ij}(r_B), & \text{if } j \in \mathcal{M}_i, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $G_{ij}(r_s) = \frac{g_l(r_s)}{g_l(r_A)+g_l(r_B)}(1-\varepsilon) + \varepsilon$ and s represents strategy A or B.

Instead of deriving a direct mathematical expression for $\mathbf{u}$, we make a first-order Taylor approximation on the Markov matrix, stationary distribution, and average frequency of strategy A with respect to the regret strength at $\beta = 0$. In other words, we consider the case of weak regret strength, i.e., $\beta \to 0$. We stress that for weak regret strength, there is little difference between the regret values for individuals, but it makes

sense to quantify the effect compared to the neutral drift during a long evolutionary process. Accordingly, we have

$$\begin{cases} \mathbf{u} & = \mathbf{u_0} + \mathbf{u_1}\beta + o(\beta^2), \\ \mathbf{P} & = \mathbf{P_0} + \mathbf{P_1}\beta + o(\beta^2), \\ \langle x_A \rangle & = \langle x_A \rangle_0 + \langle x_A \rangle_1 \beta + o(\beta^2) \\ & = \mathbf{u_0 x} + \mathbf{u_1 x} \cdot \beta + o(\beta^2), \end{cases} \tag{5}$$

where $\mathbf{u_0} = \mathbf{u}|_{\beta=0}$, $\mathbf{u_1} = \frac{d}{d\beta}\mathbf{u}|_{\beta=0}$, $\mathbf{P_0} = \mathbf{P}|_{\beta=0}$, and $\mathbf{P_1} = \frac{d}{d\beta}\mathbf{P}|_{\beta=0}$. Specifically, we have

$$\mathbf{P}_0 = \begin{cases} \frac{1}{2}(1-\varepsilon), & \text{if } j = i, \\ \frac{1}{N} \cdot \frac{1}{2}(1-\varepsilon), & \text{if } j \in \mathcal{N}_i \cup \mathcal{M}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Since $\mathbf{P_0}$ is symmetric and $\mathbf{u_0} = \mathbf{u_0 P_0}$, we obtain

$$\mathbf{u_0} = \left[\frac{1}{2^N}, ..., \frac{1}{2^N} \right], \tag{7}$$

and then we easily gain $\langle x_A \rangle_0 = \mathbf{u_0} \cdot \mathbf{x} = 1/2$.

Next, we focus on the theoretical calculation of $\langle x_A \rangle_1$ in order to directly calculate the average frequency of strategy A, $\langle x_A \rangle$. Here, according to previous works [33, 34], we have the following conclusion about the calculation of $\langle x_A \rangle_1$, given as

$$\langle x_A \rangle_1 = \mathbf{u_0 P_1 Q}, \tag{8}$$

where $\mathbf{Q} = (\mathbf{I} + \mathbf{P_0} - \mathbf{F})\mathbf{x}$, $\mathbf{I}$ is an identity matrix, and $\mathbf{F} = \mathbf{e} \cdot \mathbf{e^T}$. Here $\mathbf{e}$ is a $Z \times 1$ matrix consisting entirely of ones. Further details of the calculation for Eq. (8) are provided in Appendix A.

Furthermore, since $\mathbf{P_1} = \frac{d}{d\beta}\mathbf{P}|_{\beta=0}$, we have

$$\mathbf{P}_1 = \begin{cases} -\frac{\delta}{N} \sum\limits_{j \in \mathcal{N}_i} f(s_i), & \text{if } j \in \mathcal{N}_i, \\ \frac{\delta}{N} \left[\sum\limits_{j \in \mathcal{N}_i} f(s_i) - \sum\limits_{j \in \mathcal{M}_i} f(s_i) \right], & \text{if } j = i, \\ \frac{\delta}{N} \sum\limits_{j \in \mathcal{M}_i} f(s_i), & \text{if } j \in \mathcal{M}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here, $f(\mathbf{s}_i) = p_A(\mathbf{s}_i) - p_B(\mathbf{s}_i)$ means the payoff differences when individual l respectively adopts strategy A and strategy B in state $\mathbf{s}_i$ and $\delta = \frac{g_l'(0)}{4g_l(0)}(1-\varepsilon)$.

In addition, we have

$$p_A(\mathbf{s}_i) = \sum_{j=1}^{N} \frac{w_{lj}}{d_l}[as_l s_j + bs_l(1-s_j)], \tag{10}$$

$$p_B(\mathbf{s}_i) = \sum_{j=1}^{N} \frac{w_{lj}}{d_l}[c(1-s_l)s_j + d(1-s_l)(1-s_j)]. \tag{11}$$

Since $\mathbf{Q}$ is a $Z \times 1$ column vector, we can set $\mathbf{Q} = (q_1, ..., q_Z)^T$. We note that $\mathbf{x} = (x_A(\mathbf{s}_1), ..., x_A(\mathbf{s}_Z))$, so there are C_N^k states in which the fraction of individuals choosing strategy A is always k/N, indicating that in these states k individuals choose strategy A in the population. Here $k = 0, ..., N$, so all states $\mathbf{s}_i$ with the same number of individuals with strategy A can be merged into one category and redefined as $\mathbf{S}'_k$ $(k = 0, ..., N)$. Since $2^N = C_N^0 + \cdots + C_N^N$, we can divide the 2^N states into $N+1$ categories to simply present the jth element $q_j (j = 1, 2, ..., 2^N)$ in $\mathbf{Q}$. By setting $\alpha = \frac{1}{2}(1-\varepsilon)$ and $\gamma = \frac{1}{2}(1+\varepsilon)$, the jth element of $\mathbf{Q}$ is given by

$$q_j = \frac{1}{N}\left[k\alpha + \frac{(k+1)N - 2k}{N}\gamma\right] + \frac{k}{N} - 2^{N-1}. \tag{12}$$

By substituting Eq. (7) and Eqs. (9)-(12) into Eq. (8), we have (further details are shown in Appendix B)

$$\langle x_A \rangle_1 = \frac{2N-1-\varepsilon}{16N^2} \cdot \sum_{l=1}^{N} \frac{g'_l(0)}{g_l(0)}(1-\varepsilon)[a+b-(c+d)]. \tag{13}$$

Furthermore, according to Eq. (5), we obtain

$$\langle x_A \rangle = \frac{1}{2} + \frac{2N-1-\varepsilon}{16N^2} \cdot \sum_{l=1}^{N} \frac{g'_l(0)}{g_l(0)}(1-\varepsilon)\beta(a+b-c-d)) + o(\beta^2), \tag{14}$$

(details of the derivation are provided in Appendix B).

Based on these calculations, we then have the following conclusion under weak regret strength according to theorem 1.

Theorem 1. *(1) If $a+b > c+d$, then $\langle x_A \rangle_1 > 0$, and hence strategy A is favored over strategy B under weak regret strength, independently of random exploration rate ε, network structure, and regret function. (2) When $a+b > c+d$, $\langle x_A \rangle$ monotonously decreases as the random exploration rate increases. When $a+b < c+d$, $\langle x_A \rangle$ increases monotonously as the random exploration rate increases.*

Proof. (1) According to Eq. (14), we find that if $\langle x_A \rangle > 1/2$ holds, then strategy A can prevail over strategy B. Thus, we can conclude that if $\langle x_A \rangle_1 > 0$, strategy A is favored over strategy B under weak regret strength. Since $\frac{1}{2} + \frac{2N-1-\varepsilon}{16N^2} > 0$, $\sum_{l=1}^{N} \frac{g'_l(0)}{g_l(0)} > 0$, $1-\varepsilon > 0$, and $\beta > 0$, $\langle x_A \rangle_1 > 0$ is equivalent to $a+b > c+d$. Thus, we can conclude that once $a+b > c+d$, strategy A prevails in the population, regardless of the applied regret functions, network structures, and random exploration ε.

