Horizon-free Regret for Linear Markov Decision Process

Zihan Zhang[1], Jason D. Lee[1], Yuxin Chen[2], Simon S. Du[3]

[1] Princeton University
[2] University of Pennsylvania
[3] University of Washington

Motivation

- Horizon-free regret bound for RL
  - Positive answer for tabular MDP (Zhang et al., 2021, 2022);
  - Beyond tabular MDP?
  - Positive answer for linear MDP (**this work**).

Main Result

Theorem. For any linear MDP with totally bounded reward function, there exists an algorithm with regret bound $\tilde{O}(d^{5.5}\sqrt{K} + d^{6.5})$ w.h.p..

- Horizon-free regret bound for linear MDP
  - Without learning the transition model precisely;
  - Sharing data efficiently among different horizon.

Problem Settings

- Markov Decision Process (MDP)
  - State-action space $\mathcal{S} \times \mathcal{A}$
  - Reward function $R := \{R(s, a)\}$
  - Transition model $P =: \{P(\cdot \mid s, a)\}$
- Goal: find a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize $\mathbb{E}_{\pi} \left[\sum_{h=1}^{H} R(s_h, a_h) \right]$

Problem Settings

- Linear MDP
  - State-action space $\mathcal{S} \times \mathcal{A}$ with feature $\{\phi(s, a)\}$
  - Reward function $R := \{R(s, a)\}$ such that $R(s, a) = \left\langle \phi(s, a), \theta_r \right\rangle$
  - Transition model $P =: \{P(\cdot \,|\, s, a)\}$ such that $P(s' \,|\, s, a) = \langle \phi(s, a), \mu(s') \rangle$
  - Goal: find a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize $\mathbb{E}_\pi \left[\sum_{h=1}^{H} R(s_h, a_h) \right]$

Algorithm

- Construct confidence region for the transition kernel $\{\mu(s')\}_{s' \in \mathcal{S}}$
  - Construct confidence interval for $\left\langle \phi, \mu^\top v \right\rangle$ with least square regression
    - ϕ : possible feature vector
    - v : possible choice of optimal value function (low intrinsic dimension)
- Construct confidence region for θ_r (see VOFUL in [Zhang et al., 2021b])
- Plan optimistically according to the confidence regions

Technical Ideas

- Hardness
  - The sum of bonus due to inconsistent value functions $\{V_h^*\}_{h\in[H]}$;
    - Inconsistent value function leads to inconsistent information matrix.
- Solution
  - Dividing $[H]$ into different intervals $[H] = \cup_i \mathscr{H}_i$;
  - Prove that $\{V_h^*\}_{h\in\mathscr{H}_i}$ is nearly consistent measured by total variance.

Conclusion

- Future direction
  - Extend the results to RL with general function approximation
  - Improve the dependence on d

Thanks