Accountable Off-Policy Evaluation via a Kernelized Bellman Statistics

Yihao Feng*, Tongzheng Ren*, Ziyang Tang, Qiang Liu

The University of Texas at Austin

ICML 2020

*Equal Contribution

Off Policy RL: Medical Treatment

States ($\mathcal{S}$): Physiological condition of patients

Actions ($\mathcal{A}$): Usage of medical treatments

Rewards ($\mathcal{R}$): Health recovery of patients

- On Policy: ☹ High Risk !
- Off Policy: ☺ Safe and data efficient !

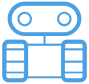
Off-Policy Evaluation (OPE)

- Given data $\mathcal{D} = \{s_i, a_i, r_i, s_i'\}_{i=1}^{n}$ collected with (unknown) behavior policy π_{BEH} to estimate the discounted total return of the target policy π:

$$\eta^{\pi} := \lim_{T\to\infty} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} \gamma^{t} r_t\right] = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t\right].$$

- Useful when on-policy data is not available.

Medical



Robotics

Recommendation

Existing OPE Estimations

Methods	Estimating η^π
Value Based	$\mathbb{E}_{x\sim\mu_0\times\pi}[Q^\pi(x)]$
Importance Sampling	$R(\tau)\prod_{t=1}^{T}\frac{\pi(a_t\|s_t)}{\pi_0(a_t\|s_t)}$
Density Based [LLTZ18]	$\mathbb{E}_{x\sim\mu}[\rho^\pi(x)r(x)]$
Doubly Robust [JL15] ,...	
Model-based [LGR+18], ...	
...	

For high-stakes decision applications, we may need more information to make sure that these estimators are reliable...

Motivation

More information for high-stake decisions:

- Data is limited, we need to quantify the uncertainty of the predictions.
- We want to determine the **lower** and the **upper** bounds for η^{π} with high probability.
- If possible, we can provide a post-hoc correction for existing OPE estimators, such that the corrected estimation is in the high confidence interval.

Existing Confidence Interval for OPE

Prior works [e.g. TTG15, HSN17]:

- Most bounds are Importance Sampling (IS) based.

☹ Behavior **aware** (need π_{BEH}).

☹ Trajectory based: data inefficient.

☹ Curse of horizon.

Our work:

- Value based bound.

☺ Behavior **agnostic** (black-box)!

☺ Transition based: data efficient!

☺ Work for long or infinite horizon MDP!

General Idea

- Value-based Estimator:

$$\eta^{\pi} = \mathbb{E}_{x \sim \mu_0 \times \pi}\left[Q^{\pi}(x) \right].$$

- Assume we have a feasible set $\mathcal{Q}_n$, such that:
  - $\mathbb{P}(Q^{\pi} \in \mathcal{Q}_n) \geq 1 - \delta$.
  - $\mathcal{Q}_n \to \{Q^{\pi}\}$ as $n \to \infty$.
- Define

$$\eta^{+}(\text{resp. } \eta^{-}) = \max_{Q \in \mathcal{Q}_n} (\text{resp. } \min_{Q \in \mathcal{Q}_n}) \ \mathbb{E}_{x \sim \mu_0 \times \pi}\left[Q(x) \right],$$

  thanks to the property of $\mathcal{Q}_n$, we have
  - $\mathbb{P}(\eta^{\pi} \in [\eta^{-}, \eta^{+}]) \geq 1 - \delta$.
  - $\eta^{-}, \eta^{+} \to \eta^{\pi}$ as $n \to \infty$.

