Application of IS in RL

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

Assume all trajectories terminate in H steps.

Data: $(s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_H, a_H, r_H)$. where $s_1 \sim d_0$, $a_1 \sim \pi_b$

Goal: estimate $J(\pi) := \mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \,\middle|\, \pi\right]$.

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

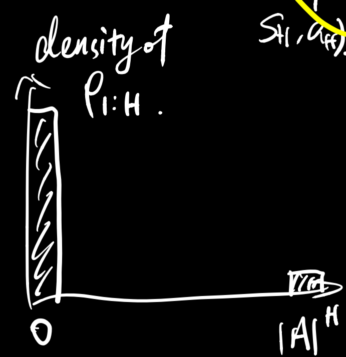
$\tau \sim p : \tau \sim \pi$. $\qquad f: \tau \mapsto \sum_{h=1}^{H} \gamma^{h-1} r_h$. $\qquad \mathbb{E}_{\tau\sim p}[f(\tau)]$

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

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

$$\frac{p(\tau)}{q(\tau)} = \frac{\cancel{d_0(s_1)} \cdot \pi(a_1|s_1) \cdot \cancel{R(r_1|s_1,a_1)} \cdot \cancel{P(s_2|s_1,a_1)} \cdot \pi(a_2|s_2) \cdots \cancel{R(r_H|s_H,a_H)}}{\cancel{d_0(s_1)} \cdot \pi_b(a_1|s_1) \cdot \cancel{R(r_1|s_1,a_1)} \cdot \cancel{P(s_2|s_1,a_1)} \cdot \pi_b(a_2|s_2) \cdots \cancel{R(r_H|s_H,a_H)}}$$

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



$\boxed{\text{Traj.-IS}}$ $\left(\sum_{h=1}^{H} \gamma^{h-1} r_h\right) \cdot \rho_{1:H}$ ← never use!

▷ Special case: π_b is unif rand. π deterministic.

only traj w/ all actions matching π gets non-zero weight (weight $= |A|^H$, w/ prob. $1/|A|^H$).

$\boxed{\text{Step. IS.}}$ $J(\pi) = \mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \,\middle|\, \pi\right] = \sum_{h=1}^{H} \gamma^{h-1} \mathbb{E}[r_h | \pi]$.

Idea: use IS to estimate $\mathbb{E}[r_h|\pi]$ for $h = 1, 2, \cdots, H$.

How to estimate $\mathbb{E}[r_1|\pi] \leftarrow \rho_1 \cdot r_1$. $\quad \left(\rho_1 = \frac{\pi(a_1|s_1)}{\pi_b(a_1|s_1)}\right)$

$\mathbb{E}[r_2|\pi] \leftarrow \rho_{1:2} \cdot r_2$.

Final: $\sum_{h=1}^{H} \gamma^{h-1} \rho_{1:h} \cdot r_h$. $\quad$ (compare: traj: $\sum_{h=1}^{H} \gamma^{h-1} \cdot \rho_{1:H} \cdot r_h$.

Alt. interpretation of per-step IS

Let $v_0 := 0$.

Claim: v_H is precisely the step-IS estimator.

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

π is target policy

unbiased for $\underline{V^{\pi}(s_{h+1})}$

unbiased. $Q^{\pi}(s_h, \pi)$. $V^{\pi}(s_h)$.

unbiased for $\underline{Q^{\pi}(s_h, \boxed{a_h})}$ π_b.

$\Rightarrow$ v_H unbiased estm of $V^{\pi}(s_0) \to J(\pi)$

Bandit DR: $\hat{R}(s,\pi) + \rho \cdot (r - \underbrace{\hat{R}(s,a)}_{\text{ideally good estimate of } r})$.

DR for RL [Jiang & Li '16]: $\hat{Q}(s,a)$: approximate Q^{π}.

$$v^{DR}_{H-h+1} = \hat{Q}(s_h,\pi) + \rho_h (r_h + \gamma v^{DR}_{H-h} - \hat{Q}(s_h,a_h)).$$

Policy Gradient (on-policy policy optimization).

$\{\pi_\theta : \theta \in \Theta\}$ policy class.

