Doubly Robust Bias Reduction in Infinite-horizon Off-policy Estimation

Ziyang Tang[1]* Yihao Feng[1]*
Lihong Li[2] Dengyong Zhou[2] Qiang Liu[1]

[1]University of Texas at Austin

[2]Google Research

Overview

- Existing (model free) methods:

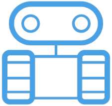
Methods	Variance	Bias
Value Based	**Low**	Large Bias
Importance sampling on trajectory	Very High	**Unbiased**
Original Doubly Robust (JL'16)	High	**Unbiased**
Density Based(LLTD'18)	**Low**	Biased
This work	**Low**	**Small Bias**

- The two unbiased estimator suffers from **the curse of horizon**(LLTD'18).

- **Off-Policy Evaluation(OPE)**: Evaluate a new policy by only using historical data.

- Widely useful when running new RL policies is costly or impossible, due to high cost, risk, or ethical/legal concerns.

Medical



Robotic

Recommendation

Problem Setting

- **Infinite Horizon OPE**: Let R^π be the average discounted reward for policy π:

$$R^\pi = \mathbb{E}_{\tau \sim \pi}\left[\frac{\sum_{t=0}^{\infty} \gamma^t r_t}{\sum_{t=0}^{\infty} \gamma^t}\right],$$

where $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, ...)$ is one trajectory from policy π.

- Two ways to rewrite the formulation of R^π:

1. **Value-based Formula:**

$$R^\pi = (1-\gamma)\sum_s \mu_0(s) V^\pi(s).$$

2. **Density-based Formula:**

$$R^\pi = \sum_s d_\pi(s) r^\pi(s),$$

Both Estimators are Biased

- Two (low variance) estimators:
  1. Value-based Estimation, find $V \approx V^\pi$, approximate R^π as
  $$R^\pi_{\mathrm{VAL}}[V] := (1-\gamma)\sum_s \mu_0(s)V(s).$$
  2. Density-based Estimation, find $\rho \approx d_\pi$(LLTD'18), approximate R^π as
  $$R^\pi_{\mathrm{DEN}}[\rho] := \sum_s \rho(s) r^\pi(s).$$
- If $V = V^\pi$, value-based estimation is unbiased; If $\rho = d_\pi$, density-based estimation is unbiased.
- In general, **both estimators are biased**!

Doubly Robust Estimation

- Our estimation: find $V \approx V^\pi$, $\rho \approx d_\pi$, approximate R^π as

$$R^\pi_{\mathrm{DR}}[V,\rho] := R^\pi_{\mathrm{VAL}}[V] + R^\pi_{\mathrm{DEN}}[\rho] - \underbrace{\sum_s \rho(s)\left(I - \gamma\mathcal{P}^\pi\right)V(s)}_{R^\pi_{\mathrm{conn}}[V,\rho]}$$

- The third term try to cancel out the "doubly worse" part.
- Double robustness:
  "if **either** $V = V^\pi$ **or** $\rho = d_\pi$ our estimator is **unbiased**."

Reduce the Bias

- Bias of value-based estimation and density-based estimation:

$$R^{\pi}_{\mathrm{VAL}}[V] - R^{\pi} = \sum_{s} d_{\pi}(s)\varepsilon_{\boldsymbol{V}}(s), \quad R^{\pi}_{\mathrm{DEN}}[\rho] - R^{\pi} = \sum_{s} \varepsilon_{\boldsymbol{\rho}}(s) r^{\pi}(s).$$

where,

$$\varepsilon_{\boldsymbol{V}}(s) = V(s) - r^{\pi}(s) - \gamma \mathcal{P}^{\pi} V(s), \qquad \varepsilon_{\boldsymbol{\rho}}(s) = \rho(s) - d_{\pi}(s)\,.$$

- Bias of doubly robust estimation:

$$R^{\pi}_{\mathrm{DR}}[V, \rho] - R^{\pi} = \sum_{s} \varepsilon_{\boldsymbol{\rho}}(s)\varepsilon_{\boldsymbol{V}}(s),$$

Optimization Framework

Primal optimization formulation of policy evaluation

$$\min_{V} \quad \underbrace{\sum_{s} (1-\gamma)\mu_0(s)V(s)}_{:=R^{\pi}_{\text{VAL}}[V]}$$

$$\text{s.t.} \quad V \geq r^{\pi} + \gamma \mathcal{P}^{\pi} V,$$

where $\mathcal{P}^{\pi}$ is an forward operator:

$$\mathcal{P}^{\pi} f(s) = \sum_{s',a} \pi(a|s) T(s'|s,a) f(s').$$

The dual formula corresponds to density learning

$$\max_{\rho \geq 0} \quad \underbrace{\sum_{s} \rho(s) r^{\pi}(s)}_{:=R^{\pi}_{\text{DEN}}[\rho]}$$

$$\text{s.t.} \quad \rho = (1-\gamma)\mu_0 + \gamma \mathcal{T}^{\pi} \rho,$$

where $\mathcal{T}^{\pi}$ is an backward operator

$$\mathcal{T}^{\pi} f(s') = \sum_{s,a} \pi(a|s) T(s'|s,a) f(s).$$

Lagrangian – Doubly Robust Estimator

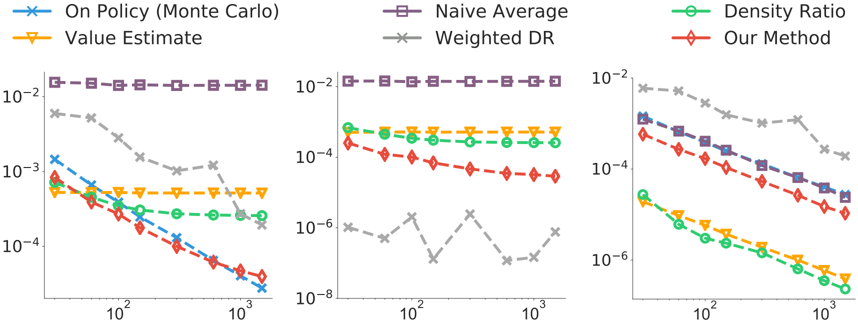
- Surprisingly, the Lagrangian function is a **Doubly Robust estimator!**

$$L(V,\rho) = (1-\gamma)\sum_s \mu_0(s)V(s) - \sum_s \rho(s)\left(V(s) - r^\pi(s) - \gamma\mathcal{P}^\pi V(s)\right)$$

$$= \underbrace{\sum_s (1-\gamma)\mu_0(s)V(s)}_{R^\pi_{\mathrm{VAL}}[V]} + \underbrace{\sum_s \rho(s) r^\pi(s)}_{R^\pi_{\mathrm{DEN}}[\rho]} - \underbrace{\sum_s \rho(s)\left(I - \gamma\mathcal{P}^\pi\right)V(s)}_{R^\pi_{\mathrm{conn}}[V,\rho]}$$

$$= R^\pi_{\mathrm{DR}}[V,\rho]$$

Taxi environment(LLTD'18).



(a) MSE (b) Bias Square (c) Variance

Thank You

References & Acknowledgment

[JL'16]
N. Jiang and L. Li.
Doubly robust off-policy value evaluation for reinforcement learning.

[LLTD'18]
Q. Liu, L. Li, Z. Tang and D. Zhou.
Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation.

Work supported in part by NSF CRII 1830161, NSF CAREER 1846421 and Google Cloud.