Importance Sampling | Off-policy Evaluation (OPE)

$s_1, a_1, s_2, a_2, \cdots, s_H, a_H$.

↳ $a_t \sim \pi_b(\cdot|s_t)$ $\Rightarrow J(\pi)$.

Fitted-Q w/ $\mathcal{F}$. $\Rightarrow \hat{Q}^\pi$. $\Rightarrow = Q^\pi(s_0, \pi)$

$f: \mathcal{X} \to \mathbb{R}$. $p, q \in \Delta(\mathcal{X})$.

Goal: estimate $\mathbb{E}_p[f]$.

Suppose ~~$X_1, X_2, \cdots, X_n \sim p$.~~ $\frac{1}{n}\sum_{i=1}^n f(x_i) \to \mathbb{E}_p[f]$.

$X_1, X_2, \cdots, X_n \sim q$.

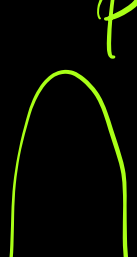
IS: $\frac{1}{n}\sum_{i=1}^n \frac{p(x_i)}{q(x_i)} f(x_i) \to \frac{p(x)}{q(x)} f(x)$.

Unbiasedness: $\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)} f(x)\right]$

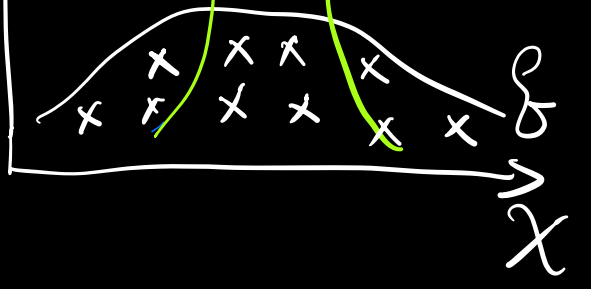
$= \sum_{x\in\mathcal{X}} q(x) \frac{p(x)}{q(x)} f(x) = \sum_x p(x) f(x) = \mathbb{E}_p[f]$.

① $q(x) > 0 \quad \forall x$. s.t. $p(x) > 0$.

② only works well



when $\frac{p(x)}{q(x)} \leq C$.



$\rho := \frac{p(x)}{q(x)}$, $\mathbb{E}_q[\rho] = 1$.

Contextual Bandits | Data: $\mathcal{D}$.

(s, a, r) : $s \sim d_0$, $a \sim \pi_b(\cdot|s)$, $r \sim R(\cdot|s,a)$

Goal: estimate $J(\pi) = \mathbb{E}_\pi[r]$.

$s \sim d_0$, $a \sim \pi(\cdot|s)$, $r \sim R(\cdot|s,a)$

$x = (s, a, r)$. p. | IS: $\frac{p(x)}{q(x)} \cdot f(x)$.

$f : (s, a, r) \mapsto r$.

$$\frac{p(x)}{q(x)} = \frac{p(s,a,r)}{q(s,a,r)} = \frac{p(s) \cdot p(a|s) \cdot p(r|s,a)}{q(s) \cdot q(a|s) \cdot q(r|s,a)}$$

$$= \frac{d_0(s) \cdot \pi(a|s) \cdot R(r|s,a)}{d_0(s) \cdot \pi_b(a|s) \cdot R(r|s,a)}$$

$$= \frac{\pi(a|s)}{\pi_b(a|s)} =: \rho.$$

$(s, a, r, \pi_b(a|s))$.

Case study: $\pi_b = \text{Unif}(A)$.

π is deterministic.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|s_i)}{\pi_b(a_i|s_i)} \cdot r_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\pi(s_i)=a_i]}{1/|A|} \cdot r_i$$

$$= \frac{1}{n/|A|} \sum_{i:\pi(s_i)=a_i} r_i.$$

$$\frac{1}{\#\{i:\pi(s_i)=a_i\}} \sum_{i:\dots} r_i$$

weighted IS.

$$\frac{\frac{1}{n}\sum \rho_i r_i}{\frac{1}{n}\sum \rho_i}.$$

$r \equiv C$.

① construct new bandit.

$r' = r - C$.

② IS in new bandit.

$J'(\pi) = 0$.

③ $J(\pi) + C$

Generalized DR (doubly robust)

assume:

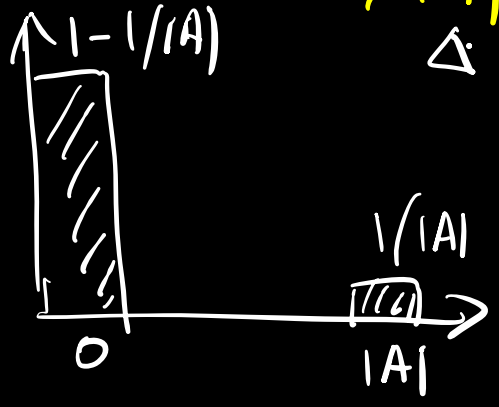
$\hat{R}(s,a)$

$$\hat{J}(\pi) = \hat{R}(s,\pi) + \rho \cdot \left(r - \hat{R}(s,a)\right)$$

Claim: $\mathbb{E}[\hat{J}(\pi)] = J(\pi)$.

$$\mathbb{E}_{a \sim \pi_b}\left[\hat{R}(s,\pi) - \rho \cdot \hat{R}(s,a)\right] = 0.$$

$$\rho = \frac{\mathbb{I}[\pi(s) = a]}{1/|A|} \begin{cases} |A|. & 1/|A|. \\ 0, & 1 - \frac{1}{|A|} \end{cases}$$



① $O\left(|A| \sqrt{\frac{1}{n} \ln \frac{1}{\delta}}\right)$. Bernstein's.

