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

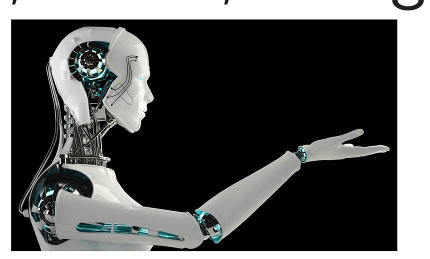
Qiang Liu[†] Lihong Li[‡] Ziyang Tang[†] Dengyong Zhou[‡]

[†]Department of Computer Science, University of Texas at Austin

[‡]Google Brain

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 π only using data from old policy π_0?
- Given trajectories $\mathcal{D}_{\pi_0} = \{\boldsymbol{\tau}^{(i)}\}_{1\le i\le m}$. where $\boldsymbol{\tau} = \{(s_t, a_t, r_t)\}_{0\le t\le T}$

  $a_t \sim \pi_0(\cdot|s_t)$
- Want to estimate "value" of the target policy π:

$$R_\pi := \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\tau}\sim\pi}[R^T(\boldsymbol{\tau})],\ R^T(\boldsymbol{\tau}) := (\sum_{t=0}^T \gamma^t r_t)/(\sum_{t=0}^T \gamma^t),$$

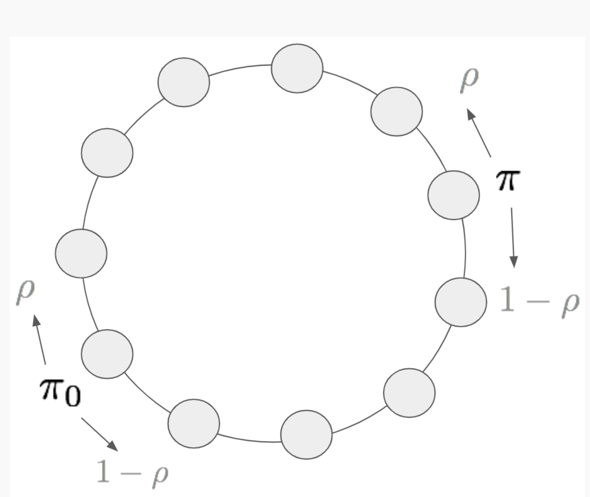
The Curse

- Importance Sampling(Basic Inverse Propensity Score estimator):

$$R_\pi^T = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_0}[\prod_{t=0}^T \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)} R^T(\boldsymbol{\tau})]. \tag{1}$$

- The Curse of Horizon: variance can grow exponentially.

Motivated Example



'Circle' MDP:
- Two actions: counterclockwise and clockwise
- Deterministic transitions

As $T\to\infty$:
- IS/DR variance goes to ∞
- But both policies visit every state equally often

The Magics

- Consider $d_\pi(s)$ as marginal visiting prob. of state s under policy π.

Rewriting:
$$R_\pi = \mathbb{E}_{s\sim d_{\pi_0}, a\sim\pi_0(\cdot|s)}\left[\frac{d_\pi(s)}{d_{\pi_0}(s)}\frac{\pi(a|s)}{\pi_0(a|s)} r(s,a)\right], \tag{2}$$

- Now importance ratio no longer depends on T.

Theorem

Define
$$L(w,f) := \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[\Delta(w;\ s,a,s')f(s')\right]. \tag{3}$$
We have
$$L(w,f) = 0,\ \forall f \iff w \propto \frac{d_\pi(s)}{d_{\pi_0}(s)}, \tag{4}$$
where $\Delta(w;\ s,a,s') = w(s)\frac{\pi(a|s)}{\pi_0(a|s)} - w(s')$

The Algorithm

Algorithm

1.Solve $\hat{w} = \min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} \hat{L}(w,f,\mathcal{D}_{\pi_0})$ (5)

2.Estimate $\hat{R}_\pi = \hat{\mathbb{E}}_{(s,a)\sim d_{\pi_0}}[\hat{w}(s)\frac{\pi(a|s)}{\pi_0(a|s)} r(s,a)]$ (6)

- Inner max: similar to GAN discriminator
- If we take $\mathcal{F}$ to be RKHS with kernel k, will have closed form solution:

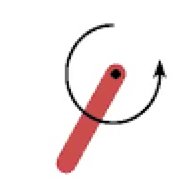
$$\max_{f\in\mathcal{F}} L(w,f)^2 = \mathbb{E}_{d_{\pi_0}}\left[\Delta(w;\ s,a,s')\Delta(w;\ \bar{s},\bar{a},\bar{s}')k(s',\bar{s}')\right] \tag{7}$$

