Incentive Design in Dynamic Systems

Vijay Gupta

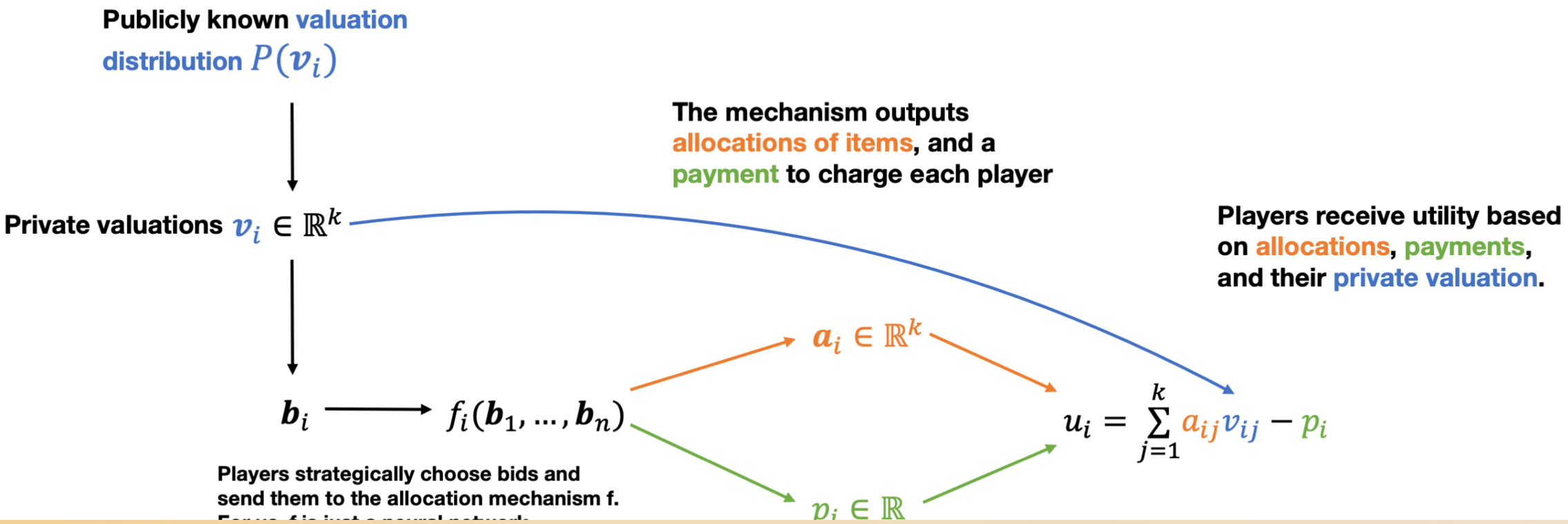
Purdue University

Joint work with Pranoy Das, Mostafa Abdelnaby, Shivam Bajaj, Yagiz Savas, Ufuk Topcu

AFOSR PI Meeting, Fall 2024

Problem Considered

- Incentive / contract / mechanism / auction design (including sequential incentives) has a long history



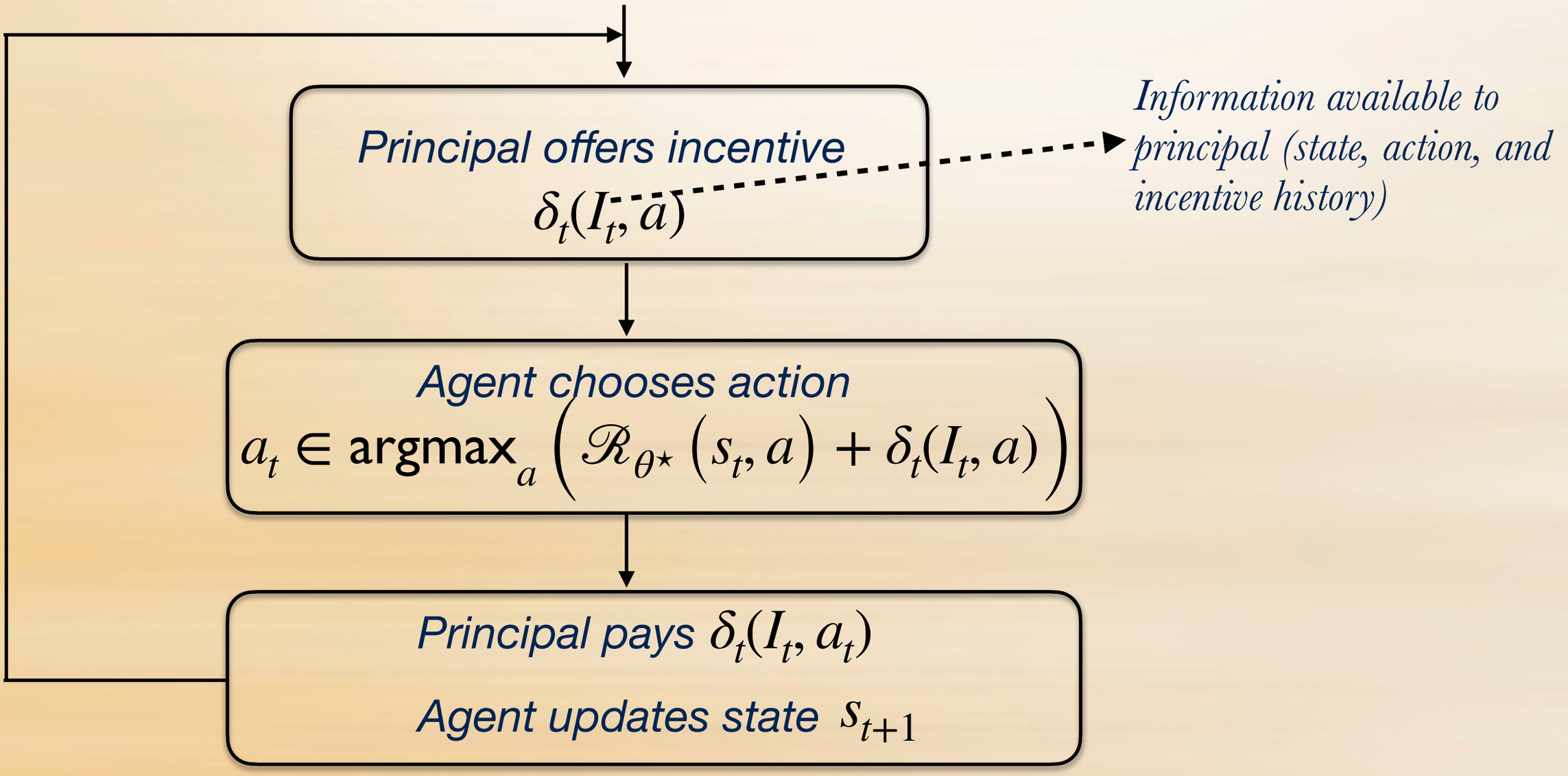
- Can we use such methods for coordinating behavior in dynamic systems?
  - Complexity of optimal contract design
  - Design of strategies for participants in resulting Markov Stackelberg games
  - Encoding objectives such as stability or robustness (as opposed to the system operator being interested in social welfare)?

Complexity of Sequential Incentive Design

Consider an agent whose behavior is modeled as an MDP with a reward function as a function of a (possibly hidden) type:

$$\mathscr{R}_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

The principal knows the reward function for each type, but not the true type.



