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

Ziyang Tang\textsuperscript{1*}  Yihao Feng\textsuperscript{1*}
Lihong Li\textsuperscript{2}  Dengyong Zhou\textsuperscript{2}  Qiang Liu\textsuperscript{1}

\textsuperscript{1}University of Texas at Austin
\textsuperscript{2}Google Research
Overview

- Existing (model free) methods:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Variance</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value Based</td>
<td>Low</td>
<td>Large Bias</td>
</tr>
<tr>
<td>Importance sampling on trajectory</td>
<td>Very High</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Original Doubly Robust (JL’16)</td>
<td>High</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Density Based (LLTD’18)</td>
<td>Low</td>
<td>Biased</td>
</tr>
<tr>
<td>This work</td>
<td>Low</td>
<td>Small Bias</td>
</tr>
</tbody>
</table>

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

- **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

Tang et al. (UT & Google)  Doubly-Robust
Infinite Horizon OPE: Let $R^\pi$ be the average discounted reward for policy $\pi$:

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

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

Two ways to rewrite the formulation of $R^\pi$:

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 $R^\pi$ as

     $$R^\pi_{\text{VAL}}[V] := (1 - \gamma) \sum_s \mu_0(s) V(s).$$

  2. **Density-based Estimation**, find $\rho \approx d_\pi$ (LLTD'18), approximate $R^\pi$ as

     $$R^\pi_{\text{DEN}}[\rho] := \sum_s \rho(s) r^\pi(s).$$

- 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 $R^\pi$ as

$$R^\pi_{\text{DR}}[V, \rho] := R^\pi_{\text{VAL}}[V] + R^\pi_{\text{DEN}}[\rho] - \sum_s \rho(s) (I - \gamma P^\pi) V(s)$$

- 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:

\[ R_{\text{VAL}}^{\pi}[V] - R^{\pi} = \sum_s d_\pi(s)\varepsilon_V(s), \quad R_{\text{DEN}}^{\pi}[\rho] - R^{\pi} = \sum_s \varepsilon_\rho(s)r_\pi(s). \]

where,

\[ \varepsilon_V(s) = V(s) - r_\pi(s) - \gamma P^{\pi} V(s), \quad \varepsilon_\rho(s) = \rho(s) - d_\pi(s). \]

- **Bias of** doubly robust estimation:

\[ R_{\text{DR}}^{\pi}[V, \rho] - R^{\pi} = \sum_s \varepsilon_\rho(s)\varepsilon_V(s), \]
Optimization Framework

Primal optimization formulation of policy evaluation

\[
\begin{align*}
\min_V & \quad \sum_s (1 - \gamma) \mu_0(s) V(s) \\
\text{s.t.} & \quad V \geq r^\pi + \gamma P^\pi V,
\end{align*}
\]

\(:= R^\pi_{\text{VAL}}[V]

where \( P^\pi \) is a forward operator:

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

The dual formula corresponds to density learning

\[
\begin{align*}
\max_{\rho \geq 0} & \quad \sum_s \rho(s) r^\pi(s) \\
\text{s.t.} & \quad \rho = (1 - \gamma) \mu_0 + \gamma T^\pi \rho,
\end{align*}
\]

\(:= R^\pi_{\text{DEN}}[\rho]

where \( T^\pi \) is a backward operator

\[
T^\pi f(s') = \sum_{s,a} \pi(a|s) T(s'|s,a) f(s).
\]
Surprisingly, the Lagrangian function is a **Doubly Robust estimator**!

\[
L(V, \rho) = (1 - \gamma) \sum_s \mu_0(s)V(s) - \sum_s \rho(s) (V(s) - r_\pi(s) - \gamma \mathcal{P}_{\pi} V(s))
\]

\[
= \sum_s (1 - \gamma)\mu_0(s)V(s) + \sum_s \rho(s)r_\pi(s) - \sum_s \rho(s)(1 - \gamma \mathcal{P}_{\pi}) V(s)
\]

\[
= R_{\text{VAL}}^{\pi}[V] + R_{\text{DEN}}^{\pi}[\rho] - R_{\text{conn}}^{\pi}[V, \rho]
\]

\[
= R_{\text{DR}}^{\pi}[V, \rho]
\]
Experimental Results

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

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