CS292F StatRL Lecture 13
OPE in Reinforcement Learning

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$

Task: design OPE methods

Evaluate fixed Target Policy $\boldsymbol{\pi}$

Via Uniform OPE

- Task 2: Offline Policy Learning. (OPL)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$

Task: design OPO methods

Find near optimal Policy $\widehat{\boldsymbol{\pi}}^*$

Recap: Lecture 12

- OPE algorithms in (Contextual) Bandits
  - DM, IS, WIS, DR, SWITCH

$\hat{r}(s,a) \to \mathbb{R}$ $\qquad \frac{\pi(a|s)}{\mu(a|s)}$

- Comparing DM and IS in Multi-armed Bandits: $\quad \frac{\pi}{\hat{\mu}}$
  - DM is asymptotically more efficient
  - IS is better.

- More generally:
  - DM is asymptotically more efficient if we assume realizability
  - IS cannot be improved when we don't

Recap: Two standard approaches

- Direct method / regression estimator

$$\hat{v}^{\pi}_{\mathrm{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\hat{r}(x_i,a)\pi(a|x_i)$$

- Importance sampling / Inverse Propensity Score /

$$\hat{v}^{\pi}_{\mathrm{IPS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}r_i$$

Recap: Combining DM and IS

- Doubly Robust Estimation
  - Remains unbiased, but limited benefits to the variance
- SWITCH
  - Introduce bias, but drastically reduce variance

This lecture

- Generalizing the bandits OPE idea to RL

Finite Horizon MDPs

finite state

finite Action

- Curse of Horizon

- Marginalized Importance Sampling

$MDP: (S, A, P, d_0, r)$

n trajectories i.i.d

i-th: $(s_1^{(i)}, a_1^{(i)}, r_1^{(i)}), (s_2^{(i)}, a_2^{(i)}, r_2^{(i)}) \ldots, (s_H^{(i)}, a_H^{(i)}, r_H^{(i)})$

$s_1^{(i)} \sim d_0(s)$

$a_h^{(i)} \sim \mu(\cdot \mid s_h^{(i)})$

$r_h^{(i)} \sim P(r_h \mid s_h^{(i)}, a_h^{(i)})$

OPE in Reinforcement Learning

- Importance sampling on the entire trajectory

$$\boxed{V^{\pi}_{\text{naive IS}} = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{h=1}^{H} r_h^{(i)} \right) \cdot \frac{d_1 \pi_1 \cancel{P_1} \cdot \pi_2 \cancel{P_2} \cdots \pi_{H-1} \cancel{P_{H-1}} \pi_H}{d_1 \mu_1 \cancel{P_1} \cdot \mu_2 \cancel{P_2} \cdots \mu_{H-1} \cancel{P_{H-1}} \mu_H}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \prod_{h=1}^{H} \rho_h^{(i)} \sum r_h^{(i)}$$

where $\rho_h^{(i)} = \dfrac{\pi(a_h^{(i)} \mid s_h^{(i)})}{\mu(a_h^{(i)} \mid s_h^{(i)})}$

- (Per-Step) Importance Sampling

$$V^{\pi}_{IS} = \sum_{h=1}^{H} \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{t=1}^{h} \rho_t^{(i)} \right) \cdot r_h^{(i)}$$

- Exercise:
  - Infinite horizon discounted version?
  - Weighted Importance Sampling Extension?

Doubly Robust OPE in Reinforcement Learning

- An alternative form for the Per-Step IS

$\gamma = 1$

$H = 2$

$$V^0_{\text{step-IS}} := 0, \text{ and for } t = 1, \ldots, H,$$

$$V^{H+1-t}_{\text{step-IS}} := \rho_t \left(r_t + \gamma V^{H-t}_{\text{step-IS}} \right).$$

$V_{IS}^{1} = \cancel{\rho_2 r_2 + \rho_1 r_1}$

$\rho_1 r_1 + \rho_1(\rho_2 r_2)$

- Given a value function approximator

$\hat{V}(s_t) = \sum \pi(a_t|s_t)\hat{Q}(s_t, a_t)$

$$V^0_{\text{DR}} := 0, \text{ and for } t = 1, \ldots, H,$$

$$V^{H+1-t}_{\text{DR}} := \widehat{V}(s_t) + \rho_t \Big(r_t + \gamma V^{H-t}_{\text{DR}} - \widehat{Q}(s_t, a_t) \Big).$$

Jiang, N., & Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In ICML 2016.