Problem Considered

For an MDP $\mathscr{M}$, a set B of target states, and a set Θ of possible agent types, synthesize an incentive sequence $\Gamma \in \Xi(\mathscr{M})$ that leads the agent $\theta^\star$ to a target state with maximum probability at minimum expected cost, i.e.,

$$\min_{\gamma \in \Gamma(\mathscr{M})} \max_{\theta \in \Theta} \mathbb{E}^{\pi^\star}\left[\sum_{t=1}^{\infty} \delta_t(I_t, A_t) \,\middle|\, \theta\right]$$

Incentive policy

$$\text{subject to: } \pi^\star = (d_1, d_2, d_3, \ldots)$$

$$\forall t \in \mathbb{N}, \forall s \in S,\ d_t(s) \in \arg\max_{a \in \mathscr{A}} \left[\mathscr{R}_{\theta^\star}(s,a) + \delta_t(I_t, a)\right]$$

$$\Pr^{\pi^\star}_{\mathscr{M}}(Reach[B]) = \max_{\pi \in \Pi(\mathscr{M})} \Pr^{\pi}_{\mathscr{M}}(Reach[B])\,.$$

Desired final state

- Myopic agent (could be relaxed to lookout over finitely many steps)
- Principal knows the MDP

Complexity Results

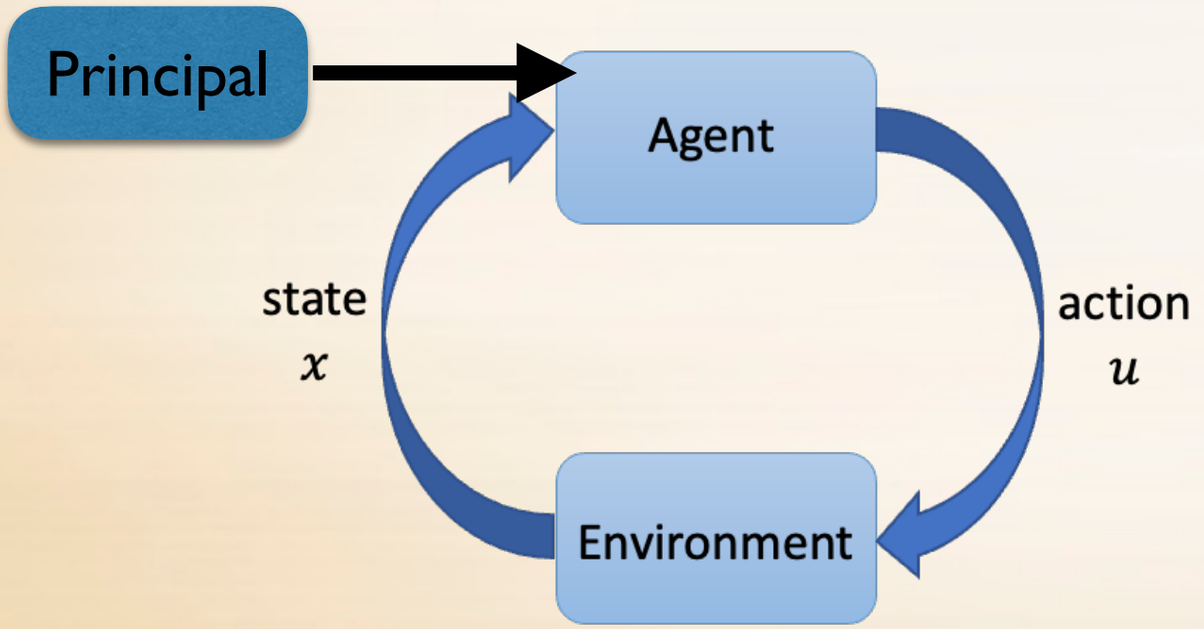
Problem	Complexity	Globally optimal solution
Behavior modification (BMP)	PSPACE-hard	—
Non-adaptive behavior modification (N-BMP)	NP-complete	MILP
Non-adaptive single-action behavior modification (NS-BMP)	NP-complete	MILP
Behavior modification of a dominant type (BMP-D)	P	LP

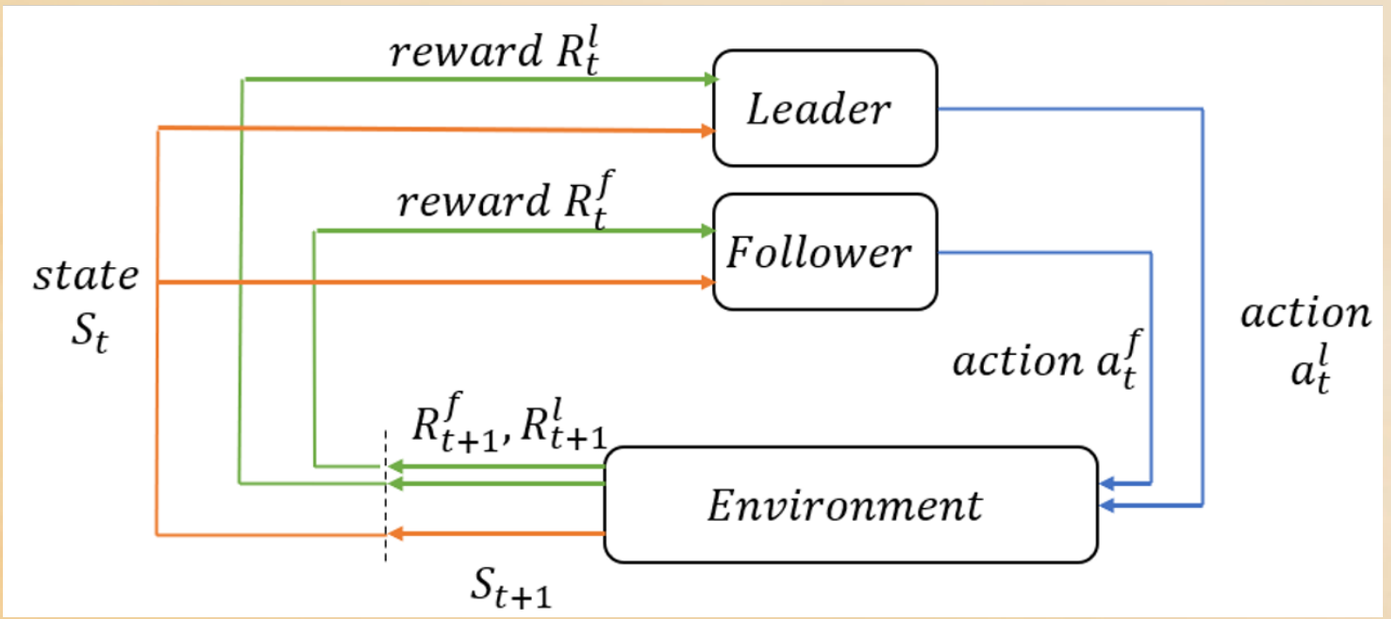
- The problem can be relaxed to finding approximately optimal solutions
- Feasible solutions can be constructed in a time polynomial in the size of the MDP
- Proof follows reduction of PSPACE-complete quantified satisfiability problem (QSAT) to BMP (similar to reduction of QSAT to POMDP optimal control problems)
- Motivates looking at learning-based solutions to the problem

Savas, Topcu, Gupta, TAC 2024

Agents Controlling a Dynamical System

- More generally, we can consider the case when both principal and agent can affect the state of the system directly

