Gap-Dependent Unsupervised Exploration for Reinforcement Learning

Jingfeng Wu, Vladimir Braverman, Lin F. Yang
Johns Hopkins University, UCLA

Problem Setup

Unsupervised RL: Task-Agnostic Exploration (TAE)

- **Reward Set**: $\mathscr{R} \subset \{r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]\}$
- **Exploration**: collect data, *w/o reward signal*
- **Planning**: given an "independent" reward $r \in \mathscr{R}$, compute a nearly optimal policy:

$$\mathbb{P}\{V_1^*(r) - V_1^\pi(r) > \epsilon\} < \delta$$

Existing Results

minimax sample complexity $\propto \tilde{\mathcal{O}}(1/\epsilon^2)$

Question (gap-TAE)

If $\mathscr{R} := \{r : \mathrm{gap}(r) \geq \rho\}$*, i.e., the possible reward induces a constant "gap", is there a faster algorithm?*

Example: Go Game, multiple winning rules,
$\mathscr{R} := \{$Chinese rule, Japanese rule, Korean rule, $\dots\}$

Algorithm

Exploration. UCBVI with two modifications:

- "reward" $\to 0$
- bonus is *clipped* (set to zero if it is small) (ρ is an input)

$$c^k(x,a) \approx \mathtt{clip}_{\frac{\rho}{H}}\left(\sqrt{\frac{H^2 \log}{N^k(x,a)}}\right) + \text{lower orders}$$

Planning. The usual UCBVI method, bonus

$$b^k(x,a) \approx \sqrt{\frac{H^2 \log}{N^k(x,a)}}$$

Theory

Finite-horizon MDP, S states, A actions, horizon length H, gap ρ for reward set $\mathscr{R}$, failure probability δ.

An Upper Bound

For the output policy π after K episodes, the error is at most

$$V_1^*(x_1) - V_1^\pi(x_1) \leq \tilde{\mathcal{O}}\left(\frac{H^3SA}{\rho K} \cdot \log\frac{1}{\delta} + \frac{H^4S^2A}{K} \cdot \log\frac{1}{\delta}\right) = \tilde{\mathcal{O}}\left(\frac{1}{K}\right)$$

where $\tilde{\mathcal{O}}$ hides $\log^2(HSAK)$ and constants.

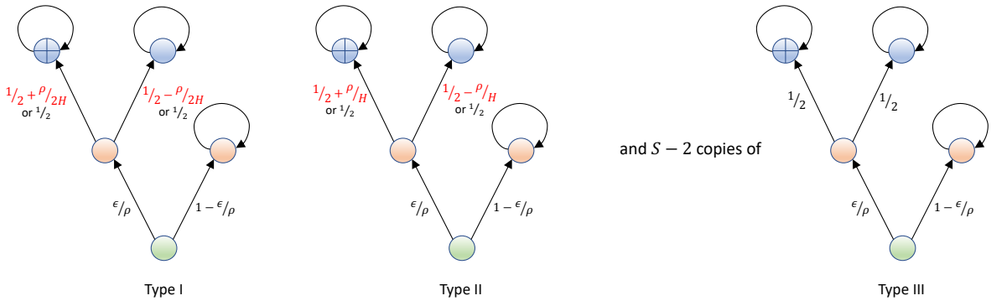
A Lower Bound

For any gap-TAE algorithm, to be (ϵ, δ)-correct needs at least K episodes where

$$\mathbb{E}[K] \geq \Omega\left(\frac{H^2SA}{\rho\epsilon} \cdot \log\frac{1}{\delta}\right) = \Omega\left(\frac{1}{\epsilon}\right)$$

Our bounds are nearly tight for ϵ (or K)

A hard instance for gap dependent task-agnostic exploration



Messages

1. gap-TAE can be faster, but is still limited

   gap-TAE $\propto \tilde{\mathcal{O}}(1/\epsilon)$ *vs. TAE* $\propto \tilde{\mathcal{O}}(1/\epsilon^2)$

2. RL vs. bandits or MDP w/ simulator: a separation in the unsupervised setting

   gap-TAE for bandit or MDP w/ simulator $\propto \tilde{\mathcal{O}}(1)$

Discussions

- An ALGO agnostic to ρ, the gap lower bound?
- Removing S^2 dependence?
- Interpolating the minimax rate and the gap-dependent rate?
- Improving H dependence?

References

[1] Simchowitz, Max, and Kevin G. Jamieson. "Non-asymptotic gap-dependent regret bounds for tabular MDPs." *Advances in Neural Information Processing Systems* 32 (2019): 1153-1162.
[2] Wu, Jingfeng, Vladimir Braverman, and Lin F. Yang. "Accommodating Picky Customers: Regret Bound and Exploration Complexity for Multi-Objective Reinforcement Learning." *arXiv preprint arXiv:2011.13034* (2020).
[3] Zhang, Xuezhou, and Adish Singla. "Task-agnostic exploration in reinforcement learning." *arXiv preprint arXiv:2006.09497* (2020).