Importance Sampling (IS) | Want to estimate $\mathbb{E}_{x\sim p}[f(x)]$.

▷ $x \sim p$. $f(x)$ is Monte-Carlo estimate. ⟵

▷ What if we only have $x \sim q$?

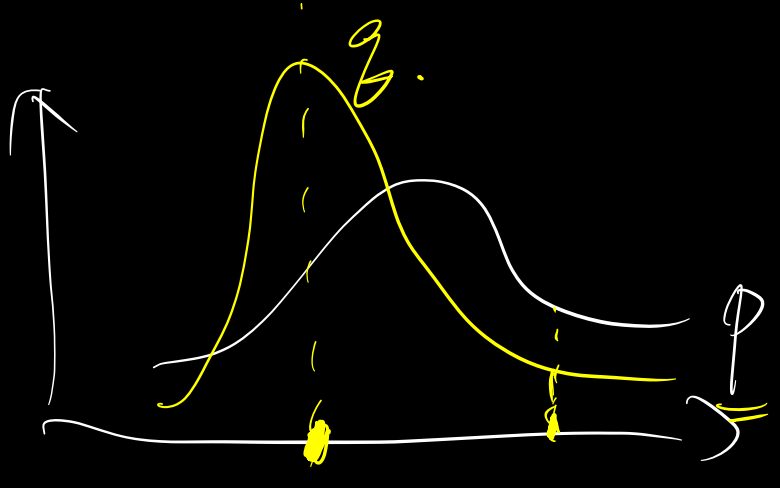
IS : $x \sim q$, $\frac{p(x)}{q(x)} \cdot f(x)$ (assume $\forall x$, s.t. $p(x)>0$. $q(x)>0$.)

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

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

$\frac{p(x)}{q(x)}$: "density ratio" / "importance weight".

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = \sum_{x\in X} q(x) \cdot \frac{p(x)}{q(x)} = 1.$$



IW "weighting"

IPS : inverse propensity scores.

Application to bandits.

behavior policy.

CB: $x \sim d_0$, $a \sim \pi_b(\cdot|x)$, $r \sim R(\cdot|x,a)$.

Off-policy evaluation: how to evaluate π.

(i.e. estimate $\mathbb{E}[r|\pi]$ using data from $\pi_b \neq \pi$.

target/eval policy.

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

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

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

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

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

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

$$= \frac{\pi(a|x)}{\pi_b(a|x)}$$

$\Rightarrow$ IS estimator for OPE: $\frac{\pi(a|x)}{\pi_b(a|x)} \cdot r$.

To apply IS in practice:

record data: $(x, a, r, \pi_b(a|x))$. $\in \mathbb{R}$

"logging/proposal prob."

Special Case

$\pi_b(a|x) = 1/|A|$. (unif policy)

$\pi(x)$ is deterministic.

IS estimator: $\frac{\mathbb{I}[\pi(x) = a]}{1/|A|} \cdot r$

- Assume $r \in [0,1]$, $\in [0, |A|]$.
- In expectation, only $n/|A|$ data pts "useful"

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

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}[\pi(x^{(i)})=a^{(i)}]}{1/|A|}\cdot r^{(i)}$$

$$=\underbrace{\frac{1}{n/|A|}}\sum_{i:\,\pi(x^{(i)})=a^{(i)}} r^{(i)}$$

expected size of $\{i:\pi(x^{(i)})=a^{(i)}\}$
(instead of actual size).

- Weighted IS / self-normalized.

replace $n/|A|$ by $|\{i:\pi(x^{(i)})=a^{(i)}\}|$.

In the general case:

replace n by $\sum_{i=1}^{n}\frac{\pi(a|x)}{\pi_b(a|x)}$.

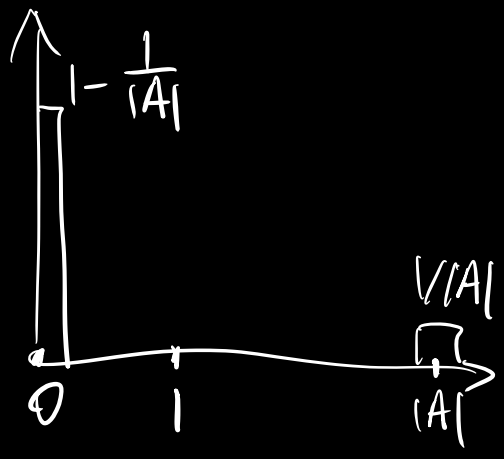
- Variance: suppose $r \in [0, 1]$.

$$\frac{\mathbb{I}[\pi(x)=a]}{1/|A|} \cdot r \in \underbrace{[0, |A|]}$$

for r.v. supported on $[0, |A|]$.

