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

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 any of the other states. Let's denote an intention to move to state s^A by a^A, to state s^B by a^B, and to state s^C by a^C. The agent does not know transition probabilities, including the distributions of rewards.

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

$$b(a|s^A) = \begin{cases} 0.5, & \text{if } a = a^A, \\ 0.25, & \text{if } a = a^B, \\ 0.25, & \text{if } a = a^C, \end{cases}$$

$$b(a|s^B) = \begin{cases} 0.25, & \text{if } a = a^A, \\ 0.5, & \text{if } a = a^B, \\ 0.25, & \text{if } a = a^C, \end{cases}$$

$$b(a|s^C) = \begin{cases} 0.25, & \text{if } a = a^A, \\ 0.25, & \text{if } a = a^B, \\ 0.5, & \text{if } a = a^C, \end{cases}$$

to generate two episodes:

episode 1:
$S_0 = s^A, A_0 = a^B, R_1 = 20, S_1 = s^B, A_1 = a^C, R_2 = 10, S_2 = s^C, A_2 = a^A, R_3 = 90, S_3 = s^A, A_3 = a^C, R_4 = 30;$

episode 2:
$S_0 = s^C, A_0 = a^A, R_1 = 50, S_1 = s^C, A_1 = a^C, R_2 = 30, S_2 = s^C, A_2 = a^B, R_3 = 10, S_3 = s^B, A_3 = a^B, R_4 = 20.$

Using the Every-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 $\pi(a^C|s) = 1$ for $s \in \{s^A, s^B\}$ and $\pi(a^C|s^C) = 1$. Assume $\gamma = 0.9$.

SOLUTION:

(a)

episode 1:

$G_0 = 20 + \gamma * 10 + \gamma^2 * 90 + \gamma^3 * 30 = 123.77$

$\delta_{0:T-1} = \frac{\pi(a^B|s^A)}{b(a^B|s^A)} * \frac{\pi(a^C|s^B)}{b(a^C|s^B)} * \frac{\pi(a^A|s^C)}{b(a^A|s^C)} * \frac{\pi(a^C|s^A)}{b(a^C|s^A)} = \frac{0}{.25} * \frac{1}{.25} * \frac{0}{.25} * \frac{1}{.25} = 0$

$G_3 = 30$

$\delta_{3:T-1} = \frac{\pi(a^C|s^A)}{b(a^C|s^A)} = \frac{1}{1/4} = 4$

episode 2:

N/A

$V_\pi(s^A) \approx \frac{1}{2}[0(123.77) + 4(30)] = 60$

(b)

episode 1:

$G_1 = 10 + 0.9 * 90 + 0.9^2 * 30 = 115.3$

$\delta_{1:T-1} = \frac{\pi(a^C|s^B)}{b(a^C|s^B)} * \frac{\pi(a^A|s^C)}{b(a^A|s^C)} * \frac{\pi(a^C|s^A)}{b(a^C|s^A)} = \frac{1}{.25} * \frac{0}{.25} * \frac{1}{.25} = 0$

episode 2:

$G_3 = 20$

$\delta_{3:T-1} = \frac{\pi(a^B|s^B)}{b(a^B|s^B)} = \frac{0}{0.5} = 0$

$V_\pi(s^B) \approx \frac{1}{2}[0(115.3) + 0(20)] = 0$

(c)

episode 1:

$G_2 = 90 + 0.9 * 30 = 117$

$\delta_{2:T-1} = \frac{\pi(a^A|s^C)}{b(a^A|s^C)} * \frac{\pi(a^C|s^A)}{b(a^C|s^A)} = \frac{0}{.25} * \frac{1}{.25} = 0$

episode 2:

$G_0 = 50 + 0.9 * 30 + 0.9^2 * 10 + 0.9^3 * 20 = 99.68$

$\delta_{0:T-1} = \frac{\pi(a^A|s^C)}{b(a^A|s^C)} * \frac{\pi(a^C|s^C)}{b(a^C|s^C)} * \frac{\pi(a^B|s^C)}{b(a^B|s^C)} * \frac{\pi(a^B|s^B)}{b(a^B|s^B)} = \frac{0}{.25} * \frac{1}{.5} * \frac{0}{.25} * \frac{0}{.5} = 0$

$G_1 = 30 + 0.9 * 10 + 0.9^2 * 20 = 55.2$

$\delta_{1:T-1} = \frac{1}{.5} * \frac{0}{.25} * \frac{0}{.5} = 0$

$G_2 = 10 + 0.9 * 20 = 28$

$\delta_{2:T-1} = \frac{0}{.25} * \frac{0}{.5} = 0$

$V_\pi(s^C) \approx \frac{1}{4}[0(117) + (0(99.68) + 0(55.2) + 0(28)] = 0$