Pessimism for Offline Linear Contextual Bandits via ℓ_p Confidence Sets

Gene Li, Cong Ma, Nathan Srebro

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

THE UNIVERSITY OF CHICAGO

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.



Many pessimistic learning rules have been proposed in theory and practice. Which one do we use?

Problem setup

Linear contextual bandits:

- Dataset $\mathscr{D} = \{(s_i, a_i, r_i)\}_{i\in[n]}$, where $r_i \sim R(s_i, a_i)$.
- Linear rewards: $r(s,a) := \mathbb{E}[R(s,a)] = \phi(s,a)^\top \theta^\star$.
- Known feature mapping $\phi(\cdot,\cdot) \in \mathbb{R}^d$.
- Known test distribution $\rho \in \Delta(\mathscr{S})$.
- Unknown parameter vector $\theta^\star \in \mathbb{R}^d$.

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

$$V(\pi) := \mathbb{E}_{s\sim\rho}[\phi(s,\pi(s))^\top \theta^\star]$$

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

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

Upper bound

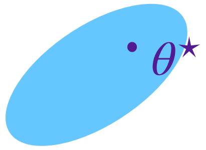
We solve the offline linear contextual bandit problem by designing learning rules based on the construction of certain ℓ_p confidence sets.

Dataset

$$\mathscr{D} = \{(s_i, a_i, r_i)\}_{i\in[n]}$$

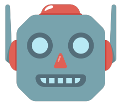
Confidence set

$$\Theta_p := \{\theta \in \mathbb{R}^d : \|\Sigma_{\mathscr{D}}^{-1/2}(\theta - \hat{\theta}_{\mathrm{OLS}})\|_p \le \beta\}$$



Learning rule

$$\hat{\pi}_p := \arg\max_{\pi} \min_{\theta\in\Theta_p} V_\theta(\pi)$$



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

$$V(\pi^\star) - V(\hat{\pi}_p) \lesssim d^{1/p} \cdot \sqrt{\frac{\log d/\delta}{n}} \cdot \|\Sigma_{\mathscr{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star(s))]\|_q.$$

"Complexity term" $\mathfrak{C}_q$

Lower bound

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

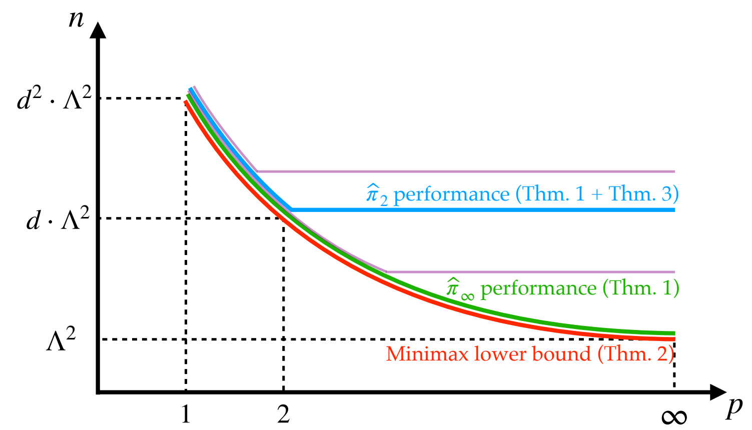
Theorem 2. Fix p, q such that $1/p + 1/q = 1$. As long as $\Lambda = \Omega(d^{1/q-1/2})$ and $n \ge d^{2/p}\Lambda^2$, we have

$$\inf_{\hat{\pi}} \sup_{\mathscr{Q}\in \mathsf{CB}_q(\Lambda)} \mathbb{E}[V_{\mathscr{Q}}^\star - V_{\mathscr{Q}}(\hat{\pi})] \gtrsim \frac{d^{1/q}}{\sqrt{n}} \cdot \Lambda^2 .$$

$\mathsf{CB}_q(\Lambda)$ is the set of all instances where $\mathfrak{C}_q \le \Lambda$.

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 $\mathsf{CB}_q(\Lambda)$, $q \ge 1$. We show this is unique to $\hat{\pi}_\infty$.



Sample complexity of $\hat{\pi}_p$ for various CB classes.

Previously proposed learning rules based on ℓ_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 ℓ_∞ pessimism for general function approximation?
- Can we provide instance-optimality guarantees?