Modeling Replicator Dynamics in Stochastic Games Using Markov Chain Method

Chuang Deng
Department of Automation
Shanghai Jiao Tong University
cdengcnc@sjtu.edu.cn

Zhihai Rong
School of Computer Science and Engineering
University of Electronic Science and Technology of China
zhihai.rong@gmail.com

Lin Wang
Department of Automation
Shanghai Jiao Tong University
wanglin@sjtu.edu.cn

Xiaofan Wang
School of Mechatronic Engineering and Automation
Shanghai University
xfwang@sjtu.edu.cn

ABSTRACT

In stochastic games, individuals need to make decisions in multiple states and transitions between states influence the dynamics of strategies significantly. In this work, by describing the dynamic process in stochastic game as a Markov chain and utilizing the transition matrix, we introduce a new method, named state-transition replicator dynamics, to obtain the replicator dynamics of a stochastic game. Based on our proposed model, we can gain qualitative and detailed insights into the influence of transition probabilities on the dynamics of strategies. We illustrate that a set of unbalanced transition probabilities can help players to overcome the social dilemmas and lead to mutual cooperation in a cooperation back state, even if the stochastic game has the same social dilemmas in each state. Moreover, we also present that a set of specifically designed transition probabilities can fix the expected payoffs of one player and make him lose the motivation to update his strategies in the stochastic game.

KEYWORDS

Multi-agent Learning; Evolutionary Game Theory; Replicator Dynamics; Stochastic Games

ACM Reference Format:
Chuang Deng, Zhihai Rong, Lin Wang, and Xiaofan Wang. 2021. Modeling Replicator Dynamics in Stochastic Games Using Markov Chain Method. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

The tragedy of the commons leads to the questions how to drive and reinforce cooperative behaviors in social and economic systems [4, 17]. These questions have been extensively explored by analyzing stylized uncooperative game theory models with some feedback mechanisms using evolutionary game theory [14, 16, 18–20]. Besides the evolutionary game theory, another method used in studying the dynamics of strategies is multi-agent reinforcement learning [1, 3, 12]. These works mainly derive and design multi-agent reinforcement learning algorithms which can lead strategies of players to the Nash equilibrium existing in the game.

For gaining qualitative and detailed insights into the learning process in repeated games, many algorithms and models have been proposed to build the bridge between evolutionary game theory and multi-agent reinforcement learning algorithms, such as FAQ-learning [10], regret minimization [11], continuous strategy replicator dynamics [5], IGA [21], IGA-WOLF [2] and so on. Besides the assumption that individuals are in the same state, how to describe strategy dynamics with the reciprocity of multiple states and separate strategies also raises the attention of researchers. Vrancx *et al.* first investigate the problem and they combine replicator dynamics and piecewise dynamics to model the learning behavior of agents in stochastic games [25]. Hennes *et al.* propose a method called state-coupled replicator dynamics which couples the multiple states directly [8]. The derived algorithms from the combination of evolutionary game theory and multi-agent reinforcement learning algorithms have been applied in multiple research fields, including femtocell power allocation [27], interdomain routing price setting [24], design of social agents [6] and design of new multi-agent reinforcement learning algorithms [7].

In this paper, by describing the dynamic process in stochastic game as a Markov chain [9, 22], we propose a new method, named state-transition replicator dynamics, to derive a set of ordinary differential equations to model the learning behavior of players in stochastic games. Moreover, based on the transition matrix of Markov chain and the derived replicator dynamics system, we demonstrate that, though players face the same social dilemmas in all states, a set of unbalanced transition probabilities between states can lead to cooperation in the cooperation back state where players have a high probability to be in after mutual cooperation. Besides, we also point out that in stochastic games, there exist several sets of specific transition probabilities which can control the expected payoffs of players. For a stateless matrix game, Press *et al.* [15] prove that if players' decisions depend on the previous actions of players participating in games, one player can unilaterally sets the payoff of the other player. In stochastic games, transition probabilities can take this work. Utilizing the analyzing method in stateless matrix game [13, 15, 23] and state-transition replicator dynamics model, we present that a set of specifically derived transition probabilities can fix the expected payoffs of one player and this player loses the motivation to update his strategies in this stochastic game.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

2 BACKGROUND

2.1 Stochastic games

Stochastic games describe how much rewards players obtain based on their strategies in multiple states. The concept of state in stochastic games refers to the environment where players interact with each other. The current state players being in and the joint action players taking not only determine the immediate payoffs players can receive in this round, but also the state they will stay in the next round. To define a stochastic game, we need to specify the set of players, the set of states, the set of actions that each player can take in each state, a transition function that describes how states change over time and the payoff function which describes how players' actions in each state affect the players' payoffs. We follow the definition in [8] to give a definition of stochastic games.

The game $G = \langle n, \boldsymbol{S}, \boldsymbol{\Omega}, \boldsymbol{z}, \tau, \pi_1, \cdots, \pi_n \rangle$ is a stochastic game with n players and k states. In each state $s \in \boldsymbol{S} = (s^1, \cdots, s^k)$, every player i has an action set $\boldsymbol{\Omega}_i(s)$ and strategy $\pi_i(s)$. In each round, every player i stays in one sate s and chooses an action a_i from action set $\boldsymbol{\Omega}_i(s)$ according to strategy $\pi_i(s)$. The payoff function $\tau(s, \boldsymbol{a}) : \prod_{i=1}^{n} \boldsymbol{\Omega}_i(s) \mapsto R^n$ maps the joint action $\boldsymbol{a} = (a_1, \cdots, a_n)$ to an immediate payoff value for each player. The transition function $z(s, \boldsymbol{a}) : \prod_{i=1}^{n} \boldsymbol{\Omega}_i(s) \mapsto \Delta^{k-1}$ determines the probabilistic state transition, where Δ^{k-1} is the $(k-1)$-simplex and $z_{s'}(s, \boldsymbol{a})$ is the transition probability from state s to s' under joint action $\boldsymbol{a}$.

In this work, we also follow the restriction proposed in [8] that all states $s \in \boldsymbol{S}$ are in an ergodic set. This restriction ensures that the game has no absorbing states.

2.2 Two-state two-player two-action stochastic games

Here, we introduce the simplest version of stochastic games: two-state two-player two-action stochastic game.

In games containing two players, player P (Q) has a strategy $\boldsymbol{p}$ ($\boldsymbol{q}$). Each player chooses between only two actions, cooperation (c) and defection (d). Thus, we can use a probability to fully define one player's strategy. For player P, the parameter p means the probability of cooperation and $\boldsymbol{p} = (p, 1-p)^{\mathrm{T}}$. Similarly, for player Q, $\boldsymbol{q} = (q, 1-q)^{\mathrm{T}}$. The payoff function can be represented by a bi-matrix (A, B), that gives the payoffs for the row player (P) in A, and the column player (Q) in B, which is shown as follow,

$$(A, B) = \begin{pmatrix} a_{cc}, b_{cc} & a_{cd}, b_{cd} \\ a_{dc}, b_{dc} & a_{dd}, b_{dd} \end{pmatrix}. \tag{1}$$

The outcome of players' joint action determines the payoffs to both players. It follows the order of actions of player P and Q, $\boldsymbol{a} = (a_P, a_Q)$, where a_P (a_Q) is the action taken by player P (Q).

