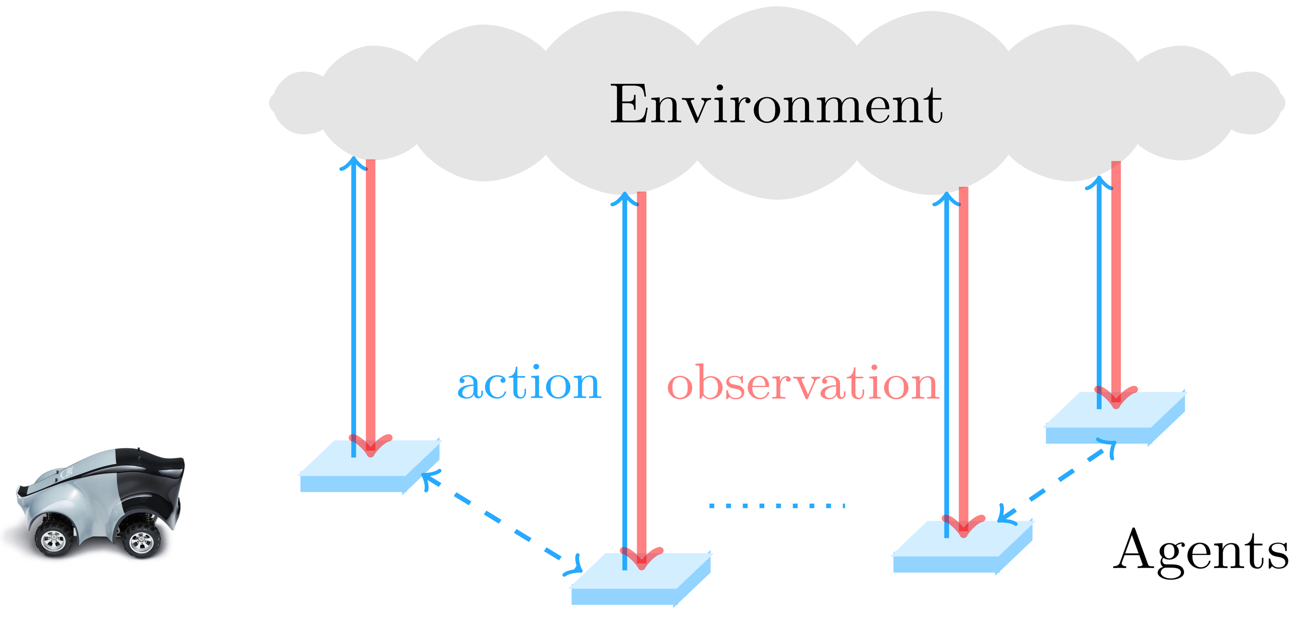
Provably Efficient Generalized Lagrangian Policy Optimization for Safe MARL

Dongsheng Ding (Penn), Xiaohan Wei (Meta), Zhuoran Yang (Yale), Zhaoran Wang (Northwestern), Mihailo R. Jovanović (USC)

