CS6700 - Reinforcement Learning - Written Assignment 2

Atul Balaji

March 18, 2019

1 Question 1

1.1 (a)

1.1.1 MC

MC estimates are the average of rewards observed from a particular state from all episodes till termination.

For state A:

$$V(A) = \frac{1+1+2}{3} = 1.33 \tag{1}$$

For state B:

$$V(B) = \frac{1+1+1}{5} = 0.6 \tag{2}$$

For state C:

$$V(C) = \frac{1+1+1+1}{5} = 0.8 \tag{3}$$

1.1.2 TD(0)

For state B:

$$V(B) = \frac{1+1+1}{5} = 0.6 \tag{4}$$

For state C:

$$V(C) = \frac{1+1+1+1}{5} = 0.8 \tag{5}$$

For state A:

$$V(A) = \frac{(0+V(B))+(0+V(C))+(1+V(C))}{3} = 1.067 \tag{6}$$

1.2 (b)

The transition probabilities can be estimated based on the possible transitions which are $A \to B$, $A \to C$, $B \to T$ and $C \to T$ where T is the terminal state.
The transition probabilities can be found as:

$$P(S_{t+1} = j \,|\, S_t = i) = \frac{n(S_{t+1} = j \,|\, S_t = i)}{n(S_t = i)} \tag{7}$$

So, we get:

Transition	Probability
$A \to B$	0.33
$A \to C$	0.67
$B \to T$	1
$C \to T$	1

We can build a matrix for rewards as follows:

	B	C	T
A	0	0.5	0
B	0	0	0.6
C	0	0	0.8

where R_{ij} denotes the average reward obtained for transition $i \to j$. Note that the transitions $B \to A$ and $C \to A$ have not been shown since they are not seen in the episodes.

1.3 (c)

The required MSE is the sum of the Mean-square errors of the three states.

1.3.1 MC

Substituting the estimates V(A), V(B) and V(C) in the equation given the question for MSE, we obtain

$$MSE = \frac{1}{3}(0.66) + \frac{1}{5}(0.912) + \frac{1}{5}(0.8) = 0.5624$$

1.3.2 TD(0)

Substituting the estimates V(A), V(B) and V(C), we obtain:

$$MSE = \frac{1}{3}(2.28) + \frac{1}{5}(0.912) + \frac{1}{5}(0.8) = 1.1024$$

Based on the MSE, Monte Carlo is truer to the training data compared to TD(0).

1.4 (d)

The TD(0) method maximizes the likelihood of the observations being obtained from the MDP. The value estimate obtained in this manner is called the certainty equivalence estimate. Thus it relies on the Markov assumption and is therefore true to the model in (b).

1.5 (e)

When the Markov assumption for the data is true, TD(0) performs better due to its assumption of Markov nature. But in a non-Markov setting, Monte Carlo method is preferred as it does not make such assumptions about the model.

2 Question 2

State is observed at time t. The action is applied to the system at time $(t + \tau)$. So, we have to update the action-value (Q) for $\hat{a_t} = a_{t-\tau}$ that actually takes place in s_t, rather than for $Q(s_t, a_t)$.

2.1 (a)

For this task, the return can be written as:

$$G_t = R_{t+\tau+1} + \gamma R_{t+\tau+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+\tau+k+1} \tag{8}$$

2.2 (b)

The TD(0) backup equation is:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+\tau+1} + \gamma V(S_{t+\tau+1}) - V(S_t)] \tag{9}$$

3 Question 3

3.1 (a)

Truncating the eligibility trace after 3 time steps results in a variant of G_t^λ.
In order to compute the updates for $V(s_1)$ we must consider the states until which the eligibility trace goes to zero. Starting from s_1, we get error terms involving s_1, s_2, s_3, s_4. This can be seen in the table below.
s_1: $e^{(t)} = [0, 1, (\gamma\lambda), (\gamma\lambda)^2, (\gamma\lambda)^3, 0]$
s_2: $e^{(t)} = [0, 0, 1, (\gamma\lambda), (\gamma\lambda)^2, (\gamma\lambda)^3]$
s_3: $e^{(t)} = [0, 0, 0, 1, (\gamma\lambda), (\gamma\lambda)^2]$
s_4: $e^{(t)} = [0, 0, 0, 0, 1, (\gamma\lambda)]$

s_5: $e^{(t)} = [0, 0, 0, 0, 0, 1]$

In general, the update equation for the value estimate $V(S_t)$ is:

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t e^{(t)} \tag{10}$$

where $e^{(t)}$ is the vector of eligibility traces at time t and δ_t is the error in $V(S_{t+1})$.

