A Pair-Approximation Method for Modelling the Dynamics of Multi-Agent Stochastic Games

Chen Chu[1,2], **Zheng Yuan**[1], **Shuyue Hu**[3*], **Chunjiang Mu**[2,4], **Zhen Wang**[2,4*]
[1] School of Statistics and Mathematics, Yunnan University of Finance and Economics
[2] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
[3] Shanghai Artificial Intelligence Laboratory
[4] School of Cybersecurity, Northwestern Polytechnical University
hushuyue@pjlab.org.cn, w-zhen@nwpu.edu.cn

Abstract

Developing a dynamical model for learning in games has attracted much recent interest. In stochastic games, agents need to make decisions in multiple states, and transitions between states, in turn, influence the dynamics of strategies. While previous works typically focus either on 2-agent stochastic games or on normal form games under an infinite-agent setting, we aim at formally modelling the learning dynamics in stochastic games under the infinite-agent setting. With a novel use of pair-approximation method, we develop a formal model for myopic Q-learning in stochastic games with symmetric state transition. We verify the descriptive power of our model (a partial differential equation) across various games through comparisons with agent-based simulation results. Based on our proposed model, we can gain qualitative and quantitative insights into the influence of transition probabilities on the dynamics of strategies. In particular, we illustrate that a careful design of transition probabilities can help players overcome the social dilemmas and promote cooperation, even if agents are myopic learners.

Introduction

Recent years have witnessed a significant gain in the capability of multi-agent reinforcement learning (MARL). However, the theory underlying MARL is still far from being well understood. One line of emergent research is to examine the evolutionary dynamics of learning in games (Bloembergen et al. 2015; Boone and Piliouras 2019; Cheung 2018; Leonardos, Piliouras, and Spendlove 2021). As Tuyls and Parsons (2007) voice, the development of theory in this direction is crucial because it will not only yield a better theoretical understanding of existing algorithms, but potentially facilitate the design of new methods, leading to practical algorithmic advancements.

In their seminal work, Tuyls et al. (2003) propose the *selection-mutation model* to formalize the dynamics of Q-learning (Watkins and Dayan 1992) with Boltzmann exploration in 2-player normal-form games. This model reveals a surprising connection between multi-agent Q-learning and the well-known *replicator dynamics* of evolutionary game theory (EGT); thus, it has inspired many works in the literature. Along this line of research, Panozzo et al. (2014) generalize this model for Q-learning that operates on extensive form games. Hennes et al. (2009) introduce a state-coupled variant of this model, extending it to 2-player stochastic games. However, these early works typically focus on two-agent interaction scenarios.

More recently, there have also been some works reported on modelling the Q-learning dynamics beyond the two-agent setting. Hu et al. (2019) leverages *mean field theory* and derives a model for the evolution of an infinite well-mixed population where Q-learning agents are randomly paired up to play 2-player normal form games. Using a similar mean field-theoretic approach, Leung et al. (2022) generalize the model to capture the stochastic effects of local and incomplete information; Chu et al. (2022) propose a variant of the model for Q-learning on regular graphs where each edge represents a 2-player normal form game between two vertices (agents).

In this paper, we aim to formally model the learning dynamics in stochastic games under the infinite-agent setting. Specifically, we consider the population structure to be a complete graph, where each myopic Q-learner occupies a node and each edge, connecting two agents, is associated with a stochastic game with symmetric state transitions. At a given time step, each agent takes an action to play against all of its neighbours along every edge that the agent is connected with. Then, depending on the joint actions of two connected agents, the state of the stochastic game transits. Although infinite-agent settings and stochastic games each have independently been studied, the learning dynamics in stochastic games with infinitely many agents have not been explored before. More importantly, this is a common and realistic setting in the MARL literature, as modern MARL algorithms often feature state transitions and a large number of agents (Ganapathi Subramanian et al. 2020; Yang et al. 2018; Long et al. 2020).

A typical approach to formalizing the evolution of an infinite well-mixed population is based on mean field theory, which approximates the effects of other agents on a focal agent with a mean field. This approach, however, is incompatible with our considered setting. Modelling the dynamics in stochastic games poses a new challenge—how should we reflect the correlation between strategic interactions and

[*] Corresponding authors.

environmental state transitions? For a focal agent, the state transition depends on the joint actions of itself and its opponent. On the other hand, at a given time step, as each agent plays the same action against different opponents in different states, its action choice, aiming to maximize the expected payoffs over different states, is affected by the state distribution. Putting these coupling effects together breaks a typical assumption of the mean field approximation which requires the effects of opponents on agents to be somewhat homogeneous.

To address this challenge, we find that the key is to track the co-evolution of the state distribution and opponent strategies, and that such information can be derived from agents' Q-values. We propose a novel use of the *pair-approximation* method that is well-known in statistical physics (Hauert and Szabó 2005). Intuitively, instead of considering the frequency of strategies as in mean field theory, the pair-approximation tracks the frequencies of strategy pairs. Thus, compared with the mean-field approximation, the pair-approximation method better captures the heterogeneous effects that arise from local interactions. In this work, we define a *pair* to be a tuple $\langle \mathbf{Q}^i, s, \mathbf{Q}^j \rangle$, where s is a state and $\mathbf{Q}^i$ and $\mathbf{Q}^j$ denote the Q-values of two agents i and j. We develop a partial differential equation to model the evolution of the probability distribution of these pairs on the state-Q-values space. From the probability distribution of these pairs, the state distribution, the opponent strategies, and consequently the population state at a given time step can be derived.

To illustrate our model, we consider different game configurations in our experiments and numerically solve the developed partial differential equation in those games. We show that across different games, initial conditions, transition rules, and algorithm parameters, our model always provides an accurate description of the Q-learning dynamics. More interestingly, our model shows that the state transition promotes the emergence of cooperation in social dilemma games by facilitating myopic Q-learning agents to learn the strategies which yield them higher long-term rewards. For example, compared with a repeated prisoner's dilemma (PD) game, agents are more willing to cooperate in a two-state game with the transition between a PD game and a stag hunt (SH) game. In other words, the dependence of states on pairwise interactions can greatly enhance cooperation even if agents are myopic.

