Importance Sampling (IS)

Problem: estimate $\mathbb{E}_{x\sim p}[f(x)]$, where $p \in \Delta(\mathcal{X})$ and $f: \mathcal{X} \to \mathbb{R}$.

MC: draw $x_1, \dots, x_n \overset{iid}{\sim} p$, estm: $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$

if f is bounded. Hoeffding's ineq applies.

(in short, will write: $x \sim p$. $f(x)$ as the estimator).

→ Example: MC policy eval. $J(\pi) := \mathbb{E}_\pi\left[\sum_{t=1}^{H} \gamma^{t-1} r_t\right]$.

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

$p \longleftrightarrow \tau \sim \pi$. (from d_0).

$f \longleftrightarrow \tau \mapsto \sum_{t=1}^{H} \gamma^{t-1} r_t$.

IS: no access to $x \sim p$. but can sample $x \sim q$ where $q \in \Delta(\mathcal{X})$.

if: $\forall x$ where $p(x) > 0$, we have $q(x) > 0$. then: ←

$x \sim q$, $\boxed{\frac{p(x)}{q(x)}} f(x)$ ← IS estimator. IPS, IW.

imp. weight/ratio, density ratio, etc.

IS: design q.

IPS, IW: $\frac{p(x)}{q(x)}$.

Claim: IS is unbiased . assume $\mathcal{X}$ is finite.

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

ideally: $\frac{p(x)}{q(x)}$ should be small. | if $\max_x \frac{p(x)}{q(x)} \le C$. $f(x) \in [-B, B]$.

↓ $[-CB, CB]$.

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

$\|\frac{p}{q}\|_\infty$ ←

$\mathbb{E}_{x\sim q}\left[\frac{p(x)^2}{q(x)^2}\right]$ ←

Example: OPE in contextual bandit

$\mathcal{S}$: contexts, $\mathcal{A}$: actions, $R: \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$.

let $d_0 \in \Delta(\mathcal{S})$ be the ctx dist.

behavior/logging policy.

Have data: $\{(s, a, r)\}$: $s \sim d_0$, $a \sim \pi_b(\cdot|s)$, $r \sim R(\cdot|s,a)$.

target/eval policy

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

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

Proof of unbiasedness: let $(s,a,r)\sim q \iff s\sim d_0, \boxed{a\sim\pi_b}, r\sim R$ (data).

let $(s,a,r)\sim p \iff s\sim d_0, \boxed{a\sim\pi}, r\sim R$

$$J(\pi) = \mathbb{E}_{(s,a,r)\sim p}[r] = \mathbb{E}^{\text{"data"}}_{(s,a,r)\sim q}\left[\frac{p(s,a,r)}{q(s,a,r)}\cdot r\right].$$

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

$$= \mathbb{E}_{(s,a,r)\sim q}\left[\frac{\cancel{d_0(s)}\cdot \pi(a|s)\cdot \cancel{R(r|s,a)}}{\cancel{d_0(s)}\cdot \pi_b(a|s)\cdot \cancel{R(r|s,a)}}\cdot r\right] = \mathbb{E}_{(s,a,r)\sim q}\left[\frac{\pi(a|s)}{\pi_b(a|s)} r\right].$$

Variance of IS: Consider special case where

- π is deterministic. $a=\pi(s)$ — Let $K=|A|$.
- π_b is unif over A. (or $\pi_b\sim U$)

IS: $a\sim U$, $\rho\cdot r$, where $\rho = \dfrac{\pi(a|s)}{\pi_b(a|s)} = \dfrac{\mathbb{I}(a=\pi(s))}{1/K}$.

Lets further assume that r is const (ind. of s,a, has no rndness). $\boxed{r_0}$

$$\mathrm{Var}[\rho r] = r_0^2\,\mathrm{Var}[\rho]$$

$$= r_0^2\left(E[\rho^2] - (E[\rho])^2\right) = r_0^2\left(E[\rho^2]-1\right).$$

$$= r_0^2\left(\mathbb{E}\left[\frac{\mathbb{I}(a=\pi(s))}{1/K^2}\right] - 1\right)$$

$$= r_0^2\left(K\,\underbrace{\mathbb{E}\left[\frac{\mathbb{I}(a=\pi(s))}{1/K}\right]} - 1\right) = \boxed{r_0^2(K-1)}$$

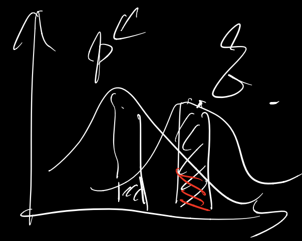
IS: $\{(s_i,a_i,r_i)\}_{i=1}^n$

$$\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{I}(a_i=\pi(s_i))}{1/K}\, r_i = \boxed{\frac{1}{n/K}}\ \boxed{\sum_{i:\,a_i=\pi(s_i)} r_i}.$$

Can we address this?

- Improvement 1: WIS. (self-normalized IPS).

In the above spec. case: $\dfrac{\sum_{i: a_i = \pi(s_i)} r_i}{\left|\{i : a_i = \pi(s_i)\}\right|}$.



General case: $\dfrac{\sum_{i=1}^{n} \rho_i \cdot r_i}{\sum_{i=1}^{n} \rho_i}$ $\quad \sum_{i=1}^{n} \rho_i \xRightarrow{\text{expe}} n$. (biased but consistent).

- Improvement 2: DR (doubly robust) (unbiased).

In the spec case: $\rho \cdot \boxed{r} \Rightarrow \hat{r}_0 + \rho\,(r - \boxed{\hat{r}_0})$

$\mathbb{E}[\rho \cdot r_0] = r_0 \cdot \mathbb{E}[\rho] = r_0$

General case: $\boxed{\hat{R}} : S \times A \to \mathbb{R}$.
arbitrary function.

$\text{DR} = \mathbb{E}_{a' \sim \pi}[\hat{R}(s,a')] + \rho \cdot (r - \hat{R}(s,a))$.

"control variate"

[Dudik, Li, Langford].

Why DR if we have good $\hat{R}$?

→ DR has low var. if $\hat{R} \approx R$.

→ DR is always unbiased even if $\hat{R}$ is poor.

→ IS is a special case of DR: $\hat{R} \equiv 0$.

regress $(s,a) \to r$. on separate data to fit $\hat{R}$.

Applications to RL