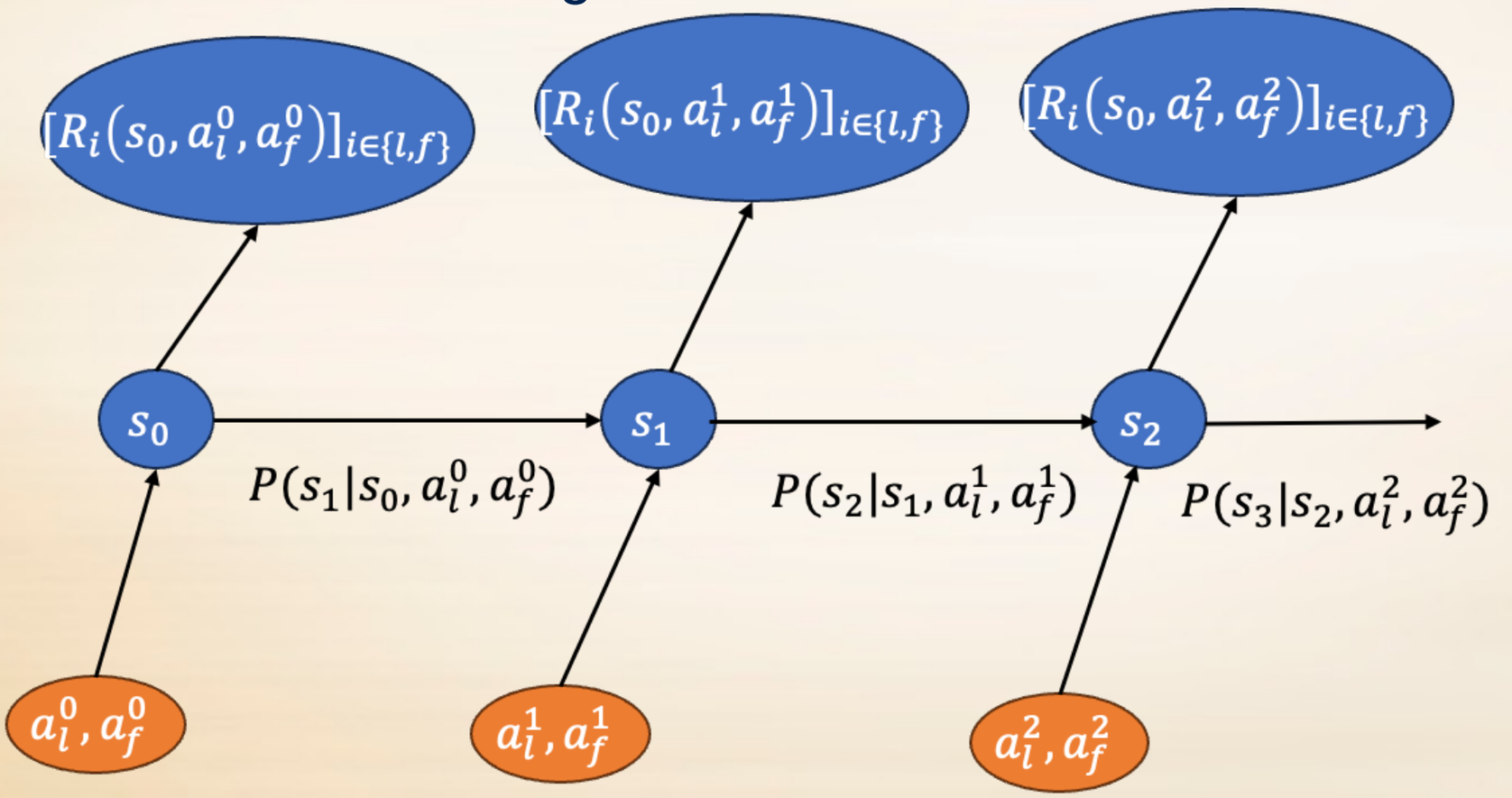




Still a Markov Stackelberg Game

- Stackelberg structure where incentive designer is the leader

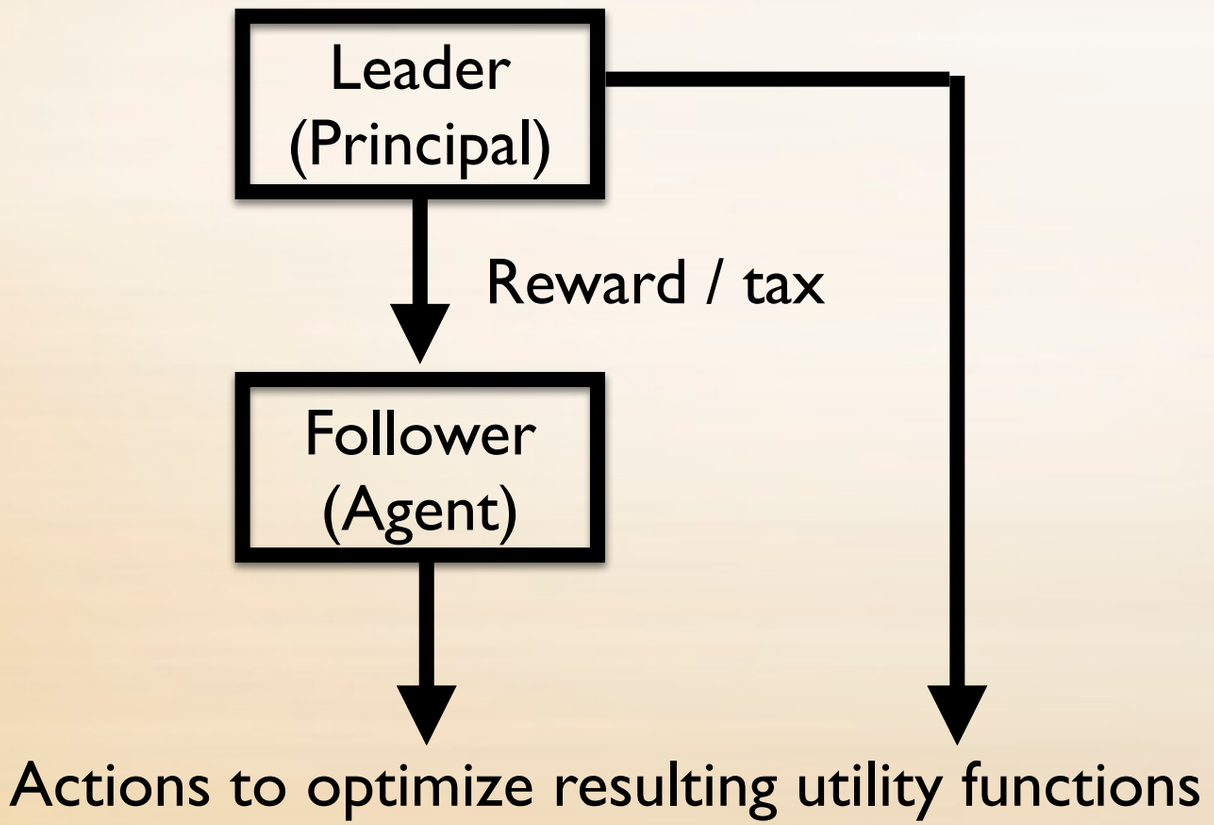
Leader announces policy before game starts



- Markov game (with Nash equilibrium among players in response to given incentive)

Learning in a Stochastic Stackelberg Game

- Incentive design seeks to change the utility of an agent to elicit a desired response



- Only recently have learning algorithms been identified even in a repeated static game setting (Fiez et al 2020)

- For stochastic setting, algorithms known only for special cases (follower plays deterministic strategy (Vorobeychik and Singh 2012), linear Markov games (Zhong et al 2021), zero-sum games (Goktas et al 2022, Metz et al 2016) …)

Value Functions

Value Functions:

$$\bar{V}_l(\pi_l, \pi_f)(s) := \mathbb{E}_{\mathcal{T}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathcal{R}_l(s^\tau, a_l^\tau, a_f^\tau) | s^0 = s\right]$$

$$\bar{V}_f(\pi_l, \pi_f)(s) := \mathbb{E}_{\mathcal{T}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathcal{R}_f(s^\tau, a_l^\tau, a_f^\tau) | s^0 = s\right]$$

$$\bar{V}_f^\lambda(\pi_l, \pi_f)(s) := \bar{V}_f(\pi_l, \pi_f)(s) + \lambda H_{\pi_f}(\pi_l, s)$$