Our results clearly suggest that even if agents apply myopic Q-learning, the effects of state transition are non-trivial. This provides theoretical evidence for the critical role of games with state transition under the infinite-agent setting, complementing previous results of learning in stochastic games under the 2-player setting (Deng et al. 2021). To aid investigations on such effects, our model can provide insights that are unable to obtain using previous models (Deng et al. 2021; Hennes, Tuyls, and Rauterberg 2009; Hu et al. 2022). Therefore, our model is an important theoretical contribution towards a better understanding of multi-agent Q-learning.

Preliminaries

In this paper, we consider an infinite well-mixed multi-agent system (MAS), where all agents are paired up to play stochastic games and learn their strategies through myopic Q-learning. We begin this section by providing a brief introduction to stochastic games. Subsequently, we present a learning framework for Q-learners playing stochastic games in a well-mixed population.

Stochastic Games

The key concept of stochastic games is that the current state and the joint action of agents not only determine the rewards agents can obtain in the current round, but also the state agents will stay in the next round. We follow the formal definition of stochastic games in (Hennes, Tuyls, and Rauterberg 2009) and change some notations for better illustration. A stochastic game with n agents and k states is defined by a tuple $< \mathcal{N}, \mathcal{S}, \mathcal{A}, z, r, \pi_1, \ldots, \pi_n >$. In each state $s \in \mathcal{S} = \{s_1, \ldots, s_k\}$, each agent $i \in \mathcal{N} = \{1, \ldots, n\}$ has an action set $\mathcal{A}_i(s)$ and strategy $\pi_i(s)$. The payoff function $r(s, \mathbf{a}) : \prod_{i=1}^n \mathcal{A}_i(s) \rightarrow \mathbb{R}^n$ maps the joint action $\mathbf{a} = (a^1, \ldots, a^n)$ in state s to a reward for each agent. The transition function $z(s, \mathbf{a}) : \prod_{i=1}^n \mathcal{A}_i(s) \rightarrow \Delta^{k-1}$ determines how the state transition occurs under current state s and joint action $\mathbf{a}$, where Δ^{k-1} is the $(k-1)$-simplex and $z_{s'}(s, \mathbf{a})$denotes the transition probability from state s to state s' under joint action $\mathbf{a}$.

We take a two-agent two-state two-action stochastic game as an example for further explanation. We consider that each state corresponds to a symmetric matrix game, thus different agents have identical action sets in a given state s, i.e., $\mathcal{A}_1(s) = \mathcal{A}_2(s) = \mathcal{A}(s)$. $\mathcal{A}(s)$ is the set of available actions for agents in s, and the action sets that agents have in different states can be different, that is $\mathcal{A}(s_1) = \{a_1, a_2\}$, $\mathcal{A}(s_2) = \{b_1, b_2\}$. A general form of a two-agent two-state two-action symmetric stochastic game can be given as follows:

$$\mathbf{M}_{s_1} = \begin{pmatrix} r_{a_1a_1} & r_{a_1a_2} \\ r_{a_2a_1} & r_{a_2a_2} \end{pmatrix}, \mathbf{M}_{s_2} = \begin{pmatrix} r_{b_1b_1} & r_{b_1b_2} \\ r_{b_2b_1} & r_{b_2b_2} \end{pmatrix},$$

$$\mathbf{T}_{s_1 \to s_2} = \begin{pmatrix} z_{s_2}(s_1, a_1, a_1) & z_{s_2}(s_1, a_1, a_2) \\ z_{s_2}(s_1, a_2, a_1) & z_{s_2}(s_1, a_2, a_2) \end{pmatrix},$$

$$\mathbf{T}_{s_1 \to s_1} = \mathbf{I}_2 - \mathbf{T}_{s_1 \to s_2},$$

$$\mathbf{T}_{s_2 \to s_1} = \begin{pmatrix} z_{s_1}(s_2, b_1, b_1) & z_{s_1}(s_2, b_1, b_2) \\ z_{s_1}(s_2, b_2, b_1) & z_{s_1}(s_2, b_2, b_2) \end{pmatrix},$$

$$\mathbf{T}_{s_2 \to s_2} = \mathbf{I}_2 - \mathbf{T}_{s_2 \to s_1},$$

where $\mathbf{M}_{s_1}$ and $\mathbf{M}_{s_2}$ are the payoff matrices of a row agent in state s_1 and s_2, respectively. $\mathbf{T}_{s_1 \to s_2}$ and $\mathbf{T}_{s_1 \to s_1}$ capture the transition probabilities in state s_1 given different joint actions, $z_{s_2}(s_1, a_i, a_j)$ denotes the transition probability from s_1 to s_2 under the joint action $\mathbf{a} = (a_i, a_j)$, where $i, j \in \{1, 2\}$. Likewise, $\mathbf{T}_{s_2 \to s_1}$ and $\mathbf{T}_{s_2 \to s_2}$ capture the transition probabilities in state s_2, where $z_{s_1}(s_2, b_i, b_j)$ denotes the transition probability from s_2 to s_1 under the joint action $\mathbf{a} = (b_i, b_j)$. $\mathbf{I}_2$ is a second-order matrix where all the elements equal 1.

Note that here we assume that the state transition is symmetric, which means the influence of the agent behaviors on the environment depends only on the actions, not on the identities of agents. Formally, the transition probabilities satisfy $z_{s_2}(s_1, a_i, a_j) = z_{s_2}(s_1, a_j, a_i)$ and $z_{s_1}(s_2, b_i, b_j) = z_{s_1}(s_2, b_j, b_i)$.