- Same idea applies to discounted reward case

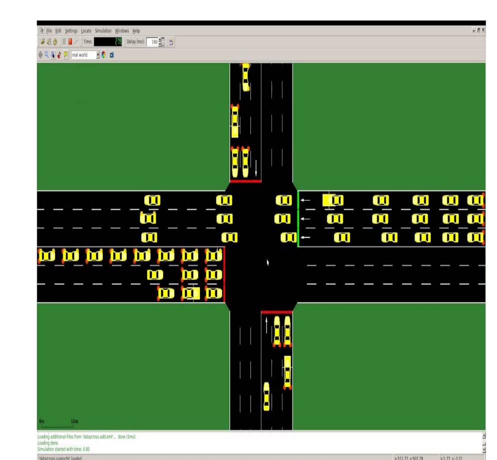
The Results

Estimate $\hat{w}$

(taxi environment)

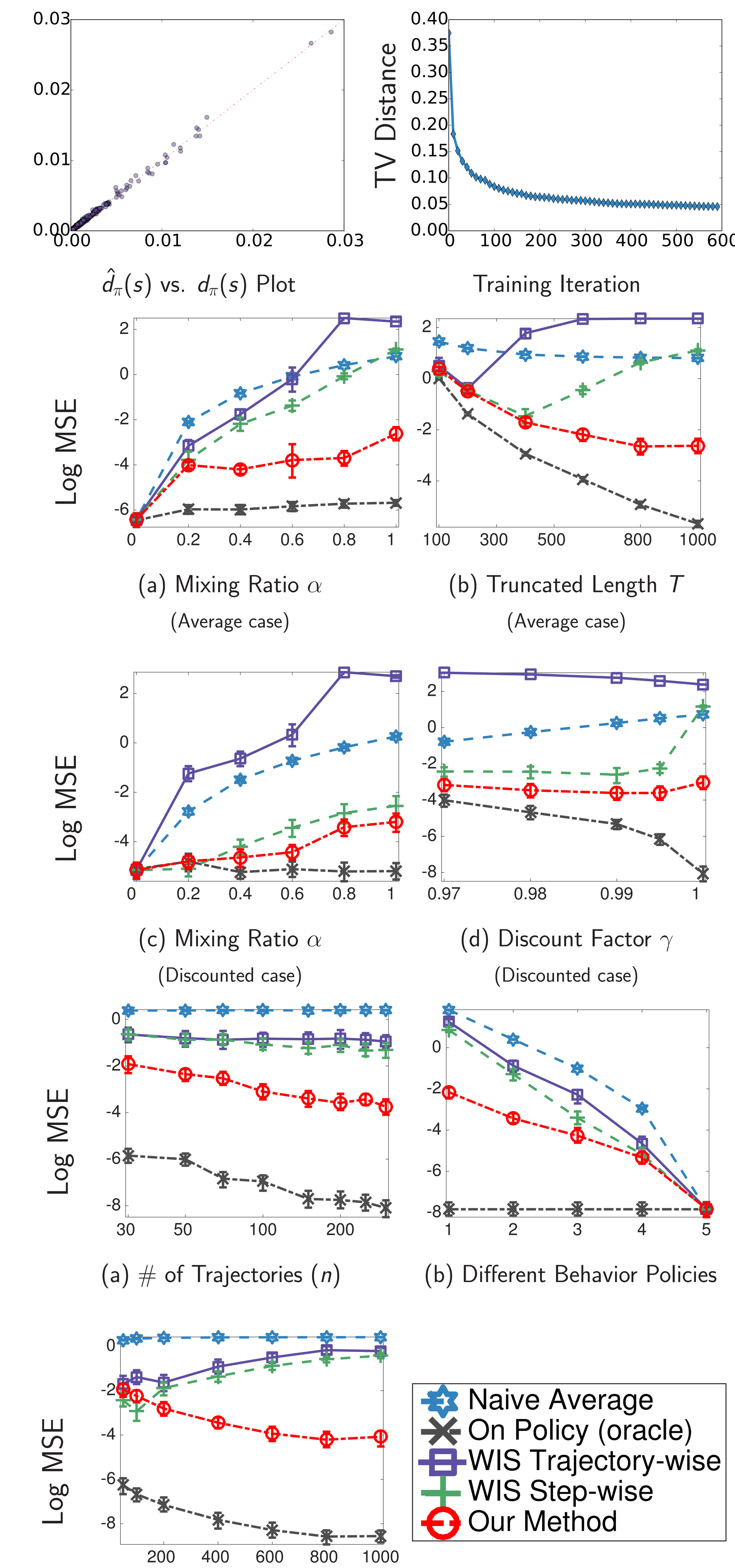


Pendulum



Traffic control

(with SUMO simulator)



$\hat{d}_\pi(s)$ vs. $d_\pi(s)$ Plot

(a) Mixing Ratio α (Average case)

(b) Truncated Length T (Average case)

(c) Mixing Ratio α (Discounted case)

(d) Discount Factor γ (Discounted case)

(a) # of Trajectories (n)

(b) Different Behavior Policies

(c) Truncated Length T

Acknowledgment

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

email: lqiang,ztang@cs.utexas.edu, lihong,dennyzhou@google.com