AI3000 : ASSIGNMENT-3 AI20BTECH11025 VAISHNAVI

Q1] Importance Sampling

(a) Dataset = $\{(a, r)\}$

$$v^{\pi}(s) = E_{\pi}[r \mid a \sim \pi]$$
$$= E_{\pi_b}\left[\frac{\pi(a|s)}{\pi_b(a|s)} \cdot r \mid a \sim \pi_b\right]$$
$$= \underbrace{\frac{\pi(a|s)}{\pi_b(a|s)}}_{\rho} \cdot r$$
[∵ There is only data pt in the dataset]

It is an unbiased est.

$$= \rho r$$

(b) $$E_{\pi_b}\left[\frac{\pi(a|.)}{\pi_b(a|.)}\right] = \sum_{a_i \in A} \frac{\pi(a_i|.)}{\pi_b(a_i|.)} \pi_b(a_i|.)$$
$$= \sum_{a_i \in A} \pi(a_i|a) = 1$$

(∵ π_b fully supports π, $\forall a \in A$ if $\pi(a) > 0 \Rightarrow \pi_b(a) > 0$]

(c) π_b is a uniformly random policy

$a \sim U \Rightarrow \pi_b(a|s) = \frac{1}{K}$ (If there are total k actions) $|A| = k$

π is a deterministic policy

Imp sampling ratio $\rho = \frac{\pi(a|.)}{\pi_b(a|.)} = \frac{\pi(a|.)}{1/K}$

$$\pi(a|s) = \begin{cases} 1 & a = \pi(s) \\ 0 & \text{otherwise} \end{cases}$$

$$\rho(s) = \begin{cases} K & a = \pi(s) \\ 0 & \text{otherwise} \end{cases}$$

$$= \mathbb{1}_{a=\pi(s)} \cdot K$$

(d) Reward function is deterministic
i.e $R(a) = r$

π_b – Uniform behaviour policy $a \sim U$
π – Deterministic target policy.

~~Variance~~

$$Var[V^\pi] = Var\left[\rho r \mid a \sim U\right]$$
$$= r^2 \, var[\rho \mid a \sim U]$$
$$= r^2 \left[var\left(\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} \mid a \sim U\right)\right]$$
$$= r^2 \left[E\left[\frac{\pi^2(a|)}{\pi_b^2(a)} \mid a \sim U\right] - \underbrace{E\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} \mid a \sim U\right]^2}_{=1 \text{ from } 1(b)} \right]$$
$$= r^2 \left[\underbrace{E\left(\left[\mathbb{1}_{a=\pi(s)} \cdot k\right]^2 \mid a \sim U\right)}_{\sum_a \frac{1}{k} \mathbb{1}_{a=\pi(s)} k^2 = \frac{1}{k} \cdot k^2 + 0 \cdot k^2 = k} - 1 \right]$$
$$= r^2 (k-1)$$

(e)

$$Var[V^\pi] = var[\rho r] = E_{\pi_b}[\rho^2 r^2] - (E[\rho r])^2$$
$$\leq E(\rho^2 r^2)$$
$$= E_{\pi_b}\left[\frac{\pi(a)}{\pi_b(a)} \cdot \frac{\pi(a)}{\pi_b(a)} r^2\right] \qquad \because r \in [0,1]$$
$$= E_\pi\left[\frac{\pi(a)}{\pi_b(a)} \cdot r^2\right] \leq E_\pi\left[\frac{\pi(a)}{\pi_b(a)}\right]$$
$$= \sum_a \pi(a) \left[\frac{\pi(a)}{\pi_b(a)}\right] \cdot \cancel{R(a)^2} = k$$

(f) Trajectory: $\tau : s_0, a_0, s_1, a_1 \ldots s_t, a_t$

$s_0 \sim \mu(s_0)$

$P \rightarrow \pi$

$Q \rightarrow \pi_b$

Prob of occurrence of τ using policy π

$$p(\tau;\pi) = \mu(s_0) \prod_{t=1}^{\infty} \pi(a_t|s_t)\, p(s_{t+1}|a_t, s_t)$$

$$\text{IS wt}: \frac{P(\tau)}{Q(\tau)} = \frac{p(\tau;\pi)}{p(\tau;\pi_b)} = \frac{\mu(s_0) \prod_{t=1}^{\infty} \pi(a_t|s_t)\, p(s_{t+1}|a_t, s_t)}{\mu(s_0) \prod_{t=1}^{\infty} \pi_b(a_t|s_t)\, p(s_{t+1}|a_t, s_t)}$$

$$= \prod_{t=1}^{\infty} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} \quad //$$