MDL Assignment 3 Part 2

SARSOP Solver

Part 1

- Dayitva Goel: 2019101005
- Prajneya Kumar: 2019114011

1 Overview of POMDP

A partially observable Markov decision process (POMDP) is a generalization of a Markov decision process (MDP). A POMDP models an agent decision process in which it is assumed that the system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state. Instead, it must maintain a probability distribution over the set of possible states, based on a set of observations and observation probabilities, and the underlying MDP.

2 Calculation

The roll number we have used is 2019114011.

$$\begin{aligned} x &= 1 - ((21 + 1)/100)) \\ &= 1 - \frac{22}{100} \\ &= 1 - 0.22 \\ &= 0.78 \end{aligned}$$

$$\begin{aligned} \text{Reward} &= 2019114011\%90 + 10 \\ &= 11 + 10 \\ &= 21 \end{aligned}$$

3 Questions

3.1 Question 1

If you know the target is in (1,0) cell and your observation is o6, what will be the initial belief state? Please submit the optimal policy file named RollNumber.policy for the POMDP taking into account the initial belief state you obtained.

Each state of the POMDP is represented as a tuple **(Agent Position, Target Position, Call)**.

Therefore, there are a total of 8 x 8 x 2 = 128 states.

The target is in (1,0) cell and since our observation is o6, the target cannot be in 1 cell neighbourhood of the agent. This limits the agent to cells (0,1), (0,2), (0,3), (1,2) and (1,3).

The number of states reduce to 5 x 1 x 2 = 10 states.

The 10 states are:

((0,1), (1,0), On)
((0,1), (1,0), Off)
((0,2), (1,0), On)
((0,2), (1,0), Off)
((0,3), (1,0), On)
((0,3), (1,0), Off)
((1,2), (1,0), On)
((1,2), (1,0), Off)
((0,3), (1,0), On)
((0,3), (1,0), Off)

The initial belief state values for above states will be 1/10 and 0 for all other states.

3.2 Question 2

If you are in (1,1) and you know the target is in your one neighborhood and is not making a call what is your initial belief state?

The agent is in (1,1) cell and since the target is in one neighbourhood, it can be present in (1,0), (1,1), (1,2) or (0,1). The target is also not making a call.

The number of states reduce to 1 x 4 x 1 = 4 states.

The 4 states are:

((1,1), (1,0), Off)
((1,1), (1,1), Off)
((1,1), (1,2), Off)
((1,1), (0,1), Off)

The initial belief state values for these states will be 1/4 and 0 for all other states.

3.3 Question 3

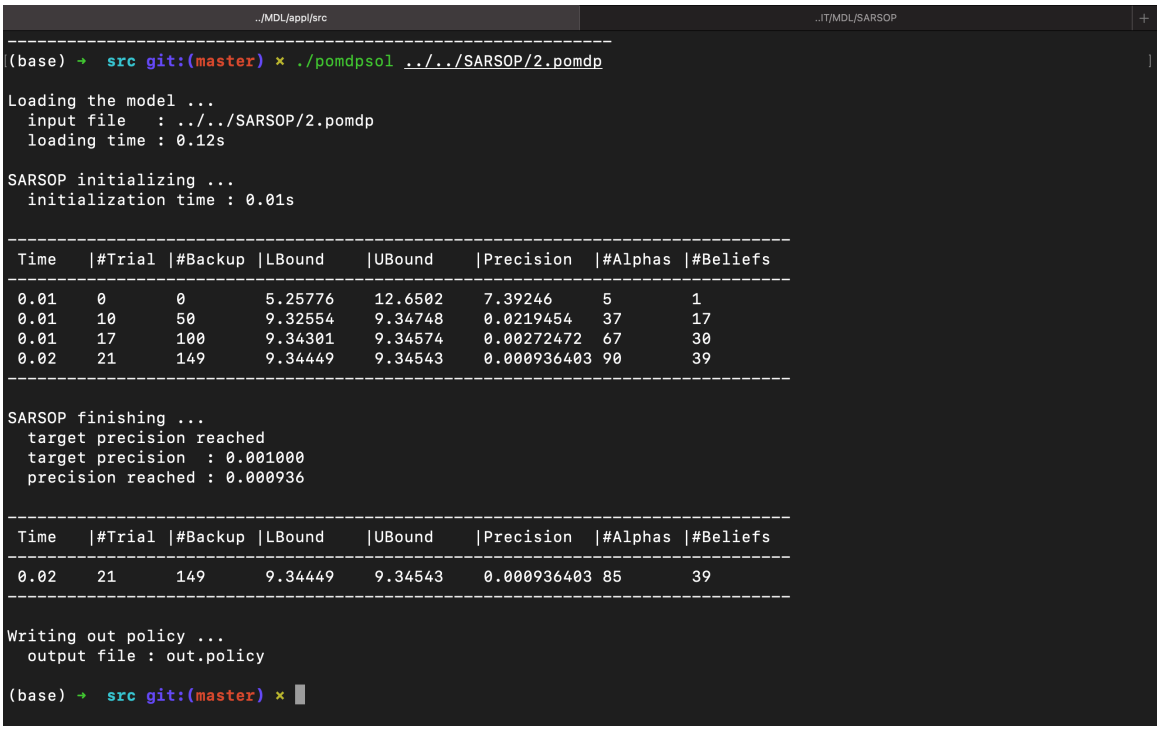
What is the expected utility for initial belief states in questions 1 and 2?

