# Modelling the Dynamics of Multi-Agent Q-learning: The Stochastic Effects of Local Interaction and Incomplete Information

**Chin-wing Leung**[1] , **Shuyue Hu**[2*] and **Ho-fung Leung**[1]
[1]The Chinese University of Hong Kong
[2] Shanghai Artificial Intelligence Laboratory
cwleung@cse.cuhk.edu.hk, shuyuehu217@gmail.com, lhf@cuhk.edu.hk

## Abstract

The theoretical underpinnings of multiagent reinforcement learning has recently attracted much attention. In this work, we focus on the generalized social learning (GSL) protocol — an agent interaction protocol that is widely adopted in the literature, and aim to develop an accurate theoretical model for the Q-learning dynamics under this protocol. Noting that previous models fail to characterize the effects of local interactions and incomplete information that arise from GSL, we model the Q-values dynamics of each individual agent as a system of stochastic differential equations (SDE). Based on the SDE, we express the time evolution of the probability density function of Q-values in the population with a Fokker-Planck equation. We validate the correctness of our model through extensive comparisons with agent-based simulation results across different types of symmetric games. In addition, we show that as the interactions between agents are more limited and information is less complete, the population can converge to an outcome that is qualitatively different than that with global interactions and complete information.

## 1 Introduction

A multiagent system (MAS) concerns multiple intelligent agents interacting in a shared environment. MASs can be used to model many complex scenarios, such as urban traffic flow [Wang *et al.*, 2021], multi-robot coordination [Campbell and Wu, 2011], distributed sensing [Lesser *et al.*, 2003] and load balancing [Keshvadi and Faghih, 2016].

Multiagent reinforcement learning (MARL) has recently received much attention [Bloembergen *et al.*, 2015; Lanctot *et al.*, 2017; Leibo *et al.*, 2017]. However, the theory underlying MARL is far from being understood. In seminal work, Tuyls *et al.* [2003; 2006] formally model the dynamics of Q-learning [Watkins and Dayan, 1992] in 2-player games and develop the selection-mutation model. Based on this model, Kianercy and Galstyan [2012] provide a comprehensive characterization of the rest point structure for different 2-player games. Gomes and Kowalczyk [2009] develop another formal model for Q-learning with $\epsilon$-greedy exploration in 2-player games. Wunder *et al.* [2010] show that when $\epsilon$-greedy exploration is applied, the Q-learning dynamics may exhibit chaotic behaviors. However, all these earlier works focus on 2-agent interaction scenarios.

More recently, there have been some works reported on modeling the Q-learning dynamics beyond the 2-agent setting. A pioneer work that we notice is that of Hu *et al.* [2019], which models the population dynamics of Q-learning agents according to the *concurrent learning* (CL) protocol under an $n$-agent setting with $n \to \infty$. Using mean field theory [Weiss, 1907], the authors capture the population dynamics by a system of three equations.

In this work, we aim at formally modelling the Q-learning dynamics in the scenarios that agents abide by the *generalized social learning* (GSL) protocol, a generalized version of the *social learning* (SL) protocol [Sen and Airiau, 2007]. Specifically, at every time step, each agent forms pairs with $m \geq 1$ randomly chosen opponents; then each pair of agents simultaneously plays a 2-player symmetric game and learns the strategy based on the reward from the game. SL and CL are the extreme cases of GSL, in which $m = 1$ and $m \to \infty$, respectively.

However, unlike CL in the work of Hu *et al.* [2019], we consider that in GSL each agent plays with a finite number ($m$) of opponents, instead of infinitely many, at each time step. That is, our work tackles the more common scenarios in which agents have only *local* and *incomplete* information about the system, instead of complete and global information as in the ideal scenario in the work of Hu *et al.* [2019].

The (generalized) social learning protocol is significant as it is widely used in the literature of research into norm and cooperation emergence [Eccles *et al.*, 2019; McKee *et al.*, 2020; Hu and Leung, 2017; Hao and Leung, 2013; Baker, 2020; Anastassacos *et al.*, 2020]. Lacking proper mathematical underpinnings, all these previous studies have had to rely on experimental studies using agent-based simulations.

By adopting the GSL protocol, we have to deal with the stochasticity in the learning process of each agent, which is not present in CL. This is because (i) an agent's $m$ opponents in games are randomly chosen and they change over time, and (ii) the exploration mechanism of Q-learning generally presupposes that the action choice is stochastic. Such

[*]Corresponding Author.

stochasticity has not been captured by the previous model of Hu *et al.* [2019], which focuses on depicting the change of expectation of Q-values under the CL protocol and thus is effectively deterministic. To properly characterize these stochastic effects on Q-learning dynamics under the GSL protocol, we carefully model the Q-values dynamics of each individual agent as a system of stochastic differential equations (SDE). Specifically, we approximate the Q-values dynamics by Gaussian moment closure method and evaluate the first and second order moments of the change in Q-values for each individual agent. As such, given $d$ available actions for each agent, we model the Q-values dynamics of each agent using a $d$-dimensional Wiener process such that the drift and diffusion of the Wiener process depend on the Q-values, the exploration mechanism and the probability density function (PDF) of Q-values in the population. As a result of the learning of individual agents, the PDF of Q-values in the population also evolve accordingly. Based on the SDE, we express the time evolution of the PDF of Q-values in the population with a Fokker-Planck equation (FPE). Therefore, our formal model for the Q-learning dynamics under the GSL protocol is given by two coupled equations: the SDE that captures the Q-values dynamics of each agent and the FPE that captures the Q-values dynamics of the entire population.

