MDL ASSIGNMENT-3 REPORT (PART - 2)

Team Members:
T.H.Arjun (Roll number: 2019111012)
Gokul Vamsi Thota (Roll number: 2019111009)

1. Here, we know that the target is in cell (1,0) and the certain observation is o6. So, for the target, row number is 1 and column number is 0. It is given that o6 is observed if and only if the target is not in the 1 cell neighbourhood of the agent. There are 10 possible such initial states (5 in which call is off and the other 5 in which call is on). Hence, the position of target and agent in the initial configuration can be given by the following set of possibilities.

	Agent		
Target(on/off)			

		Agent	
Target(on/off)			

			Agent
Target(on/off)			

Target(on/off)		Agent	

Target(on/off)			Agent

Thus, there are 5 possibilities for agent position (as shown above), there is exactly 1 possibility for the target position (given), and there are 2 possibilities for the call (can be on/off, as it doesn't depend upon the positions of agent and target). Therefore, there are a total of 10 possibilities for initial state. Hence, the agent's initial belief state would have equal probabilities for each of these 10 possibilities, therefore each of above 10 states have belief state probability as (1/10), i.e, 0.1. Rest of the states have belief state probability as 0. It can be depicted by the table below.
Note that a state can be represented by the tuple $((a_i, a_j), (t_i, t_j), c)$, where (a_i, a_j) represents position (row number, column number) of agent, (t_i, t_j) represents position (row number, column number) of target, and c represents call status (on/off).

Belief state notation for a given state	**Belief state probability for given state**
B((0, 1), (1, 0), on)	0.1
B((0, 1), (1, 0), off)	0.1
B((0, 2), (1, 0), on)	0.1
B((0, 2), (1, 0), off)	0.1
B((0, 3), (1, 0), on)	0.1
B((0, 3), (1, 0), off)	0.1
B((1, 2), (1, 0), on)	0.1
B((1, 2), (1, 0), off)	0.1
B((1, 3), (1, 0), on)	0.1
B((1, 3), (1, 0), off)	0.1
For all other $B((a_i, a_j), (t_i, t_j), c)$	0

2. It is given that we are in (1,1) and we also know that the target is in our one neighborhood and is not making a call. Hence, the position of target and agent in the initial configuration can be given by the following set of possibilities.

	Target(off)		
	Agent		

Target(off)	Agent		

	Agent	Target(off)	

	Agent, Target(off)		

Thus, there are 4 possibilities for the target position (as shown above), because the target should lie in one cell neighbourhood of the agent. There is exactly 1 possibility for the agent position (given), and there is 1 possibility for the call (given that call is off). Therefore, there are a total of 4 possibilities for initial state. Hence, the agent's initial belief state would have equal probabilities for each of these 4 possibilities, therefore each of above 4 states have belief state probability as (1/4), i.e, 0.25. Rest of the states have belief state probability as 0. It can be depicted by the table below. Note that a state can be represented by the tuple ((a_i, a_j), (t_i, t_j), c), where (a_i, a_j) represents position (row number, column number) of agent, (t_i, t_j) represents position (row number, column number) of target, and c represents call status (on/off).

Belief state notation for a given state	**Belief state probability for given state**
B((1, 1), (0, 1), off)	0.25
B((1, 1), (1, 0), off)	0.25
B((1, 1), (1, 2), off)	0.25
B((1, 1), (1, 1), off)	0.25
For all other B((a_i, a_j), (t_i, t_j), c)	0

3. We can find the expected utility for the initial belief states by running pomdpsim with the .pomdp file generated along with the.policyfile created from pomdpsol.