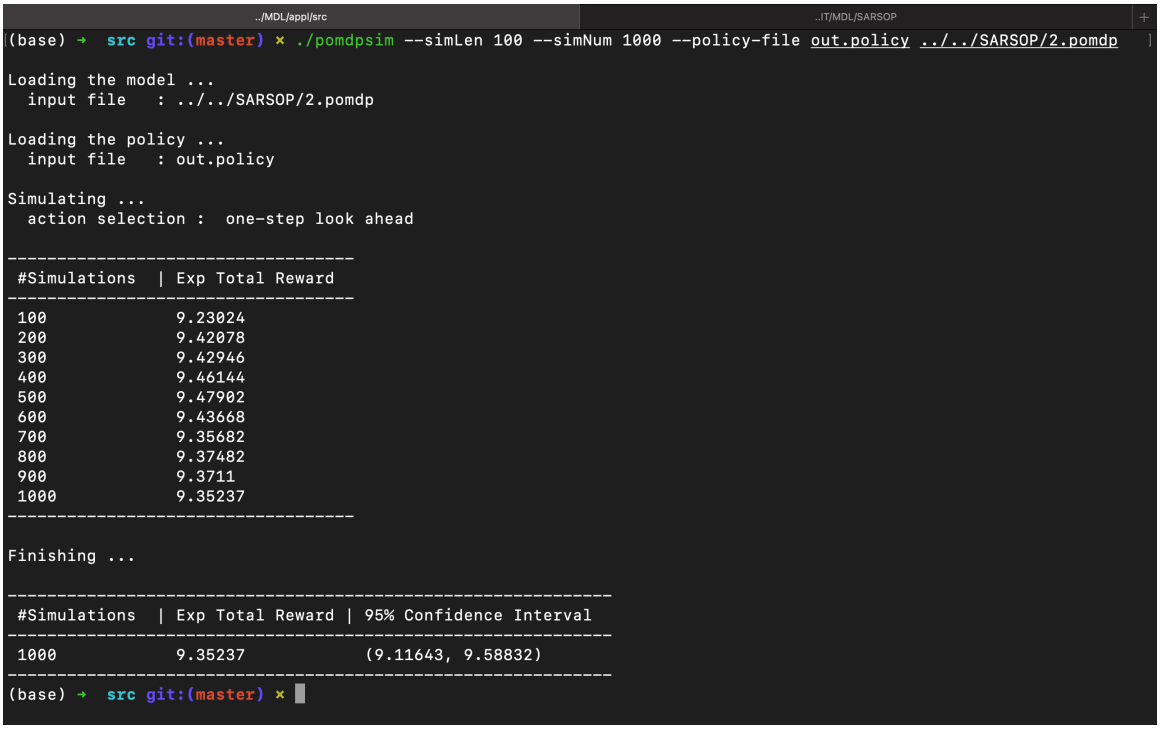
A POMDP file was created for both questions 1 and 2 and SARSOP was run on them. A policy file was created using 'pomdpsol' and then the expected utility was calculated using 'pomdpsim'.