Now for $V(s_1)$, the update is as follows.

$$\begin{aligned}V(s_1) \leftarrow V(s_1) + \alpha\{R_1 + \gamma V(s_2) - V(s_1) + \gamma\lambda\{R_2 + \gamma V(s_3) - V(s_2)\} + (\gamma\lambda)^2\{R_3 + \gamma V(s_4) - V(s_3)\}\\ +(\gamma\lambda)^3\{R_4 + \gamma V(s_5) - V(s_4)\}\}\end{aligned}$$

We can write the following identities:

$$R_1 = R_1(1-\lambda) + R_1(1-\lambda)\lambda + R_1(1-\lambda)\lambda^2 + R_1(1-\lambda)\lambda^3 + R_1\lambda^4 \tag{11}$$

$$R_2 = R_2(1-\lambda) + R_2(1-\lambda)\lambda + R_2(1-\lambda)\lambda^2 + R_2\lambda^3 \tag{12}$$

$$R_3 = R_3(1-\lambda) + R_3(1-\lambda)\lambda + R_3\lambda^2 \tag{13}$$

The expression for the value function estimates can be therefore be written as given below.

$$\begin{aligned}V(s_1) \leftarrow V(s_1) + \alpha\{(1-\lambda)\{\{\lambda^0\{R_1 + \gamma V(s_2)\} + \lambda^1\{R_1 + \gamma R_2 + \gamma^2 V(s_3)\} + \lambda^2\{R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 V(s_4)\} +\\ \lambda^3\{R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 V(s_5)\}\} + \lambda^4\{R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 V(s_5)\}\} - V(s_1)\}\end{aligned}$$

Let the required return be denoted as G^λ.

$$\implies G^\lambda = \sum_{i=1}^{4} \lambda^{i-1} G^{(i)} + \lambda^4 G^{(4)} \tag{14}$$

where $G^{(n)}$ is the n-step TD return.
This is the required λ-return that we are optimizing.

3.2 (b)

Based on part (a) where we truncate the traces after 3 time steps, we can generalize it to truncation after n time steps.

$$G^\lambda = \sum_{i=1}^{n+1} \lambda^{i-1} G^{(i)} + \lambda^{n+1} G^{(n+1)} \tag{15}$$

4 Question 4

TD methods rely on the Markov assumption. They maximize the likelihood of the of the data being obtained from the MDP. Let's consider the case of estimating the value function. The value function of a state s is

$$V^\pi(s) = E_\pi[r_{t+1} + \gamma E_\pi[r_{t+1} + \gamma r_{t+2} \dots | S_t = s, S_{t+1} = s'] | S_t = s]$$

Using Markov assumption, this reduces to

$$E_\pi[r_{t+1} + \gamma E_\pi[r_{t+1} + \gamma r_{t+2} \dots | S_t = s, S_{t+1} = s'] | S_t = s] = E_\pi[r_{t+1} + \gamma V^\pi(s') | S_t = s]$$

which is the Bellman equation, which is used in the TD update. But when the Markov assumption is not true, the future rewards depend on the past states in addition to the current state. This renders the above equations invalid. Therefore TD Learning is not suitable in non-Markov systems.
Rather than TD methods, it is preferred to use eligibility traces, since they bring out the dependence on many previous states.

5 Question 5

In the given grid-world, we have 6 states out of which T1 and T2 are terminal states, S is the start state.

		S		
T2	*	S1	S2	T1

There is a reward of +5 units when $T2$ is reached, +10 units when $T1$ is reached and any transition to $*$ results in a reward of a units.

The following actions are possible from each state:

S: down

***:** left and right

S1: left and right

S2: left and right

From state $S1$, the up action is not taken because there is no reward obtained and even from S, while going back to S1, the reward is zero. Effectively no future rewards are accumulated in these two steps.

Therefore there are totally 8 policies possible.

A particular policy will be optimal if the Bellman optimality equation is satisfied. The policies have been drawn below.

