Information Design in Multi-Agent Reinforcement Learning

Yue Lin, Wenhao Li, Hongyuan Zha, Baoxiang Wang

March 14, 2024

The Chinese University of Hong Kong, Shenzhen

Outline

Introduction

Reinforcement Learning

Reinforcement learning (RL) studies how a world-agnostic agent makes sequential decisions to maximize its utility.

- The setting of stationary world is ideal
- In real tasks other agents in the environment have their own goals and behave adaptively to the ego agent
- To thrive, one needs to influence other agents

Breakout from Atari 2600: stationary reward, stationary environment

Nonstationary, Multi-Agent Settings

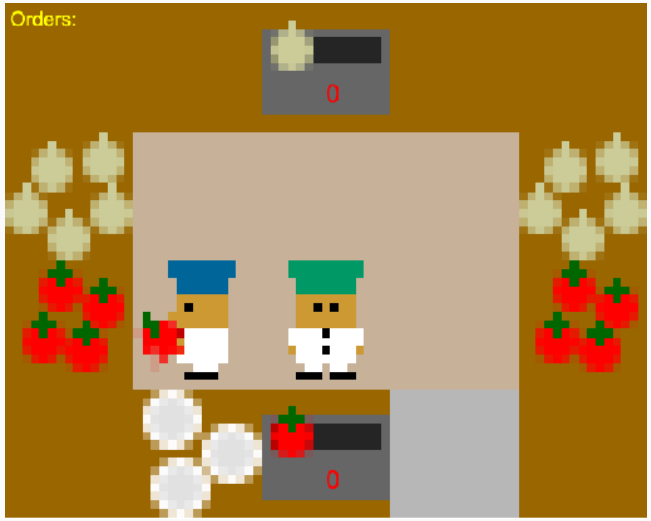
Multi-agent reinforcement learning (MARL) investigates the interaction and influence among **multiple** rational RL agents

Easy settings:

- Fully cooperative, towards consentaneous goal. No influence needed
- Two-player zero-sum. No influence possible.

Hard, much less charted settings:

- Almost all mixed-motive games.



Overcooked. An example of fully cooperative AI.

Influencing Others

Studies in computational economics have distilled two types (and two types only) of ways to directly influence a rational, self-interested agent

- **Mechanism Design** Providing tangible goods to influence the receivers' learning processes.
  Relatively easy: Reward does not affect trajectory. Reward is compulsory.
- **Information Design** Sending messages to change the receivers' posterior beliefs.
  Relatively hard: Information immediately changes transitions. Information can be ignored.

Setting of this work

There are agents i, j with observations o^i, o^j, respectively.

- *Communication*: There is a message channel between them
- *Mixed-motive*: They are rational, self-interested, towards different but not zero-sum goals
- *Informational advantage*
  - $o^i - o^j \neq \varnothing$
  - $o^i - o^j$ affects the receiver's payoff expectation

Example: 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)
- The HR knows the strategy of the professor (commitment)

Example: Recommendation Letter

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

Three examples in Recommendation Letter:

- 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 $\frac{1}{2} - \epsilon$ (for some $0 < \epsilon < \frac{1}{2}$)
  - The professor gets $\frac{2}{3} - \frac{2}{3}\epsilon$ reward;
  - The HR gets $\frac{2}{3}\epsilon$ reward.

Information Design

The insight of information design is to maximize the sender's payoff expectation, while subjecting to the incentive compatibility (IC, a.k.a. OC/obedience constraint) for the receiver:

$$\begin{aligned} \max_{\varphi} \quad & \mathbb{E}_{\varphi}[\ W^i(s,a)\ ] \\ \text{s.t.} \quad & \sum_{s} P(s)\cdot\varphi(a\mid s)\cdot\left[\ W^j(s,a)-W^j(s,a')\ \right]\geq 0, \quad \forall a,a'. \end{aligned} \tag{1}$$

W^i, W^j are payoff functions. $P(s)$ is the prior probability (e.g. students excellent/weak). $\varphi(a|s)$ is the *public* stochastic signaling scheme (e.g. recommendation).

