Homework 5

Jerry Duncan

October 27, 2020

1 Problem 1

We're given the trajectory tabularized in Table 1 for answering part a, b, and c.

Table 1: States, Actions, and Rewards for Problem 1

T	S_T	A_T	R_{T+1}
1	s_1	a_1	-8
2	s_1	a_2	-16
3	s_2	a_1	20
4	s_1	a_2	-10
5	s_2	a_1	

Part a For part a, we're using Q-Learning that is shown in Eq. 1. Also, $\alpha = 0.5$ and $\gamma = 0.5$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \text{max}_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)] \quad (1)$$

Step 1:

$$Q_0(s_1) = [0, 0]$$
$$Q_0(s_2) = [0, 0]$$
$$Q_1(s_1, a_1) = Q_0(s_1, a_1) + \alpha[R_2 + \gamma \text{max}_{a'} Q_0(s_1, a') - Q_0(s_1, a_1)]$$
$$Q_1(s_1, a_1) = 0 + 0.5[-8 + 0.5 \cdot 0 - 0] = -4$$

Step 2:

$$Q_1(s_1) = [-4, 0]$$
$$Q_1(s_2) = [0, 0]$$
$$Q_2(s_1, a_2) = Q_1(s_1, a_2) + \alpha[R_3 + \gamma \text{max}_{a'} Q_1(s_2, a') - Q_1(s_1, a_2)]$$
$$Q_2(s_1, a_2) = 0 + 0.5[-16 + 0.5 \cdot 0 - 0] = -8$$

Step 3:

$$Q_2(s_1) = [-4, -8]$$
$$Q_2(s_2) = [0, 0]$$
$$Q_3(s_2, a_1) = Q_2(s_2, a_1) + \alpha[R_4 + \gamma \text{max}_{a'} Q_2(s_1, a') - Q_2(s_2, a_1)]$$
$$Q_3(s_1, a_2) = 0 + 0.5[20 + 0.5 \cdot -4 - 0] = 9$$

Step 4:

$$Q_3(s_1) = [-4, -8]$$
$$Q_3(s_2) = [9, 0]$$
$$Q_4(s_1, a_2) = Q_3(s_1, a_2) + \alpha[R_5 + \gamma \text{max}_{a'} Q_3(s_2, a') - Q_3(s_1, a_2)]$$
$$Q_4(s_1, a_2) = -8 + 0.5[-10 + 0.5 \cdot 9 - (-8)] = -6.75$$

Final Q:

$$Q(s_1, a_1) = -4$$
$$Q(s_1, a_2) = -6.75$$
$$Q(s_2, a_1) = 9$$
$$Q(s_2, a_2) = 0$$

Part b For part b, we're using Sarsa that is shown in Eq. 2. Also, $\alpha = 0.5$ and $\gamma = 0.5$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \quad (2)$$

Step 1:

$$Q_0(s_1) = [0, 0]$$
$$Q_0(s_2) = [0, 0]$$
$$Q_1(s_1, a_1) = Q_0(s_1, a_1) + \alpha[R_2 + \gamma Q_0(S_{t+1}, A_{t+1}) - Q_0(s_1, a_1)]$$
$$Q_1(s_1, a_1) = 0 + 0.5[-8 + 0.5 \cdot 0 - 0] = -4$$

Step 2:

$$Q_1(s_1) = [-4, 0]$$
$$Q_1(s_2) = [0, 0]$$
$$Q_2(s_1, a_2) = Q_1(s_1, a_2) + \alpha[R_3 + \gamma Q_1(S_{t+1}, A_{t+1}) - Q_1(s_1, a_2)]$$
$$Q_2(s_1, a_2) = 0 + 0.5[-16 + 0.5 \cdot 0 - 0] = -8$$

Step 3:

$$Q_2(s_1) = [-4, -8]$$
$$Q_2(s_2) = [0, 0]$$
$$Q_3(s_2, a_1) = Q_2(s_2, a_1) + \alpha[R_4 + \gamma Q_2(S_{t+1}, A_{t+1}) - Q_2(s_2, a_1)]$$
$$Q_3(s_1, a_2) = 0 + 0.5[20 + 0.5 \cdot -8 - 0] = 8$$

