Name: Karma Tarap
CSCI E-89c Deep Reinforcement Learning
Part I of Assignment 5

Please consider a Markov Decision Process with $\mathcal{S} = \{s^A, s^B, s^C\}$.

Given a particular state $s \in \mathcal{S}$, the agent is allowed to either try staying there or switching to one of the "neighboring" states. Let's denote an intention to stay by 0, an attempt to move to the left by -1, and an intention to move to the right by $+1$. The agent does not know transition probabilities, including the distributions of rewards.

Suppose the agent uses the following behavior policy $b(a|s)$:

$$b(a|s^A) = \begin{cases} 0.5, & \text{if } a = 0, \\ 0.5, & \text{if } a = 1, \end{cases}$$

$$b(a|s^B) = \begin{cases} 1/3, & \text{if } a = -1, \\ 1/3, & \text{if } a = 0, \\ 1/3, & \text{if } a = +1, \end{cases}$$

$$b(a|s^C) = \begin{cases} 0.5, & \text{if } a = -1, \\ 0.5, & \text{if } a = 0, \end{cases}$$

to generate two episodes:

episode 1:
$S_0 = s^A, A_0 = +1, R_1 = 40, S_1 = s^B, A_1 = +1, R_2 = 30, S_2 = s^C, A_2 = 0, R_3 = 70;$

episode 2:
$S_0 = s^C, A_0 = -1, R_1 = 10, S_1 = s^B, A_1 = +1, R_2 = 50, S_2 = s^C, A_2 = 0, R_3 = 10.$

Using the First-Visit Monte Carlo (MC) prediction algorithm for estimating $v_\pi(s)$, please estimate

(a) $v_\pi(s^A)$,

(b) $v_\pi(s^B)$,

(c) $v_\pi(s^C)$,

where the target policy is $\pi(+1|s) = 1$ for $s \in \{s^A, s^B\}$ and $\pi(0|s^C) = 1$. Assume $\gamma = 0.9$.

Hint: Find the importance-sampling ratios $\rho_{t:(3-1)}$ for both episodes in cases (a)–(c).

SOLUTION: (a) $v_\pi(s^A) = R_1 * \frac{\pi(1|s^A)}{b(1|s^A)} + \gamma * R_2 * \frac{\pi(1|s^B)}{b(1|s^B)} + \gamma^2 * R_3 * \frac{\pi(0|s^C)}{b(0|s^C)}$

$= R_1 * \frac{\pi(1|s^A)}{b(1|s^A)} + \gamma * R_2 * \frac{\pi(1|s^B)}{b(1|s^B)} + \gamma^2 * R_3 * \frac{\pi(0|s^C)}{b(0|s^C)}$

$= 40 * \frac{1}{0.5} + 0.9 * 30 * \frac{1}{1/3} + 0.9^2 70 * \frac{1}{0.5}$

$= 274.4$

(b) $v_\pi(s^B) = \frac{(R_2 * \frac{\pi(1|s^B)}{b(1|s^B)} + \gamma * R_3 * \frac{\pi(0|s^C)}{b(0|s^C)}) + (R_2 * \frac{\pi(1|s^B)}{b(1|s^B)} + \gamma * R_3 * \frac{\pi(0|s^C)}{b(0|s^C)})}{2}$

$= \frac{(30*\frac{1}{1/3}+0.9*70*\frac{1}{0.5})+(50*\frac{1}{1/3}+0.9*10*\frac{1}{0.5})}{2} = 192$

(c) $v_\pi(s^C) = \frac{(70*2)+(10*0+50*0.9*3+10*0.9^2*2)}{2}\ \frac{140+151.2}{2} = 145.6$