An Overview of Off-Policy Policy Evaluation for Stationary and Non-Stationary Environments

Vincent Liu
Feb 2023

UNIVERSITY OF ALBERTA

Outline

- Problem setting
- A short overview of OPE
- OPE with non-stationary reward functions

Off-policy policy evaluation (OPE)

- We consider
  - Finite horizon MDP, one initial state s_0
  - Offline setting where we are given a dataset $\{(S_0^{(i)}, A_0^{(i)}, R_0^{(i)}, S_1^{(i)}, \ldots, R_{H-1}^{(i)})\}_{i=1}^{n}$ collected by π_b
- Given a target policy, the goal is to estimate the value of the policy, denoted by $J(\pi) = \mathrm{E}^{\pi}[R_0 + R_1 + \ldots + R_{H-1}]$

Why OPE?

- Applications: recommendation/search advertising

$\mathcal{A}$

		Item 1	Item 2	Item 3	
	User A	1.0		0.0	
$\mathcal{S}$	User B		1.0		
	User C	0.0	1.0		

- Given historical data, we might want to evaluate the performance of different recommendation policies (i.e., offline A/B testing)

Importance Sampling (IS)

- Recall $J(\pi) = \mathrm{E}^{\pi}[R_0 + R_1 + \ldots + R_{H-1}]$
- **IS corrects the sample return by the product of IS ratios**

$$\hat{J}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\prod_{h=0}^{H-1} \frac{\pi(A_h^{(i)} \mid S_h^{(i)})}{\pi_b(A_h^{(i)} \mid S_h^{(i)})}}_{W_i} \sum_{h=0}^{H-1} R_h^{(i)}$$

- Condition: $\pi_b(a \mid s) > 0$ if $\pi(a \mid s) > 0$
- Properties: Unbiased but high variance
- High probability bound (with probability at least $1 - \delta$, the following holds): $|\hat{J}(\pi) - J(\pi)| \leq O(\frac{A^H H \ln(1/\delta)}{n})$

Exponential in the horizon

Fitted Q Evaluation (FQE)

- **FQE aims to estimate the action value function directly**
- Choose a function class $\mathscr{F}$, a set of function $f : \mathcal{S} \times \mathscr{A} \to \mathbb{R}$
  - Initialize $q_H = 0$
  - For $h = H-1,\ldots,0$, $q_h = \arg\min_{f \in \mathscr{F}} \hat{l}_h(f, q_{h+1})$ where
    $$\hat{l}_h(f, q_{h+1}) = \frac{1}{|D_h|} \sum_{(s,a,r,s') \in D_h} (f(s,a) - r - q_{h+1}(s', \pi(s')))^2$$
  - The FQE estimate is $\hat{J}(\pi) = \mathrm{E}_{a \sim \pi(\cdot|s_0)}[q_0(s_0, a)]$
- Properties: Biased but low variance

FQE

Notation.
$d_h^\pi(s,a) = \mathbb{P}^\pi(S_h = s, A_h = a)$
$\mu_h(s,a)$ is the data distribution, i.e., $\mu_h(s,a) = \mathbb{P}^{\pi_b}(S_h = s, A_h = a)$

- Conditions:
  - Data coverage for π: $\max_{h \in [H]} \max_{s \in \mathcal{S}_h, a \in \mathcal{A}_h} \dfrac{d_h^\pi(s,a)}{\mu_h(s,a)} \leq C$
  - $\mathcal{F}$ is closed under $\mathcal{T}^\pi$: $\forall q \in \mathcal{F}, \mathcal{T}^\pi q \in \mathcal{F}$
- High probability bound (Duan, et. al., 2021):

$$|\hat{J}(\pi) - J(\pi)| \leq O(H\sqrt{C(H\mathcal{R}_n(\mathcal{F}) + H^2\frac{\log(H/\delta)}{n}})$$

$$\mathcal{R}_n(\mathcal{F}) \leq H\log(|\mathcal{F}|)/n$$

Marginalized IS/Stationary IS/DualDICE/...

- **These methods aim to estimate the visitation distribution ratio** $\hat{w}_{\pi/\mu}(s,a) \approx w_{\pi/\mu}(s,a) \doteq \dfrac{d_h^\pi(s,a)}{\mu_h(s,a)}$
  - The estimate is $\hat{J}(\pi) = \dfrac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{H-1} \hat{w}_{\pi/\mu}(S_h^{(i)}, A_h^{(i)})R_h^{(i)}$
- Many ways to estimate the ratio (*MSWL*: Liu et. al., 2018, *DualDICE*: Nachum et. al., 2019, *MWL*: Uehara et. al., 2020)
  - DualDice solves the min-max optimization (zeta)

$$\min_{\nu:S\times A\to\mathbb{R}} \max_{\zeta:S\times A\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}, a'\sim\pi(s')}\left[(\nu(s,a) - \gamma\nu(s',a'))\zeta(s,a) - \zeta(s,a)^2/2\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right]. \quad (11)$$