Mean and Variance of Doubly Robust OPE in RL

- Doubly Robust OPE in RL is unbiased

- Variance

Theorem 1. *V_{DR} is an unbiased estimator of $v^{\pi_1,H}$, whose variance is given recursively as follows:* $\forall t = 1, \ldots, H$,

$$\mathbb{V}_t\left[V_{DR}^{H+1-t}\right] = \mathbb{V}_t\left[V(s_t)\right] + \mathbb{E}_t\left[\mathbb{V}_t\left[\rho_t \Delta(s_t, a_t) \mid s_t\right]\right]$$
$$+ \mathbb{E}_t\left[\rho_t^2 \, \mathbb{V}_{t+1}\left[r_t\right]\right] + \mathbb{E}_t\left[\gamma^2 \rho_t^2 \, \mathbb{V}_{t+1}\left[V_{DR}^{H-t}\right]\right], \quad (11)$$

where $\Delta(s_t, a_t) := \widehat{Q}(s_t, a_t) - Q(s_t, a_t)$ *and* $\mathbb{V}_{H+1}\left[V_{DR}^0 \mid s_H, a_H\right] = 0.$

$$\mathbb{V}(\hat{v}_{DR}) = \sum_{h=1}^{H} \mathbb{E}_h \sum_{t=1} \left(\prod \rho_t\right)^2 \mathrm{Var}_{(s_t,a)}\left[(V(s') + r \mid s,a\right]$$

Main challenge of OPE in RL: The curse of Horizon

i.i.d trajector

$$\widehat{v}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\boxed{\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right]} r_h^{(i)}.$$

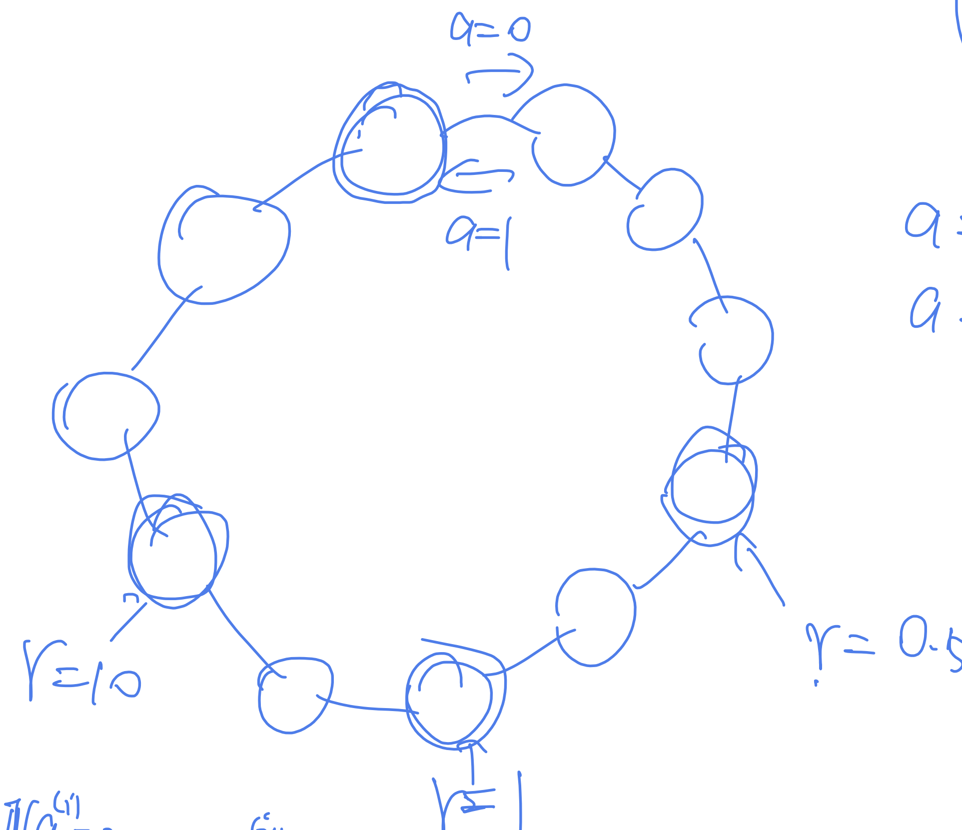
The curse of horizon. (Liu et al, 2018 NeurIPS)