When a stochastic game has two states, s^1 and s^2, the payoff matrices are given by

$$\begin{aligned} (A^1, B^1) &= \begin{pmatrix} a^1_{cc}, b^1_{cc} & a^1_{cd}, b^1_{cd} \\ a^1_{dc}, b^1_{dc} & a^1_{dd}, b^1_{dd} r \end{pmatrix}, \\ (A^2, B^2) &= \begin{pmatrix} a^2_{cc}, b^2_{cc} & a^2_{cd}, b^2_{cd} \\ a^2_{dc}, b^2_{dc} & a^2_{dd}, b^2_{dd} \end{pmatrix}. \end{aligned} \tag{2}$$

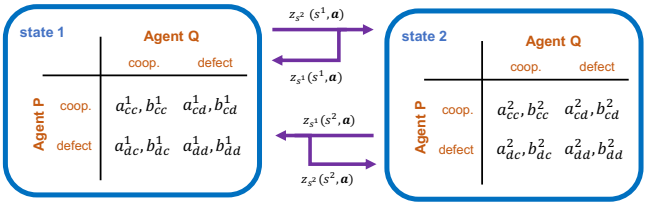

Figure 1: The framework of a two-state two-player two-action stochastic game. There are two players, P and Q. In every round, they can be in one state, s^1 or s^2. In each state, each player has a corresponding strategy and payoff matrix. The state in the next round depends on the current state, on the players' joint action and on chance, $z_{s'}(s, \boldsymbol{a})$.

A^s (B^s) determines the payoffs for player P (Q) in state s. Player P receives $\tau_P(s, \boldsymbol{a}) = a^s_{a_P, a_Q}$ while player Q gets $\tau_Q(s, \boldsymbol{a}) = b^s_{a_P, a_Q}$ for a given joint action $\boldsymbol{a} = (a_P, a_Q)$. Player P has strategies $\boldsymbol{p}^1 = (p^1, 1-p^1)^{\mathrm{T}}$ and $\boldsymbol{p}^2 = (p^2, 1-p^2)^{\mathrm{T}}$ in state s^1 and s^2, respectively. Player Q also has two strategies, $\boldsymbol{q}^1$ and $\boldsymbol{q}^2$. Since there are only two states, the transition probabilities have the relationship, $z_{s^1}(s^1, \boldsymbol{a}) = 1 - z_{s^2}(s^1, \boldsymbol{a})$ and $z_{s^2}(s^2, \boldsymbol{a}) = 1 - z_{s^1}(s^2, \boldsymbol{a})$. The framework of a two-state two-player two-action stochastic game is shown in Figure 1.

2.3 Learning automata

In this section, we give a description on how player can update his strategy with multi-agent reinforcement learning method [26]. A player can be seen as a learning automata, who updates his strategy by monitoring the reinforcement signal resulted from the action he has chosen in this round. An automata wants to learn the optimal strategy and maximizes the expected payoffs.

In this paper, we focus on finite action-set learning automata (FALA) [8]. At the beginning of a round, every player draws a random action $a(t)$ from his action set according to his strategy $\boldsymbol{\pi}(t)$. Based on the action $a(t)$, the environment responds with a reward $\tau(t)$. The automata uses this reward to update his strategy to $\boldsymbol{\pi}(t+1)$. The update rule for FALA using the linear reward-inaction (L_{R-I}) scheme is given below,

$$\pi_{a^*}(t+1) = \pi_{a^*}(t) + \begin{cases} \alpha\tau(t)\left(1 - \pi_{a^*}(t)\right) & \text{if } a(t) = a^* \\ -\alpha\tau(t)\pi_{a^*}(t) & \text{otherwise}, \end{cases} \tag{3}$$

where $\tau \in [0, 1]$. π_{a^*} means the probability of taking action a^*. The parameter $\alpha \in [0, 1]$ determines the learning rate.

In multi-state games, a player associates a dedicated learning automata to each state of the game. One LA tries to optimize the strategy in one state using the standard update rule given in Equation (3). Only a single LA is active and selects an action in each round of the game. However, the immediate reward from the environment is not directly fed back to this LA. Instead, when the LA becomes active again, i.e., next time the same state is played, it is informed about the cumulative reward gathered since the last activation and the time that has passed by [8].

The reward $\tau_i(t)$ for agent i's automaton $\mathrm{LA}_i(s)$ associated with state s is defined as

$$\tau_i(t) = \frac{\Delta r_i}{\Delta t} = \frac{\sum_{l=t_0(s)}^{t-1} r_i(l)}{t - t_0(s)}, \tag{4}$$

where $r_i(l)$ is the immediate reward for agent i in round l and $t_0(s)$ is the last occurrence function and determines when state s was visited last. The reward feedback in round t equals the cumulative reward Δr_i divided by time-frame Δt. The cumulative reward Δr_i is the sum over all immediate rewards gathered in all states beginning with round $t_0(s)$ and including the last round $t-1$. The time-frame Δt measures the number of rounds that have passed since automaton $\mathrm{LA}_i(s)$ has been active last. This means the state strategy is updated using the average reward over the interim immediate rewards.

2.4 Replicator dynamics

From the perspective of evolutionary game theory, the dynamics of strategies can be formulated as a system of differential equations. Each replicator represents one strategy. Strategies that gain above-average payoff become more likely to be played [25]. The dynamics of strategy that player i takes action a^*, π_{i,a^*}, can be written as

$$\frac{d\pi_{i,a^*}}{dt} = (f_{i,a^*} - f_{i,\pi})\pi_{i,a^*}, \tag{5}$$

where f_{i,a^*} is the payoff when player i takes action a^* and $f_{i,\pi}$ is the average payoff when player i applies strategy π.

In a two-player two-action game, the expected payoffs of cooperation and defection of player P can be written as

$$\begin{aligned} f(c, \boldsymbol{q}) &= a_{cc}q + a_{cd}(1-q) = (A\boldsymbol{q})_c, \\ f(d, \boldsymbol{q}) &= a_{dc}q + a_{dd}(1-q) = (A\boldsymbol{q})_d. \end{aligned} \tag{6}$$

and similarly we can write the expected payoff of strategy $\boldsymbol{p}$ as

$$f(\boldsymbol{p}, \boldsymbol{q}) = pf(c, \boldsymbol{q}) + (1-p)f(d, \boldsymbol{q}) = \boldsymbol{p}^{\mathrm{T}} A \boldsymbol{q}. \tag{7}$$

Thus, the two-player two-action replicator dynamics can be defined as the following system of ordinary differential equations

$$\begin{aligned} \frac{dp}{dt} &= \left[(A\boldsymbol{q})_c - \boldsymbol{p}^{\mathrm{T}} A \boldsymbol{q} \right] p, \\ \frac{dq}{dt} &= \left[(B^{\mathrm{T}}\boldsymbol{p})_c - \boldsymbol{q}^{\mathrm{T}} B^{\mathrm{T}} \boldsymbol{p} \right] q. \end{aligned} \tag{8}$$

As there are only two actions for each player, the replicator dynamics can also be written as

$$\begin{aligned} \frac{dp}{dt} &= \left[(A\boldsymbol{q})_c - (A\boldsymbol{q})_d \right] (1-p)p, \\ \frac{dq}{dt} &= \left[(B^{\mathrm{T}}\boldsymbol{p})_c - (B^{\mathrm{T}}\boldsymbol{p})_d \right] (1-q)q. \end{aligned} \tag{9}$$

3 REPLICATOR DYNAMICS IN STOCHASTIC GAMES

If the game has only one state, the joint action results in a deterministic outcome and brings a deterministic reward to the players. However, in a stochastic game, as the joint action influences the next state players being in, the average reward of joint action in one state is affected by the future rewards players can receive. Thus, how to evaluate the payoff of an action and a strategy in one state in stochastic games is a critical point to derive the replicator dynamics in stochastic games.