The *revelation principle* proves that there is an optimal signaling scheme that uses a signal space of the same size as the action space of the receiver. Therefore a in $\varphi(a\mid s)$ is both signal/action.

Incentive Compatibility

A signaling scheme φ is a Bayes correlated equilibrium (BCE) of the game if it satisfy the obedient constraints. And in this way the rational receiver has no incentive to deviate from the sender's recommendation:

$$\sum_s \mu_0(s) \cdot \varphi(a \mid s) \cdot \Big(r^j(s,a) - r^j(s,a') \Big) \geq 0, \qquad \forall a' \in A$$

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

$$\Leftrightarrow \sum_s \mu(s \mid a) \cdot \Big(r^j(s,a) - r^j(s,a') \Big) \geq 0, \qquad \forall a' \in A$$

$$\Leftrightarrow \sum_s \mu(s \mid a) \cdot r^j(s,a) \geq \sum_s \mu(s \mid a) \cdot r^j(s,a'), \qquad \forall a' \in A$$

Commitment Assumption

In Bayesian persuasion, the sender will **commit** to a signaling scheme first.

- The sender will determine its signaling scheme before the game starts and publish such scheme.
- In a repeated game where a long-term sender interacts with a sequence of short-term receivers, the commitment will naturally emerge in equilibria. This is due to the sender's need to establish its reputation for credibility.
- (Instead, RL allows for organic and repeated interactions between senders and receivers in a given environment, more closely resembling real-world scenarios.)

Challenges in Learning Information Design

- **Non-stationarity.** As the sender's signaling scheme is updated, the receiver's environment also changes.
  - Especially problematic in mixed-motive scenarios
- **Effect on episode generation.** Unlike incentive design, the communication will affect the episode generation phase of RL.
- **Persuasiveness.** The sender should provide information that the receivers are willing to respect.

Method

Markov Signaling Games

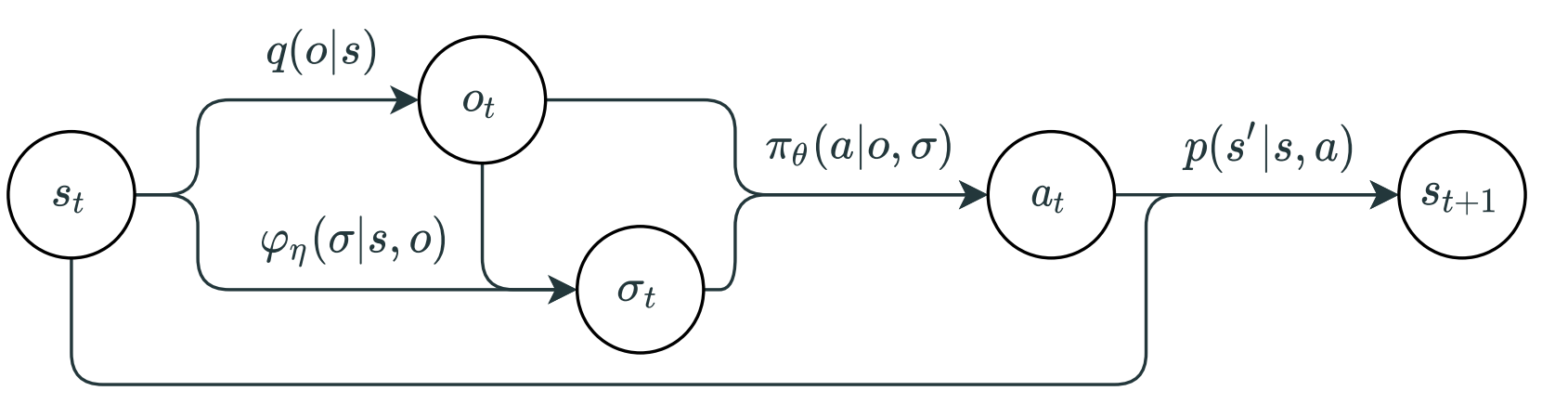


Figure 1: An illustration of a Markov signaling game. The arrows symbolize probability distributions, whereas the nodes denote the sampled results.