Discounted entropy regularization

$$\mathbb{E}_{\mathcal{T}}\left[\sum_{\tau=0}^{\infty} -\gamma^\tau \log \pi_f | s^0 = s, \pi_l\right]$$

- Widely used in MDPs and games (Haarnoja et al 2018, Schulman et al 2017, Mei et al 2020, Meritokopoulos and Sandholm 2016, Sun et al 2024, Aggarwal et al 2024, …) due to many conceptual and numerical advantages

Sources of Difficulty

Identify

Outer loop $\pi_l^* \in \arg\max_{\pi_1} \bar{V}_l(\pi_l) = V_l(\pi_l, B(\pi_l))$

where

Inner loop $B(\pi_l) = \{\pi^*_f : V_f^{\lambda}(\pi_l, \pi^*_f)(\mu) \geq V_f^{\lambda}(\pi_l, \pi_f)(\mu)\}$

Best response of follower

- The leader and follower problems are linked through the best response mapping.

- The non-smoothness and non-uniqueness of the best response mapping leads to discontinuity in the landscape of value function at the leader.

- Each level may consist of different classes of optimization problems.

- Our contribution: We provide the first model free learning algorithm that provably converges to a stationary point for the leader and an optimal best response for the follower

Proposed Algorithm

- Parametrize follower policy as softmax and leader policy as direct (softmax also possible)

$$\pi_f(a_{f,j}|s) = \frac{e^{\theta_i(s,a_{f,j})}}{\sum_{a_{f,j}\in\mathcal{A}_f} e^{\theta_f(s,a_{f,j})}}$$

$$\pi_l(a_{l,j}|s) = \theta_l(s,a_{l,j}) \text{ s.t.} \sum_{j\in\mathcal{A}_l} \pi_l(a_{l,j}|s) = 1$$

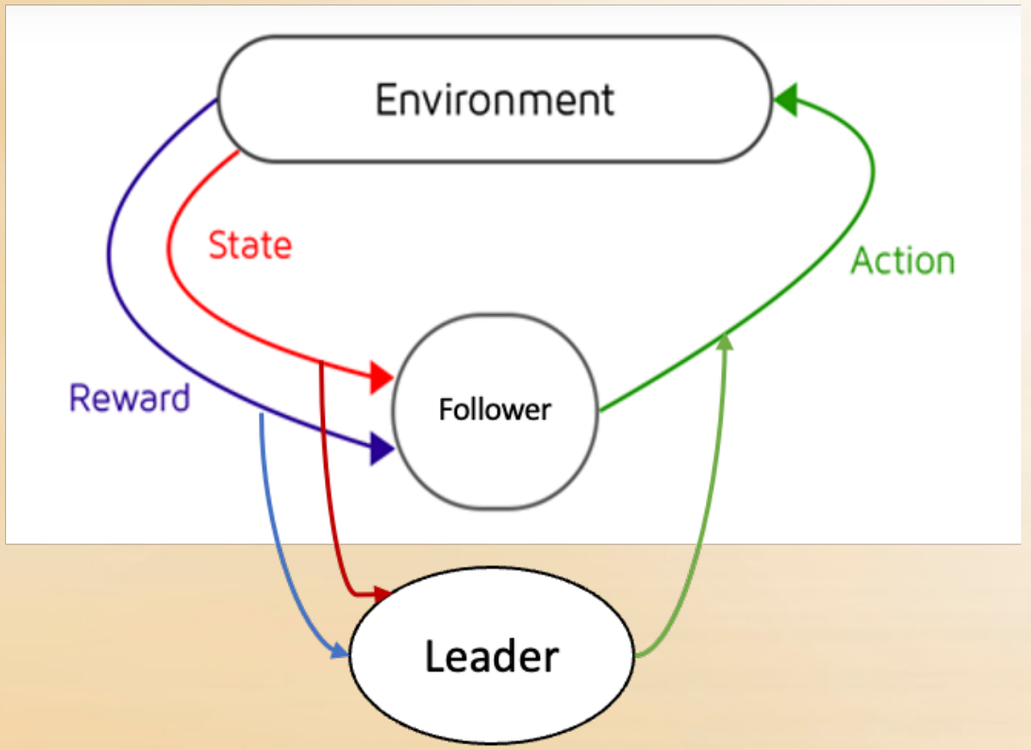
- A two loop algorithm

Algorithm 1 Loopy Direct Stackelberg Policy Gradient

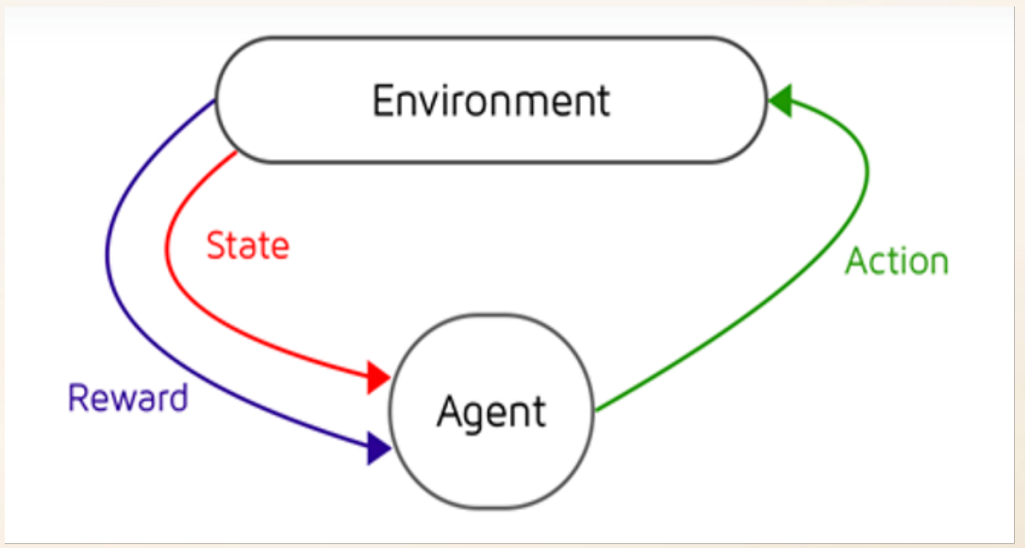
Input parameters: Step sizes $\eta_t, \forall t \in \{0,1,\cdots T\}$, $\beta_n, \forall n \in \{0,1,\cdots M\}$, distribution μ of initial states with positive support for all $s \in \mathcal{S}$.
Initialize $\theta_l^0, \theta_f^{0,0}$, $\theta_f^{0,M}$
for $t = 0...T$ **do**
  $\theta_l^{t+1} \leftarrow P_{\Delta(\mathcal{A}_l)^{|\mathcal{S}|}}\left(\theta_l^t + \eta_t \nabla_{\theta_l} \bar{V}_l(\theta_l^t, \theta_f^{t,M})(\mu)\right)$
  if $t > 0$ **then**, initialize $\theta_f^{t,0} = \theta_f^{t-1,M}$
  end if
  for $n = 0...M$ **do**
    $\theta_f^{t,n+1} \leftarrow \theta_f^{t,n} + \beta_n \nabla_{\theta_f} V_f^{\lambda}(\theta_l^t, \theta_f^{t,n})(\mu)$
  end for
end for
Output $\theta_l^* = \theta_l^{T+1}, \theta_f^* = \theta_f^{T,M}$