Var $\sim O(|A|^2)$ in worst case

$$\mathrm{Var}\left[\frac{\mathbb{I}[\pi(x)=a]}{1/|A|} \cdot r\right].$$



$$\leq \mathbb{E}\left[\frac{\mathbb{I}[\pi(x)=a]}{1/|A|^2} \cdot \underbrace{r^2}\right]$$

$$\leq |A| \cdot \mathbb{E}\left[\frac{\mathbb{I}[\pi(x)=a]}{1/|A|}\right] \cdot 1 = |A|.$$

- Further special case $r \equiv$ const.

  we know: $\forall \pi$, $J(\pi) =$ const

  this should be "easy".

  MC estimator: 0 variance.

  IS estimator: $\Theta(|A|)$ variance.

How to improve IS?

(i) WIS.

(ii) $\mathrm{Var}(IS) \leq \mathbb{E}\left[\frac{\mathbb{I}(\pi(x)=a)}{1/|A|^2} \cdot r^2\right]$.

to reduce variance: center reward around 0.

in the special case: { shift reward by const.
DPE
add const back.

General case: Doubly Robust (DR)

Given $\hat{Q}: \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. such that

$\hat{Q}(x,a) \approx \mathbb{E}[r \mid x, a]$.

DR: $\hat{Q}(x,\pi) + \frac{\pi(a|x)}{\pi_b(a|x)} \cdot \left(r - \hat{Q}(x,a)\right)$

Claim: DR is unbiased.

$$\mathbb{E}_{\pi_b}\left[\hat{Q}(x,\pi) - \frac{\pi(a|x)}{\pi_b(a|x)} \underbrace{\hat{Q}(x,a)}_{\sim \pi_b}\right] = 0.$$

IS is special case where $\hat{Q} \equiv 0$.

Application to multi-step RL

$M = (S, A, P, R, \gamma, d_0)$.

$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_H, a_H, r_H)$.

Collect data w/ $a_t \sim \pi_b$.

to eval policy π, $\mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \mid \pi\right]$.

Derivation of IS

Define $\tau \sim p$ as induced by π.

$\tau \sim q$ --- by π_b.

$$\mathop{\mathbb{E}}_{\tau \sim p}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h\right] = \mathop{\mathbb{E}}_{\tau \sim q}\left[\frac{p(\tau)}{q(\tau)} \cdot \sum \gamma^{h-1} r_h\right]$$

$$\frac{p(\tau)}{q(\tau)} = \frac{d_0(s_1) \cdot \pi(a_1|s_1) \cdot P(s_2|s_1,a_1) \cdot \pi(a_2|s_2) \cdots}{d_0(s_1) \cdot \pi_b(a_1|s_1) \cdot P(s_2|s_1,a_1) \cdot \pi_b(a_2|s_2) \cdots}$$

$$= \frac{\pi(a_1|s_1)}{\pi_b(a_1|s_1)} \cdots \frac{\pi(a_H|s_H)}{\pi_b(a_H|s_H)}.$$

$$=: \rho_1 \cdot \rho_2 \cdots \rho_H = \rho_{1:H}.$$

$\Rightarrow$ IS estimator: $\rho_{1:H} \cdot \sum_{h=1}^{H} \gamma^{h-1} r_h.$

Special case: π is deterministic

π_b is unif. rand.

$$\rho_{1:H} = \begin{cases} 0, & \exists h,\ \pi(s_h) \neq a_h. \\ |A|^H, & \text{o.w.} \end{cases}$$

(w.p. $1/|A|^H$).

"Curse of horizon"

In general. blow up is $\sim$

$$\left(\max_{s,a} \frac{\pi(a|s)}{\pi_b(a|s)} \right)^H$$

therefore, IS only works when either.
H is small, or, $\pi \approx \pi_b$.

Per-step IS

$$\mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \mid \pi\right] = \sum_{h=1}^{H} \gamma^{h-1} \mathbb{E}[r_h \mid \pi]$$

Idea: estimate $\mathbb{E}[r_h \mid \pi]$ for $h \in [H]$

when estimating $\mathbb{E}[r_h \mid \pi]$.

pretend traj is $(s_1, a_1, r_1, \cdots, s_h, a_h, r_h)$.

$\Rightarrow$ (step-wise) IS: $\sum_{h=1}^{H} \gamma^{h-1} \cdot \underline{\rho_{1:h}} r_h$.

DR rewrite step-wise IS.

recursively:

Define $v_0 := 0$. then $\forall h < H$.

$$v_{H-h+1} = \rho_h (r_h + \gamma v_{H-h})$$

"remaining time steps"

Claim: v_H = step-wise IS.

$$v_H = \rho_1 (r_1 + \gamma \rho_2 (r_2 + \gamma \rho_3 (\cdots)$$

$$= \rho_1 r_1 + \gamma \rho_{1:2} r_2 + \gamma^2 \rho_{1:3} r_3$$

$$+ \cdots + \gamma^{H-1} \rho_{1:H} r_H .$$

$$v_{H-h+1} = \rho_h (r_h + \gamma v_{H-h})$$

$V^{\pi}_{H-h+1}(s_h)$.

