MDL Assignment 3 Part B

Stella Sravanthi:2019101101
Anvita Reddy: 2019115009

We have used the roll number **2019115009** for all the calculations. We have formulated the POMDP and policies based on the given problem statement and used the SARSOP to find the optimal policy, where

1.The probability of moving in desired direction

x=1 - (((last four digits of roll number) % 30) + 1)/100
=1-(((5009)%30)+1)/100
=0.7

2.Reward for reaching the target before the call is turned off is given by (RollNumber%90 + 10)

=(2019115009%90)+10
=19+10
=29

In the POMDP, each state in represented in the notation
s_**Agent** Row-**Agent** Column_**Target** Row-**Target** Column_**Call**
And call is 0 when off and1 when on

For example:

A			
			T

And when Call is **off**, the state would be represented as s_0-0_1-3_0

For all of the pomdps and policies generated the representation of cell is given by

0 **(0,0)**	**1** **(0,1)**	**2** **(0,2)**	**3** **(0,3)**
4 **(1,0)**	**5** **(1,1)**	**6** **(1,2)**	**7** **(1,3)**

Question 1:

Given Target:(1,0)

Observation is o6: the target is not in the cell neighbourhood of the agent.

The possible positions in the case are 5 cells with equal probabilities:

(0, 1), (0, 2), (0, 3), (1, 2), (1, 3)

	A	**A**	**A**
T		**A**	**A**

Call may be on or off. (2 possibilities). Hence, the 5*2 = 10 state tuples having initial probabilities of 1/10=0.1 each and hence have the same belief state. While the rest of the states have belief states 0.

State tuples are:

1. ((0, 1), (1, 0), 0))
2. ((0, 1), (1, 0), 1))
3. ((0, 2), (1, 0), 0))
4. ((0, 2), (1, 0), 1))
5. ((0, 3), (1, 0), 0))
6. ((0, 3), (1, 0), 1))
7. ((1, 2), (1, 0), 0))
8. ((1, 2), (1, 0), 1))
9. ((1, 3), (1, 0), 0))
10. ((1, 3), (1, 0), 1))

The policy file for this question is attached as well.

Question 2:

Agent:(1,1)

Target: In neighbourhood of the agent

The possible target positions in the case are 4 cells with equal probabilities:

(1, 1), (0, 1), (1, 2), (1, 0). Here we included Agent and target in the same cell, since they are also in the neighbourhood.

	T		
T	**A/T**	**T**	

Given call is off. Hence only one possibility.

Hence, the 4*1 = 4 state tuples having initial probabilities of ¼=0.25 each.

State tuples are:

1. ((1, 1), (1, 1), 0)
2. ((1, 1), (1, 2), 0)
3. ((1, 1), (1, 0), 0)
4. ((1, 1), (0, 1), 0)

Question 3:

To get the expected rewards, we run the simulation using the policy with the following command:

For Question:1

./pomdpsim --simLen 100 --simNum 1000 --policy-file q1.policy q1.pomdp

```
Simulating ...
  action selection :  one-step look ahead

-----------------------------------
 #Simulations  | Exp Total Reward
-----------------------------------
 100            3.94684
 200            3.97511
 300            4.05981
 400            3.88819
 500            3.79712
 600            3.71463
 700            3.66255
 800            3.63119
 900            3.54862
 1000           3.50346
-----------------------------------

Finishing ...

-------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
-------------------------------------------------------------
 1000           3.50346            (3.23371, 3.77322)
-------------------------------------------------------------
```

The Expected Total reward came to be **3.50346**

For Question:2

./pomdpsim --simLen 100 --simNum 1000 --policy-file q2.policy q2.pomdp

```
Simulating ...
  action selection :  one-step look ahead

-----------------------------------
 #Simulations  | Exp Total Reward
-----------------------------------
 100             40.8707
 200             40.4097
 300             40.2569
 400             39.4455
 500             39.5031
 600             39.1152
 700             38.6957
 800             38.2785
 900             38.2771
 1000            38.4359
-----------------------------------

Finishing ...

-----------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
-----------------------------------------------------------------
 1000            38.4359            (37.5625, 39.3093)
-----------------------------------------------------------------
```

The Expected Total reward came to be **38.4359**

Question 4:
Probability(Agent in (0, 0)) = 0.4
Probability(Agent in (1, 3)) = 0.6
Target -> (0,1), (0,2), (1,1), (1,2) with equal probability, i.e 0.25

A(0.4)	**T**	**T**	
	T	**T**	**A(0.6)**

The possible Combinations are:

Agent	Target	Observation	Probability
(0,0)	(0,1)	o2	0.1(0.4*0.25)
(0,0)	(0,2)	o6	0.1
(0,0)	(1,1)	o6	0.1
(0,0)	(1,2)	o6	0.1
(1,3)	(0,1)	o6	0.15 (0.6*0.25)
(1,3)	(0,2)	o6	0.15
(1,3)	(1,1)	o6	0.15
(1,3)	(1,2)	o4	0.15

Here we see that there are only three observations o2,o4,o6. Now we can calculate the probability of each of these observations as a sum of the individual probabilities of them.

$$P\ (o2) = 0.1$$
$$P\ (o4) = 0.15$$
$$P\ (o6) = 0.75$$

These are the correct probabilities as attested by the fact that

$$P\ (o2) + P\ (o4) + P\ (o6) = 1\ .$$

Hence we can say that the most probable observation is o6, with probability 0.75

Question 5:

Given,

- the number of nodes in the tree, N
- the height of the tree (i.e., horizon of the POMDP), T
- the number of observation (here, equal to 6), |O|
- the number of actions (here, equal to 5), |A|

Where total number of nodes in a policy tree is given by

$N = \Sigma_i\ (|O|^T – 1) / (|O| - 1)$, here i range from (0,T-1)

$= \Sigma_i\ (6^T – 1) / (6-1)$

The number of policy trees, $P = |A|^N = 5^N$

Upon using the command **./pomdpsol q4.pomdp**,
for the pomdp generated in Question 4, we get

```
SARSOP initializing ...
  initialization time : 0.01s

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.01    0       0        10.7141    31.5304    20.8164     5        1
 0.01    10      51       24.6268    24.7462    0.11942     24       14
 0.01    16      101      24.7194    24.734     0.014654    34       22
 0.02    20      151      24.7223    24.7315    0.00920983  48       36
 0.03    26      205      24.7277    24.7311    0.00340053  75       48
 0.03    30      250      24.729     24.731     0.00207344  83       55
 0.04    34      301      24.7292    24.7308    0.00154891  102      69
 0.06    38      359      24.7296    24.7307    0.00111886  130      82
 0.06    41      395      24.7296    24.7306    0.000997578 160      93
-------------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000998

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.07    41      395      24.7296    24.7306    0.000997578 160      93
-------------------------------------------------------------------------------
```

Here we get the #Trial or the time horizon T=41

So, the number of nodes are:

$$N=(6^{41}-1)/ (6-1)$$

$$= 1.6040993e+31$$

The number of policy trees generated by these nodes are

$$P = |A|^{N}$$

$$= 5^{N}$$

$$=5^{1.6040993e+31}$$

It is a very large number. This happens due to the non-convergence of the number of nodes on the increment of the horizon. So, as the horizon size increases, there will be new policy trees.