Choice of $\mathcal{Q}_n$

We choose $\mathcal{Q}_n$ based on a consistent test statistics $\mathbb{D}(Q, Q^\pi)$ for the hypothesis $Q = Q^\pi$:

$$\mathcal{Q}_n = \left\{ Q \in \mathcal{F} \;\middle|\; \hat{\mathbb{D}}_n(Q, Q^\pi) \in \mathrm{CI}_{\delta,n} \right\},$$

where

- $\mathcal{F}$ is a proper function class that contains Q^π.
- $\hat{\mathbb{D}}_n(Q, Q^\pi)$ is the empirical estimate of $\mathbb{D}(Q, Q^\pi)$ with off-policy data.
- $\mathrm{CI}_{\delta,n}$ is the confidence interval of $\hat{\mathbb{D}}_n(Q^\pi, Q^\pi)$ under confidence level δ.

Choice of $\mathbb{D}(Q, Q^\pi)$

- Kernel loss [FLL19] as a test statistic:

$$\mathbb{D}_K(Q, Q^\pi) = L_K(Q) := \mathbb{E}_{x,\bar{x}\sim\mu}\left[\mathcal{R}_\pi Q(x) \cdot K(x,\bar{x}) \cdot \mathcal{R}_\pi Q(\bar{x})\right], \quad (1)$$

where $\mathcal{R}_\pi Q = \mathcal{B}_\pi Q - Q$ and

$$\mathcal{B}_\pi Q(s,a) := \mathbb{E}_{(s',a')\sim P(\cdot|s,a)\times\pi(\cdot|s')}\left[r(s,a) + \gamma Q(s',a')|s,a\right]. \quad (2)$$

- Why kernel loss?
  - Consistency:

$$L_K(\hat{Q}) = 0 \quad \Leftrightarrow \quad \hat{Q} = Q^\pi.$$

  - Can measure the deviation from $\hat{Q}$ to Q^π without the knowledge of Q^π!

Empirical Estimation and Concentration

- We can use the so-called *V-statistics* to estimate the kernel loss:

$$\hat{\mathbb{D}}_K^V(Q, Q^\pi) = \hat{L}_K^V(Q) = \frac{1}{n^2} \sum_{i,j \in [n]} \hat{\mathcal{R}}_\pi Q(x_i) \cdot K(x_i, x_j) \cdot \hat{\mathcal{R}}_\pi Q(x_j) \,.$$

Theorem (Concentration of V-statistics)

Assume the behaviour policy π_{BEH} is ergodic, then with probability $1 - \delta$,

$$\left|\hat{L}_K^V(Q) - L_K(Q)\right| \leq C\left(\frac{n-1}{n}\sqrt{\frac{\log 2/\delta}{n}} + \frac{1}{n}\right), \tag{3}$$

where C is an absolute constant that depends on the maximum reward $R_{\max}$, discounted factor γ and the maximum value of kernel K.

Choice of $\mathcal{F}$: RKHS

Consider the optimization problem:

$$\eta^+ = \max_{Q \in \mathcal{F}} \mathbb{E}_{x \sim \mu_0 \times \pi}[Q(x)], \quad \text{s.t.} \quad \hat{\mathbb{D}}_K^V(Q, Q^\pi) \le \lambda_{\delta,n},$$

where $\lambda_{\delta,n}$ is chosen via Eq (3).

- Assume $\mathcal{F}$ is a RKHS with dimension $d \le n$, can be effectively solved with convex optimization methods!
- Compatible with simple linear feature and more expressive features like neural feature.

Post-hoc Diagnosis

- Given a black-box estimate of Q^π, how can we know if it is accurate?
- Check if

$$\hat{\mathbb{D}}_K^V(\hat{Q}, Q^\pi) \leq \lambda_{\delta,n} \,.$$

  - Yes: $\hat{Q}$ is a reasonable estimate of Q^π.
  - No: $\hat{Q}$ is not a reasonable estimate of Q^π, need **correction**.

