Modelling the dynamics of multi-agent Q-learning: the stochastic effects of local interaction and incomplete information

Chin-wing Leung[1], Shuyue Hu[2], 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

A Derivation for equation (10)-(12)

Starting with

$$\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t \equiv \alpha(\frac{1}{m}\sum_{j=1}^{d} \zeta_j^t \boldsymbol{e}_k^\top \boldsymbol{U} \boldsymbol{e}_j - Q_k^t) z_k^t \tag{9}$$

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

$$\begin{aligned}
\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t] &= \alpha z_k^t(\frac{1}{m}\sum_{j=1}^{d} \mathbb{E}[\zeta_j^t] \boldsymbol{e}_k^\top U \boldsymbol{e}_j - Q_k^t) \\
&= \alpha z_k^t(\frac{1}{m}\sum_{j=1}^{d} m y_j^t \boldsymbol{e}_k^\top U \boldsymbol{e}_j - Q_k^t) \\
&= \alpha z_k^t(\boldsymbol{e}_k^\top U \boldsymbol{y}^t - Q_k^t)
\end{aligned}$$

$$\mu_k^t = \mathbb{E}[\partial_t Q_k^t] = \alpha x_k^t(\boldsymbol{e}_k^\top U \boldsymbol{y}^t - Q_k^t) \tag{10}$$

$$Var(\partial_t Q_i^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t) = 0$$

$$\begin{aligned}
Var(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t) &= \alpha^2 (z_k^t)^2 \frac{1}{m^2} Var(\sum_{j=1}^{d} \zeta_j^t \boldsymbol{e}_k^\top U \boldsymbol{e}_j) \\
&= \alpha^2 (z_k^t)^2 \frac{1}{m^2}[\sum_{j=1}^{d} Var(\zeta_j^t \boldsymbol{e}_k^\top U \boldsymbol{e}_j) + 2\sum_{j_1 \neq j_2} Cov(\zeta_{j_1}^t \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_1}), \zeta_{j_2}^t \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_2}))] \\
&= \alpha^2 (z_k^t)^2 \frac{1}{m^2}[\sum_{j=1}^{d} (\boldsymbol{e}_k^\top U \boldsymbol{e}_j)^2 m y_j^t (1 - y_j^t) - 2\sum_{j_1 \neq j_2} m y_{j_1}^t y_{j_2}^t \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_1} \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_2}] \\
&= \alpha^2 (z_k^t)^2 \frac{1}{m}[\sum_{j=1}^{d} y_j^t \boldsymbol{e}_k^\top U \circ U \boldsymbol{e}_j - \sum_{j=1}^{d} (y_j^t)^2 (\boldsymbol{e}_k^\top U \boldsymbol{e}_j)^2 \\
&\quad - 2\sum_{j_1 \neq j_2} m y_{j_1}^t y_{j_2}^t \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_1} \boldsymbol{e}_k^\top U \boldsymbol{e}_{j_2}] \\
&= \frac{1}{m}\alpha^2 (z_k^t)^2 [\boldsymbol{e}_k^\top U \circ U \boldsymbol{y}^t - (\boldsymbol{e}_k^\top U \boldsymbol{y}^t)^2]
\end{aligned}$$

$$E[Var(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t)] = \frac{1}{m}\alpha^2 x_k^t [\boldsymbol{e}_k^\top U \circ U \boldsymbol{y}^t - (\boldsymbol{e}_k^\top U \boldsymbol{y}^t)^2]$$

$$Var(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t]) = \alpha^2 (\boldsymbol{e}_k^\top U \boldsymbol{y}^t - Q_k^t)^2 x_k^t (1 - x_k^t)$$

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

$$Cov(\partial_t Q_k^t, \partial_t Q_l^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t) = 0$$

$$Cov(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t], \mathbb{E}[\partial_t Q_l^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t) = -\alpha^2 z_k^t z_l^t \frac{1}{m^2} Cov(\sum_{j=1}^d \zeta_j^t \boldsymbol{e}_k^\top U \boldsymbol{e}_j, \sum_{j=1}^d \zeta_j^t \boldsymbol{e}_l^\top U \boldsymbol{e}_j)$$

$$E[Cov(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t], \mathbb{E}[\partial_t Q_l^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t] \mid \boldsymbol{z}^t)] = 0$$

$$Cov(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t], \mathbb{E}[\partial_t Q_l^t \mid \boldsymbol{z}^t]) = -\alpha^2 (\boldsymbol{e}_k^\top U \boldsymbol{y}^t - Q_k^t)(\boldsymbol{e}_l^\top U \boldsymbol{y}^t - Q_l^t) x_k^t x_l^t$$

