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

Supplementary Material

Details about the Derivation

The definition of $T(\mathbf{P}, \mathbf{P}', t)$ in the master equation (the Equation(11) in the main paper) is given as follows:

$$\begin{aligned} & T(\mathbf{P}, \mathbf{P}', t) \\ & = \lim_{\Delta t \to 0} \frac{Pr(\mathbf{P}, t + \Delta t \mid \mathbf{P}', t) - Pr(\mathbf{P}, t \mid \mathbf{P}', t)}{\Delta t}, \end{aligned} \tag{S.1}$$

where $Pr(\mathbf{P}, t + \Delta t \mid \mathbf{P}', t)$ is the probability that a pair of agents in $\mathbf{P}'$ at time t moves to $\mathbf{P}$ in Δt , $Pr(\mathbf{P}, t \mid \mathbf{P}', t) = 0$ always holds according to the definition of $\mathbf{P}'$.

Consider that an agent pair in one position can move to different positions because they can have various joint actions, we can further distinguish the different positions $\{\mathbf{P}'\}$ from which two connected agents can take different actions to reach $\mathbf{P}$. We define the position set $\{\mathbf{P}'\}$ as $\{\mathbf{P}'_{a_1,a_1}\} \cup \{\mathbf{P}'_{a_1,a_2}\} \cup \cdots \cup \{\mathbf{P}'_{a_m,a_m}\}$, where $\mathbf{P}'_{a_i,a_j}$ is the starting position from which a focal agent taking action a_i and its opponent taking action a_j can get to $\mathbf{P}$ in Δt. Therefore, the first term of the right hand side of the master equation in the main paper can be rewritten as:

$$\begin{aligned} & \int T(\mathbf{P}, \mathbf{P}', t) p^s(\mathbf{P}', t) d\mathbf{P}' \\ & = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{\forall a_i \in \mathcal{A}} \sum_{\forall a_j \in \mathcal{A}} \int Pr(\mathbf{P}, t + \Delta t \mid \mathbf{P}'_{a_i,a_j}, t) \\ & \quad \times p^s(\mathbf{P}'_{a_i,a_j}, t) d\mathbf{P}'_{a_i,a_j}, \end{aligned} \tag{S.2}$$

where $Pr(\mathbf{P}, t+\Delta t \mid \mathbf{P}'_{a_i,a_j}, t) = x_i(\mathbf{Q}^{1'}_{t,a_i}) x_j(\mathbf{Q}^{2'}_{t,a_j})$, this is because an agent pair in position $\mathbf{P}'_{a_i,a_j} = [\mathbf{Q}^{1'}_{t,a_i}, \mathbf{Q}^{2'}_{t,a_j}]$ at time t can reach $\mathbf{P}$ at $t + \Delta t$ if and only if the focal agent chooses a_i and its opponent chooses a_j.

Note that $\mathbf{v}(\mathbf{P}'_{a_i,a_j}, a_i, a_j, t)$ can be approximated as $\mathbf{v}(\mathbf{P}, a_i, a_j, t)$, as $\Delta t \to 0$. Therefore, we have $\mathbf{P}'_{a_i,a_j} = \mathbf{P} - \mathbf{v}(\mathbf{P}, a_i, a_j, t)\Delta t$, and the integral symbol in Equation (S.2) can be removed:

$$\begin{aligned} & \int T(\mathbf{P}, \mathbf{P}', t) p^s(\mathbf{P}', t) d\mathbf{P}' \\ & = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{\forall a_i \in \mathcal{A}} \sum_{\forall a_j \in \mathcal{A}} x_i(\mathbf{Q}^1_t - \mathbf{v}(\mathbf{Q}^1_t, a_i)\Delta t) \\ & \quad \times x_j(\mathbf{Q}^2_t - \mathbf{v}(\mathbf{Q}^2_t, a_j)\Delta t) p^s(\mathbf{P} - \mathbf{v}(\mathbf{P}, a_i, a_j, t)\Delta t, t). \end{aligned} \tag{S.3}$$