Q-learning for Multi-Agent Stochastic Games

We consider a well-mixed MAS containing n agents with n tends to infinity, the structure of the well-mixed system can be described by a complete graph. Each agent occupies a vertice on the graph, and each pair of agents is connected by an edge. Note that each edge represents the state the two agents stay in.

At each time step t, all agents simultaneously choose their own actions according to their strategies. Then, each agent participates in pairwise interactions with all other agents. The game played in each interaction is determined by the state in which the agent and its opponent stay. Each agent updates its strategy after receiving the immediate reward which is averaged over its $n-1$ interactions. At the end of this time step, for each pair of agents, the state transition occurs according to the transition rule. The transition rule implies the coupling between game plays and state transitions in the environment. At time t, a pair of agents play strategies, and then the state transits in the environment, determining the game agents play at time $t+1$.

Note that an agent adopts the same action to play all the games with its opponents, which is decided at the beginning of a given time step. Using the same action along every edge is a typical assumption for learning on graphs (e.g., graphical polymatrix games (Cheung 2018)). Typical real-world examples include information sharing on social media, the movement of an arbitrary drone in an unmanned aerial vehicle swarm, and the uncontrolled intersections and lane change problems for self-driving vehicles. In these scenarios, the state transition is an "edge" property, and once an agent plays an action, this action will immediately affect all its opponents.

As the agent has to take the same action to play multiple different games rather than respond specifically to each type of game, in this paper, we assume the agents are myopic Q-learners, and do not have knowledge about game transition. The immediate reward is the only signal they can get. Therefore, each myopic agent maintains a vector of Q-values for each action, and the action sets for all states are considered identical. At time t, for any agent i in the population, if i takes action $a_j \in \mathcal{A} = \{a_1, \ldots, a_m\}$ from m available actions to interact with all other agents, and receives a reward $r_t^i(a_j)$ averaged over the $n-1$ interactions, then the j-th element in its Q-value vector $\mathbf{Q}_t^i = [Q_t^i(a_1), \ldots, Q_t^i(a_m)]^\top$ is updated as follows, while other elements remain unchanged.

$$Q_{t+1}^i(a_j) = Q_t^i(a_j) + \alpha[r_t^i(a_j) - Q_t^i(a_j)], \qquad (1)$$

where α is the learning rate. We consider the probability of action selection for each agent is generated by the Boltzmann exploration scheme. As a result, for any agent i, it has a mixed-strategy $\mathbf{x}_t^i = [x_t^i(a_1), \ldots, x_t^i(a_m)]^\top$, where $\forall a_j \in \mathcal{A}$, $x_t^i(a_j)$ is the probability that agent i takes action a_j at time t. The value of $x_t^i(a_j)$ is given by:

$$x_t^i(a_j) = \frac{e^{\tau Q_t^i(a_j)}}{\sum_{\forall a \in \mathcal{A}} e^{\tau Q_t^i(a)}}, \qquad (2)$$

where τ is the Boltzmann exploration temperature. The value of τ determines the trade-off between exploration and exploitation. If $\tau = 0$, agents will take actions randomly which means complete exploration. If $\tau \to \infty$, agents choose the action corresponding to the maximum Q-value.

The Dynamics Model of Multi-Agent Stochastic Games

In this section, we present the theoretical model of Q-learning dynamics in multi-agent stochastic games. First, we focus on an individual agent and model the dynamics of its Q-value vector. Next, we compare the commonly used mean-field approach with our pair-approximation method. Finally, we aim to accurately capture the population dynamics by modeling the evolution of the distribution of pairs.

Dynamics of Q-values for an Agent

When playing a pairwise stochastic game, the reward an agent can receive from playing against another agent depends not only on their joint action but also on their state. Consequently, the reward of a row agent, taking action a_i against its opponent with action a_j in state s, is calculated as:

$$r(a_i \mid s, a_j) = \mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j, \qquad (3)$$

where $\mathbf{M}_s$ is the payoff matrix of game played in state s, $\mathbf{e}_i$ is the unit column vector with size m (the i-th element equals 1 and the other $m-1$ elements equal 0).

Under our considered scenario, at time t, for an arbitrary agent i, given the actions of all its opponents and the corresponding games with each opponent, its immediate reward of taking action a_j is given by:

$$\begin{aligned} & r_t^i(a_j \mid \{s_t^{ih}\}_{h \in \{1,\ldots,n-1\}}, \{a_{vh}\}_{h \in \{1,\ldots,n-1\}}) \\ &= \frac{1}{n-1} \sum_{h=1}^{n-1} \mathbf{e}_j^\top \mathbf{M}_{s_t^{ih}} \mathbf{e}_{vh}, \end{aligned} \qquad (4)$$

where $\mathbf{M}_{s_t^{ih}}$ is the payoff matrix in state s_t^{ih}, s_t^{ih} is the state where agent i and its h-th opponent stay at time t, and agent i's h-th opponent takes the vh-th action a_{vh}.

Agents start to update their strategies after receiving immediate rewards. For Q-learners, only the Q-value of the taken action can be reinforced, so if agent i takes the j-th action a_j, only the j-th element of its Q-value vector will be updated according to Equation (1), while other elements remain unchanged. Consequently, for an individual agent i

who chooses the j-th action a_j at time t, the velocity of change of the j-th element of its Q-value vector is given as:

$$\begin{aligned} v_j(\mathbf{Q}_t^i, a_j) :=& Q_{t+1}^i(a_j) - Q_t^i(a_j) \\ =& \alpha[\mathbb{E}[r_t^i(a_j)] - Q_t^i(a_j)], \end{aligned} \tag{5}$$

where $\mathbb{E}[r_t^i(a_j)]$ is i's expected reward by taking action a_j. According to Equation (4), this expected reward depends on i's state distribution and opponent strategies in each state. We use a vector $\mathbf{v}(\mathbf{Q}_t^i, a_j)$ to denote the velocity of change of i's Q-value vector when it takes action a_j.

