Pessimism for Offline Linear Contextual Bandits via $\ell_p$ Confidence Sets

Gene Li, Cong Ma, Nathan Srebro

Motivation

- **Offline reinforcement learning**: For many applications, collecting new data may be costly or dangerous. Instead, our objective is to learn good policies from fixed historical data.
- **Offline data may have insufficient coverage over the state/action spaces.**
- To address this, the principle of pessimism discounts policies which are less represented by the offline dataset.

Problem setup

Linear contextual bandits:
- **Dataset** $\mathcal{D} = \{(s_t, a_t, r_t)\}_{t \in [n]}$, where $r_t \sim R(s_t, a_t)$.
- **Linear rewards**: $r(s, a) := E[R(s, a)] = \phi(s, a)^\top \theta^*$.  
- **Known feature mapping** $\phi(\cdot, \cdot) \in \mathbb{R}^d$.
- **Known test distribution** $\rho \in \Delta(\mathcal{S})$.
- **Unknown parameter vector** $\theta^* \in \mathbb{R}^d$.

Goal: find a policy $\pi: \mathcal{S} \rightarrow \mathcal{A}$ that maximizes value.

$$V(\pi) := E_{s \sim \rho, \pi}[\phi(s, \pi(s))^{\top} \theta^*]$$

We define $\pi^*(s)$ to be the policy that maximizes $V(\pi)$.

$$\pi^*(s) := \arg \max_{a \in \mathcal{A}} \phi(s, a)^{\top} \theta^*.$$  

Upper bound

We solve the offline linear contextual bandit problem by designing learning rules based on the construction of certain $\ell_p$ confidence sets.

$$\mathcal{D} = \{(s_t, a_t, r_t)\}_{t \in [n]}$$

Confidence set $\Theta_p := \{\theta \in \mathbb{R}^d : \|\Sigma^{-1/2} \phi(s_t, \phi^*(s_t))\|_p \leq \rho\}$

Learning rule $\hat{\pi}_p := \arg \max_{\pi \in \Theta_p} V_\rho(\pi)$

**Theorem 1.** Fix $p, q$ such that $1/p + 1/q = 1$. With probability at least $1 - \delta$,

$$V(\pi^*) - V(\hat{\pi}_p) \leq d^{1/p} \cdot \sqrt{\frac{\log d/\delta}{n} \cdot \|\Sigma^{-1/2} \phi(s, \pi^*(s))\|_q}.$$  

Lower bound

We prove that each $\hat{\pi}_p$ is minimax-optimal (up to log factors) over certain constrained classes of contextual bandit instances.

**Theorem 2.** Fix $p, q$ such that $1/p + 1/q = 1$. As long as $\Lambda = \Omega(d^{1/q - 1/2})$ and $n \geq d^{2/p} \Lambda^2$, we have

$$\inf_{\hat{s} \in \mathbb{C}_q} \sup_{\pi \in \mathcal{CB}_q(\Lambda)} E[V^*_\rho - V(\hat{s})] \geq \frac{d^{1/q}}{\sqrt{n}} \cdot \Lambda^2.$$  

$\mathcal{CB}_q(\Lambda)$ is the set of all instances where $\mathbb{C}_q \leq \Lambda$.

Adaptive minimax optimality

Theorem 1 and 2 show that the $\hat{\pi}_\infty$ learning rule satisfies an adaptive minimax optimality property, i.e., $\hat{\pi}_\infty$ attains minimax optimality (up to log factors) for all classes $\mathbb{C}_q(\Lambda)$, $q \geq 1$. We show this is unique to $\hat{\pi}_\infty$.

Previously proposed learning rules based on $\ell_2$ pessimism (Jin-Yang-Wang’21, Xie-Cheng-Jiang-Mineiro-Agarwal’21, Zanette-Wainwright-Brunskill’21) cannot adapt to easy instances, while $\hat{\pi}_\infty$ can!

Main takeaways

- **We make progress towards understanding how to correctly measure complexity and design algorithms for offline RL.**
- **What is the analogue of $\ell_\infty$ pessimism for general function approximation?**
- **Can we provide instance-optimality guarantees?**