In a similar way, we can rewrite the second term of the right hand side of the master equation in the main paper as:

$$\begin{aligned} & \int T(\mathbf{P}', \mathbf{P}, t) p^s(\mathbf{P}, t) d\mathbf{P}' \\ & = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{\forall a_i \in \mathcal{A}} \sum_{\forall a_j \in \mathcal{A}} x_i(\mathbf{Q}^1_t) x_j(\mathbf{Q}^2_t) p^s(\mathbf{P}, t). \end{aligned} \tag{S.4}$$

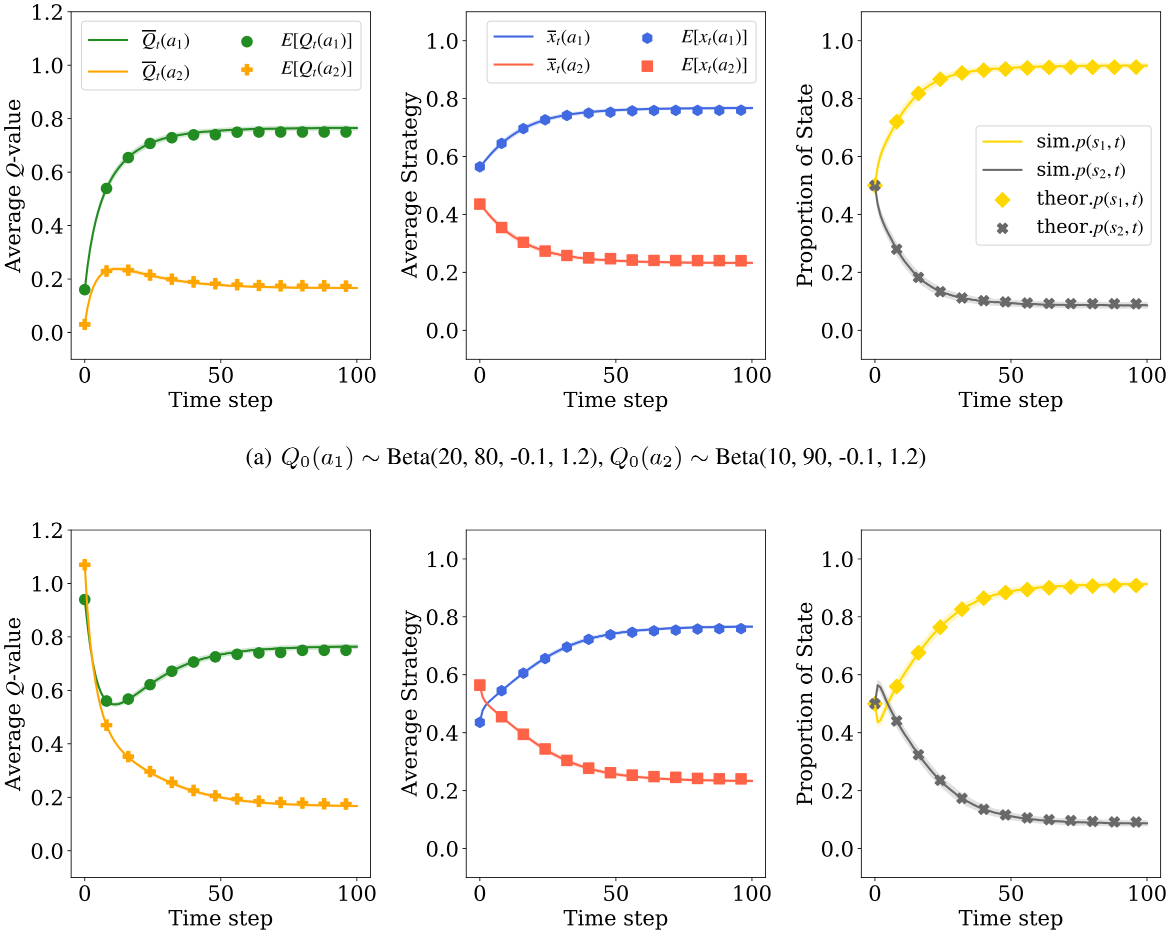
Substituting the first and second terms of the right hand side of the master equation in the main paper with Equation (S.3) and Equation (S.4), respectively, and based on the Taylor expansion at $\mathbf{P}$, we finally obtain the Equation (12) in the main paper.

