NAME: Andrew Caide

CSCI S-89c Deep Reinforcement Learning
Part I of Assignment 8

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 TD(λ) with Approximation for estimating v_π:

$$\begin{aligned}
\mathbf{z}_{-1} &\doteq (0,0)^T, \\
\mathbf{z}_t &\doteq \gamma\lambda \mathbf{z}_{t-1} + \nabla \hat{v}(S_t, \mathbf{w}_t) \quad \text{for } t \geq 0, \\
\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t) \right] \mathbf{z}_t \quad \text{for } t \geq 0,
\end{aligned}$$

where $\lambda = 0.2$, $\alpha = 0.1$, $\gamma = 0.9$, and weights $\mathbf{w}_t$ are set to zero 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:

(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	2.03	3.88	4.25	4.255
$w_{2,t}$	0	0	0.18	0.5	2.55	2.73

(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	2.03	3.88	4.55	4.26
$\hat{v}(s^B, \mathbf{w}_t)$	0	2	2.03	3.88	4.55	4.26
$\hat{v}(s^C, \mathbf{w}_t)$	0	0	~~0.18~~ $\boxed{0.18}$	0.5	2.55	2.73

$v_\pi(s) \approx \hat{v}(s, \omega) = \omega_1 \mathbb{1}_{s^A} + \omega_1 \mathbb{1}_{s^B} + \omega_2 \mathbb{1}_{s^C}$

$\boxed{t=1}$

$$Z_{t=1} = \gamma \lambda Z_0 + \nabla \hat{v}(s^A, \omega)$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\omega_{t+1} \doteq \omega_t + \alpha \left[R_{t+1} + \gamma \hat{v}(s_{t+1}, \omega_t) - \hat{v}(s_t, \omega_t) \right] Z_{t=1}$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \alpha \left[20 + 0.9\, \hat{v}(s^C, \omega_t) - 0 \right] \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= 0.1 \left[20 + 0.9 \cdot 0 \right] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$v(s) \approx \hat{v}(s, \omega) \Rightarrow$

$$\boxed{\begin{aligned} v(s^A, \omega) &= 2 + 0 + 0 \\ v(s^B, \omega) &= 0 + 2 + 0 \\ v(s^C, \omega) &= 0 + 0 + 0 \cdot 1 = 0 \end{aligned}}$$

$\boxed{t=2}$

$$Z_{t=2} = \gamma\lambda\, Z_{t=1} + \nabla \hat{v}(s^C, w)$$

$$= 0.18 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.18 \\ 1 \end{bmatrix}$$

$$w_{t=2} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \alpha\left[0 + \gamma\, \hat{v}(s^B, w_t) - \hat{v}(s^C, w_t)\right] \begin{bmatrix} 0.18 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \alpha\left[\gamma \cdot 2 - 0\right] \begin{bmatrix} 0.18 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} + 0.18 \begin{bmatrix} 0.18 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.03 \\ 0.18 \end{bmatrix}$$

$$V_\pi(S^A) = V_\pi(S^B) = 2.03$$

$$V_\pi(S^C) = 0.18$$

$\boxed{t=3}$

$$Z_{t=3} = \gamma\lambda\, Z_{t=2} + \nabla \hat{V}(S^B, w)$$

$$= 0.18 \begin{bmatrix} 0.18 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.03 + 1 \\ 0.18 + 0 \end{bmatrix} = \begin{bmatrix} 1.03 \\ 0.18 \end{bmatrix}$$

$$w_{t=3} = \begin{bmatrix} 2.03 \\ 0.18 \end{bmatrix} + \alpha\left[20 + \gamma\, \hat{v}(s^C, w_t) - \hat{v}(s^B, w_t)\right] \begin{bmatrix} 1.03 \\ 0.18 \end{bmatrix}$$

$$= \begin{bmatrix} 2.03 \\ 0.18 \end{bmatrix} + 0.1\left[20 + \gamma(0.18) - 2.03\right] \begin{bmatrix} 1.03 \\ 0.18 \end{bmatrix}$$

$$= \begin{bmatrix} 2.03 \\ 0.18 \end{bmatrix} + 1.8 \begin{bmatrix} 1.03 \\ 0.18 \end{bmatrix} = \begin{bmatrix} 2.03 + 1.85 \\ 0.18 + 0.32 \end{bmatrix} = \begin{bmatrix} 3.88 \\ 0.50 \end{bmatrix}$$

$$V_\pi(A) = V_\pi(B) = 3.88$$

$$V_\pi(C) = 0.5$$

$\boxed{t=4}$ $\quad Z_{t=4} = \gamma\lambda Z_{t=3} + \nabla\hat{V}(s^c, w) = 0.18\begin{bmatrix}1.03\\0.18\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}$

$$= \begin{bmatrix}.185\\1.03\end{bmatrix}$$

$$w_{t=4} = \begin{bmatrix}3.88\\0.5\end{bmatrix} + \alpha\left[20 + \gamma\hat{v}(s^c, w) - \hat{v}(s^c, w)\right]\begin{bmatrix}0.185\\1.03\end{bmatrix}$$

$$= \begin{bmatrix}3.88\\0.5\end{bmatrix} + \alpha\left[20 + \gamma\cdot 0.5 - 0.5\right]\begin{bmatrix}0.185\\1.03\end{bmatrix}$$

$$= \begin{bmatrix}3.88\\0.5\end{bmatrix} + 1.995\begin{bmatrix}0.185\\1.03\end{bmatrix} = \begin{bmatrix}3.88 + 0.37\\0.5 + 2.05\end{bmatrix} = \begin{bmatrix}4.25\\2.55\end{bmatrix}$$

$$V_\pi(A) = V_\pi(B) = 4.55$$

$$V_\pi(C) = 2.55$$

$\boxed{t=5}$ $\quad Z_{t=5} = \gamma\lambda Z_{t=4} + \nabla\hat{V}(s^c; w) = 0.18\begin{bmatrix}0.185\\1.03\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}$

$$= \begin{bmatrix}0.033\\1.185\end{bmatrix}$$

$$w_{t=5} = \begin{bmatrix}4.25\\2.55\end{bmatrix} + \alpha\left[0 + \gamma\hat{v}(a^B, w_t) - \hat{v}(a^c, w_t)\right]\begin{bmatrix}0.033\\1.185\end{bmatrix}$$

$$= \begin{bmatrix}4.25\\2.55\end{bmatrix} + \alpha\left[0.9(4.55) - 2.55\right]\begin{bmatrix}0.033\\1.185\end{bmatrix}$$

$$= \begin{bmatrix}4.25\\2.55\end{bmatrix} + 0.1545\begin{bmatrix}0.033\\1.185\end{bmatrix} = \begin{bmatrix}4.25 + 0.005\\2.55 + 0.18\end{bmatrix} = \begin{bmatrix}4.255\\2.73\end{bmatrix}$$