Homework 3. Partially observable Markov decision problems

We will continue with the game we developed in HW2. Consider now that the spider cannot observe in which step of the ladder it is in. However, using its legs, the spider can identify whether it is sitting on the ground, at a web, or at the top level.

Exercise

(a) Define a POMDP based on the available information. The observations are {g,w,t,e} for ground, web, top level, and empty (no observation).

(b) Compute the belief (the probability that the spider is at a given state) for the following situations:

  (a) After feeling the web with its legs.

  (b) After feeling the web with its legs and playing two turns, assuming that the spider made no observation after each step (i.e., it makes two empty observations).

  (c) After starting and playing 3 times, assuming that the spider made no observation after each step (i.e., it makes three empty observations).

(c) If the belief is:

$$\begin{bmatrix} 0.2 & 0.08 & 0.24 & 0.32 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

what is the best action to make? (Use MLS and QMDP)

Consider the following transition matrices and cost function describing the game developed in HW2.

$$\boldsymbol{P}^{play} = \begin{bmatrix} .2 & .4 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .2 & 0 & .4 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .2 & 0 & 0 & .4 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .2 & 0 & 0 & 0 & .4 & .4 & 0 & 0 & 0 & 0 & 0 & 0 \\ .2 & 0 & 0 & 0 & 0 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .2 & .4 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .2 & 0 & .4 & .4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .2 & .4 & .4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .2 & 0 & .4 & .4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .2 & 0 & 0 & .8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{P}^{stop} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The cost function is

$$C^{play} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$
$$C^{stop} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Solution:

(a) $\mathcal{Z} = \{g, w, t, e\}$

$\gamma = 0.9$, same as HW2

$$\mathbf{O} = \{\mathbf{O}^{play}, \mathbf{O}^{stop}\}$$

$$\mathbf{O}^{play} = \mathbf{O}^{stop} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Thus, the POMDP is fully specified by the tuple: $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathbf{P}, \mathbf{O}, C, \gamma)$

(b) (a) After feeling the web with its legs (observation w)

$$b_t = \delta(8) = \begin{bmatrix} 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 1 & 0. & 0. & 0. \end{bmatrix}$$

As only in state 8 the spider can feel the web with its legs, the spider is sure to be in that state. It is irrelevant the initial belief in this case (as long as the probability of being in state 8 is not zero).

(b) From the previous question we know that $b_t = \delta(8)$. After feeling the web with its legs and playing two turns (assuming the spider never felt the web with its legs again, nor felt it was at the top level), we have that

$$\begin{aligned} b_{t+1} &= \frac{b_t P^{play} diag(O(:, \mathrm{e}))}{\| b_t P^{play} diag(O(:, \mathrm{e})) \|_1} \\ &= \begin{bmatrix} 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.5 & 0.5 & 0. \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
b_{t+2} &= \frac{b_{t+1}P^{play}diag(O(:,\mathrm{e}))}{\|b_{t+1}P^{play}diag(O(:,\mathrm{e}))\|_1} \\
&= \begin{bmatrix} 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 1 & 0. \end{bmatrix}
\end{aligned}
$$

(c) After starting and playing 3 times

$$
b_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
$$

$$
\begin{aligned}
b_1 &= \frac{b_0P^{play}diag(O(:,e))}{\|b_0P^{play}diag(O(:,e))\|_1} \\
&= \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
b_2 &= \frac{b_1P^{play}diag(O(:,e))}{\|b_1P^{play}diag(O(:,e))\|_1} \\
&= \begin{bmatrix} 0 & 0 & 0.25 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
b_3 &= \frac{b_2P^{play}diag(O(:,e))}{\|b_2P^{play}diag(O(:,e))\|_1} \\
&= \begin{bmatrix} 0 & 0 & 0 & 0.25 & 0.75 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

Alternatively, we could make the following reasoning: since the agent did not fall nor reached the top in three consecutive plays, it must be either in state 3 or state 4. The sequence of the die was either 1-1-1, 1-1-2, 1-2-1, or 2-1-1, each with equal probability. Moreover the spider always succeeded in climbing the ladder the number of steps given by the die. Therefore the spider is either in state 3, with 0.25 probability, or in state 4, with 0.75 probability.

(c) To answer this question, we need first to solve the MDP, yielding the following optimal Q function:

(using the transition matrices specified in this problem statement)

```
[[3.812 4.431]
 [3.41  3.431]
```

```
[3.119 2.495]
[2.293 3.431]
[1.686 3.431]
[0.    0.   ]
[3.948 4.553]
[3.443 3.553]
[2.773 3.495]
[2.039 2.495]
[1.499 2.495]
[0.    0.   ]]
```

If the belief is

$$\begin{bmatrix} 0.2 & 0.08 & 0.24 & 0.32 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

then:

- using the MLS heuristic. The most probable state is state 3 and the best action in that case is action *play*.
- using the AV heuristic we get

  $$AV = [0.76, 0.24],$$

  and we select *play* since it has more votes
- using QMDP then

  $$bQ = [2.79, 3.41]^T$$

  action *play*.