- The variance is exponential in H!

$\mathrm{Var}(\hat{V}^{\pi}_{IS}) = \frac{1}{n}\mathrm{Var}(\quad)$

Example on Curse of Horizon

(Liu et al. 2018)



$a = 0$

$a = 1$

$\mu_0(s) = \begin{cases} 0 & \text{w.p. } p \\ 1 & \text{w.p. } 1-p \end{cases}$

$\pi(s) = 0$

$H = 200$

$$\hat{V}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left(\frac{1}{p}\right)^{h} \cdot \mathbb{1}\left(a_t^{(i)} = 0 \text{ for } \forall t \le h\right) \cdot r_h^{(i)}$$

$$\prod_{t=1}^{h} \frac{\pi_h(a|s)}{\mu_h(a|s)}$$

$$Var\left(\hat{V}^{\pi}_{IS}\right) = O\left(\left(\frac{1}{p}\right)^{H}\right)$$

From Importance Sampling to Marginalized Importance Sampling

$$\widehat{v}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\boxed{\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right]} r_h^{(i)}.$$

$\hat{d}_t^{\pi}(s) \to \underline{d_t^{\pi}(s)}$

$\hat{d}_t^{\mu}(s) \to \underline{d_t^{\mu}(s)}$

$\hat{r}_t^{\pi}(s) \to r_t^{\pi} \,|\, \boxed{s \sim d_t^{\mu}(s)}$

$$\widehat{v}^{\pi}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\boxed{\frac{\widehat{d}_t^{\pi}(s_t^{(i)})}{\widehat{d}_t^{\mu}(s_t^{(i)})}}\,\widehat{r}_t^{\pi}(s_t^{(i)}).$$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

What are some ideas for estimating the marginalized importance weight?

$$\widehat{v}^{\pi}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\boxed{\frac{\widehat{d}^{\pi}_t(s_t^{(i)})}{\widehat{d}^{\mu}_t(s_t^{(i)})}}\widehat{r}^{\pi}_t(s_t^{(i)}).$$

$$\widehat{r}^{\pi}_t(s) = \frac{1}{n_{s,t}}\sum_{i=1}^{n}\mathbb{1}(s_t^{(i)}=s)\cdot r_t^{(i)}\cdot\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}$$

$|S| \cdot$ of $r_t^{(i)}$

- Idea 1: averaging over multiple visits to the same state.

$$\cancel{\widehat{d}^{\mu}_t(s) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(s_t^{(i)}=s)}$$

$$\frac{d_t^{\pi}(s)}{d_t^{\mu}(s)} = \frac{1}{n_{t,s}}\sum_{i=1}^{n}\underbrace{\prod_{h=1}^{t}\rho_h^{(i)}(s_h^{(i)},a_h^{(i)})}_{\frac{\pi_h(a|s)}{\mu_h(a|s)}}\cdot\mathbb{1}(s_t^{(i)}=s)$$

Idea 2: Recursive estimation

$$d_t^\pi(s_t) = \sum_{s_{t-1}} P_t^\pi(s_t|s_{t-1}) d_{t-1}^\pi(s_{t-1}),$$

$P^\pi(s'|s)$ $d^\pi(s)$

Add d_t plugin

$$\widehat{d_t^\pi} = \widehat{P_t^\pi}\,\widehat{d_{t-1}^\pi},$$

$d_t(s) = \frac{1}{n} \sum \mathbb{1}(s_t^{(i)} = s)$

$$\text{where } \widehat{P}_t^\pi(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^{n} \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}) = (s_{t-1}, s_t));$$

use IS

$$\widehat{r}_t^\pi(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^{n} \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t),$$

use IS

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

Results: OPE error bound of MIS

- The MSE of MIS estimator obeys:

$\le \tau_s^2$ $\qquad\qquad$ τ_a^2

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_{\mu}\left[\frac{d_t^{\pi}(s_t)^2}{d_t^{\mu}(s_t)^2}\mathrm{Var}_{\mu}\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}(V_{t+1}^{\pi}(s_{t+1})+r_t)\middle|s_t\right]\right]+\tilde{O}(n^{-1.5})$$

$$\sum_{s_t\in S} d_t^{\mu}(s_t)\cdot\frac{d^{\pi}(s_t)^2}{d^{\mu}(s_t)^2}$$

$\le O(H)$

$$\max_{s,t}\left|\frac{d_t^{\pi}(s)}{d_t^{\mu}(s)}\right|\le \tau_s$$

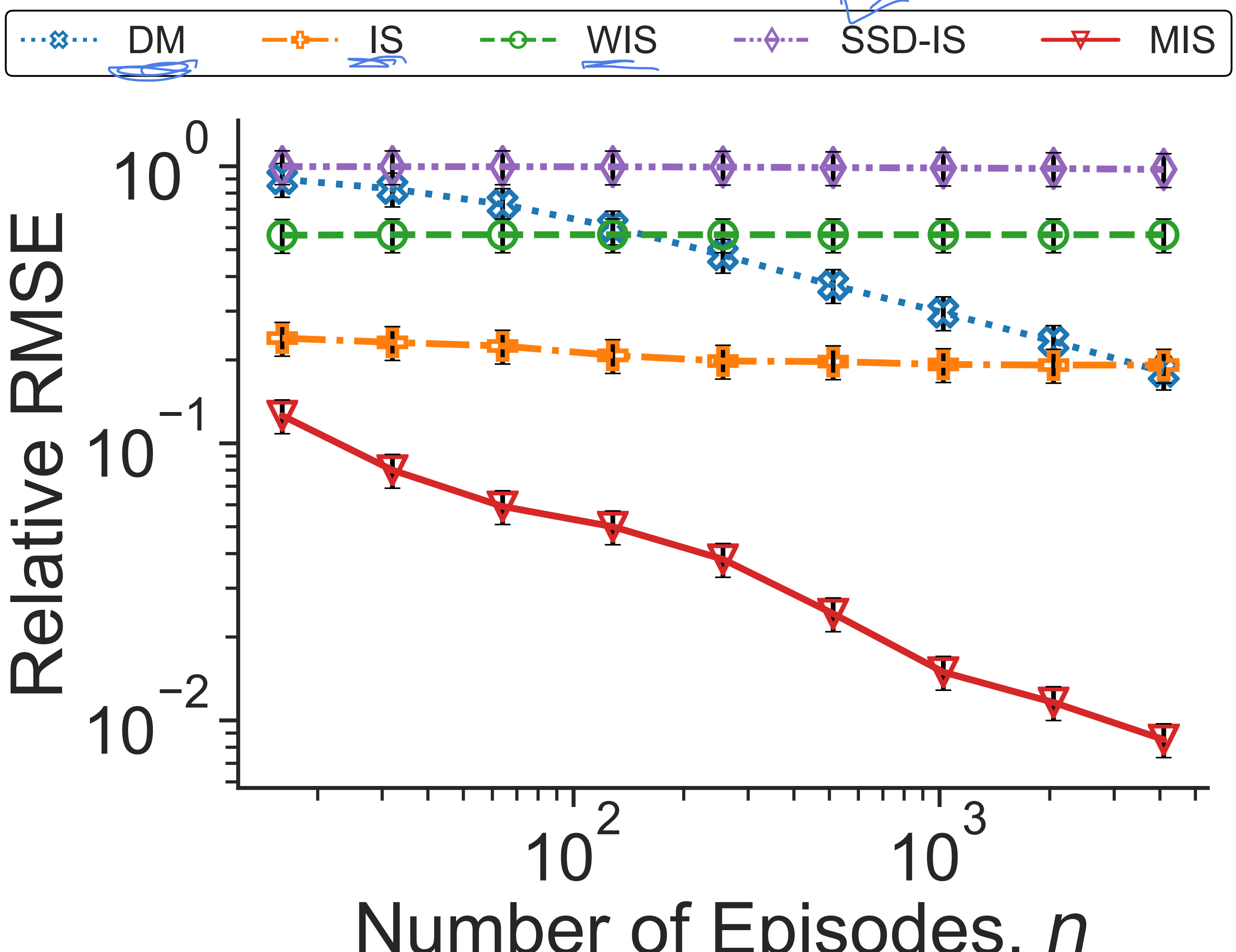
$$\max_{s,a,t}\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}\le \tau_a$$

