Machine, Data and Learning

Assignment-5

Part-2

Anirudh Palutla

2018113007

Question 1

A state s can be represented as:

```
s = (agent_position, target_position, call_active)
where agent_position = (x_{a}, y_{a})
      target_position = (x_{t}, y_{t})
      call_active = True/False
```

The observation $o6$ is observed when the target is not in the 1-neighbourhood of the agent. Then, the possible start states are those states where `agent_position` and `target_position` are not adjacent to each other.
We have the grid:

	0	**1**	**2**
0	A		A
1		T	
2	A		A

In this table, A marks the possible positions of the agent and the target is positioned at (1, 1) as mentioned in the question marked by T. These are all the possible positions of the agent which are not adjacent to the target, and hence $o6$ is observed.

Hence, the initial belief state can be given by the function:

$$b(s) = \begin{cases} 1/\eta & \text{if } \texttt{agent\_position in [(0, 0), (0, 2), (2, 0), (2, 2)]} \\ & \text{and } \texttt{target\_position == (1, 1)} \\ 0 & \text{otherwise} \end{cases}$$

Here, $\eta = 8$ as the number of such states are 8. Hence, for any state s where `agent_position` is one of [(0, 0), (0, 2), (2, 0), (2, 2)], and `target_position` is (1, 1) and `call_active` is either True or False. This gives us 8 states, whose belief value is 1/8 and the belief value for the rest of the states is 0.

Question 2

The general tuple for such a state for a is given by:

```
s = ((0, 1), target_pos, False)
```

Here, `target_pos` is one of (0, 1), (0, 0), (0, 2) or (1, 1) since it is in the 1-neighbourhood of (0, 1).
There are hence 4 states that fit the given general form. In the initial belief state, the belief value for each of these 4 states is $\frac{1}{4}$ and the belief value for the rest is 0.

Question 3

The expected utility for a belief state is given by:

$$r(b) = \sum_{s \in S} b(s) * R(s)$$

where R is the reward function for a particular state. (In this problem, reward is solely dependent on the state and not action taken)

Calculation for Q1

The reward for every state with a non-zero belief value in Q1 is equal to -1 as the agent and the target are not in the same position in any of these states. They also have the same belief value, which is $\frac{1}{8}$ and there are 8 such states. Hence:

$$\begin{aligned} r(b_1) &= 8 * (\frac{1}{8} * (-1)) \\ &= -1 \end{aligned}$$

Therefore, the expected utility value for the belief state in Q1 is -1.

Calculation for Q2

The reward for every state with non-zero belief in Q2 is also equal to -1. Although there is a state where the agent and the target are in the same position, the target is not making a call, i.e, `call_active == False`. They also have the same belief value equal to $\frac{1}{4}$ and there are 4 such states. Hence:

$$\begin{aligned} r(b_2) &= 4 * (\frac{1}{4} * (-1)) \\ &= -1 \end{aligned}$$

Therefore, the expected utility value for the belief state in Q2 is -1.

Question 4

In the first case, where agent is in (0, 1) and the target is in the four corners of the cell, we have the grid:

	0	**1**	**2**
0	T	A	T
1			
2	T		T

This gives the observations $o2$ and $o4$ with probability $\frac{1}{4}$ each and the observation $o6$ with probability $\frac{1}{2}$ when agent is in (0, 1).

In the second case, where agent is in (2, 1), we have the grid:

	0	**1**	**2**
0	T		T
1			
2	T	A	T

This also gives the observations $o2$ and $o4$ with probability $\frac{1}{4}$ each and the observation $o6$ with probability $\frac{1}{2}$ when agent is in (2, 1).
Since we have the same observations from both states, there is no need to weight the observation probabilities with the probability of the occurrence of the agent's positions.
In total, we observe $o2$ with probability $\frac{1}{4}$, $o4$ with probability $\frac{1}{4}$ and $o6$ with probability $\frac{1}{2}$.
Therefore, the observation we are most likely to observe is $o6$.

Question 5

If there are A actions, O observations and T horizon, the number of policy trees P is given by:

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

From the formula, we can see that the number of trees obtained is dependent on the horizon T. The value of P monotonously increases as the value of T increases.
In the SARSOP Solver, the horizon is not a fixed value, but depends on the precision value obtained. The program terminates when a target precision is obtained for a certain converging calculated value. Hence, the number of policy trees obtained is not a fixed value and can not be calculated without running the program.