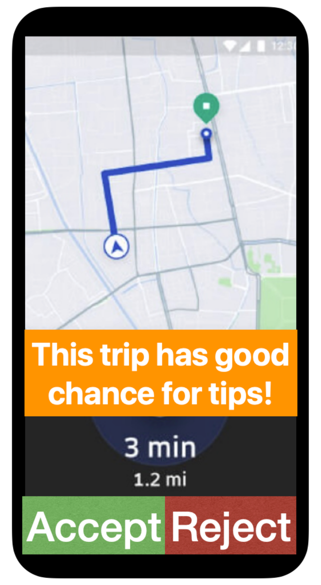
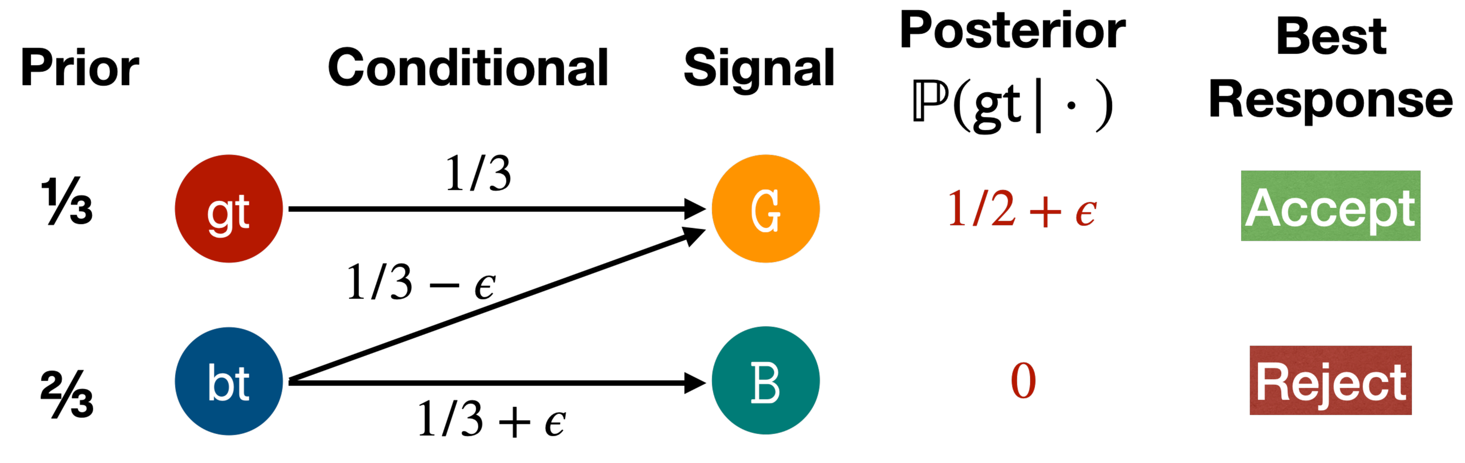
Sequential Information Design: *Markov Persuasion Process and Its Efficient Reinforcement Learning*

Jibang Wu (UVA), Zixuan Zhang (USTC), Zhe Feng (Google), Zhaoran Wang (Northwestern), Zhuoran Yang (Yale), Michael I. Jordan (Berkeley), Haifeng Xu (UVA→UChicago)

A Real-world Example of Sequential Information Design: Ride-hailing Platform





The sequential nature of this problem:

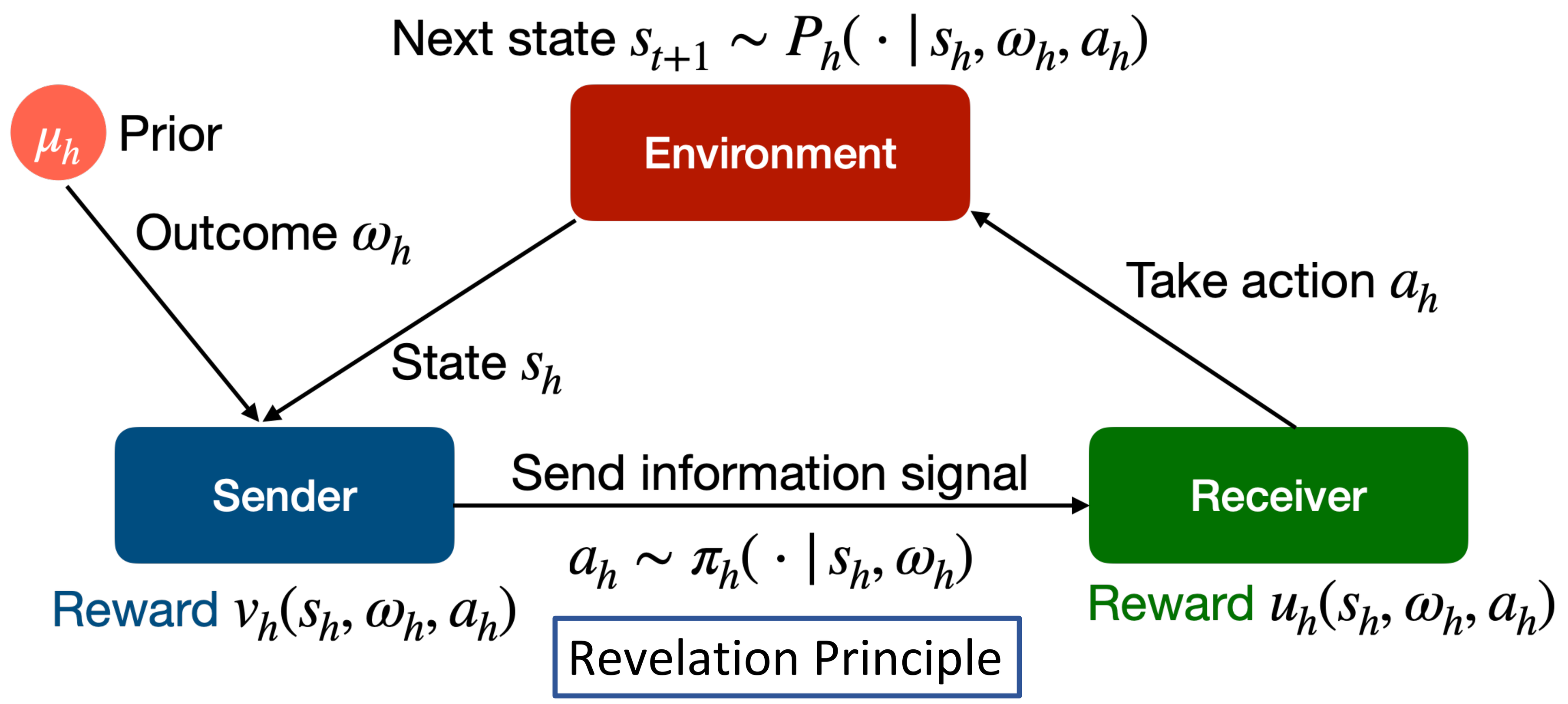
State Transition Driver supply and rider demand are affected by decisions.

Cumulative Utility Platform optimizes for the welfare in the long run.

This necessitates the combination of information design and planning.

Full Paper

MPP: the Markov Persuasion Process



Bellman equations for MPPs

Optimal signaling policy $\pi^\star = \{\pi_h^\star\}$ is greedy

w.r.t. each $Q_h^\star$ recursively defined as,

State-action function: $Q_h^\star(s,a) = v_h(s,a) + \mathbb{E}_{s_{h+1}\sim P}\left[V_{h+1}^\star(s_{h+1})\right]$

Value function: $V_h^\star(s) = \max_{\pi_h} \mathbb{E}_{\omega\sim\mu_h, a\sim\pi_h}\left[Q_h^\star(s,\omega,a)\right]$

with π_h subject to persuasiveness constraint $\pi_h \in \mathtt{Pers}(\mu_h, u_h)$

This recursion can still be solved efficiently by DP from $V_{H+1}^\star(s) = 0$.