$$\frac{1}{n}H\,\tau_s^2\,\tau_a^2\,H^2=\frac{H^3}{n}\tau_s^2\tau_a^2$$

$0\le \mathbb{E}r\le 1,\ \mathrm{Var}\, r\le 1$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

Experiment on mountain car



Challenges of the analysis

1. **Dependent data:** The data within each trajectory are not independent

2. **An annoying bias:** there is a non-zero probability that some states are not visited at all. And it affects all future estimates

3. **Error propagation** from recursive estimation

$$\hat{d}_t^{\pi_1} = \underbrace{\hat{P}_t^{\pi_1}}_{\text{error}} \underbrace{\hat{d}_{t-1}^{\pi_1}}_{\text{error}}$$

Addressing Challenge 1: Define an appropriate martingale

$h=1$ $\quad$ $h=2$ $\quad \cdots \quad$ $h=H$

traj 1 $\quad (s_1^{(1)}, a_1^{(1)}, r_1^{(1)})$ $\quad (s_2^{(1)}, a_2^{(1)}, r_2^{(1)})$ $\quad \cdots \quad$ $(s_H^{(1)}, a_H^{(1)}, r_H^{(1)})$

$\vdots$

traj n $\quad (s_1^{(n)}, a_1^{(n)}, r_1^{(n)})$ $\quad \cdots \quad$ $(s_H^{(n)}, a_H^{(n)}, r_H^{(n)})$

Conditioning on $\dfrac{n_{s,s',t}}{n_{s,a,t}} \rightarrow \hat{P}(s'|s)$

- Consider the data collection in parallel
- Group all data for time h together
- Conditioning on the number of times states are visited

Addressing Challenge 2: Fictitious estimator technique

$$\widetilde{v}^{\pi} := \sum_{t=1}^{H} \sum_{s_t} \widetilde{d}_t^{\pi}(s_t) \widetilde{r}_t^{\pi}(s_t) \quad \text{where} \quad \widetilde{d}_t^{\pi} = \widetilde{\mathbb{P}}_{t,t-1}^{\pi} \widetilde{d}_{t-1}^{\pi}$$

$$\widetilde{r}_t^{\pi}(s_t) = \begin{cases} \widehat{r}_t^{\pi}(s_t) & \text{if } n_{s_t} \geq n d_t^{\mu}(s_t)(1-\delta) \\ r_t^{\pi}(s_t) & \text{otherwise;} \end{cases}$$

true value

$$\widetilde{\mathbb{P}}_{t,t-1}^{\pi}(\cdot|s_{t-1}) = \begin{cases} \widehat{\mathbb{P}}_{t,t-1}^{\pi} & \text{if } n_{s_{t-1}} \geq n d_t^{\mu}(s_{t-1})(1-\delta) \\ \mathbb{P}_{t,t-1}^{\pi} & \text{otherwise.} \end{cases}$$