3.1 State-coupled replicator dynamics

In work [8], the authors propose an approach named state-coupled replicator dynamics to derive the replicator dynamics in stochastic games. In their work, they first need to get an average reward game of each state. For a stochastic game $G = \langle n, \boldsymbol{S}, \boldsymbol{\Omega}, z, \tau, \pi_1, \cdots, \pi_n \rangle$ and state $s \in \boldsymbol{S}$, the average reward game is defined as

$$\bar{G}(s, \pi_1 \dots \pi_n) = \left\langle n, \Omega_1(s) \dots \Omega_n(s), \bar{\tau}, \pi_1(s') \dots \pi_n(s') \right\rangle \tag{10}$$

where each player i plays a fixed strategy $\pi_i(s')$ in all states $s' \neq s$. The payoff function $\bar{\tau}_i$ is given by

$$\bar{\tau}_i(s, \boldsymbol{a}) = x_s(s, \boldsymbol{a})\tau_i(s, \boldsymbol{a}) + \sum_{s' \in S - \{s\}} x_{s'}(s, \boldsymbol{a}) F_i(s') \tag{11}$$

where

$$F_i(s') = \sum_{\boldsymbol{a}' \in \prod_{i=1}^n \Omega_i(s')} \left(\tau_i(s', \boldsymbol{a}') \prod_{i=1}^n \pi_{i,a_i'}(s') \right). \tag{12}$$

In Equation (11), $x_{s^*}(s, \boldsymbol{a})$ is a stationary distribution over all states $\boldsymbol{S}$, given the joint action $\boldsymbol{a}$ being played in state s. The distribution meets the condition that $\sum_{s^* \in \boldsymbol{S}} x_{s^*}(s, \boldsymbol{a}) = 1$ and

$$x_{s^*}(s, \boldsymbol{a}) = x_s(s, \boldsymbol{a}) q_{s^*}(s, \boldsymbol{a}) + \sum_{s' \in S - \{s\}} x_{s'}(s, \boldsymbol{a}) Z_i(s'), \tag{13}$$

where

$$Z_i(s') = \sum_{\boldsymbol{a}' \in \prod_{i=1}^n \Omega_i(s')} \left(z_{s^*}(s', \boldsymbol{a}') \prod_{i=1}^n \pi_{i,a_i'}(s') \right). \tag{14}$$

Given the average reward games of each state, the state-coupled replicator dynamics are defined by the following system of differential equations:

$$\frac{d\pi_{i,j}(s)}{dt} = \left[f_i(s, \boldsymbol{e}_j) - f_i(s, \pi_i(s)) \right] \pi_{i,j} x_s(\pi) \tag{15}$$

where $\boldsymbol{e}_j$ is the j^{th}-unit vector which means the case that player i takes action j. $f_i(s, \omega)$ is the expected payoff for a player i playing strategy ω in state s. f_i is defined as

$$f_i(s, \omega) = \sum_{j \in \Omega_i(s)} \left[\omega_j \sum_{\boldsymbol{a} \in \prod_{l \neq i} \Omega_l(s)} \left(\bar{\tau}_i(s, \boldsymbol{a}) \prod_{l \neq i} \pi_{l,a_l}(s) \right) \right] \tag{16}$$

where joint action $\boldsymbol{a} = (a_1, \cdots, a_{i-1}, j, a_{i+1}, \cdots, a_n)$. x is the stationary distribution over all states under strategy π, with

$$\sum_{s \in S} x_s(\pi) = 1 \text{ and} \tag{17}$$

$$x_s(\pi) = \sum_{s^* \in S} \left[x_{s^*}(\pi) \sum_{\boldsymbol{a} \in \prod_{i=1}^n \Omega_i(s^*)} \left(z_s(s^*, \boldsymbol{a}) \prod_{i=1}^n \pi_{i,a_i}(s^*) \right) \right] \tag{18}$$

3.2 State-transition replicator dynamics

In the derivation of state-coupled replicator dynamics, one critical step is to get the expected payoff $f_i(s, \boldsymbol{a})$ for the joint action $\boldsymbol{a}$ taken in state s. The expected payoff is given by adding the immediate payoff for joint action $\boldsymbol{a}$ in state s with the expected payoffs in all other states. Expected payoffs are then weighted by the frequency of corresponding state occurrences. As calculating the expected payoffs in other states and the frequency of state occurrences, the strategies in other states are assumed to be fixed.

In this paper, we propose a new approach to derive the average reward game and replicator dynamics in stochastic games, by the means of describing the dynamic process in stochastic game as a Markov chain. As we use the properties of Markov chain and transition matrix in the derivation of replicator dynamics, we call this approach as state-transition replicator dynamics.

One outcome of the Markov chain for stochastic game is the combination of state s and the joint action $\boldsymbol{a}$ in this state, which is written as $O_{(s,\boldsymbol{a})}$. In total, the Markov chain has number of $N = \sum_{s\in S}\prod_{i=1}^{n}|\Omega_i(s)|$ possible outcomes. For a stochastic game $G = \langle n, \boldsymbol{S}, \boldsymbol{\Omega}, z, \tau, \pi_1, \cdots, \pi_n\rangle$, transitions between outcomes in the Markov chain can be represented by a transition matrix $M(G)$ with the entries equaling to

$$P(X_{t+1} = O_{(s,\boldsymbol{a})}|X_t = O_{(s',\boldsymbol{a'})}) = z_s(s', \boldsymbol{a'})\prod_{i=1}^{n}\pi_{i,a_i}(s). \tag{19}$$

The entry of the transition matrix means the probability that the outcome moves from $O_{(s',\boldsymbol{a'})}$ to $O_{(s,\boldsymbol{a})}$. This probability is determined by the multiplication of two parts. One is the state transition probability, $z_s(s', \boldsymbol{a'})$). The other is the probability of joint action $\boldsymbol{a}$ taking place in state s, which is written as $\prod_{i=1}^{n}\pi_{i,a_i}(s)$, where $a_i \in \boldsymbol{a}$

We define the vector $\boldsymbol{v}(G)$ as the left eigenvector of the transition matrix $M(G)$ corresponding to the eigenvalue 1. This vector represents the stationary distribution of outcomes in the Markov chain. The entry $v_{(s,\boldsymbol{a})}$ of this vector is the expected frequencies to observe the outcome $O_{(s,\boldsymbol{a})}$ over the course of the stochastic game. Furthermore, we know the immediate payoffs for players given a state s and the joint action $\boldsymbol{a}$. Thus, for stochastic game G, the expected payoff of player i can then be computed by

$$f_i(G) = \sum_{s\in\boldsymbol{S},\boldsymbol{a}\in\prod_{l=1}^{n}\Omega_l(s)} v_{(s,\boldsymbol{a})}\cdot\tau_i(s,\boldsymbol{a}). \tag{20}$$

For obtaining the average reward game, we need to calculate the expected payoffs of the combinations of every state and every joint action. The joint action in state s can be seen as that each player takes a pure strategy in this state. Given that players take the joint action $\boldsymbol{a}^*$ in state s^*, the transition matrix $M^*(G)$ for this combination of state and joint action can be written as

$$P(O_{(s,\boldsymbol{a})}|O_{(s',\boldsymbol{a'})}) = \begin{cases} 0 & \text{if } s = s^*, \boldsymbol{a} \neq \boldsymbol{a}^* \\ z_s(s', \boldsymbol{a'}) & \text{if } s = s^*, \boldsymbol{a} = \boldsymbol{a}^* \\ z_s(s', \boldsymbol{a'})\prod_{i=1}^{n}\pi_{a_i}^{i}(s) & \text{if } s \neq s^* \end{cases} \tag{21}$$

With transition matrix $M^*(G)$, we can get the stationary distribution $\boldsymbol{v}^*(G)$ with entries as $v^*_{(s,\boldsymbol{a})}$. The expected payoff function $\bar{\tau}^i(s^*, \boldsymbol{a}^*)$ is given by

$$\bar{\tau}_i(s^*, \boldsymbol{a}^*) = \sum_{s\in\boldsymbol{S},\boldsymbol{a}\in\prod_{l=1}^{n}\Omega_l(s)} v^*_{(s,\boldsymbol{a})}\cdot\tau_i(s,\boldsymbol{a}). \tag{22}$$

Given the average reward games, we can use Equations (15) and (16) to derive the system of replicator dynamics. The stationary distribution over all states under strategy π equals to

$$x_s(\pi) = \sum_{\boldsymbol{a}\in\prod_{i=1}^{n}\Omega_i(s)} v_{(s,\boldsymbol{a})} \tag{23}$$

3.3 Results of strategy trajectory traces

In this section, we plot multiple strategy trajectory traces originating from learning automata as well as state-coupled and state-transition replicator dynamics based on the stochastic game presented in [8]. These results illustrate that state-transition replicator dynamics can model the learning process precisely.