In the experiments, we consider four canonical 2-player symmetric games: prisoner's dilemma, stag hunt, hawk-dove and rock-paper-scissors. We show that across different games and initial Q-value settings, our new model always provides a significantly more accurate description of the learning dynamics under the GSL protocol than the model of Hu *et al.* [2019], particularly when the number $(m)$ of opponents is small (Figure 1). Even when the number $(m)$ of opponents tends to infinity, which is equivalent to the CL protocol, our new model still has a better accuracy (Figure 2). Moreover, our model reveals that different values of $m$, which determines the completeness of information each agent has, can lead to *qualitatively* different Q-learning dynamics (Figure 4 in the appendix) and even outcomes (Figure 3).

Our results clearly indicate that the effects of local interactions and incomplete information on multiagent Q-learning are *non-trivial*. This provides theoretical evidence for the previous findings that rely on agent-based simulations [Hao and Leung, 2013; Yu *et al.*, 2013]. To aid investigations on such effects, our model lays the mathematical underpinnings and can provide insights that are unable to obtain using the previous model [Hu *et al.*, 2019]. Therefore, our model is an important theoretical contribution towards a better understanding of multiagent Q-learning.

## 2 Preliminaries

Throughout this paper, we denote a column vector $\boldsymbol{x} \in \Re^d$ as $(x_1, ..., x_d)$.

### 2.1 Symmetric Games

Conventionally, a 2-player-$d$-action normal-form game involves a row player and a column player, each of which has a set of $d$ available actions to choose from. During game plays, the two players simultaneously choose an action and receive an immediate payoff (or reward) based on their joint action choices. Formally, a 2-player-$d$-action game can be represented by two payoff matrices ($\boldsymbol{U}_{\text{row}}$ and $\boldsymbol{U}_{\text{column}}$) as follows:

$$\boldsymbol{U}_{\text{row}} = \begin{bmatrix} p_{11} & \cdots & p_{1d} \\ \vdots & \ddots & \vdots \\ p_{d1} & \cdots & p_{dd} \end{bmatrix} \quad \boldsymbol{U}_{\text{column}} = \begin{bmatrix} q_{11} & \cdots & q_{1d} \\ \vdots & \ddots & \vdots \\ q_{d1} & \cdots & q_{dd} \end{bmatrix}$$

where each element represents the immediate payoff of the row or column player given joint action choices. The game is symmetric if (i) both the players have the same set of available actions, and (ii) the resulting payoffs depend *not* on the roles of the players, but only on their joint action choices, i.e. $\boldsymbol{U}_{\text{row}} = \boldsymbol{U}_{\text{column}}^{\top}$. Let us define $\boldsymbol{U} := \boldsymbol{U}_{\text{row}} = \boldsymbol{U}_{\text{column}}^{\top}$. For any player, suppose that it chooses action $a_i$ and its opponent chooses action $a_j$, its immediate payoff denoted by $r$ is given by

$$r = \boldsymbol{e}_i^{\top} \boldsymbol{U} \boldsymbol{e}_j \tag{1}$$

where $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ represent basis vectors such that $\boldsymbol{e}_i = (0...1...0)$ with only the $i^{\text{th}}$ element equaling 1.

### 2.2 Q-Learning and Boltzmann Exploration

Q-learning is a reinforcement learning algorithm. It is defined in a Markov decision process (MDP) $\langle S, A, T, R\rangle$, where $S$ is a set of states, $A$ is a set of the available action, $T : S \times A \to \mathcal{P}(S)$ is a state transition probability function, and $R : S \times A \to \Re$ is a reward (or payoff) function. A Q-learning agent maintains a Q-value for each state-action pair $(s, a)$ to estimate the expected reward of using each action $a \in A$ under each state $s \in S$. Suppose that at a given time step $t$, the agent is in state $s$ and selects an action $a_i$, we denote the corresponding Q-value as $Q_i^t(s) := Q^t(s, a_i)$. Let $r_t$ be the immediate reward it receives as the new state becomes $s'$. The agent updates its Q-value for the state-action pair $(s, a_i)$ as follows:

$$Q_i^{t+1}(s) = (1-\alpha)Q_i^t(s) + \alpha(r_t + \gamma \max_{a_j \in A} Q_j^t(s')) \tag{2}$$

where $Q_i^{t+1}(s)$ is the updated Q-value, $\alpha \in (0, 1)$ is the learning rate, $\gamma$ is the discount factor, and $\max_{a_j \in A} Q_j^t(s')$ estimates the maximum reward after state transition. For symmetric games, since there is no state transition in every play of the games, Equation (2) is simplified to

$$Q_i^{t+1} = (1-\alpha)Q_i^t + \alpha r_t. \tag{3}$$

A Q-learning agent selects its action based on its Q-values and the exploration mechanism that it applies. Let $\boldsymbol{x}^t = (x_1^t, ..., x_d^t)$ (conventionally referred to as the *policy*) be a point in a $d-1$ probability simplex where $x_k^t$ is the probability of selecting action $a_k$ given the current Q-values at time $t$. For *Boltzmann exploration*, the probability that an agent uses an action $a_i$ at time $t$ is given by

$$x_i^t = \frac{e^{\tau Q_i^t}}{\sum_{j=1}^{d} e^{\tau Q_j^t}} \tag{4}$$

where $\tau$ is a parameter known as the inverse of the temperature. The agent is in pure exploration (randomly taking each action) when $\tau$ is 0, and in pure exploitation (taking the action with the highest $Q$-value) when $\tau \to \infty$.

