~~Persuading a Learning Agent~~
Generalized Principal-Agent Problem with a Learning Agent

Tao Lin,
Yiling Chen

Generalized Principal-Agent Problem

A general model for principal-agent interactions [1][2]. It includes:

- Contract design
- Bimatrix Stackelberg games
- Bayesian persuasion
- …

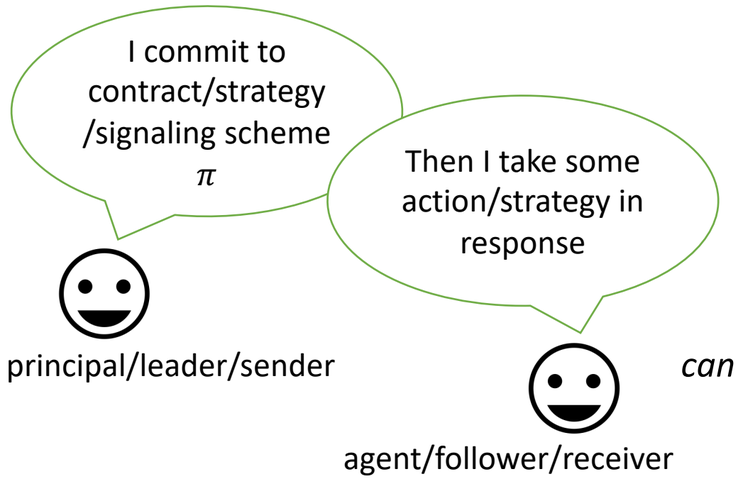
Formally,

- The principal has a convex decision space $\mathcal{X}$
- The agent has a finite action set A
- Principal's utility $u: \mathcal{X} \times A \to \mathbb{R}$
- Agent's utility $v: \mathcal{X} \times A \to \mathbb{R}$
- u, v are assumed to be linear in $x \in \mathcal{X}$
- Additionally, there is a finite set of signals/messages S

Timeline:

- The principal commits to a strategy $\pi = (p_s, x_s)_{s \in S}$, which consists of:
  - $(p_s)_{s \in S}$: a distribution over S
  - x_s: a decision associated with signal s
- The agent chooses a strategy $\rho: S \to A$
  - Best response:
    $\rho(s) \in \operatorname{argmax}_{a \in A} v(x_s, a)$
- The two players obtain (expected) utility
  $\mathbb{E}_{s \sim p}\left[u\left(x_s, \rho(s)\right)\right]$
  $\mathbb{E}_{s \sim p}\left[v\left(x_s, \rho(s)\right)\right]$

[1] Myerson (1982): Optimal Coordination Mechanism in Generalized Principal-Agent Problems.
[2] Gan, Han, Wu, Xu (2024): Generalized Principal-Agency: Contracts, Information, Games and Beyond.
[3] Deng, Schneider, Sivan (2019): Strategizing Against No-Regret Learners.



- In contract design, $x = (p_1, \dots, p_n)$ is a payment vector
- In bimatrix Stackelberg game, $x \in \mathcal{X} = \Delta([n])$ is a mixed strategy
- In Bayesian persuasion, $x \in \mathcal{X} = \Delta(\Omega)$ is a posterior belief

Additionally, π satisfies a constraint:

$\sum_{s \in S} p_s x_s \in$ convex set $\mathcal{C} \subseteq \mathcal{X}$.

In Bayesian persuasion, $\mathcal{C} = \{\mu_0\}$ where μ_0 is the prior for the state of the world.

Learning Agent

Instead of best-responding, we consider an agent who *learns* which action to take for each signal.

T rounds of interactions. In each round t,

- The principal chooses a strategy $\pi^t = (p_s^t, x_s^t)_{s \in S}$, can be *unknown* to the agent.
- Based on history, the agent chooses a (randomized) strategy $\rho^t: S \to \Delta(A)$.
- A signal $s^t \sim p^t$ is realized, the principal makes decision $x^t = x_{s^t}^t$, the agent samples action $a^t \sim \rho^t(s^t)$. The two players obtain (expected) utility
  $\mathbb{E}\left[u(x^t, a^t)\right]$ and $\mathbb{E}\left[v(x^t, a^t)\right]$

Definition: The agent's learning algorithm satisfies

- **No-contextual-regret** if
  $$\forall d: S \to A, \quad \mathbb{E}\Big[\sum_{t=1}^{T}\Big(v(x^t, d(s^t)) - v(x^t, a^t)\Big)\Big] \le \mathrm{CReg}(T) = o(T).$$
- **No-contextual-swap-regret** if
  $$\forall d: S \times A \to A, \quad \mathbb{E}\Big[\sum_{t=1}^{T}\Big(v(x^t, d(s^t, a^t)) - v(x^t, a^t)\Big)\Big] \le \mathrm{CSReg}(T) = o(T).$$

Main Results: Under some regularity conditions,

- Against a **no-contextual-regret** learning agent, the principal can obtain average utility at least $U^* - O\left(\sqrt{\frac{\mathrm{CReg}(T)}{T}}\right)$; U^* is the principal's optimal utility against a best-responding agent.
- Against a **no-contextual-swap-regret** learning agent, the principal *cannot* obtain more than $U^* + O\left(\frac{\mathrm{CSReg}(T)}{T}\right)$ (even knowing the agent's learning algorithm and using adaptive strategies).
- For *some* **no-contextual-regret** agent, the principal can obtain more than $U^* + O\left(\frac{\mathrm{CReg}(T)}{T}\right)$.

Intuition: no contextual **swap**-regret learning $\approx$ *approximately* best responding $\approx$ best responding