Inner Loop

- Start with the policy for the follower for given policy by leader



Fix leader policy



Two player stochastic game

Averaged Markov decision process

$$\mathcal{R}_f^{\pi_{\theta_l^t}}(s, a_f) := \sum_{a_l \in \mathcal{A}_l} \pi_{\theta_l^t}(a_l|s)\mathcal{R}_f(s, a_l, a_f)$$

$$\mathcal{P}^{\pi_{\theta_l^t}}(s'|s, a_f) := \sum_{a_l \in \mathcal{A}_l} \pi_{\theta_l^t}(a_l|s)\mathcal{P}(s'|s, a_l, a_f)$$

Best Response for the Follower

- Gradient ascent for an MDP has been studied (Agarwal et al. 2019)
- The optimal response (best response) exists and is unique (Geist et al 2019)
- The gradient algorithm converges asymptotically to this optimal policy (Mei et al 2020)
- The gradient of the value function satisfies non-uniform Lozasiewicz condition

$$\left\| \frac{\partial V_f^{\lambda}(\pi_{\theta_l^t}, \pi_{\theta_f^{t,n}})(\mu)}{\partial \theta_f^{t,n}} \right\|_2 \geq L(\theta_l^t, \theta_f^{t,n}) \left[V_f^{\lambda}(\pi_{\theta_l^t}, \pi^{\star}_{\theta_f^t})(\rho) - V_f^{\lambda}(\pi_{\theta_l^t}, \pi_{\theta_f^{t,n}})(\rho) \right]^{\frac{1}{2}}$$

- The value function satisfies Reverse Lozasiewicz condition

$$\left\| \frac{\partial V_f^{\lambda}(\pi_{\theta_l^t}, \pi_{\theta_f^{t,n}})(\rho)}{\partial \theta_f^{t,n}} \right\|_2 \leq \Delta \, . \left[V_f^{\lambda}(\pi_{\theta_l^t}, \pi^{\star}_{\theta_f^{t,n}})(\rho) - V_f^{\lambda}(\pi_{\theta_l^t}, \pi_{\theta_f^{t,n}})(\rho) \right] .$$

Best Response for the Follower

- Gradient ascent for an MDP has been studied (Agarwal et al. 2019)
- The optimal response (best response) exists and is unique (Geist et al 2019)
- The gradient algorithm converges asymptotically to this optimal policy (Mei et al 2020)
- At the end of M iterations of the inner loop with a constant step size

$$\left\| \theta_f^{t,M} - B(\pi_{\theta_l^t}) \right\| \leq \chi \,.\, \Delta \,.\, f(M, \pi_{\theta_l^t}, \theta_f^{t,0}, \beta),$$

$$f(M, \pi_{\theta_l^t}, \theta_f^{t,0}, \beta) = \beta \sum_{i=M}^{\infty} \exp\left[- K(\pi_{\theta_l^t}, \theta_f^{t,0}, \beta)(i-1) \right]$$

Positive and decreasing in M

- The best response function is c-Lipshcitz in leader's policy where

$$c = \lambda \, |S| \sqrt{|A_f|} \,.\, \Delta^{\star}$$

$$\left\| B(\pi_l^1) - B(\pi_l^2) \right\|_{TV} \leq \lambda \, |S| \sqrt{|A_f|} \,.\, \Delta^{\star} \left\| \pi_l^1 - \pi_l^2 \right\|_{\infty}$$

Policy of the Leader

- At the end of T iterates of the outer loop, Leader's policy satisfies

$$\min_{t\in[T]} \left\| \nabla \hat{V}_l(\pi_{\theta_l}^t, \pi_{\theta_f}^M) \right\| \leq \left(\frac{\hat{V}_l^{opt} - \hat{V}_l(\pi_l^0)}{\eta^2(\frac{1}{\eta} - \frac{L_{\hat{V}_l}}{2})(T+1)} + \frac{\mathcal{G}(T, M, \pi_{\theta_l}^0, \pi_{\theta_f}^0, \eta, \beta)}{(T+1)} \right)^{\frac{1}{2}}$$

$$\mathcal{G}(T, M, \pi_l^0, \pi_{\theta_f}^0, \eta, \beta) = (\eta L_{\hat{V}_l} \chi \,.\, \Delta)^2 \sum_{t=0}^{T} f^2(M, \pi_{\theta_l}^t, \pi_{\theta_f}^0, \beta)$$

decreasing in M and increasing in T

- Provable convergence to a stationary point for the leader and an optimal best response for the follower

- We also provide a finite sample analysis of the algorithm

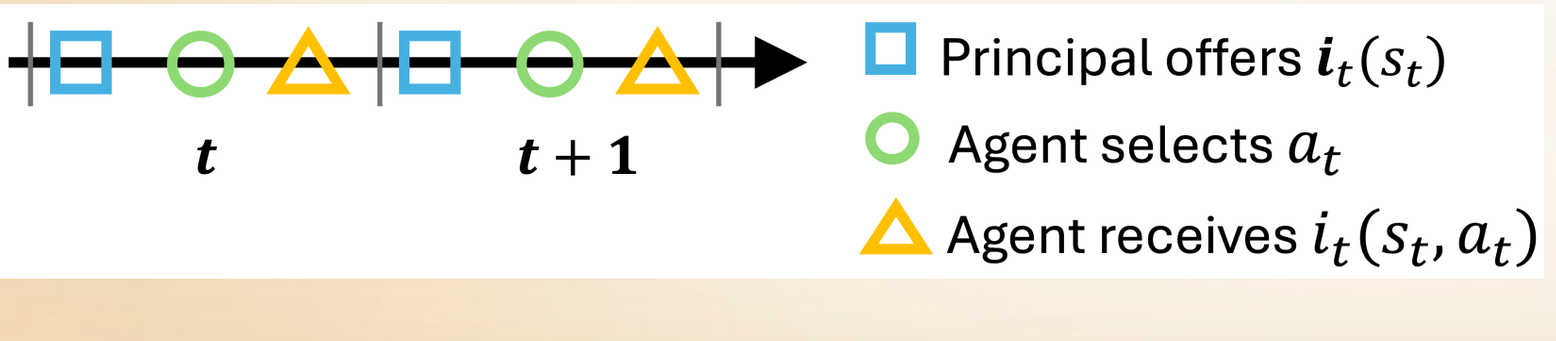
- The algorithm does not require two time scales

Das and Gupta, TAC (Submitted)

Unknown Agent Types

- So far, we had action asymmetry but no information asymmetry

- Consider the agent to have knowledge of an MDP that it acts on, but the principal to know neither the MDP nor the rewards of the agent with the timeline



- The principal offers a set of incentives

$$\mathbf{i}_t(s_t) := \{i_t(s_t, a), \forall a \in \mathscr{A}\}$$

and accumulates reward over time T

$$\sum_{t=0}^{T} (r_t^P(s_t, a_t) - i_t(s_t, a_t))$$

- The agent chooses an action

$$a_t(\mathbf{i}_t) := \textbf{argmax}_{a \in \mathscr{A}}(r_t^A(s_t, a) + i_t(s_t, a))$$

Principal Regret

- The principal does not know MDP or the reward function of the agent, but receives a noisy unbiased observation of the agent reward

- Regret for the principal

$$\Delta(\mathscr{M}, \mathscr{U}, s_0, T) = T\rho^*(\mathscr{M}) - \sum_{t=0}^{T} (r_t^P(s_t, a_t) - i_t(s_t, a_t))$$

- Can we achieve a sublinear regret?