(2) To further explore how ε influences the value of $\langle x_A \rangle$, we calculate the derivative of $\langle x_A \rangle$ with respect to ε and obtain that

$$\frac{d}{d\varepsilon}\langle x_A \rangle = \frac{\varepsilon - N}{8N^2} \cdot \sum_{l=1}^{N} \frac{g'_l(0)}{g_l(0)}\beta[a+b-(c+d)]. \tag{15}$$

□

We find that once $a+b > c+d$, $\frac{d}{d\varepsilon}\langle x_A \rangle < 0$ since $\varepsilon < N$ under weak regret strength. In this case, $\langle x_A \rangle$

decreases monotonously as the random exploration increases. While for $a+b<c+d$, $\frac{d}{d\varepsilon}\langle x_A\rangle>0$. In this situation, $\langle x_A\rangle$ increases monotonously as the random exploration increases.

Remark: From Theorem 1, we can directly determine whether strategy A can prevail over strategy B. Furthermore, we can numerically calculate the average frequency of strategy A in the population based on Eq. (14). We note that Eq. (14) is obtained by using the degree-weighted average payoff for individuals. Indeed, if we use the accumulated payoff for individuals, we find that the success condition for strategy A remains unchanged. For more details, see Appendix C.

4. Simulation Results

Algorithm 1. Calculation of $\langle x_A\rangle$

Input: Population size N, simulation runs M, T_{trans}, T_{sample}, payoff elements a,b,c,d, network $\mathcal{G}$, exploration rate ε, regret intensity β
Output: $\langle x_A\rangle$
for sim $=1$ **to** M **do**
  Initialize strategy distribution
  $x_A^{\text{sum}}\leftarrow 0$
  for $t=1$ **to** $T_{\text{trans}}+T_{\text{sample}}$ **do**
    Randomly select node i
    Compute p_i^A, p_i^B
    $p_i^*\leftarrow\max(p_i^A,p_i^B)$
    $r_i^A\leftarrow\beta(p_i^*-p_i^A)$
    $r_i^B\leftarrow\beta(p_i^*-p_i^B)$
    $P_{A\to B}\leftarrow\dfrac{g_i(r_i^A)}{g_i(r_i^A)+g_i(r_i^B)}$
    $P_{B\to A}\leftarrow\dfrac{g_i(r_i^B)}{g_i(r_i^A)+g_i(r_i^B)}$
    Draw $r\sim\mathcal{U}(0,1)$
    if $r<\varepsilon$ **then**
      Flip strategy of i
    else
      Draw $R\sim\mathcal{U}(0,1)$
      if $strategy(i)=1$ **and** $R<P_{A\to B}$ **then**
        $strategy(i)\leftarrow 0$
      else if $strategy(i)=0$ **and** $R<P_{B\to A}$ **then**
        $strategy(i)\leftarrow 1$
      end if
    end if
    if $t>T_{\text{trans}}$ **then**
      $n_A\leftarrow\sum_{j=1}^{N}strategy(j)$
      $x_A^{\text{sum}}\leftarrow x_A^{\text{sum}}+n_A/N$
    end if
  end for
  $x_A^{(\text{sim})}\leftarrow x_A^{\text{sum}}/T_{\text{sample}}$
end for
Return: $\langle x_A\rangle=\dfrac{1}{M}\sum_{\text{sim}=1}^{M}x_A^{(\text{sim})}$

In this section, we summarize the simulation results of evolutionary outcomes induced by the protocol which combines the regret function with random exploration. To verify our theoretical results, we apply four representative interaction networks. These are complete, regular, Barabási-Albert (BA) scale-free [35], and Watts-Strogatz (WS) small-world networks [36]. The regular graph is represented by a ring, where each node connects its $K=6$ nearest neighbors. The BA scale-free is constructed by using the preferential-attachment rule starting from the initial number of nodes $m_0=3$, where each new node is connected to $m=3$ existing nodes during the graph generation. The WS small-world network starts from a regular ring, where each

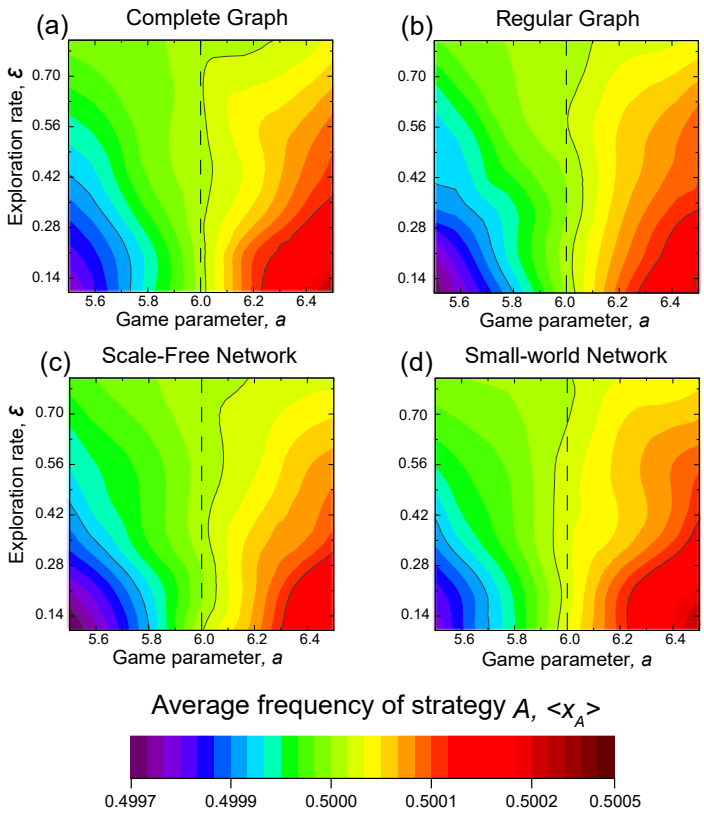

Fig. 2. The average frequency of strategy A as a function of the game parameter a and random exploration rate ε on four different networks: complete graph (a), regular graph (b), scale-free network (c), and small-world network (d). Here $b = 0$, $c = 5$, $d = 1$, the regret strength $\beta = 0.01$, and the network size for each network structure is set to $N = 100$. The black vertical dash line in each panel corresponds to the case of $a + b = c + d$.

node connects to its $K = 6$ nearest neighbors and the rewiring probability is set to $p = 0.1$. In all cases we used the same network size $N = 100$. For simplicity, we fix the values of parameters b, c, and d at 0, 5, and 1 and vary the parameter a between 5 and 7. We compute the average frequency of strategy A, $\langle x_A \rangle$, which is obtained by averaging the frequency of strategy A in each round of simulation in the last 10^6 time steps after the transient 10^6 steps. The simulation data are obtained by executing 2000 independent simulation runs. Further details are given in Algorithm 1.