s: state; o: observation; σ: signal; a: action;

Extensions of MSGs: The Sender's Action

- The sender i chooses actions $a^i \in A^i$ according to its policy $\pi^i_{\theta^i} : S \times \Sigma \to \Delta(A^i)$.
- Notably, the sender's action policy considers the signals it sends to the receivers in the same round.
- This is necessary to enable the adaptation to a variety of receiver responses induced by the dispatched signals.

Extensions of MSGs: Partial Observibility

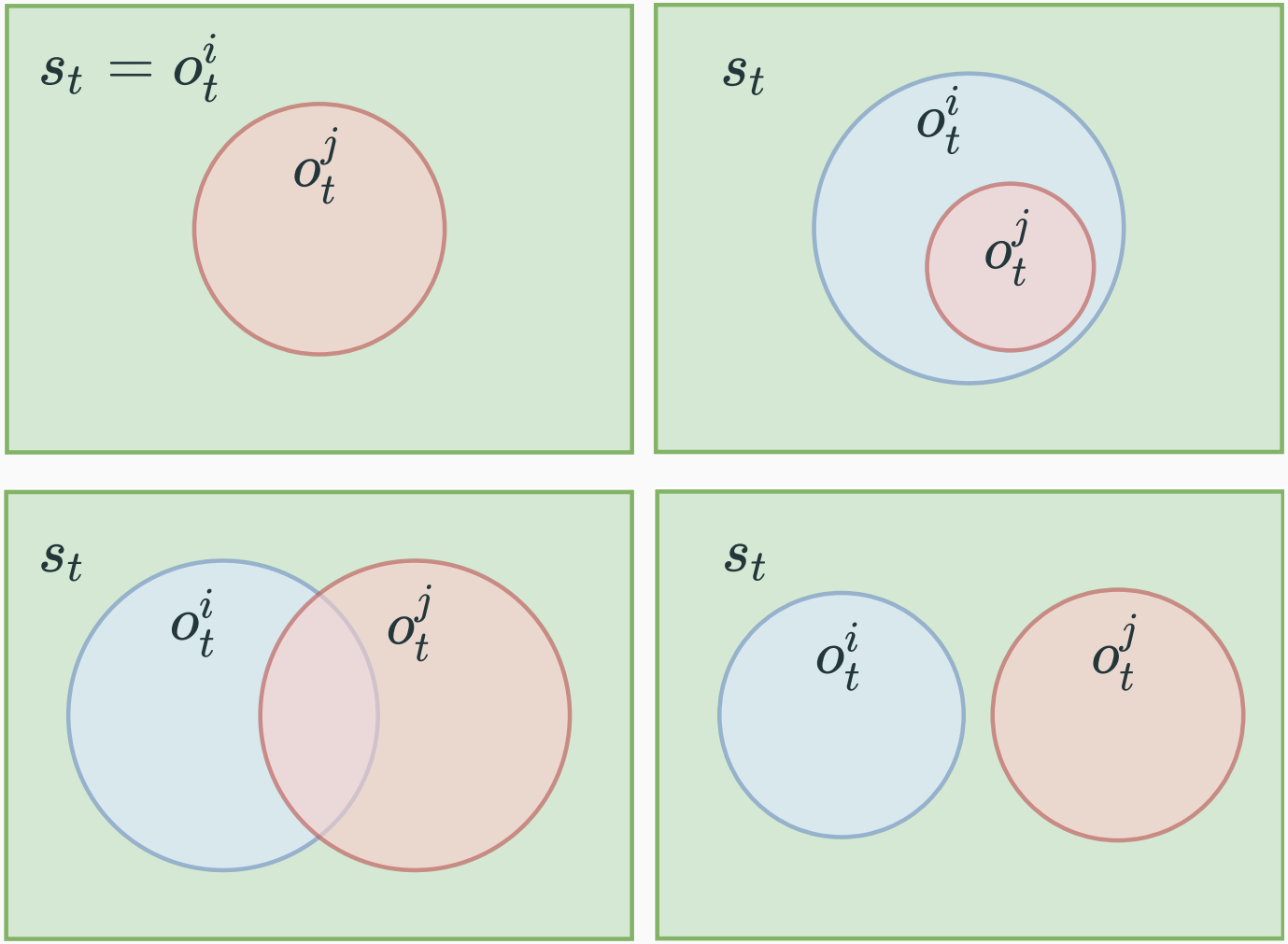


Figure 2: Sender i has informational advantage over receiver j. The information advantage is reflected by $o_t^i - o_t^j$. The receiver's observation in standard MSGs is turned to $o_t^i \cap o_t^j$ in every case.

Value Functions in MSGs

Three value functions in MSG: the state value function $V(s)$, the signal value function $Q(s, \sigma)$, the action value function $W(s, a)$.

- A new value function:
  $U^i_{\varphi,\pi}(s, \sigma, a) = \mathbb{E}_{\varphi,\pi}\left[G^i_t \mid s_t = s, \sigma_t = \sigma, a_t = a\right].$
- No direct costs are associated with the sender transmitting signals; The signals do not impact state transitions:
  $W^i_{\varphi,\pi}(s, a) = U^i_{\varphi,\pi}(s, \sigma, a).$
- Bellman Equation:
  $V^i_{\varphi,\pi}(s) = \sum_o q(o \mid s) \sum_\sigma \varphi_\eta(\sigma \mid s) \sum_a \pi_\theta(a \mid o, \sigma) \cdot U^i_{\varphi,\pi}(s, \sigma, a).$

Signaling Gradient

The introduction of communication φ further complicates this process as it involves deriving $\nabla_\eta d_{\varphi,\pi}(s)$.

