Importance Sampling

$\mathbb{E}_{x\sim p}[f(x)]$.

$\{x_i\}$ ~~iid~~ p.

$\frac{1}{n}\sum_i f(x_i)$

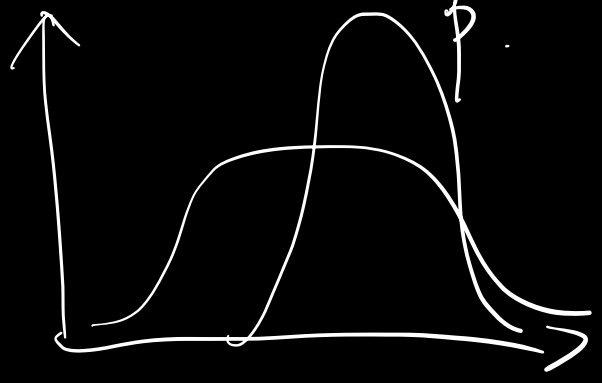
$\{x_i\}$ iid q.

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

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right] = \sum_x q(x)\cdot\frac{p(x)}{q(x)}f(x)$$

$$= \mathbb{E}_p[f].$$

Assumption: $\forall x$ s.t $p(x)>0$. we must have $q(x)>0$.



$\frac{p(x)}{q(x)} \le C. \quad \forall x.$

- $\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = 1$

App. Contextual bandits

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

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

Data: $x = (s,a,r)$. $s \sim d_0$. $a \sim \pi_b$. $r \sim R(\cdot|s,a)$. $x \sim q$.

$$J(\pi) = \mathop{\mathbb{E}}_{p}[f(x)].$$

$$(f(s,a,r) = r).$$

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

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

$$= \boxed{\frac{\pi(a|s)}{\pi_b(a|s)}} r.$$

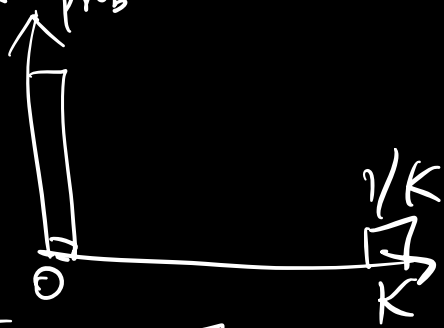
$$= \rho \cdot r.$$

Case Study: $|A| = K$. $\pi_b(a|s) = \frac{1}{K}$.

π is deterministic. $(\pi: S \to A)$.

$\frac{\mathbb{I}[\pi(s)=a]}{1/K}$. $r = \begin{cases} K \cdot r \\ = \\ 0 \end{cases}$ if $\pi(s)=a$

if $\pi(s) \neq a$. w.p. $1-1/K$.



$$\mathrm{Var}_{\beta}[\rho \cdot r] \leq \mathbb{E}_{\beta}[\rho^2 r^2]$$

$$\leq R_{max}^2 \, \mathbb{E}_{\beta}[\rho^2]$$

$$= R_{max}^2 \, \mathbb{E}_{\beta}[\rho \cdot K].$$

$$= K R_{max}^2.$$

$$\rho \cdot r \in [0, K R_{max}].$$

π_b unif. π deterministic. & r is const.

$$\mathrm{Var}[\rho r] = r^2 \mathrm{Var}[\rho].$$

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

$$= r^2 \left(\mathbb{E}[\rho] - \mathbb{E}[\rho] \right)$$

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

$$= r^2 (k-1).$$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}[a_i = \pi(x_i)]}{1/k} \cdot r_i$$

$$= \frac{1}{n/k} \cdot \sum_{i : a_i = \pi(x_i)} r_i .$$

WIS: $\frac{1}{n} \sum_{i=1}^{n} \rho_i r_i$

$$\sum_i \rho_i .$$

Doubly Robust : $\rho \cdot r$ $\left(\rho = \frac{\pi(a|s)}{\pi_b(a|s)}\right)$.

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

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

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

MDP : H-step finite horizon

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

ϕ

$s_1 \sim d_0$. $a\sim\pi$, $r_1 = R(s,a)$, $s_2 \sim P(\cdot|(s,a))$, $a_2 \sim \pi$.

$\ldots$ s_H, a_H, r_H.

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

OPE: data $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)$

$a_{1:H} \sim \pi_b$.

$\mathcal{B}$.

$$\frac{p(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)}{q(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)} \cdot \sum_{t=1}^{H} r_t .$$

$$\frac{\cancel{d_0(s_1)} \cdot \pi(a_1|s_1) \cdot \cancel{p(s_2|s_1,a_1)} \cdot \pi(a_2|s_2) \cdots}{\cancel{d_0(s_1)} \cdot \pi_b(a_1|s_1) \cdot \cancel{p(s_2|s_1,a_1)} \cdot \pi_b(a_2|s_2) \cdots}$$

$$= \prod_{t=1}^{H} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} =: \prod_{t=1}^{H} \rho_t =: \rho_{1:H}$$

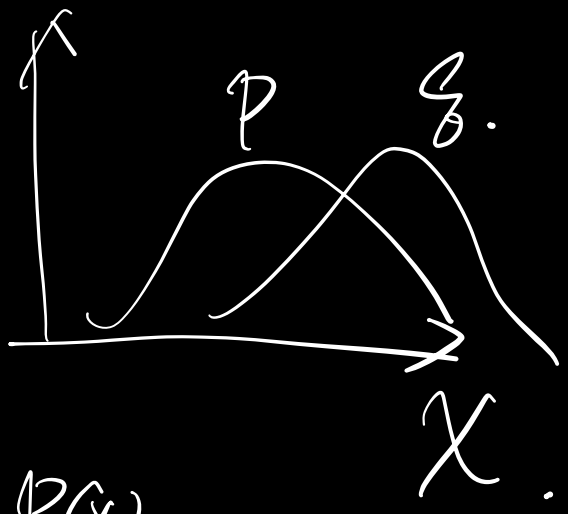
IS: $\rho_{1:H} \cdot \sum_{t=1}^{H} r_t$.