In Fig. 2, we first show how the average frequency of strategy A, $\langle x_A \rangle$ changes in dependence on the game parameter a and exploration rate ε on four different networks including the complete, regular, BA scale-free, and WS small-world networks, respectively. We observe that for each given value of random exploration rate ε, the fraction of strategy A increases with increasing the payoff value a. In addition, when $a < 6$, indicating that $a + b < c + d$, the frequency of strategy A increases as the random exploration rate ε increases. Whereas when $a > 6$, implying that $a + b > c + d$, the fraction of strategy A in the population decreases with increasing exploration rate ε. We further see that these results presented here are consistent in all four representative networks (see Fig. 2 (a)-(d)). In particular, we observe that for given values of a and ε, the fraction of strategy A is very close in these four different networks. These results suggest that

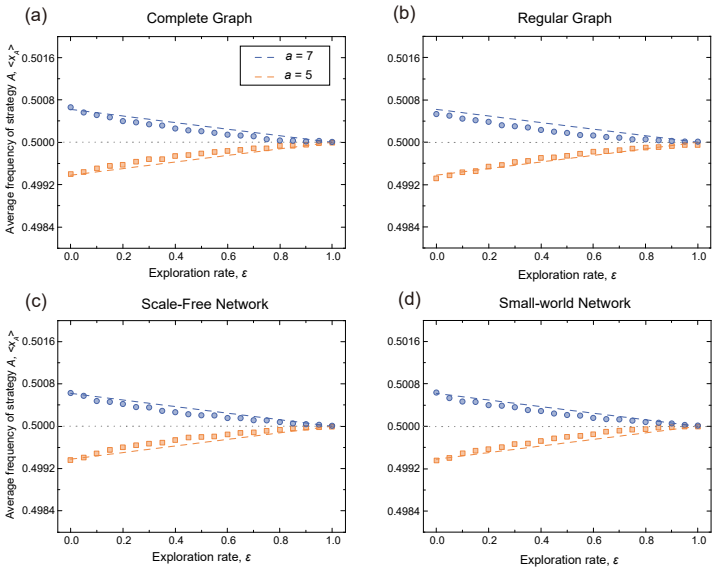

Fig. 3. The average frequency of strategy A in dependence on the random exploration rate ε on four different networks: complete graph (a), regular graph (b), scale-free network (c), and small-world network (d). In each panel, the blue and orange dashed lines respectively represent the numerical results obtained by Eq. (14) in two different game parameter scenarios: $a+b>c+d$ (here $a=7$) and $a+b<c+d$ (here $a=5$). The blue circles and orange squares respectively represent the simulation results for $a=7$ and $a=5$. The horizontal grey dotted line represents $\langle x_A\rangle=\frac{1}{2}$. Other parameter values are the same to those used in Fig. 2.

the average fraction of strategy A is independent of the network structure. Hence, we can conclude that all these simulation results confirm our theoretical predictions mentioned above.

In order to demonstrate how the random exploration rate influences the fraction of strategy A, we present the average frequency of strategy A as a function of ε for two different values of the game parameter a in the four different networks, as shown in Fig. 3. For the sake of comparison, we not only present the simulation results, but also show the theoretical predictions in these plots. We can observe that both the simulation and the theoretical results agree that the fraction of strategy A decreases with increasing ε for larger $a=7$, but is always larger than 0.5, no matter what the value of ε is. Whereas the fraction of strategy A increases as the ε value increases for smaller $a=5$, but it is always smaller than 0.5, independently of the value of ε. In addition, our simulation results agree well with the theoretical predictions in each type of network structure.

In order to provide a deeper understanding of these results, we show how the average regret values of strategy A and B are influenced by the game parameter a in four different networks including complete, regular, BA scale-free, and WS small-world graphs, as presented in Fig. 4. We find that the regret value of strategy A decreases as the game parameter a increases, while the regret value of strategy B remains unchanged with increasing a value. In addition, when the game parameter a is relatively small, the regret value r_A of strategy A is larger than the regret value r_B of strategy B. In this case, individuals prefer to choose strategy B rather than strategy A, hence strategy B is more favored over strategy A. However, when the game parameter a is continuously increased, the value of r_A becomes less than the value of r_B. Therefore, agents prefer to adopt strategy A rather than strategy B, which explains why the average fraction of strategy A increases as the game parameter a increases. On the other hand, once $a+b>c+d$, strategy A is more favored. However, when random exploration is introduced, the evolutionary advantage of strategy A can be weakened. This is because individuals have the opportunity to randomly choose the strategy rather than to choose the strategy according to the regret values of choosing strategies. In this case, the average

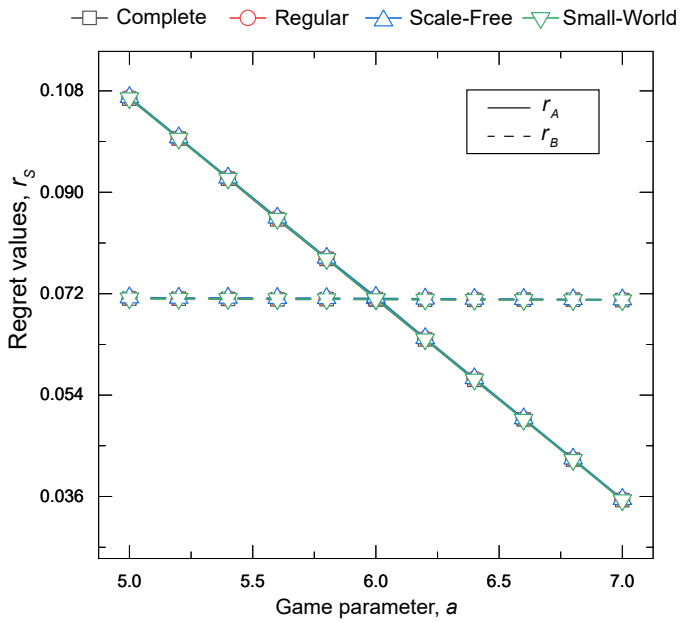

Fig. 4. The average regret value r_s as a function of the game parameter a on four different networks. Here r_s represents the average regret value of an individual choosing strategy s ($s = A, B$), and the solid (dashed) lines represent the average regret value r_A (r_B) of strategy A (B).

fraction of strategy A will decrease as the random exploration rate increases. But it is always larger than 0.5 even if the ε value is close to one. While $a + b < c + d$, strategy B prevails over strategy A. However, once random exploration is introduced, it makes individuals choose strategy B with a higher probability during the strategy update process. Hence the evolutionary advantage of strategy A is relatively decreased. In this case, the average fraction of strategy A will increase as the random exploration rate increases. But it is always less than 0.5 even if the ε value is close to one.

Our theoretical analysis shows that the condition for strategy A to prevail over strategy B is independent of the specific expressions of regret function. In order to verify this, we show how the average fraction of strategy A changes in dependence on the game parameter a for three different regret functions. The results are shown for four different network structures in Fig. 5. We observe that in each network structure we considered, the average fraction of strategy A always increases as the a value increases, irrespective of the regret functions. In particular, for $a < 6$, which makes $a + b < c + d$ satisfied, $\langle x_A \rangle$ is always less than 0.5. While for $a > 6$, which makes $a + b > c + d$ satisfied, $\langle x_A \rangle$ is always larger than 0.5. Thus, these simulation results verify our theoretical predictions well. In addition, we observe that the increase rate of the average fraction of strategy A with the parameter a is larger when the regret function is $g_{l,2}(r) = (1 + \mathrm{erf}(r))/2$. It means that the average fraction of strategy A under this function is higher than those under two other regret functions for $a > 6$. These results indicate that although the regret function does not influence the success condition of strategy A, it can affect the final fraction of strategy A in the population, which has been clearly shown in Eq. (14).

