Exercise Sheet - Multiagent decision making

In the decentralized tiger problem [1], two agents are standing in a hallway with two doors. Behind one door, there is a treasure and behind the other there is a tiger, as illustrated in Figure 1. The state describes which door the tiger is behind: left s_{left} or right s_{right}, each occurring with 0.5 probability (i.e., the initial state distribution is uniform $b_0 = (0.5, 0.5)$). Each agent can perform three actions: open the left door a_{left}, open the right door a_{right} or listen a_{list}. If either of them opens the door behind which the tiger is present, they are both attacked (equally) by the tiger receiving a penalty of -100. However, the injury sustained if they open the door to the tiger is less severe if they open that door jointly and in this case they get a reward of -50. Similarly, they receive wealth which they share equally when they open the door to the treasures in proportion to the number of agents that opened that door. Each listen action, however, also has a minor cost -1, resulting in a joint reward of -2 when both the agents decide to listen. The reward function is defined by Table 1.

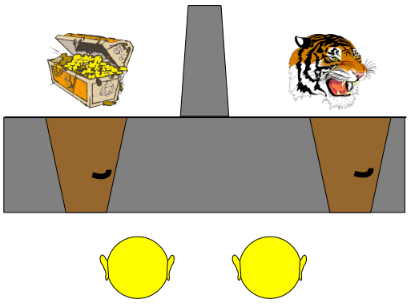

Figure 1:

	state	
joint action	s_{left}	s_{right}
$(a_{\text{list}}, a_{\text{list}})$	-2	-2
$(a_{\text{left}}, a_{\text{left}})$	-50	$+20$
$(a_{\text{right}}, a_{\text{right}})$	$+20$	-50
$(a_{\text{list}}, a_{\text{left}})$	-101	$+9$
$(a_{\text{left}}, a_{\text{list}})$	-101	$+9$
$(a_{\text{list}}, a_{\text{right}})$	$+9$	-101
$(a_{\text{right}}, a_{\text{list}})$	$+9$	-101
$(a_{\text{right}}, a_{\text{left}})$	-100	-100
$(a_{\text{left}}, a_{\text{right}})$	-100	-100

Table 1: Reward.

At every stage each agent gets an observation: they can either hear the tiger behind the left o_{left} or right door o_{right}, but each agent has a 0.85 probability of hearing it correctly (getting the right observation). Therefore, the probability that both agents get the correct observations when they listen is $0.85 \times 0.85 = 0.7225$. Moreover, the observation is informative only if both agents listen; if either agent opens a door, both agents receive an uninformative (uniformly drawn) observation and the problem resets to s_{left} or s_{right} with equal probability. At this point the problem just continues, such that the agents may be able to open the door to the treasure multiple times. The transition model is listed in Table 2.

		next state	
joint action	**state**	s_{left}	s_{right}
$(a_{\text{list}}, a_{\text{list}})$	s_{left}	1	0
$(a_{\text{list}}, a_{\text{list}})$	s_{right}	0	1
otherwise		0.5	0.5

Table 2: Transition probabilities.

1. Complete the observation probability model in Table 3.

		observations			
joint action	**state**	$(o_{\text{left}}, o_{\text{left}})$	$(o_{\text{left}}, o_{\text{right}})$	$(o_{\text{right}}, o_{\text{left}})$	$(o_{\text{right}}, o_{\text{right}})$
$(a_{\text{list}}, a_{\text{list}})$	s_{left}				
$(a_{\text{list}}, a_{\text{list}})$	s_{right}				
otherwise	$s_{\text{left}}, s_{\text{right}}$				

Table 3: Observations

2. Set the problem horizon to $h = 2$. Consider the deterministic policy π_A such that the agent listens at the first step and at the second step it opens the door according to the observation received ($o_{\text{left}} \rightarrow a_{\text{right}}$ and $o_{\text{right}} \rightarrow a_{\text{left}}$). Represent this policy as a tree and compute the value of the joint policy $\pi = (\pi_A, \pi_A)$ using the recurring formulation

$$V^{\pi}(s_t, \bar{o}_t) = \begin{cases} R(s_t, \pi(\bar{o}_t)) & \text{if t=h-1} \\ R(s_t, \pi(\bar{o}_t)) + \sum_{s_{t+1}} \sum_{o_{t+1}} \Pr(s_{t+1}, o_{t+1} | s_t, \pi(\bar{o}_t)) V^{\pi}(s_{t+1}, \bar{o}_{t+1}) & \text{otherwise} \end{cases}$$

where o_t is the joint observations at time t, $\bar{o}_t = (o_0, o_1, ..., o_t)$ is the joint observations history and o_0 is the empty joint observation. To compute the value then we use the initial belief according to

$$V(\pi) = \sum_{s_0} b_0(s_0) V^{\pi}(s_0, \bar{o}_0).$$

3. Set the horizon $h = 3$. Now consider two different policies, π_B and π_C. π_B is the extension of policy π_A defined in question 2, in which the agent always listens at the third step. Whereas we define the policy π_C by the policy tree in Figure 2. The value achieved by the joint policy (π_C, π_C) is 6.8.
Which joint policy achieves a higher value between (π_B, π_B), (π_C, π_C)?

4. How many joint policies do you need to evaluate for brute force search for horizon $h = 3$? (hint: count first the number of nodes in a policy tree, then the number of policy trees and finally the number of joint policies).

