# Apply IS to bandits

IS: $f: X \to \mathbb{R}$, $p, q \in \Delta(X)$. $\mathbb{E}_p[f] = \mathbb{E}_q\left[\frac{p}{q} \cdot f\right]$.

---

Contextual bandit: in each round,

(1) context $x \sim d_0$.

(2) learner takes action $a$

(3) reward $r \sim R(\cdot \mid x, a)$

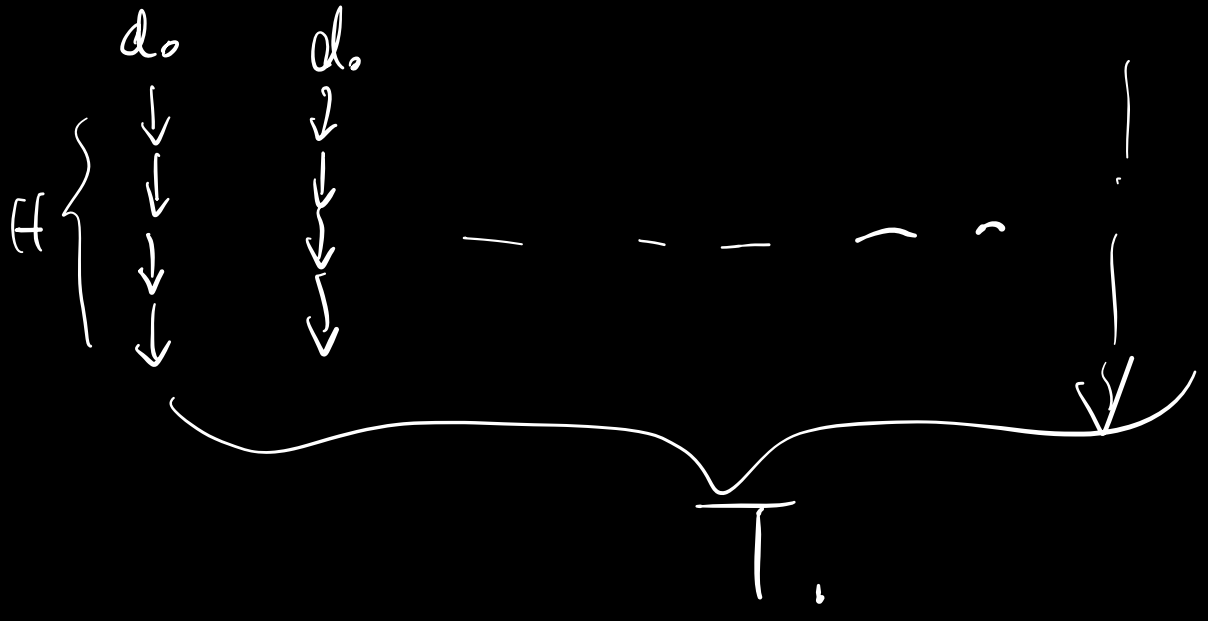
Off-policy Evaluation (OPE).

Want to estimate $J(\pi) = \mathbb{E}[r \mid \pi]$.

Data: $\{(x, a, r)\}$, $a \sim \pi_b \neq \pi$.

$\mathbb{E}_{x \sim d_0}[R(x, \pi(x))]$.

---



$(x,a,r) \sim p \iff x \sim d_0,\ a \sim \pi,\ r \sim R(\cdot|x,a)$

$$\mathbb{E}[r|\pi] = \mathbb{E}_{(x,a,r)\sim p}[r].$$

$(x,a,r) \sim q \iff x \sim d_0,\ a \sim \pi_b,\ r \sim R(\cdot|x,a)$

$$\mathbb{E}_{(x,a,r)\sim p}[r] = \mathbb{E}_{(x,a,r)\sim q}\left[\frac{p(x,a,r)}{q(x,a,r)} \cdot r\right].$$

data.

$$p(x,a,r) = p(x)\cdot p(a|x)\cdot p(r|x,a)$$
$$= d_0(x)\cdot \pi(a|x)\cdot R(r|x,a)$$

$$q(x,a,r) = d_0(x)\cdot \pi_b(a|x)\cdot R(r|x,a)$$

$$\Rightarrow \rho = \frac{\pi(a|x)}{\pi_b(a|x)}. \quad \mathbb{E}_\pi[r] = \mathbb{E}_{\pi_b}[\rho\cdot r].$$

IS.

