AI 3000 / CS 5500 : Reinforcement Learning
Assignment № 3

Due Date : 29/09/2022

Course Instructor : Easwar Subramanian 17/09/2022

Problem 1 : Importance Sampling

Consider a single state MDP with finite action space, such that $|\mathcal{A}| = K$. Assume the discount factor of the MDP γ and the horizon length to be 1. For taking an action $a \in \mathcal{A}$, let $\mathcal{R}^a(r)$ denote the unknown distribution of reward r, bounded in the range $[0, 1]$. Suppose we have collected a dataset consisting of action-reward pairs $\{(a, r)\}$ by sampling $a \sim \pi_b$, where π_b is a stochastic behaviour policy and $r \sim \mathcal{R}^a$. Using this dateset, we now wish to estimate $V^\pi = \mathbb{E}_\pi[r|a \sim \pi]$ for some target policy π. We assume that π is fully supported on π_b.

(a) Suppose the dataset consists of a single sample (a, r). Estimate V^π using importance sampling (IS). Is the obtained IS estimate of V^π is unbiased ? Explain. (2 Points)

The unbiased IS estimate of V^π is given by $\rho\, r$ where $\rho = \frac{\pi(a|s)}{\pi_b(a|s)}$. One can argue that the estimate is unbiased in the following way.

$$V^\pi(s) = \mathbb{E}_{a\sim\pi}(r) = \mathbb{E}_{a\sim\pi_b}\left(\frac{\pi(a|s)}{\pi_b(a|s)} r\right)$$

The entity $\rho\, r$ is sample estimate of the expecation in RHS

(b) Compute

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

(1 Point)

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

(c) For the case that π_b is a uniformly random policy (all K actions are equiprobable) and π a deterministic policy, provide an expression for importance sampling ratio. (1 Point)

$$\rho = \frac{1_{a=\pi(s)}}{1/K}$$

(d) For this sub-question, consider the special case when the reward r for choosing any action is identical, given by a deterministic constant r [i.e., $r = \mathcal{R}(a), \forall a \in \mathcal{A}$]. For a uniform behaviour policy π_b and a deterministic target policy π, calculate the variance of V^π estimated using importance sampling (IS) method. (5 Points)

[**Note :** Variance needs to be estimated under measure π_b]

$$\begin{aligned} V[\rho\, r|a \sim U] &= r^2 V[\rho|a \sim U] \\ &= r^2 \left(\mathbb{E}(\rho^2|a \sim U) - \mathbb{E}(\rho|a \sim U)^2\right) \\ &= r^2 \left(\mathbb{E}(\rho^2|a \sim U) - 1\right) \\ &= r^2 \left(\mathbb{E}(\left[\frac{1_{a=\pi(s)}}{1/K}\right]^2 |a \sim U) - 1\right) \\ &= r^2(K-1) \end{aligned}$$

(e) Derive an upper bound for the variance of the IS estimate of V^π for the general case when the reward distribution is bounded in the range $[0, 1]$. (3 Points)

$$V[\rho\, r|a \sim U] \leq \mathbb{E}(\rho^2 r^2|a \sim U) \leq \mathbb{E}(\rho^2 r^2|a \sim U) = K$$

(f) We now consider the case of multi-state (i.e $|\mathcal{S}| > 1$), multi-step MDP. We futher assume that $\mu(s_0)$ to be the initial start state distribution (i.e. $s_0 \sim \mu(s_0)$) where s_0 is the start state of the MDP. Let τ denote a trajectory (state-action sequence) given by, $(s_0, a_0, s_1, a_1, \cdots, s_t, a_t, \cdots)$ with actions $a_{0:\infty} \sim \pi_b$. Let Q and P be joint distributions, over the entire trajectory τ induced by the behaviour policy π_b and a target policy π, respectively. Provide a compact expression for the importance sampling weight $\frac{P(\tau)}{Q(\tau)}$. (3 Points)

[**Note :** A probablity distribution P is fully supported on another probablity distributions Q, if Q does not assign non-zero probablity to any outcome that is assigned non-zero probablity by P].

Let $\tau \sim \pi_\theta$ denote the state-action sequence given by $s_0, a_0, s_1, a_1, \cdots, s_t, a_t, \cdots$. Then, $P(\tau;\theta)$ be the probability of finding a trajectory τ with policy π

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

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

The point is that the dynamics and start state distribution gets cancelled as they don't depend on policy.