- Related works: Only one action chosen in a static environment (Ratliff and Fiez 2020, Gao et al 2022, Dogan et al 2023) or assume MDP known (Plambeck and Zenios 2000)

Proposed Algorithm

Algorithm 1: Incentivized-UCRL2

for *each epoch* $k \geq 1$ **do**
  Set $t_k = t$.
  for *all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Compute $n_k(s,a), \hat{\bar{r}}_k^A(s,a)$ and $\hat{p}_k(s'|s,a)$.
    Set $\nu_k(s,a) = 0$.
  end
  π_k = INCENTIVIZED EXTENDED VALUE ITERATION$(\hat{p}_k(s'|s,a), \hat{\bar{r}}_k^A(s,a), d_1(s,a), \frac{1}{\sqrt{t_k}})$
  while $\nu_k(s,a) < \max\{1, n_k(s,a)\}$ **do**
    Determine action $a_t^{\text{pref}} = \pi_k(s_t)$.
    Offer $\mathbf{i}_t(s_t)$.
    Obtain reward r_t^P and observe a_t and s_{t+1}.
    Give $i_t(s_t, a_t) \in \mathbf{i}_t(s_t)$.
    Set $\nu_k(s,a) = \nu_k(s,a) + 1$.
  end
end

Find empirical functions from data so far

1. Generate set of plausible MDPs
2. Find optimal policy of optimistic MDP in this set

Choose incentive for the myopic agent

Guarantees

- The algorithm converges
- The set of plausible MDPs contains the true MDP with high probability

$$\mathbb{P}\{\mathscr{M} \notin \mathbf{M}(t)\} < \frac{\delta}{15t^6}$$

- The regret is sublinear with high probability

$$\Delta(\mathscr{M}, \textit{I-UCRL2}, s, T) \leq \mathcal{O}\left(D\,|\,\mathcal{S}\,|\sqrt{|\,\mathcal{A}\,|\,T\log\left(\frac{T}{\delta}\right)}\right)$$

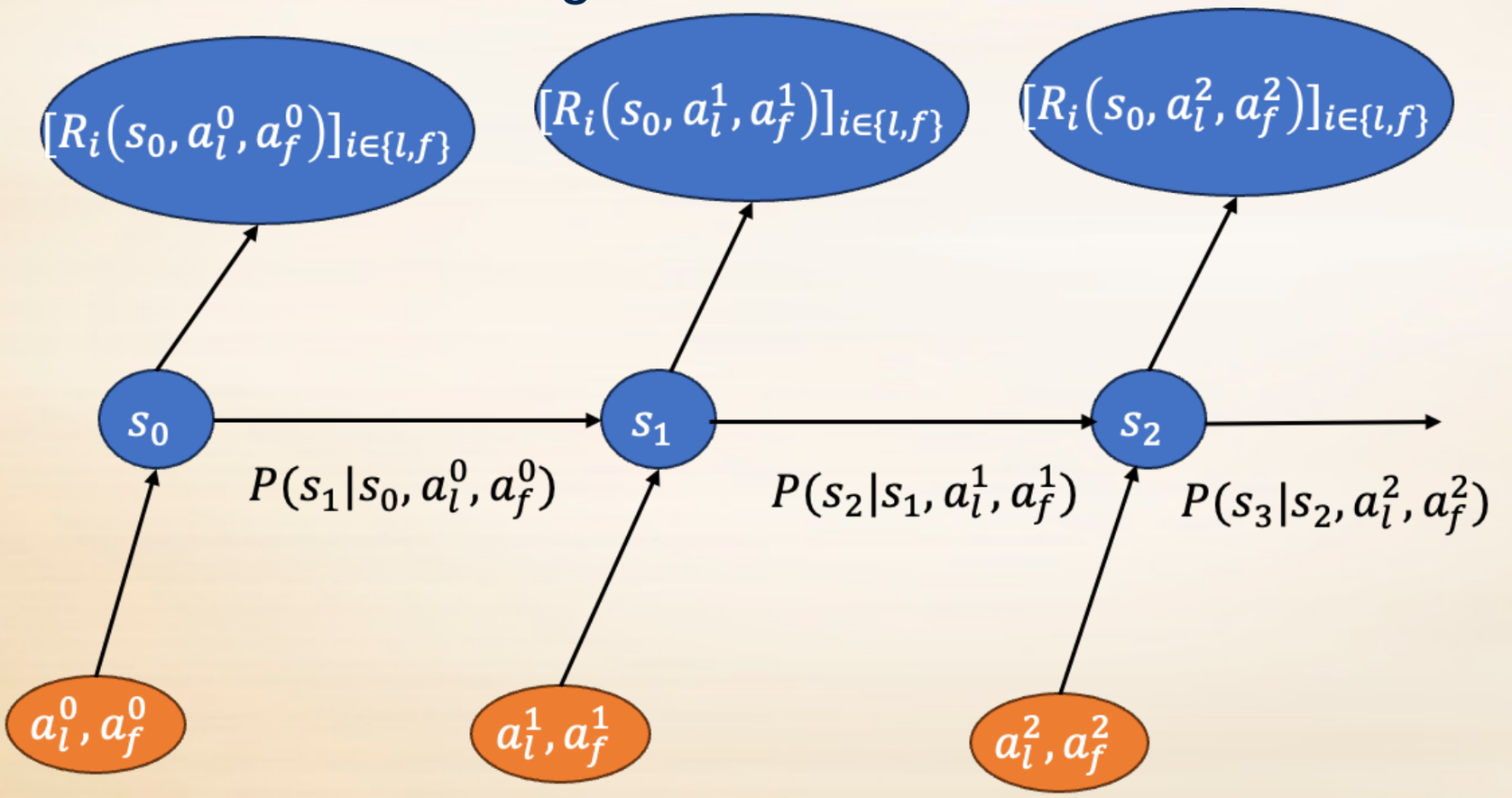
- Ongoing work: multiple agents

Bajaj and Gupta, L-CSS (Submitted)

What Game are We Considering?

- Stackelberg structure where incentive designer is the leader

Leader announces policy before game starts



- Markov game (with Nash equilibrium among players in response to given incentive)

Learning Algorithms for NE

- Consider a stochastic game

$$\langle \mathcal{N}, \mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{N}}, \mathcal{P}, (\mathcal{R}_i)_{i\in\mathcal{N}}, \gamma, \rho \rangle$$

where the agents want to maximize their expected value function

$$V_i^{\pi}(\rho) = \mathbb{E}_{s\sim\rho}\left[\mathbb{E}_{\mathcal{T}}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}_i(s_t, a_i^t, a_{-i}^t) \,|\, s_0 = s, \pi_{-i}\right]\right]$$

- Policies at NE satisfy

$$V_i^{\pi_i^*,\pi_{-i}^*}(\rho) \geq V_i^{\pi_i,\pi_{-i}^*}(\rho), \forall \pi_i \in \mathcal{S}^{\Delta(\mathcal{A}_i)}, \forall i \in \mathcal{N}$$

- Even for static games, Milionis et al 2023 proved that there exists games where any dynamical policy update process that admits a continuous flow will not converge to the set of Nash equilibria for some initial conditions.

- For stochastic cases, some convergence results available for special structures (e.g. zero-sum games (Daskalis et al 2020, Sayin et al 2022), identical interest Markov games (Sayin et al 2022), Markov potential games (Mahewshwari et al 2022, Leonardos et al 2021, Fox et al 2022)

Quantal Response Equilibrium

- Consider again the entropy regularized version of the cost

$$V^{\pi}_{\tau,i} := V^{\pi}_{i} + \tau \mathscr{H}_i(\rho, \pi)$$

with the infinite horizon discounted entropy

$$\mathscr{H}_i(\rho, \pi) := \mathbb{E}_{s_0 \in \rho, \mathbf{a}^t \sim \pi(.|s_t), s_{t+1} \sim \mathscr{P}(.|s_t, \mathbf{a}^t), t \geq 0} \left[\sum_{t=0}^{\infty} - \gamma^t \log \pi_i(a_i^t \,|\, s) \right]$$

- Leads to Quantal Response Equilibrium

$$V^{\pi_i^*, \pi_{-i}^*}_{\tau,i}(\rho) \geq V^{\pi_i, \pi_{-i}^*}_{\tau,i}(\rho), \forall \pi_i \in \mathcal{S}^{\Delta(\mathscr{A}_i)}, \forall i \in \mathscr{N} \,.$$

- Our contribution: The first algorithm to provably converge to QRE.