---

$\mathbb{E}_{\pi_b}[\rho] = 1.$ $|A| = K.$ $\pi_b = \text{Unif}$

$\pi$ deterministic.

$$\frac{\pi(a|x)}{\pi_b(a|x)} = \begin{cases} K & \text{if } a = \pi(x). \\ 0 & \text{otherwise.} \end{cases}$$

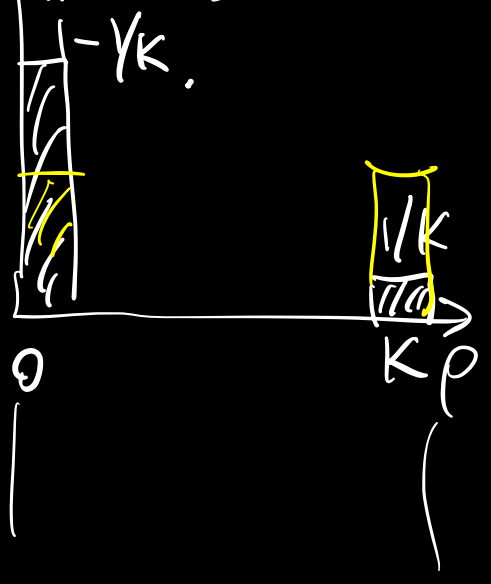
prob. mass

$$\mathrm{Var}(\rho) \le \mathop{\mathbb{E}}_{\pi_b}[\rho^2]$$

$$= \mathop{\mathbb{E}}_{\pi_b}\left[\left(\frac{\mathbb{I}[\pi(x)=a]}{1/K}\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\pi_b}\left[K \cdot \frac{\mathbb{I}[\pi(x)=a]}{1/K}\right]$$

$$= K \mathop{\mathbb{E}}_{\pi_b}[\rho] = K.$$




---

Case study: $R(\cdot|x,a)$ is deterministic and constant.

IS: $\dfrac{\mathbb{I}[\pi(x)=a]}{1/K} \cdot C$ $\quad (= C)$

$\{(x^i, a^i, r^i)\}_{i=1}^n$.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\pi(x^i)=a^i]}{1/K} \cdot C = \frac{K}{n} \cdot \sum_{i:\pi(x^i)=a^i} C.$$

$$= \frac{\sum_{i:\pi(x^i)=a^i} C}{n/K}.$$

Alt: $\frac{\sum_{i:\pi(x^i)=a^i} c}{|\{i:\pi(x^i)=a^i\}|}$.

$\rho^i = \frac{\pi(a^i|x^i)}{\pi_b(a^i|x^i)}$. IS: $\frac{1}{n}\sum_i \rho^i \cdot r^i$.

WIS: $\sum_i \frac{\rho^i r^i}{\sum_j \rho^j}$.

---

Doubly Robust.. IS: $\rho \cdot r$.

DR: Input: $\hat{R}(x,a)$.

$\hat{R}(x,\pi) + \rho \cdot (r - \hat{R}(x,a))$.

---

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

---

Remark: $\rho \cdot r = \frac{\pi(a|x)}{\pi_b(a|x)} \cdot r$.

$(x, a, r, \pi_b(a|x))$ "logging prob"

proposal

MDP: finite-horizon episodic MDP:

$(S, A, P, R, d_0, H)$.