**Algorithm 1** Generalized Social Learning Protocol

```
Input: $N, m, T, \alpha, \tau$
create and initialise $Agents[0]$ to $Agents[N-1]$
for $time = 1$ to $T$ do
    for all $agent$ in $Agents$ parallel do
        $agent.genAction()$
    end for
    $assignments = gameAssignment(N, m)$
    for all $am$ in $assignments$ parallel do
        $playGame(Agents[am[0]], Agents[am[1]])$
    end for
    for all $agent$ in $Agents$ parallel do
        $agent.train(\alpha, \tau)$ according to equation (6)
    end for
end for
```

# 3 A Formal Model of Population Dynamics for Large-Scale Multiagent Q-learning

In this section, we propose a formal model of population dynamics for large-scale multiagent systems in which the interaction protocol is a generalized version of social learning [Sen and Airiau, 2007]. We first revisit the model proposed by Hu *et al.* [2019], and point out an inadequacy of their model in this scenario. We then develop a formal model that rectifies this inadequacy.

## 3.1 The Generalized Social Learning (GSL) Protocol

*Social learning* (SL) [Sen and Airiau, 2007] is a widely adopted protocol in the multiagent reinforcement learning literature [Eccles *et al.*, 2019; McKee *et al.*, 2020; Baker, 2020; Anastassacos *et al.*, 2020]. Consider a large population of $N$ Q-learning agents, where $N$ tends to infinity. The social learning (SL) protocol prescribes that agents learn their strategies during repeated local interactions with different agents. Specifically, at each time step $t$, each agent is paired up with a randomly chosen opponent to play a 2-player symmetric game. All these games are played simultaneously and independently. Before a game play, each agent selects an action based on its Q-values and the exploration mechanism. During the game play, each pair of agents take actions simultaneously and receive immediate payoffs. After the game play, each agent updates the Q-value of the action it takes independently.

In this paper we consider *generalized social learning* (GSL), in which each agent will instead choose an action, play with $m \geq 1$ randomly chosen opponents and update the Q-value of the chosen action based on the averaged rewards. We present the pseudocode of this interaction protocol in Algorithm 1. The canonical social learning is a special case when we consider $m = 1$, while the *concurrent learning* (CL) protocol considered by Hu *et al.* [2019] is another special case with $m \to \infty$.

## 3.2 Formalizing the Q-Learning Dynamics under the Social Learning Protocol

In the concurrent learning (CL) scenario studied in Hu *et al.* [2019], each agent has *global* interactions, that is, with *all* the other (infinitely many) agents, at every time step. Hence each agent's learning dynamics can be abstracted by that of its interaction with a virtual *mean field agent*, and the population dynamics can be described by a system of deterministic equations. A consequent shortcoming in their model is that the stochasticity arising from each agent's *local* interactions with only $m$ randomly chosen opponents is not captured. As we shall show in Section 4, the population behaviour in the 'ideal' situation as predicted by their model can significantly deviate from the actual one when agents are learning through local interactions rather than global interactions, particularly when the value of $m$ is small.

**Change of Q-values with respect to time.** Adopting the notations used by Hu *et al.* [2019], we denote the vector of Q-values of an arbitrary agent by $\boldsymbol{Q}^t = (Q_1^t, \ldots, Q_d^t)$ such that each agent is represented by its Q-value vector in a $d$-dimensional space. Let $p(\boldsymbol{Q}^t) : \Re^d \to \Re$ be the probability density function (PDF) that represents the density of agents having certain Q-values at time step $t$. We denote the action distribution in the population by $\boldsymbol{y}^t = (y_1^t, ..., y_d^t)$, where $y_i^t$ is the proportion of agents that take action $a_i$ at time $t$. Given the PDF of agents at time $t$, we have

$$y_i^t := \mathbb{E}_{\boldsymbol{Q}^t}[x_i^t] = \int_{\Re^d} x_i^t p(\boldsymbol{Q}^t) d\boldsymbol{Q}^t \tag{5}$$

where $x_i^t$ is given by Equation (4) for Boltzmann exploration. The change $\Delta\boldsymbol{Q}^t$ in Q-values depends on the immediate reward jointly determined by the agent's action and its opponent's action. Let $a_i$ be the action used by the agent and $a_{j_1}, a_{j_2}, \ldots, a_{j_m}$ be the actions used by its $m$ opponents, by Equation (1) and (3), we have

$$\Delta Q_i^t := Q_i^{t+1} - Q_i^t = \alpha\left(\frac{1}{m}\sum_{h=1}^{m} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{e}_{j_h} - Q_i^t\right) \tag{6}$$

Note that $\Delta Q_k^t = 0$ for any other action $a_k$ not used. Consider the change in Q-values in continuous time, where $\partial_t Q_i^t := \frac{\partial Q_i^t}{\partial t} = \Delta Q_i^t$, we have

$$\partial_t Q_i^t = \alpha\left(\frac{1}{m}\sum_{h=1}^{m} \boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{e}_{j_h} - Q_i^t\right) \tag{7}$$

where $t \in \Re$.

