Information Design in Multi-Agent Reinforcement Learning

Yue Lin, Wenhao Li, Hongyuan Zha, Baoxiang Wang

The Chinese University of Hong Kong, Shenzhen

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Existing literature on multi-agent reinforcement learning mainly focuses on fully cooperative scenarios. However, computational economics reveals a different aspect. It shows that an agent can influence other agents in a **mixed-motive** setting by leveraging its informational advantages.

`Recommendation Letter`. A bunch of students are about to enter the job market. Among them, $\frac{1}{3}$ are excellent and the remaining are weak. A professor can observe each student's quality, while the HR cannot (informational advantage); The professor can communicate with the HR (communication); The professor's goal is to get more students employed, while the HR wants to hire only excellent students (mixed-motive).

	HR		
	hire	not hire	
pro.	1, -1	0, 0	(if stu. is weak)
	1, 1	0, 0	(else)

- Professor reveals no information. Both agents get 0 reward.
- Professor honestly reports. Both agents get $\frac{1}{3}$ reward.
- Professor honestly reports for excellent students, and lies for weak students with a probability of $\frac{1}{2}-\epsilon$ $(0<\epsilon<\frac{1}{2})$. The professor gets $\frac{2}{3}-\frac{2}{3}\epsilon$ reward, and the HR gets $\frac{2}{3}\epsilon$ reward.

Information Design

For information design, the core insight is to send messages to change the posterior beliefs of the receiver, which persuades it to take actions that benefit the sender. The process of information design can be modeled as a linear constrained optimization problem

$$\max_{\varphi} \mathbb{E}_{\varphi}[\, r^i(s,a)\,], \quad \text{s.t.} \sum_s \mu_0(s)\cdot\varphi(a\mid s)\cdot\Big[\, r^j(s,a)-r^j(s,a')\,\Big]\geq 0, \forall a,a', \tag{1}$$

where μ_0 is the common prior belief, $r^i(s,a)$ and $r^j(s,a)$ are the utility functions of the sender and the receiver respectively, $\varphi(a\mid s)$ is the sender's signaling scheme. The constraints are named obedience, satisfying which the rational receiver will follow the sender's recommendations. Because

$$\begin{aligned}
&\sum_s \mu_0(s)\cdot\varphi(a\mid s)\cdot\Big(r^j(s,a)-r^j(s,a')\Big)\geq 0, && \forall a,a'\in A\\
\Leftrightarrow &\sum_s \frac{\mu_0(s)\cdot\varphi(a\mid s)}{\sum_{s'}\mu_0(s')\cdot\varphi(a\mid s')}\cdot\Big(r^j(s,a)-r^j(s,a')\Big)\geq 0, && \forall a,a'\in A\\
\Leftrightarrow &\sum_s \mu(s\mid a)\cdot r^j(s,a)\geq \sum_s \mu(s\mid a)\cdot r^j(s,a'), && \forall a,a'\in A
\end{aligned}$$

Markov Signaling Games

To formulate this communication problem, we proposed Markov signaling games (MSGs).

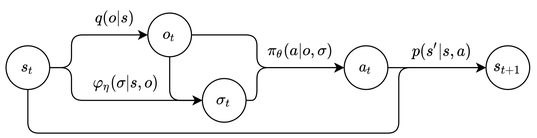
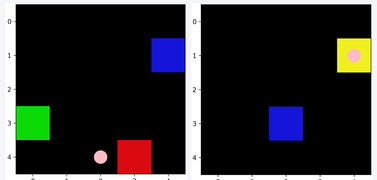


Figure 1. Illustration of the Markov signaling game. The arrows symbolize probability distributions, whereas the nodes denote the sampled variables.

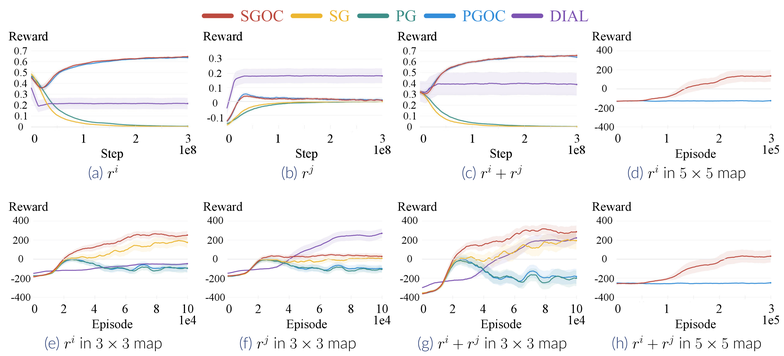
Experimental Environment: Reaching Goals



(a) Reaching Goals (b) The aligned goals

Figure 2. Maps 5×5 of `Reaching Goals`. The blue, red, and green squares represent the receiver, the sender's goal, and the receiver's goal, respectively. If the red square and the green square overlap, it will turn yellow, meaning that the goals of agents are aligned. The pink dots represent the messages sent by the sender. The sender is out of the map. Thus it can only get a reward when the receiver reaches the red goal.

Experiments



(a) r^i (b) r^j (c) r^i+r^j (d) r^i in 5×5 map

(e) r^i in 3×3 map (f) r^j in 3×3 map (g) r^i+r^j in 3×3 map (h) r^i+r^j in 5×5 map

Figure 3. Comparisons of the performance. (a-c) The results of `Recommendation Letter`. (d-h) The results of `Reaching Goals`. The rewards and penalties are amplified by 20 and 5 (12 and 3.5) respectively in 3×3 (5×5) map.

