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 = 15.973

```
❯ 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        8.5049     16.2511    7.74625     5        1
 0.01   10      51       15.8964    16.0584    0.161995    22       14
 0.01   16      105      16.027     16.039     0.011912    45       24
 0.02   20      151      16.0338    16.0381    0.0042314   50       30
 0.02   24      201      16.0362    16.0377    0.00148119  93       43
 0.04   27      257      16.0364    16.0375    0.00109203  114      60
 0.04   30      305      16.0364    16.0374    0.000975568 109      63
 0.04   30      305      16.0364    16.0374    0.000975568 109      63
---------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000976

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0.04   30      305      16.0364    16.0374    0.000975568 109      63
---------------------------------------------------------------------------

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


 /mnt/Documents/MDL_Assign/Assignment3 on  main *1 !8 ?1
❯ ./pomdpsol.file  2019111009_2019111012.pomdp
```

```
❯ ./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           16.9122
 200           16.5304
 300           16.6132
 400           16.3385
 500           16.6433
 600           16.5219
 700           16.4679
 800           16.1726
 900           16.0631
 1000          15.973
-----------------------------------

Finishing ...

-------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
-------------------------------------------------------------
 1000          15.973             (15.2228, 16.7232)
-------------------------------------------------------------
❯

 /mnt/Documents/MDL_Assign/Assignment3 on  main *1 !8 ?1
❯
```

For question 2, Expected Utility = Expected reward = 31.458

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

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

SARSOP initializing ...
  initialization time : 0.00s

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0      0       0        16.8082    43.1466    26.3384     5        1
 0.01   11      55       31.2081    31.2468    0.0386742   35       16
 0.01   16      101      31.2373    31.2451    0.00783112  63       31
 0.02   20      150      31.2412    31.2446    0.00338162  98       47
 0.03   24      200      31.2423    31.2444    0.00208511  122      58
 0.04   28      250      31.2428    31.2441    0.00128788  153      70
 0.05   31      289      31.2429    31.2439    0.000999366 183      86
---------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000999

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0.05   31      289      31.2429    31.2439    0.000999366 179      86
---------------------------------------------------------------------------

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

 /mnt/Documents/MDL_Assign/Assignment3 on  main *1 !8 ?1
```

```
  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           32.419
 200           31.1287
 300           30.6049
 400           30.7102
 500           30.9239
 600           31.1046
 700           31.0987
 800           31.1108
 900           31.2334
 1000          31.458
-----------------------------------

Finishing ...

-------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
-------------------------------------------------------------
 1000          31.458             (30.6996, 32.2164)
-------------------------------------------------------------

 /mnt/Documents/MDL_Assign/Assignment3 on  main *1 !8 ?1
❯
```

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        11.7668    27.6077    15.8408     5        1
 0.01   11      50       22.2436    22.3557    0.112061    32       15
 0.01   17      101      22.3195    22.3455    0.0260407   52       24
 0.02   22      150      22.3306    22.3415    0.0109038   65       34
 0.02   27      200      22.336     22.3405    0.00450918  76       46
 0.03   31      250      22.3378    22.3402    0.00246775  107      60
 0.05   35      307      22.3383    22.34      0.0016485   143      72
 0.06   38      353      22.3386    22.3398    0.00125481  170      82
 0.07   41      400      22.3387    22.3397    0.00103436  187      100
 0.08   42      411      22.3388    22.3397    0.000958154 187      100
---------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000958

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0.08   42      411      22.3388    22.3397    0.000958154 187      100
---------------------------------------------------------------------------

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

 /mnt/Documents/MDL_Assign/Assignment3 on  main *1 !8 ?1
❯
```

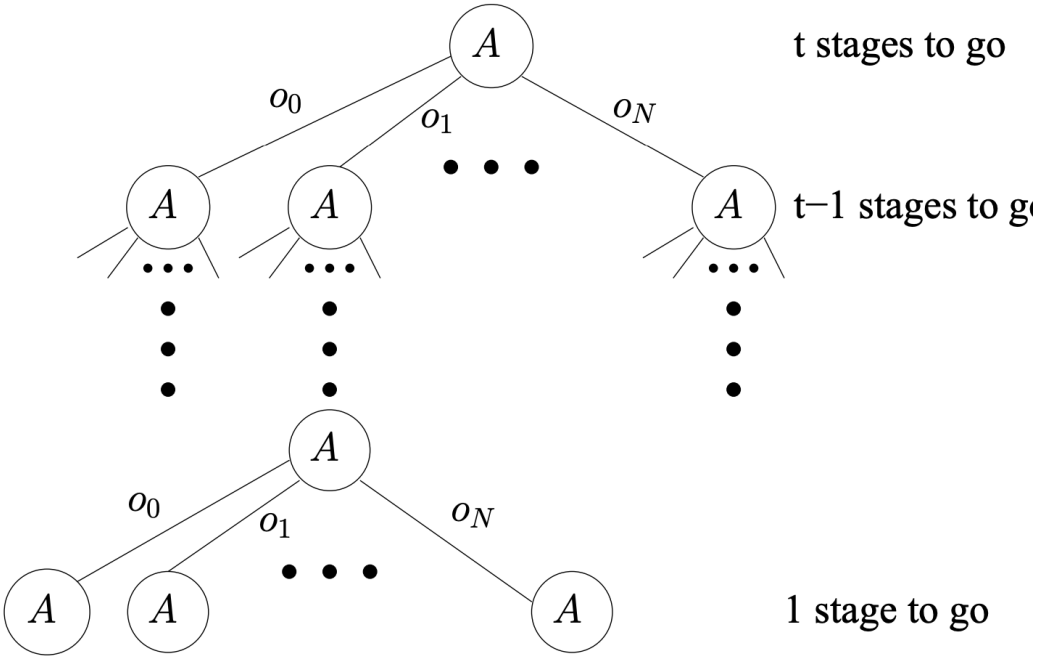
So we can see that Trials= T= 42. Substituting in formula we get:

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

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.