Off-Policy Interval Estimation with Lipschitz Value Iteration

Ziyang Tang*, Yihao Feng*, Na Zhang‡, Jian Peng†, Qiang Liu*

*UT Austin, †UIUC, ‡Tsinghua

NeurIPS 2020

Scenario

- Consider the following long term medical treatment scenario as a Markov decision process(MDP):

state s: patients physiological features.

action a: medical action, e.g. take a medicine or not; how many doses.

reward r: patient condition; side effect.

next state s': patients physiological features at the next time step.

Problem Settings: Policy Evaluation

- Goal: evaluate a new treatment policy π.
- Formally, evaluate the long term average reward R_π:

$$R_\pi = \mathbb{E}_{\tau \sim \pi}[\sum_{t=0}^{H} \gamma^t r_t],$$

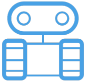
where $\tau = (s_0, a_0, r_0, \ldots, s_i, a_i, r_i, \ldots)$ is the trajectory from policy π.

- On-policy vs off-policy:
  1. **On-policy**: ☹ collect data directly by deploying the treatment policy on patients: **high risk, unethical**!!
  2. **Off-policy**: ☺ leverage historical data to do the estimation.

Hardness of Off-Policy Evaluation(OPE)

- Point estimation can be arbitrarily bad:
  1. High variance for trajectory-based methods: variance grows exponentially with the length of horizon[LLTZ18].
  2. Bias for optimization-based methods[JL15, TFL+20].
  3. Small effective sample size due to policy mismatch(distributional shift).
- In high-stakes scenarios, point estimation is not enough; need **confidence interval** as well!!

Medical



Robotics

Recommendation

Finite Samples Bellman Equations

- Formulate R^{π} with q-function:

$$R_{\pi} = \mathbb{E}_{s \sim \mu_0, a \sim \pi(\cdot|s)}[q_{\pi}(s,a)] := R_{\mu_0,\pi}[q_{\pi}].$$

- Notice that q_{π} is the **unique fixed point** of the Bellman equation:

$$q_{\pi}(x) = r(x) + \gamma \mathbb{E}_{s' = T(x), a' \sim \pi(\cdot|s')}[q_{\pi}(x')] := \mathcal{B}^{\pi} q_{\pi}(x), \quad \forall x$$

  where x is short for state, action pair s, a.

- Only get access to **finite number** of the transition operator $\mathcal{B}^{\pi}$.

- Multiple or infinite q can satisfy the finite samples Bellman equation at $x_i, \ \forall i \in [n]$.

Interval Estimation Frameworks

- Constraint on finite sample Bellman equation may lead to arbitrary large/small value on unseen region, need a model assumption $q_\pi \in \mathcal{F}$.

- Optimization framework:

$$\overline{R}_{\mathcal{F},\pi} = \sup_{q \in \mathcal{F}} \{R_{\mu_0,\pi}[q], \text{ s.t. } q(x_i) = \mathcal{B}^\pi q(x_i), \ \forall i \in [n]\}.$$

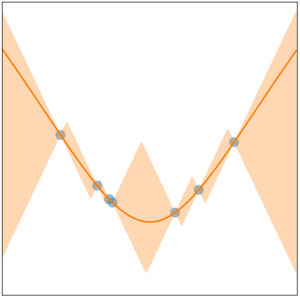
- Simplest assumption: *smoothness* assumption;
  Formally we consider the following bounded Lipschitz class:

$$\mathcal{F}_\eta = \{f : \|f\|_{Lip} \le \eta\},$$

  where $\|f\|_{Lip} := \sup_{x \neq x'} \frac{|f(x) - f(x')|}{d(x,x')}$.

A Lipschitz Regression Example

- Why is it possible to solve an infinite dimension optimization under finite sample constraints?



Consider a regression problem:

$$\overline{f}(x) = \sup_{f \in \mathcal{F}_\eta} \{ f(x), \text{ s.t. } f(x_i) = f_i, \ \forall i \in [n] \}$$

Closed form solution:

$$\overline{f}(x) = \min_{i \in [n]} \{ f_i + \eta d(x, x_i) \}$$

- Similar for the lower bound:

$$\underline{f}(x) = \max_{i \in [n]} \{ f_i - \eta d(x, x_i) \}$$

A Value Iteration Style Algorithm

- A iterative way of solving q-function

$$\overline{R}_{\mathcal{F},\pi} = \sup_{q \in \mathcal{F}} \{ R_{\mu_0,\pi}[q], \text{ s.t. } q(x_i) = \mathcal{B}^{\pi} q(x_i), \ \forall i \in [n] \}.$$

1. Plug in the last q_t as our new regression constraints $\overline{q}_{i,t+1} = \mathcal{B}^{\pi} \overline{q}_t(x_i)$.
2. Solve the new q_{t+1} as a Lipschitz regression problem:

$$\overline{q}_{t+1}(x) = \sup_{q \in \mathcal{F}_{\eta}} \{ q(x), \text{ s.t. } \ q(x_i) = \overline{q}_{i,t+1} \} = \min_{i \in [n]} \{ \overline{q}_{i,t+1} + \eta d(x, x_i) \}.$$

Theoretical Properties of Lipschitz Value Iteration (Informal)

- **Monotonicity**, with a well-defined $\overline{q}_0$, we have:

$$\overline{q}_t(x) \geq \overline{q}_{t+1}(x) \geq q_\pi(x), \ \forall x$$

- **Linear Convergence:**

$$\overline{q}_t(x) - \overline{q}_\infty(x) = \mathcal{O}(\gamma^t)$$

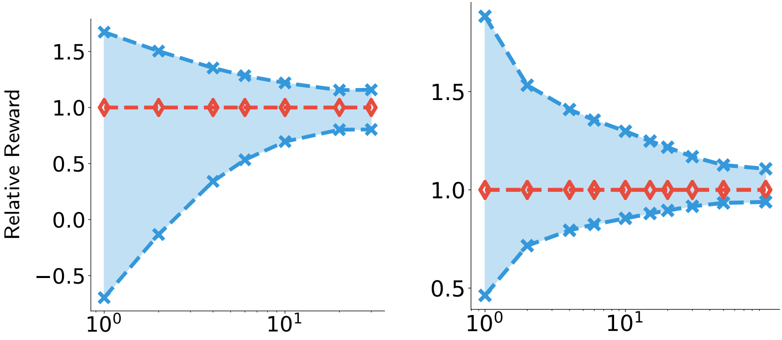
- **Tightness of bounds:**

$$\overline{q}_t(x) - \underline{q}_t(x) = \mathcal{O}(\varepsilon_{X_n})$$

where ε_{X_n} is the covering radius of data set $X_n = \{x_i\}_{i \in [n]}$, with:

$$\varepsilon_{X_n} = \sup_x \min_{i \in [n]} \{d(x, x_i)\}$$

Experimental Results



number of trajectories n_t
(a) Pendulum Environment

number of trajectories n_t
(b) HIV Environment

Thanks!

Reference I

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

[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.

[TFL+20] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In International Conference on Learning Representations, 2020.