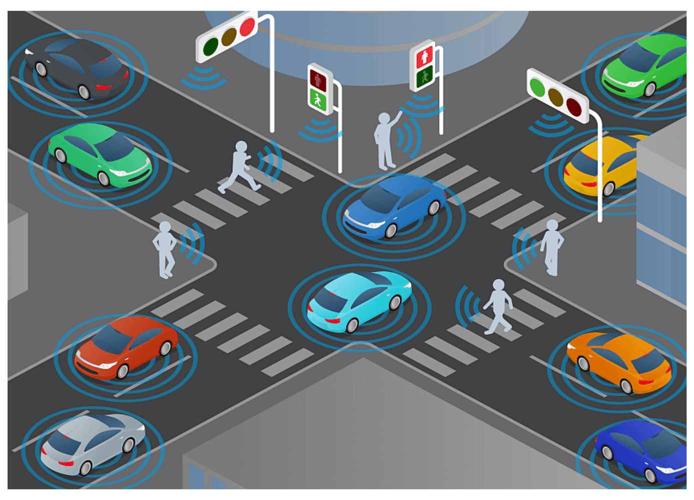
Motivation

Multi-agent sequential decision making



Trade-off { reward, profit, ... } vs. { safety, budget, fairness, ... }

Constraint-rich multi-agent systems



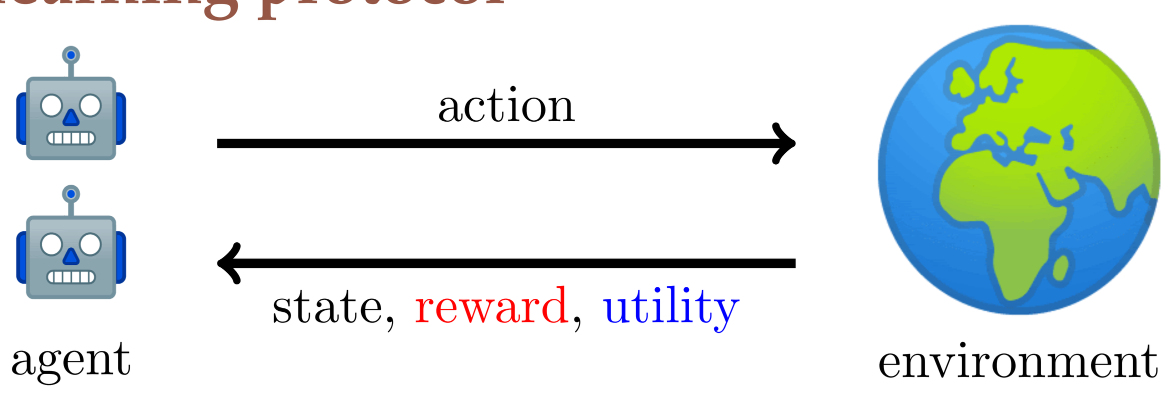
automated vehicles

satellite communication

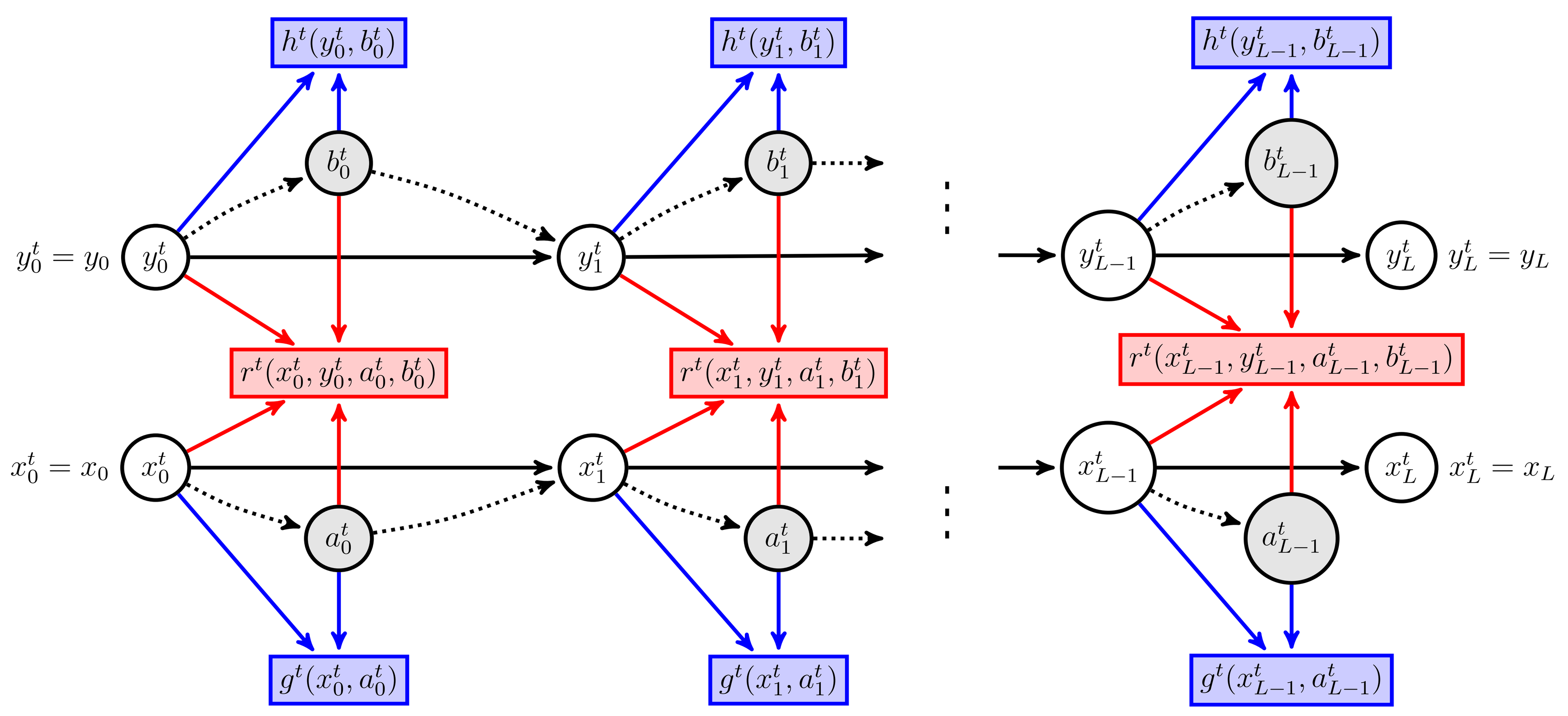
Challenges • safe exploration • efficiency

