Robust Offline Reinforcement Learning via Lipschitz Value Function

Tengyang Xie

tx10@illinois.edu

Overview

Goal: Provably Robust RL against (adversarial) perturbations on state observations.

Results:

1. On robustness definition: Using Lipschitz Q-function is sufficient against perturbations on states (done: formally proved)
2. Lipschitz policy evaluation

   (method proposed with basic guarantee. future work: detailed analysis)
3. Robust offline RL (Lipschitz policy evaluation + Policy improvement) (future work)

Lipschitz Value Function

Fact: Lipschitz (transition + reward) $\Longrightarrow$ Lipschitz Q-Function [Asadi et al., 2018]

Theorem. [Policy improvement via Lipschitz Q-function is robust against perturbations on states]
Assume $(\widehat{Q}, \pi)$ *pair satisfies (1)* $\|\widehat{Q} - Q^{\pi}\|_{\infty} = \varepsilon_1$; *(2)* $\widehat{Q}$ *is L-Lipschitz. Let* $\pi_{\widehat{Q}} \circ \nu_{\varepsilon_2}$ *be the perturbed greedy policy w.r.t.* $\widehat{Q}$ *and an* ε_2*-bounded perturbation* ν_{ε_2}*, then*

$$V^{\pi_{\widehat{Q}} \circ \nu_{\varepsilon_2}}(s) \geq V^{\pi}(s) - \frac{2\varepsilon_1 + 2L\varepsilon_2}{1-\gamma}, \ \forall s \in \mathcal{S},$$

where $\pi \circ \nu_{\varepsilon_2}(s) = \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}(\nu_{\varepsilon_2}(s), a)$.

Lipschitz Value Function

What can we learn from that Theorem?

Policy improvement still holds approximately even with adversarial perturbations on states as long as:

(i) We can approximate Q^π accurately;

(ii) Approximated Q^π is Lipschitz continuous.

A robust RL framework based on that concept:

1. Policy evaluation with ensuring Lipschitz; (next page)
2. Policy improvement according to the estimated Lipschitz Q-function. (future work)

Lipschitz Policy Evaluation

A non-parametric Lipschitz policy evaluation approach:

Given batch data $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ (all the states are perturbed by ν_{ε_2}) and target policy π.

1. Set $q_{0,i} = r_i - \gamma L d(s_i, s_i'), \forall i \in [i]$
2. For $t = 1, 2, \ldots, T$
   - 2.1 $Q_t(s, a) = \max_{i \in [n]: a_i = a} (q_{t-1,i} - Ld(s, s_i) - L\varepsilon_2), \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}$
   - 2.2 $q_{t,i} = r_i + \gamma [\widehat{T}^\pi Q_t](s_i, a_i), \forall i \in [i]$

Theorem

Let $Q = Q_T$, $T \to \infty$, we have

$$\|Q_T - Q^\pi\|_\infty \leq \frac{2L(\varepsilon_{\mathcal{S}} + \varepsilon_2)}{1 - \gamma},$$

where $\varepsilon_{\mathcal{S}} \coloneqq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \min_{i \in [n]: a_i = a} d(s, s_i)$, and ε_2 is the perturbation bound.

Reference

Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 264–273, 2018.