The payoff matrices of a two-state two-player two-action stochastic game are given by

$$\left(A^1, B^1\right) = \begin{pmatrix} 3,3 & 0,10 \\ 10,0 & 2,2 \end{pmatrix}, \left(A^2, B^2\right) = \begin{pmatrix} 4,4 & 0,10 \\ 10,0 & 1,1 \end{pmatrix}. \tag{24}$$

The transition probabilities are defined as

$$\begin{aligned} z_{s'}(s, cc) = z_{s'}(s, dd) = 0.1, \\ z_{s'}(s, cd) = z_{s'}(s, dc) = 0.9, \end{aligned} \tag{25}$$

where $s, s' \in \boldsymbol{S}$ and $s \neq s'$. Player P (Q) has strategies $\boldsymbol{p}^1$ ($\boldsymbol{q}^1$) and $\boldsymbol{p}^2$ ($\boldsymbol{q}^2$) in state s^1 and s^2, respectively. The pure stationary equilibrium reflects the strategies where one of the players defects in one state while cooperating in other and the second player does exactly the opposite.

Figure 2 presents trajectory plots for this stochastic game. The state-transition replicator dynamics can model the learning dynamics precisely, as well as the state-coupled replicator dynamics proposed in [8]. Both of these two methods describe the coupling between states. The main difference in these two methods is the approach to obtain the average reward game and the stationary distribution over all states under strategy π. State-coupled method copes with these problems by building a system of multi-variable equations and solving it. The state-transition replicator dynamics method proposed in this paper solves these problems by viewing the dynamic process in stochastic games as a Markov chain. Then, the average reward game and the stationary distribution over states can be obtained through calculating the stationary vector of the transition matrix.

4 APPLICATIONS

By viewing the stochastic game in the perspective of Markov chain and utilizing the properties of transition matrix, we can get a powerful tool to analyze and design some specific stochastic games.

4.1 Unbalanced transition probabilities between states

We first illustrate that how transition probabilities between states influence the dynamics of strategies and promote cooperation in social dilemmas. It is assumed that players take part in three alike

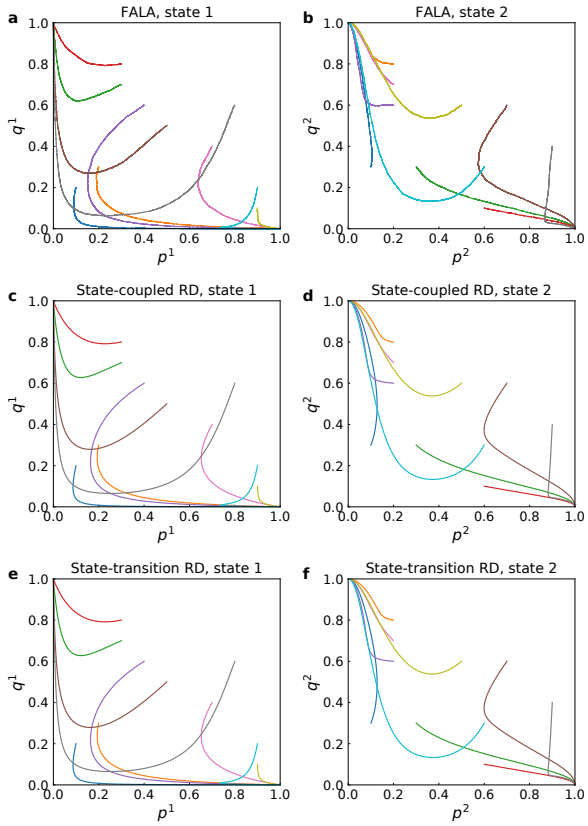

Figure 2: The strategy trajectory traces of learning automata, state-coupled and state-transition replicator dynamics. Initial strategy profiles are picked randomly in state 1 and state 2 at start.

Prisoner's Dilemma games in a stochastic game. The payoff matrices in each state are set as

$$(A^1, B^1) = (A^2, B^2) = (A^3, B^3) = \begin{pmatrix} 3,3 & 1,4 \\ 4,1 & 2,2 \end{pmatrix} \tag{26}$$

The difference between these three states is defined by the transition probabilities. State 1 is assumed to be a cooperation back state. If players cooperate mutually, they have a high probability to be in state 1 in the next round. Otherwise, they have a high probability to stay in state 2 or state 3. Here, the transition probabilities are set as $z_{s^1}(s, cc) = 0.9, z_{s^1}(s, \boldsymbol{a}) = 0.1, z_{s^2}(s, cc) = z_{s^3}(s, cc) = (1 - z_{s^1}(s, cc))/2$ and $z_{s^2}(s, \boldsymbol{a}) = z_{s^3}(s, \boldsymbol{a}) = (1 - z_{s^1}(s, \boldsymbol{a}))/2$, where $s \in \boldsymbol{S}$ and $\boldsymbol{a} \in \prod_{i=1}^{n} \Omega^i(s) \setminus (cc)$.

Figure 3 plots the strategy trajectory traces originating from learning automata and state-transition replicator dynamics in the stochastic game with unbalanced transition probabilities. It can be observed that, for most of initialization conditions shown in Figure 3, both players P and Q converge to cooperation in state s^1 and tend to defect mutually in states s^2 and s^3. The mutual defection in states s^2 and s^3 can be seen as a punishment for players if they choose defection state s^1. The temptation of defection in state s^1 can not cover the loss of leaving it. However, there still exist some initialization conditions with that players turn to be mutually defective in state s^1. With the help of state-transition replicator dynamics model, we can investigate what initialization conditions that players need to meet for the prevalence of cooperation.

For player i in a stochastic game, we define the vector $\theta_i = (a^1, \cdots, a^k)$ as the combination of actions taken by player i in each state s. For player i and state s, by setting the strategy as follow,

$$\pi_{i,a}(s) = \begin{cases} 1 & \text{if } a = a^s \\ 0 & \text{if } a \neq a^s, \end{cases} \tag{27}$$

and substituting it into the Equations (19) and (20), we can get the expected payoffs for player i with combination of actions θ_i against other players' strategies.

In stochastic game defined in this section, players have two actions, c and d, in three states. Through enumerating all the possible combinations of actions for player P and Q, we can get an expected payoff matrix shown as below,

	(c, c, c)	(c, c, d)	(c, d, c)	(c, d, d)
(c, c, c)	3.0, 3.0	2.83, 3.08	2.83, 3.08	2.0, 3.5
(c, c, d)	3.08, 2.83	2.92, 2.92	2.75, 2.75	2.25, 3.0
(c, d, c)	3.08, 2.83	2.75, 2.75	2.92, 2.92	2.25, 3.0
(c, d, d)	3.5, 2.0	3.0, 2.25	3.0, 2.25	2.5, 2.5
(d, c, c)	3.5, 2.0	2.71, 2.61	2.71, 2.61	1.3, 3.7
(d, c, d)	3.68, 1.64	3.04, 1.96	2.65, 2.35	1.75, 2.8
(d, d, c)	3.68, 1.64	2.65, 2.35	3.04, 1.96	1.75, 2.8
(d, d, d)	4.0, 1.0	3.1, 1.45	3.1, 1.45	2.2, 1.9

2.0, 3.5	1.64, 3.68	1.64, 3.68	1.0, 4.0
2.61, 2.71	1.96, 3.04	2.35, 2.65	1.45, 3.1
2.61, 2.71	2.35, 2.65	1.97, 3.04	1.45, 3.1
3.7, 1.3	2.8, 1.75	2.8, 1.75	1.9, 2.2
2.5, 2.5	2.0, 2.97	2.0, 2.97	1.1, 3.8
2.96, 2.0	2.32, 2.32	2.45, 2.45	1.55, 2.9
2.96, 2.0	2.45, 2.45	2.32, 2.32	1.55, 2.9
3.8, 1.1	2.9, 1.55	2.9, 1.55	2.0, 2.0
(d, c, c)	(d, c, d)	(d, d, c)	(d, d, d)

The entries of the matrix represent the expected payoffs for player P and Q with combination of actions θ_P and θ_Q. The first value in each entry represents the payoff for row player P and the second one for column player Q.

