Learning to Switch Among Agents in a Team

Vahid Balazadeh Meresht, Abir De, Adish Singla, Manuel Gomez-Rodriguez

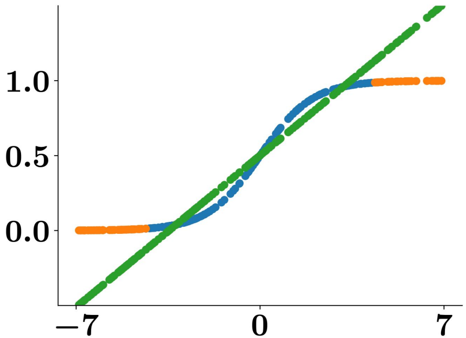
MPI-SWS

Reinforcement learning vs humans

Video games

Autonomous driving

Deploy RL agents under lower automation levels



De et al. 2020

When should we switch control?

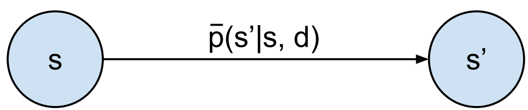
1. Level of automation
2. Number of switches
3. Unknown human policy

Separate environment and agents

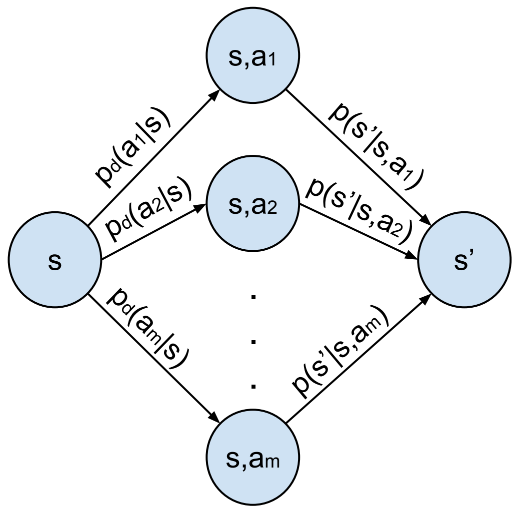
$$\pi^* = \underset{\pi}{\operatorname{argmin}}\ \mathbb{E}\left[\sum_{\tau=t}^{L} c'(s_\tau, a_\tau) + c_c(d_\tau) + c_x(d_\tau, d_{\tau-1}) \,\middle|\, s_t = s.d_{t-1} = d\right]$$
$$d_t = \pi_t(s_t, d_{t-1})$$

2-layer MDP

Standard MDP



No learning about the environment



UCRL2 with Multiple Confidence sets

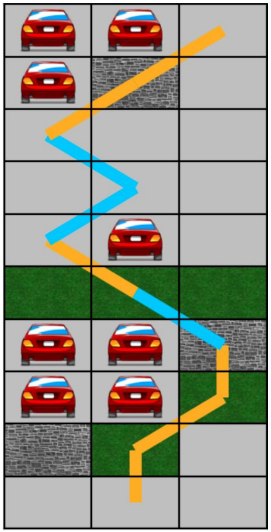
Theorem 1. *For any episode k, the optimal value function $v_t^k(s,d)$ satisfies the following recursive equation:*

$$v_t^k(s,d) = \min_{d_t \in \mathcal{D}} \left[c_{d_t}(s,d) + \min_{p_{d_t} \in \mathcal{P}^k_{\cdot \mid d_t, s, t}} \sum_{a \in \mathcal{A}} p_{d_t}(a \mid s,t) \times \left(c_e(s,a) + \min_{p \in \mathcal{P}^k_{\cdot \mid s,a,t}} \mathbb{E}_{s' \sim p(\cdot \mid s,a,t)} [v_{t+1}^k(s', d_t)] \right) \right],$$

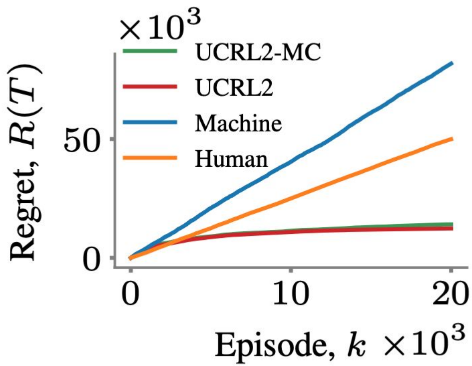
Setting	UCRL2-MC Regret	UCRL2 Regret
Single team of agents	$\tilde{\mathcal{O}}(L\|\mathcal{S}\|\sqrt{\mathcal{A}T})$	$\tilde{\mathcal{O}}(L\|\mathcal{S}\|\sqrt{\mathcal{D}T})$
Multiple teams of agents	$\tilde{\mathcal{O}}(L\|\mathcal{S}\|\sqrt{\mathcal{A}TN} + NL\sqrt{\|\mathcal{A}\|\|\mathcal{S}\|\|\mathcal{D}\|T})$	$\tilde{\mathcal{O}}(NL\|\mathcal{S}\|\sqrt{\mathcal{D}T})$

Results

- Obstacle avoidance task in a lane driving environment
- Improved regret in multiple teams of agents setting

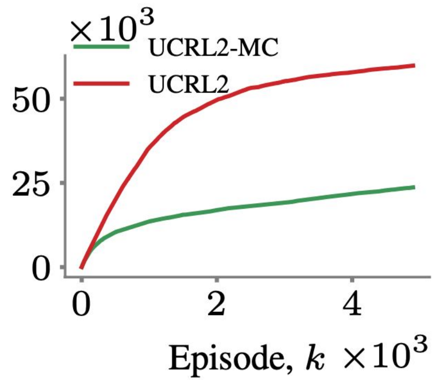


Single team of agents



(a) $c_c(\mathbb{H}) = 0.2, c_x = 0.1$

Multiple teams of agents
N = 10



(a) $c_c(\mathbb{H}) = 0.2, c_x = 0.1$