For question 1, Expected Utility = Expected reward = 19.3505

```
❯ python3 script.py > 2019111009_2019111012.pomdp
❯ ./pomdpsol.file  2019111009_2019111012.pomdp

Loading the model ...
  input file   : 2019111009_2019111012.pomdp
  loading time : 0.02s

SARSOP initializing ...
  initialization time : 0.00s

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0       0       0        9.99197    19.8535    9.86149      5        1
 0       10      53       19.5285    19.6371    0.108555     21       13
 0.01    16      101      19.6132    19.6293    0.0160505    42       19
 0.01    20      150      19.6219    19.6279    0.00598159   49       35
 0.02    24      203      19.6258    19.6274    0.00169434   59       42
 0.03    27      250      19.6258    19.6272    0.00134873   83       51
 0.04    30      300      19.626     19.6271    0.00115006   96       65
 0.05    31      317      19.6261    19.6271    0.000973678  90       71
-------------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000974

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.05    31      317      19.6261    19.6271    0.000973678  90       71
-------------------------------------------------------------------------------

Writing out policy ...
  output file : out.policy

 /mnt/Documents/MDL_Assign/Assignment3  on  main *1 !2 ?2
❯
```

```
❯ ./pomdpsim.file --simLen 100 --simNum 1000 --policy-file out.policy 2019111009_2019111012.pomdp

Loading the model ...
  input file   : 2019111009_2019111012.pomdp

Loading the policy ...
  input file   : out.policy

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

-----------------------------------
 #Simulations  | Exp Total Reward
-----------------------------------
 100             20.439
 200             20.2871
 300             20.7298
 400             20.4729
 500             19.9323
 600             19.7092
 700             19.2598
 800             19.4369
 900             19.2173
 1000            19.3505
-----------------------------------

Finishing ...

-------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
-------------------------------------------------------------
 1000            19.3505            (18.4725, 20.2285)
-------------------------------------------------------------

 /mnt/Documents/MDL_Assign/Assignment3  on  main *1 !2 ?2
❯
```

For question 2, Expected Utility = Expected reward = 37.0572

```
❯ python3 script.py > 2.pomdp
❯ ./pomdpsol.file  2.pomdp

Loading the model ...
  input file   : 2.pomdp
  loading time : 0.02s

SARSOP initializing ...
  initialization time : 0.01s

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.01    0       0        19.7792    51.6051    31.8259      5        1
 0.01    10      50       37.2073    37.2899    0.0826172    39       17
 0.01    17      101      37.2724    37.2821    0.00971064   70       30
 0.02    21      151      37.2777    37.2812    0.00351243   98       40
 0.03    24      200      37.2786    37.281     0.00241164   116      55
 0.04    27      250      37.2791    37.2807    0.00166209   142      68
 0.06    31      300      37.2794    37.2804    0.00103681   172      81
 0.06    32      313      37.2794    37.2804    0.000969246  185      86
-------------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000969

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.06    32      313      37.2794    37.2804    0.000969246  178      86
-------------------------------------------------------------------------------

Writing out policy ...
  output file : out.policy

 /mnt/Documents/MDL_Assign/Assignment3  on  main *1 !2 ?2
❯
```

```
 0.06    32      313      37.2794    37.2804    0.000969246  178      86
-------------------------------------------------------------------------------

Writing out policy ...
  output file : out.policy

❯ ./pomdpsim.file --simLen 100 --simNum 1000 --policy-file out.policy 2.pomdp

Loading the model ...
  input file   : 2.pomdp

Loading the policy ...
  input file   : out.policy

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

-----------------------------------
 #Simulations   | Exp Total Reward
-----------------------------------
 100             37.9724
 200             36.3708
 300             37.0795
 400             36.8741
 500             36.8304
 600             36.6757
 700             36.9317
 800             37.0865
 900             37.1579
 1000            37.0572
-----------------------------------

Finishing ...

-------------------------------------------------------------
 #Simulations   | Exp Total Reward | 95% Confidence Interval
-------------------------------------------------------------
 1000            37.0572            (36.2079, 37.9064)
-------------------------------------------------------------

 /mnt/Documents/MDL_Assign/Assignment3  on  main *1 !2 ?2
❯
```

4. Given that our agent is in (0,0) with probability 0.4 and in (1,3) with probability 0.6 and the target is in (0,1), (0,2), (1,1) and (1,2) with equal probability for each.

