NAME: ... KEY .....
CSCI E-89c Deep Reinforcement Learning
Part I of Assignment 7

Please consider a Markov Decision Process (MDP) 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. There is, however, some evidence that the agent gets rewards only at the entrance to s^C; and transition MDP probabilities to/from s^A appear to be same (or nearly same) as to/from s^B.

Suppose the agent chooses policy $\pi(a^A|s) = 0.05$, $\pi(a^B|s) = 0.05$, $\pi(a^C|s) = 0.90$ for all $s \in \{s^A, s^B, s^C\}$. Because of the apparent symmetry between s^A and s^B, it makes sense to assume that $v_\pi(s^A) \approx v_\pi(s^B)$ and approximate the state-values as follows:

$$v_\pi(s) \approx \hat{v}(s, \mathbf{w}) = w_1 \cdot \mathbb{1}_{(s=s_A)} + w_1 \cdot \mathbb{1}_{(s=s_B)} + w_2 \cdot \mathbb{1}_{(s=s_C)}.$$

Please notice that $\hat{v}(s^A, \mathbf{w}) = \hat{v}(s^B, \mathbf{w})$ for any choice of weights.

Assume the agent runs the Semi-gradient 1-step Temporal-Difference (TD) prediction algorithm with $\alpha = 0.1$ for estimating v_π:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[\underbrace{R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t)}_{\approx v_\pi(S_t)} - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t)$$

with $\gamma = 0.9$ and zero weights at time $t = 0$.

If the agent observes the following sequence of states, actions, and rewards:

$$\begin{aligned}
&S_0 = s^A, A_0 = a^C, R_1 = 20, \\
&S_1 = s^C, A_1 = a^B, R_2 = 0, \\
&S_2 = s^B, A_2 = a^C, R_3 = 20, \\
&S_3 = s^C, A_3 = a^C, R_4 = 20, \\
&S_4 = s^C, A_4 = a^B, R_5 = 0, \\
&S_5 = s^B,
\end{aligned}$$

find (a) weights $\mathbf{w}_t$ and (b) corresponding approximations $\hat{v}(s, \mathbf{w}_t)$ for $t = 1, 2, \ldots, 5$. Specifically, please fill the tables in below:

SOLUTION:

The gradient is

$$\begin{aligned}
\nabla \hat{v}(S_t, \mathbf{w}_t) =& \nabla \{w_1 \cdot \mathbb{1}_{(S_t=s_A)} + w_1 \cdot \mathbb{1}_{(S_t=s_B)} + w_2 \cdot \mathbb{1}_{(S_t=s_C)}\} \\
=& \left(\mathbb{1}_{(S_t=s_A)} + \mathbb{1}_{(S_t=s_B)}, \mathbb{1}_{(S_t=s_C)}\right)^T,
\end{aligned}$$

then the Semi-gradient 1-step Temporal-Difference (TD) prediction algorithm becomes

$$\begin{bmatrix} w_{t+1,1} \\ w_{t+1,2} \end{bmatrix} \doteq \begin{bmatrix} w_{t,1} \\ w_{t,2} \end{bmatrix} + \alpha \left[\underbrace{R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t)}_{\approx v_\pi(S_t)} - \hat{v}(S_t, \mathbf{w}_t) \right] \begin{bmatrix} \mathbb{1}_{(S_t = s_A)} + \mathbb{1}_{(S_t = s_B)} \\ \mathbb{1}_{(S_t = s_C)} \end{bmatrix}.$$

Then for

$$\begin{aligned} &S_0 = s^A, A_0 = a^C, R_1 = 20, \\ &S_1 = s^C, A_1 = a^B, R_2 = 0, \\ &S_2 = s^B, A_2 = a^C, R_3 = 20, \\ &S_3 = s^C, A_3 = a^C, R_4 = 20, \\ &S_4 = s^C, A_4 = a^B, R_5 = 0, \\ &S_5 = s^B, \end{aligned}$$

we have

$$\begin{bmatrix} w_{1,1} \\ w_{1,2} \end{bmatrix} \doteq \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \alpha \left[R_1 + \gamma \hat{v}(S_1, (0,0)^T) - \hat{v}(S_0, (0,0)^T) \right] \begin{bmatrix} \mathbb{1}_{(S_0 = s_A)} + \mathbb{1}_{(S_0 = s_B)} \\ \mathbb{1}_{(S_0 = s_C)} \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} w_{2,1} \\ w_{2,2} \end{bmatrix} \doteq \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \alpha \left[R_2 + \gamma \hat{v}(S_2, (2,0)^T) - \hat{v}(S_1, (2,0)^T) \right] \begin{bmatrix} \mathbb{1}_{(S_1 = s_A)} + \mathbb{1}_{(S_1 = s_B)} \\ \mathbb{1}_{(S_1 = s_C)} \end{bmatrix} = \begin{bmatrix} 2 \\ 0.18 \end{bmatrix},$$

$$\begin{bmatrix} w_{3,1} \\ w_{3,2} \end{bmatrix} \doteq \begin{bmatrix} 2 \\ 0.18 \end{bmatrix} + \alpha \left[R_3 + \gamma \hat{v}(S_3, (2,0.18)^T) - \hat{v}(S_2, (2,0.18)^T) \right] \begin{bmatrix} \mathbb{1}_{(S_2 = s_A)} + \mathbb{1}_{(S_2 = s_B)} \\ \mathbb{1}_{(S_2 = s_C)} \end{bmatrix} = \begin{bmatrix} 3.8162 \\ 0.18 \end{bmatrix},$$

$$\begin{bmatrix} w_{4,1} \\ w_{4,2} \end{bmatrix} \doteq \begin{bmatrix} 3.8162 \\ 0.18 \end{bmatrix} + \alpha \left[R_4 + \gamma \hat{v}(S_4, (3.8162,0.18)^T) - \hat{v}(S_3, (3.8162,0.18)^T) \right] \begin{bmatrix} \mathbb{1}_{(S_3 = s_A)} + \mathbb{1}_{(S_3 = s_B)} \\ \mathbb{1}_{(S_3 = s_C)} \end{bmatrix} = \begin{bmatrix} 3.8162 \\ 2.1782 \end{bmatrix},$$

$$\begin{bmatrix} w_{5,1} \\ w_{5,2} \end{bmatrix} \doteq \begin{bmatrix} 3.8162 \\ 2.1782 \end{bmatrix} + \alpha \left[R_5 + \gamma \hat{v}(S_5, (3.8162,2.1782)^T) - \hat{v}(S_4, (3.8162,2.1782)^T) \right] \begin{bmatrix} \mathbb{1}_{(S_4 = s_A)} + \mathbb{1}_{(S_4 = s_B)} \\ \mathbb{1}_{(S_4 = s_C)} \end{bmatrix} = \begin{bmatrix} 3.8162 \\ 2.3038 \end{bmatrix}.$$

(a) weights $\mathbf{w}_t = (w_{1,t}, w_{2,t})^T$:

	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
$w_{1,t}$	0	2	2	3.8162	3.8162	3.8162
$w_{2,t}$	0	0	0.18	0.18	2.1782	2.3038

(b) approximations $\hat{v}(s, \mathbf{w}_t)$:

	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
$\hat{v}(s^A, \mathbf{w}_t)$	0	2	2	3.8162	3.8162	3.8162
$\hat{v}(s^B, \mathbf{w}_t)$	0	2	2	3.8162	3.8162	3.8162
$\hat{v}(s^C, \mathbf{w}_t)$	0	0	0.18	0.18	2.1782	2.3038