Homework 4

Jerry Duncan

October 15, 2020

1 Problem 1

Part a In Table 1, I've taken the episode data we're given and converted it into table form so we can see state, actions, and rewards over time. From those we can then update our value function like so, where $\gamma = 1$ because it's episodic:

$$\begin{gathered} V(S_T) = G_T = R_{T+1} + \gamma G_{T+1} = R_{T+1} + R_{T+2} \dots \\ V(8, 19, 0) = -1 \\ V(8, 15, 0) = 0 + -1 = -1 \\ V(8, 14, 0) = 0 + 0 + -1 = -1 \\ V(10, 15, 0) = -1 \\ V(8, 20, 0) = 1 \end{gathered}$$

Table 1: Part 1a in tabular form

		1			2			3	
T	S_T	A_T	R_{T+1}	S_T	A_T	R_{T+1}	S_T	A_T	R_{T+1}
1	(8, 14, 0)	H	0	(10, 15, 0)	H	-1	(8, 20, 0)	S	1
2	(8, 15, 0)	H	0	T			T		
3	(8, 19, 0)	H	-1						
4	T								

Part b When we use DP methods, we have access to the model's dynamics function, $p(s', r|s, a)$. When we use MC methods, we often do not have access to it. In order to calculate the next policy, we either need the dynamics function and state value function or just the action-value function. Because we don't have access to the dynamics function when using MC methods, we instead calculate the action-value function so that we can use it to update our policy. This can be seen in the following functions:

$$\pi'(s) = \underset{a}{\text{argmax}} \sum_{s',r} p(s', r|s, a)[r + \gamma V_\pi(s')]$$
$$\pi'(s) = \underset{a}{\text{argmax}} Q_\pi(s, a)$$

Part c All we need to do is use Q_π instead of V_π.

$$Q(S_T, A_T) = G_T = R_{T+1} + \gamma G_{T+1} = R_{T+1} + R_{T+2} \dots$$
$$Q(8, 19, 0, H) = -1$$
$$Q(8, 15, 0, H) = 0 + -1 = -1$$
$$Q(8, 14, 0, H) = 0 + 0 + -1 = -1$$
$$Q(10, 15, 0, H) = -1$$
$$Q(8, 20, 0, S) = 1$$

2 Problem 2

Part a When using an off-policy approach, the estimation policy can be deterministic because our behavior policy can be stochastic and explore all actions and states for us. This allows our off-policy approach to find an optimal policy while using an on-policy ϵ-greedy approach can only find a near-optimal policy, due to it taking a non-greedy action part of the time.

Part b We often want to find the optimal policy for a given problem. The issue is that in order to find the optimal policy, we need to explore the environment but we'd like our final policy to be greedy. These two things are at odds, so how can we reconcile them? Through off-policy prediction — that is, we have a behavior policy that generates episodes from the environment and a target policy that we are using to find the optimal policy for the problem. In order to use the behavior policy's episodes to evaluate our target policy, we need to use importance sampling. Importance sampling is a technique for estimating expected values under one distribution (the target policy) given samples from another distribution (the behavior policy). It allows us to evaluate our target policy by weighting the returns from the behavior policy's episodes according to the relative probability of their trajectories occurring under the target policy. This is called the importance-sampling ratio, given by $\rho_{t:T(t)-1} = \Pi_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$. When we scale the returns by these ratios and do a simple average of the results, it is called ordinary importance sampling and is given by the following formula:

$$V(s) = \frac{\sum_{t \in \tau(s)} \rho_{t:T(t)-1 \cdot G_t}}{|\tau(s)|}$$

where $\tau(s)$ is the time steps at which we visited s and $|\tau(s)|$ is the total number of times we visited s. We can use this formula to accurately update our state value function while evaluating the target policy.

Part c All the calculations are contained in Table 2. This table shows states, rewards, returns, π, b, ρ, $V(S_T)_{unweighted}$, and $V(S_T)_{weighted}$ values for each step in the episode. To specifically answer the question of which entries in the state value function will change, they're given below and include the first and second visit to states to be extra clear about all changes to V:

$$\text{unweighted}$$
$$V(2,0) = \frac{0 \cdot 9}{1} = 0 \text{ first visit}$$
$$V(2,0) = \frac{0 \cdot 9 + 32 \cdot 10}{2} = 160 \text{ second visit}$$
$$V(2,1) = \frac{16 \cdot 10}{1} = 160$$
$$V(1,1) = \frac{4 \cdot 10}{1} = 40$$
$$\text{weighted}$$
$$V(2,0) = \frac{0 \cdot 9}{0} = 0 \text{ first visit}$$
$$V(2,0) = \frac{0 \cdot 9 + 32 \cdot 10}{0 + 32} = 10 \text{ second visit}$$
$$V(2,1) = \frac{16 \cdot 10}{16} = 10$$
$$V(1,1) = \frac{4 \cdot 10}{4} = 10$$

