RL Assignment-3

AI20BTECH11006

Problem 1)

a) Update rule in the algorithm is

$$V(s) \leftarrow V(s) + \alpha_t \Big[\underbrace{\frac{\pi(a|s)}{\pi_b(a|s)} r}_{\text{horizon length} = 1} - V(s) \Big]$$

Assuming initialization $V(s) = 0$

$$V(s) = \alpha_t \left[\frac{\pi(a|s)}{\pi_b(a|s)} r \right] = \alpha_1 \left[\frac{\pi(a|s)}{\pi_b(a|s)} r \right]$$

assuming $\alpha_1 = 1$

$V(s) = \frac{\pi(a|s)}{\pi_b(a|s)} r$, there is only 1 state

$$\therefore \hat{V}^{\pi} = \frac{\pi(a|s)}{\pi_b(a|s)} r$$

$$E[\hat{V}^{\pi}] = \underbrace{E_{\pi_b}}_{\text{actions are sampled from } \pi_b} \left[\frac{\pi(a|s)}{\pi_b(a|s)} r \right] = \sum_{a \in A} \frac{\cancel{\pi_b(a|s)}\, \pi(a|s)}{\cancel{\pi_b(a|s)}} r$$

$$= E_{\pi}[r]$$

$\therefore \hat{V}^{\pi}$ is an unbiased estimator of V^{π}

b)

$$\mathbb{E}_{\pi_b}\left[\frac{\pi(a|s)}{\pi_b(a|s)}\right] = \sum_{a\in A} \cancel{\pi_b(a|s)} \frac{\pi(a|s)}{\cancel{\pi_b(a|s)}}$$

$$= \sum_{a\in A} \pi(a|s) = 1$$

(c) $\pi_b(a|s) = 1/K$

$$\pi(a|s) = \begin{cases} 1 & \text{if } a = a' \\ 0 & o/w \end{cases}$$

$$IS = \frac{\pi(a|s)}{\pi_b(a|s)} = \begin{cases} K & \text{if } a = a' \\ 0 & o/w. \end{cases}$$

(d) $\hat{v}^\pi = \frac{\pi(a|s)}{\pi_b(a|s)} r$

$$\mathrm{var}(\hat{v}^\pi) = \mathbb{E}_{\pi_b}[(\hat{v}^\pi)^2] - \mathbb{E}_{\pi_b}[\hat{v}^\pi]^2$$

$$= \sum_{a\in A} \cancel{\pi_b(a|s)} \frac{\pi(a|s)^2}{\pi_b(a|s)^{\cancel{2}}} r^2 - \left(\sum_{a\in A} \cancel{\pi_b(a|s)} \frac{\pi(a|s)\, r}{\cancel{\pi_b(a|s)}}\right)^2$$

$$= K r^2 - (r)^2 = r^2(K-1)$$

(d)

From part - a

$$\hat{V}^{\pi} = \frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} r$$

Given

$R(a) = r \;\forall a$

we know that

$$Var(X) = E[X^2] - E[X]^2$$

$$Var(\hat{V}^{\pi}) = E[(\hat{V}^{\pi})^2] - E[\hat{V}^{\pi}]^2$$

$$= \sum_{a \in A} \frac{\pi(a|\cdot)^2}{\pi_b(a|\cdot)^{\cancel{2}}} r^2 \cdot \cancel{\pi_b(a|\cdot)} - \left[\sum_{a \in A} \cancel{\pi_b(a|\cdot)} \frac{\pi(a|\cdot) r}{\cancel{\pi_b(a|\cdot)}}\right]^2$$

from last subpart

$$\pi(a|\cdot) = \begin{cases} 1 & a = a' \\ 0 & o/w \end{cases}$$

$$\pi_b(a|\cdot) = \{1/K \quad \forall a \in A$$

$$= K r^2 - [r]^2 = \underbrace{r^2 (K-1)}$$

This is the variance of $\hat{V}^{\pi}$

e)

Now r is bounded between $[0, 1]$

$$\text{var}_{\pi_b}(\hat{V}^{\pi}) = E_{\pi_b}\left[\frac{\pi^2(a|s)}{\pi_b^2(a|s)} r_a^2\right] - E_{\pi_b}\left[\frac{\pi(a|s)}{\pi_b(a|s)} r_a\right]^2$$

π is a deterministic policy

$$\pi(a|\cdot) = \begin{cases} 1 & 'a = a' \\ 0 & o/w \end{cases}$$

$$\pi_b(a|\cdot) = \{1/K \quad \forall a \in A$$

$$\text{var}_{\pi_b}(\hat{V}^{\pi}) = \sum_{a \in A} \frac{\pi^2(a|s)}{\pi_b(a|s)} r_a^2 - \left(\sum_{a \in A} \pi(a|s) r_a\right)^2$$

$$= K r_{a'}^2 - r_{a'}^2 \quad [r_{a'} \text{ is a rv}]$$

$$= r_{a'}^2 (K-1)$$

$$\leq (K-1)$$

$$\because r_{a'} \in [0, 1]$$

f)

~~Let the~~ trajectory (τ) ~~length be~~ of some

~~size~~ $\cancel{t}$

Let us consider the trajectory τ till length l

$$P(\tau_l) = \begin{cases} 1 & \text{if } \forall s_i, a_i : \pi(a_i|s_i) = 1 \\ 0. & o/w \end{cases}$$

assuming a deterministic environment

$$Q(\tau_l) = K^{-l}.$$

$$\therefore \frac{P(\tau_l)}{Q(\tau_l)} = \begin{cases} 0 & \text{if } \tau_l \text{ isn't possible under } \pi \\ K^{l} & o/w \end{cases}$$

$$\frac{P(\tau)}{Q(\tau)} = \lim_{l \to \infty} \frac{P(\tau_l)}{Q(\tau_l)} = \begin{cases} \infty & \text{if } \tau \sim \pi \\ 0 & o/w \end{cases}$$