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

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

Signaling Gradient v.s. Policy Gradient

Signaling Gradient

$$\begin{aligned}\nabla_\eta V^i_{\varphi,\pi}(s) \propto \mathbb{E}_{\varphi,\pi}\left[W^i_{\varphi,\pi}(s,a)\nabla_\eta \log \pi_\theta(a \mid o,\sigma)\right] \\ + \mathbb{E}_{\varphi,\pi}\left[W^i_{\varphi,\pi}(s,a)\nabla_\eta \log \varphi_\eta(\sigma \mid s)\right]\end{aligned} \tag{3}$$

Policy Gradient

$$\nabla_\eta V^i_{\varphi,\pi}(s) \propto \ \mathbb{E}_{\varphi,\pi}\left[Q^i_{\varphi,\pi}(s,\sigma)\cdot\nabla_\eta \log \varphi_\eta(\sigma \mid s)\right] \tag{4}$$

Vanilla policy gradient will be independent of the actions taken by the receivers and is therefore biased.

Extended Obedience Constraints

Lemma
Given a joint observation o, the obedience constraints in MSGs are

$$\sum_{s} d_{\varphi,\pi}(s) \cdot \varphi_{\eta}(\sigma \mid s) \cdot \sum_{a} \Big[\pi_{\theta}(a \mid o, \sigma) - \pi_{\theta}(a \mid o, \sigma')\Big] \cdot W^{j}(s, a) \geq 0, \tag{5}$$

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

For convenience, the left-hand side of (5) is denoted as $C^{j}_{\varphi}(\sigma, \sigma')$.

In the learning context, information revelation induces *want of dictatorship*. Such want does not build trust and respect between sender and receiver and instead, and inevitably drives them to the equilibrium where signals are arbitrary and are nevertheless ignored. Counter-intuitively, the revelation principle should be removed under the learning context.

