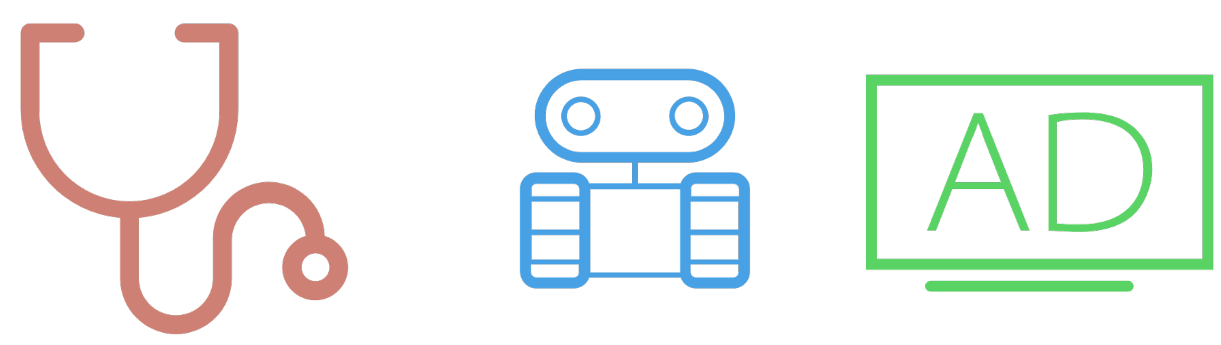
Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation

Ziyang Tang★[1] Yihao Feng★[1] Lihong Li [2] Denny Zhou[2] Qiang Liu [1]

[1]UT Austin [2]Google Research

Infinite-Horizon Off-Policy Evaluation

- **Problem**: Evaluate a target policy π given data from behavior policy π_0.
- **Wide Applications:** whenever evaluating new policies is costly or impossible, due to **high cost, risk, or ethics, legal concerns**.



Healthcare Robotic & Control Recommendation

- **Setup**: given behavior trajectories $\{s_i, a_i, s_i', r_i\} \sim \pi_0$, we want to estimate the average discounted reward of π

$$R^\pi = \mathbb{E}_{\tau\sim\pi}[\frac{\sum_{t=0}^{\infty}\gamma^t r_t}{\sum_{t=0}^{\infty}\gamma^t}] = (1-\gamma)\mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]\,.$$

Bellman Equation

- **Value function:** Define $\mathcal{P}^\pi f(s) = \sum_{a,s'}\pi(a|s)T(s'|s,a)f(s')$, we have the Bellman equation for value function

$$V^\pi = r^\pi + \gamma\mathcal{P}^\pi V^\pi.$$

- **Density function:** Define $\mathcal{T}^\pi f(s') = \sum_{a,s}\pi(a|s)T(s'|s,a)f(s)$, we have the Bellman equation for density function [1]

$$d_\pi = (1-\gamma)\mu_0 + \gamma\mathcal{T}^\pi d_\pi.$$

Basic Estimators

- **Estimation via Value Function** (*Direct Methods*):

$$R^\pi = (1-\gamma)\mathbb{E}_{s\sim\mu_0}[V^\pi(s)]\,.$$

Given a learned $\hat{V}$, we have the following value function estimator

$$\hat{R}^\pi_{\mathsf{VAL}}[\hat{V}] = \frac{(1-\gamma)}{n_0}\sum_{i=1}^{n_0}\hat{V}(s_0^{(i)}).$$

- **Estimation via State Density Function** (*IS methods*)[1]:

$$R^\pi = \mathbb{E}_{s,a\sim d_{\pi_0}}[w_{\pi/\pi_0}(s)\frac{\pi(a|s)}{\pi_0(a|s)}r(s,a)]\,,$$

where $w_{\pi/\pi_0}(s) = \frac{d_\pi(s)}{d_{\pi_0}(s)}$ is the state density ratio function.
Given a learned density ratio $\hat{w}$, we have the IS estimator

$$\hat{R}^\pi_{\mathsf{SIS}}[\hat{w}] = \frac{1}{n}\sum_{i=1}^{n}\hat{w}(s_i)\frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)}r_i.$$