Post-hoc Correction

- We want minimum manipulation of $\hat{Q}$ to keep the desired properties that $\hat{Q}$ may have.
- With a given estimator $\hat{Q}$ that need to correct, let the final estimator be

$$Q = \hat{Q} + Q_{\text{debias}},$$

and solve the following problem to get the correction term Q_{debias}:

$$\min_{Q_{\text{debias}} \in \mathcal{F}} \|Q_{\text{debias}}\|_{\mathcal{F}}, \quad \text{s.t.} \quad \hat{\mathbb{D}}_K^V(\hat{Q} + Q_{\text{debias}}, Q^{\pi}) \leq \lambda_{\delta,n}\,.$$

- If $\mathcal{F}$ is a RKHS with dimension $d \leq n$, can be solved with convex optimization methods effectively.

Experiments

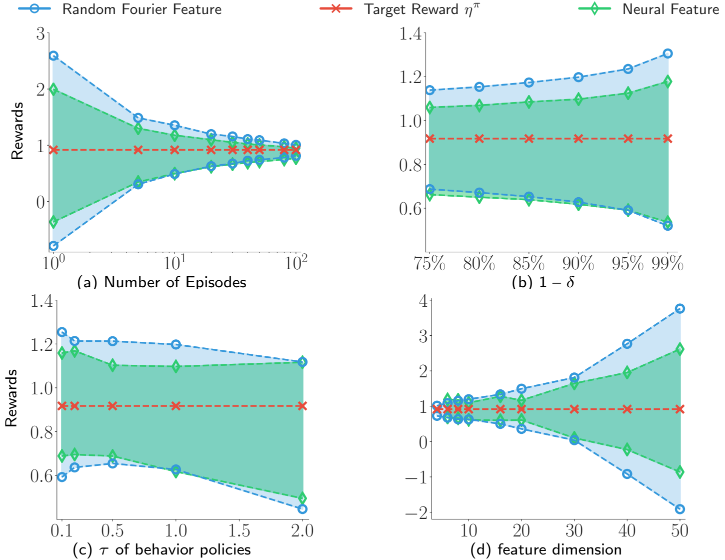
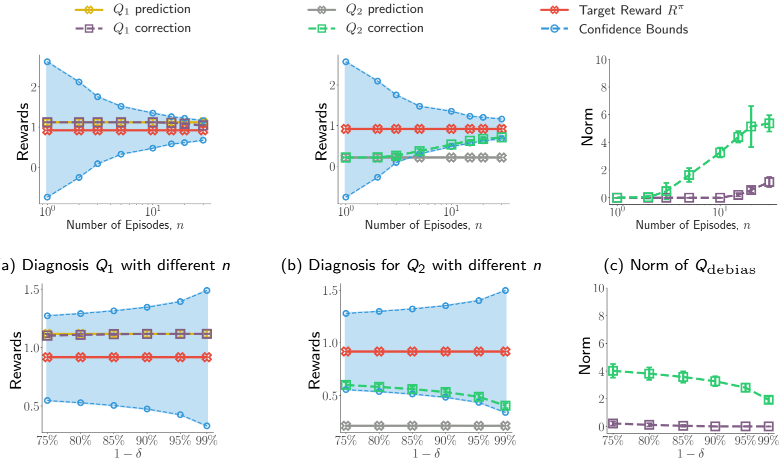


Figure: Off-policy evaluation results on Inverted-Pendulum.

Experiments



(a) Diagnosis Q_1 with different n

(b) Diagnosis for Q_2 with different n

(c) Norm of Q_{debias}

(d) Diagnosis for Q_1 with different δ

(e) Diagnosis for Q_2 with different δ

(f) Norm of Q_{debias}

Figure: Post-hoc diagnosis on Inverted-Pendulum.

Thanks!

Reference I

[FLL19] Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. In Advances in Neural Information Processing Systems, pages 15430–15441, 2019.

[HSN17] Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[JL15] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. arXiv preprint arXiv:1511.03722, 2015.

[LGR+18] Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In Advances in Neural Information Processing Systems, pages 2644–2653, 2018.

[LLTZ18] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In Advances in Neural Information Processing Systems, pages 5356–5366, 2018.

[TTG15] Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.