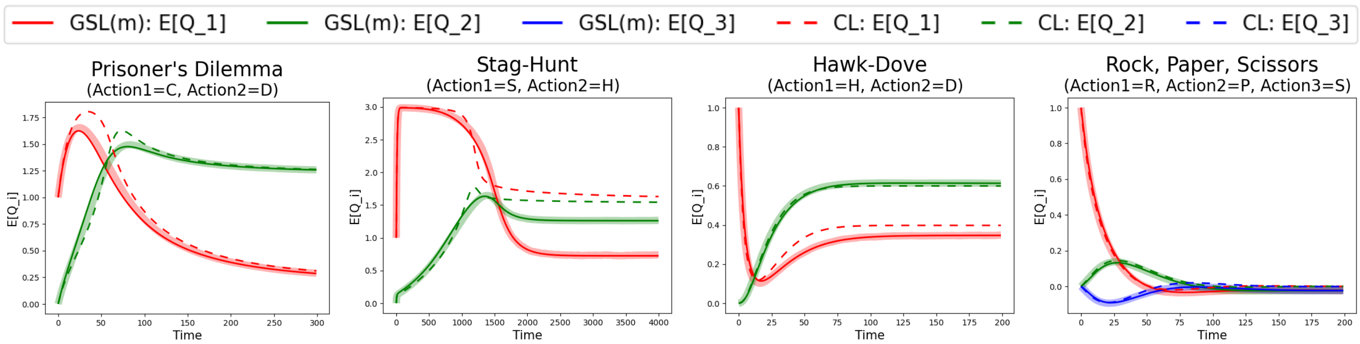
$$\begin{aligned}\sigma_{kl}^t &= Cov(\partial_t Q_k^t, \partial_t Q_l^t) \\ &= \mathbb{E}[Cov(\partial_t Q_k^t, \partial_t Q_l^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t)] + E[Cov(\mathbb{E}[\partial_t Q_k^t \mid \boldsymbol{z}^t, \boldsymbol{\zeta}^t], \mathbb{E}[\partial_t Q_l^t \mid \boldsymbol{z}^t,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

where $\boldsymbol{U} \circ \boldsymbol{U}$ represents the element-wise multiplication of matrix $\boldsymbol{U}$ and $\boldsymbol{U}$.

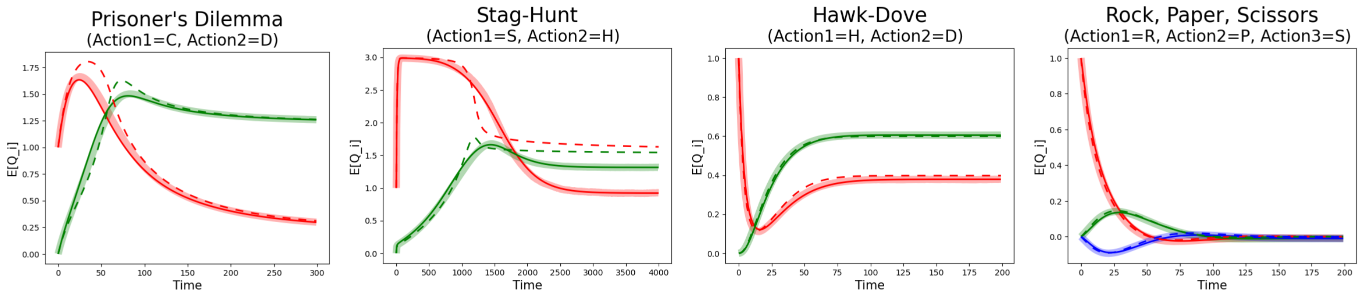
B Additional experiments results

We provide additional experiments results in the following. Figure 1 and 2 contrast the Q-learning dynamics predicted by our model (GSL) with the prediction made by the previous model (CL), the experimental settings are described in the section 4.1. We take the agent-based simulation results as the benchmark. Figure 1 presents the comparisons under the situations with a *small* value of m.[1] Figure 2 presents the comparisons under the situations where $m \to \infty$. It is clear that our model always provides more accurate descriptions on the dynamics of the expected Q-values $E[Q_k], \forall k$ in a population. Figure 3 and 4 shows the dynamics of Q-values and population policy for different m, it is clear that the value of m plays an important role on the *outcome* of multiagent Q-learning in the GSL protocol.

