Exercise Sheet - POMDPs

This exercise concerns a POMDP where the underlying states form a chain on which the agent can walk left or right. To help conceptualize, the underlying Markov chain is shown in Figure 1:

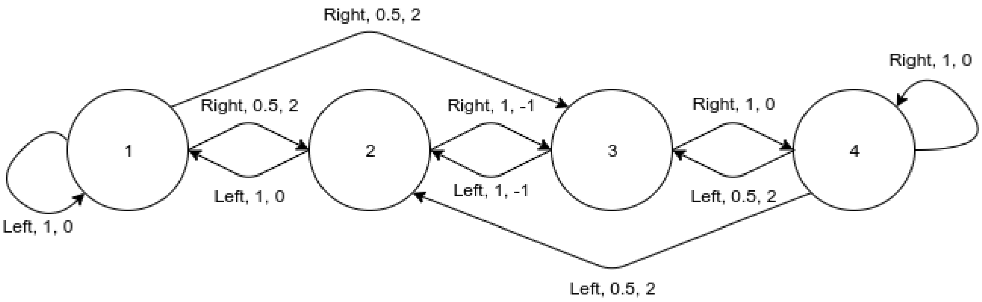

Figure 1: Underlying problem. The values along the arrow denote the action, the probability of the transition, and the immediate reward.

Formally, the POMDP can be described as folllows:

- $\mathcal{S} = \{1, 2, 3, 4\}$
- $\mathcal{A} = \{Left, Right\}$
- $\mathcal{O} = \{Green, Blue\}$
- Transitions:

```
%Transitions: P(s'|s,a) = T{a}(from, to)
T{Left} = [
[1,  0,  0,  0 ] %from S1
[1,  0,  0,  0 ] %from S2
[0,  1,  0,  0 ] %from S3
[0, .5, .5,  0 ] %from S4
];
T{Right} = [
[0, .5, .5,  0 ] %from S1
[0,  0,  1,  0 ] %from S2
[0,  0,  0,  1 ] %from S3
[0,  0,  0,  1 ] %from S4
];
```

- Observation probabilities:

```
%P(o|s') = O(s',o)
%i.e., the first row specifies the probabilities [ P(Green|S1), P(Blue|S1) ]
O = [
[ 1,  0] %to S1
[.5, .5],%to S2
[.5, .5],%to S3
[ 0,  1] %to S4
];
```

- Rewards:

```
%R(from, a)
%i.e., the first row specifies [ R(S1,Left), R(S1,Right) ]
R = [
[0,  2],
[0, -1],
[-1, 0],
[2,  0]
]
```

- $b_0 = (1, 0, 0, 0)$ is the initial belief. (I.e., we know we start in state 1)

Given this POMDP....

1. Compute the tree of all reachable beliefs by taking 2 actions.
2. For all these beliefs compute the expected immediate reward for taking action $Left$ or $Right$.
3. Now, perform backwards induction to compute $V^{\tau=2}(b_0)$ (the value of the initial belief for two timesteps to go).

Solution: For 1, we update the belief using Bayes' rule:

$$b'(s') = \frac{P(o|s') \sum_{s \in \mathcal{S}} P(s'|s,a) b(s)}{P(o|b,a)}.$$

With

$$P(o|b,a) = \sum_{s \in \mathcal{S}} b(s) \sum_{s' \in \mathcal{S}} P(s'|s,a) P(\mathrm{o}|s').$$

Starting from the initial belief $b_0(1,0,0,0)$, we first calculate the reachable beliefs by taking the action right.

From the transition function we know that we will end up in either state 2 or 3, so $b_1(1) = b_1(4) = 0$. Since for states 2 and 3 the observation probabilities are the same, we get $b_1(2) = b_1(3)$.

$$\begin{aligned} b_1(2) &= \frac{P(o|2) \sum_{s \in \mathcal{S}} P(2|s,a) b(s)}{P(o|b,a)} \\ &= \frac{0.5 \times 0.5 \times 1}{1 \times (0.5 \times 0.5 + 0.5 \times 0.5)} = 0.5 \end{aligned}$$

So after taking the action right and observing either Green or Blue, we end up in belief state $b_1 = (0, 0.5, 0.5, 0)$.

Now we calculate the reachable beliefs after taking the action right twice. We know that after taking the action right once we will end up in belief state $b_1 = (0, 0.5, 0.5, 0)$. After taking the action right again we can end up in either state 3 or state 4. Since for these states the observation functions are different, we need to take into account the observation we get.

For observation Green:

$$\begin{aligned} b_2(3) &= \frac{P(\text{Green}|3) \sum_{s \in \mathcal{S}} P(3|s,a) b_1(s)}{P(\text{Green}|b_1,a)} \\ &= \frac{0.5 \times 1 \times 0.5}{0.5 \times 1 \times 0.5 + 0.5 \times 1 \times 0} = 1 \\ b_2(4) &= 0 \end{aligned}$$

For observation Blue:

$$\begin{aligned} b_2(3) &= \frac{P(\text{Blue}|3) \sum_{s \in \mathcal{S}} P(3|s,a) b_1(s)}{P(\text{Blue}|b_1,a)} \\ &= \frac{0.5 \times 1 \times 0.5}{0.5 \times 1 \times 0.5 + 0.5 \times 1 \times 1} = 1/3 \\ b_2(4) &= \frac{P(\text{Blue}|4) \sum_{s \in \mathcal{S}} P(3|s,a) b_1(s)}{P(\text{Blue}|b_1,a)} \\ &= \frac{1 \times 1 \times 0.5}{0.5 \times 1 \times 0.5 + 0.5 \times 1 \times 1} = 2/3 \end{aligned}$$

So after taking the action right twice, we can end up in either $(0, 0, 1, 0)$ (if the second observation = Green) or in $(0, 0, 1/3, 2/3)$ (if the second observation = Blue).