Mean-Field Approximation vs Pair-Approximation

Hu et al. (2019) leverage a mean field theoretic approach to approximate the effects of other agents on a single agent as the number of opponents goes to infinity. For an arbitrary agent i, the payoff that it receives from playing repeated games with all other agents is approximated by the payoff of playing against the mean strategy $\bar{\mathbf{x}}_t$:

$$\mathbb{E}[r_t^i(a_j)] = \mathbf{e}_j^\top \mathbf{M} \bar{\mathbf{x}}_t. \tag{6}$$

While this approach has yielded significant insights into learning in repeated normal form games under the infinite-agent setting, it is incompatible with stochastic games.

In stochastic games, the strategic interactions and environmental changes jointly drive the evolution of agent behaviors. Thus, how a focal agent is influenced by all of its opponents also depends on the relationship between the focal agent and each of its opponents. However, such state distributions and even the opponent strategies in each state for different agents can be heterogeneous. Specifically, different actions of a focal agent will lead to different distributions of joint actions in all its interactions, thus resulting in different state distributions after the state transition. The existence of heterogeneity also explains why agents taking the same action may get different rewards under our considered scenario. To summarize, environmental variability leads to the effects on different agents being heterogeneous, this is contrary to the assumption of the mean-field approximation that the effects on different agents are homogeneous.

Describing the interplay between agent behaviors and the environment is the key point in modelling learning dynamics in stochastic games. Under our setting, all interactions in the population are conducted in a pairwise way. If the evolution of each pair of interactions can be described, the dynamics of the whole population can also be captured.

In EGT, the pair-approximation method from statistical physics is used to obtain the spatial dynamics. This method tracks the frequencies of strategy pairs rather than only considering the frequency of strategies. Inspired by this method, if we focus on a pair of agents and their state, the influence of the joint behavior on the state and the influence of the state on the two agents can be characterized. Here, we propose a pair-approximation method by defining the concept of *pair* as a tuple consisting of the Q-value vectors of two connected agents $[\mathbf{Q}_t^1, \mathbf{Q}_t^2]$ and their state s. Thus, a pair in the system can be denoted by $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 \rangle$, where the superscripts 1 and 2 represent the focal agent and its opponent, respectively. In fact, for Q-learners, the heterogeneity of their Q-values also implies the heterogeneity of their state distributions and that of the opponent strategies in each state. By this pair-approximation method, we can get other necessary information about the agent based on its Q-values, thus the expected reward in Equation (5) actually depends on the agent's Q-values.

Theoretical Analysis of Q-Learning Dynamics in a Multi-Agent System with Game Transition

After giving the definition of pair, the system state can be defined as the probability distribution of pairs, and $p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t)$ is the proportion of pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 \rangle$ in the population at time t. As the interactions between agents proceed, the probability distribution of the pairs will evolve. By working with the temporal evolution of the system state, we can predict the learning dynamics of agents and describe how the environmental state (i.e., the distribution of states in the population) evolves. Next, we focus on how to track the evolution of $p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t)$.

Based on the Bayes rule, we have:

$$\begin{aligned} &\frac{\partial p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right)}{\partial t} = \frac{\partial [p\left(s, t\right) p\left(\mathbf{Q}_t^1, \mathbf{Q}_t^2, t \mid s\right)]}{\partial t} \\ &= p\left(s, t\right) \frac{\partial p\left(\mathbf{Q}_t^1, \mathbf{Q}_t^2, t \mid s\right)}{\partial t} + p\left(\mathbf{Q}_t^1, \mathbf{Q}_t^2, t \mid s\right) \frac{\mathrm{d} p\left(s, t\right)}{\mathrm{d} t}. \end{aligned} \tag{7}$$

The change of a single pair involves the change of two Q-value vectors and the transition of state. As Equation (7) expresses, in order to track the evolution of the system state, we have to describe how the conditional probability distribution of Q-value vector pairs in each state space evolves, and describe the evolution of the environmental state.

At first, we focus on the evolution of the environmental state of the system. The proportion of any state s in the population at time t is denoted by $p(s, t)$, and it is obtained by:

$$p(s, t) = \int \ldots \int p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t) A1A2, \tag{8}$$

where we define $dQ_t^1(a_1) \ldots dQ_t^1(a_m)$ (resp.$dQ_t^2(a_1) \ldots dQ_t^2(a_m)$) as $A1$ (resp.$A2$).

As the changes of states in each pair cumulatively result in a change in the whole system, at time $t+1$, we have:

$$\begin{aligned} & p(s, t+1) \\ &= \sum_{\forall s' \in \mathcal{S}} \int \ldots \int p(\mathbf{Q}_t^1, s', \mathbf{Q}_t^2, t) Pr(s \mid \mathbf{Q}_t^1, s', \mathbf{Q}_t^2) A1A2 \\ &= \sum_{\forall s' \in \mathcal{S}} \int \ldots \int p(\mathbf{Q}_t^1, s', \mathbf{Q}_t^2, t) \sum_{\forall a_i \in \mathcal{A}} \sum_{\forall a_j \in \mathcal{A}} x_i(\mathbf{Q}_t^1) \\ &\quad \times x_j(\mathbf{Q}_t^2) z_s(s', a_i, a_j) A1A2, \end{aligned} \tag{9}$$

where $Pr(s \mid \mathbf{Q}_t^1, s', \mathbf{Q}_t^2)$ is the probability of transition from state s' to s for the pair $\langle \mathbf{Q}_t^1, s', \mathbf{Q}_t^2 \rangle$. $x_i(\mathbf{Q}_t^1)$ is the probability that the focal agent selects action a_i when its Q-value vector is $\mathbf{Q}_t^1$, $x_j(\mathbf{Q}_t^2)$ is the probability that the focal agent's opponent selects action a_j when its Q-value vector is $\mathbf{Q}_t^2$, $x_i(\mathbf{Q}_t^1)$ and $x_j(\mathbf{Q}_t^2)$ can be obtained by Equation (2).

