Breaking the Curse of Horizon: Infinite-horizon Off-policy Estimation

Qiang Liu\textsuperscript{1} \quad Lihong Li\textsuperscript{1} \quad Ziyang Tang\textsuperscript{1} \quad Dengyong Zhou\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Texas at Austin \quad \textsuperscript{2}Google Brain

\textbf{The Problem}

- Not always possible to deploy \& run a new RL policy because of cost, risk, ethics, or legal concerns:
  - Healthcare: treatment effect
  - Robotic \& Control: Web: recommendation, advertising, search

- Question: Can we evaluate a new policy \( \pi \) only using data from old policy \( \pi_0 \)?
- Given trajectories \( D_m = \{ \tau^{(i)} \}_{1 \leq i \leq m} \), where \( \tau = \{(s_t, a_t, r_t)\}_{0 \leq t \leq T} \)
  \( a_t \sim \pi_0(\cdot|s_t) \)
- Want to estimate "value" of the target policy \( \pi \):
  \[ R_\pi := \lim_{T \to \infty} \frac{\mathbb{E}_{\tau \sim \pi} [R^T(\tau)]}{\mathbb{E}_{\tau \sim \pi_0} [R^T(\tau)]} \]

\textbf{The Curse}

- Importance Sampling (Basic Inverse Propensity Score estimator):
  \[ R^T_\pi = \mathbb{E}_{\tau \sim \pi_0} \frac{\prod_{t=0}^T \pi(a_t|s_t)}{\pi_0(a_t|s_t)} R^T(\tau) \]  

- The Curse of Horizon: variance can grow exponentially.  

\textbf{Motivated Example}

- ‘Circle’ MDP:
  - Two actions: counterclockwise and clockwise
  - Deterministic transitions
  - As \( T \to \infty \):
    - IS/DR variance goes to \( \infty \)
    - But both policies visit every state equally often

\textbf{The Magic}

- Consider \( d_\pi(s) \) as marginal visiting prob. of state \( s \) under policy \( \pi \).
  - Rewriting:
  \[ R_\pi = \mathbb{E}_{s \sim d_\pi, a \sim \pi_0(\cdot|s)} \left[ \frac{d_\pi(s)}{\pi_0(a|s)} \pi(a|s) r(s, a) \right] \]  

- Now importance ratio no longer depends on \( T \).

\textbf{Theorem}

Define
\[ L(w, f) := \mathbb{E}_{(s, a, s') \sim d_\pi} \left[ \Delta(w; s, a, s') f(s') \right] \]
We have
\[ L(w, f) = 0, \quad \forall f \iff w \propto \frac{d_\pi(s)}{d_0(s)} \]  
where \( \Delta(w; s, a, s') = w(\pi(a|s) \pi_0(\cdot|s)) - w(s') \)

\textbf{The Algorithm}

1. Solve \( \hat{w} = \min_{w \in F} L(w, f, D_n) \)

2. Estimate \( \hat{R}_\pi = \mathbb{E}_{s \sim \pi_0, a \sim \hat{w}} \left[ \Delta(w; \pi(a|s), \pi_0(\cdot|s)) r(s, a) \right] \)

\textbf{The Results}

\textbf{Estimate} \( \hat{w} \)

- Pendulum
  - Mixing Ratio \( \alpha \) (Average case)
  - Discount Factor \( \gamma \) (Discount case)

\textbf{Traffic control (with SUMO simulator)}

\textbf{Acknowledgment}

This work is supported in part by NSF CRII 1830161. We would like to acknowledge Google Cloud for their support.