**Moment closure approximation of rate of changes of Q-values.** From this point we shall apply the moment closure method [Goodman, 1953; Whittle, 1957] to approximate $\partial_t \boldsymbol{Q}^t = (\partial_t Q_1^t, ..., \partial_t Q_d^t)$. In a nutshell, the moment closure method takes the cumulant-generating function of $\partial_t \boldsymbol{Q}^t$

$$\begin{aligned} K^t(\boldsymbol{\nu}) &= \log(\mathbb{E}[e^{\mathrm{i}\boldsymbol{\nu}^\top \partial_t \boldsymbol{Q}^t}]) \\ &= \sum_i \nu_i \kappa_i + \frac{1}{2!}\sum_{i,j} \nu_i \nu_j \kappa_{i,j} + \frac{1}{3!}\sum_{i,j,k} \nu_i \nu_j \nu_k \kappa_{i,j,k} + ... \end{aligned}$$

where $\mathrm{i} = \sqrt{-1}$, $\kappa_i$, $\kappa_{i,j}$, $\kappa_{i,j,k}$, $\ldots$ are cumulants, and sets the 3rd and higher order cumulants to zero. In other words, $\partial_t \boldsymbol{Q}^t$ is approximated by a multivariate Gaussian distribution. Hence, it is expressed by the following stochastic process

$$\partial_t \boldsymbol{Q}^t \approx \boldsymbol{\mu}^t dt + \sqrt{\boldsymbol{\Sigma}^t} d\boldsymbol{W}^t \tag{8}$$

where $\boldsymbol{\mu}^t = \mathbb{E}[\partial_t \boldsymbol{Q}^t]$ is the *mean vector*, and $\boldsymbol{\Sigma}^t = [\sigma^t_{kl}] \in \Re^{d\times d}$ is the *covariance matrix* with $\sigma^t_{kk} = \mathrm{Var}(\partial_t Q^t_k)$ being the variance of $\partial_t Q^t_k$ and $\sigma^t_{kl} = \mathrm{Cov}(\partial_t Q^t_k, \partial_t Q^t_l)$ the covariance between $\partial_t Q^t_k$ and $\partial_t Q^t_l$. $\boldsymbol{W}^t = (W^t_1, ..., W^t_d)$ is the standard $d$-dimensional Wiener process. Note that $\boldsymbol{\Sigma}^t$ is positive definite.

**Calculation of the mean vector and the covariance matrix.** For the particular setting we consider, the value of $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$ are derived in the following. From Equation (7), it is clear that $\partial_t \boldsymbol{Q}^t$ depends on the agent's action and its opponents' action. Let $a_i$ be the agent's action chosen by the policy $\boldsymbol{x}^t$ and $a_{j_1}, ..., a_{j_m}$ be the $m$ opponents' actions determined by the action distribution $\boldsymbol{y}^t$ in the population, and $\boldsymbol{z}^t = (0...1...0)$ and $\boldsymbol{\xi}^t_h = (0...1...0)$, $h = 1, ..., m$ the corresponding one-hot vectors. Let $\boldsymbol{\zeta}^t = \sum_h \boldsymbol{\xi}^t_h$, we have $\boldsymbol{z}^t|\boldsymbol{x}^t \sim \mathrm{multinomial}(1, \boldsymbol{x}^t) \in [0,1]^d$. Likewise, we also have $\boldsymbol{\zeta}^t|\boldsymbol{y}^t \sim \mathrm{multinomial}(m, \boldsymbol{y}^t) \in [0, ..., m]^d$.

Recall that at each time step, an agent interacts with $m$ randomly chosen opponents, and since the agent chooses to play action $a_i$, the Q-value of non-chosen actions are not updated $(\partial_t Q^t_k \mid z^t_k = 0) \equiv 0$. Hence by Equation (7),

$$\begin{aligned}\partial_t Q^t_k \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t &\equiv \alpha(\frac{1}{m}\sum_{h=1}^{m} \boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{e}_{j_h} - Q^t_k) z^t_k \\ &= \alpha(\frac{1}{m}\sum_{j=1}^{d} \zeta^t_j \boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{e}_j - Q^t_k) z^t_k\end{aligned} \tag{9}$$

The unconditional first and second moments of $\partial_t \boldsymbol{Q}^t$ are evaluated as

$$\mu^t_k = \mathbb{E}[\partial_t Q^t_k] = \alpha x^t_k (\boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{y}^t - Q^t_k) \tag{10}$$

$$\begin{aligned}\sigma_{kk} &= \mathrm{Var}(\partial_t Q^t_k) \\ &= \mathbb{E}[\mathrm{Var}(\partial_t Q^t_k \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[\partial_t Q^t_k \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t)] \\ &\quad + \mathrm{Var}(\mathbb{E}[\partial_t Q^t_k \mid \boldsymbol{z}^t]) \\ &= \alpha^2 (\boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{y}^t - Q^t_k)^2 x^t_k (1 - x^t_k) \\ &\quad + \frac{1}{m}\alpha^2 x^t_k [\boldsymbol{e}_k^\top \boldsymbol{U} \circ \boldsymbol{U} \boldsymbol{y}^t - (\boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{y}^t)^2]\end{aligned} \tag{11}$$

