上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Learning Adversarial Linear Mixture Markov Decision Processes with Bandit Feedback and Unknown Transition

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, Shuai Li
Shanghai Jiao Tong University
The Chinese University of Hong Kong, Shenzhen

Adversarial MDPs - Literature

- Adversarial MDPs with function approximation:
  - Adversarial linear MDPs in bandit feedback setting: $\tilde{O}\left(K^{14/15}\right)$ [Luo et al., 2021]
  - Adversarial linear mixture MDPs in full-information feedback setting: $\tilde{O}\left(\sqrt{K}\right)$ [He et al., 2022]

Question: does there exist an algorithm with $\tilde{O}\left(\sqrt{K}\right)$ regret for RL with linear function approximation and adversarial losses in bandit feedback setting?

- Jin et al. Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition. ICML, 2020.
- Luo et al. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. NeurIPS, 2021.
- He et al. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. AISTATS, 2022.

Our contribution

- A new algorithm, termed as LSUOB-REPS, for adversarial linear mixture MDPs in the bandit feedback setting
- We prove $\tilde{O}(dS^2\sqrt{K} + \sqrt{HSAK})$ regret upper bound for LSUOB-REPS
- An $\Omega(dH\sqrt{K} + \sqrt{HSAK})$ regret lower bound is also provided

Algorithm	Model	Feedback	Regret
Shifted Bandit UC-O-REPS (Rosenberg & Mansour, 2019a)	Tabular MDPs	Bandit Feedback	$\tilde{O}\left(H^{3/2}SA^{1/4}K^{3/4}\right)$
UOB-REPS (Jin et al., 2020a)	Tabular MDPs	Bandit Feedback	$\tilde{O}\left(HS\sqrt{AK}\right)$
OPPO (Cai et al., 2020)	Linear Mixture MDPs	Full-information	$\tilde{O}\left(dH^2\sqrt{K}\right)$
POWERS (He et al., 2022)	Linear Mixture MDPs	Full-information	$\tilde{O}\left(dH^{3/2}\sqrt{K}\right)$
LSUOB-REPS **(Ours)**	Linear Mixture MDPs	Bandit Feedback	$\tilde{O}\left(dS^2\sqrt{K} + \sqrt{HSAK}\right)$ $\Omega\left(dH\sqrt{K} + \sqrt{HSAK}\right)$

Adversarial MDPs - Setting

- An adversarial MDP $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=0}^{H-1}, \{\ell_k\}_{k=1}^{K}\right)$
- In each episode $k = 1,2, \dots, K$:
  - In each step $h = 0,2, \dots, H - 1$:
    - Observes state s, and takes action a $\sim \pi_k(\cdot\,|s)$
    - Then observes loss $\ell_k(s, a)$, and transits to next-state s$' \sim P_{h+1}(\cdot\,|s, a)$

Adversarial MDPs - Setting

- Particularly, transition P_h in linear mixture MDPs satisfies

$$P_h(s' \mid s, a) = \langle \phi(s' \mid s, a), \boldsymbol{\theta}_h^* \rangle$$

where ϕ is a known feature mapping and $\boldsymbol{\theta}_h^*$ is an unknown d-dimensional vector

Adversarial MDPs - Setting

- Let $\ell_k(\pi)$ be the expected loss of policy π in the k-th episode
- Learning objective: minimize the cumulative regret

$$R(K) = \sum_{k=1}^{K} \ell_k(\pi_k) - \sum_{k=1}^{K} \ell_k(\pi^*),$$

where $\pi^* \in \mathrm{argmin}_{\pi \in \Pi} \sum_{k=1}^{K} \ell_k(\pi)$ is the optimal policy in hindsight

Method

- High-level idea:
  - Maintain an ellipsoid confidence set $\mathcal{P}_{k,h}$ for P_h
  - Perform online mirror descent (OMD) over the occupancy measure space
  - Use an optimistically biased loss estimator in OMD

Key technical challenge!

Method

- To construct $\mathcal{P}_{k,h}$ and control the error of occupancy measure:
  - We do not use the *value-targeted regression* (VTR) scheme
  - Instead, we learn $\boldsymbol{\theta}_h^*$ via solving

$$\boldsymbol{\theta}_{k,h} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{k} \left[\left\langle \boldsymbol{\phi}\left(s'_{i,h+1} \mid s_{i,h}, a_{i,h}\right), \boldsymbol{\theta}\right\rangle - \boldsymbol{\delta}_{s_{i,h+1}}\left(s'_{i,h+1}\right)\right]^2 + \lambda \parallel \boldsymbol{\theta} \parallel_2^2 ,$$

where $s'_{i,h+1}$ is called as the *imaginary* next state

- Particularly, $s'_{k,h+1}$ is chosen to be

$$s'_{k,h+1} \in \operatorname{argmax}_{s\in\mathcal{S}_{h+1}} \left\|\boldsymbol{\phi}\left(s \mid s_{k,h}, a_{k,h}\right)\right\|_{\boldsymbol{M}_{k-1,h}^{-1}} ,$$

where $\boldsymbol{M}_{k,h} = \sum_{i=1}^{k} \boldsymbol{\phi}\left(s'_{i,h+1} \mid s_{i,h}, a_{i,h}\right)\boldsymbol{\phi}\left(s'_{i,h+1} \mid s_{i,h}, a_{i,h}\right)^\top + \lambda \boldsymbol{I}$ is the feature covariance matrix

Concluding Remarks

- Contribution:
  - We propose the LSUOB-REPS algorithm, for adversarial linear mixture MDPs in the bandit feedback setting, based on a new regression scheme
  - We prove $\tilde{O}\left(dS^2\sqrt{K} + \sqrt{HSAK}\right)$ regret upper bound for LSUOB-REPS
  - An $\Omega(dH\sqrt{K} + \sqrt{HSAK})$ regret lower bound is also provided
- Thank you!