NAME: Andrew Caide ....

CSCI S-89c Deep Reinforcement Learning

Part I of Assignment 9

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 that the agent runs the following algorithm with $\alpha = 0.1$ and $m = 2$ for estimating v_π:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha \sum_{t=mk}^{m(k+1)-1} \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_k) - \hat{v}(S_t, \mathbf{w}_k)\right] \nabla \hat{v}(S_t, \mathbf{w}_k), \ k = 0, 1, 2, \ldots$$

This algorithm is a modification of the Semi-gradient 1-step Temporal-Difference (TD) with the model now being trained in mini-batches of size m. Please use $\gamma = 0.9$ and zero weights for $k = 0$.

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

$$\begin{aligned}
S_0 &= s^A, A_0 = a^C, R_1 = 20, \\
S_1 &= s^C, A_1 = a^B, R_2 = 0, \\
S_2 &= s^B, A_2 = a^C, R_3 = 20, \\
S_3 &= s^C, A_3 = a^C, R_4 = 20, \\
S_4 &= s^C, A_4 = a^B, R_5 = 20, \\
S_5 &= s^C, A_5 = a^C, R_6 = 0, \\
S_6 &= s^B,
\end{aligned}$$

find (a) weights $\mathbf{w}_k$ and (b) corresponding approximations $\hat{v}(s, \mathbf{w}_k)$ for iteration step $k = 1, 2, 3$. Specifically, please fill the tables in below:

SOLUTION:

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

	$k=0$	$k=1$	$k=2$	$k=3$
$w_{1,k}$	0	2	3.8	3.8
$w_{2,k}$	0	0	2	4.122

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

	$k=0$	$k=1$	$k=2$	$k=3$
$\hat{v}(s^A, \mathbf{w}_k)$	0	2	3.8	3.8
$\hat{v}(s^B, \mathbf{w}_k)$	0	2	3.8	3.8
$\hat{v}(s^C, \mathbf{w}_k)$	0	0	2	4.122

$$w_{k+1} = w_k + \alpha \sum_{t=mk}^{m(k+1)-1} \left[R_{t+1} + \gamma \hat{v}(s_{t+1}, w_k) - \hat{v}(s_t, w_k) \right] \nabla \hat{v}(s_t, w_k)$$

$m=2,\ k=1,\ s^0 = s^A,\ s^1 = s^C,\ s^2 = s^B$

$$w_1 = w_0 + \alpha \sum_{0}^{m-1=1} \begin{bmatrix} [R_1 + \gamma \hat{v}(s_1, w_k) - \hat{v}(s_0, w_k)] \nabla \hat{v}(s_0, w_k) + \\ [R_2 + \gamma \hat{v}(s_2, w_k) - \hat{v}(s_1, w_k)] \nabla \hat{v}(s_1, w_k) \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} (20+0-0)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \\ (0+0-0)\begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{bmatrix} = 0.1 \begin{bmatrix} 20+0 \\ 0+0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$\boxed{\begin{matrix} V_\pi(A) = V_\pi(B) = 2 \\ V_\pi(C) = 0 \end{matrix}}$$

$k=2,\ m=2,\ s_2 = s^B,\ s_3 = s^C,\ s_4 = s^C$

$$w_2 = w_1 + \alpha \sum_{t=2}^{3} \begin{bmatrix} [R_3 + \gamma \hat{v}(s_3, w_k) - \hat{v}(s_2, w_k)] \nabla \hat{v}(s_2, w_k) + \\ [R_4 + \gamma \hat{v}(s_4, w_k) - \hat{v}(s_3, w_k)] \nabla \hat{v}(s_3, w_k) \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} (20+0-2)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \\ (20+0-0)\begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} + 0.1 \begin{bmatrix} 18 \\ 20 \end{bmatrix} = \begin{bmatrix} 3.8 \\ 2 \end{bmatrix}$$

$$\boxed{\begin{matrix} V_\pi(A) = V_\pi(B) = 3.8 \\ V_\pi(C) = 2 \end{matrix}}$$

Andrew Caide
Problem Set 9

$K=3,\ m=2$

$S_4 = S^C,\ S_5 = S^C,\ S_6 = S^B$

$$w_3 = w_2 + \alpha \sum_{t=4}^{5} \begin{bmatrix} [R_5 + \gamma \hat{v}(S_5, w_t) - \hat{v}(S_4, w_t)] \nabla \hat{v}(S_4, w_t) + \\ [R_6 + \gamma \hat{v}(S_6, w_t) - \hat{v}(S_5, w_t)] \nabla \hat{v}(S_5, w_t) \end{bmatrix}$$

$$= \begin{bmatrix} 3.8 \\ 2 \end{bmatrix} + \alpha \begin{bmatrix} [20 + 0.9(2) - 2] \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \\ [0 + 0.9(3.8) - 2] \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 3.8 \\ 2 \end{bmatrix} + 0.1 \left[19.8 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 1.42 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right]$$

$$= \begin{bmatrix} 3.8 + 0 + 0 \\ 2 + 1.98 + .142 \end{bmatrix} = \begin{bmatrix} 3.8 \\ 4.122 \end{bmatrix}$$

$$\boxed{\begin{array}{l} V_\pi(A) = V_\pi(B) = 3.8 \\ \qquad V_\pi(C) = 4.122 \end{array}}$$