$$\begin{aligned}\sigma_{kl} &= \mathrm{Cov}(\partial_t Q^t_k, \partial_t Q^t_l) \\ &= \mathbb{E}[\mathrm{Cov}(\partial_t Q^t_k, \partial_t Q^t_l \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t)] \\ &\quad + E[\mathrm{Cov}(\mathbb{E}[\partial_t Q^t_k \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t], \mathbb{E}[\partial_t Q^t_l \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t)] \\ &\quad + \mathrm{Cov}(\mathbb{E}[\partial_t Q^t_k \mid \boldsymbol{z}^t], \mathbb{E}[\partial_t Q^t_l \mid \boldsymbol{z}^t]) \\ &= -\alpha^2 (\boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{y}^t - Q^t_k)(\boldsymbol{e}_l^\top \boldsymbol{U} \boldsymbol{y}^t - Q^t_l) x^t_k x^t_l\end{aligned} \tag{12}$$

where $\boldsymbol{U} \circ \boldsymbol{U}$ represents the element-wise multiplication of matrix $\boldsymbol{U}$ and $\boldsymbol{U}$, $\boldsymbol{\mu}^t = (\mu^t_1, ..., \mu^t_d)$ and $\boldsymbol{\Sigma}^t = [\sigma^t_{kl}]$.

We can see from Equation (11) that the variance of $\partial_t Q^t_i$ is decomposed into two terms: the first term is induced by the randomness of the action choice of the focal agent, whereas the second term is induced by the randomness of the action choices of the opponents that are randomly selected from the population. The covariance (Equation 12) between $\partial_t Q^t_i$ and $\partial_t Q^t_j$ is purely induced by the action choice of the focal agent such that choosing an action $a_i$ forces $\partial_t Q^t_k = 0$ for all the other actions $a_k \neq a_i$.

**Time evolution of the PDF of agent density.** We are now ready to formalize the time evolution of the PDF $p(\boldsymbol{Q}^t)$ representing density of agents having certain Q-values at time step $t$ by applying the Fokker-Planck equation (FPE) [Fokker, 1914; Planck, 1917]. In general, the FPE for a $d$-dimensional stochastic process $\boldsymbol{x}^t$ depicted by the stochastic process

$$\partial_t \boldsymbol{x}^t = \boldsymbol{\mu}^t dt + \boldsymbol{\varrho}^t d\boldsymbol{W}^t$$

is

$$\frac{\partial p(\boldsymbol{x}^t)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[\mu^t_i p(\boldsymbol{x}^t)] + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}[D^t_{ij} p(\boldsymbol{x}^t)]$$

where $\boldsymbol{\mu}^t = \{\mu^t_1, \ldots, \mu^t_d\}$ is known as the *drift vector* and $\boldsymbol{D}^t = [D^t_{ij}] = \frac{1}{2}\boldsymbol{\varrho}^t \boldsymbol{\varrho}^{t\top}$ the *diffusion tensor*. Hence the FPE that describes how $p(\boldsymbol{Q}^t)$ changes over time according to Equation (8) is given by

$$\frac{\partial p(\boldsymbol{Q}^t)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial Q_i}[\mu^t_i p(\boldsymbol{Q}^t)] + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial Q_i \partial Q_j}[\sigma^t_{ij} p(\boldsymbol{Q}^t)] \tag{13}$$

Given the initial Q-values distribution of agents $p(\boldsymbol{Q}^0)$, the PDF $p(\boldsymbol{Q}^t)$ $\forall t \in \Re$ can be obtained by solving Equations (8) and (13) simultaneously. Note that the population policy $\boldsymbol{y}^t$ can then be obtained by Equation (5).

### 3.3 Discussion

In their pioneer work, Hu *et al.* [2019] assume that an agent always plays with a virtual mean field agent, whose policy is the expectation of all of the (infinitely many) other agents in the population. Hence the counterpart of Equations 8 and 9 in their model is

$$\mathbb{E}[\partial_t Q^t_i] = \alpha x^t_i (\boldsymbol{e}_i^\top \boldsymbol{U} \boldsymbol{y}^t - Q^t_i)$$

where $\boldsymbol{y}^t$ is the policy of the virtual mean field agent at time $t$. It is worth noting that the stochasticity of focal agent's interactions with $m$ opponents, which is captured by Equation (7) and approximated via the moment closure method in Equation (8), is omitted in the previous model. Specifically, the covariance matrix $\boldsymbol{\Sigma}$ in the second term of Equation (8) depicting the stochastic process of $\partial_t \boldsymbol{Q}^t$, together with the variances (Equation 11) and covariances (Equation 12) in it, are effectively summarily assumed to be zero in the previous model.

We note that such differences would be less and less significant when $m \to \infty$, as $m$ appears as a denominator in a factor in Equation (11). However, they are noticeable when $m$ is small. Particularly, the model of Hu *et al.* [2019], which assumes $m \to \infty$, is generally unable to correctly depict the population dynamics in the canonical social learning settings, in which $m = 1$. The inaccuracy keeps noticeable as long as the value of $m$ is small. The new model provides a more accurate depiction of the population dynamics, and rectifying these inaccuracies. As we will show in Section 4 (Figure 1), the new model represents the population dynamics much more correctly when $m$ is small.