Heterogeneous Initial Q-value Distribution

We consider two more cases where the initial Q-values of all actions follow different Beta distributions. In Figure S1, it can be observed that our theoretical predictions well agree with the simulation results under these different initial conditions.

Different Population Sizes

Although our theoretical derivation requires an infinite population, the population size is not a limiting factor in the application of our dynamics model. We conduct simulations under different settings of population size. As shown in Figure S2, our approach requires a sufficiently large population (e.g., more than 100 agents) to have an accurate prediction on every single simulation run. However, in order to agree with simulation results averaged over many simulation runs, our approach works even with very small population sizes (e.g., 10 agents).



(a) $Q_0(a_1) \sim$ Beta(20, 80, -0.1, 1.2), $Q_0(a_2) \sim$ Beta(10, 90, -0.1, 1.2)

(b) $Q_0(a_1) \sim$ Beta(80, 20, -0.1, 1.2), $Q_0(a_2) \sim$ Beta(90, 10, -0.1, 1.2)

Figure S1: Evolution of agent behaviors and that of the environmental state of the population under different initial Q-value distributions. We consider the two-state stochastic game with transition between a SH game and a PD game (see our main paper), and set $\alpha = 0.4$, $\tau = 2$, $n = 1000$. Unless otherwise specified, we use the same setting of the stochastic game and parameters for subsequent experiments.

Effects of Learning Parameters on the Applicability of Our Model

The learning rate α and Boltzmann exploration temperature τ have an impact on the learning process of agents. Here, we validate the applicability of our formal model to different settings of algorithm parameters and analyze the effects of these parameters on the population dynamics and the evolution of the environmental state.

In Figure S3, we further explore the effects of small and large learning rates on the evolutionary dynamics. The predictions of our theoretical model are consistent with the simulation results under different learning rates. We also find that the learning rate affects the convergence rate of the learning process of agents, and the convergence rate is faster under a larger learning rate.