Doubly Robust Estimator

- Combine the estimators $R^\pi_{\mathsf{SIS}}[\widehat{w}]$ and $R^\pi_{\mathsf{VAL}}[\widehat{V}]$, we get the "**doubly robust**" estimator

$$R^\pi_{\mathsf{DR}}[\widehat{V},\widehat{w}] = \underbrace{\sum_s r^\pi(s)d_{\pi_0}(s)\widehat{w}(s)}_{R^\pi_{\mathsf{SIS}}[\widehat{w}]} + \underbrace{(1-\gamma)\sum_s\widehat{V}(s)\mu_0(s)}_{R^\pi_{\mathsf{VAL}}[\widehat{V}]} \underbrace{-\sum_s\left(\widehat{V}(s)-\gamma\mathcal{P}^\pi\widehat{V}(s)\right)d_{\pi_0}(s)\widehat{w}(s)}_{R^\pi_{\mathsf{bridge}}[\widehat{V},\widehat{w}]}\,.$$

- **Theorem: Double Robustness**

$$R^\pi_{\mathsf{DR}}[\widehat{V},\widehat{w}] - R^\pi = \mathbb{E}_{s\sim d_{\pi_0}}\left[\varepsilon_{\widehat{w}}(s)\varepsilon_{\widehat{V}}(s)\right]\,,$$

$$\varepsilon_{\widehat{w}} = \frac{d_\pi(s)}{d_{\pi_0}(s)} - \widehat{w}(s), \qquad \varepsilon_{\widehat{V}}(s) = \widehat{V}(s) - r^\pi(s) - \gamma\mathcal{P}^\pi\widehat{V}(s)\,.$$

Double Robustness and Lagrangian Duality

Primal Problem

$$R^\pi = \min_V\left\{(1-\gamma)\sum_s\mu_0(s)V(s) \quad \text{s.t.} \quad V(s)\geq r^\pi(s)+\gamma\mathcal{P}^\pi V(s),\ \ \forall s\right\}$$

$\Downarrow$ Lagrangian multiplier $\rho(s)\geq 0$

Lagrangian function

$$L(V,\rho) = \begin{cases}(1-\gamma)\sum_s\mu_0(s)V(s)+\sum_s\rho(s)r^\pi(s)-\sum_s\rho(s)\cdot(I-\gamma\mathcal{P}^\pi)V(s)\\ \mathcal{P}^\pi \text{ and } \mathcal{T}^\pi \text{ are self-adjoint!} \Updownarrow \\ \sum_s\rho(s)r^\pi(s)+(1-\gamma)\sum_s\mu_0(s)V(s)-\sum_s(I-\gamma\mathcal{T}^\pi)\rho(s)\cdot V(s)\end{cases}$$

$L(V,\rho)$ is the DR estimator!

$$w_{\rho/\pi_0} = \rho(s)/d_{\pi_0}, \quad R^\pi_{\mathsf{DR}}[V, w_{\rho/\pi_0}] = L(V,\rho)$$

$\Uparrow$ Lagrangian multiplier $V(s)$

Dual Problem

$$R^\pi = \max_{\rho\geq 0}\left\{\sum_0\rho(s)r^\pi(s) \quad \text{s.t.} \quad \rho(s') = (1-\gamma)\mu_0(s')+\gamma\mathcal{T}^\pi\rho(s'),\ \ \forall s'\right\}$$

- **Application I:** Extension to $\gamma=1$ (Average Case)