So far, we mainly present the evolutionary dynamics of network games with regret-based learning and random exploration in the case of weak regret strength via theoretical analysis and computer simulations. Indeed, the evolutionary dynamics could be significantly influenced by the regret selection intensity. Unfortunately, theoretical analysis becomes infeasible when the regret strength is not weak. To overcome this difficulty, we perform computer simulations for the case of non-weak regret strength. Our results are sum-

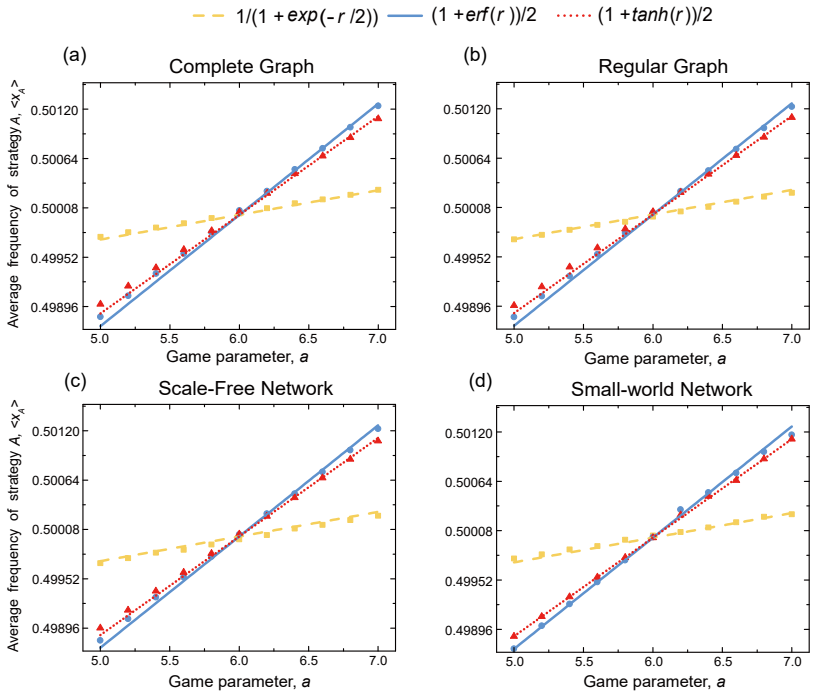

Fig. 5. The average frequency of strategy A as a function of the game parameter a for three different regret functions on the four network structures. Yellow dashed lines and squares correspond to the exponential function $g_l(r) = 1/(1+\exp(-r/2))$, blue solid lines and circles correspond to the error function $g_l(r) = (1+\mathrm{erf}(r))/2$, and red dotted lines and triangles correspond to the tangent function $g_l(r) = (1+\tanh(r))/2$. Lines represent the numerical results obtained by Eq. (14) and symbols depict the results by computer simulations. Other parameter values are the same to those in Fig. 2.

marized in Fig. 6, where we show the average fraction of strategy A as a function of the regret strength for two different values of the game parameter a for different networks. We find that the fraction of strategy A increases as the regret strength increases for larger a, which makes $a+b > c+d$ satisfied. Besides, the fraction of strategy A is always larger than 0.5 for all these values of β. In contrast, the fraction of strategy A decreases as the regret strength increases for smaller a, which makes $a+b < c+d$ satisfied. Furthermore, the fraction of strategy A is always less than 0.5 for all these values of β. These observations indicate that our main results obtained in the case of weak regret strength are still valid for non-weak regret strength. In addition, the fraction of strategy A is significantly increased under stronger regret strength when $a+b > c+d$, which implies that stronger regret strength can further enhance the evolutionary advantage of strategy A. In contrast, when $a+b < c+d$, the fraction of strategy A decreases dramatically under stronger regret strength. It shows that a stronger regret strength can further weaken the evolutionary advantage of strategy A.

5. Conclusion and Discussion

In this work, we have proposed a microscopic strategy updating protocol which combines the regret factor and random exploration in decision making, and studied how such a protocol influences the evolutionary dynamics of strategies in a structured population of individuals playing a general two-player, two-strategy game with neighbors. We have derived the mathematical condition under which strategy A can be more

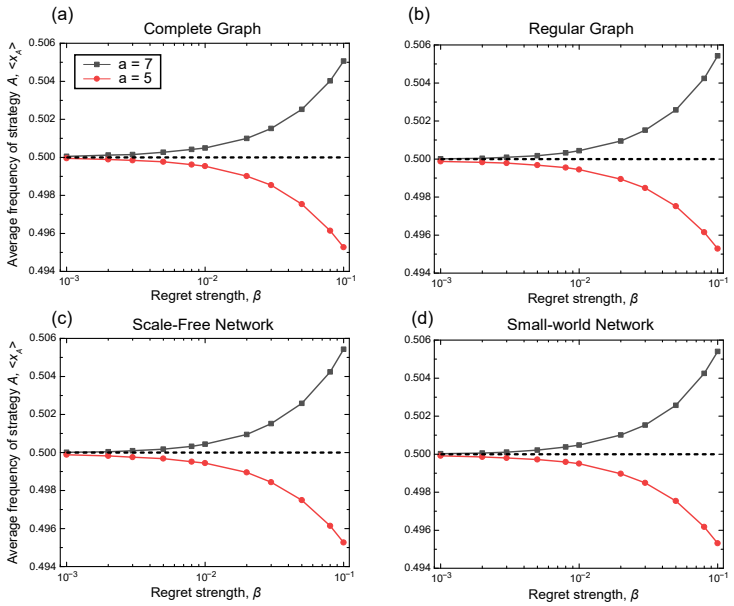

Fig. 6. The average frequency of strategy A as a function of the regret strength on the four network structures. The black squares and red circles respectively represent the simulation results for $a = 7$ and $a = 5$. Other parameter values are the same to those used in Fig. 2.

favored against strategy B. Our findings reveal that such strategy success condition shows some universality and depends only on game parameters and is independent of regret function, random exploration rate, and network structures under weak regret strength. However, the average frequency of strategy A does depend on the random exploration rate and the regret function. We find that the introduction of random exploration weakens the evolutionary advantage of strategy A when it prevails, but it enhances the evolutionary advantage of strategy A when it is not favored. In addition, we perform computer simulations on different representative networks to validate our theoretical predictions. To complete our study with further computer simulations, we have shown that increasing the regret strength further enhances the dominance of strategy A when it is more favored, but further suppresses the evolution of strategy A when it is not favored. This result indicates the dynamic of "the strong get stronger, the weak get weaker" when the regret strength is not weak.

We also demonstrate that the condition for strategy A to prevail over strategy B under our proposed learning rule is the same to that under the aspiration-based update rule in the general two-person, two-strategy game [37]. Indeed, the derived condition for strategy success is consistent with the traditional risk-dominance condition for the game [38]. However, we note that the regret value is determined by the difference between the maximum attainable payoff and the actual payoff. Indeed, the maximum possible payoff of one agent just depends on the strategy pattern and the weighted connections in the neighborhood and is time-varying during the evolutionary process. So, it is a quantity of objective existence. It is different from individual aspiration [37, 39], since individual aspiration level in a gaming environment could be subjectively conceived. In addition, in our work we consider a random exploration of strategy revisions during the evolutionary process, which is not considered previously in Ref. [37]. Although we find that the introduction of random exploration does not influence the condition of strategy success, it can alter the evolutionary outcomes of strategies, which are not reported earlier. Hence, our results shed light on how regret-based learning and random exploration shape strategy evolution in a general two-player, two-strategy game. Importantly, these findings can be conductive to the design of learning protocols involving

the regret-based learning and random exploration in gaming interaction scenarios [1, 2].