We derive a continuous-time differential equation for the evolution of $p(s,t)$ according to the method in (Tuyls, Verbeeck, and Lenaerts 2003):

$$\begin{aligned}&\frac{\mathrm{d}p\left(s,t\right)}{\mathrm{d}t}\\ &=\sum_{\forall s'\in\mathcal{S}}\int\cdots\int p(\mathbf{Q}_t^1,s',\mathbf{Q}_t^2,t)\sum_{\forall a_i\in\mathcal{A}}\sum_{\forall a_j\in\mathcal{A}}x_i(\mathbf{Q}_t^1)\\ &\quad\times x_j(\mathbf{Q}_t^2)z_s(s',a_i,a_j)A1A2\\ &\quad-\int\cdots\int p(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t)A1A2.\end{aligned}\tag{10}$$

Then, we focus on deriving how the conditional probability distribution of Q-value vector pairs in a given state space s evolves. Due to the existence of the state transition mechanism, deriving the equation that can describe the evolution of this conditional distribution is complicated. Nevertheless, under our considered scenario, after the decision-making of agents, the strategy updating and the state transition are two independent processes. The state transitions caused by the current interactions only affect the learning of agents at the next time step. Therefore, we first assume that interactions between agents at time t do not lead to any state transitions. In this way, we can derive the velocity of change of the density of agent pairs having their Q-values equal $\mathbf{P}=[\mathbf{Q}_t^1,\mathbf{Q}_t^2]$ in state space s, under the condition that the state does not transit at time t (i.e., $\frac{\partial p^s(\mathbf{P},t)}{\partial t}$). Then we consider how the change of state leads to the transition of agent pairs among different state spaces, and further deduce the state distributions of different Q-value vector pairs. Finally, we can derive the equation that can capture the evolution of $p(\mathbf{P},t\mid s)$.

We follow the method in (Wang et al. 2022) to derive the velocity of change of $p(\mathbf{P},t\mid s)$ at time t based on the assumption that there is no state transition. As we use $\frac{\partial p(\mathbf{P},t|s)}{\partial t}$ and $\frac{\partial p^s(\mathbf{P},t)}{\partial t}$ to represent the velocity of change of $p(\mathbf{P},t\mid s)$ at time t with state transition and without transition, respectively, we rewrite $p(\mathbf{P},t\mid s)$ as $p^s(\mathbf{P},t)$ here. From a spatial perspective, the state space s is a $2m$-dimensional euclidean space, where m is the size of the action set, and the agent pairs in state s occupy a position in this space according to their Q-values. The change of the density of agent pairs at any position $\mathbf{P}$ in the space from t to $t+\Delta t$ is caused by the process that agent pairs leave the position $\mathbf{P}$, and incoming agent pairs reach $\mathbf{P}$ from other positions. We denote all the positions where agent pairs may exchange with $\mathbf{P}$ as $\{\mathbf{P}'\}$. During the time interval Δt, all agents can only update their Q-value vectors once, the evolution of $p^s(\mathbf{P},t)$ can be represented by the master equation:

$$\begin{aligned}&\frac{\partial p^s\left(\mathbf{P},t\right)}{\partial t}=\lim_{\Delta t\to 0}\frac{1}{\Delta t}(p^s(\mathbf{P},t+\Delta t)-p^s(\mathbf{P},t))\\ &=\int T(\mathbf{P},\mathbf{P}',t\mid s)p^s(\mathbf{P}',t)-T(\mathbf{P}',\mathbf{P},t\mid s)p^s(\mathbf{P},t)d\mathbf{P}',\end{aligned}\tag{11}$$

where $T(\mathbf{P},\mathbf{P}',t\mid s)$ is the transition rate from position $\mathbf{P}'$ to $\mathbf{P}$ in the state space s at time t, and likewise for $T(\mathbf{P}',\mathbf{P},t\mid s)$. Note that the position set $\{\mathbf{P}'\}$ is the same for agent pairs in different state spaces but with the same Q-values $\mathbf{P}$. Because each agent interacts with an infinite number of opponents, and according to Equation (5), the change in an agent's Q-values depends on its action and Q-values. That is, for a pair of agents, the change in their Q-values is independent of their state. Specifically, when the focal agent takes action a_i and its opponent takes action a_j at time t, if we denote the change of Q-values $\mathbf{P}$ of this pair of agents as $\mathbf{v}(\mathbf{P},a_i,a_j,t)$, $\forall s\in\mathcal{S}$, we have $\mathbf{v}(\mathbf{P},a_i,a_j,t\mid s)=\mathbf{v}(\mathbf{P},a_i,a_j,t)$. Therefore, $\forall s\in\mathcal{S}$, we have $T(\mathbf{P},\mathbf{P}',t\mid s)=T(\mathbf{P},\mathbf{P}',t)$, and $T(\mathbf{P}',\mathbf{P},t\mid s)=T(\mathbf{P}',\mathbf{P},t)$.

Then, by deriving $T(\mathbf{P},\mathbf{P}',t)$ and $T(\mathbf{P}',\mathbf{P},t)$ to further deduce the master equation, we have:

$$\begin{aligned}&\frac{\partial p^s\left(\mathbf{P},t\right)}{\partial t}\\ &=-\sum_{\forall a_i\in\mathcal{A}}v_i(\mathbf{Q}_t^1,a_i)\frac{\partial[p\left(\mathbf{P},t\mid s\right)x_i(\mathbf{Q}_t^1)]}{\partial Q_t^1(a_i)}\\ &\quad-\sum_{\forall a_j\in\mathcal{A}}v_j(\mathbf{Q}_t^2,a_j)\frac{\partial[p\left(\mathbf{P},t\mid s\right)x_j(\mathbf{Q}_t^2)]}{\partial Q_t^2(a_j)}.\end{aligned}\tag{12}$$

More details about the derivation of Equation (12) are presented in our supplementary material[1].