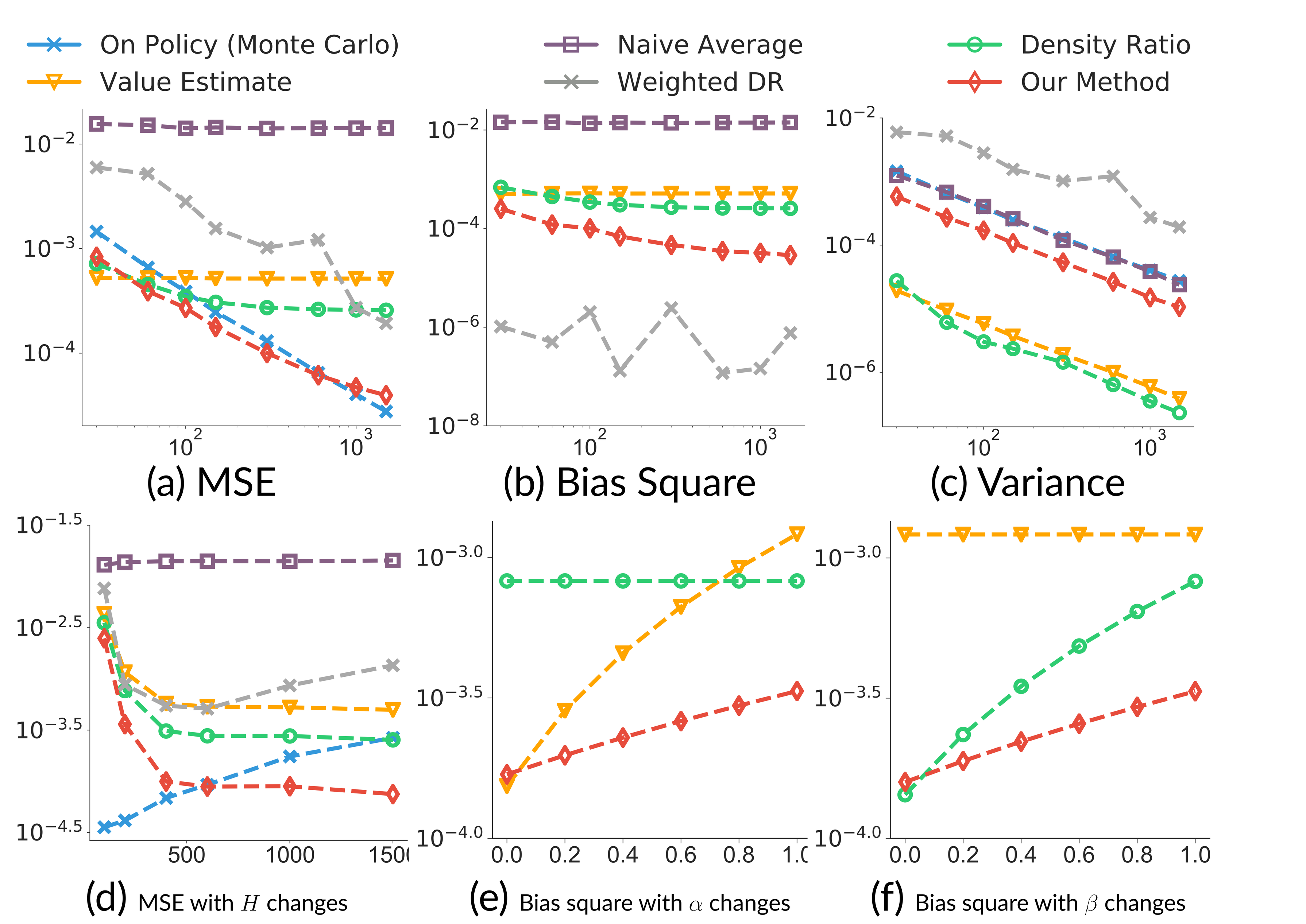
$$R^\pi_{\mathrm{DR}}[\hat{V},\hat{w}] := \frac{\mathbb{E}_{s\sim d_{\pi_0}}\left[\hat{w}(s)\left(r^\pi(s)-\hat{V}(s)+\mathcal{P}^\pi\hat{V}(s)\right)\right]}{\mathbb{E}_{s\sim d_{\pi_0}}[\hat{w}(s)]}\,.$$

- **Application II** : Extension to Q-value functions:

$$R^\pi_{\mathsf{DR}}[Q,x] = \underbrace{(1-\gamma)\mathbb{E}_{s\sim\mu_0,a\sim\pi(\cdot|s)}[Q(s,a)]}_{R^\pi_{\mathsf{VAL}}[Q]} + \underbrace{\mathbb{E}_{s,a\sim\mathcal{D}}[x(s,a)r(s,a)]}_{R^\pi_{\mathsf{SAIS}}[x]} \underbrace{-\mathbb{E}_{s,a,s'\sim\mathcal{D},a'\sim\pi(\cdot|s')}\left[x(s,a)\left(Q(s,a)-\gamma Q(s',a')\right)\right]}_{R^\pi_{\mathsf{bridge}}[Q,x]}\,.$$

Experimental Results

Taxi Environment



- **Related DR method**: contextual bandit [2], finite-horizon RL[3], concurrent work [4].
- **Acknowledgment**: This work is supported in part by NSF CRII 1830161 and NSF CAREER 1846421.

[1] Q. Liu, L. Li, Z. Tang, D. Zhou, Breaking the curse of horizon: Infinite-horizon off-policy estimation, in: Advances in Neural Information Processing Systems, 2018, pp. 5361–5371.
[2] M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML), 2011, pp. 1097–1104.
[3] N. Jiang, L. Li, Doubly robust off-policy evaluation for reinforcement learning, in: Proceedings of the 23rd International Conference on Machine Learning (ICML), 2016, pp. 652–661.
[4] N. Kallus, M. Uehara, Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes, arXiv preprint arXiv:1909.05850 (2019).