Our work also stimulates several new directions for future research. First, we find that the success condition of strategy A is independent of the regret function, but the strategy's frequency in the population depends on the regret function, as shown in Fig. 5. Hence, it would be meaningful to explore whether there exists an optimal regret function which can induce the highest frequency of strategy A in the game competition. On the other hand, we can also design the appropriate regret function for the desired outcome of strategy evolution. Besides, we emphasize that the expression of individual regret value applied in our work is relatively simple, but traditionally used. We employ the Boltzmann exploration mechanism based on the regret function to depict the regret-based learning. Indeed, there exist alternative ways to calculate individual regret value and some regret-based algorithms have already been proposed, such as the regret minimization and the $\phi - regret$ algorithms [40, 41]. To our best knowledge, these regret-based algorithms can produce different dynamics of strategy evolution via a large amount of computation. Thus, it would be interesting to consider these algorithms into our theoretical framework and explore how the key parameters involving the regret function influence the evolutionary outcomes quantitatively. Finally, we note that our focus is on the scenario of two-player, two-strategy game, although this game type is a representative class of games. Indeed, there exist other common game scenarios, such as multi-player and multi-strategy [42, 43, 44]. The game interaction mode for multi-player scenario is different from the pairwise interaction for two-player case [45, 46, 47] and the computation complexity significantly increases from two-strategy to multi-strategy situation. Hence, it will be worth exploring how regret-based learning and random exploration influence the dynamics of strategy evolution in these game scenarios for future study.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant No. 62473081) and the National Research, Development and Innovation Office (NKFIH) under Grant No. K142948.

Appendix A. Calculation details for obtaining Eq. (8)

Based on the normalized eigenvector associated with the eigenvalue 1 of the transition matrix $\mathbf{P}$, the Markov matrix has a stationary distribution $\mathbf{u} = (u_1, u_2, ..., u_i, ..., u_Z)$, where u_i represents the average fraction of time that the population spends in state i. So we calculate the average frequency of strategy A, which is described by $\langle x_A \rangle = \mathbf{u} \cdot \mathbf{x}$. Accordingly, we have

$$\mathbf{u} = \mathbf{u}\mathbf{P}, \tag{A.1}$$

$$\mathbf{u} \cdot \mathbf{e} = \mathbf{1}. \tag{A.2}$$

Here $\mathbf{e}$ is a $Z \times 1$ matrix consisting entirely of ones. Taking a Taylor expansion of $\mathbf{u}$, $\mathbf{P}$, and $\langle x_A \rangle$ with respect to β at $\beta = 0$, we have

$$\mathbf{u} = \mathbf{u_0} + \mathbf{u_1}\beta + o(\beta^2), \tag{A.3}$$

$$\mathbf{P} = \mathbf{P_0} + \mathbf{P_1}\beta + o(\beta^2), \tag{A.4}$$

$$\langle x_A \rangle = \langle x_A \rangle_0 + \langle x_A \rangle_1 \beta + o(\beta^2). \tag{A.5}$$

Here $\mathbf{u_0} = \mathbf{u}|_{\beta=0}$, $\mathbf{u_1} = \frac{d}{d\beta}\mathbf{u}|_{\beta=0}$, $\mathbf{P_0} = \mathbf{P}|_{\beta=\mathbf{0}}$, and $\mathbf{P_1} = \frac{d}{d\beta}\mathbf{P}|_{\beta=0}$. Then we gain two following formulas for $\beta = 0$ [33]

$$\mathbf{u_1}(\mathbf{I} - \mathbf{P_0}) = \mathbf{u_0}\mathbf{P_1}, \tag{A.6}$$

$$\mathbf{u_1}\mathbf{e} = \mathbf{0}. \tag{A.7}$$

Based on the above equations, we can obtain the specific form of $\mathbf{u_1}$. However, we note that $(\mathbf{I} - \mathbf{P_0})$ is not invertible since it is not full rank. Accordingly, we add a matrix $\mathbf{F}$ into the former matrix $(\mathbf{I} - \mathbf{P_0})$.

Here, $\mathbf{F} = \mathbf{e} \cdot \mathbf{e^T}$ and $\mathbf{u_1 F} = 0$, so we can get

$$\mathbf{u_1}(\mathbf{I} - \mathbf{P_0} + \mathbf{F}) = \mathbf{u_0 P_1}. \tag{A.8}$$

We can then prove that $(\mathbf{I} - \mathbf{P_0} + \mathbf{F})$ is invertible. Specifically, we set $\lambda_1, ..., \lambda_Z$ as the eigenvalues of $\mathbf{P_0}$ and $\mathbf{d}_1, ..., \mathbf{d}_Z$ as the corresponding right eigenvectors. Obviously, the matrix $\mathbf{P_0}$ is primitive (there exists a constant $\mathbf{k} > 0$, s.t. $(\mathbf{P_0})^{\mathbf{k}} > 0$) and row-stochastic, that is, a non-negative matrix where the sum of each row is 1. Thus, it has an eigenvalue $\lambda_1 = 1$ and the corresponding right eigenvector $\mathbf{d}_1 = (1, 1, ..., 1)^T$. We know that $(\mathbf{I} - \mathbf{P_0})$ has a sole eigenvalue 0 with the corresponding right eigenvector $\mathbf{d}_1$ from the properties of the primitive matrix and $\lambda_i \neq 1$ $(i = 2, ..., Z)$ [34]. Accordingly, the eigenvalues of $(\mathbf{I} - \mathbf{P_0} + \mathbf{F})$ are $\mathbf{d}_1^T \mathbf{d}_1, 1 - \lambda_2, ..., 1 - \lambda_Z$, because $\mathbf{d}_1^T \mathbf{d}_1 = Z$ and $1 - \lambda_i \neq 0$ $(i = 2, ..., Z)$. Hence, the eigenvalues of $(\mathbf{I} - \mathbf{P_0} + \mathbf{F})$ cannot be zero and are invertible.

Meanwhile, we note that $(\mathbf{I} - \mathbf{P_0} + \mathbf{F})^{-\mathbf{1}}$ is a fundamental matrix according to Ref. [48], and then for $n \to \infty$ we have

$$(\mathbf{I} - \mathbf{P_0} + \mathbf{F})^n = I = (\mathbf{I} - \mathbf{P_0} + \mathbf{F})M, \tag{A.9}$$

where $M = I + (\mathbf{P_0} - \mathbf{F}) + ... + (\mathbf{P_0} - \mathbf{F})^{n-1}$. Furthermore, following previous study [33] we take the first-order approximation for M and obtain that $M = (\mathbf{I} + \mathbf{P_0} - \mathbf{F})$. According to Eq. (A.8), we can calculate $\mathbf{u_1}$ as

$$\mathbf{u_1} = \mathbf{u_0 P_1}(\mathbf{I} + \mathbf{P_0} - \mathbf{F}). \tag{A.10}$$

We define $\mathbf{Q} = (\mathbf{I} + \mathbf{P_0} - \mathbf{F})\mathbf{x}$ and accordingly have $\langle x_A \rangle_1 = \mathbf{u_1 x} = \mathbf{u_0 P_1 Q}$.