Finally, we track the evolution of $p(\mathbf{P},t\mid s)$ by considering the occurrence of state transition on the basis of obtaining the evolution of Q-value distribution in the population. Without state transition, in state space s, the change of the proportion of agent pairs having their Q-values equal $\mathbf{P}$ from time t to $t+\Delta t$ is $\frac{\partial p^s(\mathbf{P},t)}{\partial t}\Delta t$, but due to the transition of state, only a certain proportion of agent pairs will stay in state space s, while other agent pairs will move into other different state spaces. Thus, we have:

$$\begin{aligned}&\frac{\partial p\left(\mathbf{P},t\mid s\right)}{\partial t}=\lim_{\Delta t\to 0}\frac{1}{\Delta t}(p(\mathbf{P},t+\Delta t\mid s)-p(\mathbf{P},t\mid s))\\ &=\frac{1}{p(s,t)}(\sum_{\forall s'\in\mathcal{S}}p(s',t)\frac{\partial p^{s'}\left(\mathbf{P},t\right)}{\partial t}\sum_{\forall a_i\in\mathcal{A}}\sum_{\forall a_j\in\mathcal{A}}x_i(\mathbf{Q}_t^1)\\ &\quad\times x_j(\mathbf{Q}_t^2)z_s(s',a_i,a_j)-p(\mathbf{P},t\mid s)\frac{\mathrm{d}p\left(s,t\right)}{\mathrm{d}t}).\end{aligned}\tag{13}$$

Finally, the Q-learning dynamics of multi-agent stochastic games can be modelled accurately by the Equation (2), (5), (7), (8), (10), (12) and (13).

Based on the system state, we can get more information on agent behaviors. The proportion of agents having their Q-values equal $\mathbf{Q}_t^1$ is given below:

$$p(\mathbf{Q}_t^1,t)=\sum_{\forall s\in\mathcal{S}}\int\cdots\int p(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t)A2.\tag{14}$$

The proportion of agents having their Q-value of action a_i equal $Q_t^1(a_i)$ is given below:

$$\begin{aligned}&p(Q_t^1(a_i),t)=\sum_{\forall s\in\mathcal{S}}\int\cdots\int p(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t)\\ &dQ_t^1(a_1)\dots dQ_t^1(a_{i-1})dQ_t^1(a_{i+1})\dots dQ_t^1(a_m)A2.\end{aligned}\tag{15}$$

[1] https://github.com/Zheng-YZ/AAAI2023SM

The expected Q-value of action a_i can be obtained by:

$$\mathbb{E}[Q_t(a_i)] = \int Q_t^1(a_i) p(Q_t^1(a_i), t) dQ_t^1(a_i). \quad (16)$$

Then, the expected strategy for selecting action a_i is:

$$\mathbb{E}[x_t(a_i)] = \int \dots \int p(\mathbf{Q}_t^1, t) \frac{e^{\tau Q_t^1(a_i)}}{\sum_{\forall a \in \mathcal{A}} e^{\tau Q_t^1(a)}} A1. \quad (17)$$

Experiments

In this section, we conduct experiments to validate our theoretic model and further reveal the interesting phenomena caused by state transitions.

Different Initial Conditions

We consider two different initial conditions, homogeneous and heterogeneous. For the homogeneous case, the initial Q-values of all available actions are set to 0 for all agents, and each pair of agents is in state s_1 at the beginning. For the heterogeneous case, the initial Q-values of agents follow different Beta distributions, and the initial states between agents are determined at random.

Experiments are conducted on a two-state two-action stochastic game where agents play a SH game in state s_1 and play a PD game in state s_2. The action sets and payoff matrices of the SH game and PD game are given as follows:

$$\mathcal{A}(\text{SH}) = \mathcal{A}(\text{PD}) = \{\text{cooperate C}, \text{defect D}\},$$

$$\mathbf{M}_{\text{SH}} = \begin{pmatrix} r_{\text{CC}} & r_{\text{CD}} \\ r_{\text{DC}} & r_{\text{DD}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ r & r \end{pmatrix},$$

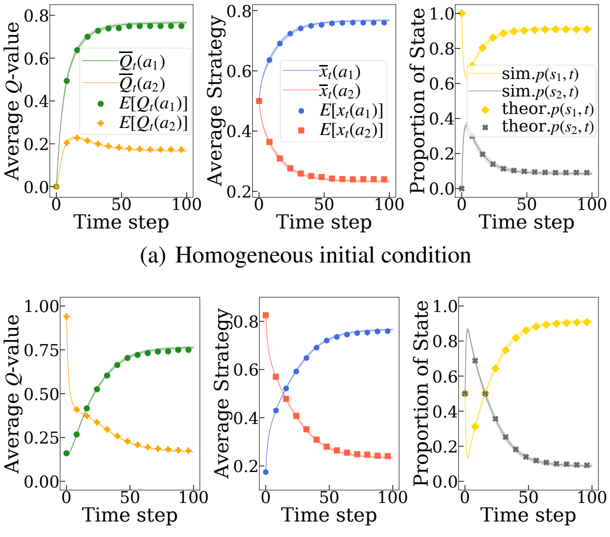
$$\mathbf{M}_{\text{PD}} = \begin{pmatrix} r_{\text{CC}} & r_{\text{CD}} \\ r_{\text{DC}} & r_{\text{DD}} \end{pmatrix} = \begin{pmatrix} 1 & -r \\ b & 0 \end{pmatrix}.$$

We set $r = 0.1$ for the SH game, and $b = 1.2$, $r = 0.1$ for the PD game. Transitions between the two states occur according to the following rule: for a pair of agents, if their current state is s_1, only their mutual defection can lead to a transition from s_1 to s_2, and if they stay in s_2 at present, only mutual cooperation can help them return to s_1.