$\frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$

$Q^{\pi}_{H-h+1}(s_h, a_h)$.

$\sim \pi_b$.

$V^{\pi}_{H-h}(s_{h+1})$

Give $\hat{Q} \approx Q^\pi$.

$$V_{H-h+1}^{DR} := \hat{Q}(s_h, \pi) + \rho_h\left(r_h + \gamma V_{H-h}^{DR} - \hat{Q}(s_h, a_h)\right)$$

cancel in expectation.

Policy Gradient $\{\pi_\theta : \theta \in \Theta\}$ $\max_\theta J(\pi_\theta) := \mathbb{E}_{d_0}[V^{\pi_\theta}(s)]$

(S)GD: compute (unbiased estm) of $\nabla_\theta J(\pi_\theta)$.

Question: can we estimate $\nabla_\theta J(\pi_\theta)$ using data alone?

Answer: yes if we have data trajectories from π_θ.

Derivation Let $\nabla J(\pi) := \nabla_\theta J(\pi_\theta)$.

Goal: $\nabla J(\pi) = \mathbb{E}_{\text{traj}\sim\pi}[\underbrace{\text{function of traj}}_{\text{"policy gradient"}}]$

"REINFORCE":

$$\nabla J(\pi) = \nabla_\theta \sum_\tau \underbrace{R(\tau)}_{\text{const.}} \cdot P^{\pi_\theta}(\tau).$$

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

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

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

$$R(\tau) = \sum_{h=1}^{H} \gamma^{h-1} r_h.$$

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

$$= \sum_{\tau} R(\tau) \cdot P^{\pi}(\tau) \cdot$$

$$\nabla \log \big(\cancel{d_0(s_1)} \cdot \pi(a_1|s_1) \cdot \cancel{P(s_2|s_1,a_1)} \cdot$$

$$\pi(a_2|s_2) \cdot \cdots \cdot \cancel{P(s_H|s_{H-1},a_{H-1})} \cdot \pi(a_H|s_H) \big) .$$

$$= \sum_{\tau} P^{\pi}(\tau) \cdot R(\tau) \cdot \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h) .$$

$$= \mathbb{E}_{\tau \sim \pi} \left[\left(\sum_{h=1}^{H} \gamma^{h-1} r_h \right) \cdot \left(\sum_{h=1}^{H} \nabla \log \pi(a_h|s_h) \right) \right] .$$

"REINFORCE"

$$f(x) \cdot (\log f(x))' = f(x) \cdot \frac{f'(x)}{f(x)} = f'(x) .$$

Remark: 1. $\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \dfrac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$ | assumes scalar θ but can be generalized.

traj-

$$\Rightarrow \text{"REINFORCE"} = \lim_{\Delta\theta \to 0} \frac{IS(\pi_{\theta+\Delta\theta}) - IS(\pi_\theta)}{\Delta\theta}$$

2. "Vanilla PG" $= \lim_{\Delta\theta \to 0} \dfrac{\text{step-}IS(\pi_{\theta+\Delta\theta}) - \text{step-}IS(\pi_\theta)}{\Delta\theta}$

$$\rightarrow = \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h) \left(\sum_{h'=h}^{H} \gamma^{h'-1} \cdot r_{h'} \right)$$

|| in expectation.

3. $$\frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d^{\pi}} \left[\underbrace{Q^{\pi}(s,a)}_{\text{"PG thm"}} \cdot \nabla \log \pi(a|s) \right] .$$

4. $$\nabla J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d^{\pi}} \left[\left(Q^{\pi}(s,a) - f(s) \right) \nabla \log \pi(a|s) \right]$$

f is arbitrary. e.g. $f(s) = V^\pi(s)$, $A^\pi(s,a)$.

"baseline": $\mathbb{E}_{(s,a)\sim d^\pi}[f(s)\nabla \log \pi(a|s)]$

$= \mathbb{E}_{s\sim d^\pi}\left(\mathbb{E}_{a\sim\pi(\cdot|s)}[f(s)\nabla \log \pi(a|s)]\right)$

$= \mathbb{E}_{s\sim d^\pi}\left[f(s)\cdot \mathbb{E}_{a\sim\pi(\cdot|s)}[\nabla \log \pi(a|s)]\right]$

$\mathbb{E}_{a\sim\pi}\left[\frac{\nabla\pi(a|s)}{\pi(a|s)}\right] = \sum_a \pi(a|s)\cdot\frac{\nabla\pi(a|s)}{\pi(a|s)}$

$= \sum_a \nabla\pi(a|s)$

$= \nabla\sum_a \pi(a|s)$

$= \nabla\cdot 1 = \vec{0}$.

5. $J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}[Q^\pi(s,a)\log\pi(a|s)]$

$Q^\pi(s,a)$ can be estimated by e.g. TD.

"actor-critic"

π ↙ ↘ Q^π