On the other hand, it should be noted that the new model provides a more accurate depiction of the population dynamics than the previous model in many cases even when the

| | $C$ | $D$ |
|---|---|---|
| $C$ | $2, 2$ | $0, 3$ |
| $D$ | $3, 0$ | $1, 1$ |

(a) Prisoner's Dilemma

| | $S$ | $H$ |
|---|---|---|
| $S$ | $3, 3$ | $0, 2$ |
| $H$ | $2, 0$ | $1, 1$ |

(b) Stag Hunt

| | $H$ | $D$ |
|---|---|---|
| $H$ | $-2, -2$ | $2, 0$ |
| $D$ | $0, 2$ | $1, 1$ |

(c) Hawk–Dove

| | $R$ | $P$ | $S$ |
|---|---|---|---|
| $R$ | $0, 0$ | $-1, 1$ | $1, -1$ |
| $P$ | $1, -1$ | $0, 0$ | $-1, 1$ |
| $S$ | $-1, 1$ | $1, -1$ | $0, 0$ |

(d) Rock, Paper, Scissors

Table 1: Payoff bi-matrices of the games considered in our experiments.

| Game | pure-strategy NE | mixed-strategy NE $(\beta, \beta)$ |
|---|---|---|
| PD | $(D, D)$ | - |
| SH | $(S, S), (H, H)$ | $\beta = (S(1/2), H(1/2))$ |
| HD | $(H, D), (D, H)$ | $\beta = (H(1/3), D(2/3))$ |
| RPS | - | $\beta = (R(1/3), P(1/3), S(1/3))$ |

Table 2: Nash equilibria of the games considered in our experiments. The mixed-strategy Nash equilibrium $(\beta, \beta)$ means that both players take each action with probability $\beta$.

value of $m$ is large, as we shall show in Section 4 (Figure 2). This is largely due to the fact that the new model does not consider the covariances (Equation 12) to be zero, which is nonetheless an implicit working assumption in the previous model. Consequently, the influence of the changes of the Q-value of an action on the changes of the Q-value of another action of the same agent is duly recognised, and it manifests in the generally nonzero covariances $\sigma_{kl}$ in the covariance matrix $\mathbf{\Sigma}$.

# 4 Experiments

In this section, we apply the model derived in Section 3 to four canonical types of 2-player symmetric games. In Section 4.1, we describe the game configurations and our settings of multiagent Q-learning. In Section 4.2, we validate that our model provides a more accurate description than the previous model [Hu *et al.*, 2019] through comparisons with agent-based simulation results across different settings. In Section 4.3, we illustrate that our model can shed light on the effects of local interactions and incomplete information on multiagent Q-learning.

## 4.1 Experimental Settings

**Game Configurations.** We consider $4$ different types of 2-player symmetric games: Prisoner's Dilemma (PD) game, Stag Hunt (SH) game, Hawk-Dove (HD) game, and Rock, Paper, Scissors (RPS) game. We present the payoff bi-matrices of these games in Table 1 and summarize the properties of these games in Table 2. Note that these games differ greatly in the number and types of Nash equilibria.

**Settings of Multiagent Q-Learning.** We consider that agents interact according to the GSL protocol (presented in Algorithm 1). For the initial Q-values of each agent at time $t = 0$, we consider two settings: (i) the Q-values of every action are $0$, which is a common setting in MARL research, and (ii) the Q-value of the first action is $1$ and the Q-values of the other actions are $0$. The default learning rate and Boltzmann temperature are $\alpha = 0.1$ and $\tau = 2$, respectively.

## 4.2 Our Model versus Previous Model

In this section, we compare the Q-learning behaviors predicted by our model with the prediction made by the previous model [Hu *et al.*, 2019] across different settings of games and initial Q-values (as described in the previous section). For validation, we take the agent-based simulation results as the benchmark. For each specific setting, we obtain the prediction of our model by solving Equation (8) and (13) with the finite volume method [Wyns and Du Toit, 2017] and obtain the simulation results by repeating $100$ independent simulation runs. For each simulation run, we consider $2,000$ agents.

Figure 1 presents the comparisons under the situations with a *small* value of $m$. It is clear that across different settings, our model always provides more accurate descriptions on the dynamics of the expected Q-values $E[Q_k], \forall k$ in a population. In particular, for Stag-Hunt and Hawk-Dove games, our model precisely predicts the average Q-values at which the population stabilizes; on the contrary, the predictions of the previous model significantly deviate from the benchmark.

Figure 2 presents the comparisons under the situations where $m \rightarrow \infty$. In the simulations, since the number of agents is finite, we consider an agent will play with all the other agents. From Figure 2, we again observe that our model more accurately describes the learning dynamics than the previous model.