MIS/SIS/DualDICE/…

- Condition:
  - Data coverage for π
  - Function approximation: $w_{\pi/\mu} \in \mathcal{F}$ and some restriction on the auxiliary function class
- High probability bound: Similar to FQE, complexity of the function classes + concentration terms $O(\sqrt{1/n})$

Per-decision IS (PDIS)

- Recall trajectory-wise IS is

$$\hat{J}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\prod_{h=0}^{H-1} \frac{\pi(A_h^{(i)} \mid S_h^{(i)})}{\pi_b(A_h^{(i)} \mid S_h^{(i)})}}_{W_i} \sum_{h=0}^{H-1} R_h^{(i)}$$

- PDIS applies the product of IS ratios for each horizon

$$\hat{J}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \sum_{h=0}^{H-1} \underbrace{\prod_{t=0}^{h} \frac{\pi(A_t^{(i)} \mid S_t^{(i)})}{\pi_b(A_t^{(i)} \mid S_t^{(i)})}}_{\rho_{0:h}^{(i)}} R_h^{(i)}$$

- Properties: Unbiased but slightly less variance

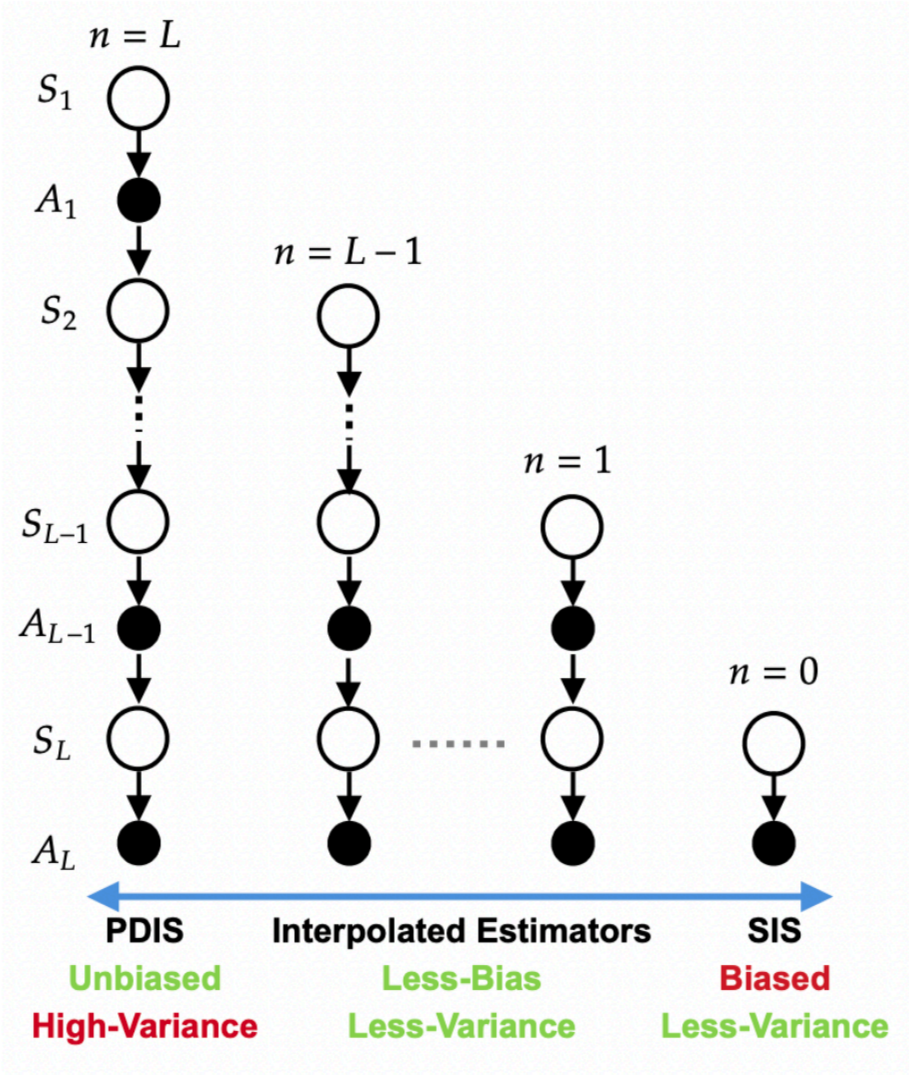
- High probability bound: still exponential in the horizon

Doubly Robust (DR)

- **DR combines PDIS and FQE**
  - Run FQE to get q
  - $\hat{J}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{H-1}\rho_{0:h}^{(i)}R_h^{(i)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{H-1}(\rho_{0:h}^{(i)}q_h(S_h^{(i)}, A_h^{(i)}) - \rho_{0:h-1}^{(i)}v_h(S_h^{(i)}))$
    where $v(s) = q(s, \pi(s))$
- Condition: $\pi_b(a\,|\,s) > 0$ if $\pi(a\,|\,s) > 0$
- Properties:
  - If q is accurate, then $q_h(S_h, A_h) \approx R_h + v_{h+1}(S_{h+1})$ and $\hat{J}(\pi) \approx v_0(s_0)$
  - Taking expectation, the terms with q and v cancel out