For the agent-based simulations, we set the population size $n = 1000$, the learning rate $\alpha = 0.4$, and the temperature $\tau = 2$ (Unless otherwise specified, the parameters are set in the same way for subsequent experiments). We run 500 simulations for each setting to smooth out the randomness. As shown in Figure 1, the dots represent the results derived from our dynamics model, the solid lines represent the mean of the results of agent-based simulations, and the shaded areas represent the standard deviation of simulation results (The results of subsequent experiments are presented in the same manner). Under different initial conditions, our model always provides accurate descriptions of the evolution of agent behaviors and that of the environmental state.

Deterministic and Probabilistic Transitions

For the above experiments, the state transition between two paired agents is driven by their joint action and current state, this is a general form of transition. Now, we investigate another case where the state transition between agents depends only on their joint action, that is the state-independent transition. Furthermore, using this state-independent transition, we further validate the applicability of our model to the deterministic and probabilistic transition rules.



(a) Homogeneous initial condition

(b) Heterogeneous initial condition

Figure 1: Evolution of agent behaviors and that of the environmental state under different initial conditions. In (b), we set $Q_0(a_1) \sim$ Beta(20, 80, -0.1, 1.2), $Q_0(a_2) \sim$ Beta(80, 20, -0.1, 1.2), the first two parameters of the Beta distribution control the shape of the probability density function, and the latter two parameters prescribe the support to be $[r_{\min}, r_{\max}]$, where $r_{\min}$ and $r_{\max}$ are the minimum and maximum payoff of the stochastic game, respectively.

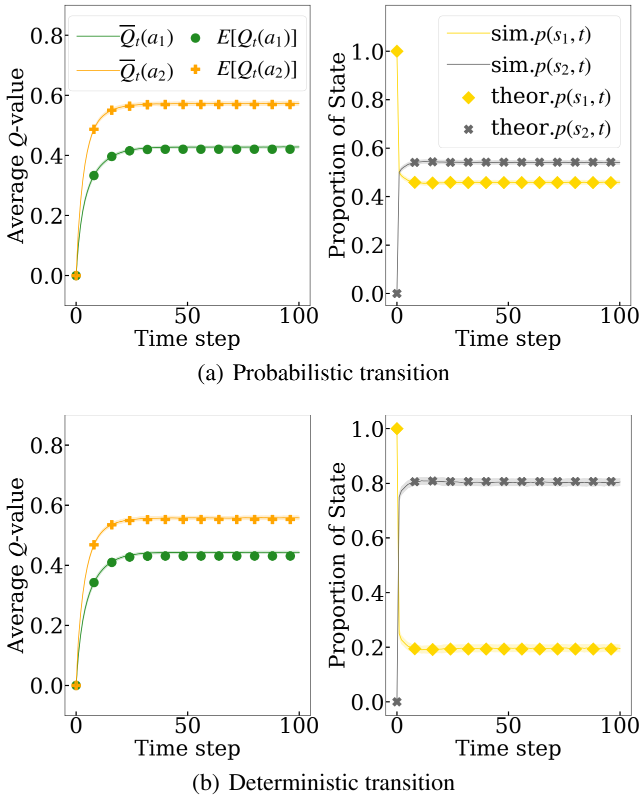
We consider another scenario where each pair of agents play a two-state PD game, and the two different states $s1$ and $s2$ correspond to two different PD games PD1 ($b = 1.5, r = 0$) and PD2 ($b = 1.2, r = 0$), respectively. The probabilistic transition rule is given by:

$$\mathbf{T}_{s_1 \to s_2} = \begin{pmatrix} 0.1 & 0.6 \\ 0.6 & 0.7 \end{pmatrix}, \mathbf{T}_{s_2 \to s_1} = \begin{pmatrix} 0.9 & 0.4 \\ 0.4 & 0.3 \end{pmatrix}.$$

The deterministic transition rule is given by:

$$\mathbf{T}_{s_1 \to s_2} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \mathbf{T}_{s_2 \to s_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We present the results in Figure 2, for different transition rules, the quantitative agreement between the results of our model and the simulation results is also notable. Moreover, for better illustration, the dynamics of the average strategies under the probabilistic (rule1) and deterministic (rule2) transitions are compared in Figure 3 (a). From Figure 3 (a) and Figure 2, it can be found that the transition probability can greatly affect the environmental state of the population. The deterministic transition can lead to a noticeable increase in the proportion of s_2 where agents have less temptation to defect, thus the cooperation level is higher than the case of probabilistic transition. Without loss of generality, we consider deterministic transition for subsequent experiments.



(a) Probabilistic transition

(b) Deterministic transition

Figure 2: Evolution of agent behaviors and that of the environmental state under different transition rules.

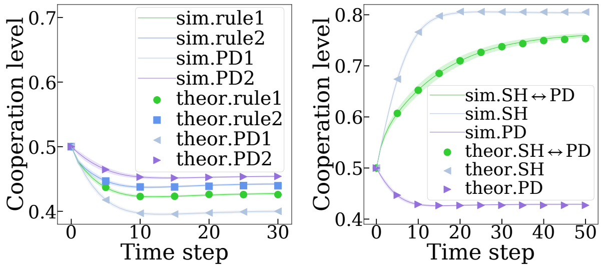
Stochastic Games vs Normal Form Games

In order to provide deeper insight into the agent behaviors under the ever-changing environment, we compare the results in Figure 1 (a) and Figure 2 with the case where agents play the repeated normal form games.

In Figure 3 (a), if the transition can occur between PD1 and PD2, regardless of whether the transition is probabilistic or deterministic, the probability that agents choose to be cooperators is higher than the case where agents play the normal form game PD1, but is lower than the case where agents play PD2 without transition. Similarly, for the stochastic game with transition between PD game and SH game, Figure 3 (b) shows that the transition mechanism significantly reinforces the positive behaviors of agents. More broadly, this suggests that game transitions, either naturally occurring or designed, help to resolve social dilemmas, such as climate change and public resource management in real life.