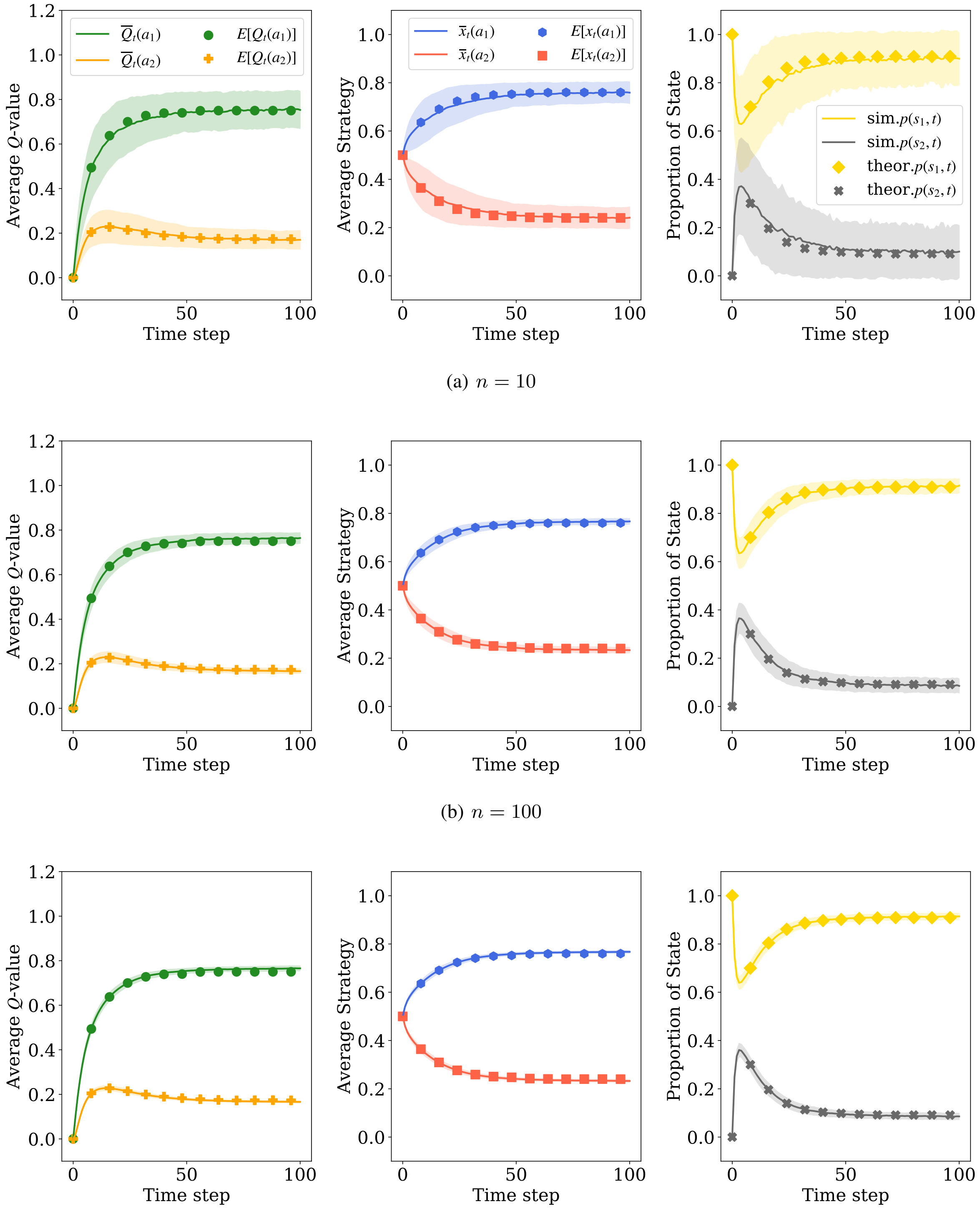
In Figure S4, our model still accurately describes the evolution of agent behaviors and that of the environmental state under different exploration temperatures. In addition, we can find that the variance of the results of multiple simulation runs increases with the increase of τ.

Two-state Rock Paper Scissors Game

The stochastic games between pairs of agents we have considered so far involve only two actions and two states. Next, we carry out simulations to verify the applicability of our formal model under the settings of larger action sets and state sets.

We conduct the experiment on a 2-state rock paper scissors (RPS) game to validate the effectiveness of our model under a two-state three-action stochastic game setting. The payoff matrix of the RPS game is given as follows:

$$\mathcal{A}(\text{RPS}) = \{\text{rock R}, \text{paper P}, \text{scissors S}\},$$