Step 4:

$$Q_3(s_1) = [-4, -8]$$
$$Q_3(s_2) = [8, 0]$$
$$Q_4(s_1, a_2) = Q_3(s_1, a_2) + \alpha[R_5 + \gamma Q_3(S_{t+1}, A_{t+1}) - Q_3(s_1, a_2)]$$
$$Q_4(s_1, a_2) = -8 + 0.5[-10 + 0.5 \cdot 8 - (-8)] = -7$$

Final Q:

$$Q(s_1, a_1) = -4$$
$$Q(s_1, a_2) = -7$$
$$Q(s_2, a_1) = 8$$
$$Q(s_2, a_2) = 0$$

Part c For part c, we're using Expected Sarsa that is shown in Eq. 3. Also, $\alpha = 0.5$, $\gamma = 0.5$, and $\epsilon = 0.1$. Under an ϵ-greedy policy, the max action has a probability of $1 - \epsilon + \frac{\epsilon}{|A|}$ being taken and all non-max actions have a probability of $\frac{\epsilon}{|A|}$. In this case, the max action has a probability of 0.95 being taken and the non-max action has a probability of 0.05 being taken.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)] \tag{3}$$

Step 1:

$$Q_0(s_1) = [0, 0]$$
$$Q_0(s_2) = [0, 0]$$
$$Q_1(s_1, a_1) = Q_0(s_1, a_1) + \alpha[R_2 + \gamma \sum_a \pi(a|S_{t+1}) Q_0(S_{t+1}, a) - Q_0(s_1, a_1)]$$
$$Q_1(s_1, a_1) = 0 + 0.5[-8 + 0.5 \cdot (0.05 \cdot 0 + 0.95 \cdot 0) - 0] = -4$$

Step 2:

$$Q_1(s_1) = [-4, 0]$$
$$Q_1(s_2) = [0, 0]$$
$$Q_2(s_1, a_2) = Q_1(s_1, a_2) + \alpha[R_3 + \gamma \sum_a \pi(a|S_{t+1}) Q_1(S_{t+1}, a) - Q_1(s_1, a_2)]$$
$$Q_2(s_1, a_2) = 0 + 0.5[-16 + 0.5 \cdot (0.05 \cdot 0 + 0.95 \cdot 0) - 0] = -8$$

Step 3:

$$Q_2(s_1) = [-4, -8]$$
$$Q_2(s_2) = [0, 0]$$
$$Q_3(s_2, a_1) = Q_2(s_2, a_1) + \alpha[R_4 + \gamma \sum_a \pi(a|S_{t+1}) Q_2(S_{t+1}, a) - Q_2(s_2, a_1)]$$
$$Q_3(s_1, a_2) = 0 + 0.5[20 + 0.5 \cdot (0.05 \cdot -8 + 0.95 \cdot -4) - 0] = 8.95$$

Step 4:

$$Q_3(s_1) = [-4, -8]$$
$$Q_3(s_2) = [8.95, 0]$$
$$Q_4(s_1, a_2) = Q_3(s_1, a_2) + \alpha[R_5 + \gamma \sum_a \pi(a|S_{t+1}) Q_3(S_{t+1}, a) - Q_3(s_1, a_2)]$$
$$Q_4(s_1, a_2) = -8 + 0.5[-10 + 0.5 \cdot (0.05 \cdot 0 + 0.95 \cdot 8.95) - (-8)] = -6.874$$

Final Q:

$$Q(s_1, a_1) = -4$$
$$Q(s_1, a_2) = -6.874$$
$$Q(s_2, a_1) = 8.95$$
$$Q(s_2, a_2) = 0$$

2 Problem 2

Part a Given the example from class, where we've already learned how long it takes to get home from our old building and we've recently moved to a new building. TD is able to adjust estimates on-the-fly so that when we take our new route home, we're able to adjust our estimate as we go and are given new information, whereas Monte Carlo requires us to get home before we can update our estimate. While it's not proven that TD is faster, this idea that TD can react to new information more quickly contributes to our intuition that it's faster to converge.