Here are the results:

```
../MDL/appl/src                                    ../IT/MDL/SARSOP
------------------------------------------------------------
(base) ➜  src git:(master) ✗ ./pomdpsol ../../SARSOP/1.pomdp

Loading the model ...
  input file   : ../../SARSOP/1.pomdp
  loading time : 0.11s

SARSOP initializing ...
  initialization time : 0.00s

--------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
--------------------------------------------------------------------------------
 0      0       0        1.05956    3.72339    2.66383     5        1
 0.01   10      50       3.64473    3.67053    0.0257963   22       15
 0.01   16      107      3.66246    3.66502    0.00255483  30       28
 0.01   19      141      3.66394    3.66477    0.000831291 55       33
--------------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000831

--------------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
--------------------------------------------------------------------------------
 0.01   19      141      3.66394    3.66477    0.000831291 53       33
--------------------------------------------------------------------------------

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

(base) ➜  src git:(master) ✗
```

```
../MDL/appl/src                                    ../IT/MDL/SARSOP
(base) ➜  src git:(master) ✗ ./pomdpsim --simLen 100 --simNum 1000 --policy-file out.policy ../../SARSOP/1.pomdp

Loading the model ...
  input file   : ../../SARSOP/1.pomdp

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

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

-----------------------------------
 #Simulations  | Exp Total Reward
-----------------------------------
 100           3.44528
 200           3.22822
 300           3.43327
 400           3.54678
 500           3.6496
 600           3.66255
 700           3.62386
 800           3.7268
 900           3.67319
 1000          3.74056
-----------------------------------

Finishing ...

------------------------------------------------------------
 #Simulations  | Exp Total Reward | 95% Confidence Interval
------------------------------------------------------------
 1000          3.74056          (3.5028, 3.97832)
------------------------------------------------------------
(base) ➜  src git:(master) ✗
```





The expected utility is given under the header 'Exp Total Reward'.

In question 1, expected utility = 3.74856

In question 2, expected utility = 9.35237

3.4 Question 4

If your 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, which observation are you most likely to observe? Explain.

Case I: If the agent is in (0,0):

Observation	Target Location	Probability
o2	(0,1)	0.4 x 0.25 = 0.1
o6	(0,2)	0.4 x 0.25 = 0.1
o6	(1,1)	0.4 x 0.25 = 0.1
o6	(1,2)	0.4 x 0.25 = 0.1

Case II: If agent is in (1,3):

Observation	Target Location	Probability
o6	(0,1)	0.6 x 0.25 = 0.15
o6	(0,2)	0.6 x 0.25 = 0.15
o6	(1,1)	0.6 x 0.25 = 0.15
o4	(1,2)	0.6 x 0.25 = 0.15

The final probabilities for all observations are:

Observation	Probability
o1	0
o2	0.1
o3	0
o4	0.15
o5	0
o6	0.75

Therefore, observation o6 is most likely.

3.5 Question 5

How many policy trees are obtained in the case of question 4, explain?

We have,
No of actions = $|A| = 5$
No of observations = $|O| = 6$
Time horizon = T
No of nodes = N

```
../MDL/appl/src                                                    ..IT/MDL/SARSOP
------------------------------------------------------------
(base) ➜  src git:(master) ✗ ./pomdpsol ../../SARSOP/4.pomdp

Loading the model ...
  input file   : ../../SARSOP/4.pomdp
  loading time : 0.12s

SARSOP initializing ...
  initialization time : 0.01s

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0.01    0       0        1.85555    7.34143    5.48587     5        1
 0.01    11      51       5.71633    5.75661    0.0402777   15       15
 0.02    16      100      5.74437    5.75337    0.00899969  42       27
 0.02    21      153      5.74795    5.75199    0.00404175  71       39
 0.03    26      200      5.75049    5.75174    0.00124482  75       45
 0.03    28      227      5.75077    5.75162    0.00085883  86       54
---------------------------------------------------------------------------

SARSOP finishing ...
  target precision reached
  target precision  : 0.001000
  precision reached : 0.000859

---------------------------------------------------------------------------
 Time   |#Trial |#Backup |LBound    |UBound    |Precision  |#Alphas |#Beliefs
---------------------------------------------------------------------------
 0.03    28      227      5.75077    5.75162    0.00085883  86       54
---------------------------------------------------------------------------

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

Hence, T = No of trials = 28

$$
\begin{aligned}
N &= \sum_{i=0}^{T-1} |O|^i \\
&= \frac{|O|^T - 1}{|O| - 1} \\
&= \frac{6^2 8 - 1}{6 - 1} \\
&= \frac{6.1409422e + 21}{5} \\
&= 1.2281884e + 21
\end{aligned}
$$

We can now calculate the number of policy trees.

$$
\begin{aligned}
\text{No of policy trees} &= |A|^N \\
&= 5^{1.2281884e+21} \\
&\approx \infty
\end{aligned}
$$