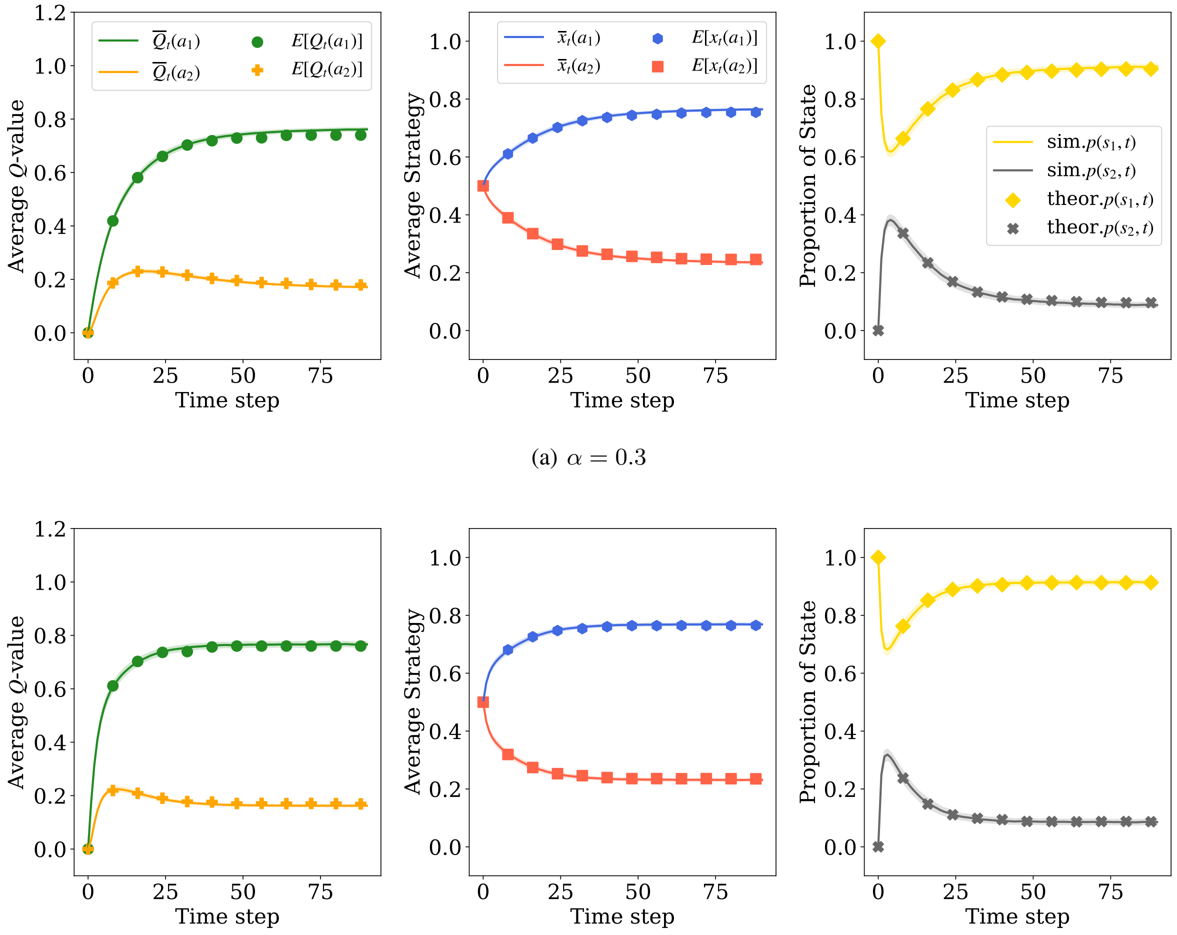


(a) $n = 10$

(b) $n = 100$

(c) $n = 500$

Figure S2: Evolution of agent behaviors and that of the environmental state of the population under different population sizes.



(a) $\alpha = 0.3$

(b) $\alpha = 0.7$

Figure S3: Evolution of agent behaviors and that of the environmental state of the population under different learning rates.

$$\mathbf{M}_{\text{RPS}} = \begin{pmatrix} r_{\text{RR}} & r_{\text{RP}} & r_{\text{RS}} \\ r_{\text{PR}} & r_{\text{PP}} & r_{\text{PS}} \\ r_{\text{SR}} & r_{\text{SP}} & r_{\text{SS}} \end{pmatrix} = \begin{pmatrix} 0 & -u & u \\ u & 0 & -u \\ -u & u & 0 \end{pmatrix},$$

where the parameter u should satisfy $u > 0$. For RPS game, $(\frac{1}{3}R, \frac{1}{3}P, \frac{1}{3}S)$ is the mixed strategy Nash equilibrium.

We use a RPS game with $u = 1$ and another RPS game with $u = 0.5$ to describe the interaction in state s_1 and s_2, respectively. We formulate the state transition rule as follows: if a pair of agents take the same actions (i.e.,(R, R), (P, P) and (S, S)), their state won't transit, but if the two agents take different actions, their state will change to another different state.

As shown in Figure S5, although agents can either face a state with high reward and high risk or a state with low reward and small loss, agents still learn to play the 2-state RPS game by making a choice randomly.

