Non-asymptotic Confidence Intervals of Off-policy Evaluation: Primal and Dual Bounds

Yihao Feng*, Ziyang Tang*, Na Zhang‡, Qiang Liu*

*UT Austin, ‡Tsinghua University

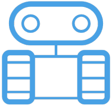
ICLR 2021

Overview: Confidence Intervals for OPE

- Prior work on **confidence intervals for off-policy evaluation**:
  - Trajectory-based concentration [TTG15, HSN17]: **curse of horizon, long interval**.
  - Bootstrap based methods [DNC+20, KN20]: Not safe (**asymptotic guarantee**); require **i.i.d. assumption** on transition.
- This work:
  - **Non-asymptotic guarantee**: safe!
  - **Weaker data assumptions**! General off-policy data collection procedures satisfy the assumption!
  - **Tight**: Shorter interval compared with trajectory-based methods.

Real World RL Applications

Medical



Robotics

Recommendation

- Knowing the performance of the policies is critical!
- **Off-Policy Evaluation**: leverage historical data to estimate target policy performance!

Policy Evaluation

- Goal: Estimate the performance of policy π: $J_\pi = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r_t]$.
- **Off-policy Evaluation**: ☺ Leverage historical data to do the estimation.
- Historical data is limited, we need to quantify the uncertainty of the estimation!
- **Confidence Interval for Off-Policy Evaluation**: Help us to make high-stake decisions!

Confidence Interval for OPE: Problem Setup

- Given data $\mathcal{D}_n = \{s_i, a_i, r_i, s_i'\}_{i=1}^n$ collected from previous data in some arbitrary way (off-policy, behavior-agnostic).

- Goal: construct a $1 - \delta$ confidence interval $[J^-,\ J^+]$ for J_π:

$$\mathbb{P}\left(J_\pi \in \left[J^-,\ J^+\right]\right) \geq 1 - \delta.$$

General Idea

- Denote $x = (s, a)$, we have the value function based estimator:

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

- 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

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

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

Feasible Sets $\mathcal{Q}_n$

- With a tight concentration inequalities for kernel loss [FLL19], we can construct the feasible sets $\mathcal{Q}_n$:

$$\mathcal{Q}_n := \left\{ q \in \mathcal{Q} \ : \quad L_{\mathcal{K}}(q; \mathcal{D}_n) \leq \sqrt{\frac{C \cdot \log(2/\delta)}{n}} \right\},$$

  where C is a computable constant, and $L_{\mathcal{K}}(q; \mathcal{D}_n)$ is the empirical estimation of kernel loss

$$L_{\mathcal{K}}(q; \mathcal{D}_n) := \sqrt{\frac{1}{n^2} \sum_{ij=1}^{n} \left(q(x_i) - \gamma q(x_i') - r_i \right) k(x_i, x_j) \left(q(x_j) - \gamma q(x_j') - r_j \right)}.$$

Primal and Dual Bounds

- We can obtain the upper bounds J^+ via

$$J^+_{\mathcal{Q}} = \sup_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\mu_0,\pi}[q], \quad \text{s.t.} \quad L_{\mathcal{K}}(q; \mathcal{D}_n) \le \varepsilon_n \right\},$$

where $\varepsilon_n = \sqrt{\frac{C \cdot \log(2/\delta)}{n}}$.

- The lower bounds can be obtained by changing the maximizing to minimization.

- **Primal bounds**: solve the optimization exactly to obtain a valid confidence interval.

Primal and Dual Bounds

- Solving primal bound needs **exact global optimum**, which is not flexible for neural approximation.

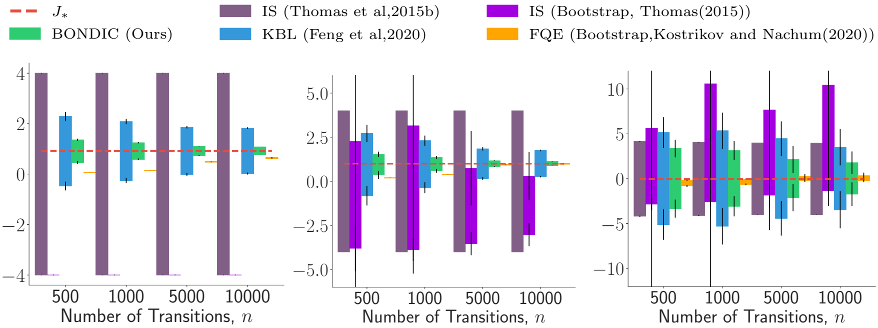
- Instead, we derive its dual bounds:

$$J^{++} := \inf_{w \in \mathcal{W}} \left\{ \mathbb{E}_{\mathcal{D}_n}[w(x)r(x)] + I(q, \mathcal{D}_n) + \varepsilon_n \cdot \|w\|_{\mathcal{W}} \right\} ,$$

  where $I(q, \mathcal{D}_n) := \sup_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{\mathcal{D}_n}[w(x)(\gamma q(x') - q(x))] + \mathbb{E}_{\mu_0,\pi}[q] \right\}$, which can be solved with closed form by assuming $\mathcal{Q}$ is an RKHS.

- **Dual Bound**: it is always a valid confidence interval even when the optimization is not solved exactly!

Experimental Results



(a) Pendulum (b) CartPole (c) Type 1 diabetes

Thanks!

A more detailed version: https://arxiv.org/pdf/2103.05741.pdf.

Reference I

[DNC+20] Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. arXiv preprint arXiv:2010.11652, 2020.

[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.

[KN20] Ilya Kostrikov and Ofir Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. arXiv preprint arXiv:2007.13609, 2020.

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