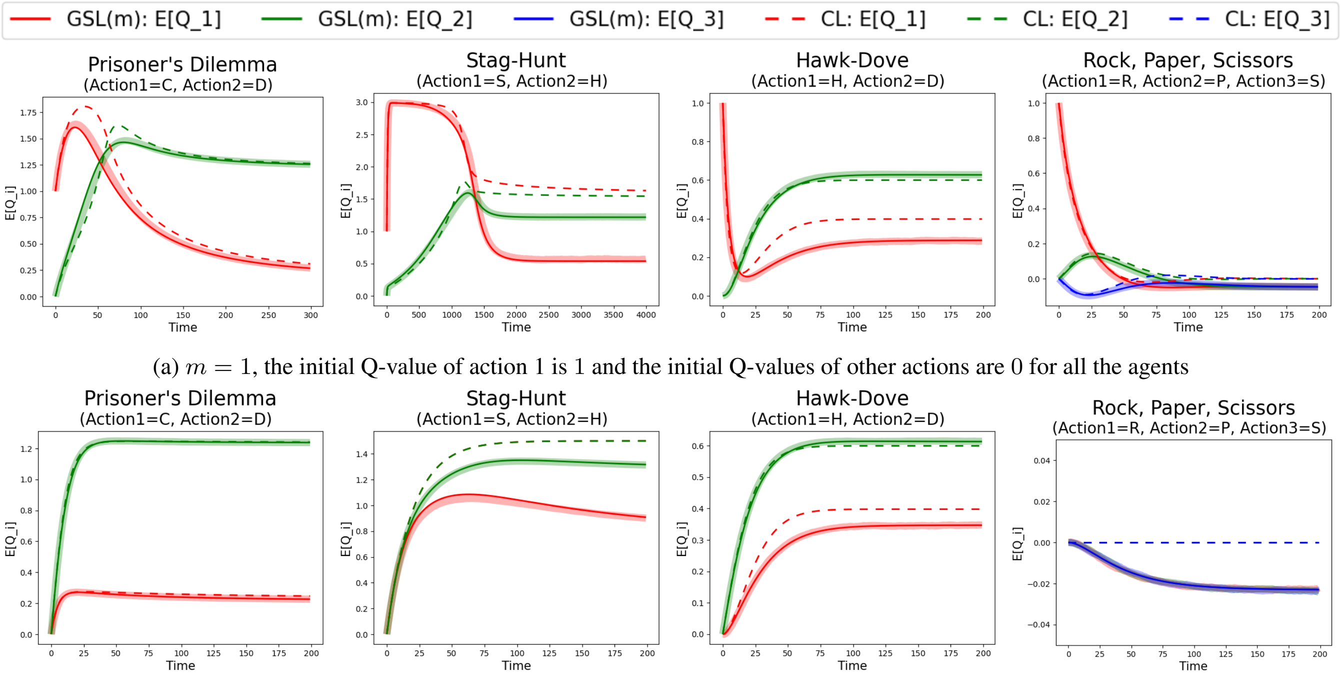
Note that the significantly better accuracy of our model against the previous model can be observed in many settings. Due to space restrictions, more comparisons under other settings are provided in the appendix. Based on all these comparisons, we validate our main theoretical claim — our model provides more accurate descriptions of the Q-learning dynamics under the GSL protocol.

## 4.3 Effects of Local Interactions and Incomplete Information

As a direct benefit of more accurate descriptions on learning dynamics, our model can provide new theoretical insights that cannot be obtained using the previous model [Hu *et al.*, 2019]. Remember that a larger value of $m$ indicates that agents have interactions with a broader range of opponents and thus have more complete information about the entire population. Figure 3 shows that the value of $m$ plays an important role on the *outcome* of multiagent Q-learning with the GSL protocol.

For Stag-Hunt games (Figure 3(a)), as the value of $m$ increases, Q-learner populations tend to stabilize with a smaller proportion of agents playing action $S$, even though these populations start with the same initial Q-values settings. More interestingly, Q-learner populations converge to the unique completely mixed-strategy Nash equilibrium as $m \rightarrow \infty$, while they do not converge to any particular Nash equilibria when $m$ is finite and small. This clearly shows that different values of $m$ can lead to *qualitatively* different learning outcomes.

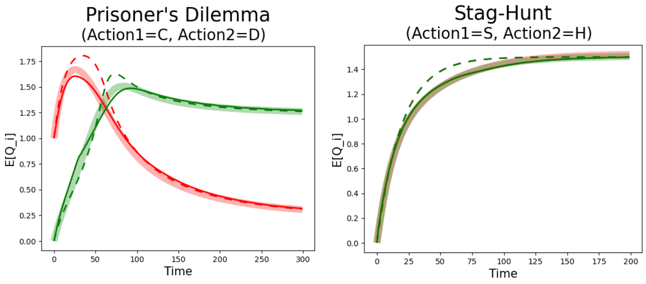
We observe similar effects of the value of $m$ on multiagent Q-learning in Hawk-Dove games. The above results suggest



(a) $m = 1$, the initial Q-value of action 1 is 1 and the initial Q-values of other actions are 0 for all the agents

(b) $m = 2$, the initial Q-value of each action is 0 for all the agents

Figure 1: With a small value of $m$, comparison among the dynamics of average Q-values predicted by our model (solid line), the previous model (dashed line), and the agent-based simulations (shaded line). In all these settings, our model better captures the qualitative and quantitative dynamics of the populations.

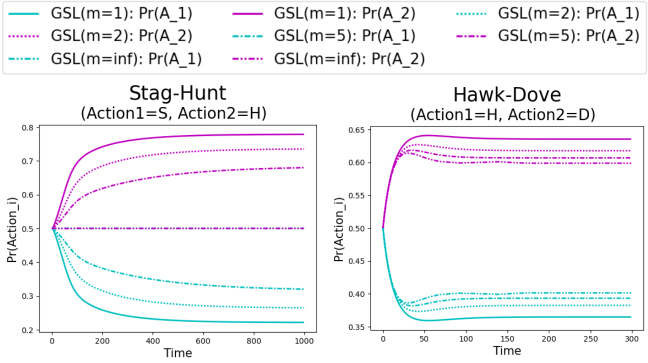


(a) $\boldsymbol{Q}^0 = (1, 0)$

(b) $\boldsymbol{Q}^0 = (0, 0)$

Figure 2: As $m \to \infty$, comparison among the dynamics of average Q-values predicted by our model (solid line), the previous model (dashed line), and the agent-based simulations (shaded line).

that it is non-trivial to investigate the effects of local interactions and incomplete information on multiagent Q-learning. To aid such investigation, our model can provide accurate theoretical insights, which is not available using the previous model.