Constrained Optimization

Information design in an MSG can be formalized as a constrained optimization problem:

$$\begin{aligned} \max_{\eta} \quad & \mathbb{E}_{\varphi,\pi}\left[V^i(s)\right] \\ \text{s.t.} \quad & C^j_\varphi(\sigma,\sigma') \geq 0, \quad \forall j,\sigma,\sigma'. \end{aligned} \tag{6}$$

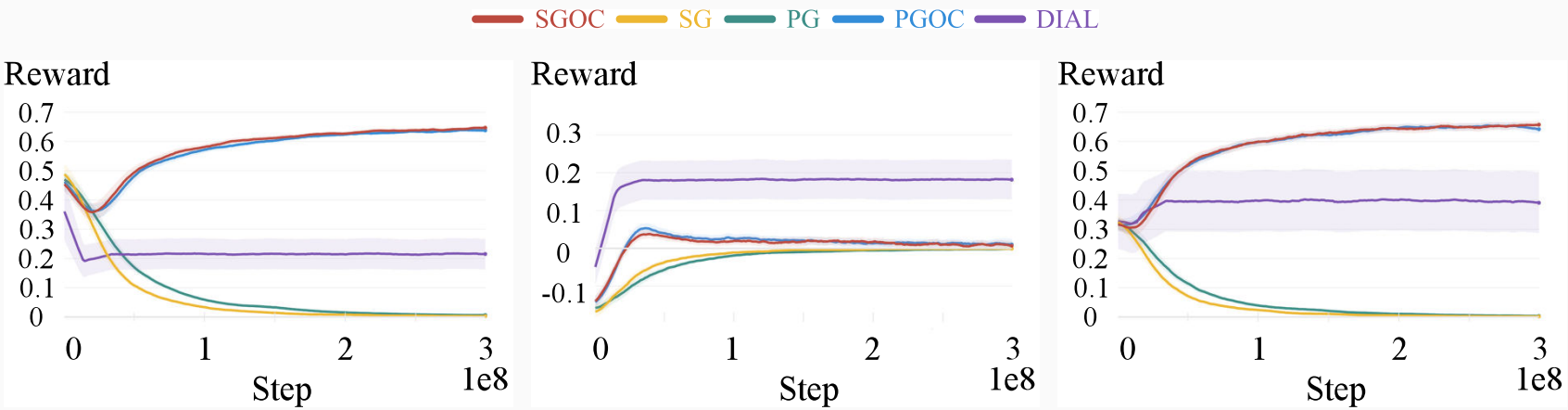
There are various methods available to solve this constrained optimization problem iteratively, e.g. the Lagrangian method.

$$\eta^{(k+1)} \leftarrow \eta^{(k)} + \nabla_\eta \mathbb{E}_{\varphi,\pi}\left[V^i(s)\right] + \sum_{j,\sigma,\sigma'} \lambda(\sigma,\sigma') \cdot \nabla_\eta \Big(C^j_\varphi(\sigma,\sigma') \Big)^-, \tag{7}$$

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

Experiments

Results on Recommendation Letter



(a) The sender's reward **(b)** The receiver's reward **(c)** Social welfare

Figure 3: Comparison of performance in the Recommendation Letter experiments. (a) The sender's rewards along the training process. (b) The receiver's rewards along the training process. (c) Social welfare is defined as the sum of the sender's reward and the receiver's reward.