② $O\left(\sqrt{\frac{1}{n/|A|} \ln \frac{1}{\delta}}\right)$. ✓.

$$\text{Var}(\rho) = \mathbb{E}[\rho^2] - (\mathbb{E}[\rho])^2$$

$$= \mathbb{E}\left[\frac{\mathbb{I}[\dots]}{1/|A|^2}\right] - 1$$

$$= |A| - 1.$$

MDP | Data: $(s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_H, a_H, r_H)$.

$a_t \sim \pi_b(\cdot|s_t)$. $s_1 \sim d_0$.

Goal: estimate $J(\pi) = \mathbb{E}_\pi\left[\sum_{t=1}^{H} r_t\right]$

$X = (s_1, a_1, r_1, \cdots, s_H, a_H, r_H)$. $f(X) = \sum_{t=1}^{H} r_t$.

p, q are dist. over X. induced by π vs. π_b.

$$\frac{p(x)}{q(x)} = \frac{P(s_1, a_1, r_1, \cdots, s_H, a_H, r_H)}{q(\cdots)\, p(r_1 \cdots)}$$

$$= \frac{P(s_1)\cdot P(a_1|s_1)\cdot P(s_2|s_1,a_1)\cdot P(a_2|s_1,a_1,r_1,s_2)\cdots}{\cdots}$$

$$= \frac{d_0(s_1)\cdot \pi(a_1|s_1)\cdot R(r_1|s_1,a_1)\cdot P(s_2|s_1,a_1)\cdots}{d_0(s_1)\cdot \pi_b(a_1|s_1)\cdot R(r_1|s_1,a_1)\cdot P(s_2|s_1,a_1)\cdots}$$

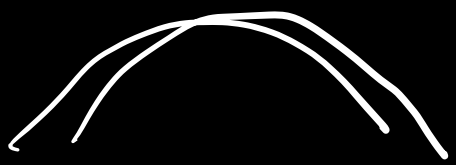
$$= \frac{\pi(a_1|s_1)\cdot \pi(a_2|s_2)\cdots \pi(a_H|s_H)}{\pi_b(a_1|s_1)\cdot \pi_b(a_2|s_2)\cdots \pi_b(a_H|s_H)} \quad \leftarrow \rho_1$$

$$= \rho_{1:H}.$$

$$\Rightarrow \left(\rho_{1:H}\right)\sum_{t=1}^{H} r_t.$$

Case Study: π_b is unif(A), π is deterministic.

$$\prod_{t=1}^{H} \frac{\mathbb{I}[\pi(s_t)=a_t]}{1/|A|} = \frac{\mathbb{I}\left[\text{"AND"}_{t=1}^{H} (\pi(s_t)=a_t)\right]}{1/|A|^H}.$$



$\rho_t \le 1+\epsilon$.

$(1+\epsilon)^H$.

$(\rho_{1:H}) \cdot \sum_{t=1}^{H} r_t$

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=1}^{H} r_t\right] = \sum_{t=1}^{H} \mathbb{E}_\pi[r_t].$$

$$\mathbb{E}_\pi[r_1] = \mathbb{E}_{\pi_b}[\rho_{1:1} r_1] \cdot \sum_{t=1}^{H} \rho_{1:H} \cdot r_t$$

$$\mathbb{E}_\pi[r_t] = \mathbb{E}_{\pi_b}[\rho_{1:t} r_t]$$

$$\Rightarrow \sum_{t=1}^{H} \rho_{1:t} \cdot r_t \quad \text{(step-wise IS)}.$$

$$J_R(\pi) + \rho_{1:t}(r_t - \hat{R}(s,a))$$

DR: $\rho_1 \cdot r_1 + \rho_{1:2} r_2 + \rho_{1:3} r_3 + \cdots$

$= \rho_1 (r_1 + \rho_2 \cdot r_2 + \rho_{2:3} r_3)$

$= \rho_1 (r_1 + \rho_2 (r_2 + \rho_3 (r_3 + \cdots + \rho_H \cdot r_H)))$

$s_H, a_H, r_H.$ $\quad \mathbb{E}[\rho_H \cdot r_H \mid s_H] = R(s_H, \pi).$

$\rho_{H-1}(r_{H-1} + \rho_H \cdot r_H)$

$\downarrow$

$R(s_H, \pi).$

$\mathbb{E}[\cdots \mid s_{H-1}, a_{H-1}] = Q^\pi(s_{H-1}, a_{H-1}).$

$V^\pi(s_t) = \rho_t (r_t + \text{estm}_{t+1})$

π $\quad$ $V^\pi(s_{t+1})$

$Q^\pi(s_t, a_t)$

π_0

$V^\pi(s_1).$

$$\hat{V}^{\pi}(s_t) + \rho_t \left(r_t + \underline{estm_{t+1}} - \hat{Q}^{\pi}(s_t, a_t) \right).$$

$estm_t$.