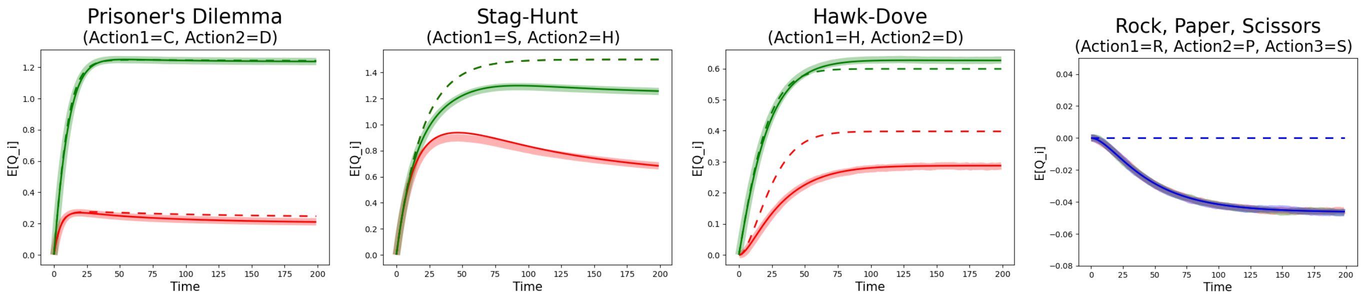
[1] When m is large , the dynamics are close to the case of $m \to \infty$.



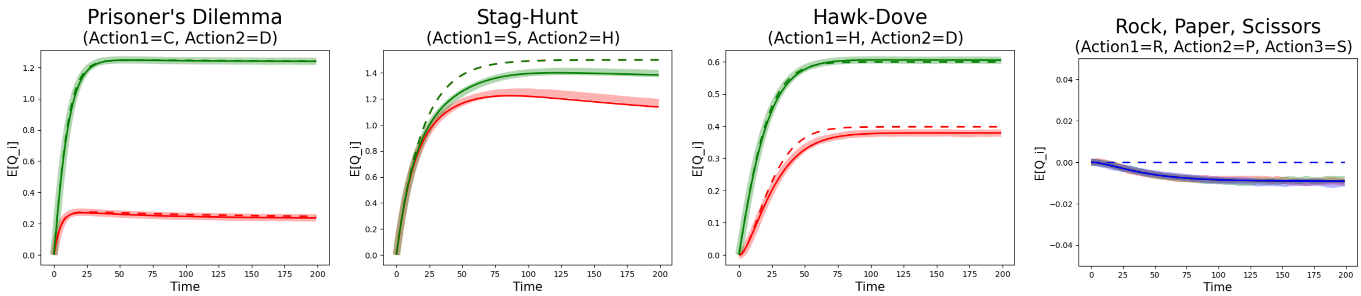
(a) $m = 2$, the initial Q-value of action 1 is 1 and the initial Q-values of other actions are 0 for all the agents



(b) $m = 5$, the initial Q-value of action 1 is 1 and the initial Q-values of other actions are 0 for all the agents