π_b unif. random: $\pi_b(a|s) = 1/K$.

deterministic π.

IS: $\left(\prod_{t=1}^{H} \frac{\mathbb{I}[\pi(s_t) = a_t]}{1/K} \right) \left(\sum_{t=1}^{H} r_t \right)$

$K \cdot \mathbb{I}[\pi(s_t) = a_t, \forall t]$.



$$\rho_t = \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} \leq C.$$

$$\rho_{1:H} \leq C^H.$$

$$\max_x \frac{p(x)}{q(x)} \geq 1.$$

$$C \leq 1 + \frac{1}{H}.$$

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

$$= \sum_{t=1}^{H} \mathbb{E}_\pi[r_t].$$

$$\mathbb{E}_\pi[r_1] = \mathbb{E}_{\pi_b}[\rho_{1:H} \cdot r_1].$$

$$= \mathbb{E}_{\pi_b}[\rho_1 \cdot r_1].$$

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

$$\mathbb{E}_\pi\left[\sum_{t=1}^{H} r_t\right] = \mathbb{E}_{\pi_b}\left[\sum_{t=1}^{H} \rho_{1:t}\, r_t\right].$$

"Step-wise" IS $\left(\sum_{t=1}^{H} \rho_{1:t} r_t\right)$, (compare to $\sum_{t=1}^{H} \rho_{1:H} r_t$).

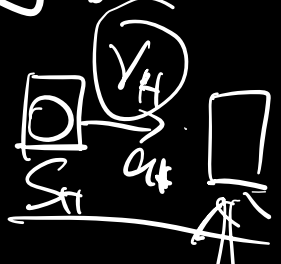
$$v_{H-t+1} := \rho_t (r_t + v_{H-t}).$$

$\frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$ — unbiased estimate of $V^{\pi}_{t+1}(s_{t+1})$.

$Q^{\pi}_t(s_t, a_t)$.

where $a_{t-1} \sim \pi_b$.

$Q^{\pi}_t(s_1, \pi) = V^{\pi}(s_1)$



$v_1 \leftarrow v_0 = 0.$

$= \rho_H \cdot r_H.$

$\hat{Q} : S \times A \to \mathbb{R}.$

$$v_{H-t+1} := \hat{Q}(s_t, \pi) + \rho_t (r_t + v_{H-t} - \hat{Q}(s_t, a_t))$$

Policy Gradient

$$\max_{\pi_\theta \in \Pi} J(\pi_\theta).$$

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\underbrace{\quad\quad}_{s_1, a_1, r_1, \cdots, s_H, a_H, r_H} \right]$$

$$\nabla_\theta J(\pi_\theta)\Big|_{\theta=\theta_0} = \mathbb{E}_{\pi_{\theta_0}}[\cdots]$$

$\nabla_\theta J(\pi_\theta)$.

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

$(s_1, a_1, r_1, \cdots, s_H, a_H, r_H) = \tau.$

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

$$= \nabla \sum_\tau \boxed{P^\pi(\tau)}\, R(\tau).$$

$$= \sum_\tau R(\tau)\, \nabla P^\pi(\tau).$$

$$= \sum_\tau P^\pi(\tau)\, R(\tau)\, \nabla \log P^\pi(\tau).$$

$$= \mathbb{E}_\pi\left[R(\tau)\, \nabla \log P^\pi(\tau) \right].$$

$$\nabla \log P^{\pi_\theta}(\tau) = \nabla \log \Big(d_0(s_1) \cdot \pi(a_1|s_1)\, P(s_2|s_1,a_1)\, \pi(a_2|s_2) \cdots \pi(a_H|s_H) \Big)$$

$$= \nabla \log d_0(s_1) + \nabla \log \pi(a_1|s_1) + \nabla \log P(s_2|s_1,a_1) + \cdots + \nabla \log \pi(a_H|s_H)$$

$$= \sum_{t=1}^{H} \nabla \log \pi(a_t|s_t).$$

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

$$= \mathbb{E}_{\pi_\theta}[\cdots].$$

"REINFORCE" $\nabla J(\pi) = \mathbb{E}_\pi \Big[\Big(\sum_{t=1}^{H} \nabla \log \pi(a_t|s_t) \Big) \Big(\sum_{t=1}^{H} \gamma_t \Big) \Big]$.

"Vanilla PG".

$$\nabla J(\pi) = \mathbb{E}_\pi \Big[\sum_{t=1}^{H} \nabla \log \pi(a_t|s_t) \sum_{t'=1}^{H} \gamma_{t'} \Big].$$

$t' = t$

$\left(Q^{\pi}(s_t, a_t) \cdot - f(s_t) \right)$.