Problem formulation

Episodic learning protocol



- $\ell = 0, \ldots, L$ – horizon
- $t = 0, \ldots, T-1$ – episode
- $(x_\ell^t, y_\ell^t), (a_\ell^t, b_\ell^t), r_\ell^t, (g_\ell^t, h_\ell^t)$ – state, action, reward, utilities
- $a_\ell^t \sim \pi^t(\cdot \,|\, x_\ell^t)$, $b_\ell^t \sim \mu^t(\cdot \,|\, y_\ell^t)$ – policies
- $x_{\ell+1}^t \sim P_1(\cdot \,|\, x_\ell^t, a_\ell^t)$, $y_{\ell+1}^t \sim P_2(\cdot \,|\, y_\ell^t, b_\ell^t)$ – independent dynamics



- $\langle q_1^t \cdot q_2^t, r^t \rangle := \mathbb{E}[\sum_{\ell=0}^{L-1} r^t(x_\ell, a_\ell, a_\ell, b_\ell)]$ – reward value
- $\langle q_1^t, g^t \rangle := \mathbb{E}[\sum_{\ell=0}^{L-1} g^t(x_\ell, a_\ell)]$ – utility value; also for $\langle q_2^t, h^t \rangle$

Constrained zero-sum Markov game

$$\underset{q_1 \in \Delta(P_1)}{\text{maximize}} \;\; \underset{q_2 \in \Delta(P_2)}{\text{minimize}} \;\; \sum_{t=0}^{T-1} \langle q_1 \cdot q_2, r^t \rangle \quad \text{subject to} \;\; \langle q_1, g \rangle + \langle q_2, h \rangle \;\le\; b$$

- r^t – adversarial
- g, h – expectations of stochastic g^t, h^t
- $(q_1^\star, q_2^\star)$ – constrained Nash equilibrium

Peformance measure

$$\text{Regret}(K) := \sum_{t=0}^{T-1} \left(\langle q_1^t \cdot q_2^\star, r^t \rangle - \langle q_1^\star \cdot q_2^t, r^t \rangle \right)$$

$$\text{Violation}(K) := \sum_{k=1}^{K} \left(\langle q_1^t, g^t \rangle + \langle q_2^t, h^t \rangle - b \right)$$