Other ways to combine multiple estimators

- MAGIC: use PDIS ratios only for horizon <= L-1, and use FQE for horizon L
- SOPE: use PDIS ratios only for horizon <= L-1, use MIS/SIS for horizon L



Summary of OPE in stationary environments

- OPE generally requires some notion of *data coverage* for the target policy
- A key question is the trade-off between <u>the bias from using a function approximation</u> and <u>the exponential variance from using product of IS ratios</u>
- **Hyperparameter/model selection for OPE is an open problem in the offline setting**
  - Some methods (FQE, DICE) requires choosing a function approximation
  - Some methods (MAGIC, SOPE) requires choosing hyperparameters

Non-stationary environments

- Consider $H = 1$
- Consider a piecewise stationary environments with known change points where the reward function changes
  - Period 1 with reward function r_1: we collect data D_1
  - ...
  - Period k with reward function r_k: we collect data D_k
  - From $D_1, \ldots, D_k$, we want to estimate
    $J_k(\pi) = \sum_{s,a} P(s)\pi(a \mid s)r_k(s, a)$

How should we reuse past data?

- IS using D_k only is unbiased (under some conditions) but has high variance -> we need more data

$$\hat{J}_{IS,k}(\pi) = \frac{1}{n_k} \sum_{(s,a,r)\in D_k} \frac{\pi(a \mid s)}{\pi_b(a \mid s)} r$$

- Naively using past data introduces large bias

$$\hat{J}_k(\pi) = \frac{1}{\sum_t n_t} \sum_{(s,a,r)\in D_1,\ldots,D_k} \frac{\pi(a \mid s)}{\pi_b(a \mid s)} r$$

- Jagerman et al. (2019) propose to decrease the weighting for the past data

$$\hat{J}_{SWIS,k}(\pi) = \frac{1}{\sum_{t=k-B}^{k} n_t} \sum_{(s,a,r)\in D_{k-B},\ldots,D_k} \frac{\pi(a \mid s)}{\pi_b(a \mid s)} r$$

  - B controls the bias-variance tradeoff, however, **the bias can still be large**
  - **Choosing this hyperparameter is hard in the offline setting!!!**

Regression-assisted DR

- Learn reward predictions $\hat{r}_{k-B}(s,a), \ldots, \hat{r}_{k-1}(s,a)$ from past data $D_{k-B}, \ldots, D_{k-1}$
- Construct feature vector $\phi_k(s,a) = (1, \hat{r}_{k-B}(s,a), \ldots, \hat{r}_{k-1}(s,a))$
- Fit a regression on top of $\phi_k(s,a)$ to predict $r_k(s,a)$ using D_k:

$$\hat{\beta}_k = \left(\sum_{(s,a)\in D_k} \frac{\pi(a\,|\,s)}{\pi_b(a\,|\,s)} \phi(s,a)\phi(s,a)^\top \right)^{-1} \left(\sum_{(s,a)\in D_k} \frac{\pi(a\,|\,s)}{\pi_b(a\,|\,s)} \phi(s,a) r_k(s,a) \right)$$

- Apply the DR estimator

$$\hat{J}_{Reg,k}(\pi) = \frac{1}{n} \sum_{s\in D_k} \sum_{a\in\mathscr{A}} \pi(a\,|\,s) \phi_k(s,a)^\top \hat{\beta}_k + \frac{1}{n} \sum_{(s,a)\in D_k} \frac{\pi(a\,|\,s)}{\pi_b(a\,|\,s)} \left(r_k(s,a) - \phi_k(s,a)^\top \hat{\beta}_k \right)$$

- Reference: Asymptotically Unbiased Off-Policy Policy Evaluation when Reusing Old Data in Nonstationary Environments, AISTATS 2023. With Yash, Philip and Martha

Properties

- Theoretical:
  - Asymptotically unbiased
  - Large-sample confidence interval:
    $P(t_y \in [\hat{t}_{Reg} - z_{\alpha/2}\hat{\mathrm{V}}(t_{Reg}), \hat{t}_{Reg} + z_{\alpha/2}\hat{\mathrm{V}}(t_{Reg})]) \to 1 - \alpha$
- Empirical:
  - The estimator is robust to the hyperparameter B

Summary of OPE in non-stationary environments

- For non-stationary OPE, a key question is the trade-off between <u>the bias from reusing old data</u> and <u>the variance from using less data</u>
  - We propose an estimator that (1) has a better trade-off, and (2) is not sensitive to its hyperparameter

- Future directions:
  - We assume we know the change points -> <u>how to learn the change points?</u>
  - We focus on short horizon tasks -> <u>how to extend to long horizon tasks (without exponential variance)?</u>