The probability of making an observation o given that the state is s is P(o|s), then the probability of making an observation is,

$$P(o) = \sum_{s} P(s) * P(o/s)$$

Case 1: Consider the agent is in (0, 0), which happens with a probability of 0.4 . Now the target is in (0,1), (0,2), (1,1) and (1,2) with an equal probability of 0.25 for each case.
The grid view possibilities are as follows:

Agent	Target		

Agent		Target	

Agent			
	Target		

Agent			
		Target	

In these possible cases of grid,

- o1 is not observed
- o2 is observed when the target is in the cell to the right of the agent's cell, here it is observed in 1 possibility out of above 4 possibilities. Hence, probability of occurrence is (0.25*1) = 0.25
- o3 is not observed.
- o4 is not observed.
- o5 is not observed.
- o6 is observed when the target is not in the 1 cell neighbourhood of the agent, it is observed in 3 possibilities out of above 4 possibilities. Hence, probability of occurrence is (0.25*3) = 0.75

Case 2: Now let us consider the case where the agent is in (1, 3), which happens with probability 0.6, and the target is in (0,1), (0,2), (1,1) and (1,2) with an equal probability of 0.25 for each possibility.

The grid view possibilities are as follows:

	Target		
			Agent

		Target	
			Agent

	Target		Agent

		Target	Agent

In these possible cases of grid,

- o1 is not observed.
- o2 is not observed.
- o3 is not observed.
- o4 is observed when the target is in the cell to the left of the agent's cell. Here it is observed exactly in 1 possibility out of above 4 possibilities. Hence, probability of occurrence is (0.25*1) = 0.25.
- o5 is not observed.
- o6 is observed when the target is not in the 1 cell neighbourhood of the agent. Here it is observed in 3 possibilities out of above 4 possibilities. Hence, probability of occurrence is (3*0.25)=0.75

So probabilities averaged across the two cases, i.e, the actual probability that a particular observation is noted, for each possible observation, is as follows:

Observation	**Probability of occurrence**
o1	0.4*0 + 0.6*0 = 0
o2	0.4*0.25 + 0.6*0 = 0.1
o3	0.4*0 + 0.6*0 = 0
o4	0.4*0 + 0.6*0.25 = 0.15
o5	0.4*0 + 0.6*0 = 0
o6	0.4*0.75 + 0.6*0.75= 0.75

So, from the table it is very clear that **o6 is most likely to be observed.**

5.

$$N = \sum_{i=0}^{T-1} |O|^{i} = \frac{|O|^{T}-1}{|O|-1}$$

We can compute number of policy trees as $|A|^{N}$ where A is the number of actions possible , O is the number of observations possible, and T is Time Horizon (or the number of steps the agent takes. Here we have O=6, A= 5. To find T ran pomdpsol command on the .pomdp file that we generated for this problem. Note that the initial beliefs of this .pomdp file is the initial beliefs of question 4. The execution of this .pomdp file on pomdpsol is given below:

```
❯ python3 script.py >4.pomdp
❯ ./pomdpsol.file  4.pomdp

Loading the model ...
  input file   : 4.pomdp
  loading time : 0.02s

SARSOP initializing ...
  initialization time : 0.00s

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0       0       0        12.4788    33.107     20.6283     5        1
 0.01    11      51       26.7225    26.8696    0.147081    14       15
 0.01    17      100      26.8256    26.8579    0.032253    40       27
 0.02    21      151      26.8399    26.8521    0.0122324   63       42
 0.03    26      200      26.8461    26.8512    0.00510004  86       52
 0.03    30      253      26.8486    26.8509    0.00236657  100      57
 0.04    33      300      26.8489    26.8508    0.00185491  120      67
 0.06    37      350      26.8493    26.8506    0.00129579  145      84
 0.07    40      401      26.8493    26.8505    0.00113929  161      93
 0.07    42      429      26.8495    26.8505    0.000998269 172      97
-------------------------------------------------------------------------------

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

-------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
-------------------------------------------------------------------------------
 0.07    42      429      26.8495    26.8505    0.000998269 172      97
-------------------------------------------------------------------------------

Writing out policy ...
  output file : out.policy

  /mnt/Documents/MDL_Assign/Assignment3  on  main *1 !2 ?3
❯
```

