Assignment - 3

1(a) π is the target policy
π_b is the behaviour policy,
the importance sampling ratio can be written as $\frac{\pi(a|s)}{\pi_b(a|s)}$

$$v^{\pi}(s) = \mathbb{E}_{a\sim\pi}(\cdot) = \mathbb{E}_{a\sim\pi_b}\left[\frac{\pi(a|s)}{\pi_b(a|s)}\,\cdot\right]$$

The unbiased IS estimate of v^{π} is given by $\frac{\pi(a|s)}{\pi_b(a|s)}$.

1(b) $$\mathbb{E}_{a\sim\pi_b}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\right] = \sum_{a\in A} \frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\,\pi_b(a|\cdot)$$

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

$$\mathbb{E}_{a\sim\pi_b}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\right] = 1$$

1(c) $\pi_b(a|\cdot) = \frac{1}{K}$ (As there are K actions which are equiprobable)

$\pi(a|\cdot) = 1_{\pi(s)=a}$ (we know that $\pi(s) = a$)

so the probability that from state s with policy π the action taken a is 1.

importance sampling ratio = $\frac{1_{\pi(s)=a}}{1/K}$

(1)(d)

$$\text{var}_{\pi}(\rho r) = \text{var}_{\pi_b}(\rho r)$$

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

$$= r^2 \, \text{var}_{\pi_b}(\rho)$$

$$= r^2 \left[\mathbb{E}_{\pi_b}(\rho^2) - \left(\mathbb{E}_{\pi_b}(\rho)\right)^2 \right]$$

$$= r^2 \left[\frac{K^2}{K} - 1 \right]$$

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

(1)(e) $\text{var}_{\pi}(r) \leq \text{var}_{\pi_b}(\rho r)$ $\left(\rho = \frac{\pi(a|.)}{\pi_b(a|)} = K\right)$

$$= r^2 \, \text{var}_{\pi_b}(\rho)$$

$$\leq r^2 \, \mathbb{E}_{\pi_b}(\rho^2)$$ (maximum value for r for possible ρ)

$$= \mathbb{E}_{\pi_b}(\rho^2)$$

$$= \frac{K^2}{K}$$

$$\text{var}_{\pi}(r) < K$$

classmate
Date
Page

2) Q. given τ is the state action sequence given by $s_0, a_0 \ldots s_t, a_t \ldots$
we can write the probability of finding the trajectory τ with policy π as

$$P(\tau/\pi) = \mu(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t) \mathbb{P}(s_{t+1}|s_t, a_t)$$

similarly, with policy π_b, the probability can be written as

$$P(\tau/\pi_b) = \mu(s_0) \prod_{t=0}^{\infty} \pi_b(a_t|s_t) \mathbb{P}(s_{t+1}|s_t, a_t)$$

we can see from the above expressions $\mu(s_0)$ and $\mathbb{P}(s_{t+1}|s_t, a_t)$ doesnot depend on the policy.

importance sampling ratio $= \dfrac{P(\tau/\pi)}{Q(\tau/\pi_b)}$

$$= \frac{\mu(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t) \mathbb{P}(s_{t+1}|s_t, a_t)}{\mu(s_0) \prod_{t=0}^{\infty} \pi_b(a_t|s_t) \mathbb{P}(s_{t+1}|s_t, a_t)}$$

$$= \prod_{t=0}^{\infty} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$$