Appendix B. Calculation details for obtaining Eq. (13)

In what follows, we provide calculation details for Eq. (13) based on Eq. (A.10). Since $\mathbf{P_0}$ is a $2^N \times 2^N$ matrix and symmetric, each element in the stationary distribution of $\mathbf{P_0}$ is identical. So we have

$$\mathbf{u_0} = \left[\frac{1}{2^N}, ..., \frac{1}{2^N} \right]. \tag{B.1}$$

Meanwhile, according to the above description, we know that $\mathbf{P_1} = \frac{d}{d\beta}\mathbf{P}|_{\beta=0}$ and $\mathbf{Q} = (\mathbf{I} + \mathbf{P_0} - \mathbf{F})\mathbf{x}$. We then calculate $\mathbf{Q} = [q_j]_{Z \times 1}$. We know that $\mathbf{S}'_k$ $(k = 0, ..., N)$ comprises of a total value of C_N^k q_j and we obtain

$$q_j = \frac{k\alpha}{N} + \frac{(k+1)N - 2k}{N^2}\gamma + \frac{k}{N} - 2^{N-1}. \tag{B.2}$$

Consequently, we calculate each element of $\mathbf{u_0 P_1} = [v_i]_{1 \times Z}$, which is similar to the calculations of each element of $\mathbf{Q} = [q_j]_{Z \times 1}$. Items with the same elements are put together and redefined as $\mathbf{S}''_k$ $(k = 0, ..., N)$, hence $\mathbf{S}''_k$ comprises a total value of C_N^k v_i. Then we have

$$v_i = \frac{\delta}{2^N \cdot N}[kf(\mathbf{s}_i{}^{k-1}) + (2k - N)f(\mathbf{s}_i{}^k) - (N - k)f(\mathbf{s}_i{}^{k+1})], \tag{B.3}$$

where $f(\mathbf{s}_i{}^k)$ represents the value of $f(\mathbf{s}_i)$ when state $\mathbf{s}_i \in \mathbf{S}_k$. Multiplying Eq. (B.3) by Eq. (B.2), we can learn that there are $N + 1$ terms in $\langle x_A \rangle_1$ and we can denote the kth item of $\langle x_A \rangle_1$ as $h(\mathbf{s}_i{}^k)$. Accordingly, we can obtain

$$\langle x_A \rangle_1 = \sum_{k=0}^{N} h(\mathbf{s}_i{}^k), \tag{B.4}$$

where

$$h(\mathbf{s}_i{}^k) = \frac{\sum_{l=1}^{N} \delta \cdot f(\mathbf{s}_i{}^k)}{2^N \cdot N} \cdot [\boldsymbol{\xi} \cdot \alpha + \boldsymbol{\zeta} \cdot \gamma + \boldsymbol{\phi} \cdot 2^{N-1} + \boldsymbol{\psi}]. \tag{B.5}$$

Here, $\sum_{l=1}^{N} \delta f(\mathbf{s}_i{}^k)$ represents the accumulated payoff difference when individual l respectively adopts strategy A and strategy B in state $\mathbf{s}_i$, and based on the property of combinatorial number the coefficients $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ in the above equation are respectively calculated as

$$\begin{aligned}
\boldsymbol{\xi} =& \frac{(k+1)^2}{N} \cdot C_N^{k+1} + \frac{2k^2 - Nk}{N} \cdot C_N^k \\
& - \frac{(N-k+1)(k+1)}{N} \cdot C_N^{k+1} \\
=& k[C_{N-1}^k + 2C_{N-1}^{k-1} - C_N^k - C_N^{k+1} + C_{N-1}^{k-2}] \\
& + C_{N-1}^k + C_N^{k-1} - C_{N-1}^{k-2} \\
=& k[C_N^k + C_N^{k-1} - C_N^k - C_N^{k-1}] \\
& + (C_{N-1}^k + C_N^{k-1} - C_{N-1}^{k-2}) \\
=& C_{N-1}^k + C_{N-1}^{k-1} \\
=& C_N^k,
\end{aligned} \tag{B.6}$$

$$\begin{aligned}
\boldsymbol{\zeta} =& \frac{(k+1)}{N}(2k-N) \cdot C_N^{k+1} - (N-k+1)\frac{k}{N} \cdot C_N^{i-1} \\
& + \frac{(i+1)(i+2)}{N} \cdot C_N^{i+1} - \frac{2}{N}[\frac{(k+1)^2}{N} \cdot C_N^{k+1} \\
& + \frac{2k^2 - Nk}{N} \cdot C_N^k - \frac{(N-i+1)(k+1)}{N} \cdot C_N^{k+1}] \\
=& 2(k+1)C_{N-1}^{k-1} - (k+1)C_N^k - kC_N^{k-1} \\
& + kC_{N-1}^{k-2} + (i+2)C_{N-1}^k - \frac{2}{N}C_N^k \\
=& k[C_N^i + C_N^{k-1} - C_N^k - C_N^{k-1}] \\
& + 2C_N^k - C_N^k - \frac{2}{N}C_N^k \\
=& (1 - \frac{2}{N})C_N^k,
\end{aligned} \tag{B.7}$$

$$\begin{aligned}
\boldsymbol{\phi} =& (N-k+1)C_N^{k-1} + (N-2k)C_N^k - (k+1)C_N^{k+1} \\
=& N(C_N^{k-1} + C_N^k) - (k-1)C_N^{k-1} \\
& - 2kC_N^k - (k+1)C_N^{k+1} \\
=& NC_{N+1}^k - N(C_{N-1}^{k-2} + 2C_{N-1}^{k-1} + C_{N-1}^k) \\
=& NC_{N+1}^k - NC_{N+1}^k \\
=& 0,
\end{aligned} \tag{B.8}$$

and

$$\begin{aligned}\boldsymbol{\psi} =&\frac{(i+1)^2}{N}\cdot C_N^{k+1}+\frac{2k^2-Nk}{N}\cdot C_N^k\\ &-\frac{(N-k+1)(k+1)}{N}\cdot C_N^{k+1}\\ =&k[C_{N-1}^k+2C_{N-1}^{k-1}-C_N^k-C_N^{k+1}+C_{N-1}^{k-2}]\\ &+C_{N-1}^k+C_N^{k-1}-C_{N-1}^{k-2}\\ =&k[C_N^k+C_N^{k-1}-C_N^k-C_N^{k-1}]\\ &+(C_{N-1}^k+C_N^{k-1}-C_{N-1}^{k-2})\\ =&C_{N-1}^k+C_{N-1}^{k-1}\\ =&C_N^k.\end{aligned} \tag{B.9}$$

Hence we further obtain the kth item of $\langle x_A\rangle_1$ as

$$\begin{aligned}h({\mathbf{s}_i}^k) &= \frac{\sum_{l=1}^N \delta\cdot f({\mathbf{s}_i}^k)}{2^N\cdot N}\cdot C_N^k\cdot[\alpha+(1-\frac{2}{N})\gamma+1)]\\ &= \frac{\sum_{l=1}^N \delta}{2^N\cdot N}\left(\frac{2N-1}{N}-\frac{1}{N}\varepsilon\right)C_N^k\cdot f({\mathbf{s}_i}^k)\\ &= \frac{2N-1-\varepsilon}{2^N\cdot N^2}\cdot\sum_{l=1}^N\delta\cdot C_N^k\cdot f({\mathbf{s}_i}^k).\end{aligned} \tag{B.10}$$

Since $\mathbf{P_0}$ is obtained when $\mathbf{P}$ is derived at $\beta=0$, each individual has a $1/2$ probability of using strategy A and individual strategy update is independent of each other. Individual strategy update does not depend on the regret value, so we quantize the description into formulas and have $\langle s_l s_j\rangle_0 = \langle s_l\rangle_0\langle s_j\rangle_0$ and $\langle s_l s_l\rangle_0 = \langle s_l\rangle_0$ for $\beta=0$. Based on these equations, we can get the correlation formula between strategies for $\beta=0$, given as

$$\langle s_l s_j\rangle_0 = \begin{cases}\frac{1}{4} & \text{if } l\neq j,\\ \frac{1}{2} & \text{if } l=j.\end{cases} \tag{B.11}$$

It means that $f(\mathbf{s}_i)$ is the same for any i. So we obtain $f({\mathbf{s}_i}^k)=f(\mathbf{s}_i)=\frac{a+b-c-d}{4}$.