We can find that there exist two pure Nash equilibria, $((c, d, d), (c, d, d))$ and $((d, d, d), (d, d, d))$. That is to say, for both of players, the best actions in states s^2 and s^3 are defection. However, in state s^1, an unstable equilibrium point exists in the perspective of evolutionary game theory. By eliminating the dominated actions, we can get a sub payoff matrix shown as follow,

$$\begin{array}{c} \\ (c, d, d) \\ (d, d, d) \end{array} \begin{array}{c} \begin{array}{cc} (c, d, d) & (d, d, d) \end{array} \\ \begin{pmatrix} 2.5, 2.5 & 1.9, 2.2 \\ 2.2, 1.9 & 2.0, 2.0 \end{pmatrix} \end{array}. \tag{28}$$

For player P, the expected payoff of cooperation in state s^1, $f(c, \boldsymbol{q}^1)$, equals $2.5q^1 + 1.9(1 - q^1)$ and the expected payoff of defection, $f(d, \boldsymbol{q}^1)$, equals $2.2q^1 + 2.0(1 - q^1)$. The dynamic direction

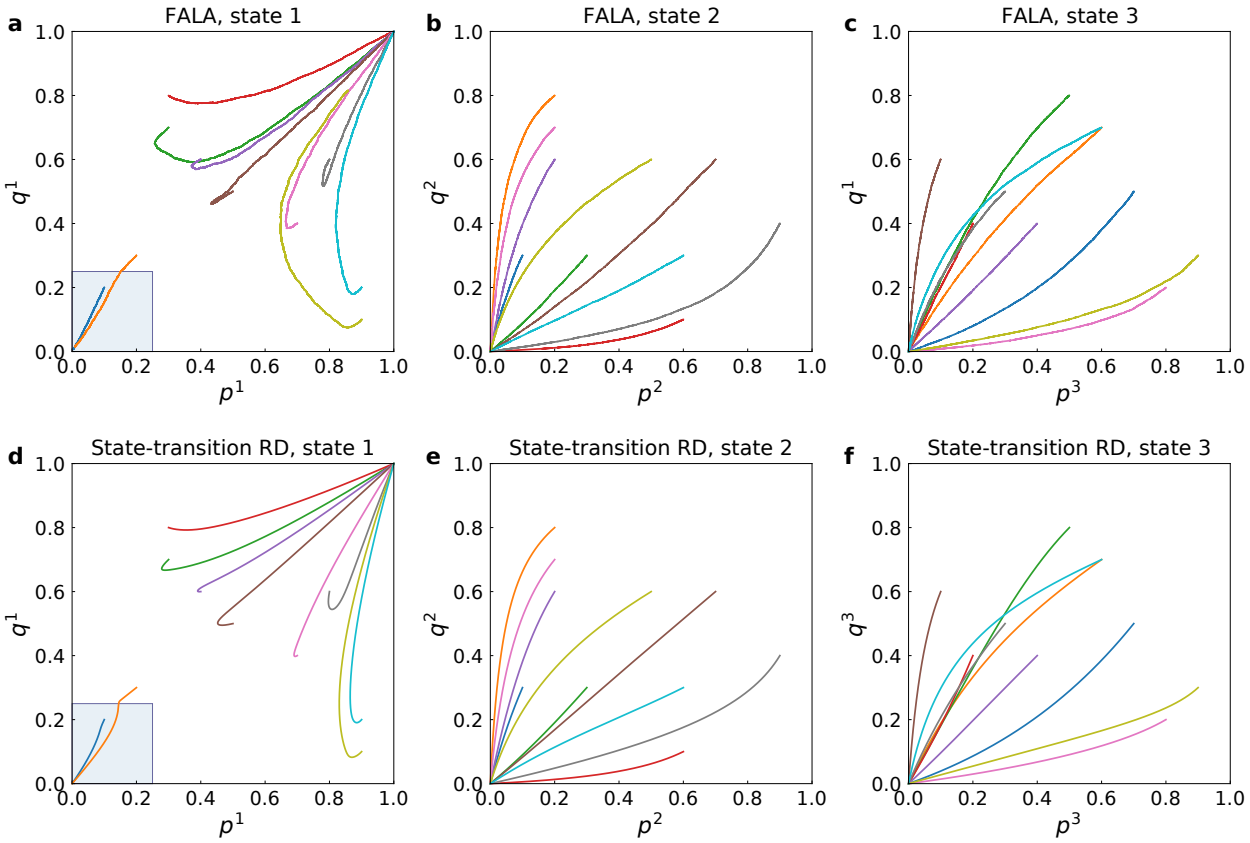

Figure 3: The strategy trajectory traces of learning automata and state-transition replicator dynamics in stochastic games with unbalanced transition probabilities. Initial strategy profiles are picked randomly in states s^1, s^2 and s^3 at start. The shadow areas in panels (a) and (d) illustrate the area into where strategies p^1 and q^1 turn to be 0 when they fall.

of strategy p^1 is determined by

$$f(c, \boldsymbol{q}^1) - f(d, \boldsymbol{q}^1) = 0.4q^1 - 0.1. \tag{29}$$

There exists a value $\hat{q}^1 = 0.25$ leading $f(c, \boldsymbol{q}^1) - f(d, \boldsymbol{q}^1)$ to be 0. When the strategy q^1 is larger than $\hat{q}^1$, player P would like to be cooperative in state s^1, and vice versa. In Figure 3 (a) and (d), there exists a shadow area where $p^1 < 0.25$ and $q^1 < 0.25$. If strategies p^1 and q^1 initially fall into this area or evolve into this area, both players P and Q would like to change to be defective.

Figure 4 illustrates the relationship between dp^1/dt and q^1 for different parameters $z_{s^1}(s, cc)$ when the strategies of players P and Q are purely defective in state s^2 and s^3. Meanwhile, the parameter $z_{s^1}(s, \boldsymbol{a})$ is fixed as 0.1 for $\boldsymbol{a} \in \prod_{i=1}^{n} \Omega^i(s) \setminus (cc)$. We can find that the unstable equilibrium point $\hat{q}^1$ turns to be smaller with the decrease of $z_{s^1}(s, cc)$. When $z_{s^1}(s, cc)$ becomes 0.5, the unstable equilibrium point vanishes. The strategy p^1 turns to be defective no matter what strategy q^1 is.

4.2 Control the payoffs of players

In this section, by extending the analysis method proposed in [15] from stateless game to stochastic game and utilizing the state-transition replicator dynamics, we demonstrate that a set of specifically derived transition probabilities in stochastic games can control the payoffs of players.