Symmetricity of the Signaling Schemes

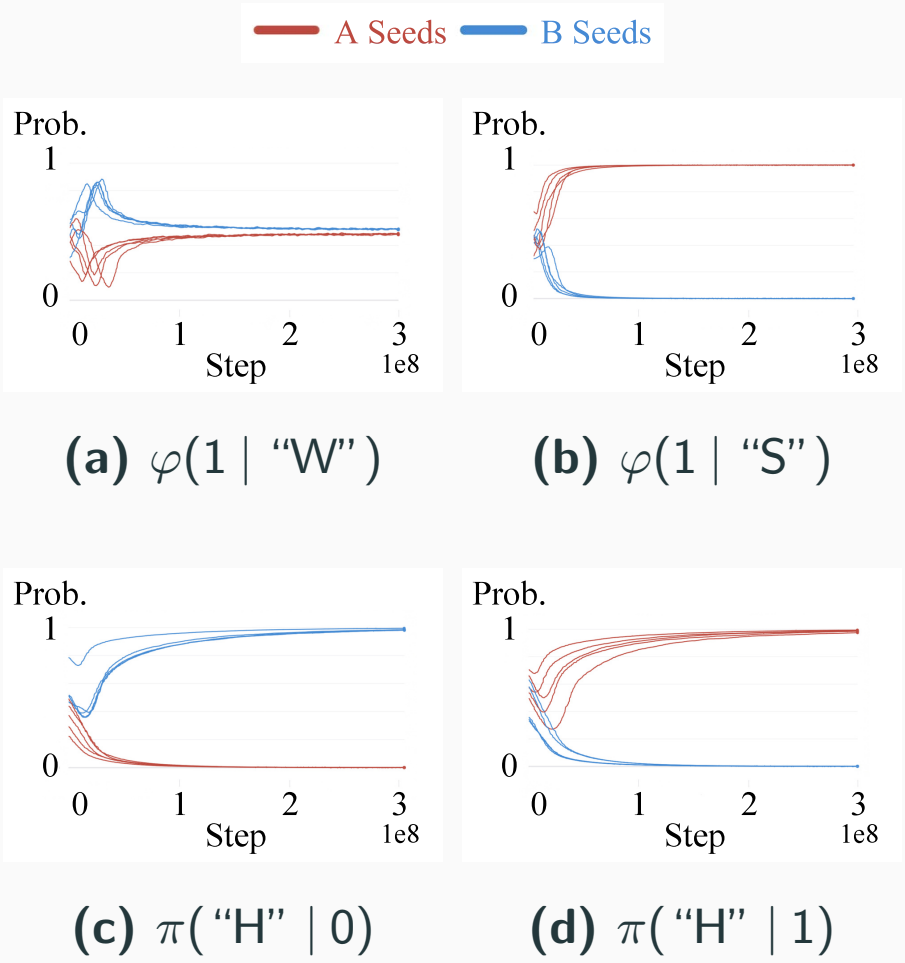
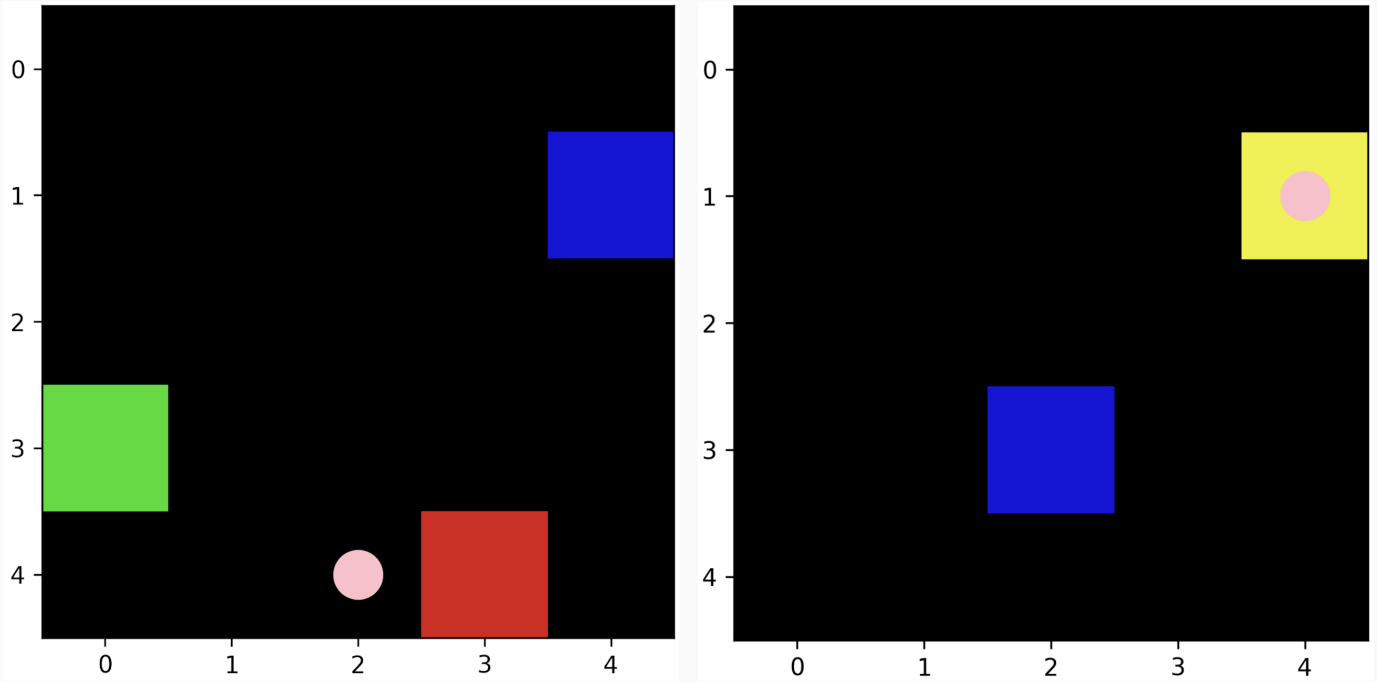