5.1 Transition Diagrams

1		*down*		
T2	*left*	*left*	*left*	T1

2		*down*		
T2	*left*	*left*	*right*	T1

3		*down*		
T2	*left*	*right*	*left*	T1

4		*down*		
T2	*left*	*right*	*right*	T1

5		*down*		
T2	*right*	*left*	*left*	T1

6		*down*		
T2	*right*	*left*	*right*	T1

7		*down*		
T2	*right*	*right*	*left*	T1

8		*down*		
T2	*right*	*right*	*right*	T1

5.2 Value functions

Policy 1

State	Value
S	$a\gamma + 5\gamma^2$
$S2$	$a\gamma + 5\gamma^2$
$S1$	$a + 5\gamma$
$*$	5

Policy 2

State	Value
S	$a\gamma + 5\gamma^2$
$S2$	10
$S1$	$a + 5\gamma$
$*$	5

Policy 3

State	Value
S	0
$S2$	0
$S1$	0
$*$	5

Policy 4

State	Value
S	$10\gamma^2$
$S2$	10
$S1$	10γ
$*$	5

Policy 5

State	Value
S	$\frac{a\gamma}{1-\gamma^2}$
$S2$	$\frac{a\gamma}{1-\gamma^2}$
$S1$	$\frac{a}{1-\gamma^2}$
$*$	$\frac{a\gamma}{1-\gamma^2}$

Policy 6

State	Value
S	$\frac{a\gamma}{1-\gamma^2}$
$S2$	10
$S1$	$\frac{a}{1-\gamma^2}$
$*$	$\frac{a\gamma}{1-\gamma^2}$

Policy 7

State	Value
S	0
$S2$	0
$S1$	0
$*$	0

Policy 8

State	Value
S	$10\gamma^2$
$S2$	10
$S1$	10γ
$*$	$10\gamma^2$

For optimality of policy π_i, $V_{\pi_i}(s) \geq max\, V_{\pi_j}(s)$ for $j \neq i$, all states s.

- From the value functions, we can say that policies 1,2,3,5,6 and 7 are not optimal for any value of γ.
- Policies 4 and 8 are optimal in two cases: **a)** $a < 0$, $\gamma > 0$; **b)** $a = 0$, $\gamma > \frac{1}{\sqrt{2}}$, in which case k = $\frac{1}{\sqrt{2}}$

Therefore, for different ranges of a, the characterization of Blackwell optimality is:
1. $a = 0 \implies k = \frac{1}{\sqrt{2}}$
2. $a < 0 \implies k = 0$

6 Question 6

The dynamics of the problem is periodic and changes every K steps. The cycle is repeated every MK steps, which means there are M separate Markov Processes (M representations are maintained), each of which run for K steps after which the process is switched. This results in the problem being Markov.
A value function based solution will work with this representation if we maintain M such value functions to be used. Value function estimates are maintained for states of each of the M Markov models.

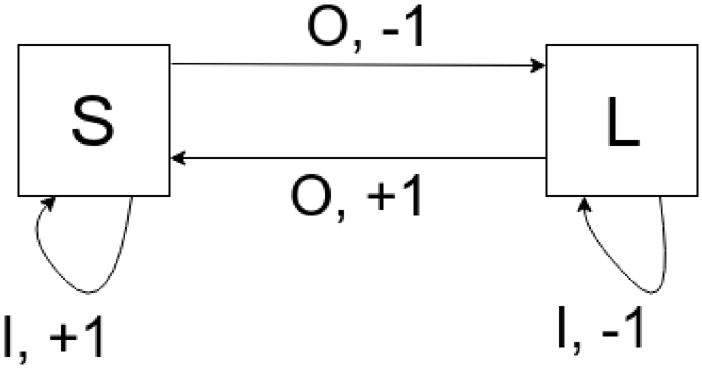
7 Question 7

Q-learning is an off-policy method in which random exploratory policies are used to generate episodes and using this, the optimal policy is learned.
It is possible to make Q-learning on-policy. The on-policy version of Q-learning is nothing but SARSA with greedy policy as the exploring policy.
SARSA with greedy policy might not converge to optimal policy since it does not explore sufficiently.
An optimal deterministic policy should exist. Therefore if we choose this as the behaviour policy, the exploration would be insufficient and weights become unbounded. So it is not possible to learn arbitrary policy value functions by following optimal policies.

8 Question 8

8.1 (a)

State-space $S = \{laughter\,(\,L\,),\ silent\,(\,S\,)\}$.
Actions $A = \{playing\ organ(O), lighting\ incense(I)\}$.
Discount factor $\gamma = 0.9$.
The **Reward function** is $R(S,O,L) = -1, R(L,O,S) = 1, R(S,I,S) = 1, R(L,I,L) = -1$ for the different transitions.