A Three-state Stochastic Game

For the experiment on a 3-state stochastic game, we consider that agents play a SH game in state s_1, play a Hawk Dove (HD) game in state s_2, and play a PD game in state s_3. The payoff matrices of the SH game and the PD game are given in our main paper, the payoff matrix of the HD game is given by:

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

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

The parameter b is set to 1.2 and r is set to 0.1 for the three different games. The mechanism of state transition is introduced by the following transition matrices:

$$\mathbf{T}_{s_1 \to s_2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{T}_{s_1 \to s_3} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\mathbf{T}_{s_2 \to s_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \mathbf{T}_{s_2 \to s_3} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{T}_{s_3 \to s_1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \mathbf{T}_{s_3 \to s_2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

When extending the stochastic game to a three-state setting, the interaction between agents becomes more complicated, the results in Figure S6 demonstrate that our model can also perform well in capturing the evolution of agent behaviors and that of the environmental state.

Our Model versus Mean-field Approximation

In Figure S7, we compare the descriptive power of the mean-field approximation (MFA) and our pair-approximation method.

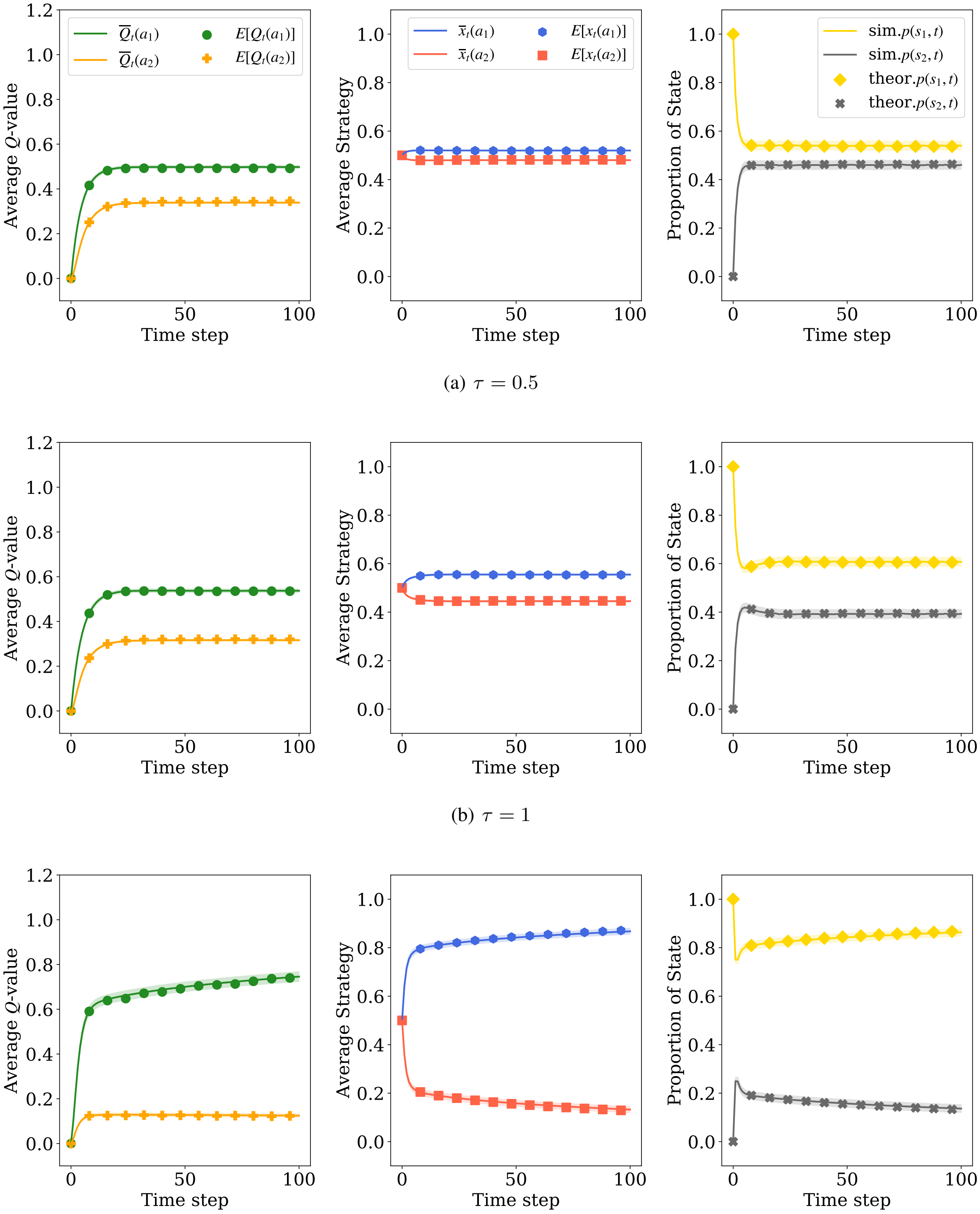
In this experiment, we consider a two-state two-action stochastic game. The state s_1 corresponds to a donation game (DG). The DG is a special form of the prisoner's dilemma game in which cooperation means offering the other agent a benefit b at a personal cost c, while defectors offer nothing. The payoff matrix of the donation game is given by:

$$\mathcal{A}(\mathrm{DG}) = \{\text{cooperate C}, \text{defect D}\},$$

$$\mathbf{M}_{\mathrm{DG}} = \begin{pmatrix} r_{\mathrm{CC}} & r_{\mathrm{CD}} \\ r_{\mathrm{DC}} & r_{\mathrm{DD}} \end{pmatrix} = \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix}.$$

In our experiment, the benefit b is set to 2 and the cost c is set to 0.5. The state s_2 is represented by a game in which all agents can only receive a negative reward $m < 0$, regardless of their actions. We set $m = -1$. The state transition rule is as follows: when agents are in s_1, only mutual cooperation can avoid a transition from s_1 to s_2, and only mutual cooperation can help agents in s_2 return to s_1.

We take the results of agent-based simulations as the benchmark, our model can accurately predict the learning dynamics and capture the evolution of the environmental state of the population, while the predictions of the mean-field approach deviate from the actual dynamics. More importantly, in this experiment, we observe that even though the two norm form games alone can not support the emergence of cooperative behavior (in the DG, agents learn to defect, while agents learn to choose cooperation and defection with equal probability in another game), the transition between the two games can significantly promote the evolution of cooperation. Therefore, the myopic reinforcement learning agents can also learn to cooperate in an ever-changing environment.



(a) $\tau = 0.5$

(b) $\tau = 1$

(c) $\tau = 10$

Figure S4: Evolution of agent behaviors and that of the environmental state of the population under different Boltzmann exploration temperatures.

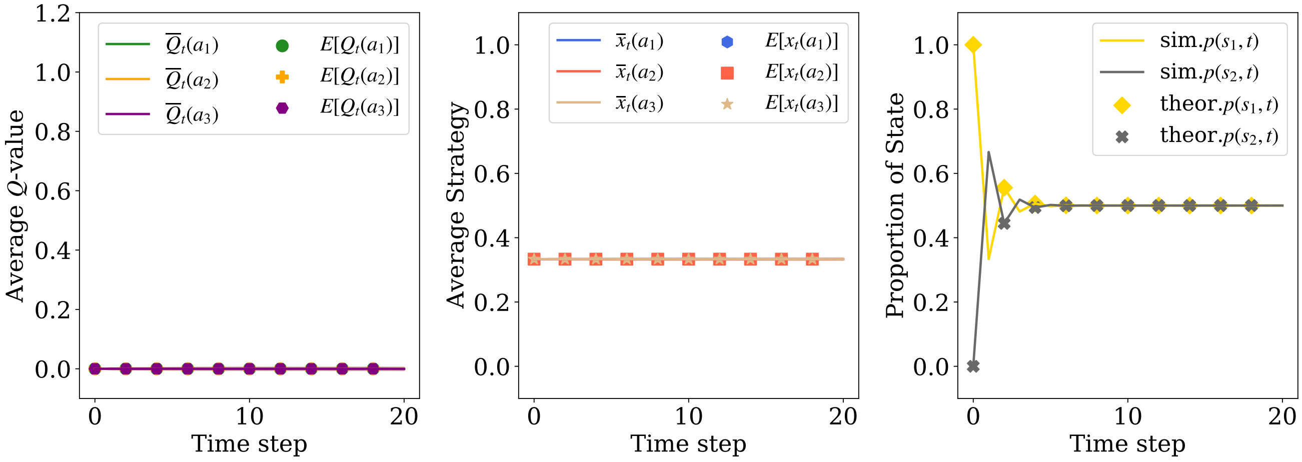


Figure S5: Evolution of agent behaviors and that of the environmental state of the population under a two-state three-action stochastic game setting