Figure 4: (a) The prob. of signaling 1 for **W**eak students. (b) The prob. of signaling 1 for **S**trong students. (c) The prob. of choosing to **H**ire when signaled 0. (d) The prob. of choosing to **H**ire when signaled 1.

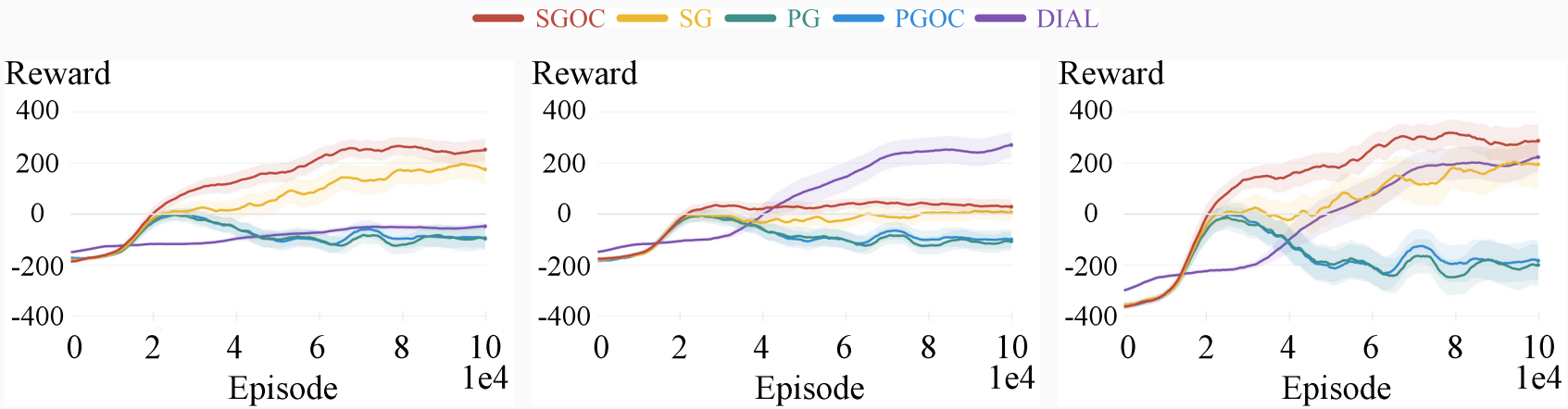
Reaching Goals



(a) Reaching Goals **(b)** The aligned goals

Figure 5: Maps 5×5 of Reaching Goals. The receiver (blue square) in the map are going to reach goals. The sender gets a reward of 30 if the receiver reaches the red square. The receiver gets a reward of 30 if it reaches the green square. The pink dot is the sent message, and the yellow square shows up when the red and green squares overlap.