$\mathbb{E}[\tilde{v}^{\pi}] = v^{\pi}$

Multiplicative Chernoff Bound

Lemma A.1 (Multiplicative Chernoff bound [Chernoff et al., 1952]).
Let X be a Binomial random variable with parameter p, n.
For any $\delta > 0$, we have that

$$\mathbb{P}[X < (1 - \delta)pn] < e^{-\frac{\delta^2 pn}{2}}$$

- Apply to our problem

$n_{s,t}$ by running $\mathcal{M}$ for n traj's,
$\sim \mathrm{Bin}(d_t^{\mathcal{M}}(s), n)$

$$P(n_{s,t} < n d_t^{\mathcal{M}}(s)(1-\delta)) \le e^{-\frac{\delta^2 n \cdot d_t^{\mathcal{M}}(s)}{2}}$$

Union bound for all $t = 1, \ldots, H$, all $s \in S$

$|H| \cdot |S|$

~~w.h.p.~~

$$P(\text{Fictitious Estimator} \neq \mathcal{M}(s))$$

$$\le |H| \cdot |S| \cdot e^{-\delta^2 n \cdot \min_{s,t} d_t^{\mathcal{M}}(s)}$$

$$\delta = \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$$

$n^{\ominus - 100}$

Address Challenge 3: Empirical / Offline version of Bellman equation of variance

$$\mathrm{Var}[\widetilde{v}^{\pi}] = \sum_{h=0}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\mathbf{1}\left(n_{s_h}\geq\frac{n d_h^{\mu}(s_h)}{(1-\delta)^{-1}}\right)\right]\mathrm{Var}_{\mu}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^{\pi}(s_{h+1}^{(1)})+r_h^{(1)})\middle| s_h^{(1)}=s_h\right].$$

$\leq (1+\epsilon)\cdot\dfrac{d^{\pi}(s_h)^2}{d_t(s)}$

$n_{s_h} = n\cdot d_t^{\mu}(s_h)$

$\mathbb{E}\left[\dfrac{1}{n d_h^{\mu}}\right]$

$\left(\dfrac{d_t^{\pi}}{d_t^{\mu}}\right)^2$ ·· t_h thm

Bounding error propagation

$$\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\mathbf{1}\left(n_{s_h} \geq \frac{n d_h^{\mu}(s_h)}{(1-\delta)^{-1}}\right)\right] \leq \frac{(1-\delta)^{-1}}{n}\left(\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)} + \mathrm{Var}\left[\widetilde{d}_h^{\pi}(s_h)\right]\right),$$

$EX^2 = \mathrm{Var}(X) + (EX)^2$

- Bounding the variance of is somewhat tedious
  - Requires use to bound the covariance

$\widetilde{d}^{\pi}$ is unbiased

$$\mathrm{Var}[\widetilde{d}_h^{\pi}(s_h)] \leq \frac{2(1-\delta)^{-1} h d_h^{\pi}(s_h)}{n}.$$

$\mathrm{Cov}(\widetilde{d}_h^{\pi}(\cdot)) = E[\mathrm{Cov}] + \mathrm{Cov}[E]$

$\in R^S$

Is MIS optimal for OPE in RL?

- It depends on the settings.
- For finite-state / infinite action space, we conjecture that it is.

(Still an open problem now.)

- For fully tabular setting, it is not optimal, at least asymptotically.

Tabular MIS

$$\widehat{v}^{\pi}_{\mathrm{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_{t}(s^{(i)}_{t})}{\widehat{d}^{\mu}_{t}(s^{(i)}_{t})}\widehat{r}^{\pi}_{t}(s^{(i)}).$$

- With a minor change to the following recursive estimation

$$\widehat{P}^{\pi}_{t}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a^{(i)}_{t-1}|s_{t-1})}{\hat{\mu}(a^{(i)}_{t-1}|s_{t-1})}\cdot \mathbf{1}((s^{(i)}_{t-1}, s^{(i)}_{t}, a^{(i)}_{t}) = (s_{t-1}, s_t, a_t));$$

$$\widehat{r}^{\pi}_{t}(s_t) = \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a^{(i)}_{t}|s_t)}{\hat{\mu}(a^{(i)}_{t}|s_t)}r^{(i)}_{t}\cdot \mathbf{1}(s^{(i)}_{t} = s_t).$$

$\Downarrow$ plug-in model based approach

A short detour: How shall we do DM in RL?