Goal: optimize $J(\pi_\theta)$. Alg: (S)GD on $J(\pi_\theta)$.

What we need: calculate stochastic gradient.

i.e. expectation $=$ $\underline{\nabla_\theta J(\pi_\theta)}$.

on-policy: $s_1, a_1, r_1, \ldots, s_H, a_H, r_H \sim \pi_\theta$.

Nontrivial: $J(\pi_\theta) \overset{=}{} V^{\pi_\theta}(s_0) = d_0^T (I - \gamma P^{\pi})^{-1} R^{\pi}$. $\leftarrow$ depends on MDP dynamics.

Abbrev: $\pi = \pi_\theta$. $\nabla = \nabla_\theta$

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

$P^\pi(\tau) = d_0(s_1)\,\pi(a_1|s_1)\,R(r_1|s_1,a_1)\cdots$.

$$\nabla J(\pi) = \nabla_\theta\left(\sum_\tau R(\tau) P^{\pi_\theta}(\tau)\right).$$

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

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

$\nabla y = y \nabla \log y$.

b/c. $\nabla \log y = \boxed{\frac{1}{y}} \nabla y$.

$$= \sum_\tau R(\tau) P^\pi(\tau)\, \nabla\Big(\log\big(d_0(s_1)\cdot \pi(a_1|s_1)\, R(r_1|s_1,a_1)\cdot P(s_2|s_1,a_1) \cdots R(r_H|s_H,a_H)\big)\Big).$$

$$= \sum_\tau R(\tau) P^\pi(\tau)\, \nabla\Big(\underbrace{\log d_0(s_1)}_{\nabla=0} + \log \pi(a_1|s_1) + \underbrace{\log \cdots}_{\nabla=0}\Big).$$

$\nabla=0$.

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

$$= \mathbb{E}\left[R(\tau)\cdot \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h) \,\Big|\, \pi\right] \leftarrow \text{"REINFORCE"}.$$

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

data π_θ. ← [Tang & Abbeel '12]?

IS. IS.

DR

$\pi \approx \pi_b$.

REINFORCE: $\left(\sum_{h=1}^{H} \gamma^{h-1} r_h\right) \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h)$. $\leftrightarrow$ $\left(\sum_{h=1}^{H} \gamma^{h-1} r_h\right)$ $\rho_{1:t+1}$.

$$\nabla J(\pi) = \nabla\left(\sum_\tau R(\tau) P^\pi(\tau)\right)$$

$$\nabla\left(\sum_{h=1}^{H} \gamma^{h-1}\, \mathbb{E}[r_h|\tau]\right).$$

$$\sum_{\tau_{pre}^h} r_h \cdot P^\pi(\tau_{pre}^h)$$

Standard PG: $\sum_{h=1}^{H} \gamma^{h-1} \nabla \log (a_h|s_h) \cdot \underbrace{\left(\sum_{h'=h}^{H} \gamma^{h'-h} r_h \right)}_{Q^\pi(s_h,a_h)}$. ← MC return

Actor-critic: estimate $\hat{Q} \approx Q^\pi$.

$$\sum_{h=1}^{H} \gamma^{h-1} \nabla \log (a_h|s_h) \hat{Q}(s_h,a_h).$$

"Baseline": $\sum \gamma^{h-1} \nabla \log (a_h|s_h) \left(\sum_{h'=h}^{H} \gamma^{h'-h} r_h - f(s_h) \right)$.

for any fixed $f = \hat{V}^\pi(s_h)$.

DR-PG [Huang & Jiang '20].

DR.

FQI: $\operatorname{argmin}_f \mathbb{E}\left(f(s,a) - r - V_{f_{k-1}}(s')\right)^2$

π^*.

Variant. $(\mathcal{V}, \Pi)$.

↑ state-function for V^*

V^π.

off-policy TD: $\operatorname{argmin}_{V \in \mathcal{V}} \mathbb{E}\left[\left(V(s) - \boxed{\frac{\pi(a|s)}{\pi_b(a|s)}} (r + V_{k-1}(s'))\right)^2\right]$

↑ V^π