RL via Optimism-Pessimism Principle

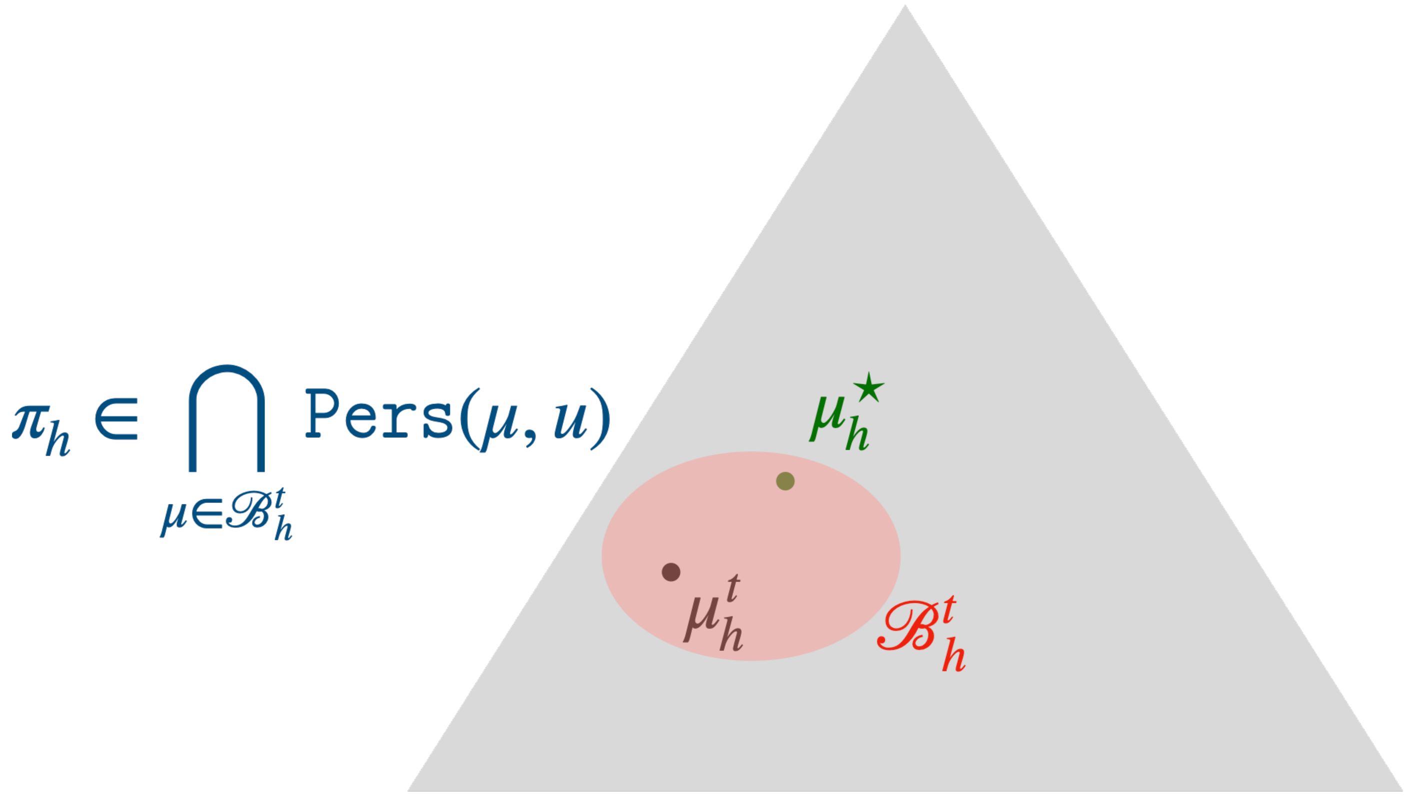
Pessimism to ensure *persuasiveness*

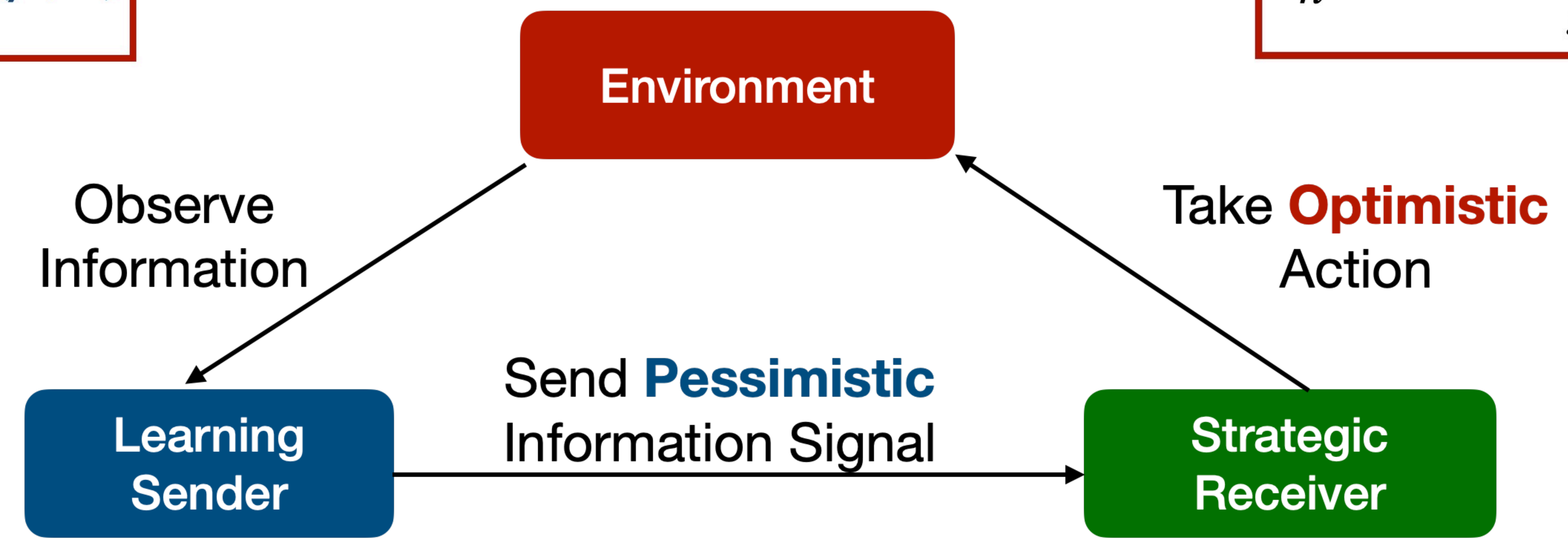
$$\pi_h^t \leftarrow \arg\max_{\pi_h} \mathbb{E}_{\mu_h,\pi_h}\left[Q_h^\star(s,\omega,a)\right] \text{ s.t. } \pi_h \in \cap_{\mu\in\mathscr{B}_h^t} \mathtt{Pers}(\mu,u)$$

Optimism to ensure *exploration*

$$\pi_h^t \leftarrow \arg\max_{\pi_h} \mathbb{E}_{\mu_h,\pi_h}\left[\tilde{Q}_h^t(s,\omega,a)\right] \text{ over } \cap_{\mu\in\mathscr{B}_h^t} \mathtt{Pers}(\mu,u)$$

$$\pi_h \in \bigcap_{\mu\in\mathscr{B}_h^t} \mathtt{Pers}(\mu,u)$$





$$\mathtt{Regret}(T) = \tilde{O}\left(\sqrt{d^3} \cdot H^2 \cdot \sqrt{T}\right)$$