For a two-state (s^1, s^2) two-player (P, Q) two-action (c, d) stochastic game G, the whole outcomes of the Markov chain can be written as a vector, $\boldsymbol{O} = ((s^1, cc), (s^1, cd), (s^1, dc), (s^1, dd), (s^2, cc), (s^2, cd), (s^2, dc), (s^2, dd))$. For simplicity, we define that the outcomes in $\boldsymbol{O}$ are labeled $1, \cdots, 8$ following the order in the vector. The probabilities of the next state being s^1 for the respective outcomes are given by a vector $\boldsymbol{z} = (z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8)$. $\boldsymbol{v}$ is the stationary vector of transition matrix M defined in Equation (19). The vector $\boldsymbol{v}$ satisfies

$$\boldsymbol{v}M = \boldsymbol{v} \text{ or } \boldsymbol{v}M' = 0, \tag{30}$$

where $M' = M - I$. I is the identity matrix with the same dimension of M. Because the transition matrix M is stochastic and has a unit eigenvalue, the matrix M' is singular, with thus zero determinant. Further, by Cramer's rule, we have

$$Adj(M')M' = det(M')I = 0, \tag{31}$$

where 0 is a zero matrix with the same dimensions of M'. $Adj(M')$ is the adjoint matrix of M'. Let m^*_{ij} be the (i, j) entry of the cofactor

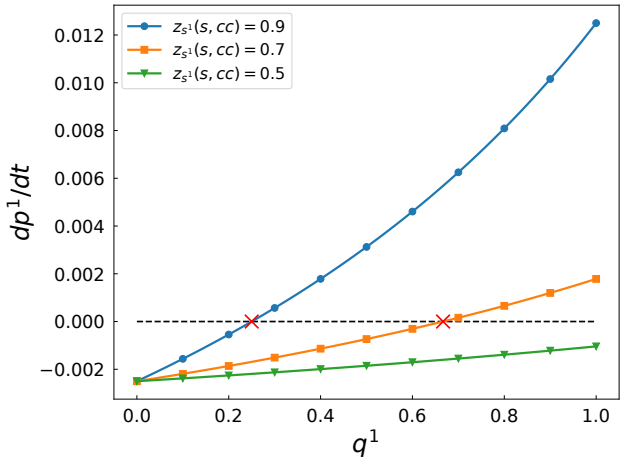

Figure 4: Gradient of strategy p^1 as a function of the strategy q^1 for different parameters $z_{s^1}(s, cc)$. The strategies of players P and Q in state s^2 and s^3 are fixed as purely defective. The parameter $z_{s^1}(s, \boldsymbol{a})$ is fixed as 0.1 for $\boldsymbol{a} \in \prod_{i=1}^{n} \Omega^i(s) \setminus (cc)$. The crosses represent the unstable equilibrium points where dp^1/dt equals 0. The value of p^1 can influence the value of dp^1/dt, but can not move the unstable equilibrium points. In this figure, we fix p^1 as 0.5. The left cross is for $z_{s^1}(s, cc) = 0.9$ and $\hat{q}^1$ equals 0.25. The right cross is for $z_{s^1}(s, cc) = 0.7$ and $\hat{q}^1$ equals $2/3$.

matrix of M', then $Adj(M')$ can be expressed as follow,

$$Adj(M') = \begin{bmatrix} m^*_{11} & m^*_{21} & m^*_{31} & m^*_{41} & m^*_{51} & m^*_{61} & m^*_{71} & m^*_{81} \\ m^*_{12} & m^*_{22} & m^*_{32} & m^*_{42} & m^*_{52} & m^*_{62} & m^*_{72} & m^*_{82} \\ m^*_{13} & m^*_{23} & m^*_{33} & m^*_{43} & m^*_{53} & m^*_{63} & m^*_{73} & m^*_{83} \\ m^*_{14} & m^*_{24} & m^*_{34} & m^*_{44} & m^*_{54} & m^*_{64} & m^*_{74} & m^*_{84} \\ m^*_{15} & m^*_{25} & m^*_{35} & m^*_{45} & m^*_{55} & m^*_{65} & m^*_{75} & m^*_{85} \\ m^*_{16} & m^*_{26} & m^*_{36} & m^*_{46} & m^*_{56} & m^*_{66} & m^*_{76} & m^*_{86} \\ m^*_{17} & m^*_{27} & m^*_{37} & m^*_{47} & m^*_{57} & m^*_{67} & m^*_{77} & m^*_{87} \\ m^*_{18} & m^*_{28} & m^*_{38} & m^*_{48} & m^*_{58} & m^*_{68} & m^*_{78} & m^*_{88} \end{bmatrix}. \tag{32}$$

From Equations (30) and (31), every row of $Adj(M')$ is proportional to $\boldsymbol{v}$. Thus, we have $\boldsymbol{v} = \rho(m^*_{18}, m^*_{28}, m^*_{38}, m^*_{48}, m^*_{58}, m^*_{68}, m^*_{78}, m^*_{88})$ for some $\rho \neq 0$.

By adding the first column, the second column and the third column to the fourth column in the matrix M', we can get

$$H = \begin{bmatrix} z_1p^1q^1 - 1 & z_1p^1\left(1-q^1\right) & z_1\left(1-p^1\right)q^1 & z_1 - 1 & (1-z_1)\,p^2q^2 \\ z_2p^1q^1 & z_2p^1\left(1-q^1\right) - 1 & z_2\left(1-p^1\right)q^1 & z_2 - 1 & (1-z_2)\,p^2q^2 \\ z_3p^1q^1 & z_3p^1\left(1-q^1\right) & z_3\left(1-p^1\right)q^1 - 1 & z_3 - 1 & (1-z_3)\,p^2q^2 \\ z_4p^1q^1 & z_4p^1\left(1-q^1\right) & z_4\left(1-p^1\right)q^1 & z_4 - 1 & (1-z_4)\,p^2q^2 \\ z_5p^1q^1 & z_5p^1\left(1-q^1\right) & z_5\left(1-p^1\right)q^1 & z_5 & (1-z_5)\,p^2q^2 - 1 \\ z_6p^1q^1 & z_6p^1\left(1-q^1\right) & z_6\left(1-p^1\right)q^1 & z_6 & (1-z_6)\,p^2q^2 \\ z_7p^1q^1 & z_7p^1\left(1-q^1\right) & z_7\left(1-p^1\right)q^1 & z_7 & (1-z_7)\,p^2q^2 \\ z_8p^1q^1 & z_8p^1\left(1-q^1\right) & z_8\left(1-p^1\right)q^1 & z_8 & (1-z_8)\,p^2q^2 \end{bmatrix}$$

$$\begin{bmatrix} (1-z_1)\,p^2\left(1-q^2\right) & (1-z_1)\left(1-p^2\right)q^2 & (1-z_1)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_2)\,p^2\left(1-q^2\right) & (1-z_2)\left(1-p^2\right)q^2 & (1-z_2)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_3)\,p^2\left(1-q^2\right) & (1-z_3)\left(1-p^2\right)q^2 & (1-z_3)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_4)\,p^2\left(1-q^2\right) & (1-z_4)\left(1-p^2\right)q^2 & (1-z_4)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_5)\,p^2\left(1-q^2\right) & (1-z_5)\left(1-p^2\right)q^2 & (1-z_5)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_6)\,p^2\left(1-q^2\right) - 1 & (1-z_6)\left(1-p^2\right)q^2 & (1-z_6)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_7)\,p^2\left(1-q^2\right) & (1-z_7)\left(1-p^2\right)q^2 - 1 & (1-z_7)\left(1-p^2\right)\left(1-q^2\right) \\ (1-z_8)\,p^2\left(1-q^2\right) & (1-z_8)\left(1-p^2\right)q^2 & (1-z_8)\left(1-p^2\right)\left(1-q^2\right) - 1 \end{bmatrix}. \tag{33}$$

where the transpose of the fourth column is denoted by $\tilde{\boldsymbol{z}} \equiv (z_1-1, z_2-1, z_3-1, z_4-1, z_5, z_6, z_7, z_8)$. It is clear that $\tilde{\boldsymbol{z}}$ depends only on the transition probabilities, but not on players' strategies.