Part b Because TD is using estimates based on state-to-state rewards, in order for it to work properly, the model must be Markovian. TD implicitly computes the estimate of the value function that would be exactly correct if the model were exactly correct. This is called the certainty-equivalence estimate because it is equivalent to assuming that the estimate of the underlying process was known with certainty rather than being approximated and the only way the underlying process can be known with certainty is if it is Markovian.

Part c Both batch methods learn on the same set of training episodes over and over again. It follows that batch MC minimizes the mean-square-error on the training set. Batch TD, on the other hand, always finds the estimates that would be exactly correct for the maximum-likelihood model of the Markov process being modeled.

3 Problem 3

Part a Using the algorithm on page 120, we need to do two things. The first is we need to change the line "A $\leftarrow$ action given by π for S" to instead generate episodes using our behavior policy, b: "A $\leftarrow$ action given by b for S". The second line we need to change is "$V(S) \leftarrow V(S)+\alpha\cdot[R+\gamma V(S')-V(S)]$" to take into account $\rho_{t:t}$ like so: "$V(S) \leftarrow V(S)+\rho_{t:t}\cdot\alpha\cdot[R+\gamma V(S')-V(S)]$".

Part b In Q-Learning, when generating episodes, we choose an action derived from Q, using some policy (say, ϵ-greedy). Then, when we go to update Q, we take the max over all the actions at the next state, $\max_a Q(S', a)$. It

then stands to reason that our episodes are generated using a behavior (ϵ-greedy) policy and our Q values are estimated using a target (greedy) policy.

Part c Using the Double Q-Learning boxed algorithm directly above the question, the only lines that would change are the updates for $Q_1(S, A)$ and $Q_2(S, A)$. For clarity, let's look at changing Q_1. Instead of using the best action from Q_1 to index into Q_2, we want the expected value of Q_2, so we sum over each action the probability we take that action times the Q of it.

$$Q_1(S, A) \leftarrow Q_1(S, A) + \alpha[R + \gamma \sum_{A'} \pi(A'|S) Q_2(S, A') - Q_1(S, A)]$$
$$Q_2(S, A) \leftarrow Q_2(S, A) + \alpha[R + \gamma \sum_{A'} \pi(A'|S) Q_1(S, A') - Q_2(S, A)]$$

4 Problem 4

Table 2: States, Actions, and Rewards for Problem 4

T	S_T	A_T	R_{T+1}
1	(0, 0)	E	0
2	(0, 1)	E	0
3	(0, 2)	N	0
4	(0, 2)	N	0
5	(0, 2)	S	0
6	(1, 2)	S	1
7	(2, 2)		

Part a For part a, we're using 1-step TD that is shown in Eq. 4. Also, $\alpha = 0.1$ and $\gamma = 0.9$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \tag{4}$$

Step 1:

$$Q(0, 0, E) = Q(0, 0, E) + \alpha[R_2 + \gamma Q(0, 1, E) - Q(0, 0, E)]$$
$$Q(0, 0, E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 - 0] = 0$$

Step 2:

$$Q(0,1,E) = Q(0,1,E) + \alpha[R_3 + \gamma Q(0,2,N) - Q(0,1,E)]$$
$$Q(0,1,E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 - 0] = 0$$

Step 3:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_4 + \gamma Q(0,2,N) - Q(0,2,N)]$$
$$Q(0,2,N) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 - 0] = 0$$

Step 4:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_5 + \gamma Q(0,2,S) - Q(0,2,N)]$$
$$Q(0,2,N) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 - 0] = 0$$

Step 5:

$$Q(0,2,S) = Q(0,2,S) + \alpha[R_6 + \gamma Q(1,2,S) - Q(0,2,S)]$$
$$Q(0,2,S) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 - 0] = 0$$

Step 6:

$$Q(1,2,S) = Q(1,2,S) + \alpha[R_7 + \gamma Q(2,2,:) - Q(1,2,S)]$$
$$Q(1,2,S) = 0 + 0.1 \cdot [1 + 0.9 \cdot 0 - 0] = 0.1$$