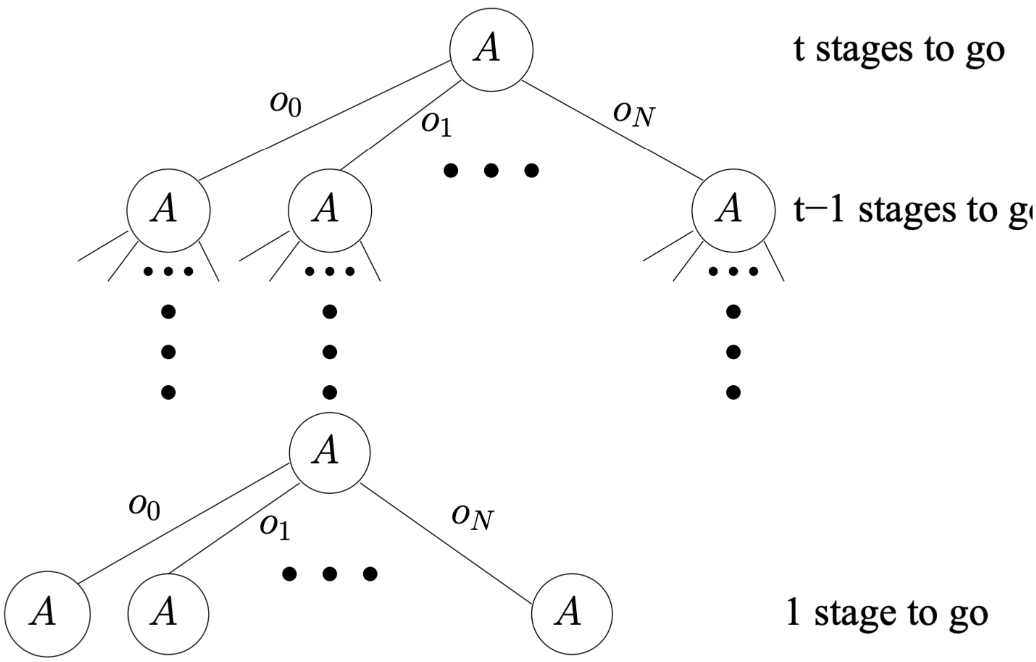
So we can see that Trials= T= 42. Substituting in formula we get:

$$N = \frac{|6|^{42} - 1}{|6| - 1} = 9.624596067967489e+31$$

No. of Policy Trees = $|A|^{N}$ = $|5|^{9.624596067967489e+31}$

Hence we get the number of policy trees as a very large finite value, mentioned above. The explanation as to how this number is obtained is given below.

Given below is a policy tree for horizon t. For each observation, there is a branch to nodes at a lower level. Each node can be labeled with any action from the set A (the set of actions).



We can see that as we increase the Time Horizon, the number of nodes do not converge easily. This is because of the explosion/divergence of number of observation possibilities which percolate from a set of action nodes, after a certain depth in the policy tree. This is attributed to the absence of an absorbing state / final state for the given POMDP model.

Hence as we increase the Time Horizon, there will be more and more policy trees tending the number of policy trees to very large numbers, which would possess very large exponential values as we have obtained above, on using the formula. This explains the reason behind such a massive number of potential policy trees.