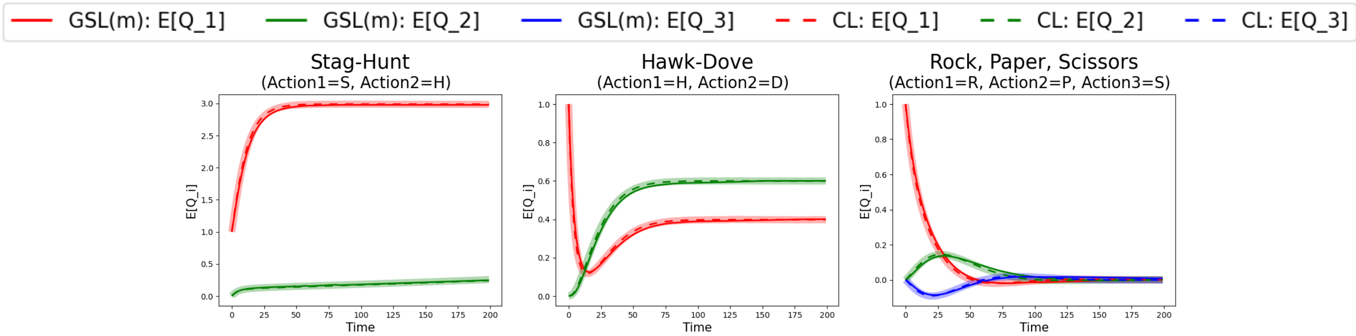


(c) $m = 1$, the initial Q-value of each action is 0 for all the agents

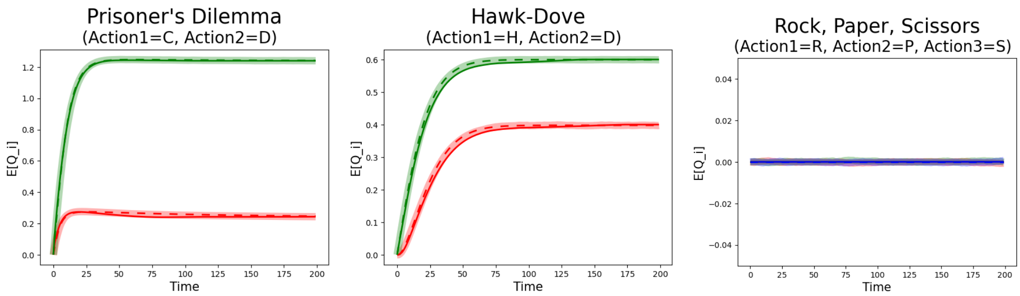


(d) $m = 5$, the initial Q-value of each action is 0 for all the agents

Figure 1: With a small value of m, comparison among the dynamics of average Q-values predicted by our model (solid line) and the previous model (dashed line), and the actual dynamics averaged over 100 runs of agent-based simulations (shaded line). In all these settings, our model better captures the qualitative and quantitative dynamics of the populations.

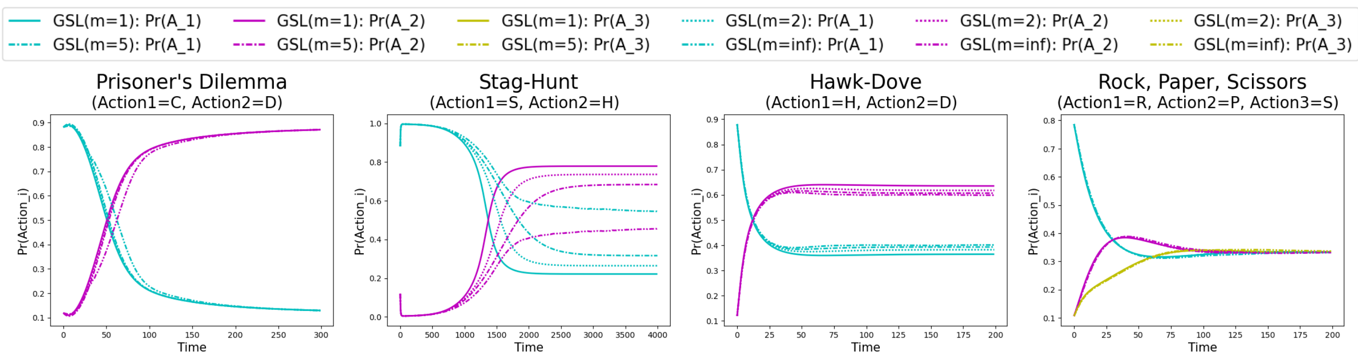


(a) $m \to \infty$, the initial Q-value of action 1 is 1 and the initial Q-values of other actions are 0 for all the agents