Let h^*_{ij} be the (i, j) entry of the cofactor matrix of H. Since the operations of additions of columns do not change the determinant of the matrix, we have $h^*_{i8} = m^*_{i8}$ for $i = 1, \cdots, 8$. Through replacing the eighth column of H by the transpose of an arbitrary eight-dimensional vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8)$, the value of the determinant of the corresponding matrix can be computed by expanding along the eighth column as follow,

$$det \begin{bmatrix} z_1p^1q^1 - 1 & z_1p^1\left(1-q^1\right) & z_1\left(1-p^1\right)q^1 & z_1 - 1 \\ z_2p^1q^1 & z_2p^1\left(1-q^1\right) - 1 & z_2\left(1-p^1\right)q^1 & z_2 - 1 \\ z_3p^1q^1 & z_3p^1\left(1-q^1\right) & z_3\left(1-p^1\right)q^1 - 1 & z_3 - 1 \\ z_4p^1q^1 & z_4p^1\left(1-q^1\right) & z_4\left(1-p^1\right)q^1 & z_4 - 1 \\ z_5p^1q^1 & z_5p^1\left(1-q^1\right) & z_5\left(1-p^1\right)q^1 & z_5 \\ z_6p^1q^1 & z_6p^1\left(1-q^1\right) & z_6\left(1-p^1\right)q^1 & z_6 \\ z_7p^1q^1 & z_7p^1\left(1-q^1\right) & z_7\left(1-p^1\right)q^1 & z_7 \\ z_8p^1q^1 & z_8p^1\left(1-q^1\right) & z_8\left(1-p^1\right)q^1 & z_8 \end{bmatrix}$$

$$\begin{bmatrix} (1-z_1)\,p^2q^2 & (1-z_1)\,p^2\left(1-q^2\right) & (1-z_1)\left(1-p^2\right)q^2 & \eta_1 \\ (1-z_2)\,p^2q^2 & (1-z_2)\,p^2\left(1-q^2\right) & (1-z_2)\left(1-p^2\right)q^2 & \eta_2 \\ (1-z_3)\,p^2q^2 & (1-z_3)\,p^2\left(1-q^2\right) & (1-z_3)\left(1-p^2\right)q^2 & \eta_3 \\ (1-z_4)\,p^2q^2 & (1-z_4)\,p^2\left(1-q^2\right) & (1-z_4)\left(1-p^2\right)q^2 & \eta_4 \\ (1-z_5)\,p^2q^2 - 1 & (1-z_5)\,p^2\left(1-q^2\right) & (1-z_5)\left(1-p^2\right)q^2 & \eta_5 \\ (1-z_6)\,p^2q^2 & (1-z_6)\,p^2\left(1-q^2\right) - 1 & (1-z_6)\left(1-p^2\right)q^2 & \eta_6 \\ (1-z_7)\,p^2q^2 & (1-z_7)\,p^2\left(1-q^2\right) & (1-z_7)\left(1-p^2\right)q^2 - 1 & \eta_7 \\ (1-z_8)\,p^2q^2 & (1-z_8)\,p^2\left(1-q^2\right) & (1-z_8)\left(1-p^2\right)q^2 & \eta_8 \end{bmatrix} \tag{34}$$

$$= \eta_1h^*_{18} + \eta_2h^*_{28} + \eta_3h^*_{38} + \eta_4h^*_{48} + \eta_5h^*_{58} + \eta_6h^*_{68} + \eta_7h^*_{78} + \eta_8h^*_{88}.$$

Recalling that $h^*_{i8} = m^*_{i8}$ for $i = 1, \cdots, 8$ and $\boldsymbol{v} = \rho(m^*_{18}, m^*_{28}, m^*_{38}, m^*_{48}, m^*_{58}, m^*_{68}, m^*_{78}, m^*_{88})$ for $\rho \neq 0$, then Equation (34) implies that $\eta_1h^*_{18}+\eta_2h^*_{28}+\eta_3h^*_{38}+\eta_4h^*_{48}+\eta_5h^*_{58}+\eta_6h^*_{68}+\eta_7h^*_{78}+\eta_8h^*_{88} = \frac{1}{\rho}(\boldsymbol{v} \cdot \boldsymbol{\eta})$. Then, we can have

$$\boldsymbol{v} \cdot \boldsymbol{\eta} = \rho D(\boldsymbol{z}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, \boldsymbol{\eta}) \tag{35}$$

where $D(\boldsymbol{z}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, \boldsymbol{\eta})$ equals to the determinant value defined in Equation (34). $\hat{\boldsymbol{p}} = (p^1, p^2)$ $(\hat{\boldsymbol{q}} = (p^1, q^2))$ represents the combination of strategies of player P (Q) in state s^1 and s^2. Player P's payoff vector is $\boldsymbol{S}_P = (a^1_{cc}, a^1_{cd}, a^1_{dc}, a^1_{dd}, a^2_{cc}, a^2_{cd}, a^2_{dc}, a^2_{dd})$, whereas player Q's

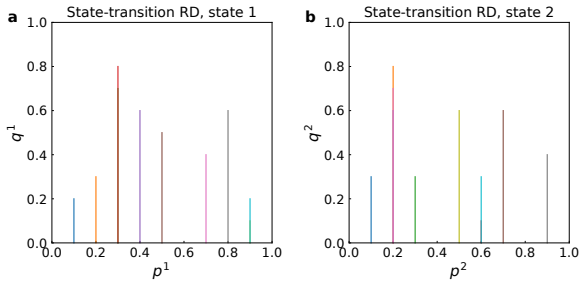

Figure 5: The strategy trajectory traces of state-transition replicator dynamics in stochastic games with ZD transition probabilities. Initial strategy profiles are picked randomly in state s^1 and s^2. Strategies p^1 and p^2 do not change and keep the initial value. Meanwhile, strategies q^1 and q^2 update to be 0 in social dilemmas.

payoff vector is $\boldsymbol{S}_Q = (b^1_{cc}, b^1_{cd}, b^1_{dc}, b^1_{dd}, b^2_{cc}, b^2_{cd}, b^2_{dc}, b^2_{dd})$. Their respective expected payoffs are

$$\begin{aligned} E_P &= \frac{\boldsymbol{v}\cdot\boldsymbol{S}_P}{\boldsymbol{v}\cdot 1} = \frac{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},\boldsymbol{S}_P)}{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},1)} \\ E_Q &= \frac{\boldsymbol{v}\cdot\boldsymbol{S}_Q}{\boldsymbol{v}\cdot 1} = \frac{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},\boldsymbol{S}_Q)}{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},1)} \end{aligned} \tag{36}$$

where 1 is the vector with all components 1.

The expected payoff for one player depends linearly on his own payoff vector. For player P, we can write his expected payoff as a linear relationship as follow,

$$\alpha E_P + \gamma = \frac{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},\alpha\boldsymbol{S}_P+\gamma 1)}{D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},1)}. \tag{37}$$

The determinant value of any matrix is zero if two of its columns are identical or one column is multiple of the other. Since that the fourth column only contains the entries of transition probabilities, the transition probabilities can equate the determinant value $D(\boldsymbol{z},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},\alpha\boldsymbol{S}_P+\gamma 1)$ to 0. This will be true if transition probabilities satisfy the case $\tilde{\boldsymbol{z}} \equiv (z_1-1, z_2-1, z_3-1, z_4-1, z_5, z_6, z_7, z_8) = \alpha\boldsymbol{S}_P+\gamma 1$, which means

$$\alpha E_P + \gamma = \frac{D(\tilde{\boldsymbol{z}},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},\alpha\boldsymbol{S}_P+\gamma 1)}{D(\tilde{\boldsymbol{z}},\hat{\boldsymbol{p}},\hat{\boldsymbol{q}},1)} = 0. \tag{38}$$

That is to say, the expected payoff of player P is fixed as a constant value, $-\frac{\gamma}{\alpha}$. We call such transition probabilities as zero-determinant (ZD) transition probabilities. Player P cannot change his expected payoff by alternating his strategies. Thus, in the stochastic game, player P has no motivation to update his strategies and will keep his initial ones. Meanwhile, his opponent player Q can update his own strategies to pursue higher expected payoffs.