Natural Policy Gradient

- For a single agent MDP, the NPG update rule for a parametrized policy is

$$\theta_{t+1} \leftarrow \theta_t + \eta^{'}(\mathscr{F}_{\rho}^{\theta})^{\dagger} \nabla_{\theta} V_{\tau}^{\pi_\theta}(\rho)$$

with the Fischer information matrix

$$\mathscr{F}_{\rho}^{\theta} := \mathbb{E}_{s\sim d_{\rho}^{\pi_\theta}, a\sim\pi_\theta(.|s)} \left[(\nabla_{\theta} \log \pi_\theta(a \mid s))(\nabla_{\theta} \log \pi_\theta(a \mid s))^{T} \right]$$

- Known to converge

$$\|Q_{\tau}^{\star} - Q_{\tau}^{t+1}\|_{\infty} \leq C_1 \gamma (1 - (1-\gamma)\eta^{'}\tau)^{n}, \forall n \geq 0$$

if learning rate chosen as $0 < \eta^{'} \leq \dfrac{(1-\gamma)}{\tau}$

- In a game, each agent has to implement this algorithm, but that requires estimating Q values

Proposed Algorithm

Algorithm 1 Best response independent Natural Policy Update

Input:Parameter d, learning rate η, initialization joint policy π^0
for $t = 0, 1, 2, \cdots$ **do**

Compute the optimal regularized Q-function $Q_{\tau,i}^{\pi_i^t *, \pi_{-i}^t}(\pi_i^t, \pi_{-i}^t)$ for every player i.
Update the policy:

$$\forall i \in \mathcal{N}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}_i : \pi_i^{(t+1)}(a|s) = \frac{1}{Z_i^{(t)}(s)} (\pi_i^{(t)}(a|s))^{(1-\eta\tau)^d} \exp\left(\eta Q_{\tau,i}^{\pi_i^t *, \pi_{-i}^t}(s,a)\right)$$

$$\text{where } Z_i^{(t)}(s) = \sum_{a' \in \mathcal{A}_i} \exp\left(\eta Q_{\tau,i}^{\pi_i^t *, \pi_{-i}^t}(s,a')\right)$$

end for

- If agents have access to the correct Q-functions, can prove that the policy converges to an equilibrium

$$QRE - gap(\pi) = \max_{i \in \mathcal{N}} \|B_i^\tau(\pi_{-i}) - \pi_i\|_\infty$$

$$\leq 2\left((1-\eta\tau)^d + 2\eta C_-(\tau) \sum_{i \in \mathcal{N}} |\mathcal{A}_i| \right)^t QRE - gap(\pi^0)$$

Proposed Algorithm

Algorithm 1 Best response independent Natural Policy Update

Input:Parameter d, learning rate η, initialization joint policy π^0

for $t = 0, 1, 2, \cdots$ **do** *approximate*

Compute the optimal regularized Q-function $Q_{\tau,i}^{\pi_i^*, \pi_{-i}^t}(\pi_i^t, \pi_{-i}^t)$ $Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}$ for every player i.

Update the policy:

$$\forall i \in \mathcal{N}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}_i : \pi_i^{(t+1)}(a|s) = \frac{1}{Z_i^{(t)}(s)} (\pi_i^{(t)}(a|s))^{(1-\eta\tau)^d} \exp\left(\eta Q_{\tau,i}^{\pi_i^*, \pi_{-i}^t}(s,a)\right) \quad Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}$$

where $Z_i^{(t)}(s) = \sum_{a' \in \mathcal{A}_i} \exp\left(\eta Q_{\tau,i}^{\pi_i^*, \pi_{-i}^t}(s,a')\right)$ $Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}$

end for

- If agents have access to the correct Q-functions, can prove that the policy converges to an equilibrium

$$QRE - gap(\pi) = \max_{i \in \mathcal{N}} \|B_i^\tau(\pi_{-i}) - \pi_i\|_\infty$$

$$\leq 2\left((1-\eta\tau)^d + 2\eta C_-(\tau) \sum_{i \in \mathcal{N}} |\mathcal{A}_i|\right)^t QRE - gap(\pi^0) + 2\frac{\phi(\eta,\tau)\delta}{\left(1 - (1-\eta\tau)^d - 2\eta C_-(\tau) \sum_{i \in \mathcal{N}} |\mathcal{A}_i|\right)}$$

Main Result

Input:Parameter d, learning rate $\eta, \eta^{'}$, initialization joint policy π^0
for $t = 0, 1, 2, \cdots$ **do**
Initialize $\forall i \in \mathcal{N}, \hat{\pi}_i^0 = \pi_i^t$
for $n_t = 0, 1, 2, \cdots M$ **do**
do Natural Policy Gradient for each agent

$$\forall i \in \mathcal{N} : \hat{\pi}_i^{(n_t+1)}(a|s) = \frac{1}{Z_i^{(n_t)}(s)} (\hat{\pi}_i^{(n_t)}(a|s))^{(1-\eta^{'}\tau)} \exp\left(\eta^{'} Q_{\tau,i}^{\hat{\pi}_i^{n_t}, \pi_{-i}^t}(s,a)\right)$$

$\forall i \in \mathcal{N}$:return Q-value associated with joint policy $(\hat{\pi}_i^{M_t}, \pi_{-i}^t)$ (Say $Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}(s,a)$) .

end for
Update the policy:

$$\forall i \in \mathcal{N}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}_i : \pi_i^{(t+1)}(a|s) = \frac{1}{Z_i^{(t)}(s)} (\pi_i^{(t)}(a|s))^{(1-\eta\tau)^d} \exp\left(\eta Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}(s,a)\right)$$

where $Z_i^{(t)}(s) = \sum_{a^{'} \in \mathcal{A}_i} \exp\left(\eta Q_{\tau,i}^{\pi_i^{M_t}, \pi_{-i}^t}(s,a^{'})\right)$

end for

- Two timescale algorithm (inner loop does NPG while outer loop updates policy)

Das and Gupta, ICML (Submitted)

Scalability

- The algorithms considered so far require the follower agents to know global state and actions, which is not scalable.