- How would you do DM in this case?
  1. Estimate MDP

$$\hat{M} = \underline{(S, A, \hat{P}, \hat{r}, \hat{d}_0)}$$

  2. Plug-in the target policy

Value iteration

$$\hat{V}^{\pi}(\hat{d}_0) = \hat{E}^{\pi}[r_1^{(1)} + r_2^{(1)} + \dots + r_H^{(1)}]$$
$$= \sum_s \hat{d}_0(s) \hat{V}_1^{\pi}(s)$$
$$= \sum_{h=1}^{H} \sum_s \hat{d}_h^{\pi}(s) \hat{r}^{\pi}(s)$$

$\hat{v}^{\pi}$

TMIS is equivalent to DM --- a model-based approach

$$\hat{V}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}^{\pi}(s_t^{(i)})}{\hat{d}^{\mu}(s_t^{(i)})}\hat{r}_t^{\pi}(s_t^{(i)}) = \sum_{s\in S}\cancel{\frac{1}{n}\sum_{i=1}^{n}}\sum_{t=1}^{H}\frac{\hat{d}_t^{\pi}(s)}{\hat{d}_t^{\mu}(s)}\hat{r}_t(s)\cdot\mathbb{1}(s_t^{(i)}=s)$$

$$\frac{1}{n}\sum_{i=1}^{n}$$

$$= \sum_{s\in S}\sum_{t=1}^{H}\hat{d}_t^{\pi}(s)\hat{r}_t(s)\,\cancel{\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(s_t^{(i)}=s)}{\hat{d}_t^{\mu}(s)}}$$

$$= \hat{V}_{DM}$$

$$\hat{d}_1(s) = \text{plug in}$$

$$\hat{d}_2^{\pi}(s') = \sum_{s}\hat{P}_1^{\pi}(s'|s)\,\hat{d}_1(s)$$

an exact transition dynamics in $\hat{M}$

MSE of the TMIS / model-based OPE estimator

- **Theorem 3.1 (Yin and W., 2020)**

$\pi(a|s) \cdot \tau_a$

$$\mathbb{E}[(\widehat{v}^{\pi}_{\text{TMIS}} - v^{\pi})^2]$$

$$\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h, a_h} \frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h \right]$$

$$+O(n^{-1.5})$$

$\leq d_h^{\pi} \cdot \tau_a$ $\quad E_a \left[\frac{\pi}{\mu}\right]^2 \mathrm{Var}_{s'}[\cdot]$

$\mathrm{Var}\left[\frac{\pi}{\mu} \cdot V_{h+1}^{\pi}\right] = E\,\mathrm{Var} + \cancel{\mathrm{Var}_{a}[\cdot]}$

Yin & W. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS-2020*

TMIS vs on-policy evaluation

Lemma 3.4. *For any policy π and any MDP.*

$\sum_t H \quad O(H^2)$

$$\mathrm{Var}_\pi\left[\sum_{t=1}^{H} r_t^{(1)}\right] = \sum_{t=1}^{H}\Bigg(\mathbb{E}_\pi\left[\mathrm{Var}\left[r_t^{(1)} + V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\right]\right] + \mathbb{E}_\pi\left[\mathrm{Var}\left[\mathbb{E}[r_t^{(1)} + V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)}, a_t^{(1)}]\Big| s_t^{(1)}\right]\right]\Bigg).$$

on policy evaluation

$\frac{1}{n}\mathrm{Var}_\pi\left[\sum r_t^{(1)}\right]$

$O(H^2)$

Combined with the previous observation:

1. TMIS has an error that is linear in H.

$\frac{H^3 \tau_a \tau_s}{n} \implies \frac{H^2 \tau_a \tau_s}{n}$

2. TMIS is better than MC even when we are doing on-policy evaluation