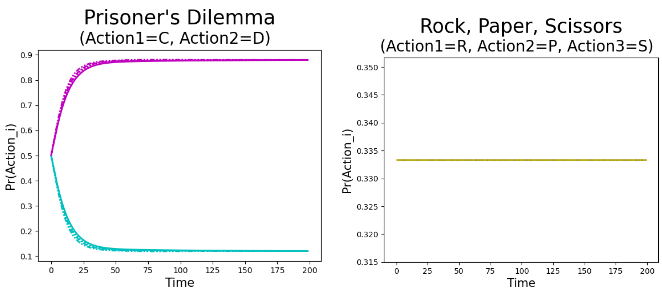


(b) $m \to \infty$, the initial Q-value of each action is 0 for all the agents

Figure 2: With a large value of m, comparison among the dynamics of average Q-values predicted by our model (solid line) and the previous model(dashed line), and the actual dynamics averaged over 100 runs of simulations (shaded line).

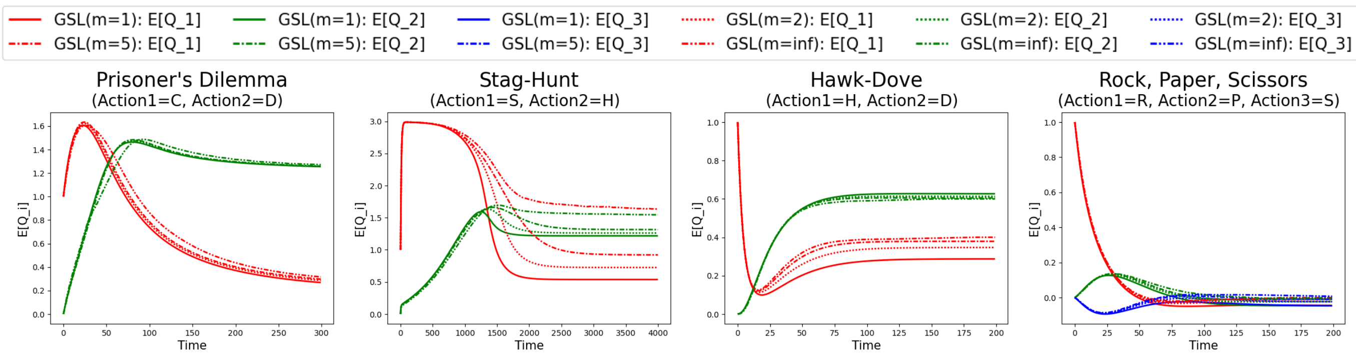


(a) the initial Q-value of action 1 is 1 and the initial Q-values of other actions are 0 for all the agents

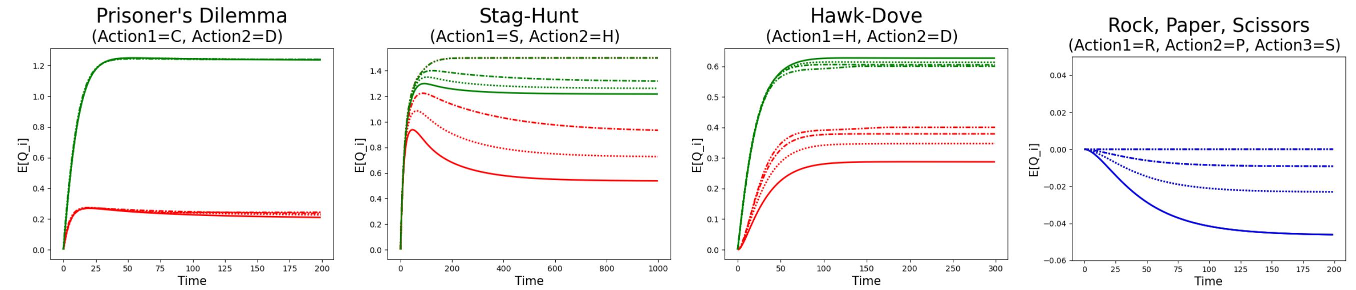


(b) the initial Q-value of each action is 0 for all the agents

Figure 3: The effects of local interactions and incomplete information on multiagent Q-learning. Our model shows that as the value of m varies, the population can stabilize at significantly different proportions of agents using each action.



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



(b) the initial Q-value of each action is 0 for all the agents

Figure 4: The effects of local interactions and incomplete information on multiagent Q-learning. Our model shows that as the value of m varies, the population will establish different Q-values dynamics.