Application to Different Population Sizes and Learning Parameters

More importantly, to better illustrate the application of our approach, we conduct more experiments under the cases of varying population sizes and learning parameters. We experiment on the above two-state game with transition between SH and PD, some of the results are presented in Figure 4, while others can be found in our supplementary material.



(a) A two-state PD game

(b) Game transition between SH and PD

Figure 3: The effects of the introduction of game transition and the transition probability on agent behaviors.

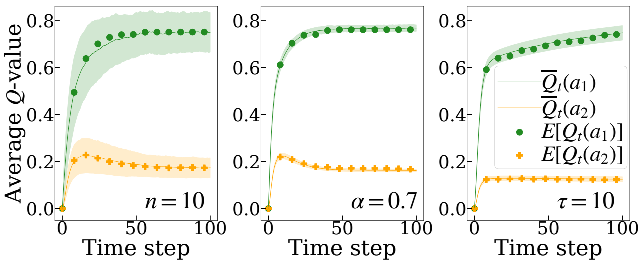


Figure 4: The performance of our method in predicting the dynamics of average Q-values under different settings.

Regarding the population size, although technically our approach requires an infinitely large population, empirically we observe that our theoretical predictions work well in small, finite populations. In Figure 4, we note that for small population sizes (e.g., $n = 10$), the result of a single simulation run can fluctuate significantly, yielding a substantial variance in the simulation results. But this is somewhat expectable, as the population is so small that the empirical distribution of Q-values in a single simulation run would inevitably deviate from the probability distribution predicted by our theory. We also varied the learning rate α and the exploration temperature τ, for these two cases, we set $n = 500$. Under these settings, our theoretical predictions always well agree with the simulation results.

We expand experiments on more complex scenarios including a two-state three-action game and a two-action three-state game. Additionally, we show our method works better than the mean-field approach through experiments. These results can be found in our supplementary material.

Conclusion

In this paper, we model the dynamics of multi-agent Q-learning in stochastic games. The proposed pair-approximation method accurately captures the influence of environmental variability on agents. The numerical experiments corroborate the descriptive power of our model and evidence the important role of state transitions in the emergence of cooperation from social dilemmas. In future work, we will extend our method to asymmetric state transitions, other graph structures, as well as other learning algorithms.

Acknowledgments

This research was supported by the National Science Fund for Distinguished Young Scholarship of China (No. 62025602), the National Key RD Program of China (No. 2018AAA0100900), the National Natural Science Foundation of China (Nos. U1803263, 11931015, 11971421, 61866039, and 62066045), Key Technology Research and Development Program of Science and Technology-Scientific and Technological Innovation Team of Shaanxi Province (Grant No. 2020TD-013), Excellent Youths Project for Basic Research of Yunnan Province(No. 202101AW070015) and the XPLORER PRIZE.

References

Bloembergen, D.; Tuyls, K.; Hennes, D.; and Kaisers, M. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53: 659–697.

Boone, V.; and Piliouras, G. 2019. From Darwin to Poincaré and von Neumann: Recurrence and cycles in evolutionary and algorithmic game theory. In *International Conference on Web and Internet Economics*, 85–99. Springer.

Cheung, Y. K. 2018. Multiplicative weights updates with constant step-size in graphical constant-sum games. *Advances in Neural Information Processing Systems*, 31.

Chu, C.; Li, Y.; Liu, J.; Hu, S.; Li, X.; and Wang, Z. 2022. A Formal Model for Multiagent Q-Learning Dynamics on Regular Graphs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 194–200.

Deng, C.; Rong, Z.; Wang, L.; and Wang, X. 2021. Modeling replicator dynamics in stochastic games using Markov chain method. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 420–428.

Ganapathi Subramanian, S.; Poupart, P.; Taylor, M. E.; and Hegde, N. 2020. Multi Type Mean Field Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 411–419.

Hauert, C.; and Szabó, G. 2005. Game theory and physics. *American Journal of Physics*, 73(5): 405–414.

Hennes, D.; Tuyls, K.; and Rauterberg, M. 2009. State-coupled replicator dynamics. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 789–796.

Hu, S.; Leung, C.-w.; and Leung, H.-f. 2019. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. *Advances in Neural Information Processing Systems*, 32.

Hu, S.; Leung, C.-W.; Leung, H.-f.; and Soh, H. 2022. The Dynamics of Q-learning in Population Games: A Physics-inspired Continuity Equation Model. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 615–623.

Leonardos, S.; Piliouras, G.; and Spendlove, K. 2021. Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality. *Advances in Neural Information Processing Systems*, 34.

Leung, C.-w.; Hu, S.; and Leung, H.-f. 2022. Modelling the Dynamics of Multi-Agent Q-learning: The Stochastic Effects of Local Interaction and Incomplete Information. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 384–390.

Long, Q.; Zhou, Z.; Gupta, A.; Fang, F.; Wu, Y.; and Wang, X. 2020. Evolutionary population curriculum for scaling multi-agent reinforcement learning. *arXiv preprint arXiv:2003.10423*.

Panozzo, F.; Gatti, N.; and Restelli, M. 2014. Evolutionary dynamics of Q-learning over the sequence form. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Tuyls, K.; and Parsons, S. 2007. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence*, 171(7): 406–416.

Tuyls, K.; Verbeeck, K.; and Lenaerts, T. 2003. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, 693–700.

Wang, Z.; Mu, C.; Hu, S.; Chu, C.; and Li, X. 2022. Modelling the Dynamics of Regret Minimization in Large Agent Populations: a Master Equation Approach. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 534–540.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8(3): 279–292.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean Field Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*, 5567–5576.