(a) $\boldsymbol{Q}^0 = (0, 0)$

(b) $\boldsymbol{Q}^0 = (0, 0)$

Figure 3: The effects of local interactions and incomplete information on multiagent Q-learning. As $m$ varies, the population can stabilize at significantly different outcomes.

## 5 Conclusion

In this work, we model the population dynamics in the scenarios of the generalized social learning protocol. We approximate the Q-values dynamics of each individual agent as a system of stochastic differential equations. The time evolution of the probability density function of Q-values is evaluated by the Fokker-Planck equation. Experimental results corroborate the correctness of our model and provide theoretical insights on the Q-learning dynamics under the GSL protocol, which are unable to obtain using the previous model.

## Acknowledgements

The work presented in this paper is partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820).

## References

[Anastassacos *et al.*, 2020] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *AAAI*, pages 7047–7054, 2020.

[Baker, 2020] Bowen Baker. Emergent reciprocity and team formation from randomized uncertain social preferences. *arXiv preprint arXiv:2011.05373*, 2020.

[Bloembergen *et al.*, 2015] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

[Campbell and Wu, 2011] Adam Campbell and Annie S Wu. Multi-agent role allocation: issues, approaches, and multiple perspectives. *AAMAS*, 22(2):317–355, 2011.

[Eccles *et al.*, 2019] Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*, 2019.

[Fokker, 1914] Adriaan Daniël Fokker. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Annalen der Physik*, 348(5):810–820, 1914.

[Gomes and Kowalczyk, 2009] Eduardo Rodrigues Gomes and Ryszard Kowalczyk. Dynamic analysis of multiagent q-learning with $\varepsilon$-greedy exploration. In *ICML*, pages 369–376, 2009.

[Goodman, 1953] Leo A Goodman. Population growth of the sexes. *Biometrics*, 9(2):212–225, 1953.

[Hao and Leung, 2013] Jianye Hao and Ho-fung Leung. The dynamics of reinforcement social learning in cooperative multiagent systems. In *IJCAI*, volume 13, pages 184–190, 2013.

[Hu and Leung, 2017] Shuyue Hu and Ho-fung Leung. Achieving coordination in multi-agent systems by stable local conventions under community networks. In *IJCAI*, pages 4731–4737, 2017.

[Hu *et al.*, 2019] Shuyue Hu, Chin-wing Leung, and Ho-fung Leung. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. In *NIPS*, pages 12125–12135, 2019.

[Keshvadi and Faghih, 2016] Sina Keshvadi and Behnam Faghih. A multi-agent based load balancing system in iaas cloud environment. *International Robotics & Automation Journal*, 1(1):1–6, 2016.

[Kianercy and Galstyan, 2012] Ardeshir Kianercy and Aram Galstyan. Dynamics of boltzmann q learning in two-player two-action games. *Physical Review E*, 85(4):041145, 2012.

[Lanctot *et al.*, 2017] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *NIPS*, pages 4190–4203, 2017.

[Leibo *et al.*, 2017] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*, pages 464–473, 2017.

[Lesser *et al.*, 2003] Victor Lesser, Charles L Ortiz Jr, and Milind Tambe. *Distributed sensor networks: A multiagent perspective*, volume 9. Springer Science & Business Media, 2003.

[McKee *et al.*, 2020] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

[Planck, 1917] M. Planck. *Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie*. Sitzungsberichte der Königlich-Preussischen Akademie der Wissenschaften zu Berlin. Reimer, 1917.

[Sen and Airiau, 2007] Sandip Sen and Stéphane Airiau. Emergence of norms through social learning. In *IJCAI*, volume 1507, page 1512, 2007.

[Tuyls *et al.*, 2003] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *AAMAS*, pages 693–700, 2003.

[Tuyls *et al.*, 2006] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *AAMAS*, 12(1):115–153, 2006.

[Wang *et al.*, 2021] Jinghui Wang, Wei Lv, Yajuan Jiang, Shuangshuang Qin, and Jiawei Li. A multi-agent based cellular automata model for intersection traffic control simulation. *Physica A: Statistical Mechanics and its Applications*, 584:126356, 2021.

[Watkins and Dayan, 1992] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[Weiss, 1907] Pierre Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.*, 6(1):661–690, 1907.

[Whittle, 1957] Peter Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 19(2):268–281, 1957.

[Wunder *et al.*, 2010] Michael Wunder, Michael L Littman, and Monica Babes. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*, pages 1167–1174. Citeseer, 2010.

[Wyns and Du Toit, 2017] Maarten Wyns and Jacques Du Toit. A finite volume–alternating direction implicit approach for the calibration of stochastic local volatility models. *International Journal of Computer Mathematics*, 94(11):2239–2267, 2017.

[Yu *et al.*, 2013] Chao Yu, Minjie Zhang, Fenghui Ren, and Xudong Luo. Emergence of social norms through collective learning in networked agent societies. In *AAMAS*, pages 475–482, 2013.