Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation

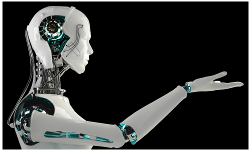
Qiang Liu† Lihong Li‡ Ziyang Tang† Dengyong Zhou‡

† Department of Computer Science, The University of Texas at Austin
‡ Google Brain (KIR)

Off-Policy Reinforcement Learning

- **Off-Policy Evaluation**: Evaluate a new policy π by only using data from old policy π_0.
- Widely useful when running new RL policies is costly or impossible, due to high cost, risk, or ethics, legal concerns:

Healthcare



Robotic & Control

Advertisement, Recommendation

"Curse of Horizon"

- **Importance Sampling (IS)**: Given trajectory $\boldsymbol{\tau} = \{s_t, a_t\}_{t=1}^{T} \sim \pi_0$,

$$R_\pi = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_0}\left[w(\boldsymbol{\tau})R(\boldsymbol{\tau})\right], \qquad \text{where} \qquad w(\boldsymbol{\tau}) = \prod_{t=0}^{T} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}$$

- The Curse of Horizon:
  - The IS weights $w(\boldsymbol{\tau})$ are **product of T terms**; T is horizon length.
  - Variance can **grow exponentially with T**.
  - **Problematic for infinite horizon problems ($T = \infty$).**

Breaking the Curse

- **Key: Apply IS on (s, a) pairs, not the whole trajectory τ:**

$$R_{\pi} = \mathbb{E}_{(s,a)\sim d_{\pi_0}}\left[w(s,a)r(s,a)\right], \quad \text{where} \quad w(s,a) = \frac{d_{\pi}(s,a)}{d_{\pi_0}(s,a)},$$

  where $d_{\pi}(s, a)$ is the *stationary / average visitation distribution of* (s, a) *under policy* π.

- **Stationary density ratio** $w(s, a)$*:*
  - *is* **NOT** *product of T terms.*
  - *can be* **small even for infinite horizon ($T = \infty$)**.
  - *But is* **more difficult to estimate**.

Main Algorithm

1. 1.Estimate density ratio by a **new minimax objective**:

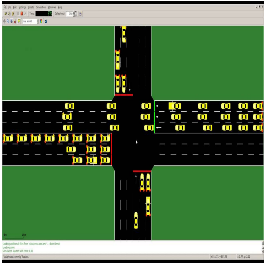
$$\hat{w} = \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} \hat{L}(w, f, \mathcal{D}_{\pi_0})$$

2. 2. Value estimation by IS:

$$\hat{R}_{\pi} = \hat{\mathbb{E}}_{(s,a)\sim d_{\pi_0}} [\hat{w}(s,a) r(s,a)]$$

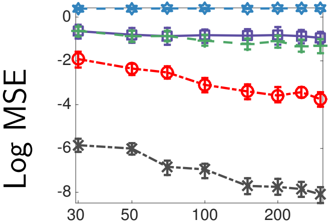
- **Theoretical guarantees** developed for the new minimax objective.
- Can be **kernelized**: Inner max has closed form if $\mathcal{F}$ is an RKHS.

Empirical Results

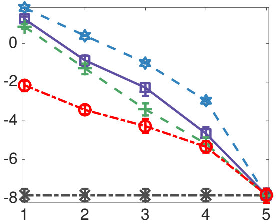


Traffic control

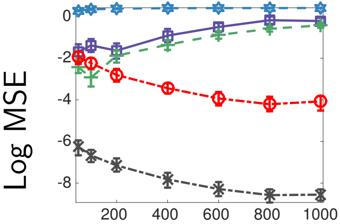
(using SUMO simulator[5])



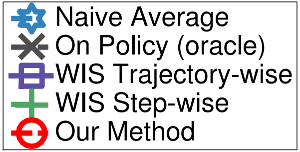
(a) # of Trajectories (n)



(b) Different Behavior Policies



(c) Truncated Length T



Thank You!

Location: **Room 210 & 230 AB; Poster #121**
Time: **Wed Dec 5th 05:00 – 07:00 PM**

References & Acknowledgment

[1] [HLR'16] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval.

[2] [JL16] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning.

[3] [LMS'15] L. Li, R. Munos, and Cs. Szepesvari. Toward minimax off-policy value estimation.

[4] [TB'16] P.S. Thomas and E. Brunskill. Data-efficient off-Policy policy evaluation for reinforcement learning.

[5] [KEBB'12] D. Krajzewicz, J.Erdmann, M.Behrisch and L.Bieker. Recent development and applications of SUMO-Simulation of Urban MObility.

Work supported in part by NSF CRII 1830161 and Google Cloud.