Part b For part b, we're using 2-step TD that is shown in Eq. 5. Also, $\alpha = 0.1$ and $\gamma = 0.9$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2}, A_{t+2}) - Q(S_t, A_t)] \quad (5)$$

Step 1:

$$Q(0,0,E) = Q(0,0,E) + \alpha[R_2 + \gamma R_3 + \gamma^2 Q(0,2,N) - Q(0,0,E)]$$
$$Q(0,0,E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 - 0] = 0$$

Step 2:

$$Q(0,1,E) = Q(0,1,E) + \alpha[R_3 + \gamma R_4 + \gamma^2 Q(0,2,N) - Q(0,1,E)]$$
$$Q(0,1,E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 - 0] = 0$$

Step 3:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_4 + \gamma R_5 + \gamma^2 Q(0,2,S) - Q(0,2,N)]$$
$$Q(0,2,N) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 - 0] = 0$$

Step 4:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_5 + \gamma R_6 + \gamma^2 Q(1,2,S) - Q(0,2,N)]$$
$$Q(0,2,N) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 - 0] = 0$$

Step 5:

$$Q(0,2,S) = Q(0,2,S) + \alpha[R_6 + \gamma R_7 + \gamma^2 Q(2,2,S) - Q(0,2,S)]$$
$$Q(0,2,S) = 0 + 0.1 \cdot [0 + 0.9 \cdot 1 + 0.9^2 \cdot 0 - 0] = 0.09$$

Step 6:

$$Q(1,2,S) = Q(1,2,S) + \alpha[R_7 + \gamma Q(2,2,:) - Q(1,2,S)]$$
$$Q(1,2,S) = 0 + 0.1 \cdot [1 + 0.9 \cdot 0 - 0] = 0.1$$

Part c For part c, we're using 4-step TD that is shown in Eq. 6. Also, $\alpha = 0.1$ and $\gamma = 0.9$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \gamma^4 Q(S_{t+4}, A_{t+4}) - Q(S_t, A_t)] \tag{6}$$

Step 1:

$$Q(0,0,E) = Q(0,0,E) + \alpha[R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 + \gamma^4 Q(0,2,S) - Q(0,0,E)]$$
$$Q(0,0,E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 + 0.9^3 \cdot 0 + 0.9^4 \cdot 0 - 0] = 0$$

Step 2:

$$Q(0,1,E) = Q(0,1,E) + \alpha[R_3 + \gamma R_4 + \gamma^2 R_5 + \gamma^3 R_6 + \gamma^4 Q(1,2,S) - Q(0,1,E)]$$
$$Q(0,1,E) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 + 0.9^3 \cdot 0 + 0.9^4 \cdot 0 - 0] = 0$$

Step 3:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_4 + \gamma R_5 + \gamma^2 R_6 + \gamma^3 R_7 + \gamma^4 Q(2,2,:) - Q(0,2,N)]$$
$$Q(0,2,N) = 0 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 0 + 0.9^3 \cdot 1 + 0.9^4 \cdot 0 - 0] = 0.0729$$

Step 4:

$$Q(0,2,N) = Q(0,2,N) + \alpha[R_5 + \gamma R_6 + \gamma^2 R_7 + \gamma^3 Q(2,2,:) - Q(0,2,N)]$$
$$Q(0,2,N) = 0.0729 + 0.1 \cdot [0 + 0.9 \cdot 0 + 0.9^2 \cdot 1 + 0.9^3 \cdot 0 - 0.0729] = 0.14661$$

Step 5:

$$Q(0,2,S) = Q(0,2,S) + \alpha[R_6 + \gamma R_7 + \gamma^2 Q(2,2,:) - Q(0,2,S)]$$
$$Q(0,2,S) = 0 + 0.1 \cdot [0 + 0.9 \cdot 1 + 0.9^2 \cdot 0 - 0] = 0.09$$

Step 6:

$$Q(1,2,S) = Q(1,2,S) + \alpha[R_7 + \gamma Q(2,2,:) - Q(1,2,S)]$$
$$Q(1,2,S) = 0 + 0.1 \cdot [1 + 0.9 \cdot 0 - 0] = 0.1$$