$s_1 \sim d_0$, $a_1 \sim \pi$, $r_1 = R(s_1, a_1)$, $s_2 \sim P(\cdot|s_1, a_1)$.

$\cdots$, $s_H$, $a_H$.

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

OPE: data $\tau = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H) \sim \pi_b$.

$p$: $\tau \sim \pi$.

$q$: $\tau \sim \pi_b$.

$f$: $\tau \mapsto R(s_1, a_1) + R(s_2, a_2) + \cdots + R(s_H, a_H)$.
$:= R(\tau)$.

$$J(\pi) = \mathbb{E}_{\tau\sim p}[R(\tau)].$$

$$= \mathbb{E}_{\tau\sim q}\left[\frac{p(\tau)}{q(\tau)} R(\tau)\right].$$

$$p(\tau) = d_0(s_1) \cdot \pi(a_1|s_1) \cdot P(s_2|s_1, a_1) \cdot \pi(a_2|s_2) \cdots$$

$$q(\tau) = d_0(s_1) \cdot \pi_b(a_1|s_1) \cdot P(s_2|s_1, a_1) \cdot \pi_b(a_2|s_2) \cdots$$

$$= \prod_{h=1}^{H} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)} =: \prod_{h=1}^{H} \rho_h =: \rho_{1:H}.$$

$\pi_b$ = unif. $\pi$ : deterministic.

$$\rho_{1:H} = \prod_{h=1}^{H} \frac{\mathbb{I}[a_h = \pi(x_h)]}{1/K} = K^H \prod_{h=1}^{H} \mathbb{I}[a_h = \pi(x_h)]$$

$$= K^H \cdot \mathbb{I}[\forall h,\ a_h = \pi(x_h)].$$

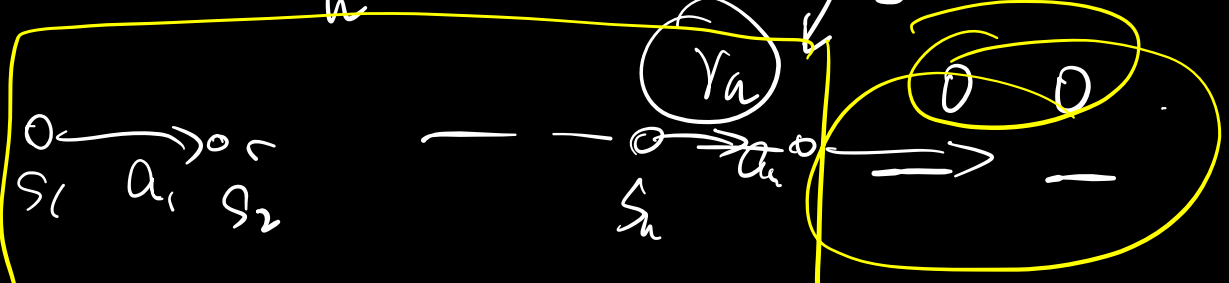
$$\prod_{h=1}^{H} \rho_h =$$

$$\left(1 + O\left(\frac{1}{H}\right)\right)^H \approx e.$$


---

$$J(\pi) = \mathbb{E}_\pi\left[\sum_h R(s_h, a_h)\right] \to R(\tau).$$

$$= \sum_h \mathbb{E}_\pi[R(s_h, a_h)].$$



$$\rho_{1:h} \cdot r_h .$$

$$\rho_{1:H} \sum_h r_h \quad , \longrightarrow \quad \sum_h \rho_{1:h} \cdot r_h .$$

$$\mathop{\mathbb{E}}_{\pi}[R(s_h, a_h)] = \mathop{\mathbb{E}}_{\pi_0}\left[\frac{p(s_h, a_h)}{q(s_h, a_h)} R(\cdot,\cdot)\right]$$

$$\frac{d_h^{\pi}(s,a)}{d_h^{\pi_0}(s,a)}$$