Importance Sampling

"target policy / eval policy"

Motivating Task: estimate $J(\pi) := \mathbb{E}_{s\sim d_0}[V^\pi(s_0)]$.

- w/ on-policy data: roll-out w/ π, $\sum_{t=1}^{H} \gamma^{t-1} r_t$.
- How about off-policy data? "Off-policy Evaluation" (OPE).

→ Solution: can use Expected SARSA to learn Q^π.

$$\rightarrow J(\pi) = \mathbb{E}_{s\sim d_0}[Q^\pi(s,\pi)].$$

Problem: tabular TD-alg does not scale; must use func. approx. → BIAS!!!

Question: unbiased estimator for OPE??

Answer: YES! Importance Sampling.

Intro to IS / Importance Weighting / Inverse Propensity Scores (IPS).

Problem: want to estimate $\mathbb{E}_{x\sim p}[f(x)]$.

MC: $x_i \overset{iid}{\sim} p$. $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Setup: can't sample from p, but can sample from q.

IS: $x_i \overset{iid}{\sim} q$. $\boxed{\frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i) \xrightarrow{\text{unbiased}} \mathbb{E}_{x\sim p}[f(x)]}$.

$x\sim q$, $\frac{p(x)}{q(x)} f(x)$

Proof of unbiasedness: want to show. $\mathbb{E}[\text{estimator}] = \text{what you want}$.

(Assume the domain of x is finite).

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

Remarks:

- MC is a special case: $p = q$. $\frac{p(x)}{q(x)} = 1$.
- $\frac{p(x)}{q(x)}$: "importance weight/ratio" converts $q \to p$.
- $\mathbb{E}_q\left[\frac{p(x)}{q(x)}\right] = 1$.

Application to Contextual Bandits (CB).

Episodic RL w/ $H=1$: in every "episode":

- $x \sim d_0$. (x is often called context).
- agent takes action $a \sim \underline{\pi_b}$ (behavior policy).
- receives reward $r \sim R(\cdot \mid x, a)$.

OPE: what is $J(\pi) := \mathbb{E}[r \mid \pi]$ for $\pi \neq \pi_b$

Idea:
- each data point (x, a, r).
- func of interest $(x,a,r) \mapsto r$.
- distribution of interest: $(x,a,r) \sim \pi$.

more formally: $x \sim d_0$, $a \sim \pi(\cdot \mid x)$, $r \sim R(\cdot \mid x, a)$

estimand $J(\pi) = \mathbb{E}_{x\sim d_0,\, a\sim\pi(\cdot\mid x),\, r\sim R(\cdot\mid x,a)}[r]$.

$(x,a,r) \sim$ "p"

Actual data is sampled from a different dist. "q".

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

$\Rightarrow$ Directly apply IS:

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

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

Remarks:

- π_b must be stochastic in general, π can be stoch or deterministic
- Do not need full knowledge of π_b, "logging probability".
  all you need is to record $(x, a, r, \pi_b(a|x))$.
- Useful to consider special case when π_b is uniform & π is deterministic

  importance weight $\boxed{\rho := \frac{\pi(a|x)}{\pi_b(a|x)}} = \frac{\mathbb{I}[a = \pi(x)]}{1/|A|}$.

  $$= \begin{cases} |A|, & \text{if } a = \pi(x). \\ 0, & \text{o.w.} \end{cases}$$

  therefore, roughly $1/|A|$ of the data is useful.
- IS works the best when $\pi \approx \pi_b$.

Application to TD(0):

- Standard TD(0) learns V^π from on-policy data

  e.g. (s, a, r, s'), require $\boxed{a \sim \pi}$

  form 1-step target: $\underline{r + \gamma \cdot V(s')} \xleftarrow{\text{update}} V(s)$

- Question: can we do TD(0) using off-policy data.

  e.g. $\underline{(s, a, r, s')}$, $\underline{a \sim \pi_b \neq \pi}$.

  $$\frac{\pi(a|s)}{\pi_b(a|s)} \times \left(r + \gamma \cdot V(s')\right) \xleftarrow{\text{updata}} V(s)$$

$$\mathbb{E}\left[\frac{\pi(a|s)}{\pi_b(a|s)} \times \left(r + \gamma V(s')\right)\right] = \sum_a \cancel{\pi_b(a|s)} \cdot \frac{\pi(a|s)}{\cancel{\pi_b(a|s)}} \cdot \mathbb{E}\left[r + \gamma V(s') \,\middle|\, a\right].$$

$$= \sum_a \pi(a|s)\, \mathbb{E}\left[r + \gamma V(s') \,\middle|\, a\right] = \mathbb{E}_{a \sim \pi}\left[r + \gamma V(s')\right].$$

Application to OPE in MDPs.

- Data traj. $\boxed{(s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_H, a_H, r_H)}$ "τ" :=

  where $\forall t$, $a_t \sim \pi_b$.