Solution:

1. The observation model is given by Table 4.
2. The policy π_A is represented by the policy tree in Figure 3.
   The value $V(\pi)$ of the joint policy $\pi = \pi_A, \pi_A$ is computed as

$$V(\pi) = 0.5\, V^{\pi}(s_{\text{left}}, o_0) + 0.5\, V^{\pi}(s_{\text{right}}, o_0).$$

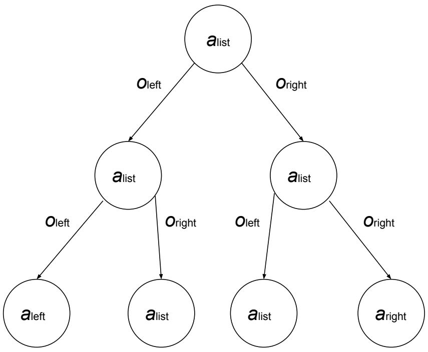

Figure 2: Representation of the policy π_C.

		observations			
joint action	**state**	$(o_{\text{left}}, o_{\text{left}})$	$(o_{\text{left}}, o_{\text{right}})$	$(o_{\text{right}}, o_{\text{left}})$	$(o_{\text{right}}, o_{\text{right}})$
$(a_{\text{list}}, a_{\text{list}})$	s_{left}	0.7225	0.1275	0.1275	0.0225
$(a_{\text{list}}, a_{\text{list}})$	s_{right}	0.0225	0.1275	0.1275	0.7225
otherwise	$s_{\text{left}}, s_{\text{right}}$	0.25	0.25	0.25	0.25

Table 4: Observation probabilities

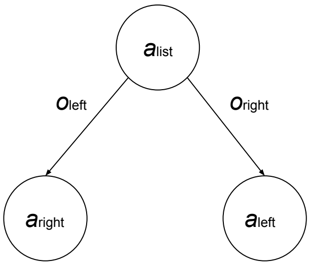

Figure 3: Representation of the policy π_A.

Since the problem and the policies are symmetric, $V^\pi(s_{\text{left}}, o_0) = V^\pi(s_{\text{right}}, o_0)$, thus it suffices to compute $V^\pi(s_{\text{left}}, o_0)$.

$$
\begin{aligned}
V^\pi(s_{\text{left}}, o_0) =& R(s_{\text{left}}, \pi(\bar{o}_0)) + \sum_{s_1}\sum_{o_1} \text{Pr}(s_1, o_1 | s_0, \pi(\bar{o}_0)) R(s_1, \pi(\bar{o}_1)) \\
=& R(s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) + \sum_{s_1}\sum_{o_1} \text{Pr}(s_1, o_1 | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_1, \pi(\bar{o}_1)) \\
=& -2 + \sum_{o_1} \text{Pr}(s_{\text{left}}, o_1 | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_{\text{left}}, \pi(\bar{o}_1)) \\
=& -2 + \text{Pr}(s_{\text{left}}, (o_{\text{left}}, o_{\text{left}}) | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_{\text{left}}, \pi((o_{\text{left}}, o_{\text{left}}))) \\
&+ \text{Pr}(s_{\text{left}}, (o_{\text{left}}, o_{\text{right}}) | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_{\text{left}}, \pi((o_{\text{left}}, o_{\text{right}}))) \\
&+ \text{Pr}(s_{\text{left}}, (o_{\text{right}}, o_{\text{left}}) | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_{\text{left}}, \pi((o_{\text{right}}, o_{\text{left}}))) \\
&+ \text{Pr}(s_{\text{left}}, (o_{\text{right}}, o_{\text{right}}) | s_{\text{left}}, (a_{\text{list}}, a_{\text{list}})) R(s_{\text{left}}, \pi((o_{\text{right}}, o_{\text{right}}))) \\
=& -2 + 0.7225 \times 20 + 0.1275 \times (-100) + 0.1275 \times (-100) + 0.0225 \times (-50) \\
=& -14.175
\end{aligned}
$$

3. The value of the joint policy (π_B, π_B) is obtained by adding the last step reward -2 to the value of 2 horizon policy (π_A, π_A), that is $V((\pi_B, \pi_B)) = -16.175$. Therefore the policy (π_C, π_C) performs better (and it is actually the optimal policy).
4. Consider a general Multiagent problem with n agents where $|O|$ is the cardinality of the one agent observations space and $|A|$ the number of one agent actions. We have that a policy tree for horizon h has a $|O|^t$ nodes at

level t. Therefore the total number of nodes is given by

$$|O|^0 + |O|^1 + \dots |O|^{h-1} = \sum_{t=0}^{h-1} |O|^t = \frac{|O|^h - 1}{|O| - 1}.$$

For each nodes in the policy tree there are $|A|$ possible actions. Thus the number of policies for one agent are $|A|^{\frac{|O|^h-1}{|O|-1}}$. Considering all the possible combinations of these policies for the n agents we will then obtain that the number of joint policies is

$$|A|^{\frac{|O|^h-1}{|O|-1}} \times |A|^{\frac{|O|^h-1}{|O|-1}} \times \dots \times |A|^{\frac{|O|^h-1}{|O|-1}} = \Pi_{i=1}^{n} |A|^{\frac{|O|^h-1}{|O|-1}} = |A|^{n\frac{|O|^h-1}{|O|-1}}.$$

So in our case, this would correspond to 4.78×10^6 joint policies.

▲

References

[1] Nair, Ranjit and Tambe, Milind and Yokoo, Makoto and Pynadath, David and Marsella, Stacy. *Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings*, IJCAI, 2003.