8.2 (b)

8.2.1 Policy iteration

- According to the question, we assume an initial estimate of policy π such that $\pi(laughing) = \pi(silent) = I \implies \pi_0 = \{L : I,\ S : I\}$.
- The transition probabilities can be written in matrix form as (with first row and column being L and second being S): $p_{\pi_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, r_{\pi_0} = \begin{bmatrix} -1 \\ +1 \end{bmatrix}, V_{\pi_0} = (I - \gamma p_{\pi_0})^{-1} r_{\pi_0} = \begin{bmatrix} -10 \\ 10 \end{bmatrix}$
- Now, the improved policy can be written as:

$$\pi_1(L) = \underset{a}{argmax}\{O:\ 1 + 0.9(10), I : -1 + 0.9(-10)\} = \underset{a}{argmax}\{O : 10, I : -10\} = O$$

$$\pi_1(S) = \underset{a}{argmax}\{O:\ -1 + 0.9(-10), I : 1 + 0.9(10)\ \ = \underset{a}{argmax}\{O : -10, I : 10\} = I$$

$\implies \pi_1 = \{L : O,\ S : I\}$.

- The second iteration of policy iteration can be written as: $p_{\pi_1} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, r_{\pi_1} = \begin{bmatrix} +1 \\ +1 \end{bmatrix}, V_{\pi_1} = (I - \gamma p_{\pi_1})^{-1} r_{\pi_1} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$

- So, the improved policy is:

$$\pi_2(L) = \underset{a}{argmax}\{O:\ 1 + 0.9(10), I: -1 + 0.9(10)\} = \underset{a}{argmax}\{O: 10, I: 8\} = O$$

$$\pi_2(S) = \underset{a}{argmax}\{O:\ -1 + 0.9(10), I: 1 + 0.9(10)\} = \underset{a}{argmax}\{O: 8, I: 10\} = I$$

$\implies \pi_2 = \{L: O,\ S: I\}$.

- $\pi_1 = \pi_2$. Therefore, policy iteration converged to the optimal policy $\pi^* = \{L: O,\ S: I\}$.
- The optimal value function is $V^* = V_{\pi_1} = \{L: 10, S: 10\}$.

8.2.2 Value iteration

- Let us assume an initial estimate of the value function, $V_0 = \{L: 0, S: 0\}$. The value iteration formula is:

$$v_{k+1}(s) = \underset{a}{max} E[R_{t+1} + \gamma v_k(S_{t+1})] \tag{16}$$

- $V_1(L) = \underset{a}{max}\{O: 1 + 0.9(0), I: -1 + 0.9(0)\} = 1$.
  $V_1(S) = \underset{a}{max}\{O: -1 + 0.9(0), I: 1 + 0.9(0)\} = 1$.
- $V_2(L) = \underset{a}{max}\{O: 1 + 0.9(1), I: -1 + 0.9(1)\} = 1.9$.
  $V_2(S) = \underset{a}{max}\{O: -1 + 0.9(1), I: 1 + 0.9(1)\} = 1.9$.
- $V_3(L) = \underset{a}{max}\{O: 1 + 0.9(1.9), I: -1 + 0.9(1.9)\} = 2.71$.
  $V_3(S) = \underset{a}{max}\{O: -1 + 0.9(1.9), I: 1 + 0.9(1.9)\} = 2.71$.
- $V_4(L) = \underset{a}{max}\{O: 1 + 0.9(2.71), I: -1 + 0.9(2.71)\} = 3.439$.
  $V_4(S) = \underset{a}{max}\{O: -1 + 0.9(2.71), I: 1 + 0.9(2.71)\} = 3.439$.
- The optimal value function V^* converges to $\{L: 10, S: 10\}$, since it is the infinite sum of a GP with first term 1 and common ratio 0.9.

8.3 (c)

Now we know that $V^* = \{L: 10, S: 10\}$. Thus, the optimal action-value estimates $Q^*(s, a)$ can be found using V^*:
$Q^*(L, O) = 1 + 0.9V^*(S) = 10$
$Q^*(L, I) = -1 + 0.9V^*(S) = 8$
$Q^*(S, O) = -1 + 0.9V^*(L) = 8$
$Q^*(S, I) = 1 + 0.9V^*(L) = 10$

8.4 (d)

The optimal policy must be followed as it ensures that the house remains silent. At the start, the laughing sound is heard, therefore the organ should be played, after which the house becomes silent. Then, the organ should be stopped and incense must be lighted indefinitely. This will ensure that the house will forever be silent.