- want to estimate $J(\pi) := \mathbb{E}\left[\underline{\sum_{t=1}^{H} \gamma^{t-1} r_t} \,\middle|\, \forall t, \underline{a_t \sim \pi}\right]$.

Apply IS: let p be the dist. of τ if $\forall t$, $\underline{a_t \sim \pi}$.

let q be the dist of τ if $\forall t$, $a_t \sim \pi_b$.

IS estimator:

$$\frac{p(\tau)}{q(\tau)} \cdot \sum_{t=1}^{H} \gamma^{t-1} r_t.$$

$$p(\tau) = \cancel{d_0(s_1)} \cdot \pi(a_1|s_1) \cdot \cancel{R(r_1|s_1,a_1)}\, \cancel{P(s_2|s_1,a_1)} \cdot \pi(a_2|s_2) \cdot \cancel{R(r_2|s_2,a_2)}$$
$$\cancel{P(s_3|s_2,a_2)} \cdot \cdots \cancel{P(s_H|s_{H-1},a_{H-1})} \cdot \pi(a_H|s_H) \cdot \cancel{R(r_H|s_H,a_H)}$$

$$q(\tau) = \cancel{d_0(s_1)} \cdot \pi_b(a_1|s_1) \cdot \cancel{R(r_1|s_1,a_1)}\, \cancel{P(s_2|s_1,a_1)}\, \pi_b(a_2|s_2) \cdot \cdots$$

$$\Rightarrow \frac{p(\tau)}{q(\tau)} = \frac{\pi(a_1|s_1) \cdot \overbrace{\boxed{\pi(a_2|s_2)}}^{\rho_2} \cdots \pi(a_H|s_H)}{\pi_b(a_1|s_1) \cdot \boxed{\pi_b(a_2|s_2)} \cdots \pi_b(a_H|s_H)}$$

$$= \rho_1 \cdot \rho_2 \cdot \rho_3 \cdot \cdots \cdot \rho_H = \underline{\rho_{1:H}}.$$

Remarks:

- Special case: π_b is unif. random. & π is deterministic.

$$\rho_t = \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} = \frac{\mathbb{I}[a_t = \pi(s_t)]}{1/|A|} = \begin{cases} 0, & \text{o.w.} \\ |A|, & \text{if } a_t = \pi(s_t) \end{cases}$$

$$\Rightarrow \rho_{1:H} \cdot \sum_{t=1}^{H} \gamma^{t-1} r_t = \begin{cases} \underline{0}, & \text{if } \exists t, \text{ s.t. } a_t \neq \pi(s_t). \\ |A|^H \cdot \sum_{t=1}^{H} \gamma^{t-1} r_t, & \text{if } \underbrace{\forall t,\ a_t = \pi(s_t)}_{\text{occurs w/ prob. } 1/|A|^H} \end{cases}$$

Improvement to (per-traj) IS.

Idea: want to estimate $\mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_t \,\middle|\, \pi\right]$.

$$= \sum_{t=1}^{H} \gamma^{t-1} \mathbb{E}[r_t | \pi]$$

Estimate $\mathbb{E}[r_t|\pi]$: states & actions after timestep t is irrelevant.

"PDIS"
"step-wise IS"

$$\rho_{1:t} \cdot r_t \Rightarrow \sum_{t=1}^{H} \gamma^{t-1} \rho_{1:t} \cdot r_t.$$

Application to multi-step TD.

$s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, s_{t+2}$.

- on-policy 2-step TD. $(s, \boxed{a}, r, s', a', \boxed{r', s''})$

$a, a' \sim \pi$.

2-step TD target: $r + \gamma \cdot r' + \gamma^2 V(s'')$ $\xleftarrow{\text{update}}$ $V(s)$.

- off-policy? $(s, a, r, s', a', r', s'')$. where $a, a' \sim \pi_b$

importance weighted target: $\dfrac{\pi(a|s) \cdot \pi(a'|s')}{\pi_b(a|s) \cdot \pi_b(a'|s')} \cdot (r + \gamma \cdot r' + \gamma^2 V(s''))$

Improve? $\dfrac{\pi(a|s)}{\pi_b(a|s)} \cdot r + \dfrac{\pi \cdot \pi}{\pi_b \cdot \pi_b} \cdot \gamma(r' + \gamma V(s'))$.

$$\mathbb{E}_{\substack{a \sim \pi \\ a' \sim \pi}} [r + \gamma r' + \gamma^2 V(s'')]$$

$$= \mathbb{E}_{\substack{a \sim \pi \\ a' \sim \pi}} [r] + \gamma \mathbb{E}_{\substack{a \sim \pi \\ a' \sim \pi}} [r' + \gamma V(s'')].$$

$$= \mathbb{E}_{a \sim \pi} [r] + \cdots.$$