Summing up all terms of $\langle x_A\rangle_1$, we can get

$$\begin{aligned}\langle x_A\rangle_1 &= \sum_{k=0}^N h({\mathbf{s}_i}^k)\\ &= \sum_{k=0}^N\frac{2N-1-\varepsilon}{2^N\cdot N^2}\cdot\sum_{l=1}^N\delta\cdot C_N^k\cdot f(\mathbf{s}_i)\\ &= \frac{2N-1-\varepsilon}{16N^2}\cdot\sum_{l=1}^N\frac{g_l'(0)}{g_l(0)}(1-\varepsilon)(a+b-c-d)).\end{aligned} \tag{B.12}$$

Appendix C. Average frequency of strategy A under accumulated payoff

In the part, we provide the calculation details for the average frequency of strategy A under accumulated payoffs. To be specific, the accumulated payoff of individual l in state $\mathbf{s}_i$ is described as

$$p_A(\mathbf{s}_i) = \sum_{j=1}^{N} w_{lj} \cdot [a s_l s_j + b s_l (1 - s_j)], \tag{C.1}$$

$$p_B(\mathbf{s}_i) = \sum_{j=1}^{N} w_{lj} \cdot [c(1 - s_l) s_j + d(1 - s_l)(1 - s_j)]. \tag{C.2}$$

Based on Eq. (B.11), we obtain

$$f(\mathbf{s}_i^k) = f(\mathbf{s}_i) = p_A(\mathbf{s}_i) - p_B(\mathbf{s}_i) = \sum_{j=1}^{N} w_{lj} \cdot f(\mathbf{s}_i) \tag{C.3}$$

$$= \sum_{j=1}^{N} w_{lj} \cdot \frac{a + b - c - d}{4} = d_l \cdot \frac{a + b - c - d}{4}. \tag{C.4}$$

Furthermore, we get $h({\mathbf{s}_i}^k)$ as

$$\begin{aligned} h({\mathbf{s}_i}^k) &= \frac{\sum_{l=1}^{N} \delta \cdot f({\mathbf{s}_i}^k)}{2^N \cdot N} \cdot [\boldsymbol{\xi} \cdot \alpha + \boldsymbol{\zeta} \cdot \gamma + \boldsymbol{\phi} \cdot 2^{N-1} + \boldsymbol{\psi}] \\ &= \frac{\sum_{l=1}^{N} d_l \cdot \delta \cdot f({\mathbf{s}_i}^k)}{2^N \cdot N} \cdot C_N^k \cdot [\alpha + (1 - \frac{2}{N})\gamma + 1)] \\ &= \frac{\sum_{l=1}^{N} d_l \cdot \delta}{2^N \cdot N} \left(\frac{2N - 1}{N} - \frac{1}{N}\varepsilon \right) C_N^k \cdot f({\mathbf{s}_i}^k) \\ &= \frac{2N - 1 - \varepsilon}{2^N \cdot N^2} \cdot \sum_{l=1}^{N} d_l \cdot \delta \cdot C_N^k \cdot f({\mathbf{s}_i}^k). \end{aligned} \tag{C.5}$$

We can then obtain $\langle x_A \rangle_1$ under accumulated payoffs as

$$\begin{aligned} \langle x_A \rangle_1 &= \sum_{k=0}^{N} h({\mathbf{s}_i}^k) \\ &= \sum_{k=0}^{N} \frac{2N - 1 - \varepsilon}{2^N \cdot N^2} \cdot \sum_{l=1}^{N} d_l \cdot \delta \cdot C_N^k \cdot f(\mathbf{s}_i) \\ &= \frac{2N - 1 - \varepsilon}{N^2} \cdot \sum_{l=1}^{N} d_l \cdot \delta \cdot f(\mathbf{s}_i) \\ &= \frac{2N - 1 - \varepsilon}{N^2} \cdot \sum_{l=1}^{N} d_l \cdot \frac{g_l'(0)}{4 g_l(0)} (1 - \varepsilon) \cdot \frac{a + b - c - d}{4} \\ &= \frac{2N - 1 - \varepsilon}{16 N^2} \cdot \sum_{l=1}^{N} d_l \cdot \frac{g_l'(0)}{g_l(0)} (1 - \varepsilon)(a + b - c - d). \end{aligned} \tag{C.6}$$

Finally, according to Eq. (5), we obtain that the average frequency of strategy A under weak regret

strength is given as

$$\langle x_A \rangle = \frac{1}{2} + \frac{2N-1-\varepsilon}{16N^2} \cdot \sum_{l=1}^{N} d_l \frac{g_l'(0)}{g_l(0)} (1-\varepsilon)\beta(a+b-c-d) + o(\beta^2).$$

Thus if the accumulated payoff is considered for individuals, the success condition for strategy A is still $a+b > c+d$.

References

[1] S. Fatima, N.R. Jennings, M. Wooldridge, Learning to resolve social dilemmas: A survey, J. Artif. Intell. Res. 79 (2024) 895-969.

[2] D. Bloembergen, K. Tuyls, D. Hennes, M. Kaisers, Evolutionary dynamics of multi-agent learning: A survey, J. Artif. Intell. Res. 53 (2015) 659-697.

[3] S. Tan, Y. Wang, Y. Chen, Z. Wang, Evolutionary dynamics of collective behavior selection and drift: Flocking, collapse, and oscillation, IEEE Trans. Cybern. 47 (2017) 1694-1705.

[4] D. Madeo, C. Mocenni, Game interactions and dynamics on networked populations, IEEE Trans. Autom. Control 60 (2015) 1801-1810.

[5] T.A. Han, Emergent behaviours in multi-agent systems with evolutionary game theory, AI Commun. 35 (2022) 327-337.

[6] S. Tan, Y. Wang, J. L, Analysis and control of networked game dynamics via a microscopic deterministic approach, IEEE Trans. Autom. Control 61 (2016) 4118-4124.

[7] J. Zhang, Y. Zhu, Z. Chen, Evolutionary game dynamics of multiagent systems on multiple community networks, IEEE Trans. Syst. Man Cybern. Syst. 50 (2020) 4513-4529.

[8] G. Como, F. Fagnani, L. Zino, Imitation dynamics in population games on community networks, IEEE Trans. Control Netw. Syst. 8 (2021) 65-76.

[9] P. Ramazi, M. Cao, Asynchronous decision-making dynamics under best-response update rule in finite heterogeneous populations, IEEE Trans. Autom. Control 63 (2018) 742-751.