Value Functions and Bellman Equations in MSGs

- The sender's state value function: $V^i_{\varphi,\pi}(s)=\mathbb{E}_{\varphi,\pi}\left[G^i_t\mid s_t=s\right]$, where $G^i_t=\sum_{k=t}^{\infty}\gamma^{k-t}r^i_{k+1}$;
- The sender's signal value function: $Q^i_{\varphi,\pi}(s,\sigma)=\mathbb{E}_{\varphi,\pi}\left[G^i_t\mid s_t=s,\sigma_t=\sigma\right]$;
- The action value function: $U^i_{\varphi,\pi}(s,\sigma,a)=\mathbb{E}_{\varphi,\pi}\left[G^i_t\mid s_t=s,\sigma_t=\sigma,a_t=a\right]$;
- The marginal action value function: $W^i_{\varphi,\pi}(s,a)=\mathbb{E}_{\varphi,\pi}\left[G^i_t\mid s_t=s,a_t=a\right]=U^i_{\varphi,\pi}(s,\sigma,a)$.

- $V^i_{\varphi,\pi}(s)=\sum_o q(o\mid s)\sum_\sigma\varphi_\eta(\sigma\mid s)\sum_a\pi_\theta(a\mid o,\sigma)\cdot U^i_{\varphi,\pi}(s,\sigma,a)$;
- $U^i_{\varphi,\pi}(s,\sigma,a)=R^i(s,a)+\gamma\sum_{s'}p(s'\mid s,a)\cdot V^i_{\varphi,\pi}(s')$.

Signaling Gradient

The proposed signaling gradient is utilized to compute the gradient of the sender's long-term expected payoff w.r.t. its signaling scheme parameters. It explicitly takes into account the chain of the receiver's policy and thus alleviate the non-stationarity between agents.

Similar to the case of the policy gradient (PG), the relationship between state visitation frequency and (φ,π) cannot be explicitly written.

Lemma 4.1. Given a signaling scheme φ_η of the sender and an action policy π_θ of the receiver in an MSG $\mathcal{G}$, the gradient of the sender's value function $V^i_{\varphi,\pi}(s)$ w.r.t. the signaling parameter η is

$$\nabla_\eta V^i_{\varphi,\pi}(s)\propto\ \mathbb{E}_{\varphi,\pi}\left[W^i_{\varphi,\pi}(s,a)\cdot\left[\nabla_\eta\log\pi_\theta(a\mid o,\sigma)+\nabla_\eta\log\varphi_\eta(\sigma\mid s,o)\right]\right]. \tag{2}$$

This RL technique allows for organic and repeated interactions between far-sighted agents in a given environment, which lifts the commitment assumption in canonical information design.

Extended Obedience Constraints

As an analogous of (1), the prior of such information is then the occupancy measure $d_{\varphi,\pi}(s)$ of the state condition on the current signaling scheme and action policy. The payoff function $w^j(s,a)$ corresponds to the action value function $W^j_{\varphi,\pi}(s,a)$ in MSGs.

Lemma 4.2. Given a receiver's observation o, the extended obedience constraints (3) in MSGs yield the same optimum as the obedience constraints in (1).

$$\sum_s d_{\varphi,\pi}(s)\cdot\varphi_\eta(\sigma\mid s,o)\cdot\sum_a\left[\pi_\theta(a\mid o,\sigma)-\pi_\theta(a\mid o,\sigma')\right]\cdot W^j_{\varphi,\pi}(s,a)\geq 0, \tag{3}$$

for all $\sigma,\sigma'\in\Sigma$.

These extended constraints (denoted as $C_\varphi\left(\sigma,\sigma'\right)$) remove the revelation principle analysis from the obedience constraints, thereby reverting the sender's behavior from "action recommending" to "signal sending".

Solving the Constrained Optimization Problem in MSGs

The self-interested sender attempts to optimize its payoff expectation in an MSG while satisfying the extended obedience constraints. This optimization problem is

$$\max_\eta \mathbb{E}_{\varphi,\pi}\left[V^i_{\varphi,\pi}(s)\right],\ \text{s.t.}\quad C_\varphi(\sigma,\sigma')\geq 0,\quad \forall\sigma,\sigma'. \tag{4}$$

Since we are employing a learning-based approach, it is necessary to calculate the gradient $\nabla_\eta C_\varphi(\sigma,\sigma')$. The gradient is estimated using the biased sampling method as below.

$$\nabla_\eta\hat{C}_\varphi(\sigma,\sigma')=\frac{1}{T}\sum_{s_t\in\tau}\left[\sum_a\left(\pi_\theta(a\mid o_t,\sigma)-\pi_\theta(a\mid o_t,\sigma')\right)\cdot W^j_{\varphi,\pi}(s_t,a)\cdot\nabla_\eta\varphi_\eta(\sigma\mid s_t,o_t)\right], \tag{5}$$

where τ is a sampled trajectory with T timesteps, and σ' is randomly sampled, instead of being sampled from the signaling scheme.

Taking the Lagrangian method as an example, The update of the signaling scheme parameters $\eta^{(k)}$ for the k-th iteration is

$$\eta^{(k+1)}\leftarrow\eta^{(k)}+\nabla_\eta\mathbb{E}_{\varphi,\pi}\left[V^i_{\varphi,\pi}(s)\right]+\sum_{\sigma,\sigma'}\lambda_{\sigma,\sigma'}\cdot\nabla_\eta\left(\hat{C}_\varphi(\sigma,\sigma')\right)^-, \tag{6}$$

where $\lambda_{\sigma,\sigma'}$ denotes the non-negative Lagrangian multipliers (predefined as hyperparameters), and $(\cdot)^-=\min\{0,\cdot\}$.