- However, if inherent structure in the game, scalablility can be ensured at the cost of some performance degradation

$$P(s' \mid s, a) = \prod_{i=1}^{n} P(s_i' \mid s_{\mathcal{N}_i}, a_{\mathcal{N}_i})$$

$$r_i : \mathcal{S}_{\mathcal{N}_i} \times \mathcal{A}_{\mathcal{N}_i} \to [0, r_{max}]$$

- Players execute independent natural policy gradient based only on k-hop neighbors

$$\theta^{t+1}_{i, s_{\mathcal{N}_i^\kappa}, a_i} = \theta^{t}_{i, s_{\mathcal{N}_i^\kappa}, a_i} + \frac{\eta}{1-\gamma} A_i^{\pi_{\theta_{i,\kappa}}}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})$$

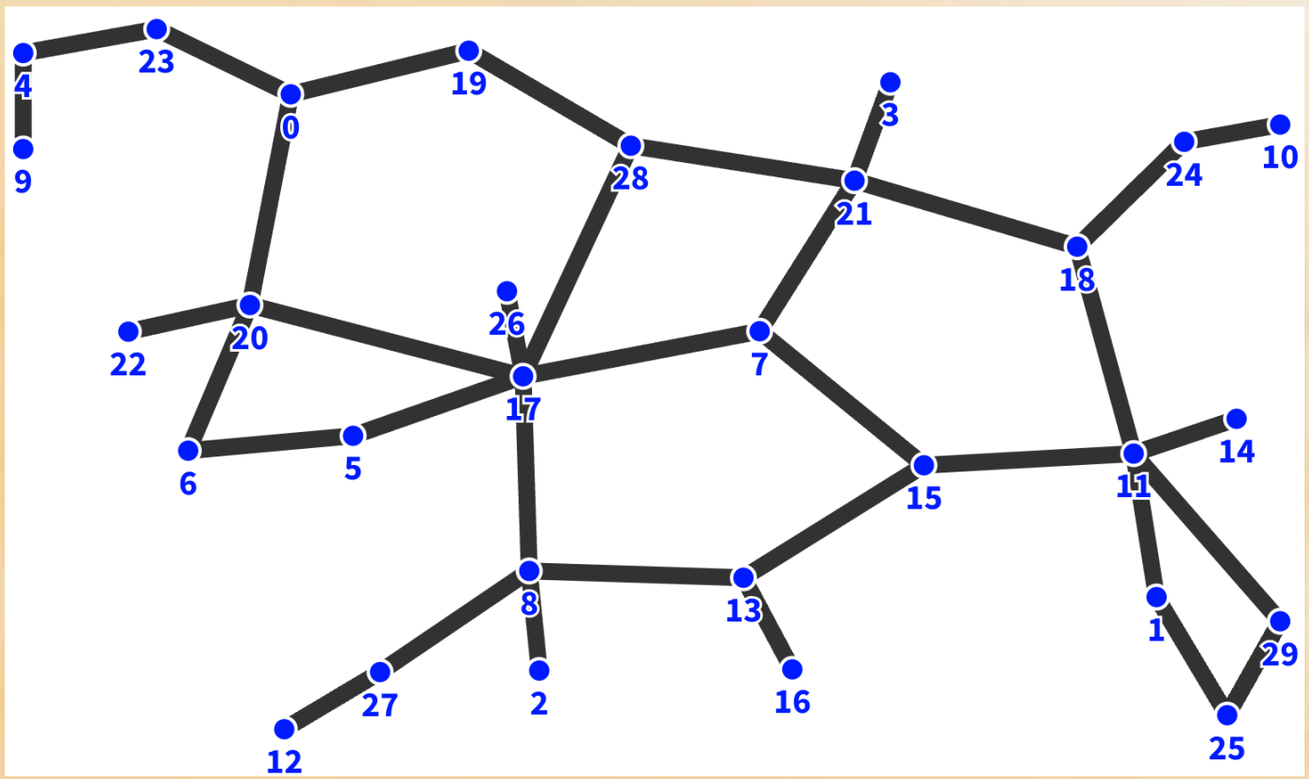
- Theorem: The modified independent natural policy gradient algorithm converges to an ϵ-equilibrium policy where

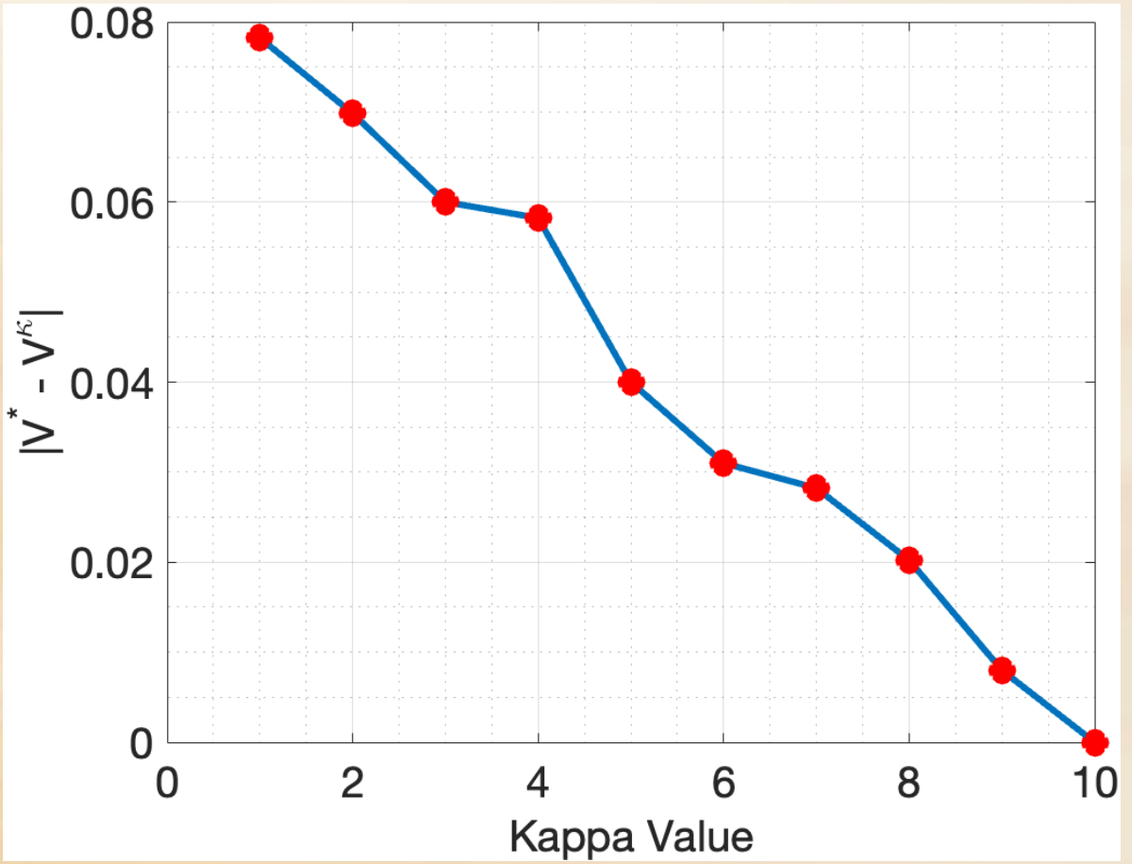
$$\epsilon = \frac{r_{max}}{1-\gamma} \gamma^{\kappa+1} .$$

Scalability

- Job balancing example with 30 agents and reward based on deviation from average load

$$r_i(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) = \begin{cases} \dfrac{1}{\left| s_i - \frac{1}{|\mathcal{N}_i^\kappa|} \sum_{j \in \mathcal{N}_i^\kappa} s_j \right| + \epsilon} & \text{if } s_i \neq \frac{1}{|\mathcal{N}_i^\kappa|} \sum_{j \in \mathcal{N}_i^\kappa} s_j \\ 1 & \text{if } s_i = \frac{1}{|\mathcal{N}_i^\kappa|} \sum_{j \in \mathcal{N}_i^\kappa} s_j \end{cases}$$





Abdelnaby and Gupta, L-CSS (Submitted)

Conclusion

- Incentive / contract / mechanism / auction design (including sequential incentives) has a long history

- Can we use such methods for coordinating behavior in dynamic systems?
  - Complexity of optimal contract design
    - In general, computationally complex
  - Design of strategies for participants in resulting Markov Stackelberg games
    - Learning algorithms for Markov Stackelberg games with or without model known to principal
    - Learning algorithms for Nash equilibrium in Markov games
    - Scalable algorithms
- Encoding objectives such as stability or robustness (as opposed to the system operator being interested in social welfare)?

gupta869@purdue.edu