[10] T.A. Han, L.M. Pereira, F.C. Santos, T. Lenaerts, To regulate or not: A social dynamics analysis of an idealised AI race, J. Artif. Intell. Res. 69 (2020) 881-921.

[11] R. Merhej, F.P. Santos, F.S. Melo, F.C. Santos, Cooperation and learning dynamics under wealth inequality and diversity in individual risk, J. Artif. Intell. Res. 74 (2022) 733-764.

[12] X. Wang, L. Zhou, A. McAvoy, A. Li, Imitation dynamics on networks with incomplete information, Nat. Commun. 14 (2023) 7453.

[13] Y. Meng, S.P. Cornelius, Y.-Y. Liu, A. Li, Dynamics of collective cooperation under personalised strategy updates, Nat. Commun. 15 (2024) 3125.

[14] Q. Wang, X. Chen, N. He, A. Szolnoki, Evolutionary dynamics of population games with an aspiration-based learning rule, IEEE Trans. Neural Netw. Learn. Syst. 36 (2025) 8387-8400.

[15] Z. Zeng, M. Feng, A. Szolnoki, Evolutionary dynamics with self-interaction learning in networked systems, IEEE Trans. Netw. Sci. Eng. 13 (2026) 296-313.

[16] Y. Wu, J. Zhang, X. Li, Q-learning promotes the evolution of fairness and generosity in the ultimatum game, Chaos, Solitons & Fractals 200 (2025) 116984.

[17] J. Li, C. Wu, D. Han, On evolution of agent behavior under limited gaming time with reinforcement learning, Chaos, Solitons & Fractals 194 (2025) 116166.

[18] Z. An, H. Liu, Y. Wang, Cooperation dynamics driven by reinforcement learning with interactive diversity in structured populations, Chaos, Solitons & Fractals 201 (2025) 117308.

[19] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, MIT Press, Cambridge, 1998.

[20] S. Leonardos, G. Piliouras, Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory, Artif. Intell. 304 (2022) 103653.

[21] Y. Shi, Z. Rong, Analysis of Q-learning-like algorithms through evolutionary game dynamics, IEEE Trans. Circuits Syst. II Express Briefs 69 (2022) 2463-2467.

[22] T. Ren, X.J. Zeng, Reputation-based interaction promotes cooperation with reinforcement learning, IEEE Trans. Evol. Comput. 28 (2024) 1177-1188.

[23] J. Li, H. Zhang, S. Ke, J. Huang, N. Chen, X. Shen, Non-cooperative multi-agent reinforcement learning exploiting population dynamics, IEEE Trans. Netw. Sci. Eng. 13 (2026) 684-700.

[24] Y. Xu, D. Zhao, T. Perc Benko, C. Xia, M. Perc, Reinforcement learning can be a double-edged sword for cooperation on higher-order networks, IEEE Trans. Syst. Man Cybern. Syst. (2026). https://doi.org/10.1109/TSMC.2025.3624366

[25] B. Pi, L.-J. Deng, M. Feng, M. Perc, J. Kurths, Dynamic evolution of complex networks: A reinforcement learning approach applying evolutionary games to community structure, IEEE Trans. Pattern Anal. Mach. Intell. 47 (2025) 8563-8582.

[26] Z. Yuan, G. Jiang, S. Hu, M. Perc, C. Chu, J. Liu, Dynamics of Q-learning in networked stochastic games, IEEE Trans. Neural Netw. Learn. Syst. (2026). https://doi.org/10.1109/TNNLS.2025.3641365

[27] Q. Su, H. Wang, Y. Xia, L. Wang, A multi-agent reinforcement learning framework for exploring dominant strategies in iterated and evolutionary games, Nat. Commun. 17 (2026) 490.

[28] A. Blum, Y. Monsour, Learning, regret minimization and equilibria, Cambridge University Press, Cambridge, 2007.
[29] D. Hu, S. Hu, C. Mu, S. Fan, C. Chu, J. Liu, Z. Wang, Regret minimization in population network games: Vanishing heterogeneity and convergence to equilibria, IEEE Trans. Neural Netw. Learn. Syst. 36 (2025) 20146-20156.
[30] P. Ladosz, L. Weng, M. Kim, H. Oh, Exploration in deep reinforcement learning: A survey, Inf. Fusion 85 (2022) 1-22.
[31] K. Chakroun, D. Mathar, A. Wiehler, F. Ganzer, J. Peters, Dopaminergic modulation of the exploration-exploitation trade-off in human decision-making, eLife 9 (2020) e51260.
[32] B. Allen, G. Lippner, Y.-T. Che, B. Fotouhi, N. Momeni, S.-T. Yau, M.A. Nowak, Evolutionary dynamics on any population structure, Nature 544 (2017) 227-230.
[33] M.C. Couto, S. Giaimo, C. Hilbe, Introspection dynamics: A simple model of counterfactual learning in asymmetric games, New J. Phys. 24 (2022) 063010.
[34] R.A. Horn, C.R. Johnson, Matrix analysis, Cambridge University Press, Cambridge, 1985.
[35] A.-L. Barabsi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509-512.
[36] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440-442.
[37] L. Zhou, B. Wu, J. Du, L. Wang, Aspiration dynamics generate robust predictions in heterogeneous populations, Nat. Commun. 12 (2021) 3250.
[38] M.A. Nowak, Evolutionary dynamics, Harvard University Press, Cambridge, 2006.
[39] X. Chen, L. Wang, Promotion of cooperation induced by appropriate payoff aspirations in a small-world networked game, Phys. Rev. E 77 (2008) 017103.
[40] G. Piliouras, M. Rowland, S. Omidshafiei, R. Elie, D. Hennes, J. Connor, K. Tuyls, Evolutionary dynamics and ϕ-regret minimization in games, J. Artif. Intell. Res. 74 (2022) 1125-1158.
[41] Y. Bai, C. Jin, S. Mei, Z. Song, T. Yu, Efficient ϕ-regret minimization in extensive-form games via online mirror descent, Adv. Neural Inf. Process. Syst. 35 (2022) 22313-22325.
[42] L. Zhou, B. Wu, V.V. Vasconcelos, L. Wang, Simple property of heterogeneous aspiration dynamics: Beyond weak selection, Phys. Rev. E 98 (2018) 062124.
[43] C. Wang, M. Perc, A. Szolnoki, Evolutionary dynamics of any multiplayer game on regular graphs, Nat. Commun. 15 (2024) 5349.
[44] D. Wang, P. Yi, G. Yan, F. Fu, Evolutionary dynamics of pairwise and group cooperation in heterogeneous social networks, IEEE Trans. Netw. Sci. Eng. (2026). https://doi.org/10.1109/TNSE.2025.3647918
[45] L. Shi, Z. He, C. Shen, J. Tanimoto, Enhancing social cohesion with cooperative bots in societies of greedy mobile individuals, PNAS Nexus 3 (2024) 223.
[46] C. Shen, Z. He, L. Shi, Z. Wang, J. Tanimoto, Prosocial punishment bots breed social punishment in human players, J. R. Soc. Interface 21 (2024) 20240019.
[47] Z. Si, Z. He, C. Shen, J. Tanimoto, Cooperative bots exhibit nuanced effects on cooperation across strategic frameworks, J. R. Soc. Interface 22 (2025) 20240427.
[48] C.M. Grinstead, J.L. Snell, Introduction to probability, American Mathematical Society, Providence, 2012.