Table 2: Tabularized 2c

T	0	1	2	3	4
S_T	(2, 0)	(2, 0)	(2, 1)	(1, 1)	T
A_T	W	E	N	N	
R_{T+1}	-1	0	0	10	
G_T	9	10	10	10	
$\pi(A_T \vert S_T)$	0	0.5	1	1	
$b(A_T \vert S_T)$	0.25	0.25	0.25	0.25	
$\rho_{t:T(t)-1}$	$\frac{0 \cdot 0.5 \cdot 1 \cdot 1}{0.25^4} = 0$	$\frac{0.5 \cdot 1 \cdot 1}{0.25^3} = 32$	$\frac{1 \cdot 1}{0.25^2} = 16$	$\frac{1}{0.25} = 4$	
$V(S_T)_{unweighted}$	$\frac{0 \cdot 9}{1} = 0$	$\frac{0 \cdot 9 + 32 \cdot 10}{2} = 160$	$\frac{16 \cdot 10}{1} = 160$	$\frac{4 \cdot 10}{1} = 40$	
$V(S_T)_{weighted}$	$\frac{0 \cdot 9}{0} = 0$	$\frac{0 \cdot 9 + 32 \cdot 10}{0 + 32} = 10$	$\frac{16 \cdot 10}{16} = 10$	$\frac{4 \cdot 10}{4} = 10$	

3 Problem 3

Part a For this problem I've converted the board state into coordinates. Top left is (0, 0) and bottom right is (1, 2). I've converted the MC state, action, and reward chain to tabular form in Table 3. I've also calculated G_T going backwards, like in class.

Table 3: Part 3a in tabular form

T	S_T	A_T	R_{T+1}	G_T
0	(0, 0)	E	0	$0 + 0.5 \cdot 5 = 2.5$
1	(0, 1)	S	0	$0 + 0.5 \cdot 10 = 5$
2	(1, 1)	E	0	$0 + 0.5 \cdot 20 = 10$
3	(1, 2)	N	20	20
4	T			

Using the G_T values calculated in Table 3, we can update our Q_π values.

$$\begin{aligned}
Q(S_T, A_T) &= Q(S_T, A_T) + \frac{1}{n}[G_T - Q(S_T, A_T)] \\
Q(0, 0, E) &= 8 + \frac{1}{10}[2.5 - 8] = 7.45 \\
Q(0, 1, S) &= 4 + \frac{1}{10}[5 - 4] = 4.1 \\
Q(1, 1, E) &= 10 + \frac{1}{10}[10 - 10] = 10 \\
Q(1, 2, N) &= 20 + \frac{1}{10}[20 - 20] = 20
\end{aligned}$$

Part b When using ordinary importance sampling, all we need to do is change our Q update function and calculate the appropriate ρ values. In Table 4 we can see the addition of π, b, and ρ for each timestep. Using those values, we can update our Q_π values like so:

$$Q(S_T, A_T) = Q(S_T, A_T) + \frac{1}{n}[\rho G_T - Q(S_T, A_T)]$$
$$Q(0, 0, E) = 8 + \frac{1}{10}[0 \cdot 2.5 - 8] = 7.2$$
$$Q(0, 1, S) = 4 + \frac{1}{10}[0 \cdot 5 - 4] = 3.6$$
$$Q(1, 1, E) = 10 + \frac{1}{10}[2 \cdot 10 - 10] = 11$$
$$Q(1, 2, N) = 20 + \frac{1}{10}[1 \cdot 20 - 20] = 20$$

Table 4: Tabularized 3b

T	0	1	2	3	4
S_T	(0, 0)	(0, 1)	(1, 1)	(1, 2)	T
A_T	E	S	E	N	
R_{T+1}	0	0	0	20	
G_T	$0 + 0.5 \cdot 5 = 2.5$	$0 + 0.5 \cdot 10 = 5$	$0 + 0.5 \cdot 20 = 10$	20	
$\pi(A_T \| S_T)$	1	0	1	1	
$b(A_T \| S_T)$	0.5	0.5	0.5	1	
$\rho_{t:T(t)-1}$	$\frac{1 \cdot 0 \cdot 1 \cdot 1}{0.5^3 \cdot 1} = 0$	$\frac{0 \cdot 1 \cdot 1}{0.5^2 \cdot 1} = 0$	$\frac{1 \cdot 1}{0.5 \cdot 1} = 2$	$\frac{1}{1} = 1$	

4 Problem 4

Part a Under Initialize, add a line that says: "$N(s) \leftarrow 0$ for all $s \in S$". In the looping over each step of the episode, under "Unless S_t appears in ...", replace "Append G to $Returns(S_t)$" with "$N(s) \leftarrow N(s) + 1$" and replace "$V(S_t) \leftarrow$ average($Returns(S_t)$)" with "$V(S_t) \leftarrow V(S_t) + \frac{1}{N(s)}[G - V(S_t)]$".

Part b The first thing to note is that in the boxed algorithm for off-policy MC control, $\pi(S_T)$ is deterministic. This means that it is 1 for one action and 0 for all other actions. The boxed algorithm implements a check for this where it exits the inner loop if $A_T \neq \pi(S_T)$ because $\pi(A_T|S_T)$ would be zero and all previous actions before A_T would have a weight of 0 and not matter (since we're going through the episode backwards). If $A_T = \pi(S_T)$

then $\pi(A_T|S_T) = 1$. This is why in the update for W, we see $\frac{1}{b(A_T|S_T)}$. In summary, we short circuit the loop if the action taken by b and by π aren't the same because it would cause W to be zero for every action before A_T. We use a coefficient of $\frac{1}{b(A_T|S_T)}$ because the only way we reach that line of the algorithm is if $\pi(A_T|S_T) = 1$.