Similarly, we can calculate the reachable beliefs for the other action sequences. We end up with:

- Initial belief: $(1, 0, 0, 0)$.
- Possible beliefs after 1 action:

$$(1, 0, 0, 0) \text{ left}$$
$$(0, 0.5, 0.5, 0) \text{ right}$$

- Possible beliefs after 2 actions:

$$(1, 0, 0, 0), \text{ left, left}$$
$$(0, 0.5, 0.5, 0), \text{ left, right}$$
$$(0, 0, 1, 0), \text{ right, right (Green)}$$
$$(0, 0, 1/3, 2/3), \text{ right, right (Blue)}$$
$$(0, 1, 0, 0), \text{ right, left (Blue)}$$
$$(2/3, 1/3, 0, 0), \text{ right, left (Green)}$$

2: The expected immediate reward is calculated according to

$$R(b, a) = \sum_{s \in \mathcal{S}} R(s, a) b(s).$$

Giving us:

$$(1, 0, 0, 0), \text{ Left: } 0, \text{ Right:} 2$$
$$(0, 0.5, 0.5, 0), \text{ Left: } -0.5, \text{ Right: } -0.5$$
$$(0, 0, 1, 0), \text{ Left: } -1, \text{ Right: } 0$$
$$(0, 0, 1/3, 2/3), \text{ Left: } 1, \text{ Right: } 0$$
$$(0, 1, 0, 0), \text{ Left: } 0, \text{ Right: } -1$$
$$(2/3, 1/3, 0, 0), \text{ Left: } 0, \text{ Right: } 1$$

3: We calculate V^τ as:

$$V^\tau(b) = \max_a [R(b, a) + \sum_o P(o|b, a) V^{\tau-1}(b_a^o)]$$

where b_a^o is the updated belief after starting from b, taking action a and getting observation o.

We have $V^{\tau=0}(b) = 0$ for all possible beliefs, since we only get a reward by taking an action, and if $\tau = 0$ we can not take any more actions. When $\tau = 1$ we therefore have: $V^{\tau=1}(b) = \max_a R(b, a)$. At this point we have taken one action so we can be in belief state $(1, 0, 0, 0)$ or $(0, 0.5, 0.5, 0)$.

$$V^{\tau=1}((1, 0, 0, 0)) = 2$$
$$V^{\tau=1}((0, 0.5, 0.5, 0)) = -0.5$$

Then $V^{\tau=2}((1, 0, 0, 0)) = \max_a [R(b, a) + \sum_o P(o|b, a) V^{\tau-1}(b_a^o)]$.

To figure out which action maximizes this term, we can first calculate $Q^{\tau=2}((1, 0, 0, 0), \text{left})$ and $Q^{\tau=2}((1, 0, 0, 0), \text{right})$.

For action left:

$$\begin{aligned} Q^{\tau=2}((1, 0, 0, 0), \text{left}) &= R((1, 0, 0, 0), \text{left}) + \sum_o P(o|(1, 0, 0, 0), \text{left}) V^{\tau-1}(b_a^o) \\ &= 0 + P(\text{Green}|(1, 0, 0, 0), \text{left}) V^{\tau-1}(b_{\text{left}}^{\text{Green}}) + P(\text{Blue}|(1, 0, 0, 0), \text{left}) V^{\tau-1}(b_{\text{left}}^{\text{Blue}}) \\ &= 0 + 1 V^{\tau=1}((1, 0, 0, 0)) + 0 \\ &= 2 \end{aligned}$$

And for action right:

$$\begin{aligned} Q^{\tau=2}((1, 0, 0, 0), \text{right}) &= R((1, 0, 0, 0), \text{right}) + \sum_o P(o|(1, 0, 0, 0), \text{right}) V^{\tau-1}(b_a^o) \\ &= 2 + P(\text{Green}|(1, 0, 0, 0), \text{right}) V^{\tau-1}(b_{\text{right}}^{\text{Green}}) + P(\text{Blue}|(1, 0, 0, 0), \text{right}) V^{\tau-1}(b_{\text{right}}^{\text{Blue}}) \\ &= 2 + 0.5 V^{\tau=1}((0, 0.5, 0.5, 0)) + 0.5 V^{\tau=1}((0, 0.5, 0.5, 0)) \\ &= 2 + 0.5 \times -0.5 + 0.5 \times -0.5 \\ &= 1.5 \end{aligned}$$

So we get $V^{\tau=2}((1, 0, 0, 0)) = \max_a [R(b, a) + \sum_o P(o|b, a) V^{\tau-1}(b_a^o)] = 2$