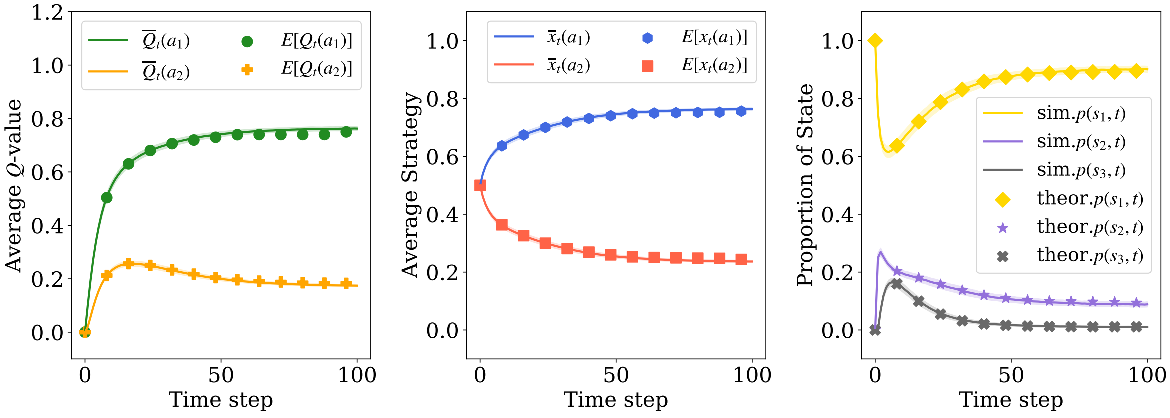


Figure S6: Evolution of agent behaviors and that of the environmental state of the population under a three-state two-action stochastic game setting.

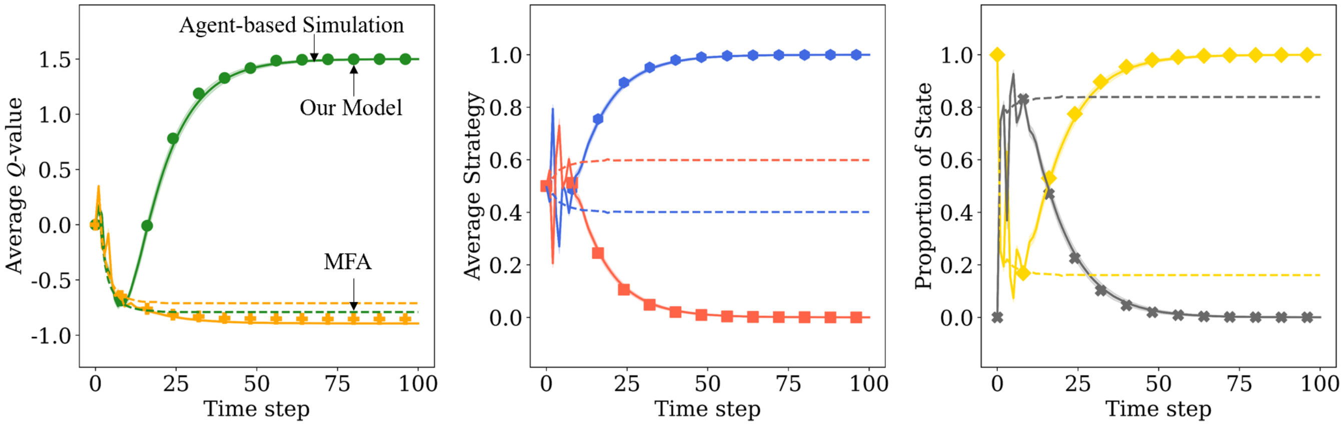


Figure S7: Comparison among the population dynamics predicted by our model (dots), mean-field approximation (dashed line), and the agent-based simulations (solid line).We set $\alpha = 0.7$ and $\tau = 5$.