- q_1^t / q_2^t – occupancy measures induced by policies π^t / μ^t

Algorithm design

One-episode constrained minimax problem

$$\underset{q_1 \in \Delta(P_1)}{\text{maximize}} \;\; \underset{q_2 \in \Delta(P_2)}{\text{minimize}} \;\; \langle q_1 \cdot q_2, r^{t-1} \rangle \quad \text{subject to} \;\; \langle q_1, g^{t-1} \rangle + \langle q_2, h^{t-1} \rangle \;\le\; b$$

- $L^t(q_1, q_2; \lambda)$ – generalized Lagrangian

$$\begin{aligned} = \; & \langle q_1 \cdot q_2, r^{t-1} \rangle && \text{minimax obj.} \\ & + \lambda \left(\langle q_1, g^{t-1} \rangle + \langle \hat{q}_2^t, h^{t-1} \rangle - b \right) && \text{min's vio.} \\ & - \lambda \left(\langle \hat{q}_1^t, g^{t-1} \rangle + \langle q_2, h^{t-1} \rangle - b \right) && \text{max's vio.} \end{aligned}$$

Online mirror descent primal-dual step

$$\hat{q}^t = \underset{q_1 \in \hat{\Delta}_1}{\arg\min} \; \underset{q_2 \in \hat{\Delta}_2}{\arg\max} \; L^t(q_1, q_2, \lambda^{t-1}) + \underbrace{\frac{1}{\eta} D_{\text{KL}}(q, \hat{q}^{t-1})}_{\text{KL regularization}}$$

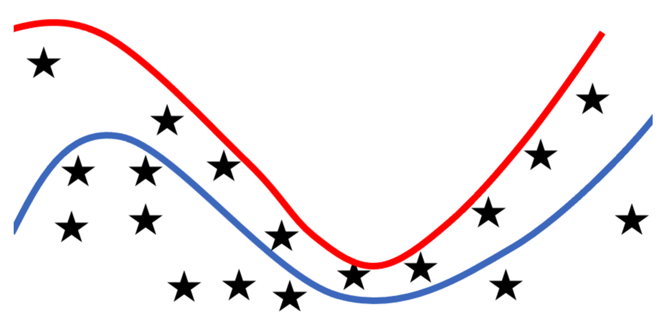
$$\lambda^t = \max \left(\lambda^{k-1} + \underbrace{\left(\langle \hat{q}_1^t, g^{t-1} \rangle + \langle \hat{q}_2^t, h^{t-1} \rangle - b \right)}_{\text{violations}}, \; 0 \right)$$

- compete for rewards
- cooperate for constraints

Optimistic estimation of $\hat{\Delta}_1$, $\hat{\Delta}_2$

$$\hat{P}_1 \longleftarrow \underbrace{\bar{P}_1}_{\text{exploit}} + \underbrace{\text{UCB}_1}_{\text{explore}}$$

$$\hat{P}_2 \longleftarrow \underbrace{\bar{P}_2}_{\text{exploit}} + \underbrace{\text{UCB}_2}_{\text{explore}}$$



$$\hat{\Delta}_i = \text{Linear constraint} \left(\bar{P}_i, \text{UCB}_i \right)$$

Theoretical guarantee

Constrained Markov games with independent dynamics

$$\text{Regret}(K), \text{Violation}(K) = \tilde{O}\left((|X| + |Y|)\, L \sqrt{T(|A| + |B|)} \right)$$

- T – # episodes;
- L – horizon length
- $|X| + |Y|$, $|A| + |B|$ – state/action space sizes
- no sampling assumptions & adversarial reward function
- applicable to side constraint case and single-controller case

Reference

[1] D. Ding, X. Wei, Z. Yang, Z. Wang, M. Jovanovic, "Provably Efficient Generalized Lagrangian Policy Optimization for Safe Multi-Agent Reinforcement Learning", *arXiv:2306.00212* (a long version with appendices).