Let us consider the following payoff matrices representing Prisoner's Dilemma games in a stochastic game

$$(A^1, B^1) = \begin{pmatrix} 3,3 & 1,4 \\ 4,1 & 2,2 \end{pmatrix}, (A^2, B^2) = \begin{pmatrix} 8,8 & 6,9 \\ 9,6 & 7,7 \end{pmatrix}. \tag{39}$$

With parameters $\alpha = 0.2$ and $\gamma = -1.0$, we can obtain the ZD transition probabilities $\tilde{\boldsymbol{z}} = (0.6, 0.2, 0.8, 0.4, 0.6, 0.8, 0.2, 0.4)$ and the value $E_P = -\frac{\gamma}{\alpha} = 5$. As shown in Figure 5, player P does not change his strategies, but keeps his initialized strategies. Meanwhile, player Q turns to be fully defective to pursue higher expected payoffs in the social dilemmas. State-transition replicator dynamics can represent the theoretical analysis perfectly.

5 CONCLUSION

In this paper, we propose a new approach named state-transition replicator dynamics for analyzing the dynamics of multi-agent learning in multi-state stochastic games. By describing the dynamic process in stochastic game as a Markov chain and utilizing the properties of transition matrix, we obtain a set of replicator dynamics to model the learning process in stochastic games.

Based on our approach and model, we have shown that transition probabilities have a significant influence on the dynamics of strategies. If the transition probabilities between states are unbalanced and there exists a cooperation back state, cooperative behaviors can prevail in this state when the strategies of players meet some conditions which can be derived by state-transition replicator dynamics. Moreover, it has also been proven in this paper that a set of specific transition probabilities can control the expected payoffs of players in some stochastic games. ZD transition probabilities can fix the expected payoff of an player as a constant value. No matter what strategies he applies, his expected payoffs do not change. Thus, such player has no motivation to update his strategies. Meanwhile, the opponent can update his own strategies to be fully defective and get an extortionate payoff in the social dilemmas.

In this paper, we consider the stochastic games where players have separate strategies in different states. Their strategies do not depend on the previous actions taken by players. That is no-memory strategy. However, many previous works have been shown that memory-one strategy, such as TFT, WSLS and ZD strategies have a great impact on the evolution of cooperation. We think that it is straight to extend our method to investigate the dynamics of memory-one strategies in stochastic games. If a player can make decisions based on the previous actions, what is the dominant strategies in stochastic games and what kind of transition probabilities can influence the dynamics of strategies are worth to be investigated.

REFERENCES

[1] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research* 53 (aug 2015), 659–697. https://doi.org/10.1613/jair.4818

[2] Michael Bowling and Manuela Veloso. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136, 2 (apr 2002), 215–250. https://doi.org/10.1016/s0004-3702(02)00121-2

[3] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (mar 2008), 156–172. https://doi.org/10.1109/tsmcc.2007.913919

[4] Thomas Dietz, Elinor Ostrom, and Paul C. Stern. 2003. The Struggle to Govern the Commons. *Science* 302, 5652 (dec 2003), 1907–1912. https://doi.org/10.1126/science.1091015

[5] Aram Galstyan. 2011. Continuous strategy replicator dynamics for multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems* 26, 1 (sep 2011), 37–53. https://doi.org/10.1007/s10458-011-9181-6

[6] Roman Gorbunov, Emilia Barakova, and Matthias Rauterberg. 2013. Design of social agents. *Neurocomputing* 114 (aug 2013), 92–97. https://doi.org/10.1016/j.neucom.2012.06.046

[7] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas,

Edgar Duèñez Guzmán, and Karl Tuyls. 2020. *Neural Replicator Dynamics: Multiagent Learning via Hedging Policy Gradients.* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 492–501.

[8] Daniel Hennes, Karl Tuyls, and Matthias Rauterberg. 2009. State-Coupled Replicator Dynamics. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2* (Budapest, Hungary) *(AAMAS '09)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 789–796.

[9] Christian Hilbe, Štěpán Šimsa, Krishnendu Chatterjee, and Martin A. Nowak. 2018. Evolution of cooperation in stochastic games. *Nature* 559, 7713 (jul 2018), 246–249. https://doi.org/10.1038/s41586-018-0277-x

[10] Michael Kaisers and Karl Tuyls. 2010. Frequency Adjusted Multi-Agent Q-Learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1* (Toronto, Canada) *(AAMAS '10)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 309–316.

[11] Tomas Klos, Gerrit Jan van Ahee, and Karl Tuyls. 2010. Evolutionary Dynamics of Regret Minimization. In *Machine Learning and Knowledge Discovery in Databases*. Springer, Springer Berlin Heidelberg, 82–96. https://doi.org/10.1007/978-3-642-15883-4_6

[12] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*. Elsevier, 157–163. https://doi.org/10.1016/b978-1-55860-335-6.50027-1

[13] Alex McAvoy and Martin A. Nowak. 2019. Reactive learning strategies for iterated games. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475, 2223 (mar 2019), 20180819. https://doi.org/10.1098/rspa.2018.0819

[14] Martin A. Nowak. 2006. Five Rules for the Evolution of Cooperation. *Science* 314, 5805 (dec 2006), 1560–1563. https://doi.org/10.1126/science.1133755

[15] William H Press and Freeman J Dyson. 2012. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences* 109, 26 (may 2012), 10409–10413. https://doi.org/10.1073/pnas.1206569109

[16] David G. Rand and Martin A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17, 8 (aug 2013), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

[17] Axelrod Robert. 1984. *The evolution of cooperation.* Basic Books, New York.

[18] Bettina Rockenbach and Manfred Milinski. 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 7120 (dec 2006), 718–723. https://doi.org/10.1038/nature05229

[19] Francisco C. Santos, Marta D. Santos, and Jorge M. Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454, 7201 (jul 2008), 213–216. https://doi.org/10.1038/nature06940

[20] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. 2010. Social learning promotes institutions for governing the commons. *Nature* 466, 7308 (jul 2010), 861–863. https://doi.org/10.1038/nature09203

[21] Satinder Singh, Michael Kearns, and Yishay Mansour. 2000. Nash Convergence of Gradient Dynamics in General-Sum Games. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence* (Stanford, California) *(UAI'00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 541–548.

[22] Qi Su, Alex McAvoy, Long Wang, and Martin A. Nowak. 2019. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences* 116, 51 (nov 2019), 25398–25404. https://doi.org/10.1073/pnas.1908936116

[23] Mohammad A. Taha and Ayman Ghoneim. 2020. Zero-determinant strategies in repeated asymmetric games. *Appl. Math. Comput.* 369 (mar 2020), 124862. https://doi.org/10.1016/j.amc.2019.124862

[24] Peter Vrancx, Pasquale Gurzi, Abdel Rodriguez, Kris Steenhaut, and Ann Nowé. 2015. A Reinforcement Learning Approach for Interdomain Routing with Link Prices. *ACM Transactions on Autonomous and Adaptive Systems* 10, 1 (mar 2015), 1–26. https://doi.org/10.1145/2719648

[25] Peter Vrancx, Karl Tuyls, and Ronald Westra. 2008. Switching Dynamics of Multi-Agent Learning. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Estoril, Portugal) *(AAMAS '08)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 307–313.

[26] Peter Vrancx, Katja Verbeeck, and Ann Nowe. 2008. Decentralized Learning in Markov Games. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38, 4 (aug 2008), 976–981. https://doi.org/10.1109/tsmcb.2008.920998

[27] Wenbo Wang, Pengda Huang, Peizhao Hu, Jing Na, and Andres Kwasinski. 2016. Learning in Markov Game for Femtocell Power Allocation with Limited Coordination. In *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6. https://doi.org/10.1109/glocom.2016.7841950