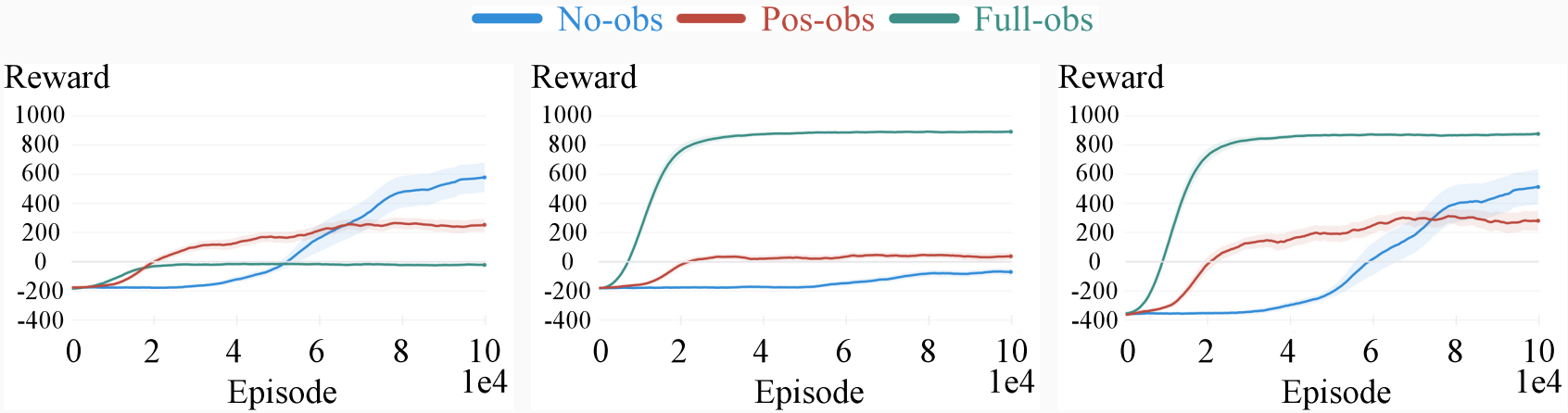
Results on Reaching Goals



(a) The sender's reward **(b)** The receiver's reward **(c)** Social welfare

Figure 6: Comparisons of performance in the Reaching Goals experiments. Once the receiver reaches a goal, the corresponding agent will receive a reward of 20. And the distance penalties are amplified 50-fold. (a) The sender's rewards along the training process. (b) The receiver's rewards along the training process. (c) Social welfare is defined as the sum of the sender's reward and the receiver's reward.

Different Observation of the Receiver



(a) The sender's reward **(b)** The receiver's reward **(c)** Social welfare

Figure 7: Performance comparisons in `Reaching Goals`. Once the receiver reaches a goal, the corresponding agent will receive a reward of 20. And the distance penalties are amplified 50-fold. (a) The sender's rewards along the training process. (b) The receiver's rewards along the training process. (c) Social welfare is defined as the sum of the sender's reward and the receiver's reward.

Future Works

- **Partial Observability of the Sender.** In some scenarios, there is information that the receivers know but the sender does not know. The sender needs to estimate it.
- **Hyper Gradient for Signaling Gradient.** Similar to the LIO, the second-order gradients can also be computed. The influence of the hyper gradient is left for future work.
- **Multiple Senders.** There co-exist the Stackelberg game between multiple senders and the Stackelberg game between senders and receivers. The game between senders needs to be formalized and analyzed.

Thank You!

Questions are very welcome!

Paper: https://arxiv.org/abs/2305.06807
Code: https://github.com/YueLin301/InformationDesignMARL

Baoxiang Wang (bxiangwang@gmail.com)
The Chinese University of Hong Kong, Shenzhen